SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 1–28

POINTWISE ERROR ESTIMATES FOR FINITE ELEMENT SOLUTIONS OF THE STOKES PROBLEM*

HONGSEN CHEN†

Abstract. In this paper pointwise error estimates for general finite element approximations of the Stokes problem are established on quasi-uniform grids in R^N. The results obtained in this paper improve and extend the existing error estimates in the maximum norm for the Stokes problem. The new pointwise error estimates exhibit a more local dependence of the errors on the true solution and as a by-product provide logarithm-free bounds for all errors except the error of the velocity approximation of the lowest order.

Key words. finite element method, pointwise error estimate, Stokes problem

AMS subject classifications. Primary, 65N30, 65N15, 65N12, 76D07; Secondary, 41A25, 35B45, 35J20

DOI. 10.1137/S0036142903438100

1. Introduction.

This paper is devoted to new pointwise error estimates of finite element approximations of the Stokes equations on general quasi-uniform meshes in R^N. The results in this paper represent an improvement on and extension of the existing maximum norm error estimates found in Durán, Nochetto, and Wang [6], which were obtained for two-dimensional problems. Our analysis is based on the technique developed recently by Schatz [15, 16] for the finite element method for second order elliptic problems (see also Schatz and Wahlbin [17]). In contrast to the traditional approach for proving error estimates in the maximum norm with the weighted function method (Scott [19], Natterer [10], Rannacher and Scott [14], etc.), the new method relies on the availability of local error estimate in energy norm for the underlying finite element discretization. The results in [15] indicate a more localized dependence of the errors on the derivatives of the true solution. As a consequence of these estimates error expansion inequalities have been derived and applied to superconvergence and extrapolation and a posteriori estimates (see [16, 8]). The aim of this paper is to extend the new technique from elliptic problems to the Stokes problem, which has a saddle point nature and requires a more careful investigation.

It is well known that the conforming finite element approximation $(\mathbf{u}_h, p_h)$ to the true solution $(\mathbf{u}, p)$ of the Stokes problem admits the following optimal error estimates in energy and L^2 norms (Brezzi and Fortin [4], Girault and Raviart [7]):

$$\|\mathbf{u}-\mathbf{u}_h\|_{H^1(\Omega)} + \|p-p_h\|_{L^2(\Omega)} \le C\Big(\inf_{\mathbf{v}\in\mathbf{V}_h} \|\mathbf{u}-\mathbf{v}\|_{H^1(\Omega)} + \inf_{q\in W_h} \|p-q\|_{L^2(\Omega)}\Big)$$

and

$$\|\mathbf{u}-\mathbf{u}_h\|_{L^2(\Omega)} \le Ch\Big(\inf_{\mathbf{v}\in\mathbf{V}_h} \|\mathbf{u}-\mathbf{v}\|_{H^1(\Omega)} + \inf_{q\in W_h} \|p-q\|_{L^2(\Omega)}\Big),$$

where $\mathbf{V}_h$ is the finite element subspace for velocity unknown and W_h is the finite element subspace for pressure unknown. The spaces $\mathbf{V}_h$ and W_h are assumed to satisfy

*Received by the editors November 23, 2003; accepted for publication (in revised form) April 11, 2005; published electronically February 8, 2006.
http://www.siam.org/journals/sinum/44-1/43810.html

†Institute for Scientific Computation, Texas A&M University, College Station, TX 77840 (hchen@isc.tamu.edu).

a certain condition (e.g., the inf-sup condition) to ensure the solvability and uniqueness of the solution of the resulting finite element system. Under the assumption that there exists a locally constructed projection operator $\Pi^h : \ H_0^1(\Omega)^N \to \mathbf{V}^h$ satisfying

$$(\nabla \cdot (\mathbf{v} - \Pi^h \mathbf{v}), q_h) = 0 \quad \forall \mathbf{v} \in H_0^1(\Omega)^N, \ q_h \in W_h,$$

the following error estimates in the maximum norm have been derived for the two-dimensional Stokes problem (Durán, Nochetto, and Wang [6]):

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^\infty(\Omega)} \le Ch \ln \frac{1}{h} B(\mathbf{u}, p), \tag{1.1}$$

$$\|p - p_h\|_{L^\infty(\Omega)} \le C \left(\ln \frac{1}{h} \right)^{1/2} B(\mathbf{u}, p), \tag{1.2}$$

where

$$B(\mathbf{u}, p) = \inf_{\mathbf{v} \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}\|_{W^{1,\infty}(\Omega)} + \inf_{q \in W_h} \|p - q\|_{L^\infty(\Omega)}.$$

Similar error estimates were derived by using weighted inf-sup conditions instead of the local projection operator in Durán and Nochetto [5]. The proofs of these error estimates are based on the estimates in weighted Sobolev norms for some so-called regularized Green's functions and their finite element approximations. This is the standard approach for proving error estimates in the maximum norm for finite element methods. Nevertheless, with this approach it is difficult to generalize these results to the case of higher space dimensions.

Following the approach developed in [15], some new pointwise error estimates for the finite element solutions of the Stokes problem will be derived in this paper. Our new pointwise error estimates for both the velocity and pressure approximations take the following form: for any $z \in \bar{\Omega}$,

$$|\nabla(\mathbf{u} - \mathbf{u}_h)(z)| + |(p - p_h)(z)| \le C \left(\ln \frac{1}{h} \right)^{\bar{s}} B(\mathbf{u}, p, z, s), \quad 0 \le s \le r, \tag{1.3}$$

$$|(\mathbf{u} - \mathbf{u}_h)(z)| \le Ch \left(\ln \frac{1}{h} \right)^{\bar{\bar{s}}} B(\mathbf{u}, p, z, s), \quad 0 \le s \le r - 1, \tag{1.4}$$

where

$$B(\mathbf{u}, p, z, s) = \inf_{\mathbf{v} \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \inf_{q \in W^h} \|p - q\|_{L^\infty(\Omega),z,s},$$

and $\| \cdot \|_{W^{1,\infty}(\Omega),z,s}$ and $\| \cdot \|_{L^\infty(\Omega),z,s}$ are weighted Sobolev norms with the weight function $\sigma_z(x)^s = (h/(|x - z| + h)^s$, and $\bar{s} = 0$ if $0 \le s < r$ and $\bar{s} = 1$ if $s = r$; $\bar{\bar{s}} = 0$ if $0 \le s < r - 1$ and $\bar{\bar{s}} = 1$ if $s = r - 1$. Here, r is the order of approximation of the finite element spaces in the H^1 norm for the velocity and the L^2 norm for the pressure (in most cases, r equals the degree of polynomials used in space $\mathbf{V}^h$).

Notice that estimates (1.4) and (1.1) coincide when $r = 1$, which corresponds to the case of the lowest order finite element spaces because the only possible value of s is zero. Nevertheless, (1.4) indeed improve (1.1) for $r > 1$. Estimates (1.4) and (1.3) are sharper than (1.1) and (1.2) in the sense that (1.4) and (1.3) imply (1.1) and (1.2) but not vice versa. Because of the weighted norms on the bounds of (1.4) and (1.3), they indicate a more local dependence of the errors at z on the true solution

$(\mathbf{u}, p)$ near z than (1.1) and (1.2) do. The larger the r, the higher local properties these estimates provide. Besides the more localized dependence on the true solution, estimates (1.4) and (1.3) also represent sharper bounds for the errors. They provide logarithm-free error bounds for the velocity approximation when $r > 1$ and for the approximations and the derivatives of the velocity and the pressure for any $r \geq 1$. In fact, (1.4) and (1.3) imply the following estimates in the maximum norm:

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^\infty(\Omega)} \leq Ch\left(\ln\frac{1}{h}\right)^{\bar{r}} B(\mathbf{u}, p),$$
$$\|\mathbf{u} - \mathbf{u}_h\|_{W^{1,\infty}(\Omega)} + \|p - p_h\|_{L^\infty(\Omega)} \leq CB(\mathbf{u}, p).$$

Here, $\bar{r} = 0$ if $r > 1$ and $\bar{r} = 1$ if $r = 1$.

The new estimates (1.4) and (1.3) are stated in Theorems 4.2, 5.1, and 6.1 and are proved under abstract assumptions on the finite element spaces. These hypotheses include the quasi-uniformity of the partitions, the approximation properties, the inverse properties, local L^2 error estimates for the Stokes finite element solutions, and the scaling property. Note that the superapproximation property and the inf-sup (stability) condition or the existence of locally orthogonal projection operator are not assumed explicitly. Instead, local error estimates for the velocity in the H^1 norm and the pressure in the L^2 norm are assumed, which can be proved under the assumptions of superapproximation and the presence of local stability or local orthogonal projection operator (see Arnold and Liu [3]).

We end this introduction with a brief outline of the rest of this paper. In section 2, we discuss some preliminaries and introduce notation and assumptions. Section 3 contains estimates of a priori type for the Stokes problem and error estimates for some special auxiliary problems. In section 4, we prove the pointwise error estimate for the pressure approximation. The pointwise error estimates for the velocity and the derivatives of the velocity are proved in sections 5 and 6, respectively.

2. Preliminaries and notation. For the sake of simplicity, we consider the following Stokes problem with a homogeneous Dirichlet boundary condition:

$$\begin{aligned} -\nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\Omega \subset R^N$ $(N \geq 2)$ is an open subset with a smooth boundary $\partial\Omega$, $\mathbf{u}$ and p are unknown functions called velocity and pressure of the fluid, respectively, $\mathbf{f}$ is a given function, and $\nu > 0$ is a constant representing the viscosity of the fluid. Without loss of generality, we assume that $\nu = 1$.

We shall use the standard notation for Sobolev spaces and their norms (see, e.g., Adams [1]). For nonnegative integer j and real number $1 \leq t \leq \infty$ and subdomain $D \subset \Omega$, denote the Sobolev spaces by $W^{j,t}(D) = \left\{v : \ \|v\|_{W^{j,t}(D)} < \infty\right\}$ with

$$\|v\|_{W^{j,t}(D)} = \left(\sum_{i=0}^{j} |v|^t_{W^{i,t}(D)}\right)^{1/t}, \quad |v|_{W^{i,t}(D)} = \left(\sum_{|\alpha|=i}\int_D |\partial^\alpha v(x)|^t dx\right)^{1/t}.$$

Let $W_0^{j,t}(D)$ denote the completion of $C_0^\infty(D)$ according to the norm $\|\cdot\|_{W^{j,t}(D)}$, where $C_0^\infty(D)$ represents the space of functions with continuous derivatives of arbitrary order and compact supports in D. We also adopt the usual notation

$$H^j(D) = W^{j,2}(D), \quad H_0^j(D) = W_0^{j,2}(D), \quad L^t(D) = W^{0,t}(D).$$

Denote by $(\cdot,\cdot)_D$ the standard inner product in $L^2(D)$ given by $(u,v)_D = \int_D uv\,dx$. When $D=\Omega$, we write $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$. The subspace of $L^2(D)$ consisting of functions of zero mean values is denoted by $L_0^2(D)$, i.e.,

$$L_0^2(D) = \left\{ q \in L^2(D) : \int_D q(x)\,dx = 0 \right\}.$$

For $j \geq 0$ and $1 \leq t < \infty$, the negative Sobolev norm $\|\cdot\|_{W^{-j,t}(D)}$ is defined as follows:

$$\|v\|_{W^{-j,t}(D)} = \sup_{\varphi \in C_0^\infty(D)} \frac{(v,\varphi)}{\|\varphi\|_{W^{j,t'}(D)}},$$

where t' is the conjugate of t, i.e., $1/t + 1/t' = 1$. We will make no distinction in notation between the Sobolev norms for scalar and vector valued functions. For instance, the norm of $\mathbf{u}$ in

$$H^1(D)^N = \overbrace{H^1(D) \times \cdots \times H^1(D)}^{N}$$

is denoted by $\|\mathbf{u}\|_{H^1(D)}$.

The solution $(\mathbf{u}, p)$ of (2.1) satisfies the following weak formulation:

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) - b(\mathbf{v},p) &= (\mathbf{f},\mathbf{v}) \quad \forall\ \mathbf{v} \in H_0^1(\Omega)^N, \\ b(\mathbf{u},q) &= 0 \qquad \forall\ q \in L_0^2(\Omega), \end{aligned} \tag{2.2}$$

where $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bilinear forms defined by

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\,dx, \quad b(\mathbf{v},q) = \int_\Omega \nabla\cdot\mathbf{v}\,q\,dx.$$

To define the finite element solution of (2.2), let $\mathbf{V}^h \subset H_0^1(\Omega)^N$ and $W^h \subset L_0^2(\Omega)$ be two families of finite-dimensional subspaces with a parameter $h \in (0,1)$. The finite element approximation $(\mathbf{u}_h, p_h) \in \mathbf{V}^h \times W^h$ is defined to be the solution of

$$\begin{aligned} a(\mathbf{u}_h,\mathbf{v}) - b(\mathbf{v},p_h) &= (\mathbf{f},\mathbf{v}) \quad \forall\ \mathbf{v} \in \mathbf{V}^h, \\ b(\mathbf{u}_h,q) &= 0 \quad \forall\ q \in W^h. \end{aligned} \tag{2.3}$$

By (2.2) and (2.3), the following error equation holds:

$$\begin{aligned} a(\mathbf{u}-\mathbf{u}_h,\mathbf{v}) - b(\mathbf{v},p-p_h) &= 0 \quad \forall\ v \in \mathbf{V}^h, \\ b(\mathbf{u}-\mathbf{u}_h,q) &= 0 \quad \forall\ q \in W^h. \end{aligned} \tag{2.4}$$

We point out that the pointwise error estimates obtained in this paper are all based on the error equation (2.4).

For any subset $D \subset \Omega$, let $\mathbf{V}^h(D)$ denote the set of restrictions of functions in $\mathbf{V}^h$ on D, and $W^h(D)$ the set of restrictions of functions in W^h on D. We also introduce the notation for balls in R^N. Denote by $B_d(x)$ the ball of radius $d>0$ and centered at $x \in R^N$.

Throughout this paper, $\kappa > 0$ denotes a constant. Below, we list the assumptions that will be used by our results.

A.0 (*quasi-uniform partition*). The domain Ω is decomposed into N_h subdomains:

$$\bar{\Omega} = \bigcup_{j=1}^{N_h} \bar{K}_j, \text{ so that } W^h(K_j) \subset C^\infty(K_j),\ j = 1, \ldots, N_h.$$

Furthermore, there are constants $C_0 > 0$ and $\tau_0 > 0$ such that for any $\tau \in (0, \tau_0)$, and $1 \le j \le N_h$ and $x \in \bar{K}_j$, there exists an $\bar{x} \in K_j$ such that

$$B_{\tau h}(\bar{x}) \subset K_j \quad \text{and} \quad |x - \bar{x}| \le C_0 \tau h.$$

A.1 (*approximation properties*). There exist $r \ge 1$ and two linear operators

$$\Pi^h :\ H^1(\Omega)^N \to \mathbf{V}^h \quad \text{and} \quad Q^h :\ L^2(\Omega) \to W^h$$

such that for any $D_1 \subset D_2 \subset \Omega$ with $\operatorname{dist}(D_1, \partial D_2 \backslash \partial\Omega) \ge \kappa h$,

(i) for $0 \le i \le 1 \le j \le r+1$, $2 \le t \le \infty$, and any $\mathbf{v} \in W^{j,t}(D_2)^N$,

$$\|\mathbf{v} - \Pi^h \mathbf{v}\|_{W^{i,t}(D_1)} \le C h^{j-i} |\mathbf{v}|_{W^{j,t}(D_2)};$$

if $N < t \le \infty$,

$$\|\mathbf{v} - \Pi^h \mathbf{v}\|_{W^{1,\infty}(D_1)} \le C h^{r-N/t} \|\mathbf{v}\|_{W^{1+r,t}(D_2)};$$

(ii) for $0 \le i \le j \le r$, $2 \le t \le \infty$, and any $q \in W^{j,t}(D_2)$,

$$\|q - Q^h q\|_{W^{i,t}(D_1)} \le C h^{j-i} |q|_{W^{j,t}(D_2)};$$

if $N < t \le \infty$,

$$\|q - Q^h q\|_{L^\infty(D_1)} \le C h^{r-N/t} \|q\|_{W^{r,t}(D_2)}.$$

A.2 (*inverse properties*). Let $D_1 \subset D_2 \subset \Omega$ with $\operatorname{dist}(D_1, \partial D_2 \backslash \partial\Omega) \ge \kappa h$. Then, for any $1 \le s \le t \le \infty$, $i = 0, 1$, $j \ge 0$ and $\mathbf{v} \in \mathbf{V}^h$, $q \in W^h$,

$$\begin{aligned}
\|\mathbf{v}\|_{W^{i,t}(D_1)} &\le C h^{-N(1/s-1/t)-i-j} \|\mathbf{v}\|_{W^{-j,s}(D_2)}, \\
\|q\|_{L^t(D_1)} &\le C h^{-N(1/s-1/t)-j} \|q\|_{W^{-j,s}(D_2)}, \\
\|q\|_{W^{1,\infty}(K_j)} &\le C h^{-1} \|q\|_{L^\infty(K_j)}, \quad 1 \le j \le N_h.
\end{aligned}$$

A.3 (*local L^2 error estimate*). Let $(\mathbf{v}, q) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ and $(\mathbf{v}_h, q_h) \in \mathbf{V}^h \times W^h$ satisfy

$$\begin{aligned}
a(\mathbf{v} - \mathbf{v}_h, \varphi) - b(\varphi, q - q_h) &= 0 \quad \forall\ \varphi \in \mathbf{V}^h, \\
b(\mathbf{v} - \mathbf{v}_h, \psi) &= 0 \quad \forall\ \psi \in W^h.
\end{aligned}$$

If $D_1 \subset D_2 \subset \Omega$ with $d = \operatorname{dist}(D_1, \partial D_2 \backslash \partial\Omega) > 0$, then there holds

$$\begin{aligned}
&\|\mathbf{v} - \mathbf{v}_h\|_{H^1(D_1)} + \|q - q_h\|_{L^2(D_1)} \\
&\quad \le C h^r \left(\|\mathbf{v}\|_{H^{r+1}(D_2)} + \|q\|_{H^r(D_2)} \right) \\
&\qquad + C \left(\|\mathbf{v} - \mathbf{v}_h\|_{W^{-t,1}(D_2)} + \|q - q_h\|_{W^{-t-1,1}(D_2)} \right),
\end{aligned}$$

where $t \ge 0$, and the constant $C > 0$ may be dependent on t and d but is independent of h and $\mathbf{v}$ and q.

A.4 (*scaling property*). Let $x_0 \in \bar{\Omega}$ and $d \geq \kappa h$. The linear transformation $y = (x - x_0)/d$ takes the domain $B_d(x_0) \cap \Omega$ into a new domain $\hat{B}_1(x_0)$, $\mathbf{V}^h(B_d(x_0))$ into a new function space $\hat{\mathbf{V}}^{h/d}(\hat{B}_1(x_0))$, and $W^h(B_d(x_0))$ into a new space $\hat{W}^{h/d}(\hat{B}_1(x_0))$. Then, $\hat{\mathbf{V}}^{h/d}(\hat{B}_1(x_0))$ and $\hat{W}^{h/d}(\hat{B}_1(x_0))$ satisfy A.1 and A.2 with h replaced by h/d. The constants occurring in A.1 and A.2 remain unchanged, in particular, independent of d.

Assumptions A.1, A.2, and A.4 are very standard (see [15, 17, 18]). The results in A.3 can be found in Arnold and Liu [3] under additional assumptions such as the availability of the local inf-sup condition or the orthogonal projection operator. Although the results in [3] were obtained for two-dimensional problems, it is easily seen that they can be extended to higher dimensions. Assumption A.0 is used only for the existence of the *regularized* Dirac delta function in the derivation of the pointwise error estimate of the pressure approximation (see section 4 for details).

Throughout this paper we assume that Assumptions A.0, A.1, A.2, A.3, and A.4 are satisfied and use the letter C for generic constants.

Before we end this section, let us introduce some notation about the weighted norms: Following the notation of [15], for a fixed $z \in \bar{\Omega}$, a real number s, and arbitrary $x \in R^N$, define the weight function

$$\sigma_{z,h}^s(x) = \left(\frac{h}{|x - z| + h}\right)^s. \tag{2.5}$$

Clearly, $\sigma_{z,h}^s(x) = \mathcal{O}(1)$ if $s > 0$ and $|x - z| = \mathcal{O}(h)$, $\sigma_{z,h}^s(x) = \mathcal{O}(h^s)$ if $|x - z| = \mathcal{O}(1)$. For $1 \leq t \leq \infty$ and fixed z, we define the following weighted norms:

$$\|\varphi\|_{L^t(\Omega),z,s} = \|\sigma_{z,h}^s \varphi\|_{L^t(\Omega)}, \tag{2.6}$$

$$\|\varphi\|_{W^{1,t}(\Omega),z,s} = \|\varphi\|_{L^t(\Omega),z,s} + \|\nabla\varphi\|_{L^t(\Omega),z,s}. \tag{2.7}$$

It is straightforward to extend this notation to vector valued functions.

3. A priori estimates. In this section, we shall collect and show some results on the regularities of the Stokes problem. These results include the global and local estimates of the higher order derivatives of the true solutions and the global and local error estimates for the finite element approximations of some special auxiliary problems.

The first result is the following global a priori estimates for the Stokes equations.

LEMMA 3.1. *Suppose Ω and $(\mathbf{v}, \mu) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ are sufficiently smooth. Then, for $k \geq -1$, there holds*

$$\|\mathbf{v}\|_{H^{k+2}(\Omega)} + \|\mu\|_{H^{k+1}(\Omega)} \leq C\left(\|\mathbf{f}\|_{H^k(\Omega)} + \|g\|_{H^{k+1}(\Omega)}\right), \tag{3.1}$$

where $\mathbf{f} = -\Delta\mathbf{v} + \nabla\mu$, $g = \nabla \cdot \mathbf{v}$, and $C > 0$ is a constant independent of $\mathbf{v}$, μ.

Proof. The proof can be found in Ladyzhenskaya [9] and Agmon, Douglis, and Nirenberg [2]. □

Next, we show a special local estimate of the solution of the Stokes problem, which is based on the pointwise bounds for Green's tensors of the Stokes problem.

LEMMA 3.2. *Suppose $D_0 \subset D_1 \subset \Omega$ are sufficiently smooth domains with $d = \operatorname{dist}(D_0, \partial D_1 \backslash \partial\Omega) > 0$, and $\mathbf{f} \in L^\infty(\Omega)$, $g \in C_0^\infty(\Omega)$, $\zeta \in C_0^\infty(\Omega)$ with $(1, \zeta) = 1$. Let $(\mathbf{v}, \mu) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ be the solution of*

$$-\Delta\mathbf{v} + \nabla\mu = \mathbf{f}, \quad \nabla \cdot \mathbf{v} = g - (1, g)\zeta.$$

If $\mathbf{f}$ *and* g *vanish in* D_1*, then for* $k \geq -1$ *and* $1 \leq t \leq \infty$*, there holds*

$$\begin{aligned}(3.2)\qquad &\|\mathbf{v}\|_{W^{k+2,\infty}(D_0)} + \|\mu\|_{W^{k+1,\infty}(D_0)} \\ &\quad \leq Cd^{-N/t-k}\left(\|\mathbf{f}\|_{L^t(\Omega)} + d^{-1}\|g\|_{L^t(\Omega)}\right) + C\|g\|_{L^1(\Omega)}\|\zeta\|_{W^{k+2,\infty}(\Omega)}.\end{aligned}$$

Proof. For any $x \in D_0$, let $(\mathbf{G}_x^j, Q_x^j)$, $j = 1, \ldots, N$, and $(\mathbf{G}_x^{N+1,\beta}, Q_x^{N+1,\beta})$ be Green's tensor of the Stokes problem (2.1) (see Solonnikov [20], Ladyzhenskaya [9], Stupelis [21], Odqvist [13]) so that $\mathbf{v}$ and μ can be expressed in terms of the components of Green's tensor through the following integrals:

$$(3.3)\qquad \mathbf{v}(x) = \sum_{j=1}^{N}\int_\Omega \mathbf{G}_x^j(y)f_j(y)\,dy + \sum_{|\beta|\leq k+2}\int_\Omega \mathbf{G}_x^{N+1,\beta}(y)\partial_y^\beta \tilde{g}(y)\,dy,$$

$$(3.4)\qquad \mu(x) = \sum_{j=1}^{N}\int_\Omega Q_x^j(y)f_j(y)\,dy + \sum_{|\beta|\leq k+2}\int_\Omega Q_x^{N+1,\beta}(y)\partial_y^\beta \tilde{g}(y)\,dy.$$

Here, $\tilde{g} = g - (1, g)\zeta$, f_j are the components of $\mathbf{f}$, and $\mathbf{G}_x^j$, Q_x^j, $\mathbf{G}_x^{N+1,\beta}$, and $Q_x^{N+1,\beta}$ satisfy the inequalities

$$(3.5)\qquad \begin{aligned}&|\partial_x^{\alpha_x}\partial_y^{\alpha_y}\mathbf{G}_x^j(y)| \leq C\varphi_0(|x-y|; 2-N-|\alpha_x|-|\alpha_y|),\\ &|\partial_x^{\alpha_x}\partial_y^{\alpha_y}Q_x^j(y)| \leq C\varphi_0(|x-y|; 1-N-|\alpha_x|-|\alpha_y|),\\ &|\partial_x^{\alpha_x}\partial_y^{\alpha_y}\mathbf{G}_x^{N+1,\beta}(y)| \leq C\varphi_0(|x-y|; 2-N-|\alpha|+k+1),\\ &|\partial_x^{\alpha}\partial_y^{\beta}Q_x^{N+1,\beta}(y)| \leq C\varphi_0(|x-y|; 1-N-|\alpha|+k+1),\end{aligned}$$

where $j = 1, \ldots, N$, $|\beta| \leq k+2$, and

$$\varphi_0(|x-y|;\tau) = \begin{cases}|x-y|^\tau & \text{if} \quad \tau < 0,\\ 1 + |\ln|x-y|| & \text{if} \quad \tau = 0,\\ 1 & \text{if} \quad \tau > 0.\end{cases}$$

Differentiating (3.4), using the assumption that $\mathbf{f}$ and g vanish in D_1, and integrating by parts, we have

$$\begin{aligned}(3.6)\qquad \partial_x^\alpha\mu(x) = &\sum_{j=1}^{N}\int_{\Omega\setminus D_1}\partial_x^\alpha Q_x^j(y)f_j(y)\,dy + (1,g)\sum_{|\beta|\leq k+2}\int_\Omega \partial_x^\alpha Q_x^{N+1,\beta}(y)\partial^\beta\zeta(y)\,dy\\ &+ \sum_{|\beta|\leq k+2}(-1)^{|\beta|}\int_{\Omega\setminus D_1}\partial_y^\beta\partial_x^\alpha Q_x^{N+1,\beta}(y)g(y)\,dy.\end{aligned}$$

We shall estimate each summation in (3.6) separately. According to (3.5), we have for each j and $|\alpha| \geq 1$,

$$(3.7)\qquad \int_{\Omega\setminus D_1}\partial_x^\alpha Q_x^j(y)f_j(y)\,dy \leq C\int_{\Omega\setminus D_1}|x-y|^{1-N-|\alpha|}|f_j(y)|\,dy.$$

Let t' be the conjugate of t, i.e., $1/t'+1/t=1$. Then, by (3.7) and Hölder's inequality, there holds for $|\alpha|=k+1$

$$\begin{aligned}(3.8)\quad \int_{\Omega\setminus D_1}\partial_x^\alpha Q_x^j(y)f_j(y)\,dy &\le C\left(\int_{\Omega\setminus D_1}|x-y|^{(1-N-|\alpha|)t'}dy\right)^{1/t'}\|f_j\|_{L^t(\Omega)}\\ &\le C\left(\int_d^\infty \rho^{(1-N-|\alpha|)t'}\rho^{N-1}\,d\rho\right)^{1/t'}\|f_j\|_{L^t(\Omega)}\\ &\le Cd^{-N/t-k}\|f_j\|_{L^t(\Omega)}.\end{aligned}$$

Similarly, for the integrals in the second summation of (3.6), we have for $|\alpha|=k+1$

$$\begin{aligned}(3.9)\quad \int_{\Omega\setminus D_1}\partial_y^\beta\partial_x^\alpha Q_x^{N+1,\beta}(y)g(y)\,dy &\le C\left(\int_{\Omega\setminus D_1}|x-y|^{(-N-|\alpha|)t'}dy\right)^{1/t'}\|g\|_{L^t(\Omega)}\\ &\le Cd^{-N/t-k-1}\|g\|_{L^t(\Omega)}.\end{aligned}$$

For the third summation in (3.6), for $|\alpha|=k+1$ there holds

$$\begin{aligned}(3.10)\quad &(1,g)\sum_{|\beta|\le k+2}\int_\Omega \partial_x^\alpha Q_x^{N+1,\beta}(y)\partial^\beta\zeta(y)\,dy\\ &\le C\|g\|_{L^1(\Omega)}\int_\Omega|x-y|^{1-N}\,dy\|\zeta\|_{W^{k+2,\infty}(\Omega)}\le C\|g\|_{L^1(\Omega)}\|\zeta\|_{W^{k+2,\infty}(\Omega)}.\end{aligned}$$

Substituting (3.10), (3.9), and (3.8) into (3.6), we obtain

$$\begin{aligned}(3.11)\quad \|\mu\|_{W^{k+1,\infty}(D_0)} &\le Cd^{-N/t-k}\left(\|\mathbf{f}\|_{L^t(\Omega)}+d^{-1}\|g\|_{L^t(\Omega)}\right)\\ &\quad +C\|g\|_{L^1(\Omega)}\|\zeta\|_{W^{k+2,\infty}(\Omega)}.\end{aligned}$$

Reasoning in the same way, we can also bound $\|\mathbf{v}\|_{W^{k+2,\infty}(D_0)}$ by the right-hand side of (3.11). This completes the proof. ∎

We now turn to the error estimates for the finite element approximation of the Stokes problem in energy and L^2 norms. First of all, we state the following well-known global error estimate.

LEMMA 3.3. *Let* $(\mathbf{u},p)$ *and* $(\mathbf{u}_h,p_h)$ *satisfy* (2.4). *Then,*

$$(3.12)\quad \|\mathbf{u}-\mathbf{u}_h\|_{H^1(\Omega)}+\|p-p_h\|_{L^2(\Omega)}\le C\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{H^1(\Omega)}+\inf_{q\in W^h}\|p-q\|_{L^2(\Omega)}\right).$$

Proof. The proof is found in Temam [22] or Girault and Raviart [7]. ∎

The results in the next lemma are error estimates of the pressure approximation in some special negative Sobolev norms. They are used for the error estimates of the finite element approximations of the auxiliary Stokes problem constructed during the proofs of our main results in other sections.

LEMMA 3.4. *Suppose* $(\mathbf{u},p)$ *and* $(\mathbf{u}_h,p_h)$ *satisfy* (2.4). *Then,*

$$(3.13)\quad \|p-p_h\|_{H^{-r}(\Omega)}\le Ch^r\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{H^1(\Omega)}+\inf_{q\in W^h}\|p-q\|_{L^2(\Omega)}\right),$$

$$(3.14)\quad \|p-p_h\|_{H^{-r-1-[N/2]}(\Omega)}\le Ch^r\left(\|\mathbf{u}-\mathbf{u}_h\|_{W^{1,1}(\Omega)}+\|p-p_h\|_{L^1(\Omega)}\right),$$

where $[N/2]$ *denotes the maximum integer less than or equal to* $N/2$.

Proof. For any $\psi \in C_0^\infty(\Omega)$, let $(\mathbf{w}, \lambda) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ satisfy

$$-\Delta \mathbf{w} + \nabla \lambda = 0, \quad \nabla \cdot \mathbf{w} = \psi - \frac{1}{|\Omega|}(1, \psi) \quad \text{in } \Omega.$$

Then, some simple manipulations give

(3.15)
$$\begin{aligned}(p - p_h, \psi) &= \Big(p - p_h, \psi - \frac{1}{|\Omega|}(1, \psi)\Big) \;=\; b(\mathbf{w}, p - p_h) \\ &= a(\mathbf{u} - \mathbf{u}_h, \Pi^h \mathbf{w} - \mathbf{w}) - b(\Pi^h \mathbf{w} - \mathbf{w}, p - p_h) - b(\mathbf{u} - \mathbf{u}_h, Q^h \lambda - \lambda).\end{aligned}$$

By the Cauchy–Schwarz inequality and the approximation properties in A.1, we have

$$\begin{aligned}(p - p_h, \psi) &\le C \left(\|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)}\right) \\ &\quad \times \left(\|\mathbf{w} - \Pi^h \mathbf{w}\|_{H^1(\Omega)} + \|\lambda - Q^h \lambda\|_{L^2(\Omega)}\right) \\ &\le C h^r \left(\|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)}\right) \\ &\quad \times \left(\|\mathbf{w}\|_{H^{1+r}(\Omega)} + \|\lambda\|_{H^r(\Omega)}\right).\end{aligned} \tag{3.16}$$

Using (3.1) of Lemma 3.1,

$$\|\mathbf{w}\|_{H^{1+r}(\Omega)} + \|\lambda\|_{H^r(\Omega)} \le C \left\| \psi - \frac{1}{|\Omega|}(1, \psi) \right\|_{H^r(\Omega)} \le C \|\psi\|_{H^r(\Omega)}$$

and estimate (3.12) of Lemma 3.3 in inequality (3.16), we prove (3.13). To show (3.14), we use (3.15), Hölder's inequality, and the approximation properties in A.1:

$$\begin{aligned}(p - p_h, \psi) &\le C \left(\|\mathbf{u} - \mathbf{u}_h\|_{W^{1,1}(\Omega)} + \|p - p_h\|_{L^1(\Omega)}\right) \\ &\quad \times \left(\|\mathbf{w} - \Pi^h \mathbf{w}\|_{W^{1,\infty}(\Omega)} + \|\lambda - Q^h \lambda\|_{L^\infty(\Omega)}\right) \\ &\le C h^r \left(\|\mathbf{u} - \mathbf{u}_h\|_{W^{1,1}(\Omega)} + \|p - p_h\|_{L^1(\Omega)}\right) \\ &\quad \times \left(\|\mathbf{w}\|_{W^{1+r,\infty}(\Omega)} + \|\lambda\|_{W^{r,\infty}(\Omega)}\right).\end{aligned} \tag{3.17}$$

From the Sobolev embedding theorem (Adam [1]) and (3.1) of Lemma 3.1,

$$\begin{aligned}\|\mathbf{w}\|_{W^{1+r,\infty}(\Omega)} + \|\lambda\|_{W^{r,\infty}(\Omega)} &\le C \left(\|\mathbf{w}\|_{H^{2+r+[N/2]}(\Omega)} + \|\lambda\|_{H^{1+r+[N/2]}(\Omega)}\right) \\ &\le C \|\psi\|_{H^{1+r+[N/2]}(\Omega)}\end{aligned} \tag{3.18}$$

and estimate (3.17), the desired estimate (3.14) follows. This completes the proof of Lemma 3.4. ∎

For the local error estimates, the results will be stated and used for special subdomains of Ω. Without loss of generality we assume *throughout this paper* that $\operatorname{diam}(\Omega) \le 1$. Let

$$d_j = 2^{-j} \quad \text{for } j = 0, 1, 2, \ldots$$

and for any fixed $z \in \Omega$ set

$$\begin{aligned}\Omega_j &= \{x \in \Omega : \; d_{j+1} < |x - z| < d_j\}, \\ \Omega_j^{(1)} &= \{x \in \Omega : \; d_{j+2} < |x - z| < d_{j-1}\}, \\ \Omega_j^{(2)} &= \{x \in \Omega : \; d_{j+3} < |x - z| < d_{j-2}\}, \\ \Omega_j^{(3)} &= \{x \in \Omega : \; d_{j+4} < |x - z| < d_{j-3}\}.\end{aligned} \tag{3.19}$$

Assumption A.3, along with Assumption A.4, takes the following form on the subdomains Ω_j.

LEMMA 3.5. *Let* $(\mathbf{u}, p)$ *and* $(\mathbf{u}_h, p_h)$ *satisfy* (2.4). *If* $d_j \geq \kappa h$, *then*

$$\begin{aligned}(3.20)\quad &\|\mathbf{u}-\mathbf{u}_h\|_{H^1(\Omega_j)} + \|p-p_h\|_{L^2(\Omega_j)}\\ &\leq Ch^r\big(\|\mathbf{u}\|_{H^{r+1}(\Omega_j^{(1)})} + \|p\|_{H^r(\Omega_j^{(1)})}\big)\\ &\quad + Cd_j^{-N/2-1-t}\big(\|\mathbf{u}-\mathbf{u}_h\|_{W^{-t,1}(\Omega_j^{(1)})} + d_j^{1-t_1}\|p-p_h\|_{W^{-t-t_1,1}(\Omega_j^{(1)})}\big),\end{aligned}$$

where $t \geq 0$, $t_1 = 0, 1$.

Proof. Without loss of generality, we assume $z = 0$ and Ω is the unit ball centered at $z = 0$. Then,

$$\Omega_j = \{x \in R^N :\ d_{j+1} < |x| < d_j\}, \quad \Omega_j^{(1)} = \{x \in R^N :\ d_{j+2} < |x| < d_{j-1}\}.$$

Introduce a new variable $\tilde{x} = x/d_{j-1}$. The regions $\Omega_j^{(1)}$ and Ω_j are transferred into regions $\tilde{D}_1$ and $\tilde{D}_0$, respectively,

$$\tilde{D}_0 = \{\tilde{x} \in R^N :\ 1/4 < |\tilde{x}| < 1/2\}, \quad \tilde{D}_1 = \{\tilde{x} \in R^N :\ 1/8 < |\tilde{x}| < 1\}.$$

Then, $\operatorname{dist}(\tilde{D}_0, \partial\tilde{D}_1) = 1/2$. Set

$$\begin{aligned}\tilde{\mathbf{u}}(\tilde{x}) = u(\tilde{x}d_{j-1}), \quad \tilde{\mathbf{u}}_h(\tilde{x}) = \mathbf{u}(\tilde{x}d_{j-1}),\\ \tilde{p}(\tilde{x}) = p(\tilde{x}d_{j-1})d_{j-1}, \quad \tilde{p}_h(\tilde{x}) = p_h(\tilde{x}d_{j-1})d_{j-1}.\end{aligned}$$

Then,

$$\begin{aligned}(3.21)\quad (\tilde{\nabla}(\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h), \tilde{\nabla}\tilde{\mathbf{v}}) - (\tilde{\nabla}\cdot\tilde{\mathbf{v}}, \tilde{p}-\tilde{p}_h) = 0 \quad \forall\ \tilde{v} \in \tilde{V},\\ (\tilde{\nabla}\cdot(\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h), \tilde{q}) = 0 \quad \forall\ \tilde{q} \in \tilde{W},\end{aligned}$$

where $\tilde{\mathbf{V}}$ and $\tilde{W}$ are the transferred spaces of $\mathbf{V}^h$ and W^h. We have

$$\begin{aligned}(3.22)\quad &\|\mathbf{u}-\mathbf{u}_h\|_{H^1(\Omega_j)} + \|p-p_h\|_{L^2(\Omega_j)}\\ &\leq d_j^{N/2-1}\big(\|\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h\|_{H^1(\tilde{D}_0)} + \|\tilde{p}-\tilde{p}_h\|_{L^2(\tilde{D}_0)}\big).\end{aligned}$$

Using Assumption A.3 on problem (3.21), we have for $t \geq 0$ and $t_1 = 0$ or 1,

$$\begin{aligned}(3.23)\quad &\|\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h\|_{H^1(\tilde{D}_0)} + \|\tilde{p}-\tilde{p}_h\|_{L^2(\tilde{D}_0)}\\ &\leq C(h/d_j)^r\big(|\tilde{\mathbf{u}}|_{H^{r+1}(\tilde{D}_1)} + |\tilde{p}|_{H^r(\tilde{D}_1)}\big)\\ &\quad + C\big(\|\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h\|_{W^{-t,1}(\tilde{D}_1)} + \|\tilde{p}-\tilde{p}_h\|_{W^{-t-t_1,1}(\tilde{D}_1)}\big).\end{aligned}$$

Noting that

$$|\tilde{\mathbf{u}}|_{H^{1+r}(\tilde{D}_1)} + |\tilde{p}|_{H^r(\tilde{D}_1)} \leq Cd_j^{-N/2+1+r}\big(|\mathbf{u}|_{H^{1+r}(\Omega_j^{(1)})} + |p|_{H^r(\Omega_j^{(1)})}\big)$$

and

$$\begin{aligned}&\|\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_h\|_{W^{-t,1}(\tilde{D}_1)} + \|\tilde{p}-\tilde{p}_h\|_{W^{-t-t_1,1}(\tilde{D}_1)}\\ &\quad \leq Cd_j^{-N-t}\big(\|\mathbf{u}-\mathbf{u}_h\|_{W^{-t,1}(\Omega_j^{(1)})} + d_j^{1-t_1}\|p-p_h\|_{W^{-t-t_1,1}(\Omega_j^{(1)})}\big),\end{aligned}$$

and using (3.22) and (3.23), we complete the proof. ∎

LEMMA 3.6. *For $\rho \in L^\infty(\Omega)$, let $(\Phi, \lambda) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ be the solution of*

$$-\Delta\Phi + \nabla\lambda = \rho, \quad \nabla \cdot \Phi = 0,$$

and let $(\Phi_h, \lambda_h) \in \mathbf{V}^h \times W^h$ be the finite element approximation of (Φ, λ). If $\mathrm{supp}(\rho) \subset B_{2\kappa h}(z)$ and $\|\rho\|_{L^2(B_{2\kappa h}(z))} \le Ch^{-N/2}$, then

$$\|\Phi - \Phi_h\|_{H^1(B_{Mh}(z))} + \|\lambda - \lambda_h\|_{L^2(B_{Mh}(z))} \le Ch^{1-N/2}, \tag{3.24}$$

and for $j = 0, 1, \ldots, J$, $t_1 = 0$ or 1,

$$\begin{aligned} &\|\Phi - \Phi_h\|_{H^1(\Omega_j)} + \|\lambda - \lambda_h\|_{L^2(\Omega_j)} \\ &\quad \le Cd_j^{-N/2-1}\big(\|\Phi - \Phi_h\|_{L^1(\Omega_j^{(1)})} + d_j^{1-t_1}\|\lambda - \lambda_h\|_{W^{-t_1,1}(\Omega_j^{(1)})}\big) \\ &\qquad + Ch^r d_j^{-N/2+1-r}. \end{aligned} \tag{3.25}$$

Proof. Using Lemmas 3.3 and 3.1, we obtain

$$\begin{aligned} &\|\Phi - \Phi_h\|_{H^1(B_{Mh}(z))} + \|\lambda - \lambda_h\|_{L^2(B_{Mh}(z))} \\ &\quad \le Ch\left(\|\Phi\|_{H^2(\Omega)} + \|\lambda\|_{H^1(\Omega)}\right) \ \le \ Ch\|\rho\|_{L^2(B_{2\kappa h}(z))}, \end{aligned}$$

which implies (3.24) according to the assumption on ρ. To prove (3.25), we use Lemma 3.5 to get

$$\begin{aligned} &\|\Phi - \Phi_h\|_{H^1(\Omega_j)} + \|\lambda - \lambda_h\|_{L^2(\Omega_j)} \\ &\quad \le Ch^r\left(\|\Phi\|_{H^{r+1}(\Omega_j^{(1)})} + \|\lambda\|_{H^r(\Omega_j^{(1)})}\right) \\ &\qquad + Cd_j^{-N/2-1}\left(\|\Phi - \Phi_h\|_{L^1(\Omega_j^{(1)})} + d_j^{1-t_1}\|\lambda - \lambda_h\|_{W^{-t_1,1}(\Omega_j^{(1)})}\right). \end{aligned} \tag{3.26}$$

In view of Lemma 3.2 and the assumption on ρ, it follows that

$$\begin{aligned} &\|\Phi\|_{h^{r+1}(\Omega_j^{(1)})} + \|\lambda\|_{H^r(\Omega_j^{(1)})} \le Cd_j^{N/2}\left(\|\Phi\|_{W^{r+1,\infty}(\Omega_j^{(1)})} + \|\lambda\|_{W^{r,\infty}(\Omega_j^{(1)})}\right) \\ &\quad \le Cd_j^{-N/2+1-r}\|\rho\|_{L^1(B_{2\kappa h}(z))} \le Cd_j^{-N/2+1-r}h^{N/2}\|\rho\|_{L^2(B_{2\kappa h}(z))} \\ &\quad \le Cd_j^{-N/2+1-r}. \end{aligned}$$

Substituting this into (3.26) implies (3.25). Therefore, the proof is complete. ∎

LEMMA 3.7. *Let $(\Phi, \lambda) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ be the solution of*

$$-\Delta\Phi + \nabla\lambda = 0, \quad \nabla \cdot \Phi = g - (1, g)\zeta,$$

and let $(\Phi_h, \lambda_h) \in \mathbf{V}^h \times W^h$ be the finite element approximation of (Φ, λ), where $\zeta \in C_0^\infty(\Omega)$ is a fixed function satisfying $(1, \zeta) = 1$, and

$$g \in C_0^\infty(B_{2\kappa h}(z)) \quad \text{and} \quad \|g\|_{L^2(B_{2\kappa h}(z))} \le Ch^{-N/2}.$$

Then

$$\|\Phi - \Phi_h\|_{H^1(B_{Mh}(z))} + \|\lambda - \lambda_h\|_{L^2(B_{Mh}(z))} \le Ch^{-N/2}, \tag{3.27}$$

and for $j = 0, 1, \ldots, J$, $t \geq 0$, *and* $t_1 = 0$ *or* 1,

$$\begin{aligned}(3.28)\quad &\|\Phi - \Phi_h\|_{H^1(\Omega_j)} + \|\lambda - \lambda_h\|_{L^2(\Omega_j)} \\ &\leq Cd_j^{-N/2-1-t}\left(\|\Phi - \Phi_h\|_{W^{-t,1}(\Omega_j^{(1)})} + d_j^{1-t_1}\|\lambda - \lambda_h\|_{W^{-t_1-t,1}(\Omega_j^{(1)})}\right) \\ &\quad + Ch^r d_j^{-N/2-r}.\end{aligned}$$

Proof. The proof is analogous to that of Lemma 3.6. In view of Lemmas 3.3 and 3.1, it follows that

$$\begin{aligned}&\|\Phi - \Phi_h\|_{H^1(B_{Mh}(z))} + \|\lambda - \lambda_h\|_{L^2(B_{Mh}(z))} \\ &\quad \leq C\left(\|\Phi\|_{H^1(\Omega)} + \|\lambda\|_{L^2(\Omega)}\right) \leq C\|g - (1,g)\zeta\|_{L^2(B_{2\kappa h}(z))} \\ &\quad \leq C\|g\|_{L^2(B_{2\kappa h}(z))} \leq Ch^{-N/2},\end{aligned}$$

which proves (3.27). To prove (3.28), we use Lemma 3.5,

$$\begin{aligned}(3.29)\quad &\|\Phi - \Phi_h\|_{H^1(\Omega_j)} + \|\lambda - \lambda_h\|_{L^2(\Omega_j)} \\ &\leq Ch^r\left(\|\Phi\|_{H^{r+1}(\Omega_j^{(1)})} + \|\lambda\|_{H^r(\Omega_j^{(1)})}\right) \\ &\quad + Cd_j^{-N/2-1-t}\left(\|\Phi - \Phi_h\|_{W^{-t,1}(\Omega_j^{(1)})} + d_j^{1-t_1}\|\lambda - \lambda_h\|_{W^{-t_1-t,1}(\Omega_j^{(1)})}\right).\end{aligned}$$

By Lemma 3.2 and the assumption on g, it follows that

$$\begin{aligned}&\|\Phi\|_{H^{r+1}(\Omega_j^{(1)})} + \|\lambda\|_{H^r(\Omega_j^{(1)})} \leq Cd_j^{N/2}\left(\|\Phi\|_{W^{r+1,\infty}(\Omega_j^{(1)})} + \|\lambda\|_{W^{r,\infty}(\Omega_j^{(1)})}\right) \\ &\quad \leq Cd_j^{-N/2+1-r}d_j^{-1}\|g - (1,g)\zeta\|_{L^1(B_{2\kappa h}(z))} \leq Cd_j^{-N/2-r}h^{N/2}\|g\|_{L^2(B_{2\kappa h}(z))} \\ &\quad \leq Cd_j^{-N/2-r}.\end{aligned}$$

Substituting this into (3.29), we obtain (3.28) and thus complete the proof. ∎

Now, we have the last lemma of this section.

LEMMA 3.8. *Assume that* $\varphi \in C_0^\infty(\Omega_j^{(1)})^N$, $\psi \in C_0^\infty(\Omega_j^{(1)})$ *satisfying*

$$\|\varphi\|_{L^\infty(\Omega_j^{(1)})} \leq 1, \quad \|\psi\|_{W^{1,\infty}(\Omega_j^{(1)})} \leq 1,$$

and $(\mathbf{v}, \mu) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ *is the solution of the problem*

$$-\Delta\mathbf{v} + \nabla\mu = \varphi, \quad \nabla\cdot\mathbf{v} = \psi - (1,\psi)\zeta,$$

where $\zeta \in C_0^\infty(\Omega)$ *is a function such that* $(1,\zeta) = 1$. *Then, for any* $\mathbf{e}_1 \in H_0^1(\Omega)$ *and* $e_2 \in L_0^2(\Omega)$, *there holds*

$$\begin{aligned}(3.30)\quad &\left|a(\mathbf{e}_1, \mathbf{v} - \Pi^h\mathbf{v}) - b(\mathbf{v} - \Pi^h\mathbf{v}, e_2) - b(\mathbf{e}_1, \mu - Q^h\mu)\right| \\ &\quad \leq Ch^r d_j^{1-r}\left(\|\mathbf{e}_1\|_{W^{1,1}(\Omega)} + \|e_2\|_{L^1(\Omega)}\right) \\ &\qquad + Chd_j^{N/2}\left(\|\mathbf{e}_1\|_{H^1(\Omega_j^{(2)})} + \|e_2\|_{L^2(\Omega_j^{(2)})}\right).\end{aligned}$$

Proof. Let us start with the following decomposition:

$$(3.31)\qquad a(\mathbf{e}_1, \mathbf{v} - \Pi^h\mathbf{v}) - b(\mathbf{v} - \Pi^h\mathbf{v}, e_2) - b(\mathbf{e}_1, \mu - Q^h\mu) = I_1 + I_2,$$

where

$$I_1 = a_{\Omega\setminus\Omega_j^{(2)}}(\mathbf{e}_1, \mathbf{v}-\Pi^h\mathbf{v}) - b_{\Omega\setminus\Omega_j^{(2)}}(\mathbf{v}-\Pi^h\mathbf{v}, e_2) - b_{\Omega\setminus\Omega_j^{(2)}}(\mathbf{e}_1, \mu - Q^h\mu),$$
$$I_2 = a_{\Omega_j^{(2)}}(\mathbf{e}_1, \mathbf{v}-\Pi^h\mathbf{v}) - b_{\Omega_j^{(2)}}(\mathbf{v}-\Pi^h\mathbf{v}, e_2) - b_{\Omega_j^{(2)}}(\mathbf{e}_1, \mu - Q^h\mu).$$

Here, $a_D(\cdot,\cdot)$ and $b_D(\cdot,\cdot)$ stand for the respective bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ in which the integrals are computed on the subdomain D. By the Cauchy–Schwarz inequality and the approximation assumption A.1, it follows that

$$\begin{aligned} I_1 &\le C\left(\|\mathbf{e}_1\|_{W^{1,1}(\Omega\setminus\Omega_j^{(2)})} + \|e_2\|_{L^1(\Omega\setminus\Omega_j^{(2)})}\right) \\ &\quad \times C\left(\|\mathbf{v}-\Pi^h\mathbf{v}\|_{W^{1,\infty}(\Omega\setminus\Omega_j^{(2)})} + \|\mu - Q^h\mu\|_{L^\infty(\Omega\setminus\Omega_j^{(2)})}\right) \\ &\le C\left(\|\mathbf{e}_1\|_{W^{1,1}(\Omega\setminus\Omega_j^{(2)})} + \|e_2\|_{L^1(\Omega\setminus\Omega_j^{(2)})}\right) \\ &\quad \times Ch^r\left(\|\mathbf{v}\|_{W^{1+r,\infty}(\Omega\setminus\Omega_j^{(1)})} + \|\mu\|_{W^{r,\infty}(\Omega\setminus\Omega_j^{(1)})}\right). \end{aligned}$$

This, along with Lemma 3.2 and the assumption on φ and ψ, yields

$$\begin{aligned} I_1 &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1\|_{W^{1,1}(\Omega\setminus\Omega_j^{(2)})} + \|e_2\|_{L^1(\Omega\setminus\Omega_j^{(2)})}\right) \\ &\quad \times \left(\|\varphi\|_{L^\infty(\Omega_j^{(1)})} + d_j^{-1}\|\psi\|_{L^\infty(\Omega_j^{(1)})}\right) \\ &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1\|_{W^{1,1}(\Omega)} + \|e_2\|_{L^1(\Omega)}\right). \end{aligned} \tag{3.32}$$

Here, we have used the fact that ψ vanishes on $\partial\Omega_j^{(1)}$ and, therefore, the following estimate holds:

$$\|\psi\|_{L^\infty(\Omega_j^{(1)})} \le Cd_j\|\psi\|_{W^{1,\infty}(\Omega_j^{(1)})} \le Cd_j. \tag{3.33}$$

According to the Cauchy–Schwarz inequality and A.1, I_2 is estimated as follows:

$$\begin{aligned} I_2 &\le C\left(\|\mathbf{e}_1\|_{H^1(\Omega_j^{(2)})} + \|e_2\|_{L^2(\Omega_j^{(2)})}\right) \\ &\quad \times C\left(\|\mathbf{v}-\Pi^h\mathbf{v}\|_{H^1(\Omega_j^{(2)})} + \|\mu - Q^h\mu\|_{L^2(\Omega_j^{(2)})}\right) \\ &\le C\left(\|\mathbf{e}_1\|_{H^1(\Omega_j^{(2)})} + \|e_2\|_{L^2(\Omega_j^{(2)})}\right) \\ &\quad \times Ch\left(\|\mathbf{v}\|_{H^2(\Omega_j^{(3)})} + \|\mu\|_{H^1(\Omega_j^{(3)})}\right). \end{aligned}$$

Using Lemma 3.1,

$$\begin{aligned} I_2 &\le Ch\left(\|\mathbf{e}_1\|_{H^1(\Omega_j^{(2)})} + \|e_2\|_{L^2(\Omega_j^{(2)})}\right)\left(\|\varphi\|_{L^2(\Omega_j^{(1)})} + \|\psi\|_{H^1(\Omega_j^{(1)})}\right) \\ &\le Chd_j^{N/2}\left(\|\mathbf{e}_1\|_{H^1(\Omega_j^{(2)})} + \|e_2\|_{L^2(\Omega_j^{(2)})}\right). \end{aligned} \tag{3.34}$$

Using (3.32) and (3.34) in (3.31), we conclude (3.30). The proof is complete. ∎

4. Pressure error. In this section we derive the pointwise error estimates for the pressure approximation. The main result of this section is in Theorem 4.2. For our purpose, throughout this section, we denote by δ a fixed nonnegative function which satisfies $\delta \in C_0^\infty(B_1(0)$ and $\int_{R^N} \delta(x)\,dx = 1$. We shall construct a *regularized delta function* $\delta_{z,h}$, following an approach suggested by Durán, Nochetto, and Wang [6].

LEMMA 4.1. *There exists a $\tau \in (0, \tau_0)$ such that for any $q \in W^h$ and $z \in \bar{\Omega}$, there holds the following inequality:*

$$|q(z)| \le 2\,|(\delta_{z,h}, q)|\,,$$

where

$$\delta z,_h = (\tau h)^{-N}\delta\left(\frac{x-\bar{z}}{\tau h}\right) \quad \textit{and} \quad |z-\bar{z}| \le Ch.$$

Proof. For any $q \in W^h$ and $z \in \bar{\Omega}$, we have $z \in \bar{K}_j$ for some $1 \le j \le N_h$ and $y \in \bar{K}_j$ such that $\|q\|_{L^\infty(K_j)} = |q(y)|$. For any $\tau \in (0, \tau_0)$, according to A.0, choose $\bar{z} \in K_j$ so that $B_{\tau h}(\bar{z}) \subset K_j$ and $|y - \bar{z}| < C_0\tau h$. Set $\delta_{z,h} = (\tau h)^{-N}\delta(\frac{x-\bar{z}}{\tau h})$. By the mean value theorem of calculus, there is a $\bar{y} \in B_{\tau h}(\bar{z})$ such that

$$(q, \delta_{z,h}) = \int_{B_{\tau h}(\bar{z})} q\delta_{z,h}\,dx = q(\bar{y}).$$

Using the triangular inequality and the inverse properties in A.2, it follows that

$$\begin{aligned} \|q\|_{L^\infty(K_j)} = |q(y)| \;&\le\; |q(\bar{y})| + |q(y) - q(\bar{y})| \\ &\le |q(\bar{y})| + 2C_0\tau h\|\nabla q\|_{L^\infty(K_j)} \\ &\le |(q, \delta_{z,h})| + C_1\tau\|q\|_{L^\infty(K_j)}. \end{aligned} \tag{4.1}$$

The constant C_1 in (4.1) depends only on C_0 and the inverse properties. Choosing any $\tau \le \tau_1 = \frac{1}{2C_1}$, we obtain

$$\|q\|_{L^\infty(K_j)} \le 2|(q, \delta_{z,h})|.$$

In view of $|q(z)| \le \|q\|_{L^\infty(K_j)}$, we complete the proof. ∎

The main result of this section is the following pointwise error estimate for the pressure approximation.

THEOREM 4.2. *Suppose $(\mathbf{u}, p)$ and $(\mathbf{u}_h, p_h)$ satisfy (2.4). Let $z \in \bar{\Omega}$. Then there exists a constant $C > 0$ independent of $z, \mathbf{u}, p, h$ such that for $0 \le s \le r$ there holds*

$$|(p - p_h)(z)| \le C\left(\ln\frac{1}{h}\right)^{\bar{s}}\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \inf_{q\in W^h}\|p-q\|_{L^\infty(\Omega),z,s}\right), \tag{4.2}$$

where $\bar{s} = 0$ if $0 \le s < r$ and $\bar{s} = 1$ if $s = r$.

Proof. Let us start with a triangular inequality: for any $q_h \in W^h$,

$$|(p - p_h)(z)| \le |(p - q_h)(z)| + |(q_h - p_h)(z)|. \tag{4.3}$$

Obviously, the first term in (4.3) is bounded by $\|p - q_h\|_{L^\infty(\Omega),z,s}$. For the second term of (4.3), we shall use the result of Lemma 4.1. In fact, according to Lemma 4.1, there exists a $\tau \in (0, \tau_0)$ such that

$$(q_h - p_h)(z) \le 2|(q_h - p_h, \delta_{z,h})|,$$

where

$$\delta_{z,h}(x) = (\tau h)^{-N}\delta((x-\bar{z})/(\tau h)) \quad \text{for some } |z-\bar{z}| \le C_0\tau h. \tag{4.4}$$

Note that $\|\delta_{z,h}\|_{L^1(B_{\tau h}(\bar{z})),z,-s} \le C$ and $(p-q_h,\delta_{z,h}) \le C\|p-q_h\|_{L^\infty(\Omega),z,s}$. From (4.3), it follows that

$$|(p-p_h)(z)| \le C\|p-q_h\|_{L^\infty(\Omega),z,s} + 2|(p-p_h,\delta_{z,h})|. \tag{4.5}$$

Choose a function $\zeta \in C_0^\infty(\Omega)$ satisfying $(1,\zeta)=1$. Then, $\tilde{\delta}_{z,h} \equiv \delta_{z,h} - \zeta \in L_0^2(\Omega) \cap C_0^\infty(\Omega)$ and

$$|(p-p_h,\delta_{z,h})| \le |(p-p_h,\tilde{\delta}_{z,h})| + |(p-p_h,\zeta)|. \tag{4.6}$$

The second term in (4.6) can be estimated by the negative norm of $p-p_h$:

$$\begin{aligned} |(p-p_h,\zeta)| \;&\le\; \|p-p_h\|_{H^{-r}(\Omega)}\|\zeta\|_{H^r(\Omega)} \\ &\le Ch^r\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{H^1(\Omega)} + \inf_{q\in W^h}\|p-q\|_{L^2(\Omega)}\right) \\ &\le C\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \inf_{q\in W^h}\|p-q\|_{L^\infty(\Omega),z,s}\right), \end{aligned} \tag{4.7}$$

where $0 \le s \le r$. We shall employ a duality argument in order to estimate the first term in (4.6). Let $(\Phi_z^{(1)},\lambda_z^{(1)}) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ denote the unique solution of the Stokes problem

$$\begin{aligned} a(\mathbf{v},\Phi_z^{(1)}) - b(\mathbf{v},\lambda_z^{(1)}) &= 0 \quad \forall\, \mathbf{v}\in H_0^1(\Omega)^N, \\ b(\Phi_z^{(1)},q) &= (\tilde{\delta}_{z,h},q) \quad \forall\, q\in L_0^2(\Omega), \end{aligned} \tag{4.8}$$

and let $(\Phi_{z,h}^{(1)},\lambda_{z,h}^{(1)}) \in \mathbf{V}^h \times W^h$ be the corresponding finite element approximation of $(\Phi_z^{(1)},\lambda_z^{(1)})$. Namely, the following error equations are satisfied:

$$\begin{aligned} a(\mathbf{v},\Phi_z^{(1)}-\Phi_{z,h}^{(1)}) - b(\mathbf{v},\lambda_z^{(1)}-\lambda_{z,h}^{(1)}) &= 0 \quad \forall\, \mathbf{v}\in\mathbf{V}^h, \\ b(\Phi_z^{(1)}-\Phi_{z,h}^{(1)},q) &= 0 \quad \forall\, q\in W^h. \end{aligned} \tag{4.9}$$

By using (4.8), (4.9), and (2.4), the first term of (4.6) is represented in terms of the errors $\Phi_{z,h}^{(1)}-\Phi_z^{(1)}$ and $\lambda_z^{(1)}-\lambda_{z,h}^{(1)}$ as follows:

$$\begin{aligned} &(p-p_h,\tilde{\delta}_{z,h}) = b(\Phi_z^{(1)},p-p_h) \\ &= a(\mathbf{u}-\mathbf{v},\Phi_{z,h}^{(1)}-\Phi_z^{(1)}) - b(\mathbf{u}-\mathbf{v},\lambda_{z,h}^{(1)}-\lambda_z^{(1)}) - b(\Phi_{z,h}^{(1)}-\Phi_z^{(1)},p-q), \end{aligned}$$

where $\mathbf{v}\in\mathbf{V}^h$ and $q\in W^h$ are arbitrary functions. Hence, by Hölder's inequality, we have

$$\begin{aligned} |(p-p_h,\tilde{\delta}_{z,h})| \le C\Big(&\|\Phi_z^{(1)}-\Phi_{z,h}^{(1)}\|_{W^{1,1}(\Omega),z,-s} + \|\lambda_z^{(1)}-\lambda_{z,h}^{(1)}\|_{L^1(\Omega),z,-s}\Big) \\ &\times \left(\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \|p-q\|_{L^\infty(\Omega),z,s}\right), \end{aligned}$$

which together with (4.7), (4.6), (4.5), and the result of Lemma 4.3 proves (4.2). The proof of Theorem 4.2 is complete. □

Next, we show the estimates for $\Phi_z^{(1)} - \Phi_{z,h}^{(1)}$ and $\lambda_z^{(1)} - \lambda_{z,h}^{(1)}$ in the weighted $W^{1,1}(\Omega)$ and $L^1(\Omega)$ norms, which have been used in the proof of Theorem 4.2.

LEMMA 4.3. *Let* $(\Phi_z^{(1)}, \lambda_z^{(1)})$ *and* $(\Phi_{z,h}^{(1)}, \lambda_{z,h}^{(1)})$ *satisfy* (4.8) *and* (4.9). *Then, there is a constant* $C > 0$ *such that for* $0 \le s \le r$,

$$\|\Phi_z^{(1)} - \Phi_{z,h}^{(1)}\|_{W^{1,1}(\Omega),z,-s} + \|\lambda_z^{(1)} - \lambda_{z,h}^{(1)}\|_{L^1(\Omega),z,-s} \le C\left(\ln\frac{1}{h}\right)^{\bar{s}}, \tag{4.10}$$

where $\bar{s} = 0$ *if* $0 \le s < r$ *and* $\bar{s} = 1$ *if* $s = r$.

Proof. Let $M > 1$ be a real number large enough so that the ball $B_{Mh}(z)$ contains the support of function $\delta_{z,h}$ defined by (4.4). M will be further determined later in this proof. Let J be an integer such that $Mh = 2^{-J}$. Then, $J \le C\ln(1/h)$. Set $\mathbf{e}_1^{(1)} = \Phi_z^{(1)} - \Phi_{z,h}^{(1)}$ and $e_2^{(1)} = \lambda_z^{(1)} - \lambda_{z,h}^{(1)}$. Then, in view of $\Omega = B_{Mh}(z)\bigcup(\bigcup_{j=0}^J \Omega_j)$ and Hölder's inequality, it follows that

$$\begin{aligned}
&\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(1)}\|_{L^1(\Omega),z,-s} \\
&\quad \le CM^{N/2+s}h^{N/2}(\|\mathbf{e}_1^{(1)}\|_{H^1(B_{Mh}(z))} + \|e_2^{(1)}\|_{L^2(B_{Mh}(z))}) \\
&\qquad + C\sum_{j=0}^J d_j^{s+N/2}h^{-s}(\|\mathbf{e}_1^{(1)}\|_{H^1(\Omega_j)} + \|e_2^{(1)}\|_{L^2(\Omega_j)}).
\end{aligned} \tag{4.11}$$

By applying (3.27) and (3.28) in Lemma 3.7 with $t = 0$, $t_1 = 1$, $g = \delta_{z,h}$, it follows from inequality (4.11) that

$$\begin{aligned}
\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(1)}\|_{L^1(\Omega),z,-s} &\le CM^{N/2+s} + C\sum_{j=0}^J (h/d_j)^{r-s} \\
&\quad + C\sum_{j=0}^J d_j^{s-1}h^{-s}\left(\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right) \\
&\le CM^{N/2+s} + C\Theta(r-s) + L_1 + L_2,
\end{aligned} \tag{4.12}$$

where Θ, L_1, and L_2 are defined, respectively, by

$$\Theta(\gamma) = \sum_{j=0}^J (h/d_j)^\gamma,$$

$$L_1 = Ch^{-1}\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega),z,1-s}, \quad L_2 = \sum_{j=0}^J d_j^{s-1}h^{-s}\|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})}.$$

We note that since $d_j = 2^{-j}$ and $J \le C\ln\frac{1}{h}$, there holds

$$\Theta(\gamma) = \sum_{j=1}^J \left(\frac{h}{d_j}\right)^\gamma \le \begin{cases} \ln\frac{1}{h} & \text{if } \gamma = 0, \\ \dfrac{1}{M^\gamma(1-2^{-\gamma})} & \text{if } \gamma > 0. \end{cases} \tag{4.13}$$

As one can see, the first two terms in (4.12) are already bounded by $C(\ln 1/h)^{\bar{s}}$. It suffices to estimate $L_1 + L_2$. To this end, we observe that

$$L_1 + L_2 \le CM^{N/2+s} + C\sum_{j=0}^J d_j^{s-1}h^{-s}\left(\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right), \tag{4.14}$$

which is obtained through the same procedure as is (4.11). We shall estimate the norms $\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})}$ and $\|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})}$ in the summation of (4.14). Recall the definitions of norms in $L^1(\Omega_j^{(1)})$ and $W^{-1,1}(\Omega_j^{(1)})$,

$$\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})} = \sup_{\substack{\varphi\in C_0^\infty(\Omega_j^{(1)})^N \\ \|\varphi\|_{L^\infty(\Omega_j^{(1)})}=1}} (\mathbf{e}_1^{(1)}, \varphi) \tag{4.15}$$

and

$$\|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})} = \sup_{\substack{\psi\in C_0^\infty(\Omega_j^{(1)}) \\ \|\psi\|_{W^{1,\infty}(\Omega_j^{(1)})}=1}} (e_2^{(1)}, \psi). \tag{4.16}$$

To estimate $\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})}$, for any $\varphi \in C_0^\infty(\Omega_j^{(1)})^N$ satisfying $\|\varphi\|_{L^\infty(\Omega)} = 1$, let $\mathbf{w}_1 \in H_0^1(\Omega)^N$ and $\lambda_1 \in L_0^2(\Omega)$ be the solution of

$$-\Delta \mathbf{w}_1 + \nabla \lambda_1 = \varphi, \quad \nabla \cdot \mathbf{w}_1 = 0 \quad \text{in } \Omega. \tag{4.17}$$

Then, some simple computations yield

$$\begin{aligned}(\mathbf{e}_1^{(1)}, \varphi) &= a(\mathbf{e}_1^{(1)}, \mathbf{w}_1) - b(\mathbf{e}_1^{(1)}, \lambda_1) \\ &= a(\mathbf{e}_1^{(1)}, \mathbf{w}_1 - \Pi^h \mathbf{w}_1) - b(\mathbf{w}_1 - \Pi^h \mathbf{w}_1, e_2^{(1)}) - b(\mathbf{e}_1^{(1)}, \lambda_1 - Q^h \lambda_1).\end{aligned} \tag{4.18}$$

Using Lemma 3.8 for the right-hand side of (4.18), we have

$$\begin{aligned}(\mathbf{e}_1^{(1)}, \varphi) \le Ch^r d_j^{1-r} &\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)} + \|e_2^{(1)}\|_{L^1(\Omega)} \right) \\ &+ Chd_j^{N/2} \left(\|\mathbf{e}_1^{(1)}\|_{H^1(\Omega_j^{(2)})} + \|e_2^{(1)}\|_{L^2(\Omega_j^{(2)})} \right).\end{aligned} \tag{4.19}$$

On the other hand, for each $\psi \in C_0^\infty(\Omega_j^{(1)})$ with unit norm $\|\psi\|_{W^{1,\infty}(\Omega)} = 1$, let $(\mathbf{w}_2, \lambda_2) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ denote the unique solution of the following auxiliary Stokes problem:

$$-\Delta \mathbf{w}_2 + \nabla \lambda_2 = 0, \quad \nabla \cdot \mathbf{w}_2 = \psi - (1, \psi)\zeta \quad \text{in } \Omega.$$

Here, $\zeta \in C_0^\infty(\Omega)$ is a fixed function satisfying $(1, \zeta) = 1$. Then, some straightforward manipulations yield

$$\begin{aligned}(e_2^{(1)}, \psi) &= (e_2^{(1)}, \psi - (1, \psi)\zeta) + (1, \psi)(e_2^{(1)}, \zeta) \\ &= b(\mathbf{w}_2, e_2^{(1)}) + (1, \psi)(e_2^{(1)}, \zeta) \\ &= a(\mathbf{e}_1^{(1)}, \Pi^h \mathbf{w}_2 - \mathbf{w}_2) + b(\mathbf{w}_2 - \Pi^h \mathbf{w}_2, e_2^{(1)}) + b(\mathbf{e}_1^{(1)}, \lambda_2 - Q^h \lambda_2) \\ &\quad + (1, \psi)(e_2^{(1)}, \zeta).\end{aligned}$$

Applying (3.14) of Lemma 3.4,

$$\left\|e_2^{(1)}\right\|_{H^{-1-r-[N/2]}(\Omega)} \le Ch^r \left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)} + \|e_2^{(1)}\|_{L^1(\Omega)} \right),$$

and Lemma 3.8, we obtain

$$
\begin{aligned}
(e_2^{(1)},\psi) &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right)\\
&\quad + Chd_j^{N/2}\left(\|\mathbf{e}_1^{(1)}\|_{H^1(\Omega_j^{(2)})}+\|e_2^{(1)}\|_{L^2(\Omega_j^{(2)})}\right)\\
&\quad + C\|e_2^{(1)}\|_{H^{-1-r-[N/2]}(\Omega)}\|\zeta\|_{H^{1+r+[N/2]}(\Omega)}\\
&\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right)\\
&\quad + Chd_j^{N/2}\left(\|\mathbf{e}_1^{(1)}\|_{H^1(\Omega_j^{(2)})}+\|e_2^{(1)}\|_{L^2(\Omega_j^{(2)})}\right).
\end{aligned}
\tag{4.20}
$$

For the second terms of (4.19) and (4.20), we apply (3.28) of Lemma 3.7 with $t_1=0$, $t=0$ and obtain

$$
\begin{aligned}
&\|\mathbf{e}_1^{(1)}\|_{H^1(\Omega_j^{(2)})}+\|e_2^{(1)}\|_{L^2(\Omega_j^{(2)})}\\
&\quad\le Cd_j^{-N/2-1}\left(\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(3)})}+d_j\|e_2^{(1)}\|_{L^1(\Omega_j^{(3)})}\right)+Ch^r d_j^{-N/2-r}.
\end{aligned}
\tag{4.21}
$$

Substituting (4.21) into (4.19) and (4.20) and taking into account the definitions of norms (4.15) and (4.16), we conclude that

$$
\begin{aligned}
&\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(1)})}+\|e_2^{(1)}\|_{W^{-1,1}(\Omega_j^{(1)})}\\
&\quad\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right)\\
&\qquad + Ch^{1+r}d_j^{-r}+Chd_j^{-1}\|\mathbf{e}_1^{(1)}\|_{L^1(\Omega_j^{(3)})}+Ch\|e_2^{(1)}\|_{L^1(\Omega_j^{(3)})}.
\end{aligned}
\tag{4.22}
$$

Thus, combining (4.22) and (4.14), we have

$$
\begin{aligned}
L_1+L_2 &\le CM^{N/2+s}+C\Theta(r-s)\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right)\\
&\quad + C\Theta(1+r-s)+C\Theta(1)L_1+C\Theta(1)\|e_2^{(1)}\|_{L^1(\Omega),z,-s}.
\end{aligned}
\tag{4.23}
$$

Notice that $\Theta(\gamma)\to 0$ as $M\to\infty$ for any fixed $\gamma>0$ according to (4.13). Choose M large enough so that $C\Theta(1)\le 1/2$. Then, the term $C\Theta(1)L_1$ on the right-hand side of (4.23) is absorbed into the left-hand side and we arrive at

$$
\begin{aligned}
L_1+L_2 &\le CM^{N/2+s}+C\Theta(1+r-s)+C\Theta(1)\|e_2^{(1)}\|_{L^1(\Omega),z,-s}\\
&\quad + C\Theta(r-s)\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right).
\end{aligned}
\tag{4.24}
$$

The substitution of (4.24) into (4.12) results in

$$
\begin{aligned}
&\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(1)}\|_{L^1(\Omega),z,-s}\\
&\quad\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(1+r-s)+C\Theta(1)\|e_2^{(1)}\|_{L^1(\Omega),z,-s}\\
&\qquad + C\Theta(r-s)\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right).
\end{aligned}
\tag{4.25}
$$

The particular case when $s=0$ in inequality (4.25) implies

$$
\begin{aligned}
\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)} &\le CM^{N/2}+C\Theta(r)+C\Theta(1+r)+C\Theta(1)\|e_2^{(1)}\|_{L^1(\Omega)}\\
&\quad + C\Theta(r)\left(\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)}+\|e_2^{(1)}\|_{L^1(\Omega)}\right),
\end{aligned}
$$

which, choosing M large enough, results in

$$\|\mathbf{e}_1^{(1)}\|_{W^{1,1}(\Omega)} + \|e_2^{(1)}\|_{L^1(\Omega)} \le C. \tag{4.26}$$

Using (4.26) in (4.25), we get (4.10). This completes the proof of Lemma 4.3. ∎

5. Velocity error. This section is devoted to deriving the pointwise error estimate for the velocity error $\mathbf{u} - \mathbf{u}_h$. The result is stated in Theorem 5.1.

THEOREM 5.1. *Suppose* $(\mathbf{u}, p)$ *and* $(\mathbf{u}_h, p_h)$ *satisfy* (2.4). *Let* $z \in \bar{\Omega}$. *Then there exists a constant* $C > 0$ *independent of* $z, \mathbf{u}, p, h$ *such that for* $0 \le s \le r-1$ *there holds*

(5.1)

$$|(\mathbf{u} - \mathbf{u}_h)(z)| \le Ch \left(\ln \frac{1}{h}\right)^{\bar{\bar{s}}} \left(\inf_{\mathbf{v} \in \mathbf{V}^h} \|\mathbf{u} - \mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \inf_{q \in W^h} \|p - q\|_{L^\infty(\Omega),z,s} \right),$$

where $\bar{\bar{s}} = 0$ *if* $0 \le s < r-1$ *and* $\bar{\bar{s}} = 1$ *if* $s = r-1$.

Proof. The triangular inequality, the inverse properties in A.2, and the approximation properties in A.1 yield for any $\mathbf{v} \in \mathbf{V}^h$

$$\begin{aligned}
|(\mathbf{u} - \mathbf{u}_h)(z)| &\le |(\mathbf{u} - \mathbf{v})(z)| + Ch^{-N/2} \|\mathbf{v} - \mathbf{u}_h\|_{L^2(B_{2\kappa h}(z))} \\
&\le C\|\mathbf{u} - \mathbf{v}\|_{L^\infty(B_{2\kappa h}(z))} + Ch^{-N/2} \|\mathbf{u} - \mathbf{u}_h\|_{L^2(B_{2\kappa h}(z))} \\
&\le Ch\|\mathbf{u}\|_{W^{1,\infty}(B_{3\kappa h}(z))} + Ch^{-N/2} \|\mathbf{u} - \mathbf{u}_h\|_{L^2(B_{2\kappa h}(z))} \\
&\le Ch\|\mathbf{u}\|_{W^{1,\infty}(\Omega),z,s} + Ch^{-N/2} \|\mathbf{u} - \mathbf{u}_h\|_{L^2(B_{2\kappa h}(z))}.
\end{aligned}$$

Replacing $\mathbf{u}$ by $\mathbf{u} - \mathbf{v}$, we obtain

$$|(\mathbf{u} - \mathbf{u}_h)(z)| \le Ch\|\mathbf{u} - \mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + Ch^{-N/2} \|\mathbf{u} - \mathbf{u}_h\|_{L^2(B_{2\kappa h}(z))}. \tag{5.2}$$

We shall estimate the second term in (5.2) by employing a duality argument. To this end, define a function

$$\rho(x) = \begin{cases} h^{-N/2}(\mathbf{u} - \mathbf{u}_h)(x) / \|\mathbf{u} - \mathbf{u}_h\|_{L^2(E_{2\kappa h}(x))} & \text{for } x \in B_{2\kappa h}(z), \\ 0 & \text{elsewhere} \end{cases}$$

and let $(\Phi_z^{(2)}, \lambda_z^{(2)}) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ denote the unique solution of

$$\begin{aligned}
a(\mathbf{v}, \Phi_z^{(2)}) - b(\mathbf{v}, \lambda_z^{(2)}) &= (\rho, \mathbf{v}) \quad \forall\, \mathbf{v} \in H_0^1(\Omega)^N, \\
b(\Phi_z^{(2)}, q) &= 0 \quad \forall\, q \in L_0^2(\Omega),
\end{aligned} \tag{5.3}$$

and let $(\Phi_{z,h}^{(2)}, \lambda_{z,h}^{(2)}) \in \mathbf{V}^h \times W^h$ be the corresponding finite element approximation of $(\Phi_z^{(2)}, \lambda_z^{(2)})$ so that

$$\begin{aligned}
a(\mathbf{v}, \Phi_z^{(2)} - \Phi_{z,h}^{(2)}) - b(\mathbf{v}, \lambda_z^{(2)} - \lambda_{z,h}^{(2)}) &= 0 \quad \forall\, \mathbf{v} \in \mathbf{V}^h, \\
b(\Phi_z^{(2)} - \Phi_{z,h}^{(2)}, q) &= 0 \quad \forall\, q \in W^h.
\end{aligned} \tag{5.4}$$

Then, for any $\mathbf{v} \in \mathbf{V}^h$ and $q \in W^h$, some simple manipulations using (5.3), (2.4), and (5.4) and Hölder's inequality result in

$$\begin{aligned}
&h^{-N/2}\|\mathbf{u} - \mathbf{u}_h\|_{L^2(E_{2\kappa h}(z))} = (\rho, \mathbf{u} - \mathbf{u}_h) = a(\mathbf{u} - \mathbf{u}_h, \Phi_z^{(2)}) - b(\mathbf{u} - \mathbf{u}_h, \lambda_z^{(2)}) \\
&= a(\mathbf{u} - \mathbf{v}, \Phi_z^{(2)} - \Phi_{z,h}^{(2)}) - b(\mathbf{u} - \mathbf{v}, \lambda_z^{(2)} - \lambda_{z,h}^{(2)}) - b(\Phi_z^{(2)} - \Phi_{z,h}^{(2)}, p - q) \\
&\le C \left(\|\Phi_z^{(2)} - \Phi_{z,h}^{(2)}\|_{W^{1,1}(\Omega),z,-s} + \|\lambda_z^{(2)} - \lambda_{z,h}^{(2)}\|_{L^1(\Omega),z,-s} \right) \\
&\quad \times C \left(\|\mathbf{u} - \mathbf{v}\|_{W^{1,\infty}(\Omega),z,s} + \|p - q\|_{L^\infty(\Omega),z,s} \right),
\end{aligned}$$

which, along with (5.2) and Lemma 5.2, yields the desired estimate (5.1). Thus, the proof is complete. ∎

The result of the next lemma is used in the proof of Theorem 5.1. It is a sharp error estimate for the solution of the auxiliary Stokes problem (5.3) and (5.4) in the weighted $W^{1,1}(\Omega)$ and $L^1(\Omega)$ norms.

LEMMA 5.2. *Let $(\Phi_z^{(2)}, \lambda_z^{(2)})$ be the solution of* (5.3) *and let $(\Phi_{z,h}^{(2)}, \lambda_{z,h}^{(2)})$ be the corresponding finite element approximation satisfying* (5.4). *Then, there is a constant $C > 0$ such that for $0 \le s \le r-1$,*

$$\|\Phi_z^{(2)} - \Phi_{z,h}^{(2)}\|_{W^{1,1}(\Omega),z,-s} + \|\lambda_z^{(2)} - \lambda_{z,h}^{(2)}\|_{L^1(\Omega),z,-s} \le Ch \left(\ln \frac{1}{h}\right)^{\bar{\bar{s}}}, \tag{5.5}$$

where $\bar{\bar{s}} = 0$ if $0 \le s < r-1$ and $\bar{\bar{s}} = 1$ if $s = r-1$.

Proof. As before, let $M > 1$ be a real number, whose value will be determined later in the proof, and let J be an integer such that $Mh = 2^{-J}$. Then, $J \le C \ln \frac{1}{h}$. Set $\mathbf{e}_1^{(2)} = \Phi_z^{(2)} - \Phi_{z,h}^{(2)}$ and $e_2^{(2)} = \lambda_z^{(2)} - \lambda_{z,h}^{(2)}$. Then, analogous to inequality (4.11), we have

$$\begin{aligned} &\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2\|_{L^1(\Omega),z,-s} \\ &\quad \le CM^{N/2+s} h^{N/2} (\|\mathbf{e}_1^{(2)}\|_{H^1(B_{Mh}(z))} + \|e_2^{(2)}\|_{L^2(B_{Mh}(z))}) \\ &\qquad + C \sum_{j=0}^{J} d_j^{s+N/2} h^{-s} (\|\mathbf{e}_1^{(2)}\|_{H^1(\Omega_j)} + \|e_2^{(2)}\|_{L^2(\Omega_j)}). \end{aligned} \tag{5.6}$$

The norms on the right-hand side of (5.6) can be estimated by using (3.24) and (3.25) in Lemma 3.6 with $t_1 = 1$. Consequently,

$$\begin{aligned} \|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(2)}\|_{L^1(\Omega),z,-s} &\le CM^{N/2+s} h + Ch \sum_{j=0}^{J} (h/d_j)^{r-1-s} \\ &\quad + C \sum_{j=0}^{J} d_j^{s-1} h^{-s} \left(\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right) \\ &\le CM^{N/2+s} h + Ch\Theta(r-1-s) + L_3 + L_4, \end{aligned} \tag{5.7}$$

where

$$L_3 = Ch^{-1} \|\mathbf{e}_1^{(2)}\|_{L^1(\Omega),z,1-s}, \quad L_4 = \sum_{j=0}^{J} d_j^{s-1} h^{-s} \|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})}.$$

We are now in a position to estimate $L_3 + L_4$. It follows immediately from the definitions of L_3 and L_4 that

$$L_3 + L_4 \le CM^{N/2+s} h + C \sum_{j=0}^{J} d_j^{s-1} h^{-s} \left(\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right). \tag{5.8}$$

To estimate $\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(1)})}$ and $\|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})}$ for each $0 \le j \le J$, recall the definitions of the norms

$$\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(1)})} = \sup_{\substack{\varphi \in C_0^\infty(\Omega_j^{(1)})^N \\ \|\varphi\|_{L^\infty(\Omega_j^{(1)})} = 1}} (\mathbf{e}_1^{(2)}, \varphi), \tag{5.9}$$

$$\|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})} = \sup_{\substack{\psi\in C_0^\infty(\Omega_j^{(1)})\\ \|\psi\|_{W^{1,\infty}(\Omega_j^{(1)})}=1}} (e_2^{(2)},\psi). \tag{5.10}$$

For each $\varphi \in C_0^\infty(\Omega_j^{(1)})^N$ satisfying $\|\varphi\|_{L^\infty(\Omega)} = 1$, similar to (4.19), we have

$$\begin{aligned}(\mathbf{e}_1^{(2)},\varphi) &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right)\\ &\quad + Chd_j^{N/2}\left(\|\mathbf{e}_1^{(2)}\|_{H^1(\Omega_j^{(2)})} + \|e_2^{(2)}\|_{L^2(\Omega_j^{(2)})}\right),\end{aligned} \tag{5.11}$$

and for each $\psi \in C_0^\infty(\Omega_j^{(1)})$ satisfying $\|\psi\|_{W^{1,\infty}(\Omega)} = 1$, similar to (4.20), we have

$$\begin{aligned}(e_2^{(2)},\psi) &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right)\\ &\quad + Chd_j^{N/2}\left(\|\mathbf{e}_1^{(2)}\|_{H^1(\Omega_j^{(2)})} + \|e_2^{(2)}\|_{L^2(\Omega_j^{(2)})}\right).\end{aligned} \tag{5.12}$$

Using (3.25) of Lemma 3.7 with $t_1 = 0$, the last terms of (5.12) and (5.11) are bounded as follows:

$$\begin{aligned}&\|\mathbf{e}_1^{(2)}\|_{H^1(\Omega_j^{(2)})} + \|e_2^{(2)}\|_{L^2(\Omega_j^{(2)})}\\ &\le Cd_j^{-N/2-1}\left(\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(3)})} + d_j\|e_2^{(2)}\|_{L^1(\Omega_j^{(3)})}\right) + Ch^r d_j^{-N/2+1-r}.\end{aligned} \tag{5.13}$$

Substituting (5.11), (5.12), and (5.13) into (5.9) and (5.10), we obtain

$$\begin{aligned}&\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(2)}\|_{W^{-1,1}(\Omega_j^{(1)})}\\ &\le Ch^r d_j^{1-r}\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right)\\ &\quad + Ch^{1+r}d_j^{1-r} + Chd_j^{-1}\|\mathbf{e}_1^{(2)}\|_{L^1(\Omega_j^{(3)})} + Ch\|e_2^{(2)}\|_{L^1(\Omega_j^{(3)})}.\end{aligned} \tag{5.14}$$

Thus, using (5.14) in (5.8), we obtain

$$\begin{aligned}L_3 + L_4 &\le CM^{N/2+s}h + C\Theta(r-s)\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right)\\ &\quad + Ch\Theta(r-s) + C\Theta(1)L_3 + C\Theta(1)\|e_2^{(2)}\|_{L^1(\Omega),z,-s}.\end{aligned} \tag{5.15}$$

Because of (4.13), we can choose M large enough so that the term $C\Theta(1)L_3$ becomes so small that it can be absorbed into the left-hand side of (5.15). Therefore,

$$\begin{aligned}L_3 + L_4 &\le CM^{N/2+s}h + Ch\Theta(r-s) + C\Theta(1)\|e_2^{(2)}\|_{L^1(\Omega),z,-s}\\ &\quad + C\Theta(r-s)\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right).\end{aligned} \tag{5.16}$$

From (5.16) and (5.7), it follows that

$$\begin{aligned}&\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(2)}\|_{L^1(\Omega),z,-s}\\ &\le CM^{N/2+s}h + Ch\Theta(r-1-s) + Ch\Theta(r-s) + C\Theta(1)\|e_2^{(2)}\|_{L^1(\Omega),z,-s}\\ &\quad + C\Theta(r-s)\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)} + \|e_2^{(2)}\|_{L^1(\Omega)}\right).\end{aligned}$$

Choosing M sufficiently large such that $\Theta(1)$ is small enough so that the term $C\Theta(1)\|e_2^{(2)}\|_{L^1(\Omega),z,-s}$ on the right-hand side can be absorbed into the left-hand side, we get

$$\begin{aligned}(5.17)\qquad &\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(2)}\|_{L^1(\Omega),z,-s}\\ &\quad\le CM^{N/2+s}h+Ch\Theta(r-1-s)+Ch\Theta(r-s)\\ &\qquad+C\Theta(r-s)\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)}+\|e_2^{(2)}\|_{L^1(\Omega)}\right).\end{aligned}$$

The special case of (5.17) when $s=0$ is

$$\begin{aligned}\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)}+\|e_2^{(2)}\|_{L^1(\Omega)}&\le CM^{N/2+s}h+Ch\Theta(r-1)+Ch\Theta(r)\\ &\quad+C\Theta(r)\left(\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)}+\|e_2^{(2)}\|_{L^1(\Omega)}\right),\end{aligned}$$

which, by choosing M large enough, gives

$$\begin{aligned}\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega)}+\|e_2^{(2)}\|_{L^1(\Omega)}&\le CM^{N/2+s}h+Ch\Theta(r-1)+Ch\Theta(r),\\ &\le Ch+Ch\Theta(r-1).\end{aligned}$$

Combining this with (5.17), we obtain

$$\begin{aligned}&\|\mathbf{e}_1^{(2)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(2)}\|_{L^1(\Omega),z,-s}\\ &\quad\le CM^{N/2+s}h+Ch\Theta(r-1-s)+Ch\Theta(r-s)+Ch\Theta(r-1)\Theta(r-s)\\ &\quad\le Ch\left(\ln\frac{1}{h}\right)^{\bar{s}}.\end{aligned}$$

This completes the proof of Lemma 5.2. ∎

6. Gradient of velocity error. This section is devoted to the pointwise estimate for the derivatives of the velocity error. The main result is stated in Theorem 6.1.

THEOREM 6.1. *Suppose* $(\mathbf{u},p)$ *and* $(\mathbf{u}_h,p_h)$ *satisfy* (2.4). *Let* $z\in\bar{\Omega}$. *Then there exists a constant* $C>0$ *independent of* $z,\mathbf{u},p,h$ *such that for* $0\le s\le r$ *there holds*

$$(6.1)\qquad |\nabla(\mathbf{u}-\mathbf{u}_h)(z)|\le C\left(\ln\frac{1}{h}\right)^{\bar{s}}\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s}+\inf_{q\in W^h}\|p-q\|_{L^\infty(\Omega),z,s}\right),$$

where $\bar{s}=0$ *if* i $0\le s<r$ *and* $\bar{s}=1$ *if* $s=r$.

Proof. To begin with, using A.1 and A.2 and following a similar procedure as in deriving (5.2), we easily arrive at

$$\begin{aligned}(6.2)\qquad \left|\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h)(z)\right|&\le\left\|\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h)\right\|_{L^\infty(B_{2\kappa h}(z))}\\ &\quad+Ch^{-N/2-1}\left\|\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h)\right\|_{H^{-1}(B_{2\kappa h}(z))}.\end{aligned}$$

By a duality argument and integration by parts,

$$
\begin{aligned}
(6.3)\qquad & h^{-N/2-1}\left\|\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h)\right\|_{H^{-1}(B_{2\kappa h}(z))}\\
&= \sup_{\substack{\varphi\in C_0^\infty(B_{2\kappa h}(z))^N\\ \|\varphi\|_{H^1(B_{2\kappa h}(z))}=1}}\left(h^{-N/2-1}\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h),\varphi\right)\\
&= \sup_{\substack{\varphi\in C_0^\infty(B_{2\kappa h}(z))^N\\ \|\varphi\|_{H^1(B_{2\kappa h}(z))}=1}}\left(\mathbf{u}-\mathbf{u}_h,-h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right).
\end{aligned}
$$

Set

$$
(6.4)\qquad \rho_1=-h^{-N/2-1}\frac{\partial\varphi}{\partial x_i},
$$

let $(\Phi_z^{(3)},\lambda_z^{(3)})\in H_0^1(\Omega)^N\times L_0^2(\Omega)$ denote the solution of

$$
(6.5)\qquad \begin{aligned} a(\mathbf{v},\Phi_z^{(3)})-b(\mathbf{v},\lambda_z^{(3)})&=(\rho_1,\mathbf{v})\quad\forall\ \mathbf{v}\in H_0^1(\Omega)^N,\\ b(\Phi_z^{(3)},q)&=0\quad\forall\ q\in L_0^2(\Omega),\end{aligned}
$$

and let $(\Phi_{z,h}^{(3)},\lambda_{z,h}^{(3)})\in\mathbf{V}^h\times W^h$ be the finite element approximation of $(\Phi_z^{(3)},\lambda_z^{(3)})$ so that

$$
(6.6)\qquad \begin{aligned} a(\mathbf{v},\Phi_z^{(3)}-\Phi_{z,h}^{(3)})-b(\mathbf{v},\lambda_z^{(3)}-\lambda_{z,h}^{(3)})&=0\quad\forall\ \mathbf{v}\in\mathbf{V}^h,\\ b(\Phi_z^{(3)}-\Phi_{z,h}^{(3)},q)&=0\quad\forall\ q\in W^h.\end{aligned}
$$

Then, for any $\mathbf{v}\in\mathbf{V}^h$ and $q\in W^h$, arguing in the same way as before, we obtain

$$
\begin{aligned}
&\left(\mathbf{u}-\mathbf{u}_h,h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right)=(\rho_1,\mathbf{u}-\mathbf{u}_h)\\
&\quad=a(\mathbf{u}-\mathbf{u}_h,\Phi_z^{(3)})-b(\mathbf{u}-\mathbf{u}_h,\lambda_z^{(3)})\\
&\quad=a(\mathbf{u}-\mathbf{v},\Phi_z^{(3)}-\Phi_{z,h}^{(3)})-b(\mathbf{u}-\mathbf{v},\lambda_z^{(3)}-\lambda_{z,h}^{(3)})-b(\Phi_z^{(3)}-\Phi_{z,h}^{(3)},p-q),
\end{aligned}
$$

which, along with Hölder's inequality and (6.8) of Lemma 6.2, yields

$$
\begin{aligned}
(6.7)\qquad &\left(\mathbf{u}-\mathbf{u}_h,h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right)\\
&\le C\left(\|\Phi_z^{(3)}-\Phi_{z,h}^{(3)}\|_{W^{1,1}(\Omega),z,-s}+\|\lambda_z^{(3)}-\lambda_{z,h}^{(3)}\|_{L^1(\Omega),z,-s}\right)\\
&\quad\times C\left(\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s}+\|p-q\|_{L^\infty(\Omega),z,s}\right)\\
&\le C\left(\ln\frac{1}{h}\right)^{\bar s}\left(\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s}+\|p-q\|_{L^\infty(\Omega),z,s}\right).
\end{aligned}
$$

Estimate (6.7), together with (6.2) and (6.3), shows

$$
\begin{aligned}
&\left|\frac{\partial}{\partial x_i}(\mathbf{u}-\mathbf{u}_h)(z)\right|\\
&\quad\le C\left(\ln\frac{1}{h}\right)^{\bar s}\left(\inf_{\mathbf{v}\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}\|_{W^{1,\infty}(\Omega),z,s}+\inf_{q\in W^h}\|p-q\|_{L^\infty(\Omega),z,s}\right),
\end{aligned}
$$

which proves (6.1). ∎

In the rest of this section, we prove the error estimates for the solutions of (6.5) and (6.6) in the weighted $W^{1,1}$ and L^1 norms, which have been used in the proof of Theorem 6.1.

LEMMA 6.2. *Let* $(\Phi_z^{(3)}, \lambda_z^{(3)})$ *and* $(\Phi_{z,h}^{(3)}, \lambda_{z,h}^{(3)})$ *satisfy* (6.5) *and* (6.6). *Then, there is a constant* $C > 0$ *such that for* $0 \le s \le r$,

$$\|\Phi_z^{(3)} - \Phi_{z,h}^{(3)}\|_{W^{1,1}(\Omega),z,-s} + \|\lambda_z^{(3)} - \lambda_{z,h}^{(3)}\|_{L^1(\Omega),z,-s} \le C\left(\ln\frac{1}{h}\right)^{\bar{s}}, \tag{6.8}$$

where $\bar{s} = 0$ *if* $0 \le s < r$ *and* $\bar{s} = 1$ *if* $s = r$.

Proof. The proof of (6.8) follows closely that of (4.10). Let M and J be as before and set $\mathbf{e}_1^{(3)} = \Phi_z^{(3)} - \Phi_{z,h}^{(3)}$ and $e_2^{(3)} = \lambda_z^{(3)} - \lambda_{z,h}^{(3)}$. Then, analogous to (4.11), there holds

$$\begin{aligned}\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(3)}\|_{L^1(\Omega),z,-s} &\\ \le CM^{N/2+s}h^{N/2}(\|\mathbf{e}_1^{(3)}\|_{H^1(B_{Mh}(z))} + \|e_2^{(3)}\|_{L^2(B_{Mh}(z))}) &\\ + C\sum_{j=0}^{J} d_j^{s+N/2}h^{-s}(\|\mathbf{e}_1^{(3)}\|_{H^1(\Omega_j)} + \|e_2^{(3)}\|_{L^2(\Omega_j)}).&\end{aligned} \tag{6.9}$$

This time, we shall not use (3.24) and (3.25) in Lemma 3.24 for the norms in (6.9) as we did for (5.6). Instead, (3.12) and (3.20) combined with (6.23) of Lemma 6.3 will be used. As a matter of fact, using (3.12) of Lemma 3.3, we obtain

$$\begin{aligned}\|\mathbf{e}_1^{(3)}\|_{H^1(B_{Mh}(z))} + \|e_2^{(3)}\|_{L^2(B_{Mh}(z))} &\\ \le Ch\left(\|\Phi_z^{(3)}\|_{H^2(\Omega)} + \|\lambda_z^{(3)}\|_{H^1(\Omega)}\right) \le Ch\|\rho_1\|_{L^2(\Omega)} \le Ch^{-N/2}.&\end{aligned} \tag{6.10}$$

Applying (3.20) with $t_1 = 1$, $t = 0$ in Lemma 3.5 and (6.23) in Lemma 6.3, we have

$$\begin{aligned}\|\mathbf{e}_1^{(3)}\|_{H^1(\Omega_j)} + \|e_2^{(3)}\|_{L^2(\Omega_j)} &\\ \le Ch^r\left(\|\Phi_z^{(3)}\|_{H^{1+r}(\Omega_j^{(1)})} + \|\lambda_z^{(3)}\|_{H^r(\Omega_j^{(3)})}\right) &\\ + Cd_j^{-N/2-1}\left(\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right) &\\ \le Ch^r d_j^{-N/2-r} + Cd_j^{-N/2-1}\left(\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right).&\end{aligned} \tag{6.11}$$

Substituting (6.10) and (6.11) into (6.9), we obtain

$$\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega),z,-s} + \|e_2^{(3)}\|_{L^1(\Omega),z,-s} \le CM^{N/2+s} + C\Theta(r-s) + L_5 + L_6, \tag{6.12}$$

where L_5 and L_6 are defined by

$$L_5 = Ch^{-1}\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega),z,1-s}, \quad L_6 = \sum_{j=0}^{J} d_j^{s-1}h^{-s}\|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}.$$

Since

$$L_5 + L_6 \le CM^{N/2+s} + C\sum_{j=0}^{J} d_j^{s-1}h^{-s}\left(\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right), \tag{6.13}$$

we shall estimate the norms $\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})}$ and $\|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}$, which are defined through

$$\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} = \sup_{\substack{\varphi\in C_0^\infty(\Omega_j^{(1)})^N \\ \|\varphi\|_{L^\infty(\Omega_j^{(1)})}=1}} (\mathbf{e}_1^{(3)}, \varphi), \tag{6.14}$$

$$\|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})} = \sup_{\substack{\psi\in C_0^\infty(\Omega_j^{(1)}) \\ \|\psi\|_{W^{1,\infty}(\Omega_j^{(1)})}=1}} (e_2^{(3)}, \psi). \tag{6.15}$$

For each $\varphi \in C_0^\infty(\Omega_j^{(1)})^N$ with unit norm $\|\varphi\|_{L^\infty(\Omega)} = 1$, and each $\psi \in C_0^\infty(\Omega_j^{(1)})$ with unit norm $\|\psi\|_{W^{1,\infty}(\Omega)} = 1$, following (4.19) and (4.20), we have

$$\begin{aligned}(\mathbf{e}_1^{(3)}, \varphi) \le{}& Ch^r d_j^{1-r} \left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)} + \|e_2^{(3)}\|_{L^1(\Omega)}\right) \\ &+ Chd_j^{N/2} \left(\|\mathbf{e}_1^{(3)}\|_{H^1(\Omega_j^{(2)})} + \|e_2^{(3)}\|_{L^2(\Omega_j^{(2)})}\right)\end{aligned} \tag{6.16}$$

and

$$\begin{aligned}(e_2^{(3)}, \psi) \le{}& Ch^r d_j^{1-r} \left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)} + \|e_2^{(3)}\|_{L^1(\Omega)}\right) \\ &+ Chd_j^{N/2} \left(\|\mathbf{e}_1^{(3)}\|_{H^1(\Omega_j^{(3)})} + \|e_2^{(3)}\|_{L^2(\Omega_j^{(2)})}\right).\end{aligned} \tag{6.17}$$

Applying (3.20) with $t_1 = 0$ in Lemma 3.5 and (6.23) in Lemma 6.3,

$$\begin{aligned}&\|\mathbf{e}_1^{(3)}\|_{H^1(\Omega_j)} + \|e_2^{(3)}\|_{L^2(\Omega_j)} \\ &\quad\le Ch^r \left(\|\Phi_z^{(3)}\|_{H^{1+r}(\Omega_j^{(1)})} + \|\lambda_z^{(3)}\|_{H^r(\Omega_j^{(3)})}\right) \\ &\qquad + Cd_j^{-N/2-1} \left(\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + d_j\|e_2^{(3)}\|_{L^1(\Omega_j^{(1)})}\right) \\ &\quad\le Ch^r d_j^{-N/2-r} + Cd_j^{-N/2-1} \left(\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + d_j\|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})}\right).\end{aligned} \tag{6.18}$$

Substituting (6.18) into (6.16) and (6.17) and then using (6.16) and (6.17) in (6.14) and (6.15), we obtain

$$\begin{aligned}&\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(1)})} + \|e_2^{(3)}\|_{W^{-1,1}(\Omega_j^{(1)})} \\ &\le Ch^r d_j^{1-r} \left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)} + \|e_2^{(3)}\|_{L^1(\Omega)}\right) \\ &\quad + Ch^{1+r} d_j^{-r} + Chd_j^{-1}\|\mathbf{e}_1^{(3)}\|_{L^1(\Omega_j^{(3)})} + Ch\|e_2^{(3)}\|_{L^1(\Omega_j^{(3)})},\end{aligned}$$

which, combined with (6.13), yields

$$\begin{aligned}L_5 + L_6 \le{}& CM^{N/2+s} + \Theta(r-s) \left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)} + \|e_2^{(3)}\|_{L^1(\Omega)}\right) \\ &+ C\Theta(1+r-s) + C\Theta(1)L_5 + C\Theta(1)\|e_2^{(3)}\|_{L^1(\Omega),z,-s}.\end{aligned} \tag{6.19}$$

Because of (4.13), we can choose M large enough so that

$$\begin{aligned}L_5 + L_6 \le{}& CM^{N/2+s} + C\Theta(1+r-s) + C\Theta(1)\|e_2^{(3)}\|_{L^1(\Omega),z,-s} \\ &+ \Theta(r-s) \left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)} + \|e_2^{(3)}\|_{L^1(\Omega)}\right).\end{aligned} \tag{6.20}$$

Using (6.20) and (6.12), we arrive at

$$\begin{aligned}(6.21)\quad &\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(3)}\|_{L^1(\Omega),z,-s}\\ &\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(1+r-s)+C\Theta(1)\|e_2^{(3)}\|_{L^1(\Omega),z,-s}\\ &\quad+\Theta(r-s)\left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)}+\|e_2^{(3)}\|_{L^1(\Omega)}\right).\end{aligned}$$

Choosing M sufficiently large to cancel the term $C\Theta(1)\|e_2^{(3)}\|_{L^1(\Omega),z,-s}$ on the right-hand side of (6.21), we conclude

$$\begin{aligned}(6.22)\qquad &\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(3)}\|_{L^1(\Omega),z,-s}\\ &\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(1+r-s)\\ &\quad+\Theta(r-s)\left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)}+\|e_2^{(3)}\|_{L^1(\Omega)}\right).\end{aligned}$$

Taking $s=0$ in (6.21), we have

$$\begin{aligned}\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)}+\|e_2^{(3)}\|_{L^1(\Omega)}&\le CM^{N/2}+C\Theta(r)+C\Theta(1+r)\\ &\quad+\Theta(r)\left(\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)}+\|e_2^{(3)}\|_{L^1(\Omega)}\right),\end{aligned}$$

which, when M is chosen sufficiently large, implies

$$\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega)}+\|e_2^{(3)}\|_{L^1(\Omega)}\le CM^{N/2}+C\Theta(r)+C\Theta(1+r)\le C.$$

Substituting this into (6.21), we get

$$\begin{aligned}&\|\mathbf{e}_1^{(3)}\|_{W^{1,1}(\Omega),z,-s}+\|e_2^{(3)}\|_{L^1(\Omega),z,-s}\\ &\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(1+r-s)+\Theta(r-s)\\ &\le C\left(\ln\frac{1}{h}\right)^{\bar{s}},\end{aligned}$$

which is the desired (6.8). The proof is complete. □

LEMMA 6.3. *Suppose ρ_1 is defined by (6.4) for some $\varphi\in C_0^\infty(B_{2\kappa h}(z))^N$ satisfying $\|\varphi\|_{H^1(B_{2\kappa h}(z))}=1$ and $(\mathbf{v},\lambda)\in H_0^1(\Omega)^N\times L_0^2(\Omega)$ is the solution of*

$$-\Delta\mathbf{v}+\nabla\lambda=\rho_1,\quad \nabla\cdot\mathbf{v}=0\quad \text{in }\ \Omega.$$

Then, there holds

$$(6.23)\qquad \|\mathbf{v}\|_{H^{1+r}(\Omega_j^{(1)})}+\|\lambda\|_{H^r(\Omega_j^{(1)})}\le Cd_j^{-N/2-r}.$$

Proof. We shall start with the integral representation (3.3) and (3.4) for the solution of the Stokes problem. For $x\in\Omega_j^{(1)}$, it follows that

$$(6.24)\qquad \mathbf{v}(x)=\sum_{t=1}^N\int_\Omega \mathbf{G}_x^t(y)\rho_1(y)\,dy,\quad \lambda(x)=\sum_{t=1}^N Q_x^t(y)\rho_1(y)\,dy.$$

In view of (6.4) and integrations by parts,

$$(6.25)\qquad \mathbf{v}(x)=-h^{-N/2-1}\sum_{t=1}^N\int_\Omega \mathbf{G}_x^t(y)\frac{\partial\varphi}{\partial y_i}\,dy=h^{-n/2-1}\sum_{t=1}^N\int_{B_{2\kappa h}(z)}\varphi\frac{\partial\mathbf{G}_x^t(y)}{\partial y_i}\,dy.$$

Differentiating (6.25) and using Hölder's inequality and estimates (3.5), we obtain for multiple index $|\alpha| \le 1 + r$

$$\begin{aligned} \partial_x^\alpha \mathbf{v}(x) &= h^{-N/2-1} \sum_{t=1}^{N} \int_{B_{2\kappa h}(z)} \varphi \partial_{y_i} \partial_x^\alpha \mathbf{G}_x^t(y)\, dy \\ &\le C h^{-N/2-1} d_j^{1-N-|\alpha|} h^{N/2} \|\varphi\|_{L^2(B_{2\kappa h}(z))} \\ &\le C d_j^{-N-r} \|\nabla \varphi\|_{L^2(B_{2\kappa h}(z))} \le C d_j^{-N-r}. \end{aligned} \tag{6.26}$$

Here, we have used the facts that $\|\varphi\|_{H^1(B_{2\kappa h}(z))} = 1$ and $\varphi \in C_0^\infty(B_{2\kappa h}(z))$, which imply

$$\|\varphi\|_{L^2(B_{2\kappa h}(z))} \le Ch \|\nabla \varphi\|_{L^2(B_{2\kappa h}(z))} \le Ch.$$

Thus, integrating (6.26) yields

$$\|\mathbf{v}\|_{H^{1+r}(\Omega_j^{(1)})} \le C d_j^{-N/2-r}. \tag{6.27}$$

With the same procedure, we show

$$\|\lambda\|_{H^r(\Omega_j^{(1)})} \le C d_j^{-N/2-r}. \tag{6.28}$$

Consequently, (6.27) and (6.28) imply (6.23). The proof is complete. ∎

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–722.

[3] D. N. Arnold and X. Liu, *Local error estimates for finite element discretizations of the Stokes equations*, Math. Modeling Numer. Anal., 29 (1995), pp. 367–389.

[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[5] R. G. Durán and R. H. Nochetto, *Weighted inf-sup condition and pointwise error estimates for the Stokes problem*, Math. Comp., 54 (1990), pp. 63–79.

[6] R. G. Durán, R. H. Nochetto, and J. Wang, *Sharp maximum norm error estimates for finite element approximations of the Stokes problem in* 2-*D*, Math. Comp., 51 (1988), pp. 491–506.

[7] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[8] W. Hoffmann, A. H. Schatz, L. B. Wahlbin, and G. Wittum, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes.* I. *A smooth problem and globally quasi-uniform meshes*, Math. Comp., 70 (2001), pp. 897–909.

[9] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, 1969.

[10] F. Natterer, *Über die punktwise konvergenz finiter elemente*, Numer. Math., 25 (1975), pp. 67–77.

[11] J. A. Nitsche, *L_∞ convergence of finite element approximations*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 6060, Springer-Verlag, New York, 1977, pp. 261–274.

[12] J. A. Nitche and A. H. Schatz, *Interior estimates for Ritz-Galerkin methods*, Math. Comp., 28 (1974), pp. 937–958.

[13] F. K. G. Odqvist, *Über die randwertaufgaben der Hydrodynamik zäher flüssigkeiten*, Math. Zs., 32 (1930), pp. 329–375.

[14] R. Rannacher and L. R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.
[15] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* I. *Global estimates*, Math. Comp., 67 (1998), pp. 877–899.
[16] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* II. *Interior estimates*, SIAM J. Numer. Anal., 38 (2000), pp. 1269–1293.
[17] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.
[18] A. H. Schatz and L. B. Wahlbin, *Maximum norm estimates in the finite element method on plane polygonal domains. Part* I, Math. Comp., 32 (1978), pp. 73–109.
[19] R. Scott, *Optimal L_∞ estimates for the finite element methods on irregular grids*, Math. Comp., 30 (1976), pp. 681–697.
[20] V. Solonnikov, *On Green's matrices for elliptic boundary value problems*, I, Proc. Steklov Inst. Math., 110 (1970), pp. 123–170.
[21] L. Stupelis, *Navier-Stokes Equation in Irregular Domains*, Kluwer Academic, Norwell, MA, 1995.
[22] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1984.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 29–40

ON THE LONG-TIME STABILITY OF THE IMPLICIT EULER SCHEME FOR THE TWO-DIMENSIONAL NAVIER–STOKES EQUATIONS*

F. TONE† AND D. WIROSOETISNO†‡

Abstract. In this paper we study the stability for all positive time of the fully implicit Euler scheme for the two-dimensional Navier–Stokes equations. More precisely, we consider the time discretization scheme and with the aid of the discrete Gronwall lemma and the discrete uniform Gronwall lemma we prove that the numerical scheme is stable.

Key words. Navier–Stokes equations, discrete Gronwall lemmas, implicit Euler scheme

AMS subject classifications. 65M12, 76D05

DOI. 10.1137/040618527

1. Introduction.

Let $\Omega \subset \mathbb{R}^2$ be an open bounded set with boundary $\partial\Omega$ of class C^2. The Navier–Stokes equations of viscous incompressible fluids are

$$u_t + (u \cdot \nabla)u - \nu\Delta u + \nabla p = f, \tag{1.1}$$

$$\operatorname{div} u = 0, \tag{1.2}$$

where $u = (u_1, u_2)$ is the velocity, p is the pressure, ν is the kinematic viscosity, and f represents body forces applied to the fluid. We complete these equations with the initial condition

$$u(x, 0) = u_0(x), \tag{1.3}$$

with $u_0 : \Omega \to \mathbb{R}^2$ being given, and with the nonslip boundary condition

$$u = 0 \quad \text{on } \partial\Omega. \tag{1.4}$$

In the notation described below, system (1.1)–(1.4) can be written as the functional evolution equation

$$u_t + \nu Au + B(u, u) = f, \quad u(0) = u_0. \tag{1.5}$$

In the two-dimensional case under consideration, the solution to the Navier–Stokes equations is known to be smooth for all time (cf. [13]). The velocity u is bounded uniformly for all time by

$$|u(t)|^2_{L^2(\Omega)^2} \le \mathrm{e}^{-\nu\lambda_1 t}|u_0|^2_{L^2(\Omega)^2} + c\big(1 - \mathrm{e}^{-\nu\lambda_1 t}\big)|f|^2_{L^\infty(\mathbb{R}_+;L^2(\Omega)^2)}, \tag{1.6}$$

where λ_1 is the first eigenvalue of the Stokes operator A, and we have assumed that $f \in L^\infty(\mathbb{R}_+; L^2(\Omega)^2)$. Furthermore, using techniques based on the uniform Gronwall

*Received by the editors November 8, 2004; accepted for publication (in revised form) May 10, 2005; published electronically February 8, 2006. This work was partially supported by the National Science Foundation under grant NSF-DMS-0305110 and by the Research Fund of Indiana University.
http://www.siam.org/journals/sinum/44-1/61852.html

†The Institute for Scientific Computing and Applied Mathematics, Indiana University, Bloomington, IN 47405-7106 (ftone@indiana.edu).

‡Current address: Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, United Kingdom (djoko.wirosoetisno@durham.ac.uk).

lemma (cf. [12]), one can bound u uniformly in $H_0^1(\Omega)$ for all $t \geq 0$ by a function which depends on the initial condition

$$|u(t)|^2_{H_0^1(\Omega)^2} \leq K(|u_0|_{H_0^1(\Omega)^2}, |f|_{L^\infty(\mathbb{R}_+;L^2(\Omega)^2)}). \tag{1.7}$$

This dependence on the initial data can be dropped when one considers sufficiently large time, $t \geq T_c(|u_0|_{L^2(\Omega)^2}, |f|_{L^\infty(\mathbb{R}_+;L^2(\Omega)^2)})$, giving

$$|u(t)|^2_{H_0^1(\Omega)^2} \leq K(|f|_{L^\infty(\mathbb{R}_+;L^2(\Omega)^2)}) \quad \forall\, t \geq T_c. \tag{1.8}$$

In this paper we consider a time discretization of (1.5) using the fully implicit Euler scheme

$$\frac{u^n - u^{n-1}}{k} + \nu A u^n + B(u^n, u^n) = f^n, \quad u^0 = u_0, \tag{1.9}$$

where

$$f_n = \frac{1}{\Delta t} \int_{(n-1)\Delta t}^{n\Delta t} f(t)\, dt, \tag{1.10}$$

and seek to obtain similar bounds on $|u^n|_{H_0^1(\Omega)^2}$.

Before we proceed further, we note that a related result for the linearized implicit Euler scheme

$$\frac{u^n - u^{n-1}}{k} + \nu A u^n + B(u^{n-1}, u^n) = f^n, \quad u^0 = u_0, \tag{1.11}$$

is proved in [7]. A different approach for the linearized implicit Euler scheme for the case without forcing term appears in [3].

Important background information on different computational methods can be found in some of the books and articles available in the literature. On finite elements, see, e.g., [4], [6]; on finite differences and finite elements, [9], [13]; on spectral methods, [1], [5].

For the mathematical setting of the problem, we consider the following spaces:

$$V = \{v \in H_0^1(\Omega)^2, \operatorname{div} v = 0\}, \tag{1.12}$$

$$H = \{v \in L^2(\Omega)^2, \operatorname{div} v = 0,\ v \cdot n = 0 \text{ on } \partial\Omega\}, \tag{1.13}$$

where n is the unit outward normal on $\partial\Omega$. The space V is endowed with the scalar product

$$((u, v)) = \sum_{i,j=1}^{2} \int_\Omega \frac{\partial u_i}{\partial x_j}(x) \frac{\partial v_i}{\partial x_j}(x)\, dx \tag{1.14}$$

and with the corresponding norm

$$\|u\| = ((u, u))^{1/2}, \tag{1.15}$$

and H is endowed with the scalar product and the norm of $L^2(\Omega)^2$, denoted by $(\cdot, \cdot)$ and $|\cdot|$.

We denote by A the linear continuous operator from V into V' such that

$$\langle Au, v\rangle_{V',V} = ((u, v)) \quad \forall\, u, v \in V. \tag{1.16}$$

The domain of A in H is denoted by $D(A)$ and, using the regularity theory for the Stokes equation (see, for instance, [13]), one can show that

$$D(A) = H^2(\Omega)^2 \cap V. \tag{1.17}$$

We have the following inclusions:

$$D(A) \subset V \subset H, \tag{1.18}$$

and the so-called Poincaré inequality holds true:

$$|u| \le \frac{1}{\sqrt{\lambda_1}} \|u\| \quad \forall\, u \in V, \tag{1.19}$$

where $\lambda_1 > 0$ is the first eigenvalue of the Stokes operator A.

As is well known, the form (1.5) of the Navier–Stokes equations was derived by Leray [8], using the weak formulation of the Navier–Stokes equations. The latter is obtained by multiplying (1.1) by a test function $v \in V$ and integrating by parts over Ω, using Green's formula, viz.,

$$\frac{d}{dt}(u(t), v) + \nu((u(t), v)) + b(u(t), u(t), v) = (f(t), v) \quad \forall\, v \in V, \tag{1.20}$$

where

$$b(u, v, w) = \sum_{i,j=1,2} \int_\Omega u_i(x) \frac{\partial v_j}{\partial x_i}(x) w_j(x)\, dx. \tag{1.21}$$

The form b is trilinear continuous on $H^1(\Omega)^2$ and enjoys the following properties:

$$|b(u, v, w)| \le c_b |u|^{1/2} |Au|^{1/2} \|v\| |w| \quad \forall\, u \in D(A),\ v \in V,\ w \in H, \tag{1.22}$$

$$|b(u, v, w)| \le c_b |u|^{1/2} \|u\|^{1/2} \|v\| |w|^{1/2} \|w\|^{1/2} \quad \forall\, u, v, w \in V, \tag{1.23}$$

$$b(u, v, v) = 0 \quad \forall\, u, v \in V, \tag{1.24}$$

the last equation implying

$$b(u, v, w) = -b(u, w, v) \quad \forall\, u, v, w \in V. \tag{1.25}$$

Using b, we define the bilinear operator B from $V \times V$ into V' by

$$\langle B(u, v), w \rangle_{V',V} = b(u, v, w) \quad \forall\, u, v, w \in V. \tag{1.26}$$

For more details about the functional spaces $D(A)$, V, and H as well as the operators A, B, and b, the reader is referred to, e.g., [2], [11], and [13].

2. H^1 stability and the main result. Throughout the paper, we assume that $f \in L^\infty(\mathbb{R}_+; H)$ and we set $|f|_\infty := |f|_{L^\infty(\mathbb{R}_+;H)}$. We adopt the following convention: c_i denotes constants that depend only on the parameters such as λ_1, ν, etc.; K_i depend in addition on $u(t_*)$ at some specified time t_* and on the forcing f; κ_i are bounds on the timestep k and may depend on u_0 and f.

In proving the main result, we will need a couple of preliminary lemmas. We begin with an analogue of (1.6), proved in almost the same way (see, e.g., [12, p. 109]).

LEMMA 2.1. *For every $k > 0$, we have*

$$|u^n|^2 \le (1+\nu\lambda_1 k)^{-n}|u_0|^2 + [1-(1+\nu\lambda_1 k)^{-n}]\frac{|f|_\infty^2}{\nu^2\lambda_1^2} \quad \forall\, n \ge 0, \tag{2.1}$$

and there exists $K_1 = K_1(|u_0|, |f|_\infty)$ such that

$$|u^n|^2 \le K_1 \quad \forall\, n \ge 0, \tag{2.2}$$

and

$$\nu \sum_{j=i}^{n} k\|u^j\|^2 \le K_1 + (n-i+1)k\,\frac{|f|_\infty^2}{\nu\lambda_1} \quad \forall\, i = 1, \dots, n. \tag{2.3}$$

Proof. Taking the scalar product of (1.9) with $2ku^n$ in H and using the relation

$$2(\varphi-\psi,\varphi) = |\varphi|^2 - |\psi|^2 + |\varphi-\psi|^2 \quad \forall \varphi,\psi \in H, \tag{2.4}$$

and the skew property (1.24), we obtain

$$|u^n|^2 - |u^{n-1}|^2 + |u^n - u^{n-1}|^2 + 2\nu k\|u^n\|^2 = 2k(f^n, u^n). \tag{2.5}$$

Using the Cauchy–Schwarz inequality and the Poincaré inequality (1.19), we majorize the right-hand side of (2.5) by

$$2k|f^n||u^n| \le \frac{2k}{\sqrt{\lambda_1}}|f^n|\|u^n\| \le \nu k\|u^n\|^2 + \frac{k}{\nu\lambda_1}|f^n|^2. \tag{2.6}$$

Relations (2.5) and (2.6) imply

$$|u^n|^2 - |u^{n-1}|^2 + |u^n - u^{n-1}|^2 + \nu k\|u^n\|^2 \le \frac{k}{\nu\lambda_1}|f^n|^2. \tag{2.7}$$

Using again the Poincaré inequality (1.19), we find from (2.7)

$$|u^n|^2 \le \frac{1}{\alpha}|u^{n-1}|^2 + \frac{k}{\alpha\nu\lambda_1}|f^n|^2, \tag{2.8}$$

where

$$\alpha = 1 + \nu\lambda_1 k. \tag{2.9}$$

Using (2.8) recursively, we find

$$\begin{aligned} |u^n|^2 &\le \frac{1}{\alpha^n}|u^0|^2 + \frac{k}{\nu\lambda_1}\sum_{i=1}^{n}\frac{1}{\alpha^i}|f^{n+1-i}|^2 \\ &\le (1+\nu\lambda_1 k)^{-n}\,|u_0|^2 + \frac{|f|_\infty^2}{\nu^2\lambda_1^2}[1-(1+\nu\lambda_1 k)^{-n}], \end{aligned} \tag{2.10}$$

which proves (2.1); (2.1) easily implies (2.2) with

$$K_1(|u_0|, |f|_\infty) := |u_0|^2 + \frac{1}{\nu^2\lambda_1^2}|f|_\infty^2. \tag{2.11}$$

Now adding up (2.7) with n from i to m and dropping some terms, we find

$$\begin{aligned}\nu k \sum_{j=i}^{m} \|u^j\|^2 &\le |u^{i-1}|^2 + \frac{k}{\nu\lambda_1} \sum_{j=i}^{m} |f^j|^2 \\ &\le K_1 + \frac{|f|_\infty^2}{\nu\lambda_1}(m-i+1)k,\end{aligned} \tag{2.12}$$

which is just (2.3) with n in place of m. ∎

Corollary 2.2. *If*

$$0 < k \le \frac{1}{\nu\lambda_1} =: \kappa_1, \tag{2.13}$$

then

$$|u^n|^2 \le 2\rho_0^2 \quad \forall\, nk \ge T_0(|u_0|, |f|_\infty) := \frac{4}{\nu\lambda_1} \ln\left(\frac{|u_0|}{\rho_0}\right), \tag{2.14}$$

where $\rho_0 := |f|_\infty/(\nu\lambda_1)$.

Proof. From the bound (2.1) on $|u^n|^2$, we infer that

$$|u^n|^2 \le \big(1 + \nu\lambda_1 k\big)^{-n} |u_0|^2 + \rho_0^2,$$

and using assumption (2.13) on k and the fact that $1 + x \ge \exp(x/2)$ if $x \in (0,1)$, we obtain

$$|u^n|^2 \le \exp\Big(-nk\frac{\nu\lambda_1}{2}\Big)|u_0|^2 + \rho_0^2.$$

For $nk \ge T_0$, the above inequality implies conclusion (2.14) of the corollary. ∎

We now seek to obtain uniform bounds on u^n in V similar to those obtained in H (see (2.2)). To this end, we first derive bounds on a finite interval of time (see Proposition 2.5). We then repeatedly use these together with (a discrete uniform Gronwall) Lemma 2.6 on successive intervals to arrive at the desired uniform bounds.

We begin with some preliminary inequalities. Taking the scalar product of (1.9) with $2kAu^n$ in H, we obtain

$$\begin{aligned}&\|u^n\|^2 - \|u^{n-1}\|^2 + \|u^n - u^{n-1}\|^2 + 2\nu k|Au^n|^2 \\ &\quad + 2kb(u^n, u^n, Au^n) = 2k(f^n, Au^n).\end{aligned} \tag{2.15}$$

Using property (1.22) of the trilinear form b and recalling (2.2), we have the following bound of the nonlinear term:

$$\begin{aligned}2kb(u^n, u^n, Au^n) &\le 2\,c_b\,k\,|u^n|^{1/2}\|u^n\||Au^n|^{3/2} \\ &\le \frac{\nu k}{2}|Au^n|^2 + \frac{27c_b^4}{2\nu^3}K_1 k\|u^n\|^4.\end{aligned} \tag{2.16}$$

We bound the right-hand side of (2.15) by Cauchy–Schwarz,

$$\begin{aligned}2k(f^n, Au^n) &\le 2k|f^n||Au^n| \\ &\le \frac{\nu k}{2}|Au^n|^2 + \frac{2}{\nu}k|f^n|^2.\end{aligned} \tag{2.17}$$

Relations (2.15)–(2.17) imply

$$\|u^n\|^2 - \|u^{n-1}\|^2 + \|u^n - u^{n-1}\|^2 + \nu k|Au^n|^2 \leq \frac{27c_b^4}{2\nu^3} K_1 k \|u^n\|^4 + \frac{2}{\nu} k|f^n|^2, \tag{2.18}$$

from which we obtain

$$0 \leq c_2 K_1 k \|u^n\|^4 - \|u^n\|^2 + \|u^{n-1}\|^2 + c_3 k|f|_\infty^2\,, \tag{2.19}$$

where

$$c_2 = \frac{27c_b^4}{2\nu^3} \quad \text{and} \quad c_3 = \frac{2}{\nu}. \tag{2.20}$$

Unlike (2.7), (2.19) does not (directly) provide a useful bound for $\|u^n\|$, so we proceed to show that (2.19) does give a proper bound for $\|u^n\|$ if the timestep k is sufficiently small.

LEMMA 2.3. *Suppose that* $0 < k \leq \kappa_1$ *and assume that, for some* n*, we have*

$$c_2 K_1 k \left(K_2 \|u^{n-1}\|^2 + c_4 |f|_\infty^2\right) \leq \frac{1}{5}, \tag{2.21}$$

where $K_2(|u_0|, |f|_\infty) = 2 + 4c_b^2 K_1/\nu^2$ *and* $c_4 = 4/(\nu^2 \lambda_1)$*. Then* (2.19) *implies*

$$\|u^n\|^2 \leq \|u^{n-1}\|^2 \left[1 + c_5 K_1 k \left(\|u^{n-1}\|^2 + k|f|_\infty^2\right)\right] + c_6 k |f|_\infty^2 \tag{2.22}$$

for some constants c_5 *and* c_6*.*

Proof. Relation (2.19) implies either

$$\|u^n\|^2 \leq \frac{1 - \sqrt{\Delta_{n-1}}}{2c_2 K_1 k} \tag{2.23}$$

or

$$\|u^n\|^2 \geq \frac{1 + \sqrt{\Delta_{n-1}}}{2c_2 K_1 k}, \tag{2.24}$$

where

$$\Delta_{n-1} = 1 - 4c_2 K_1 k (\|u^{n-1}\|^2 + c_3 k |f|_\infty^2) > 0 \quad \text{by (2.13) and (2.21).} \tag{2.25}$$

We now show that (2.21) excludes (2.24). Indeed, taking the scalar product of (1.9) with $2k(u^n - u^{n-1})$ in H, we obtain

$$\begin{aligned} &2|u^n - u^{n-1}|^2 + \nu k \|u^n\|^2 - \nu k \|u^{n-1}\|^2 + \nu k \|u^n - u^{n-1}\|^2 \\ &\quad + 2k\, b(u^n, u^n, u^n - u^{n-1}) = 2k\,(f^n, u^n - u^{n-1}). \end{aligned} \tag{2.26}$$

Using properties (1.23), (1.24), and (1.25) of the trilinear form b and recalling (2.2), we bound the nonlinear term as

$$\begin{aligned} 2kb(u^n, u^n, u^n - u^{n-1}) &= -2kb(u^n, u^n, u^{n-1}) \\ &\leq 2c_b k |u^n| \|u^n\| \|u^{n-1}\| \\ &\leq \frac{\nu}{2} k \|u^n\|^2 + \frac{2c_b^2}{\nu} K_1 k \|u^{n-1}\|^2. \end{aligned} \tag{2.27}$$

We bound the right-hand side of (2.26) using Cauchy–Schwarz,

$$\begin{aligned} 2k(f^n, u^n - u^{n-1}) &\le 2k|f^n||u^n - u^{n-1}| \\ &\le \frac{2}{\sqrt{\lambda_1}} k|f^n|\|u^n - u^{n-1}\| \\ &\le \frac{\nu}{2} k\|u^n - u^{n-1}\|^2 + \frac{2}{\nu\lambda_1} k|f^n|^2. \end{aligned} \tag{2.28}$$

Relations (2.26)–(2.28) imply

$$\begin{aligned} 2|u^n - u^{n-1}|^2 + \frac{\nu}{2} k\|u^n\|^2 - \left(\nu + \frac{2c_b^2}{\nu} K_1\right) k\|u^{n-1}\|^2 \\ + \frac{\nu}{2} k\|u^n - u^{n-1}\|^2 \le \frac{2}{\nu\lambda_1} k|f^n|^2, \end{aligned} \tag{2.29}$$

from which we obtain

$$\|u^n\|^2 \le K_2\|u^{n-1}\|^2 + c_4|f|_\infty^2, \tag{2.30}$$

and using hypothesis (2.21) we find

$$2c_2K_1k\|u^n\|^2 \le 2c_2K_1k\left(K_2\|u^{n-1}\|^2 + c_4|f|_\infty^2\right) < 1, \tag{2.31}$$

which contradicts (2.24). Therefore, (2.19) implies (2.23) and hence

$$\begin{aligned} \|u^n\|^2 &\le \frac{1 - \left[1 - 4c_2K_1k\left(\|u^{n-1}\|^2 + c_3k|f|_\infty^2\right)\right]^{1/2}}{2c_2K_1k} \\ &= 2\frac{\|u^{n-1}\|^2 + c_3k|f|_\infty^2}{1 + \sqrt{1-x}}, \end{aligned} \tag{2.32}$$

where

$$x = 4c_2K_1k(\|u^{n-1}\|^2 + c_3k|f|_\infty^2).$$

Since $x \le 4/5$ by (2.21) and

$$\frac{2}{1+\sqrt{1-x}} \le 1 + \frac{x}{2} \quad \text{if} \quad 0 \le x \le \frac{4}{5},$$

relation (2.32) implies, under assumption (2.21), that

$$\|u^n\|^2 \le \left(\|u^{n-1}\|^2 + c_3k|f|_\infty^2\right)\left[1 + 2c_2K_1k\left(\|u^{n-1}\|^2 + c_3k|f|_\infty^2\right)\right]. \tag{2.33}$$

Using (2.21) once again, (2.33) immediately implies (2.22). ∎

In order to obtain estimates on a finite interval of time, we will inductively use Lemma 2.3, together with the following result, which was proved in [10] and which we repeat here for convenience.

LEMMA 2.4. *Given $k > 0$, an integer $n_* > 0$, and positive sequences ξ_n, η_n, and ζ_n such that*

$$\xi_n \le \xi_{n-1}(1 + k\eta_{n-1}) + k\zeta_n \quad \textit{for } n = 1, \dots, n_*, \tag{2.34}$$

we have, for any $n \in \{2, \dots, n_*\}$,

$$\xi_n \le \xi_0 \exp\left(\sum_{i=0}^{n-1} k\eta_i\right) + \sum_{i=1}^{n-1} k\zeta_i \exp\left(\sum_{j=i}^{n-1} k\eta_j\right) + k\zeta_n\,. \tag{2.35}$$

Proof. Using (2.34) recursively, we derive

$$\xi_n \le \xi_0 \prod_{i=0}^{n-1}(1+k\eta_i) + \sum_{i=1}^{n} k\zeta_i \prod_{j=i}^{n-1}(1+k\eta_j)$$

with the convention that $\prod_{j=\alpha}^{\beta} r_j = 1$ for $\beta < \alpha$. Using the fact that $1 + x \le \mathrm{e}^x$ for all $x \in \mathbb{R}$, the conclusion of the lemma follows. □

PROPOSITION 2.5 (estimates on a finite interval). *Let* $T > 0$ *and let* $K_3(\cdot,\cdot,\cdot)$ *be the function, monotonically increasing in all its arguments, given in* (2.47). *Suppose the timestep* k *is such that*

$$k \le \min\{\kappa_1, \kappa_2(|u_0|, |f|_\infty), \kappa_3(\|u^0\|, |f|_\infty, T)\}, \tag{2.36}$$

where κ_1 *is given by* (2.13), *and*

$$\kappa_2(|u_0|, |f|_\infty) = \frac{1}{10c_2c_4K_1|f|_\infty^2}, \tag{2.37}$$

$$\kappa_3(\|u^0\|, |f|_\infty, T) = \frac{1}{10c_2K_1K_2K_3(\|u^0\|, |f|_\infty, T)}. \tag{2.38}$$

Then (i) *relation* (2.22) *holds for all* $n = 1, \dots, N := \lfloor T/k \rfloor$, *and* (ii)

$$\|u^n\|^2 \le K_3(\|u^0\|, |f|_\infty, nk) \quad \forall\, n = 1, \dots, N := \lfloor T/k \rfloor. \tag{2.39}$$

Proof. Let $T > 0$ and k be such that hypothesis (2.36) is satisfied. We will use induction on n.

Since $\|u^0\|^2 \le K_3(\|u^0\|, |f|_\infty, 0)$, (2.37) and (2.38) imply that condition (2.21) of Lemma 2.3 is satisfied for $n = 1$,

$$c_2K_1k(K_2\|u^0\|^2 + c_4|f|_\infty^2) \le \frac{1}{10} + \frac{1}{10} \le \frac{1}{5}. \tag{2.40}$$

By the same lemma, we have

$$\|u^1\|^2 \le \|u^0\|^2\left[1 + c_5K_1k\left(\|u^0\|^2 + k|f|_\infty^2\right)\right] + c_6k|f|_\infty^2. \tag{2.41}$$

Now assume that (2.21) holds for $n = 1, \dots, m$ for some $m \le N$. Then by Lemma 2.3, (2.22) holds for $n = 1, \dots, m$; furthermore, we can bound $\|u^m\|$ as follows. We write the stepwise bound (2.22) in Lemma 2.3 in the form

$$\xi_n \le \xi_{n-1}(1 + k\eta_{n-1}) + k\zeta, \tag{2.42}$$

where

$$\xi_n = \|u^n\|^2, \quad \eta_n = c_5K_1(\|u^n\|^2 + k|f|_\infty^2), \quad \text{and} \quad \zeta = c_6|f|_\infty^2. \tag{2.43}$$

Our intention is to apply (the discrete Gronwall) Lemma 2.4. So we compute for $i > 0$, using (2.3),

$$\sum_{j=i}^{m-1} k\eta_j = c_5 K_1 \sum_{j=i}^{m-1} k\left(\|u^j\|^2 + k|f|_\infty^2\right) \le c_7 K_1\big[K_1 + (m-i)k|f|_\infty^2\big]; \tag{2.44}$$

similarly, for $i = 0$,

$$\sum_{j=0}^{m-1} k\eta_j = c_5 K_1 \sum_{j=0}^{m-1} k\left(\|u^j\|^2 + k|f|_\infty^2\right) \le c_7 K_1\big(K_1 + mk|f|_\infty^2\big) + c_5 K_1 k\|u^0\|^2. \tag{2.45}$$

We note that, using (2.38) and recalling that $K_2 \ge 2$, the last term can be bounded as

$$c_5 K_1 k\|u^0\|^2 \le \frac{c_5\|u^0\|^2}{10c_2K_2K_3(\|u^0\|,|f|_\infty,T)} \le \frac{c_5}{10c_2K_2}\frac{\|u^0\|^2}{K_3(\|u^0\|,|f|_\infty,0)} \le \frac{c_5}{20c_2}. \tag{2.46}$$

The middle term in (2.35) here is

$$\sum_{i=1}^{m-1} k\zeta\,\exp\left(\sum_{j=i}^{m-1} k\eta_j\right) \le c_6|f|_\infty^2 \sum_{i=1}^{m-1} k\exp(c_7K_1^2 + c_7K_1(m-i)k|f|_\infty^2) \le c_6|f|_\infty^2\exp(c_7K_1^2)mk\;\exp(c_7K_1\,mk\,|f|_\infty^2).$$

The following bound on $\|u^m\|^2$ then follows from (2.35):

$$\begin{aligned}\|u^m\|^2 &\le \|u^0\|^2\exp(c_7K_1\,|f|_\infty^2\,mk)\;\exp(c_7K_1^2 + c_5/(20c_2))\\ &\quad + 2c_6|f|_\infty^2\exp(c_7K_1^2)\,mk\;\exp(c_7K_1\,|f|_\infty^2\,mk)\\ &=: K_3(\|u^0\|,|f|_\infty,mk).\end{aligned} \tag{2.47}$$

We note that the bound K_3 depends on the initial discrete value through its norm $\|u^0\|$ and also on m, but this latter dependence is only through the time mk. We also note the dependence of K_3 on $|u_0|$ through K_1, but K_1 bounds all $|u^n|^2$.

It is now clear that, given the hypothesis of the proposition, the timestep k satisfies condition (2.21) as long as $m \le \lfloor T/k \rfloor$, completing the proof. ∎

Now, since Proposition 2.5 gives a bound on $\|u^n\|^2$ that is valid on a finite time interval only, we are going to extend the result to infinite time by repeatedly applying it and the following (discrete uniform Gronwall) lemma, which is a slightly more general version of the discrete uniform Gronwall lemma of Shen [10].

LEMMA 2.6. *Given $k > 0$, positive integers n_1, n_2, n_* such that $n_1 < n_*$, $n_1 + n_2 + 1 \le n_*$, positive sequences ξ_n, η_n, and ζ_n such that*

$$\xi_n \le \xi_{n-1}(1 + k\eta_{n-1}) + k\zeta_n \quad \textit{for } n = n_1, \dots, n_*, \tag{2.48}$$

and given the bounds

$$\sum_{n=n'}^{n'+n_2} k\eta_n \le a_1(n_1,n_*), \quad \sum_{n=n'}^{n'+n_2} k\zeta_n \le a_2(n_1,n_*), \quad \sum_{n=n'}^{n'+n_2} k\xi_n \le a_3(n_1,n_*) \tag{2.49}$$

for any n' satisfying $n_1 \le n' \le n_* - n_2$, we have

$$\xi_n \le \left(\frac{a_3(n_1,n_*)}{kn_2} + a_2(n_1,n_*) \right) \mathrm{e}^{a_1(n_1,n_*)} \tag{2.50}$$

for any n such that $n_1 + n_2 + 1 \le n \le n_*$.

Proof. Let n_3 and n_4 be such that $n_1 \le n_3 - 1 \le n_4 \le n_2 + n_3 - 1 \le n_* - 1$. Using (2.48) recursively, we derive

$$\xi_{n_2+n_3} \le \xi_{n_4} \prod_{i=n_4}^{n_3+n_2-1} (1+k\eta_i) + \sum_{i=n_4+1}^{n_3+n_2} k\zeta_i \prod_{j=i}^{n_2+n_3-1} (1+k\eta_j) \tag{2.51}$$

with the convention that $\prod_{j=\alpha}^{\beta} r_j = 1$ for $\beta < \alpha$. Using the fact that $1 + x \le \mathrm{e}^x$ for all $x \in \mathbb{R}$, and recalling the first two assumptions in (2.49), we obtain

$$\xi_{n_2+n_3} \le (\xi_{n_4} + a_2) e^{a_1}.$$

Multiplying this inequality by k, summing n_4 from $n_3 - 1$ to $n_2 + n_3 - 2$, and using the third assumption in (2.49) gives the conclusion (2.50) of the lemma. ∎

We are now in a position to give the main result, that is, to derive a uniform bound for $\|u^n\|$ for all $n \ge 1$.

THEOREM 2.7. *Let $u_0 \in V$, $f \in L^\infty(\mathbb{R}_+; H)$, and u^n be the solution of the numerical scheme* (1.9)*. Also, let $r \ge 4\kappa_1$ be arbitrarily fixed and let k be such that*

$$k \le \min\{\kappa_1, \kappa_2(|u_0|, |f|_\infty), \kappa_3(\|u_0\|, |f|_\infty, T_0 + r), \kappa_3(\rho_1, |f|_\infty, r)\}, \tag{2.52}$$

where $\kappa_1 = 1/(\nu\lambda_1)$ was defined in (2.13)*, $\kappa_2(\cdot,\cdot)$ and $\kappa_3(\cdot,\cdot,\cdot)$ are given in Proposition* 2.5*, T_0, the time of entering an absorbing ball for $|u^n|$, is given by* (2.14)*, and $\rho_1(|f|_\infty, r)$ is given in* (2.57)*.*

Then we have

$$\|u^n\|^2 \le K_5(\|u_0\|, |f|_\infty) \quad \forall\, n \ge 1, \tag{2.53}$$

where $K_5(\cdot,\cdot)$ is a continuous function defined on $\mathbb{R}^2_+$, increasing in both arguments. Moreover,

$$\|u^n\|^2 \le K_4(|f|_\infty) \quad \forall\, n \ge N_0 + N_r := \lfloor T_0/k \rfloor + \lfloor r/k \rfloor, \tag{2.54}$$

i.e., $\|u^n\|$ is bounded independently of u_0 beyond $N_0 + N_r$.

Proof. Let $r \ge 4\kappa_1$ be arbitrarily fixed and let k be such that (2.52) holds.

The idea for deriving a uniform bound for $\|u^n\|^2$ for all $n \ge 1$ is as follows:

(i) Applying first Proposition 2.5 on $(0, T_0 + r)$ (that is, for $n = 1, \ldots, N_0 + N_r$), we get an upper bound for $\|u^n\|$ for $n = 1, \ldots, N_0 + N_r$; applying Lemma 2.6, we show that $\|u^{N_0+N_r}\|^2 \le \rho_1^2$, where $\rho_1(|f|_\infty, r)$ is defined in (2.57).

(ii) Iterating Proposition 2.5 and Lemma 2.6, at each step $i \ge 2$, we show that for all $n = N_0 + (i-1)N_r + 1, \ldots, N_0 + iN_r$, $\|u^n\|^2$ is bounded by $K_3(\|u^{N_0+(i-1)N_r}\|,$

$|f|_\infty, r)$; using the estimate on $\|u^{N_0+(i-1)N_r}\|$ from the previous step, we obtain that $\|u^n\|^2$ is bounded independently of the initial value for all $n = N_0 + (i-1)N_r + 1, \dots, N_0 + iN_r$ for every $i \geq 2$ (and thus for all $n \geq N_0 + N_r$).

We now proceed to give a rigorous proof of the theorem.

Noting that, by hypothesis, k satisfies condition (2.36) of Proposition 2.5 with $T = T_0 + r$, we first apply Proposition 2.5 and obtain that (2.22) holds for all $n = 1, \dots, N_0 + N_r$, and

$$\|u^n\|^2 \leq K_3(\|u^0\|, |f|_\infty, nk) \quad \forall\, n = 1, \dots, N_0 + N_r. \tag{2.55}$$

At this point we know that for k satisfying hypothesis (2.52),

$$\|u^n\|^2 \leq \|u^{n-1}\|^2[1 + c_5K_1k(\|u^{n-1}\|^2 + k|f|_\infty^2)] + c_6k|f|_\infty^2 \quad \forall\, n = 1, \dots, N_0 + N_r, \tag{2.56}$$

and we apply (the discrete uniform Gronwall) Lemma 2.6 with $\xi_n = \|u^n\|^2$, $\eta_n = c_5K_1(\|u^n\|^2 + k|f|_\infty^2)$, $\zeta_n = c_6|f|_\infty^2$, $n_1 = N_0 + 1$, $n_2 = N_r - 2$, and $n_* = N_0 + N_r$ to obtain a bound for $\|u^{N_0+N_r}\|$. In computing the sums $a_1(n_1, n_*)$, $a_2(n_1, n_*)$, and $a_3(n_1, n_*)$ that appear there, we note that since all those sums are taken for $n \geq N_0$ and since, by hypothesis, k satisfies condition (2.13) of Corollary 2.2, we can replace K_1, the bound on $|u^n|^2$, by $2\rho_0^2$, whenever the former appears. For every $n' = N_0 + 1, N_0 + 2$, we compute, using (2.3) and (2.14) for the first and last lines,

$$2c_5\rho_0^2 \sum_{n=n'}^{n'+n_2} \big(k\|u^n\|^2 + k^2|f|_\infty^2\big) \leq c_8\rho_0^2(\rho_0^2 + r|f|_\infty^2),$$

$$c_6 \sum_{n=n'}^{n'+n_2} k|f|_\infty^2 \leq c_6\, r|f|_\infty^2,$$

$$\sum_{n=n'}^{n'+n_2} k\|u^n\|^2 \leq c_9(\rho_0^2 + r|f|_\infty^2).$$

Using the conclusion (2.50) of Lemma 2.6 and the fact that $r \geq 4\kappa_1$, we obtain

$$\begin{aligned} \|u^{N_0+N_r}\|^2 &\leq \big[2c_9\big(\rho_0^2/r + |f|_\infty^2\big) + c_6\, r|f|_\infty^2\big] \exp\big(c_8\rho_0^2(\rho_0^2 + r|f|_\infty^2)\big) \\ &=: \rho_1(|f|_\infty; r)^2. \end{aligned} \tag{2.57}$$

Now, since by hypothesis $k \leq \kappa_3(\rho_1, |f|_\infty, r)$ and since $\kappa_3(\cdot,\cdot,\cdot)$ is a decreasing function of its arguments, we can regard $u^{N_0+N_r}$ as our initial data and apply Proposition 2.5 with $T = r$. We obtain that relation (2.22) holds for all $n = N_0 + N_r + 1, \dots, N_0 + 2N_r$, and

$$\|u^n\|^2 \leq K_3(\|u^{N_0+N_r}\|, |f|_\infty, N_rk) \quad \forall\, n = N_0 + N_r + 1, \dots, N_0 + 2N_r. \tag{2.58}$$

Thanks to (2.57) and to the fact that $K_3(\cdot,\cdot,\cdot)$ is an increasing function of all its arguments, we have

$$\|u^n\|^2 \leq K_3(\rho_1, |f|_\infty, N_rk) \quad \forall\, n = N_0 + N_r + 1, \dots, N_0 + 2N_r. \tag{2.59}$$

Applying again Lemma 2.6 with $n_1 = N_0 + N_r + 1$, $n_2 = N_r - 2$, and $n_* = N_0 + 2N_r$, we obtain

$$\|u^{N_0+2N_r}\|^2 \leq \rho_1^2\,. \tag{2.60}$$

Iterating Proposition 2.5 and Lemma 2.6 and reasoning as above, we arrive at

$$\|u^n\|^2 \le K_3(\rho_1, |f|_\infty, r) =: K_4(|f|_\infty) \quad \forall\, n \ge N_0 + N_r, \tag{2.61}$$

and recalling (2.55), we conclude

$$\begin{aligned} \|u^n\|^2 &\le \max\{K_3(\|u_0\|, |f|_\infty, T_0 + r), K_4(|f|_\infty)\} \\ &=: K_5(\|u_0\|, |f|_\infty) \quad \forall\, n \ge 1, \end{aligned} \tag{2.62}$$

thus proving the theorem. □

Acknowledgments. The authors thank Prof. R. Temam for suggesting the problem and for his help in the course of this work; they also thank the referees for helpful comments.

REFERENCES

[1] C. Bernardi and Y. Maday, *Approximations Spectrales de Problèmes aux Limites Elliptiques*, Math. Appl. (Berlin) 10, Springer-Verlag, Paris, 1992.

[2] C. Foias, O. Manley, R. Rosa, and R. Temam, *Navier–Stokes Equations and Turbulence*, Encyclopedia Math. Appl. 83, Cambridge University Press, Cambridge, UK, 2001.

[3] T. Geveci, *On the convergence of a time discretization scheme for the Navier–Stokes equations*, Math. Comp., 53 (1989), pp. 43–53.

[4] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, New York, 1979.

[5] D. Gottlieb and S. Orszag, *Numerical Analysis of Spectral Methods, Theory and Applications*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 26, SIAM, Philadelphia, 1977.

[6] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier-Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[7] N. Ju, *On the global stability of a temporal discretization scheme for the Navier-Stokes equations*, IMA J. Numer. Anal., 22 (2002), pp. 577–597.

[8] J. Leray, *Etude de diverses équations intégrales non linéaires et de quelques problèmes que pose l'hydrodynamique*, J. Math. Pures Appl., 12 (1933), pp. 1–82.

[9] M. Marion and R. Temam, *Navier–Stokes equations. Theory and approximation*, in Handbook of Numerical Analysis, Vol. VI, North-Holland, Amsterdam, 1998, pp. 503–688.

[10] J. Shen, *Long time stabilities and convergences for the fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.

[11] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, 2nd ed., CBMS-NSF Regional Conf. Ser. in Appl. Math. 66, SIAM, Philadelphia, 1995.

[12] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Appl. Math. Sci. 68, Springer-Verlag, New York, 1997.

[13] R. Temam, *Navier–Stokes Equations: Theory and Numerical Analysis*, AMS Chelsea, Providence, RI, 2001 (reprint of the 1984 edition).

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 41–54

ON THE PROBLEM OF TESTING THE STRUCTURE OF A MATRIX BY DISPLACEMENT OPERATIONS*

ALBRECHT BÖTTCHER†

Abstract. This paper deals with the problem of testing whether a large matrix X has a prescribed structure by looking at the magnitude of a displacement matrix $D(X)$ associated with the structure. We provide parameters on the basis of which one can judge whether the problem is well-conditioned or ill-conditioned. It turns out that even for very general structures it is the minimal eigenvalues of positive definite and banded Toeplitz matrices that are the most important of these parameters.

Key words. structured matrix, displacement matrix, conditioning, computer verification, Toeplitz matrix, Toeplitz-plus-Hankel, extreme eigenvalues

AMS subject classifications. 15A24, 47B35, 65F22, 65Q05

DOI. 10.1137/040620035

1. Introduction.

Assume our machine has computed and stored an $n \times n$ matrix $X = (x_{ij})_{i,j=1}^n$ and we want to know whether it is a Toeplitz matrix. We could proceed as follows. We let U be the $n \times n$ forward-shift matrix,

$$U = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

and we compute the so-called displacement matrix $XU - UX$, which equals

$$\begin{pmatrix} x_{12} & x_{13} & \cdots & x_{1,n-1} & 0 \\ x_{22} - x_{11} & x_{23} - x_{12} & \cdots & x_{2n} - x_{1,n-1} & -x_{1n} \\ x_{32} - x_{21} & x_{33} - x_{22} & \cdots & x_{3n} - x_{2,n-1} & -x_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n2} - x_{n-1,1} & x_{n3} - x_{n-1,2} & \cdots & x_{nn} - x_{n-1,n-1} & -x_{n-1,n} \end{pmatrix}.$$

(This and more general displacement matrices $XU - VX$ were employed in [6, 9] for other purposes.) Let $D(X)$ denote the lower-left $(n-1) \times (n-1)$ submatrix of $XU - UX$. Clearly, X is Toeplitz if and only if $D(X)$ is the zero matrix. As testing whether a numerically computed quantity equals zero is a critical issue, we test whether $D(X)$ is small, say whether $\|D(X)\|_2 < \varepsilon$, where $\|\cdot\|_2$ denotes the Frobenius norm. Does this imply that X is close to a Toeplitz matrix?

We will show that the answer is in the negative theoretically but gives rise to optimism practically.

*Received by the editors December 2, 2004; accepted for publication (in revised form) May 13, 2005; published electronically February 8, 2006.
http://www.siam.org/journals/sinum/44-1/62003.html
†Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany (aboettch@mathematik.tu-chemnitz.de).

Let us consider an explicit example. Put $\omega_n = \exp(2\pi i/n)$, $x_j = \omega_n^j$, and $X = \operatorname{diag}(x_1, \dots, x_n)$. Then

$$\begin{aligned}\|D(X)\|_2^2 &= |x_1 - x_2|^2 + |x_2 - x_3|^2 + \cdots + |x_{n-1} - x_n|^2 \\ &= (n-1)|\omega_n - 1|^2 = 4(n-1)\sin^2\frac{\pi}{n}.\end{aligned}$$

Let $\mathcal{T}_n$ be the set of all $n \times n$ Toeplitz matrices. It is easily seen that

$$\operatorname{dist}_2^2(X, \mathcal{T}_n) := \min_{T \in \mathcal{T}_n} \|X - T\|_2^2 = \sum_{j=1}^n \left| x_j - \frac{1}{n}\sum_{k=1}^n x_k \right|^2.$$

Since $\sum x_k = 0$, it follows that $\operatorname{dist}_2^2(X, \mathcal{T}_n) = \sum |x_j|^2 = n$. Consequently, if n is large, then $\|D(X)\|_2$ is small and $\operatorname{dist}_2(X, \mathcal{T}_n)$ is large. This is what we mean by saying that theoretically the answer to the above question is in the negative.

On the other hand, we will prove the following two theorems. Let $\mathbf{K}$ stand for $\mathbf{R}$ or $\mathbf{C}$ and $M_n(\mathbf{K})$ for the $n \times n$ matrices with entries in $\mathbf{K}$. We equip $M_n(\mathbf{K})$ with the Frobenius norm. Further let $\mathcal{T}_n(\mathbf{K})$ denote the set of all Toeplitz matrices in $M_n(\mathbf{K})$. The probability of an event E will be denoted by $P(E)$.

THEOREM 1.1. *We have*

$$\max_{X \notin \mathcal{T}_n(\mathbf{K})} \frac{\operatorname{dist}_2(X, \mathcal{T}_n(\mathbf{K}))}{\|D(X)\|_2} = \frac{1}{2\sin\frac{\pi}{2n}} \sim \frac{n}{\pi}.$$

THEOREM 1.2. *Take X randomly from the unit sphere of $M_n(\mathbf{K})$ with the uniform distribution. Put $\operatorname{dist}_2(X, \mathcal{T}_n(\mathbf{K}))/\|D(X)\|_2 = 0$ if $\|D(X)\|_2 = 0$. Then*

$$P\left(\frac{\operatorname{dist}_2(X, \mathcal{T}_n(\mathbf{K}))}{\|D(X)\|_2} > 10\right) < \frac{13}{n^2} \quad \textit{for} \quad n \geq 10.$$

Theorem 1.1 reveals that if $\|D(X)\|_2 = \varepsilon$, then $\operatorname{dist}_2(X, \mathcal{T}_n(\mathbf{K}))$ is at most about $n\varepsilon/\pi$. This linear growth prevents $n\varepsilon/\pi$ from becoming an astronomic number if ε and n are appropriately adapted. Moreover, Theorem 1.2 tells us that the worst-case situation of Theorem 1.1 is a very rare event for matrices of large sizes. These two conclusions make precise our statement that practically and optimistically the answer to the question raised above is in the affirmative.

The paper is organized as follows. In section 2, we pose and study the problem in the general setting. Section 3 is devoted to what we call string structures. Several concrete string structures are then examined in sections 4–7. In sections 8 and 9, we tackle the Toeplitz-plus-Hankel structure with two different tools. The conclusions are formulated in section 10.

2. The general setting. We assume that the matrix structure we are interested in can be characterized by at most n^2 linear equations for the entries of the matrix. Thus, let $D : M_n(\mathbf{K}) \to M_n(\mathbf{K})$ be a linear operator. We put $\operatorname{Ker} D = \{Y \in M_n(\mathbf{K}) : D(Y) = 0\}$. Given $X \in M_n(\mathbf{K})$, we define $\operatorname{dist}_2(X, \operatorname{Ker} D)$ as $\min_{Y \in \operatorname{Ker} D} \|X - Y\|_2$. We want to determine

$$\max_{X \notin \operatorname{Ker} D} \frac{\operatorname{dist}_2(X, \operatorname{Ker} D)}{\|D(X)\|_2}, \tag{2.1}$$

where $X \notin \operatorname{Ker} D$ is an abbreviation for $X \in M_n(\mathbf{K}) \setminus \operatorname{Ker} D$. Notice that in section 1, $D(X)$ was an $(n-1) \times (n-1)$ matrix. Extending this matrix by a zero row and a zero column we obtain an $n \times n$ matrix, which puts us into the present context.

It will be convenient to change language slightly. Namely, let $i_1 : M_n(\mathbf{K}) \to \mathbf{K}^{n^2}$ and $i_2 : M_n(\mathbf{K}) \to \mathbf{K}^{n^2}$ be two stackings of the entries of matrices in $M_n(\mathbf{K})$ to columns of length n. The concrete choice of i_1 and i_2 may depend on D. Clearly, there is a unique linear operator $\nabla : \mathbf{K}^{n^2} \to \mathbf{K}^{n^2}$ such that $D = i_2^{-1} \circ \nabla \circ i_1$. We freely identify ∇ with an $n^2 \times n^2$ matrix. The Frobenius norm on $M_n(\mathbf{K})$ becomes the ℓ^2 norm $\|\cdot\|$ on $\mathbf{K}^{n^2}$, and for $x \in \mathbf{K}^{n^2}$ and a closed subset F of $\mathbf{K}^{n^2}$ we define $\operatorname{dist}(x, F) = \min_{f \in F} \|x - f\|$. Obviously, (2.1) coincides with

$$\max_{x \notin \operatorname{Ker} \nabla} \frac{\operatorname{dist}(x, \operatorname{Ker} \nabla)}{\|\nabla x\|}. \tag{2.2}$$

For example, to test whether $X = (x_{ij})_{i,j=1}^3$ is of the form

$$\begin{pmatrix} a & b & a \\ b & a & c \\ a & d & a \end{pmatrix} \tag{2.3}$$

with certain numbers $a, b, c, d \in \mathbf{K}$, we may compute $\|D(X)\|_2^2$ with

$$D(X) = \begin{pmatrix} x_{11} - x_{13} & x_{13} - x_{22} & x_{22} - x_{31} \\ x_{31} - x_{33} & 0 & x_{12} - x_{21} \\ 0 & 0 & 0 \end{pmatrix}$$

or $\|\nabla x\|^2$ with

$$\nabla x = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{13} \\ x_{22} \\ x_{31} \\ x_{33} \\ x_{12} \\ x_{21} \\ x_{23} \\ x_{32} \end{pmatrix}.$$

The following result is well known. It follows from the equality $I - P_{\operatorname{Ker} D} = D^+ D$, where $P_{\operatorname{Ker} D}$ is the orthogonal projection onto $\operatorname{Ker} D$ and D^+ is the Moore–Penrose inverse of D. For the reader's convenience, we cite it with a self-contained proof.

THEOREM 2.1. *If $s_{\min}^+$ is the smallest nonzero singular value of D, then*

$$\max_{X \notin \operatorname{Ker} D} \frac{\operatorname{dist}_2(X, \operatorname{Ker} D)}{\|D(X)\|_2} = \frac{1}{s_{\min}^+}. \tag{2.4}$$

Proof. We consider ∇ instead of D. Put $N = n^2$. Suppose ∇ has exactly k zero singular values and let $s_{k+1} \le \cdots \le s_N$ be the nonzero singular values. Denote by P_k and Q_k the projections on $\mathbf{K}^N$ that replace, respectively, the last $N - k$ and first k coordinates with zero. Let $\nabla = USV^*$ with $S = \operatorname{diag}(0, \ldots, 0, s_{k+1}, \ldots, s_N)$ be the singular value decomposition. Obviously, $\operatorname{Ker} \nabla = \operatorname{Ran} V P_k$, $\operatorname{Ran} \nabla^* = \operatorname{Ran} V Q_k$.

For $x \in \mathbf{K}$, put $z = V^*x$. Then $x = VP_kz + VQ_kz$ is the decomposition of x corresponding to the orthogonal decomposition $\mathbf{K}^N = \operatorname{Ker}\nabla \oplus \operatorname{Ran}\nabla^*$. It follows that $\operatorname{dist}^2(x, \operatorname{Ker}\nabla) = \|VQ_kz\|^2 = \|Q_kz\|^2$ and $\|\nabla x\|^2 = \|USVV^*Q_z\|^2 = \|SQ_kz\|^2$, which shows that (2.2) is $\|Q_kz\|/\|SQ_kz\| \le 1/s_{k+1}$, with equality for $x = V^*e_{k+1}$, where e_{k+1} is the column whose $(k+1)$st entry is 1 and whose remaining entries are zero. □

THEOREM 2.2. *Consider the random variable $\xi = \|D(X)\|^2/\operatorname{dist}_2^2(X, \operatorname{Ker}D)$, where X is drawn from the uniform distribution on the unit sphere of $M_n(\mathbf{K})$ with the ℓ^2 norm and $\|D(X)\|^2/\operatorname{dist}_2^2(X, \operatorname{Ker}D) := +\infty$ for $X \in \operatorname{Ker}D$. Suppose $\dim\operatorname{Ker}D = k < n^2$. Then the expected value $E\xi$ and the variance $\sigma^2\xi$ satisfy*

$$E\xi = \frac{\|D\|_2^2}{n^2-k}, \quad \sigma^2\xi \le \frac{2\,\|D\|_\infty^4}{n^2-k},$$

respectively, where $\|\cdot\|_\infty$ denotes the spectral norm.

Proof outline. We proceed as in [3], where the result was proved with $\|X\|_2^2$ and n^2 in place of $\operatorname{dist}_2^2(x, \operatorname{Ker}D)$ and n^2-k, respectively. Thus, we turn again to ∇, put $N = n^2$, and use the singular value distribution $\nabla = USV^*$ from the proof of Theorem 2.1. We have to compute $E(\xi^j)$ for $j = 1, 2$. For the sake of simplicity, suppose $\mathbf{K} = \mathbf{R}$. With S_N and B_N denoting the unit sphere and the unit ball of $\mathbf{R}^N$,

$$\begin{aligned}
E(\xi^j) &= \frac{1}{|S_N|}\int_{S_N} \frac{\|\nabla x\|^{2j}}{\operatorname{dist}^{2j}(x, \operatorname{Ker}\nabla)}\,d\sigma(x) = \frac{1}{|B_N|}\int_{B_N} \frac{\|\nabla x\|^{2j}}{\operatorname{dist}^{2j}(x, \operatorname{Ker}\nabla)}\,dx \\
&= \frac{1}{|B_N|}\int_{B_N} \frac{\|SV^*x\|^{2j}}{\operatorname{dist}^{2j}(x, \operatorname{Ran}VP_k)}\,dx = \frac{1}{|B_N|}\int_{B_N} \frac{\|Sy\|^{2j}}{\operatorname{dist}^{2j}(y, \operatorname{Ran}P_k)}\,dy \\
&= \frac{1}{|B_N|}\int_{B_k}\left(\int_{B_{N-k}(r_k)} \frac{(s_{k+1}^2y_{k+1}^2 + \cdots + s_N^2y_N^2)^j}{(y_{k+1}^2 + \cdots + y_N^2)^j}\,dy_{k+1}\ldots dy_N\right)dy_1\ldots dy_k,
\end{aligned}$$

where $r_k = \sqrt{1 - y_1^2 - \cdots - y_k^2}$ and $B_{N-k}(r_k)$ is the ball with the radius r_k. After the substitution $y_i = r_kz_i$ for $i = k+1, \ldots, N$, the inner integral becomes an integral over the ball of radius 1, and this integral was computed in [3]. We are left with

$$\int_{B_k} (1 - y_1^2 - \cdots - y_k^2)^{(N-k)/2} dy_1\ldots dy_k,$$

which can be shown to be $\pi^{k/2}\Gamma((N-k)/2+1)/\Gamma(N/2+1)$ (see [4, No. 676.8(a)] or [13, No. 3.3.2.1]). Putting all pieces together we arrive at the asserted formulas for $E\xi$ and $\sigma^2\xi = E(\xi^2) - (E\xi)^2$. □

COROLLARY 2.3. *Under the hypothesis of Theorem 2.2,*

$$P\left(\frac{\operatorname{dist}_2(X, \operatorname{Ker}D)}{\|D(X)\|_2} > \frac{1}{\varepsilon}\right) \le \frac{2(n^2-k)\|D\|_\infty^4}{(\|D\|_2^2 - (n^2-k)\varepsilon^2)^2}$$

whenever $0 < \varepsilon^2 < \|D\|_2^2/(n^2-k)$.

Proof. The probability in question is $P(\xi < \varepsilon^2) \le P(|\xi - E\xi| > E\xi - \varepsilon^2)$, and Chebyshev's inequality along with Theorem 2.2 shows that the last probability does not exceed

$$\frac{\sigma^2\xi}{(E\xi - \varepsilon^2)^2} \le \frac{2}{n^2-k}\,\frac{\|D\|_\infty^4}{(\|D\|_2^2/(n^2-k) - \varepsilon^2)^2}.$$ □

3. String structures. We start with a partition

$$\{1,\dots,n\}\times\{1,\dots,n\} = L_1\cup\cdots\cup L_M \tag{3.1}$$

of the index set into pairwise disjoint sets $L_1,\dots,L_M$, called strings, and an ordering of the elements of each string. The number of elements in L_m will be denoted by ℓ_m. We further associate a polynomial

$$a_m(x) = a_0^{(m)} + a_1^{(m)}x + \cdots + a_{r_m}^{(m)}x^{r_m} \tag{3.2}$$

with each string and require that $0\le r_m\le \ell_m - 1$. For a matrix $X=(x_{ij})_{i,j=1}^n$, we label the entries with indices in L_m by $x_1^{(m)},\dots,x_{\ell_m}^{(m)}$ (following the ordering of L_m). We say that X has the string structure specified by (3.1) and (3.2) (and the selected orderings in the strings) if the entries of X satisfy the difference equations

$$a_0^{(m)}x_k^{(m)} + a_1^{(m)}x_{k+1}^{(m)} + \cdots + a_{r_m}^{(m)}x_{k+r_m}^{(m)} = 0 \quad (k=1,\dots,\ell_m - r_m) \tag{3.3}$$

for each m. To test whether a given matrix X has this structure, we compute

$$\sum_{m=1}^{M}\sum_{k=1}^{\ell_m - r_m}\left|a_0^{(m)}x_k^{(m)} + a_1^{(m)}x_{k+1}^{(m)} + \cdots + a_{r_m}^{(m)}x_{k+r_m}^{(m)}\right|^2 \tag{3.4}$$

and check whether this is smaller than ε^2. The number of left-hand sides of (3.3) (= the number of terms in sum (3.4)) is $\sum_{m=1}^M(\ell_m - r_m) = n^2 - \sum_{m=1}^M r_m \le n^2$. We denote by $D : M_n(\mathbf{K})\to M_n(\mathbf{K})$ any linear operator that computes the left-hand sides of (3.3) and arranges them in an $n\times n$ matrix, setting the remaining entries of the matrix zero if $\sum r_m\ge 1$. Clearly, (3.4) is just $\|D(X)\|_2^2$.

Example 3.1. Let $L_m = \{(i,j) : i-j = m\}$, $m = -(n-1),\dots,n-1$. The circumstance that the strings are not labeled from 1 to $2n-1$ but from $-(n-1)$ to $n-1$ clearly causes no problems. We order the indices (i,j) in L_m by increasing i. The structure obtained in this way requires that the entries of a matrix satisfy a difference equation along each diagonal of the matrix. For $r_m = 0$ and $a_0^{(m)} = 0$, this is no requirement. If $r_m = 0$ and $a_0^{(m)} = 1$ for all m, then the only matrix with the structure is the zero matrix. In the case where $r_m = 1$ and $a_m(x) = 1-x$ for all m, we arrive at the set of all Toeplitz matrices. In the case where $r_m = 2$ and $a_m(x) = 1-2x+x^2$ for all m, we have the set of all matrices whose entries on each diagonal form an arithmetic progression. The case where $r_m = 1$ and $a_m(x) = \alpha_m - x$ for all m corresponds to the matrices whose entries on the mth diagonal are $c_m, c_m\alpha_m,\dots,c_m\alpha_m^{n-|m|-1}$ with some $c_m\in\mathbf{K}$.

Example 3.2. The choice $L_m = \{(i,j) : i+j = m\}$, $m = 2,\dots,2n$, produces structures of the Hankel type. With $a_m(x) = 1-x$ for all m, we obtain the pure Hankel matrices. Letting $L_1 = \{(i,j) : i+j \text{ is odd}\}$ and $L_2 = \{(i,j) : i+j \text{ is even}\}$ we get chessboard structures. Pure chessboard matrices result from $a_1(x) = a_2(x) = 1-x$. The class of matrices of the form cI is characterized by the partition

$$\{1,\dots,n\}\times\{1,\dots,n\} = \{(1,1),\dots,(n,n)\}\cup\bigcup_{i\ne j}\{(i,j)\}$$

with the polynomial $1-x$ on $\{(1,1),\dots,(n,n)\}$ and the polynomial 1 on the singletons $\{(i,j)\}$. The lower-triangular matrices arise from

$$\{1,\dots,n\}\times\{1,\dots,n\} = \{(i,j) : i\ge j\}\cup\bigcup_{i<j}\{(i,j)\},$$

the zero polynomial on $\{(i,j) : i \ge j\}$, and the polynomial 1 on the singletons $\{(i,j)\}$ with $i < j$. Finally, for matrices of the form (2.3) the partition of $\{1,2,3\} \times \{1,2,3\}$ into strings is $\{(1,1),(1,3),(2,2),(3,1),(3,3)\} \cup \{(1,2),(2,1)\} \cup \{(2,3)\} \cup \{(3,2)\}$, and the polynomials are $1-x$ on the strings of lengths 5 and 2 and are the zero polynomials on the two singletons.

Given polynomial (3.2), we define the function b_m on the complex unit circle $\mathbf{T}$ by $b_m(t) = |a_m(t)|^2$ $(t \in \mathbf{T})$ and expand b_m into its Fourier series,

$$b_m(t) = \sum_{k=-r_m}^{r_m} b_k^{(m)} t^k \quad (t = e^{i\theta} \in \mathbf{T}).$$

We denote by $T(b_m)$ the infinite Toeplitz matrix $(b_{i-j}^{(m)})_{i,j=1}^{\infty}$ and by $T_R(b_m)$ the principal $R \times R$ section $(b_{i-j}^{(m)})_{i,j=1}^{R}$ of $T(b_m)$. The matrices $T_R(b_m)$ are all positive definite. We denote by $\lambda_{\min}(T_R(b_m))$ and $\lambda_{\max}(T_R(b_m))$ the minimal and maximal eigenvalues of $T_R(b_m)$, respectively.

Theorem 2.1 and Corollary 2.3 allow us to compute the maximum of (2.1) and to estimate stochastically the ratio occurring in (2.1) in terms of the quantities $s_{\min}^+$, $\|D\|_2^2$, and $\|D\|_\infty^2$. The following result provides us with these quantities in the case of string structures.

THEOREM 3.3. *For every string structure,*

$$\left(s_{\min}^+\right)^2 = \min_{1 \le m \le M} \lambda_{\min}(T_{\ell_m - r_m}(b_m)),$$

$$\|D\|_2^2 = \sum_{m=1}^{M} (\ell_m - r_m) \sum_{k=1}^{r_m} |a_k^{(m)}|^2, \quad \|D\|_\infty^2 = \max_{1 \le m \le M} \lambda_{\max}(T_{\ell_m - r_m}(b_m)).$$

Proof. After stacking matrices in $M_n(\mathbf{K})$ string by string (and following the ordering within the strings) to columns in $\mathbf{K}^{n^2}$, the operator D becomes an $n^2 \times n^2$ block diagonal matrix $\nabla = \operatorname{diag}(B_1, \ldots, B_M)$ with $\ell_m \times \ell_m$ matrices B_m. The first $\ell_m - r_m$ rows of B_m are the $(\ell_m - r_m) \times \ell_m$ Toeplitz matrix whose first row is $(a_0^{(m)}, a_1^{(m)}, \ldots, a_{r_m}^{(m)}, 0, \ldots, 0)$ and whose first column is $(a_0^{(m)}, 0, \ldots, 0)^\top$. The last r_m rows of B_m are zero. This implies the asserted formula for $\|D\|_2^2$. A straightforward computation gives

$$B_m B_m^* = \begin{pmatrix} T_{\ell_m - r_m}(b_m) & 0 \\ 0 & O_{r_m} \end{pmatrix},$$

where O_{r_m} is the $r_m \times r_m$ zero matrix. This yields the asserted expressions for $s_{\min}^+$ and $\|D\|_\infty^2$. ∎

4. Toeplitz, Hankel, and chessboard structures. Let L_m be as in Example 3.1 and choose $a_m(x) = 1 - x$ for all m. The corresponding structure is the Toeplitz structure. The functions b_m are all $b_m(t) = |t-1|^2$. It is well known (see, e.g., [2, 5, 7]) that the eigenvalues of $T_R(b_m)$ are $\lambda_j = 2 + 2\cos\frac{\pi j}{R+1}$ $(j = 1, \ldots, R)$. Thus, Theorem 3.3 with $\ell_m = n - |m|$ and $r_m = 1$ for $m = -(n-1), \ldots, n-1$ yields

$$\|D\|_2^2 = 2\sum(\ell_m - r_m) = 2n^2 - 2\sum r_m = 2n^2 - 2(2n-1) = 2(n-1)^2,$$

$$\|D\|_\infty^2 = 2 + 2\cos\frac{\pi}{n} = 4\cos^2\frac{\pi}{2n}, \quad \left(s_{\min}^+\right)^2 = 2 - 2\cos\frac{\pi}{n} = 4\sin^2\frac{\pi}{2n}.$$

Theorem 1.1 is now immediate from Theorem 2.1, while Theorem 1.2 follows from Corollary 2.3 and the fact that

$$\frac{2n^2\|D\|_\infty^4}{(\|D\|_2^2 - n^2/100)^2} \le \frac{2n^2 \cdot 4^2}{(2(n-1)^2 - n^2/100)^2} < \frac{13}{n^2} \quad \text{for} \quad n \ge 10.$$

In the Hankel case we take the partition considered in Example 3.2 and arrive at the same results as in the Toeplitz case. In particular, Theorems 1.1 and 1.2 remain literally true with $\mathcal{T}_n(\mathbf{K})$ replaced by the set of all Hankel matrices in $M_n(\mathbf{K})$.

To give another illustration, let us consider chessboard matrices. In this case $L_1, L_2, a_1(x) = a_2(x) = 1 - x$ are as in Example 3.2. The functions b_1, b_2 are again $b_1(t) = b_2(t) = |t-1|^2$, but now $\ell_1 = [n^2/2]$ and $\ell_2 = [(n^2+1)/2]$, where $[q]$ denotes the integral part of q. Thus,

$$\begin{aligned}
\|D\|_2^2 &= \left(\left[\frac{n^2}{2}\right] - 1\right) \cdot 2 + \left(\left[\frac{n^2+1}{2}\right] - 1\right) \cdot 2 = 2(n^2-2), \\
\|D\|_\infty^2 &= 4\cos^2 \frac{\pi}{2[n^2/2 + 1/2]} \sim 4\cos^2 \frac{\pi}{n^2}, \\
\left(s_{\min}^+\right)^2 &= 4\sin^2 \frac{\pi}{2[n^2/2 + 1/2]} \sim 4\sin^2 \frac{\pi}{n^2} \sim \frac{4\pi^2}{n^4},
\end{aligned}$$

which shows that (2.4) increases asymptotically as $n^2/(2\pi)$. From Corollary 2.3 we deduce that nevertheless

$$P\left(\frac{\operatorname{dist}_2(X, \operatorname{Ker} D)}{\|D(X)\|_2} > 10\right) < \frac{33}{n^2} \quad \text{for} \quad n \ge 10.$$

5. Symmetric matrices. These come from the partition

$$\{1, \dots, n\} \times \{1, \dots, n\} = \bigcup_{i<j} \{(i,j), (j,i)\} \cup \bigcup_i \{(i,i)\}$$

with $a_{ij}(x) = 1 - x$ on the doubletons and $a_i(x) = 0$ on the singletons. Theorem 3.3 gives $(s_{\min}^+)^2 = \lambda_{\min}(T_1(|t-1|^2)) = 2$, and hence (2.4) equals $\sqrt{2}/2$ for all n.

6. Vandermonde-like structures. We start with any partition (3.1) and any ordering of the strings. In the classic Vandermonde case, L_m is $\{(i,m) : 1 \le i \le n\}$ and is ordered by increasing i. Let $a_m(x) = \alpha_m - x$. We get

$$b_m(t) = |\alpha_m - t|^2 = 1 + |\alpha_m|^2 - \alpha_m t^{-1} - \overline{\alpha_m} t.$$

The eigenvalues of the tridiagonal Toeplitz matrix $T_R(b_m)$ are

$$1 + |\alpha_m|^2 + 2|\alpha_m| \cos \frac{\pi j}{R+1} \quad (j = 1, \dots, R)$$

(see again [2, 5, 7]). Consequently, by Theorem 3.3,

$$\left(s_{\min}^+\right)^2 = \min_m \left(1 + |\alpha_m|^2 - 2|\alpha_m| \cos \frac{\pi}{\ell_m}\right) = \min_m \left((1 - |\alpha_m|)^2 + 4|\alpha_m| \sin^2 \frac{\pi}{2\ell_m}\right).$$

Thus, if there is a $\delta > 0$ such that $|\alpha_m| \notin (1-\delta, 1+\delta)$, then (2.4) is not greater than $1/\delta$. The ratio (2.4) becomes large only if there is an m such that $|\alpha_m|$ is close to 1 and, at the same time, ℓ_m is large. Note that in the classic case $\ell_m = n$ for all m.

7. Polynomially ill-conditioned structures. We know that (2.4) increases asymptotically as n/π for the Toeplitz and Hankel structures and as $n^2/(2\pi)$ for the chessboard structure. In this section we provide string structures for which (2.4) increases as a polynomial of arbitrarily prescribed degree.

For the sake of definiteness, take partition (3.1) with $L_m = \{(i,m) : 1 \le i \le n\}$ with the ordering induced by increasing i. For each m, let $a_m(x)$ be a polynomial of degree $r_m \le n-1$ and define b_m by $b_m(t) = |a_m(t)|^2$. The asymptotic behavior of the extreme eigenvalues of $T_R(b_m)$ has been studied by many authors, including [5, 8, 10, 11, 12, 15, 21, 22, 23]. These results imply the following (a full proof of which is also given in [2]). If 2γ is the maximal order of the zeros of b_m on $\mathbf{T}$, then there exist constants $0 < C_1 < C_2 < \infty$ depending only on b_m such that

$$C_1 \frac{1}{R^{2\gamma}} \le \lambda_{\min}(T_R(b_m)) \le C_2 \frac{1}{R^{2\gamma}}$$

for all R. Consequently, Theorems 2.1 and 3.3 imply that the ratio (2.4) increases as $n^{\max(r_1,\ldots,r_n)}$.

To be more concrete, let $a_m(x) = (1-x)^r$ for all m. Thus, we require that the entries of each column are the values of a polynomial of degree $r-1$ at $1,\ldots,n$. We have $b_m(t) = |t-1|^{2r}$ for all m. The function $b_m(t) = |t-1|^{2r}$ has exactly one zero $t=1$ on $\mathbf{T}$, and the order of this zero is $2r$. From what was said in the preceding paragraph we deduce that (2.4) increases as n^r. In the special case at hand it is even known from the works cited above that $R^{2r}\lambda_{\min}(T_R(b_m))$ converges to a limit c_r as $R \to \infty$. The limiting constants are rapidly increasing ($c_1 = \pi^2 = 9.8696$, $c_2 \approx 500$, $c_3 \approx 61529$) and c_r can be shown to coincide with the minimal eigenvalue of the differential operator $(-1)^r u^{(2r)}$ on $(0,1)$ with the boundary conditions $u^{(j)}(0) = u^{(j)}(1) = 0$ for $0 \le j \le r-1$ (see [1, 12, 14, 22, 23]). From Theorem 3.3 we also deduce that

$$\|D\|_2^2 = n(n-r)\sum_{k=0}^{r}\binom{r}{k}^2 = \binom{2r}{r} n(n-r),$$
$$\|D\|_\infty^2 = \lambda_{\max}(T_{n-r}(|t-1|^{2r})) \le 2^{2r}$$

(note that $\|T_R(b)\|_\infty \le \max_{t\in\mathbf{T}} |b(t)|$ for every b). Consequently, Corollary 2.3 implies that nonetheless

$$P\left(\frac{\operatorname{dist}_2(X, \operatorname{Ker} D)}{\|D(X)\|_2} > 10\right) \le \frac{2n^2 \cdot 2^{4r}}{\left(\binom{2r}{r} n(n-r) - n^2/100\right)^2} = O\left(\frac{1}{n^2}\right).$$

8. Toeplitz-plus-Hankel matrices I. A Toeplitz-plus-Hankel matrix (T+H matrix, for short) is a matrix of the form $(t_{i-j} + h_{i+j})_{i,j=1}^n$ with $t_k, h_k \in \mathbf{K}$. The T+H structure is not a string structure in the sense of section 3, and hence it is not as easy to detect as a string structure—for example, with unskilled eyes it is not trivial to decide whether

$$\begin{pmatrix} 2.9 & 7.3 & -1.9 \\ 5.4 & 0.3 & 0.7 \\ 5.2 & -1.2 & 1.9 \end{pmatrix}$$

is T+H or not—and, moreover, Theorem 3.3 is not applicable to T+H matrices.

Heinig and Rost [6] observed that in the T+H case one may consider the displacement matrix $XW - WX$, where $W = U + U^\top$ and U is the shift matrix introduced in section 1. One can show that $X \in M_n(\mathbf{K})$ is T+H if and only if the

central $(n-2)\times(n-2)$ submatrix of $XW-WX$ is zero. Consequently, we define $D(X)\in M_n(\mathbf{K})$ as the matrix that results from $XW-WX$ by replacing the first and last rows and the first and last columns with zero. Throughout this section it will be convenient to emphasize the dependence on n. We therefore write W_n and D_n for W and D, respectively. Thus, $X\in M_n(\mathbf{K})$ is T+H if and only if $D_n(X)=0$.

After appropriate stacking, D_n becomes an $n^2\times n^2$ matrix

$$\nabla_n = I\otimes W_n - W_n\otimes I - R_n,$$

where $\otimes$ denotes the Kronecker product. The matrix R_n stems from deleting the first and last rows and columns of XW_n-W_nX. We take advantage of the two lucky circumstances that W_n is Hermitian and that R_n is an $n^2\times n^2$ matrix whose rank is at most $4n-2=o(n^2)$ to prove the following.

THEOREM 8.1. *In the Toeplitz-plus-Hankel case,*

$$\lim_{n\to\infty}\max_{X\notin\operatorname{Ker} D_n}\frac{\operatorname{dist}_2(X,\operatorname{Ker} D_n)}{\|D_n(X)\|_2}=\infty.$$

Proof. We employ the general fact that if $\{A_n\}$ and $\{B_n\}$ are two sequences of $n^2\times n^2$ matrices such that $\operatorname{rank}(A_n-B_n)=o(n^2)$ and if $s_j(A_n)$ and $s_j(B_n)$ are the singular values of A_n and B_n, respectively, then

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{j=1}^{n^2}(\varphi(s_j(A_n))-\varphi(s_j(B_n)))=0$$

for every compactly supported continuous function φ on $\mathbf{R}$ (see [19, 20]).

We apply this result to $A_n=\nabla_n$ and $B_n=I\otimes W_n-W_n\otimes I$. Given an interval $(a,b)\subset\mathbf{R}$, we denote by $\alpha_n(a,b)$ and $\beta_n(a,b)$ the number of singular values of A_n and B_n in (a,b) (multiplicities taken into account). The eigenvalues of W_n are $\lambda_j=2\cos\frac{\pi j}{n+1}$ $(j=1,\dots,n)$, and hence the eigenvalues of B_n are $\lambda_j-\lambda_k$ $(j,k=1,\dots,n)$. Since W_n and thus B_n is Hermitian, it follows that the singular values of B_n are

$$s_{jk}=|\lambda_j-\lambda_k|=2\left|\cos\frac{\pi j}{n+1}-\cos\frac{\pi k}{n+1}\right|=4\left|\sin\frac{\pi(j-k)}{2n+2}\sin\frac{\pi(j+k)}{2n+2}\right|,$$

where $j,k=1,\dots,n$. Fix $\varepsilon\in(0,1)$. A little thought reveals that

$$\lim_{n\to\infty}\frac{\beta_n(2\varepsilon,3\varepsilon)}{(n+1)^2}=\frac{1}{4\pi^2}\,|G_\varepsilon|,$$

where $|G_\varepsilon|$ is the area of the region

$$G_\varepsilon=\left\{(x,y)\in(0,2\pi)^2 : 2\varepsilon<4\left|\sin\frac{x-y}{2}\sin\frac{x+y}{2}\right|<3\varepsilon\right\}.$$

Clearly, $|G_\varepsilon|>0$ whenever $\varepsilon<2$. Thus, we have

$$\beta_n(2\varepsilon,3\varepsilon)=\frac{|G_\varepsilon|}{4\pi^2}\,n^2+o(n^2)\quad\text{with}\quad|G_\varepsilon|>0.$$

Let $\varphi : \mathbf{R} \to [0,1]$ be a compactly supported continuous function which is identically 1 in $(2\varepsilon, 3\varepsilon)$ and identically 0 outside $(\varepsilon, 4\varepsilon)$. Then

$$\begin{aligned}\alpha_n(\varepsilon, 4\varepsilon) &\geq \sum_{j=1}^{n^2} \varphi(s_j(A_n)) = \sum_{j=1}^{n^2} \varphi(s_j(B_n)) + o(n^2) \\ &\geq \beta_n(2\varepsilon, 3\varepsilon) + o(n^2) = \frac{|G_\varepsilon|}{4\pi^2}\, n^2 + o(n^2),\end{aligned}$$

and since $|G_\varepsilon| > 0$, we conclude that $\alpha_n(\varepsilon, 4\varepsilon) \to \infty$ as $n \to \infty$.

Now pick a large number $C > 0$. By what was just proved, there exists an $n_0 = n_0(C)$ such that $\nabla_n = A_n$ and thus D_n also has a singular value in $(\frac{1}{4C}, \frac{1}{C})$ for all $n \geq n_0$. It follows that the smallest nonzero singular value of D_n does not exceed $1/C$ for $n \geq n_0$. Theorem 2.1 therefore gives

$$\max_{X \notin \operatorname{Ker} D_n} \frac{\operatorname{dist}_2(X, \operatorname{Ker} D_n)}{\|D_n\|_2} > C \quad \text{for} \quad n \geq n_0(C).$$

As $C > 0$ was arbitrary, we arrive at the assertion. □

THEOREM 8.2. *In the Toeplitz-plus-Hankel case,*

$$\|D_n\|_2^2 = 4(n-2)^2, \quad \|D_n\|_\infty \leq 4, \quad \lim_{n\to\infty} \|D_n\|_\infty = 4.$$

Proof. We may write $D_n(X) = P(XW_n - W_nX)$, where $P : M_n(\mathbf{K}) \to M_n(\mathbf{K})$ is the projection on the central $(n-2) \times (n-2)$ matrix. This implies that $\|D_n\|_\infty \leq \|P\|_\infty(\|W_n\|_\infty + \|W_n\|_\infty) = 1 \cdot (1+1) = 4$. To prove that $\|D_n\|_\infty \to 4$, we proceed as in the proof of Theorem 8.1. Namely, $\beta_n(4-3\varepsilon, 4-2\varepsilon) = \frac{1}{4\pi^2}\,|H_\varepsilon| n^2 + o(n^2)$, where $|H_\varepsilon| > 0$ is the area of the region

$$H_\varepsilon = \left\{(x,y) \in (0,2\pi)^2 : 4 - 3\varepsilon < 4\left|\sin\frac{x-y}{2}\sin\frac{x+y}{2}\right| < 4 - 2\varepsilon\right\},$$

which implies that $\alpha_n(4-4\varepsilon, 4-\varepsilon) \to \infty$ as $n \to \infty$. Consequently, D_n has a singular value in $(4-4\varepsilon, 4-\varepsilon)$ if only n is large enough, which shows that $\|D_n\|_\infty \to 4$.

We know that D_n is unitarily similar to the matrix ∇_n introduced in the proof of Theorem 8.1. The matrix ∇_n has $4n-4$ zero rows, and the remaining $(n-2)^2$ rows have $1, 1, -1, -1$, and $n^2 - 4$ zeros as entries. This gives $\|D_n\|_2^2 = 4(n-2)^2$. □

COROLLARY 8.3. *We have*

$$P\left(\frac{\operatorname{dist}_2(X, \operatorname{Ker} D_n)}{\|D_n(X)\|_2} > 10\right) < \frac{79}{n^2} \quad \textit{for} \quad n \geq 10.$$

Proof. From Corollary 2.3 and Theorem 8.2 we deduce that the probability considered is at most $2n^2 \cdot 4^4/(4(n-2)^2 - n^2/100)^2$, which is smaller than $79/n^2$ for $n \geq 10$. □

9. Toeplitz-plus-Hankel matrices II. To estimate the growth rate in Theorem 8.1, we have to exploit heavier machinery. Let $Q = (-\pi, \pi)^2$. For a function $f \in L^\infty(Q)$, the quarter-plane Toeplitz operator $T^{(2)}(f)$ is defined on $\ell^2(\mathbf{N} \times \mathbf{N})$ by

$$\big(T^{(2)}(f)x\big)_{ij} = \sum_{k,\ell=1}^{\infty} f_{i-k,j-\ell}\, x_{k\ell} \quad (i,j \geq 1),$$

where the numbers $f_{k\ell}$ are the Fourier coefficients of f,

$$f_{k\ell} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x,y) e^{-ikx} e^{-i\ell y}\, dx\, dy.$$

The compression of $T^{(2)}(f)$ to the square $\{1,\dots,n\}\times\{1,\dots,n\}$ is denoted by $T_n^{(2)}(f)$. Writing

$$f(x,y) = \sum_{k,\ell=-\infty}^{\infty} f_{k,\ell} e^{ikx} e^{i\ell y} = \left(\sum_{k=-\infty}^{\infty} f_{k\ell} e^{ikx} \right) e^{i\ell y} =: \sum_{k=-\infty}^{\infty} f_\ell(e^{ix}) e^{i\ell y}, \tag{9.1}$$

one can identify $T_n^{(2)}(f)$ with $(T_n(f_{i-j}))_{i,j=1}^n$, which is a block Toeplitz matrix with Toeplitz blocks. If f is nonnegative and not identically zero, then $T_n^{(2)}(f)$ is positive definite. We denote by $\lambda_{\min}(T_n^{(2)}(f))$ the smallest eigenvalue of $T_n^{(2)}(f)$ and by $s_{\min}^+(n)$ the smallest nonzero singular value of D_n.

THEOREM 9.1. *For $n \ge 3$,*

$$(s_{\min}^+(n))^2 = \lambda_{\min}\big(T_{n-2}^{(2)}(f)\big) \quad \textit{with} \quad f(x,y) = 16 \sin^2 \frac{x+y}{2} \sin^2 \frac{x-y}{2}.$$

Proof. Let ∇_n be as in the preceding section. One can show that

$$\nabla_n \nabla_n^* = J \begin{pmatrix} F_n & 0 \\ 0 & O_{4n-4} \end{pmatrix} J \quad \text{with} \quad F_n = (T_{n-2}(f_{i-j}))_{i,j=1}^{n-2},$$

where J is a permutation matrix and $f_0(t) = 4+t^2+t^{-2}$, $f_1(t) = f_{-1}(t) = -2(t+t^{-1})$, $f_2(t) = f_{-2}(t) = 1$, $f_\ell(t) = 0$ for $|\ell| \ge 3$. By (9.1), we may identify F_n with $T_{n-2}^{(2)}(f)$ for

$$\begin{aligned} f(x,y) &= \sum_{\ell=-2}^{2} f_\ell(e^{ix}) e^{i\ell y} = 4 + 2\cos 2x - 2(2\cos x)(2\cos y) + 2\cos 2y \\ &= 4(\cos x - \cos y)^2 = 16 \sin^2 \frac{x+y}{2} \sin^2 \frac{x-y}{2}. \qquad \square \end{aligned}$$

The following is a significant refinement of Theorem 8.1.

THEOREM 9.2. *There exist constants $0 < C_1 < C_2 < \infty$ such that for all $n \ge 1$,*

$$C_1\, n^2 \le \max_{X \notin \operatorname{Ker} D_n} \frac{\operatorname{dist}_2(X, \operatorname{Ker} D_n)}{\|D_n(X)\|_2} \le C_2\, n^2.$$

Proof outline. By Theorems 2.1 and 9.1, we have to prove that

$$C_3/n^4 \le \lambda_{\min}(T_n^{(2)}(f)) \le C_4/n^4$$

with certain constants $0 < C_3 < C_4 < \infty$. Papers [16, 21, 23] pertain to that question, but they are restricted to the case where $f \ge 0$ has only a single zero and are therefore not applicable to the case at hand. We will instead benefit from the special structure of f.

We identify $\mathbf{K}^n$ with the set $\mathcal{P}_n$ of all trigonometric polynomials of the form $\varphi(x) = \varphi_0 + \varphi_1 e^{ix} + \cdots + \varphi_{n-1} e^{i(n-1)x}$ $(x \in (-\pi,\pi))$. The ℓ^2 norm on $\mathbf{K}^n$ induces

the norm $\|\varphi\|^2 = \int_{-\pi}^{\pi} |\varphi(x)|^2\,dx/(2\pi)$ on $\mathcal{P}_n$. Accordingly, $\ell^2(\{1,\dots,n\}\times\{1,\dots,n\})$ may be identified with $\mathcal{P}_n \otimes \mathcal{P}_n$. For $\varphi \in \mathcal{P}_n \otimes \mathcal{P}_n$,

$$\big(T_n^{(2)}(f)\varphi,\varphi\big) = 16\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sin^2\frac{x+y}{2}\,\sin^2\frac{x-y}{2}\,|\varphi(x,y)|^2\,\frac{dx\,dy}{4\pi^2},$$

and using periodicity and substituting $x+y=2\xi$, $x-y=2\eta$, one can show that this is not smaller than

$$C_5\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sin^2\xi\,\sin^2\eta\,|\varphi(\xi+\eta,\xi-\eta)|^2\,d\xi\,d\eta = C_5((A_{2n}\otimes A_{2n})\widetilde{\varphi},\widetilde{\varphi}), \tag{9.2}$$

where $C_5>0$ is some constant, $A_{2n}=T_{2n}(\sin^2 x)$, and $\widetilde{\varphi}\in\mathcal{P}_{2n}\otimes\mathcal{P}_{2n}$ is defined by $\widetilde{\varphi}(\xi,\eta)=\varphi(\xi+\eta,\xi-\eta)$. The smallest eigenvalue of A_{2n} is $\sin^2\frac{\pi}{2n+2}$, and hence (9.2) is at least

$$C_5\left(\sin^2\frac{\pi}{2n+2}\right)^2\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|\varphi(\xi+\eta,\xi-\eta)|^2\,d\xi\,d\eta \ge \frac{C_3}{n^4}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|\varphi(x,y)|^2\,\frac{dx\,dy}{4\pi^2}$$

with some constant $C_3>0$. In summary, $(T_n^{(2)}(f)\varphi,\varphi)\ge(C_3/n^4)\|\varphi\|^2$, which implies that $\lambda_{\min}(T_n^{(2)}(f))\ge C_3/n^4$.

The reverse inequality can be proved as in [2, pp. 37–41]. Namely, let $n=6m+k$ with $1\le k\le 6$, define $p_m\in\mathcal{P}_{3m+1}$ by $p_m(x)=(1+e^{ix}+\dots+e^{imx})^3$, and put $\varphi(x,y)=p_m(x+y)p_m(x-y)$. Then $\varphi\in\mathcal{P}_{6m+1}\otimes\mathcal{P}_{6m+1}\subset\mathcal{P}_n\otimes\mathcal{P}_n$ and one can show that

$$\|\varphi\|^2 \ge C_6\left(\int_{-\pi}^{\pi}|p_m(\xi)|^2\,d\xi\right)^2 \ge C_7\,m^{10},$$
$$\|f\varphi\|^2 \le C_8\left(\int_{-\pi}^{\pi}\sin^4\frac{x}{2}\,|p_m(x)|^2\,dx\right)^2 \le C_9\,m^2,$$

where C_6, C_7, C_8, and C_9 are constants in $(0,\infty)$. This implies (again see [2, p. 38]) that

$$1/\lambda_{\min}^2\big(T_n^{(2)}(f)\big) = \|(T_n^{(2)}(f))^{-1}\|_\infty \ge \frac{\|\varphi\|^2}{\|f\varphi\|^2} \ge \frac{C_7}{C_9}\,m^8,$$

whence $\lambda_{\min}(T_n^{(2)}(f)) = O(1/m^4) = O(1/n^4)$.

As pointed out by one of the referees, an alternative proof of the reverse inequality is as follows. From the last line of the proof of Theorem 9.1 we infer that

$$f(x,y)\le 16\left(\frac{x+y}{2}\right)^2\left(\frac{x-y}{2}\right)^2 = (x^2-y^2)^2 \le x^4+y^4 =: g(x,y).$$

Consequently, by a monotonicity argument (for which, see, for example, [16, p. 116]), the minimal eigenvalue of $T_n^{(2)}(f)$ does not exceed the minimal eigenvalue of $T_n^{(2)}(g)$, and the latter is known to be asymptotically equal to a constant times $1/n^4$ (see [16, 21, 22, 23]). □

10. Conclusions. We have provided parameters that allow us to estimate how far a matrix X may be from a given structure $\operatorname{Ker} D$ if the Frobenius norm of the displacement matrix $D(X)$ is small. In this way we can, for each structure, estimate

the conditioning of the problem of testing membership in this structure by checking whether the displacement matrix is small.

In practice, one would perhaps work with the norm $n \max |x_{ij}|$ instead of the Frobenius norm. The treatment of this norm is theoretically more difficult and we have not embarked on this question. However, a consideration of examples shows that the basic qualitative (though not quantitative) results on the conditioning of the problem seem to be independent of the choice of the norm.

We have seen that the more ill-conditioned the string structures, the larger the maximal string length. (The maximal string length is n for $n \times n$ Toeplitz or Hankel matrices and about $n^2/2$ for $n \times n$ chessboard matrices.) In these cases it is advisable to replace consideration of $\|D(X)\|_2$ by the direct computation of $\operatorname{dist}_2(X, \operatorname{Ker} D)$, which is an easy task if an orthonormal basis in $\operatorname{Ker} D$ is available. If, for example, we want to test whether a matrix is constant along prescribed strings, we should replace the entries of each string by their arithmetic mean and test whether the distance of this matrix to the original matrix is small.

The conditioning of checking the Toeplitz-plus-Hankel structure by having recourse to the displacement matrix $XW - WX$ increases as the square of the matrix dimension. The search for alternative tests of the Toeplitz-plus-Hankel structure is therefore desirable.

More generally, as also observed by the referees, the question of testing whether a given matrix has a prescribed structure deserves a further and careful consideration. The message of this paper is that the naive use of displacement operations may at least theoretically be dangerous. We have left open the question of how to do it in the right way. Furthermore, we have restricted ourselves to structures for which the corresponding displacement operations are known or easy to guess. Other important structures, such as locally Toeplitz sequences [18] or generalized locally Toeplitz sequences [17], cannot be tackled by displacement operations in an obvious way. These and related matters await further investigation.

Acknowledgments. I thank Peter Benner and Harold Widom for useful discussions on several aspects of the topic. I am also greatly indebted to the referees for their competent and valuable remarks.

REFERENCES

[1] A. Böttcher, *The constants in the asymptotic formulas by Rambour and Seghier for inverses of Toeplitz matrices*, Integral Equations Operator Theory, 50 (2004), pp. 43–55.

[2] A. Böttcher and S. Grudsky, *Toeplitz Matrices, Asymptotic Linear Algebra, and Functional Analysis*, Hindustan Book Agency, New Delhi, Birkhäuser-Verlag, Basel, 2000.

[3] A. Böttcher and S. Grudsky, *The norm of the product of a large matrix and a random vector*, Electron. J. Probab., 8 (2003).

[4] G. M. Fichtenholz, *Differential- und Integralrechnung*, Deutscher-Verlag der Wissenschaften, Berlin, 1972.

[5] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[6] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Akademie-Verlag, Berlin, Birkhäuser-Verlag, Basel, 1984.

[7] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[8] M. Kac, W. L. Murdock, and G. Szegö, *On the eigenvalues of certain Hermitian forms,* J. Ration. Mech. Anal., 2 (1953), pp. 767–800.

[9] T. Kailath, S.-Y. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[10] S. V. Parter, *On the extreme eigenvalues of truncated Toeplitz matrices*, Bull. Amer. Math. Soc., 67 (1961), pp. 191–196.

[11] S. V. Parter, *Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations*, Trans. Amer. Math. Soc., 99 (1961), pp. 153–192.
[12] S. V. Parter, *On the extreme eigenvalues of Toeplitz matrices*, Trans. Amer. Math. Soc., 100 (1961), pp. 263–276.
[13] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series, Vol. 1, Elementary Functions*, Gordon & Breach, New York, 1986.
[14] P. Rambour and A. Seghier, *Formulas for the inverses of Toeplitz matrices with polynomially singular symbols*, Integral Equations Operator Theory, 50 (2004), pp. 83–114.
[15] S. Serra Capizzano, *On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima/maxima*, BIT, 36 (1996), pp. 135–142.
[16] S. Serra Capizzano, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.
[17] S. Serra Capizzano, *Generalized locally Toeplitz sequences: Spectral analysis and application to discretized partial differential equations,* Linear Algebra Appl., 366 (2003), pp. 371–402.
[18] P. Tilli, *Locally Toeplitz matrices: Spectral theory and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.
[19] E. E. Tyrtyshnikov, *Influence of matrix operations on the distribution of eigenvalues and singular values of Toeplitz matrices,* Linear Algebra Appl., 207 (1994), pp. 225–249.
[20] E. E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[21] H. Widom, *On the eigenvalues of certain Hermitian operators*, Trans. Amer. Math. Soc., 88 (1958), pp. 491–522.
[22] H. Widom, *Extreme eigenvalues of translation kernels*, Trans. Amer. Math. Soc., 100 (1961), pp. 252–262.
[23] H. Widom, *Extreme eigenvalues of N-dimensional convolution operators*, Trans. Amer. Math. Soc., 106 (1963), pp. 391–414.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 55–65

CONVERGENCE OF MULTISTEP TIME DISCRETIZATIONS OF NONLINEAR DISSIPATIVE EVOLUTION EQUATIONS*

ESKIL HANSEN†

Abstract. Global error bounds are derived for multistep time discretizations of fully nonlinear evolution equations on infinite dimensional spaces. In contrast to earlier studies, the analysis presented here is not based on linearization procedures but on the fully nonlinear framework of logarithmic Lipschitz constants and nonlinear semigroups. The error bounds reveal how the contractive or dissipative behavior of the vector field, governing the evolution, and the properties of the multistep method influence the convergence. A multistep method which is consistent of order p is proven to be convergent of the same order when the vector field is contractive or strictly dissipative, i.e., of the same order as in the ODE-setting. In the contractive context it is sufficient to require strong zero-stability of the method, whereas strong A-stability is sufficient in the dissipative case.

Key words. nonlinear evolution equations, logarithmic Lipschitz constants, dissipative maps, multistep methods, stability, convergence

AMS subject classifications. 65J15, 65M12

DOI. 10.1137/040610362

1. Introduction.

The evolution equation

$$\dot{u} = f(u), \quad u(0) = \eta,$$

where $u : [0, \infty) \to X$ and the vector field f is a dissipative map on the Banach space X, has received much attention as this type of equation is found in a wide range of applications, e.g., advection-diffusion-reaction processes. In the early 1970s the theory of nonlinear semigroups made it possible to characterize the solutions of evolution equations with fully nonlinear vector fields; see [1, 2, 21]. Shortly thereafter, multistep time discretizations of such evolution equations were analyzed in the literature; see [11, 12, 16], where some stability and convergence results are derived in a Hilbert space context. Studies of multistep discretizations on infinite dimensional spaces have predominantly considered linear vector fields, e.g., [6, 14], where the analysis is based on analytic semigroups (by definition linear). The same analysis is also used when combining multistep and Galerkin methods; this is reviewed in [19]. It is not until just recently that the fully nonlinear setting has been addressed; see, e.g., [5, 13]. Here, bounds of the global error are derived by expressing the vector field as a sum of its linearization and a nonlinear residual, which enables the usage of the linear theory. The aim of this paper is therefore to derive global error bounds in the setting of fully nonlinear problems on infinite dimensional spaces, which are not necessarily linearizable, and to obtain a qualitative understanding of how the mathematical setting and the numerical methods influence such bounds.

2. Preliminaries.

This paper is based on the theory of logarithmic Lipschitz constants, which was developed in [9, 17, 18]. Below follows a short summary of the theory. Assume that X is a real valued Banach space, with norm $\|\cdot\|_X$, and f is a

*Received by the editors June 22, 2004; accepted for publication (in revised form) June 13, 2005; published electronically February 8, 2006.
http://www.siam.org/journals/sinum/44-1/61036.html
†Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden (eskil@maths.lth.se).

nonlinear map on X with domain $D(f)$ and range $R(f)$. The Lipschitz constants of f on X are defined as follows.

DEFINITION 2.1. *For u, $v \in D(f)$ define the* lub *and* glb *Lipschitz constants of f on X by*

$$L_X[f] := \sup_{u \neq v} \frac{\|f(u) - f(v)\|_X}{\|u - v\|_X}, \quad l_X[f] := \inf_{u \neq v} \frac{\|f(u) - f(v)\|_X}{\|u - v\|_X}.$$

The basic properties of the Lipschitz constants are given in Proposition 2.2.

PROPOSITION 2.2. *Assume that $R(g) \subseteq D(f)$ in property* 4 *and $D(f) \bigcap D(g) \neq \emptyset$ in properties* 3 *and* 5. *Then,*

1. $L_X[f] \geq 0$,
2. $L_X[\alpha f] = |\alpha| L_X[f]$,
3. $L_X[f + g] \leq L_X[f] + L_X[g]$,
4. $L_X[fg] \leq L_X[f] L_X[g]$,
5. $l_X[f] - L_X[g] \leq l_X[f + g]$.

In order to mimic an inner product space, introduce semi-inner products on X induced by the norm $\|\cdot\|_X$.

DEFINITION 2.3. *Let $u, v \in X$ and define the left $(\cdot,\cdot)_X^-$ and right $(\cdot,\cdot)_X^+$ semi-inner products on X as*

$$(u, v)_X^{\pm} := \|u\|_X \lim_{\varepsilon \to 0\pm} \frac{\|u + \varepsilon v\|_X - \|u\|_X}{\varepsilon}.$$

The semi-inner products exist as they are Gateaux differentials of the norm $\|\cdot\|_X$; see [4]. If the norm is induced by an inner product $\langle\cdot,\cdot\rangle_X$, then $(\cdot,\cdot)_X^{\pm} = \langle\cdot,\cdot\rangle_X$. Like the true inner product, the semi-inner products satisfy the relation $(u, u)_X^{\pm} = \|u\|_X^2$ and the Cauchy–Schwarz inequalities

$$-\|u\|_X \|v\|_X \leq (u, v)_X^{\pm} \leq \|u\|_X \|v\|_X.$$

Next, the logarithmic Lipschitz constants are introduced.

DEFINITION 2.4. *For u, $v \in D(f)$, define the* lub *and* glb *logarithmic Lipschitz constants of f on X as*

$$M_X^{\pm}[f] := \sup_{u \neq v} \frac{(u - v, f(u) - f(v))_X^{\pm}}{\|u - v\|_X^2}, \quad m_X^{\pm}[f] := \inf_{u \neq v} \frac{(u - v, f(u) - f(v))_X^{\pm}}{\|u - v\|_X^2}.$$

Some of the basic properties of the logarithmic Lipschitz constants are given in Proposition 2.5.

PROPOSITION 2.5. *Assume that $D(f) \bigcap D(g) \neq \emptyset$ in property* 4. *Then,*

1. $m_X^{\pm}[-f] = -M_X^{\mp}[f]$,
2. $-L_X[f] \leq m_X^-[f] \leq m_X^+[f] \leq l_X[f]$,
3. $m_X^{\pm}[\alpha f] = \alpha m_X^{\pm}[f]$, $\alpha \geq 0$,
4. $m_X^-[f] + m_X^{\pm}[g] \leq m_X^{\pm}[f + g]$.

The first and second properties in Proposition 2.5 imply that it is preferable to use $m_X^+[f]$ and $M_X^-[f]$, since they correspond to the weakest requirements.

LEMMA 2.6. *If $l_X[f] > 0$, then f is injective and $L_X[f^{-1}] = l_X[f]^{-1}$.*

Proof. Definition 2.1 trivially yields that f is injective when $l_X[f] > 0$ and $f^{-1} : R(f) \to D(f)$ is bijective. Let $u_1, u_2 \in D(f)$, and $v_1 := f(u_1)$, $v_2 := f(u_2)$; then

$$l_X[f]^{-1} = \sup_{u_1 \neq u_2} \frac{\|u_1 - u_2\|_X}{\|f(u_1) - f(u_2)\|_X} = \sup_{v_1 \neq v_2} \frac{\|f^{-1}(v_1) - f^{-1}(v_2)\|_X}{\|v_1 - v_2\|_X} = L_X[f^{-1}]. \qquad \square$$

Lemma 2.6 together with the inequality $m_X^+[f] \le l_X[f]$ yields the following corollary.

COROLLARY 2.7. *If $m_X^+[f] > 0$, then f is injective and $L_X[f^{-1}] \le m_X^+[f]^{-1}$.*

3. Direct product spaces. Introduce the Banach space X^N, i.e., the direct product of N real valued Banach spaces, equipped with the norm $\|\cdot\|_{X,p}$. Elements $U \in X^N$ are denoted as $U = (U_1, \ldots, U_N)^T$, and the norm $\|\cdot\|_{X,p}$ on X^N is defined as

$$\|U\|_{X,p} := \left(\sum_{i=1}^{N} \|U_i\|_X^p \right)^{1/p},$$

when $p < \infty$ and $\|U\|_{X,\infty} := \max_{1 \le i \le N} \|U_i\|_X$. In the error analysis only two types of maps on X^N are needed. To every map f on X we relate the map $\mathcal{F} : D(f)^N \to X^N$ defined as

$$(\mathcal{F}(U))_i := f(U_i) \quad \text{for} \quad i = 1, \ldots, N.$$

The lemma below now follows trivially from the definition of $\|\cdot\|_{X,p}$.

LEMMA 3.1. *$L_{X,p}[\mathcal{F}] = L_X[f]$ and $l_{X,p}[\mathcal{F}] = l_X[f]$.*

Furthermore, to every real matrix $A = \{a_{ij}\}_{i,j=1}^N$ we relate the linear map $\mathcal{A} : X^N \to X^N$ defined as

$$(\mathcal{A}U)_i := \sum_{j=1}^{N} a_{ij} U_j \quad \text{for} \quad i = 1, \ldots, N.$$

LEMMA 3.2. *If $X = l_p(\mathbb{N})$, where $1 \le p < \infty$, equipped with the usual p-norm $\|\cdot\|_p$, then $L_{p,p}[\mathcal{A}] = L_p[A]$ and $l_{p,p}[\mathcal{A}] = l_p[A]$.*

Proof. Every element $u \in l_p(\mathbb{N})$ is the limit of a sequence $\{u^j\}_{j \ge 1}$ where $u^j \in X_j := \{v \in l_p(\mathbb{N}) : v_i = 0,\ \forall i > j\}$, as $\lim_{j \to \infty} \sum_{i>j} |u_i|^p = 0$ when $p < \infty$. Hence,

$$L_{p,p}[\mathcal{A}] = \sup_{j \ge 1} \sup_{\{U \in X_j^N :\, \|U\|_{p,p} = 1\}} \|A \otimes I_{j \times j} U\|_{p,p} \le L_p[A] \sup_{j \ge 1} L_p[I_{j \times j}] = L_p[A].$$

The equality is now obtained as $L_{p,p}[\mathcal{A}] \ge \sup_{\{u \in X_1^N :\, \|u\|_p = 1\}} \|Au\|_p = L_p[A]$. The other equality is obtained by the same procedure. ∎

COROLLARY 3.3. *If there exists a linear isometric imbedding of X into $l_p(\mathbb{N})$, where $1 \le p < \infty$, then $L_{X,p}[\mathcal{A}] \le L_p[A]$ and $l_{X,p}[\mathcal{A}] \ge l_p[A]$.*

Proof. Let the linear map ϕ be the isometric imbedding of X into $l_p(\mathbb{N})$, i.e., $\|\phi u\|_p = \|u\|_X$, and define $\Phi : X^N \to l_p^N(\mathbb{N})$ by $(\Phi U)_i := \phi U_i$ for $i = 1, \ldots, N$. It is easily seen that Φ is an isometric imbedding of X^N into $l_p^N(\mathbb{N})$ and the linearity of ϕ yields that $\Phi\mathcal{A} = \mathcal{A}\Phi$. Hence, by Lemma 3.2,

$$\|\mathcal{A}U\|_{X,p} = \|\Phi\mathcal{A}U\|_{p,p} = \|\mathcal{A}\Phi U\|_{p,p} \le L_{p,p}[\mathcal{A}] \|\Phi U\|_{p,p} = L_p[A] \|U\|_{X,p},$$

which implies that $L_{X,p}[\mathcal{A}] \le L_p[A]$. The other inequality in this corollary follows in the same fashion. ∎

Note that the inequalities related to the Lipschitz constants in Corollary 3.3 can be replaced by equalities if the imbedding is surjective, and in the case of X being a separable Hilbert space the corollary is valid with $p = 2$, as these spaces are linearly and isometrically imbedded into $l_2(\mathbb{N})$. Next, consider the semi-inner products $(\cdot,\cdot)_{X,p}^{\pm}$

on X^N generated by the norm $\|\cdot\|_{X,p}$ for which we introduce the concept of *proper* semi-inner products.

DEFINITION 3.4. *The semi-inner products* $(\cdot,\cdot)^{\pm}_{X,p}$ *are called* proper *if* $M^{\pm}_{X,p}[\mathcal{F}] = M^{\pm}_X[f]$ *and* $m^{\pm}_{X,p}[\mathcal{F}] = m^{\pm}_X[f]$.

LEMMA 3.5. *The semi-inner products* $(\cdot,\cdot)^{\pm}_{X,1}$ *and* $(\cdot,\cdot)^{\pm}_{X,2}$ *are proper.*

Proof. By Definition 2.3 and the construction of the norm $\|\cdot\|_{X,p}$ it follows that

$$(U,V)^{\pm}_{X,1} = \|U\|_{X,1}\sum_{i=1}^{N}\frac{(U_i,V_i)^{\pm}_X}{\|U_i\|_X} \quad \text{and} \quad (U,V)^{\pm}_{X,2} = \sum_{i=1}^{N}(U_i,V_i)^{\pm}_X.$$

The desired equalities are now obtained via these representations together with Definition 2.4. ∎

4. Problem setting. Let X be a real valued Banach space and consider the nonlinear evolution equation

$$\dot{u} = f(u), \quad u(0) = \eta \in D(f), \tag{4.1}$$

where $u : [0,\infty) \to X$ and the vector field f is a nonlinear map on X such that $M^{-}_X[f] < \infty$; i.e., f is dissipative [1], and

$$R(I - hf) = X \quad \forall h > 0 \text{ such that } hM^{-}_X[f] < 1. \tag{4.2}$$

Example 4.1. Define the evolution triple (V, X, V^*) by $V \subset X = X^* \subset V^*$ where $X := L_2[0,1]$ and $V := W^{1,r}_0[0,1]$, with $2 \le r < \infty$. Next, consider the perturbed r-Laplacian $\Delta_r : C^{\infty}_0[0,1] \to X$ defined as

$$\Delta_r : u \mapsto \partial_x(|\partial_x u|^{r-2}\partial_x u) + su,$$

with $s < 0$, and its energetic extension $\Delta_{E,r} : V \to V^*$, i.e.,

$$\Delta_{E,r} : u \mapsto \int_0^1 -(|\partial_x u|^{r-2}\partial_x u)\partial_x(\cdot) + su(\cdot)\mathrm{d}x.$$

Then, the map $f : D \to X$, with $D := \Delta^{-1}_{E,r}(X^*)$ and $\langle f(u), \cdot\rangle_X = \Delta_{E,r}(u)$ for all $u \in D$, fulfills (4.2), $L_X[f] = \infty$, and $M^{\pm}_X[f] \le s$; see sections 26.5 and 31.5 in [21] for details and generalizations.

Before the main propositions are stated, one needs the following definitions [2, 3].

DEFINITION 4.2. *A function* $u : [0,\infty) \to X$ *is said to be a strong solution of* (4.1) *on* $[0,\infty)$ *if* $u \in W^{1,1}([0,\infty), X)$, $u(0) = \eta$, *and* (4.1) *is satisfied a.e. on* $(0,\infty)$.

DEFINITION 4.3. *Let* Ω *be a closed subset of* X. *Then the set* $Q(\omega,\Omega)$ *of* ω*-type semigroups is defined as the collection of maps* $S : [0,\infty)\times\Omega \to \Omega$ *such that*

1. $S^{t+\tau}(u) = S^t S^\tau(u)$ *for all* $u \in \Omega$, $t,\tau \ge 0$,
2. $S^0(u) = u$ *for all* $u \in \Omega$,
3. $S^{(\cdot)}(u) \in C([0,\infty),\Omega)$ *for all* $u \in \Omega$,
4. $L_X[S^t(\cdot)] \le e^{t\omega}$ *for all* $t \ge 0$.

Now, in the setting of (4.1), one may apply the standard nonlinear semigroup theory for dissipative maps; see [1, 3, 21] for proofs and further results.

PROPOSITION 4.4. *For all* $u \in \overline{D(f)}$ *and* $t \ge 0$, *the evolution map*

$$e^{tf}(u) := \lim_{n\to\infty}\left(I - \frac{t}{n}f\right)^{-n}(u)$$

exists, and $e^{(\cdot)f} \in Q(M^{-}_X[f], \overline{D(f)})$.

PROPOSITION 4.5. *If X is a reflexive Banach space, then $e^{(\cdot)f}(\eta)$ is the unique strong solution of* (4.1) *on* $[0,\infty)$.

By comparing the hypothesis of Propositions 4.4 and 4.5, we note that the evolution map can be well defined even if (4.1) does not have a strong solution. This may occur if $\eta \in \overline{D(f)}\backslash D(f)$ or if X is not a reflexive Banach space. Thus, the most general numerical setting is obtained if one considers approximations of the evolution map, i.e., finding a sequence $\{u_i\}$ in X for every fixed $t \in [0,\infty)$ and $\eta \in \overline{D(f)}$ such that $e^{tf}(\eta) = \lim_{i\to\infty} u_i$, rather then approximating the strong solution of (4.1), which is not even necessarily well defined for all $t \in (0,\infty)$.

The multistep approximation $u_n \in X$ of $e^{t_n f}(\eta)$ is defined by the difference equation

$$\text{(4.3)} \qquad \begin{cases} h^{-1}\displaystyle\sum_{i=0}^{k} \alpha_{k-i} u_{n-i} = \sum_{i=0}^{k} \beta_{k-i} f(u_{n-i}), & n \geq 1, \\ \\ u_n = e^{t_n f}(\eta), & n = 1-k, \dots, 0, \end{cases}$$

where $t_n := (n+k-1)h$, $\alpha_k \neq 0$, and α_0 or $\beta_0 \neq 0$. To every multistep method relate the real polynomials ρ and σ, where

$$\rho(\xi) := \sum_{i=0}^{k} \alpha_i \xi^i \quad \text{and} \quad \sigma(\xi) := \sum_{i=0}^{k} \beta_i \xi^i.$$

Denote the sets of roots to ρ and σ by $\{\xi_i^\rho\}_{i=1}^k$ and $\{\xi_i^\sigma\}_{i=1}^k$, respectively. In order for the approximation u_n to be well defined one needs the following assumption:

(A1) $e^{t_n f}(\eta) \in D(f)$ for $n = 1-k, \dots, 0$.

This holds, e.g., for all $t \geq 0$ when $\eta \in D(f)$ and X is a Hilbert space; see [21].

THEOREM 4.6. *If* (A1) *holds and the multistep method (ρ,σ) is implicit, i.e., $\beta_k/\alpha_k > 0$, and $h\beta_k/\alpha_k M_X^-[f] < 1$, then there exists a unique solution to* (4.3).

Proof. Assume that u_{n-k} to u_{n-1} are known elements in $D(f)$. As

$$m_X^+\left[I - h\frac{\beta_k}{\alpha_k} f\right] \geq 1 - h\frac{\beta_k}{\alpha_k} M_X^-[f] > 0,$$

one has, by (4.2) and Corollary 2.7, that $(I - h\beta_k/\alpha_k f)^{-1} : X \to D(f)$ is a bijection and

$$u_n = \left(I - h\frac{\beta_k}{\alpha_k} f\right)^{-1} \left(\frac{1}{\alpha_k} \sum_{i=1}^{k} (-\alpha_{k-i} I + h\beta_{k-i} f)(u_{n-i})\right).$$

By (A1), $u_n \in D(f)$ for $n = 1-k, \dots, 0$, and the proof of the theorem now follows by induction. ∎

Note that the implicitness of the method is a necessity for the existence of a solution to (4.3), as an explicit method with $u_n \in D(f)$ may yield $u_{n+1} \in X\backslash D(f)$ and u_{n+2} is therefore not generally well defined.

DEFINITION 4.7. *A multistep method (ρ,σ) is said to fulfill the stability property*

(S1) *if $|\xi_i^\rho| < 1$ for $i = 1, \dots, k-1$, and $\xi_k^\rho = 1$;*

(S2) *if $|\xi_i^\sigma| < 1$ for $i = 1, \dots, k$;*

(S3) *if $\mathrm{Re}\{\rho(\xi)/\sigma(\xi)\} > 0$ for $|\xi| > 1$.*

The terminology of stability properties varies somewhat in the literature, but most commonly properties (S1), (S2), and (S3) are referred to as strong zero-stability, A_∞-stability, and A-stability, respectively. Methods fulfilling both (S2) and (S3) are usually referred to as strongly A-stable.

DEFINITION 4.8. *A multistep method (ρ, σ) is said to be consistent of order p if*

$$\rho(1) = 0 \quad \textit{and} \quad \sum_{i=0}^{k} \alpha_i i^q = q \sum_{i=0}^{k} \beta_i i^{q-1} \quad \textit{for} \quad q = 1, \ldots, p.$$

The classical ODE-analysis of the global error

$$e_n := e^{t_n f}(\eta) - u_n, \quad n \geq 1,$$

is often based on the order, with respect to h, of the local residual

$$w_n := \sum_{i=0}^{k} (h^{-1}\alpha_{k-i} I - \beta_{k-i} f) \left(e^{t_{n-i} f}(\eta) \right), \quad n \geq 1.$$

The order is usually obtained by a Taylor expansion. If the same approach is to be used here, the following additional assumptions on the evolution map are sufficient:

(A2) $\frac{\mathrm{d}}{\mathrm{d}t}(e^{tf}(\eta)) = f(e^{tf}(\eta))$ for all $t \geq 0$,

(A3) $e^{(\cdot)f}(\eta) \in C^{p+1}([0, \infty), X)$.

These assumptions imply the standard result related to the order of the local residual stated below.

THEOREM 4.9. *If* (A1)–(A3) *hold and the multistep method (ρ, σ) is consistent of order p, then*

$$\|w_n\|_X = O(h^p) \quad \forall n \geq 1.$$

Proof. Assumptions (A1)–(A3) imply that $e^{(\cdot)f}(\eta)$ can be written as a p-order Taylor expansion, since integration by parts is possible in the context of Bochner integrals. Thus, the proof follows as in the ODE-setting; see [7]. ∎

Note that (A1)–(A3) hold if f is a linear map and $\eta \in D(f^{p+1})$; see [15].

5. Global error analysis. The aim is now to derive bounds of the global error by formulating (4.3) and the definition of the local residual as equations in X^N. To this end, introduce the nonlinear map $\mathcal{F} : D(f)^N \to X^N$ related to the vector field f on X and the linear maps $\mathcal{P}$, $\mathcal{S} : X^N \to X^N$ related to the $N \times N$-Toeplitz matrices

$$P := \begin{pmatrix} \alpha_k & \ldots & \alpha_0 & & \\ & \ddots & & & \\ & & \alpha_k & \ldots & \alpha_0 \\ & & & \ddots & \\ & & & & \alpha_k \end{pmatrix} \quad \text{and} \quad S := \begin{pmatrix} \beta_k & \ldots & \beta_0 & & \\ & \ddots & & & \\ & & \beta_k & \ldots & \beta_0 \\ & & & \ddots & \\ & & & & \beta_k \end{pmatrix}.$$

Next, define the nonlinear map $\mathcal{H} : D(f)^N \to X^N$ as

$$\mathcal{H}(U) := h^{-1}\mathcal{P}U - \mathcal{S}\mathcal{F}(U) + V_0,$$

where V_0 is a constant vector in X^N with k nonzero elements given by the initial data. Thus, the numerical approximations $U := (u_N, \ldots, u_1)^T$, the analytic solutions

$V := (e^{t_N f}(\eta), \ldots, e^{t_1 f}(\eta))^T$, and the local residuals $W := (w_N, \ldots, w_1)^T$ fulfill the equations $\mathcal{H}(V) = W$ and $\mathcal{H}(U) = 0$. Hence, $\mathcal{H}(V) - \mathcal{H}(U) = W$; i.e.,

$$h^{-1}\mathcal{P}(V - U) - \mathcal{S}(\mathcal{F}(V) - \mathcal{F}(U)) = W. \tag{5.1}$$

It is possible to derive bounds for the global errors $E := V - U$ from (5.1) if the topology imposed on X^N is correctly chosen. Before the bounds are derived, one needs to relate the stability properties presented in Definition 4.7 to the linear maps $\mathcal{P}$ and $\mathcal{S}$.

LEMMA 5.1. *There exist positive constants C_ρ and C_σ such that if the multistep method (ρ, σ) fulfills*

1. (S1)*, then $L_{X,p}[\mathcal{P}^{-1}] \le C_\rho N$;*
2. (S2) *and is implicit, then $L_{X,p}[\mathcal{S}^{-1}] \le C_\sigma$ for all $N \ge 1$.*

Proof. As P is an upper triangular Toeplitz matrix of size $N \times N$ with one $(\alpha_k \neq 0)$ to $k+1$ nonzero diagonals, the map $\mathcal{P}$ can be represented as

$$\mathcal{P} = \alpha_k \prod_{i=1}^{k} (I - \xi_i^\rho \mathcal{E}), \tag{5.2}$$

where $\mathcal{E}$ is the linear map related to the left shift matrix

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

Since $\mathcal{E}$ is nilpotent of order N, one has that

$$(I - \xi_i^\rho \mathcal{E})^{-1} = \sum_{n=0}^{N-1} (\xi_i^\rho)^n \mathcal{E}^n.$$

Hence, $\mathcal{P}^{-1} : X^N \to X^N$ is well defined and

$$L_{X,p}[\mathcal{P}^{-1}] \le \frac{1}{|\alpha_k|} \prod_{i=1}^{k} \sum_{n=0}^{N-1} |\xi_i^\rho|^n L_{X,p}[\mathcal{E}]^n.$$

The first assertion is now obtained as $L_{X,p}[\mathcal{E}] = 1$, and the method (ρ, σ) fulfills stability property (S1); i.e., $|\xi_i^\rho| < 1$ for $i = 1, \ldots, k-1$ and $\xi_k^\rho = 1$. The second assertion follows by making the same factorization of $\mathcal{S}$ as made for $\mathcal{P}$ in the previous part, which is possible as the method is implicit $(\beta_k \neq 0)$, and by using the fact that the method (ρ, σ) fulfills stability property (S2), i.e., $|\xi_i^\sigma| < 1$ for $i = 1, \ldots, k$. ∎

LEMMA 5.2. *If X is a separable Hilbert space and the implicit multistep method (ρ, σ) fulfills* (S2) *and* (S3)*, then $m^{\pm}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] \ge 0$.*

Proof. Corollary 3.3, Lemma 5.1, and the factorization of $\mathcal{P}$ in (5.2) yields that $L_{X,2}[\mathcal{S}^{-1}\mathcal{P}] \le L_2[S^{-1}P] \le C_\sigma |\alpha_k| 2^k$ for all $N \ge 1$. Hence,

$$\begin{aligned} M^+_{X,2}[\mathcal{S}^{-1}\mathcal{P}] &= \sup_{\{U \in X^N : \|U\|_{X,2} = 1\}} \lim_{\varepsilon \to 0+} \frac{\|(\mathcal{I} + \varepsilon \mathcal{S}^{-1}\mathcal{P})U\|_{X,2} - \|U\|_{X,2}}{\varepsilon} \\ &\le \lim_{\varepsilon \to 0+} \frac{L_{X,2}[\mathcal{I} + \varepsilon \mathcal{S}^{-1}\mathcal{P}] - 1}{\varepsilon} \le \lim_{\varepsilon \to 0+} \frac{L_2[I + \varepsilon S^{-1}P] - 1}{\varepsilon} \\ &= \max\{\lambda : \lambda \in \operatorname{spec}((S^{-1}P + P^T S^{-T})/2)\}. \end{aligned}$$

The last equality follows by the well-known theory of logarithmic norms; see [7]. Thus, by Proposition 2.5,

$$m^{\pm}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] \geq -M^{+}_{X,2}[-\mathcal{S}^{-1}\mathcal{P}] \geq \min\{\lambda : \lambda \in \operatorname{spec}((S^{-1}P + P^T S^{-T})/2)\}.$$

Using the analysis of Toeplitz matrices on $l_2(\mathbb{N})$ one has the bound

$$\min\{\lambda : \lambda \in \operatorname{spec}((S^{-1}P + P^T S^{-T})/2)\} \geq \min_{|\xi|=1} \operatorname{Re}\{\rho(\xi)/\sigma(\xi)\}.$$

See [20] for general theory and [10] for the applications in connection to multistep methods. As the method (ρ, σ) fulfills stability properties (S2) and (S3), the function $r : \xi \mapsto \operatorname{Re}\{\rho(\xi)/\sigma(\xi)\}$ is continuous for $|\xi| \geq 1$ and $r(\xi) > 0$ for all $|\xi| > 1$. Hence, $r(\xi) \geq 0$ for $|\xi| = 1$. Combining these results gives the bound

$$m^{\pm}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] \geq \min_{|\xi|=1} \operatorname{Re}\{\rho(\xi)/\sigma(\xi)\} \geq 0,$$

which concludes the proof. ∎

It is now possible to derive a discrete L_2-bound of the global error for an implicit multistep approximation on Hilbert spaces, when the vector field f is (strictly) dissipative but otherwise unbounded.

THEOREM 5.3. *Consider the approximation of $e^{t_N f}(\eta)$, using the implicit multistep method (ρ, σ) on the separable Hilbert space X. If* (A1)–(A3) *hold, $M^{-}_{X}[f] < 0$, and the method (ρ, σ) is consistent of order p and fulfills* (S2) *and* (S3)*, then*

$$\left(h \sum_{i=1}^{N} \|e_i\|^2_X \right)^{1/2} \leq C \frac{\sqrt{t_N}}{-M^{-}_{X}[f]} h^p,$$

where the positive constant C is independent of N, $M^{-}_{X}[f]$, and t_N.

Proof. Equip X^N with the norm $\|\cdot\|_{X,2}$. Note that this norm induces the proper semi-inner products $(\cdot,\cdot)^{\pm}_{X,2}$; see Lemma 3.5. The hypotheses imposed on X and the method (ρ, σ) together with Lemmas 5.1 and 5.2 imply that $\mathcal{S}^{-1}$ is well defined, $L_{X,2}[\mathcal{S}^{-1}] \leq C_\sigma$, and $m^{\pm}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] \geq 0$. It is therefore possible to apply the functional $(E, \mathcal{S}^{-1}(\cdot))^{+}_{X,2}$ to (5.1), which yields the bound

$$\begin{aligned}
(E, \mathcal{S}^{-1}W)^{+}_{X,2} &= (E, h^{-1}\mathcal{S}^{-1}\mathcal{P}(V - U) - [\mathcal{F}(V) - \mathcal{F}(U)])^{+}_{X,2} \\
&\geq m^{+}_{X,2}[h^{-1}\mathcal{S}^{-1}\mathcal{P} - \mathcal{F}]\|E\|^2_{X,2} \\
&\geq (h^{-1} m^{-}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] - M^{-}_{X}[f])\|E\|^2_{X,2}.
\end{aligned}$$

Assume that $M^{-}_{X}[f] < 0$ and use the right Cauchy–Schwarz inequality on the term $(E, \mathcal{S}^{-1}W)^{+}_{X,2}$; then

$$\|E\|_{X,2} \leq \frac{L_{X,2}[\mathcal{S}^{-1}]}{h^{-1} m^{-}_{X,2}[\mathcal{S}^{-1}\mathcal{P}] - M^{-}_{X}[f]} \|W\|_{X,2} \leq \frac{C_\sigma}{-M^{-}_{X}[f]} \|W\|_{X,2}.$$

The desired bound is now obtained by Theorem 4.9 together with the inequality

$$\|W\|_{X,2} \leq \sqrt{N} \max_{1 \leq i \leq N} \|w_i\|_X \leq \sqrt{\frac{t_N}{h}} \max_{1 \leq i \leq N} \|w_i\|_X. \qquad \blacksquare$$

TABLE 5.1
Numerically observed convergence orders.

Δx	IE	BDF2
2^{-7}	$p = 1.0426$	$p = 1.9868$
2^{-8}	$p = 1.0426$	$p = 1.9867$
2^{-9}	$p = 1.0426$	$p = 1.9867$

It is easy to see how the mathematical problem influences the global error bound: the more dissipative the vector field is, i.e., the smaller $M_X^-[f]$, the smaller is the error. One important application of Theorem 5.3 is in the context of nonlinear parabolic PDEs where $X = L_2(\Omega)$. We illustrate this by the numerical example below.

Example 5.4. Consider the evolution governed by the dissipative extension $f : D \to L_2[0,1]$ of the perturbed r-Laplacian

$$\Delta_r : u \mapsto \partial_x(|\partial_x u|^{r-2}\partial_x u) + su$$

presented in Example 4.1. The evolution equation is discretized in space by finite differences on the equidistant grid $\{0, \Delta x, 2\Delta x, \ldots, K\Delta x, 1\}$, i.e., we consider the evolution governed by the vector field $f_{\Delta x} : X_{\Delta x} \to X_{\Delta x}$ defined as

$$(f_{\Delta x}(u))^i := \frac{1}{\Delta x}\left(g\left(\frac{u^{i+1}-u^i}{\Delta x}\right) - g\left(\frac{u^i - u^{i-1}}{\Delta x}\right)\right) + su^i \quad \text{for} \quad i = 1, \ldots, K,$$

where $g = |\cdot|^{r-2}(\cdot)$ and $u^0 = u^{K+1} = 0$. Furthermore, $X_{\Delta x} = (\mathbb{R}^K, \|\cdot\|_{\Delta x})$ with

$$\|u\|_{\Delta x} := \left(\Delta x \sum_{i=1}^{K} |u^i|^2\right)^{1/2}.$$

As Theorem 5.3 requires the multistep method (ρ, σ) to fulfill stability property (S3), we are restricted to methods which are consistent of order one or two, e.g., the implicit Euler $(\xi - 1, \xi)$ and the BDF2 $(3\xi^2/2 - 2\xi + 1/2, \xi^2)$ methods; see [8]. In Table 5.1 the numerically observed convergence orders are given for these multistep discretizations applied to the evolution equation

$$\dot{u} = f_{\Delta x}(u), \quad u(0) = \eta_{\Delta x},$$

with $\eta^i_{\Delta x} = 5e^{-1/i\Delta x(1-i\Delta x)}$, $\{r, s\} = \{3, -1\}$, and $t_N = 0.3$. In the implementation of both methods, the solution of (4.3) was approximated by using Newton's method with an exact Jacobian which is well defined as g is differentiable. The second starting value for the BDF2 method was obtained by an implicit Euler step. The errors were estimated by computing the numerical solutions for $h/t_N \in [h_1, h_2]$ and comparing with the one obtained with $h/t_N = h_0$. The parameters $\{h_0, h_1, h_2\}$ were chosen as $\{2^{-13}, 2^{-10}, 2^{-7}\}$ and $\{2^{-15}, 2^{-13}, 2^{-10}\}$ for the implicit Euler and the BDF2 implementations, respectively. As seen in Table 5.1, the observed convergence orders are independent of Δx and in agreement with Theorem 5.3.

In Theorem 5.3 we obtained the same convergence orders in the discrete L_2-norm as presented in [10] for ODEs, but it is also possible to derive error bounds in the infinity-norm on arbitrary Banach space if one considers approximations of an evolution governed by a Lipschitz continuous vector field on a restricted time interval.

THEOREM 5.5. *Approximate $e^{t_N f}(\eta)$, where $t_N \in (0, 1/(C_1 L_X[f]))$, by using the implicit multistep method (ρ, σ). If* (A1)–(A3) *hold, $h\beta_k/\alpha_k M_X^-[f] < 1$, and the method (ρ, σ) is consistent of order p and fulfills* (S1)*, then*

$$\max_{1 \le i \le N} \|e_i\|_X \le \frac{C_2 t_N}{1 - C_1 t_N L_X[f]} h^p,$$

where the positive constants C_1 and C_2 are independent of N, $L_X[f]$, and t_N.

Proof. Equip X^N with the norm $\|\cdot\|_{X,\infty}$. The hypotheses on the method (ρ, σ) and Lemma 5.1 give that the map $\mathcal{P}^{-1}$ is well defined and that $L_{X,\infty}[\mathcal{P}^{-1}] \le C_\rho N$. Furthermore, factorizing $\mathcal{S}$ as in (5.2) yields that $L_{X,\infty}[\mathcal{S}] \le |\beta_k| 2^k$. Thus, one can apply the functional $\|\mathcal{P}^{-1}(\cdot)\|_{X,\infty}$ to (5.1) and the following bound is obtained:

$$\begin{aligned} \|\mathcal{P}^{-1} W\|_{X,\infty} &= \|h^{-1}(V - U) - \mathcal{P}^{-1}\mathcal{S}(\mathcal{F}(V) - \mathcal{F}(U))\|_{X,\infty} \\ &\ge l_{X,\infty}[h^{-1}\mathcal{I} - \mathcal{P}^{-1}\mathcal{S}\mathcal{F}]\|E\|_{X,\infty} \\ &\ge (h^{-1} - L_{X,\infty}[\mathcal{P}^{-1}] L_{X,\infty}[\mathcal{S}] L_X[f])\|E\|_{X,\infty}. \end{aligned}$$

Assume that $C_1 t_N L_X[f] < 1$, with $C_1 := C_\rho |\beta_k| 2^k$; then

$$\|E\|_{X,\infty} \le \frac{L_{X,\infty}[\mathcal{P}^{-1}]}{h^{-1} - C_1 N L_X[f]} \|W\|_{X,\infty} \le \frac{C_\rho t_N}{1 - C_1 t_N L_X[f]} \|W\|_{X,\infty},$$

and the desired bound follows by Theorem 4.9. ∎

Note how the global error decreases and the time interval $(0, 1/(C_1 L_X[f]))$ increases as the contractive behavior of the vector field increases, i.e., as $L_X[f]$ tends to zero.

6. Conclusions. Using the theory of logarithmic Lipschitz constants, global error bounds have been derived in the setting of multistep time discretizations of fully nonlinear evolution equations on infinite dimensional spaces. The bounds clearly reveal how the contractive or dissipative behavior of the vector field, governing the evolution, and the properties of the multistep method influence the convergence. Furthermore, the obtained convergence results are closely related to the ones found in the ODE-context.

Acknowledgment. The author would like to thank Gustaf Söderlind for introducing the author to the topic of this paper and for all the inspiring discussions.

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leiden, The Netherlands, 1976.
[2] M. G. CRANDALL, *Nonlinear semigroups and evolution governed by accretive operators*, Proc. Sympos. Pure Math., 45 (1986), pp. 305–337.
[3] M. G. CRANDALL AND T. M. LIGGETT, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
[4] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.
[5] C. GONZÁLEZ, A. OSTERMANN, C. PALENCIA, AND M. THALHAMMER, *Backward Euler discretization of fully nonlinear parabolic problems*, Math. Comp., 71 (2002), pp. 125–145.
[6] C. GONZÁLEZ AND C. PALENCIA, *Stability of time-stepping methods for abstract time-dependent parabolic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 973–989.
[7] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations.* I. *Nonstiff Problems*, 2nd ed., Springer-Verlag, Berlin, 1993.
[8] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.

[9] I. Higueras and G. Söderlind, *Logarithmic norms and nonlinear DAE stability*, BIT, 42 (2002), pp. 823–841.
[10] F. Iavernaro and F. Mazzia, *Convergence and stability of multistep methods solving nonlinear initial value problems*, SIAM J. Sci. Comput., 18 (1997), pp. 270–285.
[11] O. Nevanlinna, *On the numerical integration of nonlinear initial value problems by linear multistep methods*, BIT, 17 (1977), pp. 58–71.
[12] O. Nevanlinna, *On the convergence of difference approximations to nonlinear contraction semigroups in Hilbert spaces*, Math. Comp., 32 (1978), pp. 321–334.
[13] A. Ostermann, M. Thalhammer, and G. Kirlinger, *Stability of linear multistep methods and applications to nonlinear parabolic problems*, Appl. Numer. Math., 48 (2004), pp. 389–407.
[14] C. Palencia, *Stability of rational multistep approximations of holomorphic semigroups*, Math. Comp., 64 (1995), pp. 591–599.
[15] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[16] M. Le Roux, *Méthodes multipas pour des équations paraboliques non linéaires*, Numer. Math., 35 (1980), pp. 143–162.
[17] G. Söderlind, *On nonlinear difference and differential equations*, BIT, 24 (1984), pp. 667–680.
[18] G. Söderlind, *Bounds on nonlinear operators in finite-dimensional Banach spaces*, Numer. Math., 50 (1986), pp. 27–44.
[19] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, New York, 1997.
[20] H. Widom, *Toeplitz matrices*, in Studies in Real and Complex Analysis, I. I. Hirschman, ed., Mathematical Association of America, Buffalo, NY, 1965, pp. 179–201.
[21] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* II/*B. Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 66–81

SEMI-ITERATIVE REGULARIZATION IN HILBERT SCALES*

HERBERT EGGER†

Abstract. In this paper we investigate the regularizing properties of semi-iterative regularization methods in Hilbert scales for linear ill-posed problems and perturbed data.

It is well known that standard Landweber iteration can be remarkably accelerated by polynomial acceleration methods leading to *optimal speed of convergence*, which can be obtained by several efficient two-step methods, e.g., the ν-methods by Brakhage. It was observed earlier that a similar speed of convergence, i.e., similar iteration numbers yielding optimal convergence rates, can be obtained if Landweber iteration is performed in Hilbert scales.

We show that a combination of both ideas allows for a further acceleration, yielding optimal convergence rates with only the square root of iterations as compared to the ν-methods or Landweber iteration in Hilbert scales. The theoretical results are illustrated by several examples and numerical tests, including a comparison to the method of conjugate gradients.

Key words. inverse problems, regularization, Hilbert scales, semi-iterative methods

AMS subject classifications. 15A29, 65F10, 65F22, 65J22

DOI. 10.1137/040617285

1. Introduction.

In this paper, we study inverse problems of the form

$$Tx = y, \tag{1.1}$$

where $T : \mathcal{X} \to \mathcal{Y}$ is a linear bounded operator between infinite-dimensional Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with range $\mathcal{R}(T) \subset \mathcal{Y}$. It is well known (see, e.g., [6]) that the Moore–Penrose inverse $T^\dagger$, which is defined on $\mathcal{D}(T^\dagger) = \mathcal{R}(T) + \mathcal{R}(T)^\perp$ and maps data $y \in \mathcal{D}(T^\dagger)$ onto the best approximate solution $x^\dagger$, is unbounded if $\mathcal{R}(T)$ is not closed, and hence the solution of (1.1) is ill-posed; in particular, even for $y \in \mathcal{D}(T^\dagger)$ a solution of (1.1) does not depend continuously on the right-hand side and thus has to be regularized.

Especially for large-scale problems, iterative regularization algorithms have turned out to be an attractive alternative to Tikhonov regularization, which is probably the most well-known regularization method (see, e.g., [4, 7]). Application of Landweber iteration (cf. [12]),

$$x_k = x_{k-1} + \omega T^*(y - Tx_{k-1}), \qquad k \geq 1, \tag{1.2}$$

with $0 < \omega < 2/\|T^*T\|$ to the solution of inverse problems has been investigated intensively in the literature (see, e.g., [1, 4, 9] and the references cited there). Formally, (1.2) corresponds to Richardson iteration (successive approximation) applied to the normal equation

$$T^*Tx = T^*y.$$

*Received by the editors October 20, 2004; accepted for publication (in revised form) July 25, 2005; published electronically February 8, 2006. This work was supported by the Austrian Science Foundation (FWF) under grant SFB F013, "Numerical and Symbolic Scientific Computing," project F1308.

http://www.siam.org/journals/sinum/44-1/61728.html

†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Science, Altenbergerstrasse 69, A-4040 Linz, Austria (herbert.egger@oeaw.ac.at).

If $y \in \mathcal{D}(T^\dagger)$, then the iterates x_k converge to $T^\dagger y$; if, however, only perturbed data y^δ with a known upper bound on the noise level

$$\|y - y^\delta\| \le \delta \tag{1.3}$$

are known and $y^\delta \notin \mathcal{D}(T^\dagger)$, which is most probable if (1.1) is ill-posed and $\mathcal{R}(T)$ is not closed, then $\|x_k\|$ tends to infinity.

Iterative methods are turned into regularization algorithms by stopping the iteration after an adequate number k_* of steps. Besides a priori stopping rules, which require knowledge of the smoothness of $x^\dagger - x_0$ in terms of spaces $\mathcal{R}((T^*T)^\mu)$, the discrepancy principle (cf. [4, 14])

$$\|y^\delta - Tx_{k_*}\| \le \tau\delta < \|y^\delta - Tx_k\|, \qquad 0 \le k < k_*, \tag{1.4}$$

with $\tau > 1$ has turned out to be an appropriate a posteriori stopping rule yielding optimal convergence rates for Landweber iteration; i.e., if

$$x^\dagger - x_0 \in \mathcal{R}((T^*T)^\mu), \qquad \mu > 0, \tag{1.5}$$

then stopping by the discrepancy principle (1.4) yields (see, e.g., [4])

$$\|x_{k_*}^\delta - x^\dagger\| = o(\delta^{\frac{2\mu}{2\mu+1}}) \quad \text{and} \quad k_* = O(\delta^{-\frac{2}{2\mu+1}}). \tag{1.6}$$

The main drawback of Landweber iteration is its slow performance, i.e., a large number of iterations needed to obtain the optimal convergence rates (1.6). In order to accelerate the solution process, several *semi-iterative methods* (polynomial acceleration methods) have been proposed and analyzed in the framework of regularization (see, e.g., [8] for an overview). In our numerical experiments, we will use Brakhage's ν-methods [2], for which the number of iteration can be bounded by

$$k_* \sim \delta^{-\frac{1}{2\mu+1}} \quad \text{for} \quad 0 < \mu \le \nu - \tfrac{1}{2} \tag{1.7}$$

in case of stopping according to the discrepancy principle (1.4) (see Theorem 2.1). Thus optimal convergence rates can be obtained with approximately the square root of iterations than needed for ordinary Landweber iteration. Note that in contrast to Landweber iteration, the ν-methods show a saturation phenomenon; i.e., the optimal rates and (1.7) hold only for $\mu \le \nu$, respectively, $\mu \le \nu - 1/2$, if the iteration is stopped according to (1.4).

Regularization in Hilbert scales was introduced by Nătterer [15] in the framework of Tikhonov regularization for linear problems. More recently, the Hilbert scale approach has been investigated also for more general regularization methods for linear problems (see [4, 19]) by means of spectral theory. Originally, Hilbert scales were used to increase the range of optimal convergence for Tikhonov regularization [15, 17] and Landweber iteration for nonlinear inverse problems [18]. In [3], the case $s < 0$ (undersmoothing) was investigated in more detail for Landweber iteration, and it was shown that the application of Hilbert scales can be understood as preconditioning in this case, i.e., the number of iterations needed to get optimal convergence can be reduced, essentially, to (1.7). Additionally, the results in [3] were derived under more general than the usual assumptions for regularization in Hilbert scales (cf. section 3).

The aim of this paper is to show that the combination of polynomial acceleration methods with the Hilbert scale approach leads to a further acceleration of iterative

regularization methods, yielding optimal rates of convergence with stopping indices bounded by $k_* = O(\delta^{-\frac{1}{2(2\mu+1)}})$. In case of mildly ill-posed problems, i.e., if the singular values σ_n of the operator T decay like $O(n^{-\alpha})$ for some $0 < \alpha < 1$, the performance may even be better than that of the conjugate gradient method in the standard spaces, where we have $k_* \leq \delta^{-\frac{1}{(2\mu+1)(1+\alpha)}}$ (cf. [4, Theorem 7.14]). The faster convergence of a Hilbert scale ν-method is also illustrated numerically in Examples 5.1 and 5.3. At least formally, preconditioning in Hilbert scales can also be applied to the conjugate gradient method; a rigorous analysis of such a method will be the subject of a subsequent article.

The paper is organized as follows: in sections 2 and 3, we briefly repeat the main results on convergence of semi-iterative regularization methods and regularization in Hilbert scales. The convergence analysis for semi-iterative methods in Hilbert scales is presented in section 4. We conclude with numerical examples comparing the proposed method to the standard Landweber iteration and ν-methods, Landweber iteration in Hilbert scales, and the conjugate gradient method.

2. Accelerated Landweber methods. While for Landweber iteration (1.2) only information about the last iterate x_{k-1}^δ is used to construct the new approximation x_k^δ, semi-iterative methods make use of all the approximations for $T^\dagger y$ obtained so far: a basic step of a semi-iterative method has the form (cf. [4, 8])

$$\begin{aligned} x_k^\delta &= \mu_{1,k} x_{k-1}^\delta + \cdots + \mu_{k,k} x_0^\delta + \omega_k T^*(y^\delta - T x_{k-1}^\delta), \quad k \geq 1, \\ &\textstyle\sum_{i=1}^k \mu_{i,k} = 1, \qquad \omega_k \neq 0, \end{aligned} \tag{2.1}$$

where we set $x_0^\delta = x_0$. The iterates defined by (2.1) with y^δ replaced by the true data y will be denoted by x_k. Obviously $x_k^\delta - x_0 \in \mathcal{K}_k(T^*T, T^*(y^\delta - Tx_0))$, where

$$\mathcal{K}_k(T^*T, p) := \operatorname{span}\{p, T^*Tp, \dots, (T^*T)^{k-1}p\}$$

denotes the kth Krylov subspace of T^*T with respect to p. Consequently, there exist polynomials $g_k(\lambda)$ and $r_k(\lambda) := 1 - \lambda g_k(\lambda)$ of degree $(k-1)$, respectively, k, such that

$$x_k - x^\dagger = r_k(T^*T)(x_0 - x^\dagger) \qquad \text{and} \qquad x_k^\delta - x_k = g_k(T^*T)T^*(y^\delta - y).$$

In other words, the approximation error $x_k - x^\dagger$ is determined by the *residual polynomials* r_k, while the propagated data error $x_k^\delta - x_k$ is determined by the *iteration polynomials* g_k. Of particular importance are methods whose residual polynomials r_k form an orthogonal sequence with respect to some positive weight function. In this case, the residual polynomials satisfy a three-term recurrence, which also carries over to the iterates; i.e., there exist sequences μ_k and ω_k such that

$$x_k^\delta = x_{k-1}^\delta + \mu_k(x_{k-1}^\delta - x_{k-2}^\delta) + \omega_k T^*(y^\delta - T x_{k-1}^\delta), \quad k \geq 1,$$

with $x_0^\delta = x_{-1}^\delta = x_0$. A specific instance of such methods are the ν-methods by Brakhage [2], which are defined by $\mu_1 = 0$, $\omega_1 = (4\nu+2)/(4\nu+1)$, and

$$\begin{aligned} \mu_k &= \frac{(k-1)(2k-3)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)}, \\ \omega_k &= 4\frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+2\nu-1)}, \qquad k > 1. \end{aligned}$$

Each ν-method has *optimal speed of convergence* for $x^\dagger - x_0 \in \mathcal{R}((T^*T)^\mu)$ with $0 \le \mu \le \nu$; i.e., its residual polynomials satisfy

$$\|\lambda^\mu r_k(\lambda)\|_{\mathcal{C}[0,1]} = O(k^{-2\mu}) \tag{2.2}$$

for $\mu \le \nu$. In fact, it seems that the estimate (2.2) is the best possible in terms of powers of k for semi-iterative methods with real orthogonal polynomials, cf. [8].

The following theorem (cf. [4, Theorem 6.11]) guarantees convergence rates of optimal order for semi-iterative regularization methods satisfying (2.2) if they are equipped with appropriate stopping rules.

THEOREM 2.1. *Let $y \in \mathcal{R}(T)$, and let the residual polynomials r_k satisfy* (2.2) *for some $\mu_0 > 0$. Then the semi-iterative method* (2.1) *is a regularization method of optimal order for $T^\dagger y \in \mathcal{R}((T^*T)^\mu)$ with $0 < \mu \le \mu_0 - 1/2$, provided the iteration is stopped with $k_* = k_*(\delta, y^\delta)$ according to the discrepancy principle* (1.4) *with fixed $\tau > \sup_{k\in\mathcal{N}} \|r_k\|_{\mathcal{C}[0,1]}$. In this case we have $k_* = O(\delta^{-\frac{1}{2\mu+1}})$ and $\|x_k^\delta - x^\dagger\| = O(\delta^{\frac{2\mu}{2\mu+1}})$.*

Note that even $o(\cdot)$ can be derived for the error $\|x_k^\delta - x^\dagger\|$ (see [4]).

3. Regularization in Hilbert scales. Before we recall some results on regularization in Hilbert scales, we briefly repeat the definition of a Hilbert scale (see [11]): let L be a densely defined, unbounded, self-adjoint and strictly positive operator in $\mathcal{X}$. Then $(\mathcal{X}_s)_{s\in\mathbb{R}}$ denotes the Hilbert scale induced by L if $\mathcal{X}_s$ is the completion of $\bigcap_{k=0}^\infty D(L^k)$ with respect to the Hilbert space norm $\|x\|_s := \|L^s x\|_{\mathcal{X}}$; obviously $\|x\|_0 = \|x\|_{\mathcal{X}}$ (see [11] or [4, section 8.4] for details).

Regularization in Hilbert scales was introduced by Natterer [15] in order to improve convergence rates for Tikhonov regularization. In [18], Landweber iteration for nonlinear problems, which exhibits saturation phenomena similar to that of Tikhonov regularization (i.e., optimal convergence only for $x^\dagger - x_0 \in \mathcal{R}((T^*T)^\mu)$, $\mu \le 1/2$), has been shifted to Hilbert scales (with $s > 0$) in order to overcome the restriction $\mu \le 1/2$.

In [3], the application of the Hilbert scale approach to iterative regularization methods has been investigated from a different point of view: there, the emphasis is on the case $s < 0$, in which the Hilbert scale operator L^{-2s} appearing in the modified Landweber iteration

$$x_{k+1}^\delta = x_k^\delta + L^{-2s}T^*(y^\delta - Tx_k^\delta), \qquad k \ge 0, \tag{3.1}$$

acts as a preconditioner for the adjoint operator T^*. As a consequence, the operator $L^{-2s}T^*T$ in the preconditioned normal equation

$$L^{-2s}T^*Tx = L^{-2s}T^*y^\delta \tag{3.2}$$

has a smaller degree of ill-posedness than T^*T, while being self-adjoint in $\mathcal{X}_s$. For a finite-dimensional approximation, this implies smaller condition numbers of the iteration operators $(L^{-2s}T^*T)$, which in turn yields a faster decrease of the residual and allow to stop the residual earlier.

For a convergence rates analysis of iterative regularization methods in Hilbert scales we will need the following assumption (cf. [3]).

ASSUMPTION 3.1. *For T and L as above assume that*

(A1) *$Tx = y$ has a solution $x^\dagger$.*

(A2) *$\|Tx\| \le \overline{m}\|x\|_{-a}$ for all $x \in \mathcal{X}$ and some $a > 0, \overline{m} > 0$. Moreover, the extension of T to $\mathcal{X}_{-a}$ (again denoted by T) is injective.*

(A3) *For some $s \ge -a$ let $B := TL^{-s}$ be such that $\|B\|_{\mathcal{X},\mathcal{Y}} \le 1$.*

The following result, taken from [3], draws some conclusions from Assumption 3.1, which will be needed for the subsequent convergence analysis.

PROPOSITION 3.2. *Let Assumption* 3.1 *hold. Then condition* (A2) *is equivalent to*

$$\mathcal{R}(T^*) \subset \mathcal{X}_a \qquad \text{and} \qquad \|T^* w\|_a \le \overline{m}\|w\| \quad \text{for all} \quad w \in \mathcal{Y}.$$

Moreover, for all $\nu \in [0,1]$ *it holds that* $\mathcal{D}((B^*B)^{-\frac{\nu}{2}}) = \mathcal{R}((B^*B)^{\frac{\nu}{2}}) \subset \mathcal{X}_{\nu(a+s)}$ *and*

$$\begin{aligned} \|(B^*B)^{\frac{\nu}{2}} x\| &\le \overline{m}^{\nu} \|x\|_{-\nu(a+s)} \qquad \text{for all} \quad x \in \mathcal{X}, \\ \|(B^*B)^{-\frac{\nu}{2}} x\| &\ge \overline{m}^{-\nu} \|x\|_{\nu(a+s)} \qquad \text{for all} \quad x \in \mathcal{D}((B^*B)^{-\frac{\nu}{2}}). \end{aligned} \tag{3.3}$$

Furthermore, (3.5) *is equivalent to*

$$\mathcal{X}_{\tilde a} \subset \mathcal{R}(T^*) \quad \text{and} \quad \|T^* w\|_{\tilde a} \ge \underline{m}\|w\| \quad \text{for all} \quad w \in \mathcal{N}(T^*)^\perp \text{ with } T^* w \in \mathcal{X}_{\tilde a},$$

and if (3.5) *holds, then it follows for all* $\nu \in [0,1]$ *that* $\mathcal{X}_{\nu(\tilde a+s)} \subset \mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{D}((B^*B)^{-\frac{\nu}{2}})$ *holds and*

$$\begin{aligned} \|(B^*B)^{\frac{\nu}{2}} x\| &\ge \underline{m}^{\nu} \|x\|_{-\nu(\tilde a+s)} \qquad \text{for all} \quad x \in \mathcal{X}, \\ \|(B^*B)^{-\frac{\nu}{2}} x\| &\le \underline{m}^{-\nu} \|x\|_{\nu(\tilde a+s)} \qquad \text{for all} \quad x \in \mathcal{X}_{\nu(\tilde a+s)}. \end{aligned}$$

Usually, for the analysis of regularization methods in Hilbert scales, a stronger condition than (A2) is used, namely (cf., e.g., [15, 17])

$$\|Tx\| \sim \|x\|_{-a} \qquad \text{for all} \quad x \in \mathcal{X}, \tag{3.4}$$

where the number a can be interpreted as the *degree of ill-posedness*. However, if $s \le 0$, an estimate from below (possibly in a weaker norm), e.g.,

$$\|Tx\| \ge \underline{m}\|x\|_{-\tilde a} \quad \text{for all} \quad x \in \mathcal{X} \quad \text{and some} \quad \tilde a \ge a,\ \underline{m} > 0, \tag{3.5}$$

is only needed to interpret the smoothness condition on $x^\dagger - x_0$ required for the convergence analysis in terms of the Hilbert scale $\{\mathcal{X}_s\}_{s\in\mathbb{R}}$: if (3.4) holds, then it follows from Proposition 3.2 that $\mathcal{X}_{\nu a} = \mathcal{R}((T^*T)^{\nu/2})$ for $|\nu| \le 1$ (see also Remark 3.7 below).

Before we come to our convergence analysis, we briefly discuss the connection of regularization in Hilbert scales to regularization in standard spaces and we point out the appropriate smoothness conditions for $x^\dagger - x_0$.

Remark 3.3. Regularization in Hilbert scales, e.g., (3.1), amounts to standard regularization when considering (the extension of) T as an operator on $\mathcal{X}_s$. The adjoint with respect to the spaces $\mathcal{X}_s$ and $\mathcal{Y}$ is denoted by $T^\sharp = L^{-2s}T^*$, where T^* denotes the adjoint with respect to $\mathcal{X}$ and $\mathcal{Y}$. Hence (3.2) are the normal equations for $Tx = y$ with $T : \mathcal{X}_s \to \mathcal{Y}$. Using the notation $z = L^s x$ and $B = TL^{-s}$, one sees that (1.1) is equivalent to

$$Bz = y, \qquad L^s x = z, \tag{3.6}$$

where in the second problem L^s maps from $\mathcal{X}_s$ to $\mathcal{X}$ and thus is an isomorphism. Applying standard regularization theory to (3.6) yields

$$z_k^\delta = g_k(B^*B)B^* y^\delta, \quad \text{respectively,} \quad x_k^\delta = L^{-s} g_k(B^*B)B^* y^\delta. \tag{3.7}$$

Moreover, the usual source condition for (3.6) reads

$$z^\dagger - z_0 = (B^*B)^\mu w, \quad \text{or equivalently,} \quad x^\dagger - x_0 = L^{-s}(B^*B)^\mu w \tag{3.8}$$

for some $w \in \mathcal{X}$. Due to (3.8), we call $x^\dagger - x_0 = L^{-s}(B^*B)^\mu w$ also a *source condition* below.

As can be seen from the previous remark, the following shifted Hilbert scale will play an important role in the convergence analysis of iterative regularization methods in Hilbert scales, in particular for the formulation of appropriate source, respectively, smoothness conditions.

DEFINITION 3.4. *Let a, $s > -a$, and B be as in Assumption* 3.1. *We define the shifted Hilbert scale $\{\mathcal{X}_r^s\}_{r\in\mathbb{R}}$ by*

$$\begin{aligned} &\mathcal{X}_r^s := \mathcal{D}((B^*B)^{\frac{s-r}{2(a+s)}} L^s) \quad \textit{equipped with the norm} \\ &\|\!|x|\!\|_r := \|(B^*B)^{\frac{s-r}{2(a+s)}} L^s x\|_{\mathcal{X}}. \end{aligned} \tag{3.9}$$

Remark 3.5. Note, that for $s \neq 0$ the spaces $\mathcal{X}_r^s$ form no Hilbert scale over $\mathcal{X}$ in general. In particular, $\mathcal{X}_{-r}^s$ is usually not the dual space of $\mathcal{X}_r^s$. Nevertheless, the spaces $\mathcal{X}_r^s$ have some properties (interpolation, embedding) that justify the notion of *shifted Hilbert scale* (see Proposition 3.6 below). To see that the spaces $\mathcal{X}_u^s$ are natural source sets for (1.1) considered over $\mathcal{X}_s$, observe that

$$\mathcal{X}_u^s = \{L^{-s}(B^*B)^{\frac{u-s}{2(a+s)}} w,\ w \in \mathcal{X}\} = \{(T^\sharp T)^{\frac{u-s}{2(a+s)}} w_s,\ w_s \in \mathcal{X}_s\},$$

where the second equality follows directly for $\frac{u-s}{2(a+s)}$ being integer. For arbitrary indices it follows by interpolation and Proposition 3.2.

The next proposition (cf. [3, Proposition 3]) summarizes the main properties of the shifted Hilbert scale (3.9), which are needed for the convergence analysis below, and clarifies the relation between the spaces $\mathcal{X}_s$ and $\mathcal{X}_r^s$.

PROPOSITION 3.6. *Let Assumption* 3.1 *hold and let $(\mathcal{X}_r^s)_{r\in\mathbb{R}}$ be defined as in Definition* 3.4. *Then the following hold:*

(i) *For $p < q$, the spaces $\mathcal{X}_q^s$ are continuously embedded in $\mathcal{X}_p^s$, i.e., for $x \in \mathcal{X}_q^s$,*

$$\|\!|x|\!\|_p \leq \gamma^{p-q} \|\!|x|\!\|_q,$$

where γ is such that

$$\langle (B^*B)^{-\frac{1}{2(a+s)}} x, x\rangle \geq \gamma \|x\|^2 \quad \textit{for all} \quad x \in \mathcal{D}((B^*B)^{-\frac{1}{2(a+s)}}).$$

(ii) *The interpolation inequality*

$$\|\!|x|\!\|_q \leq \|\!|x|\!\|_p^{\frac{r-q}{r-p}} \|\!|x|\!\|_r^{\frac{q-p}{r-p}}, \qquad p < q < r,$$

holds for all $x \in \mathcal{X}_r^s$.

(iii) *For $s \leq r \leq a + 2s$,*

$$\|x\|_r \leq \overline{m}^{\frac{r-s}{a+s}} \|\!|x|\!\|_r \qquad \textit{for all} \quad x \in \mathcal{X}_r^s \subset \mathcal{X}_r. \tag{3.10}$$

In particular, if $-a/2 \leq s \leq 0$, we obtain

$$\|x\|_0 \leq \overline{m}^{\frac{-s}{a+s}} \|\!|x|\!\|_0 \qquad \textit{for all} \quad x \in \mathcal{X}_0^s \subset \mathcal{X}_0.$$

Moreover,

$$\|\!|x|\!\|_{-a} = \|Tx\| \qquad \textit{for all} \quad x \in \mathcal{X}.$$

(iv) *If, in addition,* (3.5) *is satisfied, then the following estimates hold for* $s \le r \le a + 2s$ *with* $p = s + \frac{r-s}{a+s}(\tilde{a} + s)$*:*

$$\|x\|_p \ge \underline{m}^{\frac{r-s}{a+s}} \|\!|x|\!\|_r \qquad \text{for all} \quad x \in \mathcal{X}_p \subset \mathcal{X}_r^s.$$

Proof. We prove only (3.10) in detail: by definition of the Hilbert scale norm $\|\cdot\|_s$ in (3.9), and with (3.3), we obtain for $x \in \mathcal{X}_r^s \cap \mathcal{X}_r$

$$\|x\|_r = \|L^{r-s} L^s x\| \le \overline{m}^{\frac{r-s}{a+s}} \|(B^*B)^{\frac{s-r}{2(a+s)}} L^s x\| = \|\!|x|\!\|_r.$$

This implies the space inclusion $\mathcal{X}_r^s \subset \mathcal{X}_r$. The other assertions follow similarly from [4, Proposition 8.19] and Proposition 3.2. ∎

Before we begin our convergence analysis, we discuss the assertions of the previous proposition in more detail.

Remark 3.7. Under Assumption 3.1, and for the range $-a \le r \le s$, we have $\mathcal{X}_r \subset \mathcal{X}_r^s$, while for $s \le r \le a + 2s$ the opposite inclusion $\mathcal{X}_r^s \subset \mathcal{X}_r$ holds. If the stronger condition (3.4) holds instead of (A2), then the reverse statements are valid; i.e., both inclusions together yield $\mathcal{X}_r^s = \mathcal{X}_r$ for $-a \le r \le a + 2s$. The restriction $r \le a + 2s$ amounts to $\nu \le 1$ in Proposition 3.2 and can be relaxed if, e.g., the stronger condition $\|(T^*T)^\eta x\| \le \|x\|_{-2\eta a}$ holds for some $\eta > 1/2$, or if L and T commute (cf. [19]), in which case the previous estimate holds for all $\eta > 0$.

If (3.4) holds, we also obtain by Proposition 3.2 and Definition 3.4 that

$$\|(T^*T)^{-\frac{u}{2a}} x\| \sim \|x\|_u = \|L^{u-s} L^s x\| \sim \|\!|x|\!\|_u$$

for $-a \le u \le \min\{a, a+2s\}$. Here, the first estimate holds for $|u| \le a$ and the second for $|\frac{u-s}{a+s}| \le 1$, which is $-a \le u \le a + 2s$.

4. Convergence rates for iterative regularization methods in Hilbert scales. We start by citing a convergence rate result for Landweber iteration in Hilbert scales derived in [3].

THEOREM 4.1. *Let Assumption* 3.1 *hold and* $-a/2 \le s \le 0$*. Additionally, assume* $x^\dagger - x_0 \in \mathcal{X}_u^s$*, i.e.,*

$$x^\dagger - x_0 = L^{-s}(B^*B)^{\frac{u-s}{2(a+s)}} w \tag{4.1}$$

for some $w \in \mathcal{X}$ *and* $u > 0$*. Then*

$$\|x_k^\delta - x^\dagger\| \le c(\delta k^{\frac{a}{2(a+s)}} + k^{-\frac{u}{2(a+s)}} \|\!|x^\dagger - x_0|\!\|_u).$$

If the iteration (3.1) *is stopped according to the a priori rule* $k^* \sim (\|w\|\delta^{-1})^{\frac{2(a+s)}{a+u}}$*, then*

$$\|x_k^\delta - x^\dagger\| = O(\|w\|^{\frac{a}{a+u}} \delta^{\frac{u}{a+u}}).$$

If, alternatively, the iteration is stopped according to the discrepancy principle (1.4)*, then*

$$k_* \sim \delta^{-\frac{2(a+s)}{a+u}} \quad \text{and} \quad \|x_k^\delta - x^\dagger\| = O(\delta^{\frac{u}{a+u}}). \tag{4.2}$$

Remark 4.2. It was mentioned in [3] that, if the usual condition (3.4) holds instead of (A2), then for $0 < u \le a + 2s$ these rates are optimal, i.e., the best possible worst-case error bounds under the given source condition (cf. Remarks 3.3 and 3.5).

Observe that for $s < 0$, the stopping index of the Hilbert scale method is smaller than the one for Landweber iteration; e.g., for $x^\dagger - x_0 \in \mathcal{R}((T^*T)^{\frac{u}{2a}}) \cap \mathcal{X}_u^s$ and $s = -\frac{a}{2}$ the preconditioned iteration yields approximately the square root of iterations compared to standard Landweber. As can be seen from (3.1), the preconditioned iterations are in general not well defined as iterations on $\mathcal{X}$ for $s < -\frac{a}{2}$ and arbitrary $y^\delta \in \mathcal{Y}$.

The rest of this section is devoted to the derivation of a corresponding convergence rate result for general semi-iterative regularization methods in Hilbert scales. Recall the connection of Hilbert scale regularization with standard regularization over $\mathcal{X}_s$; cf. Remark 3.3. The formulae in (3.7) then allow the following closed-form representations of the approximation error and the propagated data error:

$$z_k - z^\dagger = r_k(B^*B)(z_0 - z^\dagger) \qquad \text{and} \qquad z_k^\delta - z_k = g_k(B^*B)B^*(y^\delta - y). \tag{4.3}$$

Noting that $z = L^s x$, this yields

$$\begin{aligned} x_k - x^\dagger &= L^{-s} r_k(B^*B) L^s (x_0 - x^\dagger), \\ x_k^\delta - x_k &= L^{-s} g_k(B^*B) B^*(y^\delta - y). \end{aligned} \tag{4.4}$$

If the residual polynomials are generated by (2.1), the preconditioned iterates x_k^δ can be assembled by the iteration

$$\begin{gathered} x_k^\delta = \mu_{1,k} x_{k-1}^\delta + \cdots + \mu_{k,k} x_0 + \omega_k L^{-2s} T^*(y^\delta - T x_k^\delta), \quad k \geq 1, \\ \textstyle\sum_{i=1}^k \mu_{i,k} = 1, \qquad \omega_k \neq 0. \end{gathered} \tag{4.5}$$

Note that in practice (4.5) is typically a two- or three-term recurrence. The only difference between the preconditioned iteration (4.5) and the standard method (2.1) is that the residuals $T^*(y^\delta - Tx_k)$ are preconditioned by L^{-2s}.

We are now in the position to state and prove the main results.

PROPOSITION 4.3. *Let Assumption* 3.1 *hold with* $-a/2 \leq s \leq 0$, *and let* x_k^δ *be defined by the semi-iterative method* (4.5) *satisfying* (2.2) *for some* $\mu_0 > 0$. *Additionally, assume that* $x^\dagger - x_0 \in \mathcal{X}_u^s$; *i.e.,* (4.1) *holds for some* $w \in \mathcal{X}$ *and* $0 < u \leq 2(a+s)\mu_0$. *Then*

$$\|x_k^\delta - x^\dagger\| \leq C_u(\delta k^{\frac{a}{(a+s)}} + k^{-\frac{u}{(a+s)}} \|w\|). \tag{4.6}$$

Proof. Using the source condition (4.1) and the representation (4.4), we obtain with (3.3) that

$$\begin{aligned} \|x_k - x^\dagger\| &= \|L^{-s} r_k(B^*B)(B^*B)^{\frac{u-s}{2(a+s)}} w\| \\ &\leq c\, \|(B^*B)^{\frac{u}{2(a+s)}} r_k(B^*B)\|\, \|w\|. \end{aligned}$$

By spectral theory and (2.2) this yields for $0 < u \leq 2(a+s)\mu_0$ the estimate

$$\|x_k - x^\dagger\| \leq c_u k^{-\frac{u}{(a+s)}} \|w\|$$

for the approximation error. Similarly, the propagated data error can be estimated by

$$\begin{aligned} \|x_k^\delta - x_k\| = \|L^{-s} g_k(B^*B) B^*(y^\delta - y)\| &\leq c\,\delta\, \|(B^*B)^{\frac{a+2s}{2(a+s)}} g_k(B^*B)\| \\ &\leq c\delta \|\lambda^{\frac{a+2s}{2(a+s)}} g_k(\lambda)\|_{\mathcal{C}[0,1]}. \end{aligned}$$

Next, we derive an estimate for $\|\lambda^\mu g_k(\lambda)\|_{\mathcal{C}[0,1]}$: since $r_k(\lambda) = 1 - \lambda g_k(\lambda)$, we obtain for $0 \le \mu \le 1$ that

$$\begin{aligned}\lambda^\mu g_k(\lambda) &= \lambda^{\mu-1}(1 - r_k(\lambda)) \\ &= [\lambda^{-1}(1 - r_k(\lambda))]^{1-\mu}[1 - r_k(\lambda)]^\mu.\end{aligned}$$

Now, by the mean value theorem, one can find a $\tilde\lambda \in [0,1]$ such that

$$\lambda^{-1}(1 - r_k(\lambda)) = -r_k'(\tilde\lambda),$$

which, together with Markov's inequality ($|r_k'(\lambda)| \le 2k^2$) and $|r_k(\lambda)| \le 2$ for $\lambda \in [0,1]$, yields

$$\lambda^\mu g_k(\lambda) \le 2k^{2(1-\mu)} \quad \text{for} \quad \lambda \in [0,1].$$

Since $-a/2 \le s \le 0$ by assumption, we obtain $0 \le \frac{a+2s}{2(a+s)} \le 1$ and thus by the previous estimates $\|x_k^\delta - x_k\| \le 2\,c\,\delta\,k^{\frac{a}{a+s}}$. ∎

Proposition 4.3 guarantees convergence if k_* is chosen such that $\delta k_*^{\frac{a}{a+s}} \to 0$ and $k_* \to \infty$ with $\delta \to 0$. In order to get convergence rates in terms of δ one has to bound the number of iterations k_* in terms of δ appropriately.

THEOREM 4.4. *Let the assumptions of Proposition* 4.3 *be satisfied and x_k^δ be generated by the method* (4.5) *with residual polynomials r_k satisfying* (2.2) *for some $\mu_0 > 0$. If the iteration is stopped according to the a priori stopping rule $k_* = O(\delta^{\frac{a+s}{a+u}})$, then*

$$\|x_k^\delta - x^\dagger\| = O(\delta^{\frac{u}{a+u}})$$

for $x^\dagger - x_0 \in \mathcal{X}_u^s$ with $0 < u \le 2(a+s)\mu_0$.

If, alternatively, the iteration is stopped according to the discrepancy principle (1.4)*, then*

$$k_* = O(\delta^{\frac{a+s}{a+u}}) \quad \textit{and} \quad \|x_k^\delta - x^\dagger\| = O(\delta^{\frac{u}{a+u}}) \tag{4.7}$$

for $x^\dagger - x_0 \in \mathcal{X}_u^s$ with $0 < u \le 2(a+s)\mu_0 - a$.

Proof. The first result follows immediately from Proposition 4.3. For the second, observe that by (1.4) it follows that for $k < k_*$

$$(\tau - 1)\,\delta \le \|T(x_k^\delta - x_k)\| + \|T(x_k - x^\dagger)\|.$$

Similarly as in the proof of Proposition 4.3, one obtains

$$\|T(x_k^\delta - x_k)\| = \|Bg_k(B^*B)B^*(y^\delta - y)\| \le c\,\delta.$$

Hence, for τ sufficiently large and some $C > 0$,

$$C\delta \le \|T(x_k - x^\dagger)\| \le k^{-\frac{a+u}{a+s}} \|\!|x^\dagger - x_0|\!\|_u,$$

where the last inequality holds for $u \le 2(a+s)\mu_0 - a$. This already yields the bound on k_*. Next, observe that by (1.4) it follows that $\|T(x_{k_*} - x^\dagger)\| \le c\delta$. Application of the interpolation inequality yields

$$\begin{aligned}\|x_k - x^\dagger\| &\le c\,\|T(x_k - x^\dagger)\|^{\frac{u}{a+u}}\,\|w\|^{\frac{a}{a+u}} \\ &\le C\delta^{\frac{u}{a+u}}\,\|w\|^{\frac{a}{a+u}},\end{aligned}$$

which, together with $\|x_{k_*}^\delta - x_{k_*}\| \le c\, k_*^{\frac{a}{a+s}}\delta$ (cf. the proof of Proposition 4.3) and the bound on k_*, yields the a posteriori rate. ∎

Remark 4.5. For $s = 0$, Theorem 4.4 coincides with Theorem 2.1. As in [3], the rate (4.7) can also be proven in the stronger norm $\|\!|x_{k_*}^\delta - x^\dagger|\!\|_0$. Furthermore, it is possible to show that $\|\!|x_k^\delta - x^\dagger|\!\|_u$ stays bounded for $k \le k_*$. Together with $\|\!|x_{k_*}^\delta - x^\dagger|\!\|_{-a} = \|T(x_{k_*}^\delta - x^\dagger)\| = O(\delta)$ and interpolation arguments, one then obtains the rates

$$\|\!|x_k^\delta - x^\dagger|\!\|_r = O(\delta^{\frac{u-r}{a+u}}) \qquad \text{for} \quad -a \le r \le u \tag{4.8}$$

in intermediate norms. If additionally the stronger condition (3.4) holds, then by Proposition 3.6 and Remark 3.7, the spaces $\mathcal{X}_u^s$ and $\mathcal{X}_u$ coincide with equivalent norms for $-a \le u \le a + 2s$, and the rates in (4.8) hold for $\|x_k^\delta - x^\dagger\|_r$ with $-a \le r \le u$ (cf. [10]) and under the usual source condition for regularization in Hilbert scales, namely $x^\dagger - x_0 \in \mathcal{X}_u$. Using the improved a posteriori stopping rule given in [4, section 6], the rates (4.7) even hold for $0 < u \le 2(a+s)\mu_0$, as in the case of a priori stopping with $k_* = c\,\delta^{\frac{a+s}{a+u}}$.

To see the effect of preconditioning, consider $s = -a/2$, and $x^\dagger - x_0 \in \mathcal{R}((T^*T)^\mu) \cap \mathcal{X}_u^s$, where $u = 2a\mu$. Then the following bounds on the stopping index hold: $k_* = O(\delta^{\frac{2}{2\mu+1}})$ for Landweber iteration, $k_* = O(\delta^{\frac{1}{2\mu+1}})$ for ν-methods (with $\nu \ge \mu + 1/2$) or Landweber iteration in Hilbert scales, and $k_* = O(\delta^{\frac{1}{2(2\mu+1)}})$ for the Hilbert scale ν-methods (with $\nu \ge \mu + 1$). For $\mu = 1/2$ and $\delta = 0.01$, this amounts to a reduction of the iteration numbers from 10,000 to 100 to 10, when switching from Landweber iteration to the ν-methods and then to their Hilbert scale versions. This remarkable acceleration will be demonstrated in several numerical examples below.

5. Examples and numerical tests. In this section we present several examples, where the conditions of Assumption 3.1 are satisfied and thus the results of section 4 are applicable. We compare the performance of the proposed Hilbert scale ν-methods with standard Landweber iteration and ν-methods, Landweber iteration in Hilbert scales, and the method of conjugate gradients. For our numerical tests, we use very fine discretizations (by standard piecewise linear finite elements) in order to ensure that discretization errors can be neglected.

As a first example we consider the identification of a source term from distributed measurements.

Example 5.1. Let Ω be a bounded domain in $\mathbb{R}^n$, $n = 2, 3$, with sufficiently smooth boundary (e.g., $\partial\Omega \in \mathcal{C}^{1,1}$ or $\partial\Omega \in \mathcal{C}^{0,1}$ and Ω convex) or let Ω be a parallelepiped. Consider the operator $T : L_2(\Omega) \to L_2(\Omega)$ defined by $Tf = u$, with

$$Au := -\nabla \cdot (q\nabla u) + p \cdot \nabla u + cu = f, \qquad u|_{\partial\Omega} = 0, \tag{5.1}$$

and given sufficiently smooth parameters q, p, and c. Assume that A is uniformly elliptic; then a solution to (5.1) has improved regularity, i.e., $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $\|u\|_{H^2} \sim \|f\|_{L_2}$. Let $\mathcal{X}_2 = H^2(\Omega) \cap H_0^1(\Omega)$, with $L^2 u = -\Delta u$, define the Hilbert scale $\{\mathcal{X}_s\}_{s\in\mathbb{R}}$ over $\mathcal{X} = L_2(\Omega)$. Then we have $T \sim \mathcal{X}_2$, and thus Assumption 3.1 holds with $a = 2$. Moreover, the stronger condition (3.4) holds.

For our numerical tests, we set $\Omega = [0,1]^2$, $q = c = 1$, and $p = 0$. We choose $s = -a/2 = -1$ for preconditioning and $\nu = 2$ for the ν-methods; note that $\nu \ge 3/2$ is necessary to apply Theorem 4.4 for $u = 2a\mu = 1/2$ in our case. In our experiment, we try to identify the function

$$f^\dagger = (\pi^2 + 1)\sin(\pi x) + (4\pi^2 + 1)\sin(2\pi y)$$

TABLE 5.1
Iteration numbers for Landweber iteration (lw), the ν-method (nu), Landweber iteration in Hilbert scales (hs), the proposed Hilbert scale ν-method (hsnu), and the conjugate gradient algorithm (cg); see Example 5.1.

$\delta/\|u\|$	$k_*(lw)$	$k_*(nu)$	$k_*(hs)$	$k_*(hsnu)$	$k_*(cg)$
0.016	86	24	11	8	6
0.008	240	42	18	10	8
0.004	725	75	29	14	13
0.002	2150	130	51	18	19
0.001	6080	219	87	24	28

TABLE 5.2
Iteration error $e_{k_*} = \|f^\delta_{k_*} - f^\dagger\|$ for Landweber iteration (lw), the ν-method (nu), Landweber iteration in Hilbert scales (hs), the proposed Hilbert scale ν-method (hsnu), and the conjugate gradient algorithm (cg); see Example 5.1.

$\delta/\|u\|$	$e_{k_*}(lw)$	$e_{k_*}(nu)$	$e_{k_*}(hs)$	$e_{k_*}(hsnu)$	$e_{k_*}(cg)$
0.016	0.328403	0.33310	0.34812	0.34210	0.32499
0.008	0.283369	0.28547	0.29510	0.28842	0.27948
0.004	0.240430	0.24163	0.25057	0.24648	0.23434
0.002	0.205203	0.20668	0.21525	0.21041	0.20361
0.001	0.175605	0.17691	0.18289	0.17889	0.17137

corresponding to $u = \sin(\pi x) + \sin(2\pi y)$. As a starting value we choose $f_0 = 0$. With this setting, we have $f^\dagger \in \mathcal{R}((T^*T)^\mu)$ for all $0 \le \mu < 1/8$, and thus one would expect the iteration numbers $k_* \sim \delta^{-8/5}$ for Landweber iteration, $k_* \sim \delta^{-4/5}$ for Landweber iteration in Hilbert scales and the ν-methods, and $k_* \sim \delta^{-2/5}$ for the Hilbert scale ν-method. For $\Omega = [0,1]^2$, the singular values of T behave like $\sigma_n = O(n^{-1})$. Thus, the stopping index for the conjugate gradient method can be bounded by (apply Theorem 7.14 in [4] with $\alpha = 1$)

$$k_*(cg) \le c\delta^{-\frac{1}{(2\mu+1)(1+\alpha)}} = c\delta^{-\frac{1}{2(2\mu+1)}}, \tag{5.2}$$

which is the same bound as that for the proposed Hilbert scale ν-method. Finally, the error should behave like $\|f^\delta_{k_*} - f^\dagger\| \sim \delta^{1/5}$ for all methods.

The numerically observed iteration numbers listed in Table 5.1 yield the rates $k_* = \delta^{-1.54}$ for Landweber iteration, $k_* = \delta^{-0.80}$ for the 2-method, $k_* = \delta^{-0.75}$ for Landweber iteration in Hilbert scales, and $k_* = \delta^{-0.40}$ for the proposed Hilbert scale 2-method. As expected, the iteration numbers for conjugate gradients and the Hilbert scale ν-method are of the same order. Table 5.2 lists the iteration error $e_{k_*} = \|f^\delta_{k_*} - f^\dagger\|$ for our numerical test. The corresponding convergence rates are $e_{k_*} \sim \delta^{0.22}$ for Landweber iteration and $e_{k_*} \sim \delta^{0.23}$ for the other methods. Note that for $\Omega \subset \mathbb{R}^3$, the Hilbert scale method should theoretically outperform the conjugate gradient algorithm, since there we have only $\alpha = 2/3$ in (5.2) yielding $k_*(cg) \sim \delta^{-\frac{3}{5(2\mu+1)}}$, while the estimate for the semi-iterative method in Hilbert scales is still $k_* \sim \delta^{-\frac{1}{2(2\mu+1)}}$.

Regularization in Hilbert scales was originally investigated under the stronger condition (3.4), which is satisfied in Example 5.1. However, in the case $s \le 0$, condition (A2) suffices to obtain the appropriate convergence rates. In the following example, only a weaker estimate from below (3.5) holds. Note that due to Proposition 3.6, the source condition $x^\dagger - x_0 \in \mathcal{X}_u^s$ can still be interpreted in terms of the Hilbert scale $\{\mathcal{X}_s\}_{s\in\mathbb{R}}$.

TABLE 5.3
Iteration numbers k_* and error $e_{k_*} = \|x^\delta_{k_*} - x^\dagger\|$ for the 2-method (nu), the proposed Hilbert scale 2-method (hsnu), and the conjugate gradient algorithm (cg); see Example 5.2.

$\delta/\|y\|$	$k_*(nu)$	$e_{k_*}(nu)$	$k_*(hsnu)$	$e_{k_*}(hsnu)$	$k_*(cg)$	$e_{k_*}(cg)$
0.016	60	0.42317	11	0.38092	5	0.38866
0.008	100	0.37634	15	0.33851	6	0.33448
0.004	178	0.32480	21	0.28897	8	0.30918
0.002	313	0.28257	28	0.25193	11	0.26304
0.001	541	0.24696	36	0.22362	14	0.23447

Example 5.2. As a model problem for linear inverse problems, we consider the following Fredholm integral equation of the first kind: let $T : L_2[0,1] \to L_2[0,1]$ be defined by

$$(Tx)(s) = \int_0^1 s^{1/2} k(s,t) x(t) dt, \tag{5.3}$$

with the standard Green's kernel

$$k(s,t) = \begin{cases} s(1-t), & s < t, \\ t(1-s), & t \le s. \end{cases}$$

Noting that

$$(T^* y)(t) = (1-t) \int_0^t s^{3/2} y(s) ds + t \int_t^1 s^{1/2} (1-s) y(s) ds,$$

we obtain

$$\mathcal{R}(T^*) = \{w \in H^2[0,1] \cap H_0^1[0,1] : t^{-1/2} w''(t) \in L_2[0,1]\}.$$

As a Hilbert scale operator, we choose $L^2 x = -x''$, i.e.,

$$L^s x := \sum_{n=1}^{\infty} (n\pi)^s \langle x, x_n \rangle x_n, \qquad x_n := \sqrt{2} \sin(n\pi \cdot). \tag{5.4}$$

This choice yields $\mathcal{R}(T^*) \subsetneq \mathcal{X}_2 := H^2[0,1] \cap H_0^1[0,1]$; additionally, $\mathcal{R}(T^*) \supset \mathcal{X}_{2.5} := \{w \in H^{2.5}[0,1] \cap H_0^1[0,1] : \rho^{-1/2} w'' \in L_2[0,1]\}$, with $\rho(t) = t(1-t)$. By Theorem 11.7 in [13], we have $\|w\|^2_{2.5} \sim \|w''\|^2_{H^{1/2}} + \|\rho^{-1/2} w''\|^2_{L_2}$, and thus

$$\underline{m} \, \|x\|_{-2.5} \le \|Tx\| \le \overline{m} \, \|x\|_{-2};$$

see [3] for details. As a numerical test, we consider the reconstruction of the function

$$x^\dagger(s) = 2t - \operatorname{sign}(2t-1) - 1$$

and choose $s = -1$ and $x_0 = 0$. In Table 5.3, we report the iteration numbers obtained for the classical ν-method, the one in Hilbert scales, and for the conjugate gradient algorithm. The corresponding stopping indices behave like $k_* \sim \delta^{-0.8}$ for the ν-method, $k_* \sim \delta^{-0.43}$ for the Hilbert scale ν-method, and $k_* \sim \delta^{-0.38}$ for the conjugate gradient algorithm; the convergence rates are $e_{k_*} \sim \delta^{0.2}$ for all examples. Again, the values are almost exactly the ones predicted by the theory ($\mu = 1/8$).

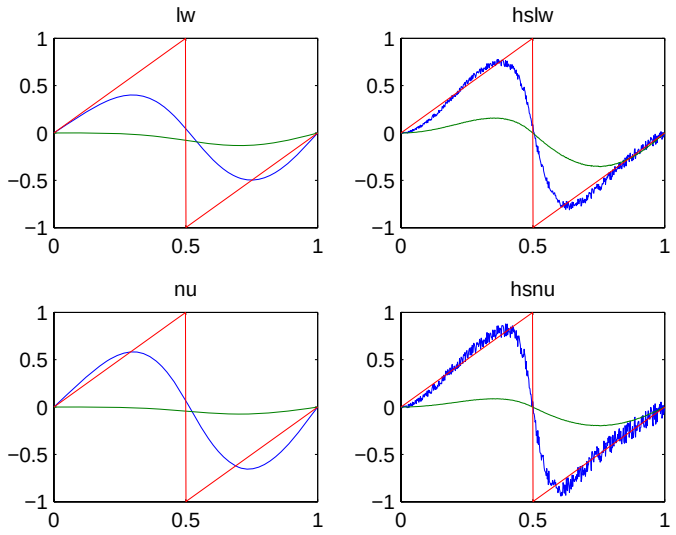


FIG. 5.1. *Iterates x_k^δ and true solution $x^\dagger$ after 1 and 14 iterations ($k_* = 14$ is the stopping index of the Hilbert scale ν-method) for a noise level $\delta = 1\%$; see Example 5.2.*

As a consequence of the preconditioning in Hilbert scales, the updates and iterates of the Hilbert scale iterations are less smooth than those of the standard iterations. Therefore, one may conjecture that especially nonsmooth parts of a solution can be reconstructed faster than without preconditioning. This behavior is illustrated in Figure 5.1. Note that the oscillations that can be seen for the Hilbert scale iterates are small in the norm of $\mathcal{X} = L_2$.

In the next example, we study the problem of *transmission computerized tomography* (see [16]).

Example 5.3. Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a compact domain with spatially varying density f. In a simple physical model the relative intensity loss along a distance Δx is assumed to satisfy

$$\frac{\Delta I}{I} = f(x)\Delta x.$$

Denoting by $I_1(\theta, s)$ and $I_0(\theta, s)$ the intensities of the X-ray beams measured at the detector and emitter connected by the line parameterized by the distance to the origin s and the direction θ and located outside of the domain Ω, one gets

$$(Rf)(\theta, s) := \int_{x\cdot\theta=s} f(x)dx = -\log \frac{I_1(\theta, s)}{I_0(\theta, s)} = g(\theta, s) \tag{5.5}$$

for $w \in \mathbb{R}^2$, $\|w\| = 1$, and $t > 0$. Determining the unknown density f from measurements of the intensity drop $g(\theta, s)$ corresponds to inversion of the *Radon transform*. By [16, Theorem 5.1], we know that for each α there exist positive constants $c(\alpha, n)$ and $C(\alpha, n)$ such that for $f \in C_0^\infty(\Omega^n)$,

$$c(\alpha, n)\|f\|_{H_0^\alpha(\Omega^n)} \le \|Rf\|_{H^{\alpha+(n-1)/2}(Z)} \le C(\alpha, n)\|f\|_{H_0^\alpha(\Omega^n)},$$

with $\Omega^n \subset \mathbb{R}^n$ denoting the unit ball, and Z the cylinder $S^{n-1} \times \mathbb{R}$. This implies (3.4) for an appropriate choice of spaces; e.g., for $\mathcal{X} = L_2(\Omega^n)$ and $\mathcal{Y} = L_2(Z)$, we see that the Radon transform behaves like differentiation of order one $\frac{1}{2}$ in dimension $n = 2$, and like 1 times differentiation in dimension $n = 3$.

If Ω is a circle with radius r and $f(\theta, s) = f(s)$, and consequently $g(\theta, s) = g(s)$, are radially symmetric, then (5.5) can be reduced to the solution of an *Abel integral equation* of the first kind (see [16]), whose solution we investigate numerically below.

Let $T : L_2[0,1] \to L_2[0,1]$ be defined by

$$(Tx)(s) := \frac{1}{\sqrt{\pi}} \int_0^s \frac{x(t)}{\sqrt{s-t}} dt, \tag{5.6}$$

with data y and "true" solution $x^\dagger = T^\dagger y$. One can show that $(T^2 x)(s) = \int_0^s x(t)dt$, and thus inverting T amounts to differentiation of half order; more precisely (cf. [5]),

$$\mathcal{R}(T) \subset H^r[0,1] \qquad \text{for all} \quad 0 \le r < 1/2.$$

Let the Hilbert scale operator L be defined by

$$L^{2s} x = \sum_{n=0}^{\infty} \lambda_n^s \langle x, x_n \rangle x_n \quad \text{with} \quad x_n(t) = \sqrt{2}\sin(\lambda_n t), \quad \lambda_n = (n + 1/2)\pi, \tag{5.7}$$

with $\mathcal{X} = L_2[0,1]$ and $\mathcal{X}_2 = \{x \in H^1[0,1] : x(0) = 0\}$. Then $\mathcal{R}(T) \subset \mathcal{X}_r$ holds for all $0 < r < 1$ and $0 < a < 1$, and the choice $-1/2 < s = -a/2$ is possible. Thus, the iteration can be preconditioned with L^{-a}, which corresponds to differentiation of fractional order and can be realized efficiently via (5.7) and the fast Fourier transform.

In the numerical test we set $s = -1/2$ (which is the limiting case of allowed choices) and try to identify the unknown density

$$x^\dagger(s) = 2t - \operatorname{sign}(2t - 1) - 1.$$

The iterations are started with $x_0 = 0$. In this setting we have $x^\dagger \in \mathcal{R}((T^*T)^\mu)$ for all $0 \le \mu < 1/2$, and thus we can expect the iteration numbers $k_* \sim \delta^{-1}$ for Landweber iteration, $k_* \sim \delta^{-1/2}$ for the ν-method and Landweber iteration in Hilbert scales, and $k_* \sim \delta^{-1/4}$ for the proposed Hilbert scale ν-method. The stopping index for the conjugate gradient algorithm is bounded by $k_* \sim \delta^{-1/3}$. As mentioned in the introduction, the bound for the Hilbert scale ν-method is stronger than that for the conjugate gradient algorithm if the singular values σ_n of T decay no faster than $n^{-\alpha}$ with some $0 < \alpha < 1$, which is the case here.

The numerically realized rates for the stopping index are $k_* \sim \delta^{-1.0}$ for the Landweber iteration, $k_* \sim \delta^{-0.53}$ for the 2-method, $k_* \sim \delta^{-0.44}$ for the Landweber iteration in Hilbert scales, and $k_* \sim \delta^{-0.4}$ for the conjugate gradient method, and these rates are in good accordance with the theoretically predicted ones. The two Hilbert scales ν-methods yield $k_* \sim \delta^{-0.48}$ for $\nu = 1$ and $k_* \sim \delta^{-0.3}$ for $\nu = 2$. Note that due to the restriction on the qualification μ_0 of the method used in Theorem 4.4, one has to choose

$$\nu \ge \frac{u - s}{2(a + s)} + \frac{1}{2} = 2$$

in order to get an optimal number of iteration and convergence rates for the Hilbert scale ν-method stopped with the discrepancy principle (1.4). This explains the higher

TABLE 5.4
Iteration numbers k_* for the Landweber iteration (lw), the 2-method (nu), Landweber iteration in Hilbert scales (hs), the proposed Hilbert scale ν-methods (hs1, hs2), and the conjugate gradient algorithm (cg); see Example 5.3.

$\delta/\|u\|$	$k_*(lw)$	$k_*(nu)$	$k_*(hs)$	$k_*(hs1)$	$k_*(hs2)$	$k_*(cg)$
0.016	37	16	9	7	6	6
0.008	75	24	12	9	8	8
0.004	146	33	15	14	10	10
0.002	300	48	21	19	12	14
0.001	643	71	31	26	15	19

TABLE 5.5
Iteration numbers k_* for the Landweber iteration (lw), the 2-method (nu), Landweber iteration in Hilbert scales (hs), the proposed Hilbert scale ν-method (hsnu), and the conjugate gradient algorithm (cg); see Example 5.4.

$\delta/\|u\|$	$k_*(lw)$	$k_*(nu)$	$k_*(hs)$	$k_*(hsnu)$	$k_*(cg)$
0.016	20	12	6	4	3
0.008	50	21	8	6	5
0.004	377	56	28	13	5
0.002	723	74	53	17	5
0.001	1116	88	80	21	5

number of iterations needed for the Hilbert scale 1-method; cf. Table 5.4. Finally, for all examples, the iteration error $e_{k_*} = \|x_{k_*}^\delta - x^\dagger\|$ decreases approximately like $\delta^{0.4}$ in accordance to the predicted rate $\delta^{\frac{2\mu}{2\mu+1}}$.

In the final example, we investigate the performance of iterative regularization methods for an exponentially ill-posed problem, namely the *backwards heat equation* and compare the numerical results obtained in [3] to those for ν-methods in Hilbert scales and the conjugate gradient algorithm.

Example 5.4. Consider $Tx = y$ with operator $T : L_2[0,1] \to L_2[0,1]$ defined by $(Tg)(x) = y(x) = u(x,\bar t)$ for some $\bar t > 0$ and

$$-u_t + qu_{xx} = 0, \qquad u(0,t) = u(1,t) = 0, \qquad u(x,0) = g.$$

Let L^s be defined by (5.4). Then we have

$$\|T^* y\|_r \le c(r)\|y\|_0 \qquad \text{for all} \quad r < 2.5,$$

but no estimate from below (3.5) exists.

We consider the numerical reconstruction and compare the numerically observed convergence rates and iteration numbers for the example

$$g^\dagger(x) = 2x - \operatorname{sign}(2x-1) - 1$$

and set $g_0 = 0$. For preconditioning we set $s = -1$, and thus L^{-2s} amounts to twice differentiation. Note, that for exponentially ill-posed problems only a logarithmic convergence rate can be expected under the weak source-condition of our example.

The stopping indices for Example 5.4 listed in Table 5.5 are bounded by $k_* \sim \delta^{-1.54}$ for Landweber iteration, $k_* \sim \delta^{-0.75}$ for the ν-method, $k_* \sim \delta^{-1.02}$ for Landweber iteration in Hilbert scales, and $k_* \sim \delta^{-0.63}$ for the Hilbert scale ν-method. According to Theorem 7.14 in [4], the stopping index for the conjugate gradient method can be bounded by $k(\delta, y^\delta) \le c\,(1 + \log \frac{1}{\delta})$ for exponentially ill-posed problems, i.e., if the singular values σ_n of T decay like $O(q^n)$ with some $q < 1$, which explains the almost

constant iteration numbers for the conjugate gradient method in our numerical test. The numerically observed convergence rates are approximately $\|x_{k_*}^\delta - x^\dagger\| \sim \delta^{0.05}$ for all methods.

REFERENCES

[1] A. B. Bakushinskii and A. V. Goncharskii, *Iterative Methods for Solving Ill-posed Problems*, Nauka, Moscow, 1989 (in Russian).

[2] H. Brakhage, *On ill-posed problems and the method of conjugate gradients*, in Inverse and Ill-posed Problems, H. W. Engl and C. W. Groetsch, eds., Academic Press, New York, 1987, pp. 165–175.

[3] H. Egger and A. Neubauer, *Preconditioning Landweber iteration in Hilbert scales*, Numer. Math., (2005), to appear.

[4] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problem*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[5] R. Gorenflo and S. Vesella, *Abel integral equations: Analysis and applications*, in Lecture Notes in Math. 1461, Springer-Verlag, Berlin, 1991.

[6] C. W. Groetsch, *Generalized Inverses of Linear Operators*, Marcel Dekker, New York, 1977.

[7] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.

[8] M. Hanke, *Accelerated Landweber iterations for the solution of ill-posed equations*, Numer. Math., 60 (1991), pp. 341–373.

[9] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[10] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Problems*, in preparation.

[11] S. G. Krein and J. I. Petunin, *Scales of Banach spaces*, Russian Math. Surveys, 21 (1966), pp. 85–160.

[12] L. Landweber, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624.

[13] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications: Volume* I, Springer, Berlin, 1972.

[14] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl., 7 (1966), pp. 414–417.

[15] F. Natterer, *Error bounds for Tikhonov regularization in Hilbert scales*, Appl. Anal., 18 (1984), pp. 29–37.

[16] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner, Stuttgart, 1986.

[17] A. Neubauer, *Tikhonov regularization of nonlinear ill-posed problems in Hilbert scales*, Appl. Anal., 46 (1992), pp. 59–72.

[18] A. Neubauer, *On Landweber iteration for nonlinear ill-posed problems in Hilbert scales*, Numer. Math., 85 (2000), pp. 309–328.

[19] U. Tautenhahn, *Error estimates for regularization methods in Hilbert scales*, SIAM J. Numer. Anal., 33 (1996), pp. 2120–2130.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 82–101

STABILIZATION OF LOW-ORDER MIXED FINITE ELEMENTS FOR THE STOKES EQUATIONS*

PAVEL B. BOCHEV†, CLARK R. DOHRMANN‡, AND MAX D. GUNZBURGER§

Abstract. We present a new family of stabilized methods for the Stokes problem. The focus of the paper is on the lowest order velocity-pressure pairs. While not LBB compliant, their simplicity and attractive computational properties make these pairs a popular choice in engineering practice. Our stabilization approach is motivated by terms that characterize the LBB "deficiency" of the unstable spaces. The stabilized methods are defined by using these terms to modify the saddle-point Lagrangian associated with the Stokes equations. The new stabilized methods offer a number of attractive computational properties. In contrast to other stabilization procedures, they are parameter free, do not require calculation of higher order derivatives or edge-based data structures, and always lead to symmetric linear systems. Furthermore, the new methods are unconditionally stable, achieve optimal accuracy with respect to solution regularity, and have simple and straightforward implementations. We present numerical results in two and three dimensions that showcase the excellent stability and accuracy of the new methods.

Key words. Stokes equations, stabilized mixed methods, equal-order interpolation, inf-sup condition

AMS subject classifications. 76D05, 76D07, 65F10, 65F30

DOI. 10.1137/S0036142905444482

1. Introduction.

Despite the fact that they violate the LBB [10] stability condition, low-order velocity-pressure pairs remain a popular practical choice in mixed finite element approximation of incompressible materials; see, e.g., [29] and the references cited therein. This popularity results from factors such as local mass conservation for the lowest order conforming pair (piecewise linear, bilinear or trilinear C^0 velocities, and piecewise constant pressures), simple and uniform data structures for the lowest equal order pair (piecewise linear, bilinear or trilinear C^0 velocities *and* pressures), and algebraic problems with manageable sizes and small bandwidths in three dimensions for both pairs. The latter are of paramount importance in engineering applications, where geometry resolution requires very fine meshes and higher order elements can quickly lead to intractable algebraic problems in three space dimensions; see [27] for an example setting.

To counteract the lack of LBB stability, low-order pairs are usually supplemented by stabilization or postprocessing procedures that remove spurious pressure modes. Unlike penalty methods (see [16, 22, 24, 25]) for which the goal is to uncouple the

*Received by the editors June 21, 2004; accepted for publication (in revised form) May 5, 2005; published electronically February 8, 2006. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/sinum/44-1/44448.html

†Computational Mathematics and Algorithms Department, Sandia National Laboratories, Mail Stop 1110, Albuquerque, NM 87185-1110 (pbboche@sandia.gov).

‡Structural Dynamics Research Department, Sandia National Laboratories, Mail Stop 0847, Albuquerque, NM 87185-0847 (crdohrm@sandia.gov).

§School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu). This author was supported in part by CSRI, Sandia National Laboratories under contract 18407.

pressure and velocity, stabilized methods aim to relax the continuity equation so as to allow application of LBB incompatible spaces. *Consistently stabilized* methods (see, e.g., [1, 2, 3, 5, 15, 20, 21]) accomplish this by using the residual of the momentum equation in the added stabilization terms. However, for low-order pairs, pressure and velocity derivatives in this residual term either vanish or are poorly approximated, causing difficulties in the application of consistent stabilization. One possible remedy is to reformulate the Stokes problem as a first-order system so that the momentum residual contains only first-order terms [5]. This, of course, leads to more unknowns and larger problems to solve. A second approach is to reconstruct the higher order derivatives [23] or to replace the Laplace operator by a discrete operator [8]. In either case, computation of a global L^2 projection may be required.

It is possible to stabilize unstable velocity pressure pairs without using residuals. One example, motivated by fractional step algorithms for time-dependent problems, is the pressure gradient projection (PGP) method (see [6, 7, 13]) and the related local pressure gradient stabilization (LPS) method [4]. In both methods the compressibility constraint is relaxed by subtracting the discontinuous pressure gradient from its projection onto a piecewise polynomial space. The difference is that PGP projects the pressure gradient onto the continuous velocity space and gives rise to a globally coupled problem, while LPS assumes nested spaces and projects the gradient onto an element patch space, which leads to local problems. However, it is clear that both methods are not appropriate for pairs with constant pressure elements.

Other examples of nonresidual stabilization are the local and global pressure jump formulations for the bilinear-constant pair [28, 29]. In these methods, the constraint is relaxed by using the jumps of the discontinuous pressure across element interfaces. Application of pressure jump stabilization requires edge-based data structures, and in the case of the local formulation, subdivision of the mesh into patches. Stabilization of the bilinear-constant pair is also considered in [26] where, instead of pressure jumps, local projections onto 2×2 macroelements are employed to relax the continuity equation.

In this paper, we analyze a new, nonresidual-based approach to the stabilization of low-order mixed finite element discretizations of the Stokes equations, further developing the idea of polynomial-pressure-projection–based stabilization that was presented and studied computationally in [14]. The starting point for the analysis of the method is a lower bound for a discrete negative seminorm of the pressure gradient which quantifies the LBB "deficiency" of an unstable pair. We show that the LBB "deficiency" admits a representation in terms of operators with suitable range spaces. This very general characterization opens up a possibility for stabilizing the mixed Stokes equations in a manner that is independent of the space dimension and the shape of the finite elements and also does not require choosing any mesh-dependent parameters.

Our approach differs from existing residual and nonresidual stabilization techniques in several important aspects. Most notably, the new methods do not require approximation of derivatives, specification of mesh-dependent parameters, or nonstandard data structures. Furthermore, our methods are unconditionally stable, optimally accurate, and always lead to symmetric problems. Their implementation relies on operators whose actions can be evaluated locally at the element level using standard finite element techniques. As a result, an existing code can be easily modified to handle the new stabilization procedures.

The paper is organized as follows. The remainder of this section introduces the notation used throughout the paper. Section 2 reviews the mixed variational formulation of the Stokes problem and a weaker form of the LBB stability condition that holds for the spaces of interest to us. The new method is formulated in section 3. Sections 4 and

5 deal with the stability and the error analysis, respectively, of the new methods while section 6 is a succinct summary of some implementation details. The paper concludes with section 7 in which the results of a series of numerical experiments are collected.

1.1. Nomenclature. In what follows, Ω denotes a simply connected bounded domain in $\mathbb{R}^d$, $d = 2, 3$, with a Lipschitz continuous boundary Γ. Throughout the paper, we employ the standard notation $H^l(\Omega)$, $\|\cdot\|_l$, $(\cdot,\cdot)_l$, $l \geq 0$, for the Sobolev spaces of all functions having square integrable derivatives up to order l on Ω, and the standard Sobolev norm and inner product, respectively. When $l = 0$ we will write $L^2(\Omega)$ instead of $H^0(\Omega)$ and drop the index from the inner product designation. $H_0^l(\Omega)$ will denote the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_l$ and $L_0^2(\Omega)$ will denote the space of all square integrable functions with vanishing mean. Spaces consisting of vector-valued functions will be denoted in boldface. Throughout the paper we use C to denote a generic positive constant whose value may change from place to place but that remains independent of the mesh parameter h.

In this paper, we formulate methods for the Stokes equations that use pressure and velocity finite element spaces defined with respect to the same partition $\mathcal{T}_h$ of Ω into finite elements Ω_e. For instance, Ω_e can be a hexahedron or a tetrahedron in three dimensions, or a triangle or a quadrilateral in two dimensions. The boundary $\partial\Omega_e$ of an element consists of faces γ_f. In two dimensions, each γ_f is an edge; in three dimensions, γ_f can be triangles or quadrilaterals. We assume that each face is oriented by selecting a normal direction $\mathbf{n}_f$. The set of all interior faces will be denoted by Γ_h. The norm

$$\|u\|_{\Gamma_h} = \left(\sum_{\gamma_f \in \Gamma_h} \int_{\gamma_f} u^2 \, dS \right)^{1/2} \tag{1.1}$$

will prove useful in what follows.

Our main focus is on low-order velocity and pressure pairs. For simplicial elements, we consider the affine finite element families

$$P_1 = \left\{ u^h \in C^0(\Omega) \mid u^h|_{\Omega_e} \in \mathcal{P}_1(\Omega_e) \; \forall \, \Omega_e \in \mathcal{T}_h \right\}, \tag{1.2}$$

where $\mathcal{P}_1(\Omega_e)$ is the space of linear polynomials on Ω_e. For quadrilateral and hexahedral elements we consider the space

$$Q_1 = \left\{ u^h \in C^0(\Omega) \mid u^h|_{\Omega_e} = \widehat{u}^h \circ F^{-1}; \; \widehat{u}^h \in \mathcal{Q}_1(\widehat{\Omega}_e) \right\}, \tag{1.3}$$

where $\widehat{\Omega}_e$ is a reference element, $F : \widehat{\Omega}_e \mapsto \Omega_e$ is a bilinear or a trilinear mapping, and $\mathcal{Q}_1(\Omega_e)$ is the space of all polynomials on $\widehat{\Omega}_e$ whose degree does not exceed 1 in each coordinate direction. Note that unless Ω_e is a parallelogram or a parallelepiped, u^h is not a piecewise polynomial function. For convenience, in what follows we will use the symbol R_1 to represent both kinds of finite element spaces. In accordance with our earlier convention, vector-valued finite element spaces will be denoted in boldface, e.g., $\mathbf{R}_1$. A well-known approximation result (see [18, p. 217]) is that for every $u \in H^2(\Omega)$, there exists a function $u^h \in R_1$ such that

$$\|u - u^h\|_0 + h^{1/2}\|u - u^h\|_{\Gamma_h} + h\|u - u^h\|_1 \leq Ch^2\|u\|_2 \, . \tag{1.4}$$

In addition to the C^0 spaces R_1, we will also need the piecewise constant space

$$R_0 = \left\{ q^h \in L^2(\Omega) \mid q^h|_{\Omega_e} \in \mathcal{P}_0(\Omega_e) \; \forall \, \Omega_e \in \mathcal{T}_h \right\}, \tag{1.5}$$

where $\mathcal{P}_0(\Omega_e)$ is a constant polynomial space on Ω_e. In (1.5), $\mathcal{T}_h$ can be a simplicial or a nonsimplicial partition of Ω into finite elements. The space R_0 has the following approximation property (see [18, p. 102]): for every $q \in H^1(\Omega)$, there exists $q^h \in R_0$ such that

$$\|q - q^h\|_0 \le Ch\|q\|_1 \,. \tag{1.6}$$

Finite element functions satisfy a number of useful inverse inequalities [12]. In particular, we will use the standard inverse inequality

$$\|\nabla q^h\|_0 \le C_I h^{-1}\|q^h\|_0 \tag{1.7}$$

that holds under some mild assumptions on $\mathcal{T}_h$ for all functions in R_1, and the inverse inequality for R_0 functions

$$\|[q^h]\|_{\Gamma_h} \le C_I h^{-1/2}\|q^h\|_0 \quad \forall\, q^h \in R_0 \,, \tag{1.8}$$

where $[q^h]$ denotes the jump of $q^h \in R_0$.

2. Mixed finite element methods for the Stokes problem. We consider the incompressible Stokes problem[1]

$$-\lambda\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega, \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad \Omega \tag{2.2}$$

along with the homogeneous velocity boundary condition

$$\mathbf{u} = 0 \quad \text{on } \Gamma \,. \tag{2.3}$$

The mixed variational form of (2.1)–(2.3) is to seek $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that

$$Q(\mathbf{u}, p; \mathbf{v}, q) = F(\mathbf{v}, q) \quad \forall\, (\mathbf{v}, q) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \,, \tag{2.4}$$

where

$$F(\mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega \,,$$

$$Q(\mathbf{u}, p; \mathbf{v}, q) = A(\mathbf{u}, \mathbf{v}) + B(\mathbf{v}, p) + B(\mathbf{u}, q) \,, \tag{2.5}$$

$$A(\mathbf{u}, \mathbf{v}) = \lambda \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v} d\Omega \,, \qquad \text{and} \qquad B(\mathbf{v}, p) = -\int_\Omega p\nabla \cdot \mathbf{v} d\Omega \,. \tag{2.6}$$

[1] We work with a nondimensional form of the Stokes problem. The dimensional form of the Stokes equation has the form

$$-\mu\Delta\mathbf{u} + \nabla p = \rho\mathbf{f} \,,$$

where μ is the given (dynamic) viscosity, ρ is the given fluid density, and $\mathbf{f}$ is a given body force per unit mass. We choose a reference speed u_{ref}, length ℓ_{ref}, and density ρ_{ref} which we use to, respectively, nondimensionalize the velocity $\mathbf{u}$, the position vector $\mathbf{x}$, and the density ρ. We then arrive at (2.1) by nondimensionalizing the pressure p using $\rho_{ref}u_{ref}^2$ and $\rho\mathbf{f}$ using $\rho_{ref}u_{ref}^2/\ell_{ref}$. Then, in (2.1), the nondimensional parameter $\lambda = \mu/(\rho_{ref}\ell_{ref}u_{ref})$ is the inverse of the "Reynolds number."

Solely for the sake of keeping the presentation simple, we make the assumption that μ is constant. For nonconstant μ, the viscous term in the Stokes equations (2.1) is given by $\nabla \cdot (\lambda(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$, where λ is no longer constant, and, in (2.6), the first bilinear form is given by $A(\mathbf{u}, \mathbf{v}) = 2\int_\Omega \lambda D(\mathbf{u}) : D(\mathbf{v}) d\Omega$, where $D(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + (\nabla\mathbf{v})^T)$. Only minor modifications in the analyses are needed.

The mixed variational equation (2.4) is the first-order optimality condition for the saddle-point $(\mathbf{u}, p)$ of the Lagrangian functional

$$L(\mathbf{v}, q) = \frac{\lambda}{2}\int_\Omega |\nabla \mathbf{v}|^2\, dx - \int_\Omega q\nabla\cdot\mathbf{v}\, dx - \int_\Omega \mathbf{f}\cdot\mathbf{v}\, dx\,. \tag{2.7}$$

To define a mixed finite element method for the Stokes problem (2.1)–(2.2), we restrict (2.4) to a pair of finite elements subspaces $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$ and $S^h \subset L_0^2(\Omega)$. A stable and accurate solution of (2.4), or equivalently, a stable and accurate approximation of the saddle-point of (2.7), requires that $\mathbf{V}^h$ and S^h satisfy the discrete inf-sup condition

$$\sup_{\mathbf{v}^h\in\mathbf{V}^h,\,\mathbf{v}^h\neq\mathbf{0}} \frac{B(p^h,\mathbf{v}^h)}{\|\mathbf{v}^h\|_1} \geq \gamma\|p^h\|_0 \quad \forall\, p^h \in S^h \tag{2.8}$$

with $\gamma > 0$ independent of h; see [18, 19].

In this paper, we will formulate stabilized mixed methods for the lowest equal order C^0 pair

$$\mathbf{V}^h = \mathbf{R}_1 \cap \mathbf{H}_0^1(\Omega) \quad \text{and} \quad S^h = R_1 \cap L_0^2(\Omega)\,, \tag{2.9}$$

and for the lowest order conforming pair

$$\mathbf{V}^h = \mathbf{R}_1 \cap \mathbf{H}_0^1(\Omega) \quad \text{and} \quad S^h = R_0 \cap L_0^2(\Omega)\,. \tag{2.10}$$

For simplicial elements, (2.10) is the unstable linear-constant pair that provides a textbook example for an overconstrained velocity space; see [19, p. 23]. For quadrilateral elements, it is the bilinear-constant pair that exhibits the notorious checkerboard pressure mode. A common misconception is that once this mode is taken care of, the bilinear-constant pair can be safely used. However, in [9] it is shown that this is not the case and that, in fact, for this pair the constant γ in (2.8) is of order h. The pair (2.9) is an additional classical example of unstable velocity-pressure pairs; see, [19, pp. 21–25].

2.1. Weak inf-sup bounds. In this section, we show that the unstable velocity-pressure pairs (2.9) and (2.10) satisfy a weaker form of the inf-sup condition (2.8). This condition identifies terms that can be used to stabilize the mixed method. To state the relevant form of the weaker inf-sup condition, we first review some results of [17, 30, 31] specialized to (2.9) and (2.10).

LEMMA 2.1. *Let $\mathbf{V}^h$ and S^h be the spaces defined in* (2.9). *Then, there exist positive constants C_1 and C_2 such that*

$$\sup_{\mathbf{v}^h\in\mathbf{V}^h} \frac{\displaystyle\int_\Omega p^h\nabla\cdot\mathbf{v}^h\, d\Omega}{\|\mathbf{v}^h\|_1} \geq C_1\|p^h\|_0 - C_2 h\|\nabla p^h\|_0 \qquad \forall\; p^h \in S^h\,. \tag{2.11}$$

Proof. By the definition of S^h, every $p^h \in S^h$ also belongs to $L_0^2(\Omega)$. As a result, there exists $\mathbf{w} \in \mathbf{H}_0^1(\Omega)$ such that

$$\int_\Omega p^h\nabla\cdot\mathbf{w}\, d\Omega \geq \widetilde{C}_1\|p^h\|_0\|\mathbf{w}\|_1\,. \tag{2.12}$$

Let $\mathbf{w}^h$ denote the interpolant of $\mathbf{w}$ out of $\mathbf{V}^h$. Then, from (1.4)

$$\|\mathbf{w}-\mathbf{w}^h\|_0 + h^{1/2}\|\mathbf{w}-\mathbf{w}^h\|_{\Gamma_h} \le Ch\|\mathbf{w}\|_1 \quad \text{and} \quad \|\mathbf{w}^h\|_1 \le C\|\mathbf{w}\|_1\,. \tag{2.13}$$

Using (2.12), (2.13), and the fact that all elements of S^h are C^0 functions,

$$\begin{aligned}
\frac{\left|\int_\Omega p^h \nabla\cdot\mathbf{w}^h\,d\Omega\right|}{\|\mathbf{w}^h\|_1} &\ge \frac{\left|\int_\Omega p^h\nabla\cdot\mathbf{w}^h\,d\Omega\right|}{C\|\mathbf{w}\|_1}\\
&= \frac{\left|\int_\Omega p^h\nabla\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega + \int_\Omega p^h\nabla\cdot\mathbf{w}\,d\Omega\right|}{C\|\mathbf{w}\|_1}\\
&\ge \frac{\int_\Omega p^h\nabla\cdot\mathbf{w}\,d\Omega}{C\|\mathbf{w}\|_1} - \frac{\left|\int_\Omega \nabla p^h\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega\right|}{C\|\mathbf{w}\|_1}\\
&\ge \frac{\widetilde{C}_1}{C}\|p^h\|_0 - \frac{\|\nabla p^h\|_0\|\mathbf{w}-\mathbf{w}^h\|_0}{C\|\mathbf{w}\|_0} \ge C_1\|p^h\|_0 - C_2h\|\nabla p^h\|_0\,.
\end{aligned} \tag{2.14}$$

Then, since

$$\sup_{\mathbf{v}^h\in\mathbf{V}^h,\,\mathbf{v}^h\neq\mathbf{0}} \frac{\int_\Omega p^h\nabla\cdot\mathbf{v}^h\,d\Omega}{\|\mathbf{v}^h\|_1} \ge \frac{\left|\int_\Omega p^h\nabla\cdot\mathbf{w}^h\,d\Omega\right|}{\|\mathbf{w}^h\|_1}\,, \tag{2.15}$$

the lemma is proved. □

For the velocity-pressure pair (2.10), the discontinuity of the pressure space necessitates some minor modifications in the statement and proof of the weak inf-sup condition.

LEMMA 2.2. *Let $\mathbf{V}^h$ and S^h be the spaces defined in* (2.10). *Then, there exist positive constants C_1 and C_2 such that*

$$\sup_{\mathbf{v}^h\in\mathbf{V}^h} \frac{\displaystyle\int_\Omega p^h\nabla\cdot\mathbf{v}^h\,d\Omega}{\|\mathbf{v}^h\|_1} \ge C_1\|p^h\|_0 - C_2h^{1/2}\|[p^h]\|_{\Gamma_h} \qquad \forall\ p^h\in S^h\,. \tag{2.16}$$

Proof. The pressure space S^h defined in (2.10) is also a subspace of $L^2_0(\Omega)$. Thus, there exists a $\mathbf{w}\in\mathbf{H}^1_0(\Omega)$ and a $\mathbf{w}^h\in\mathbf{V}^h$ that satisfy (2.12) and (2.13). Proceeding as in Lemma 2.1, we find that

$$\begin{aligned}
\frac{\left|\int_\Omega p^h\nabla\cdot\mathbf{w}^h\,d\Omega\right|}{\|\mathbf{w}^h\|_1} &\ge \frac{\left|\int_\Omega p^h\nabla\cdot\mathbf{w}^h\,d\Omega\right|}{C\|\mathbf{w}\|_1}\\
&\ge \frac{\int_\Omega p^h\nabla\cdot\mathbf{w}\,d\Omega}{C\|\mathbf{w}\|_1} - \frac{\left|\int_\Omega p^h\nabla\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega\right|}{C\|\mathbf{w}\|_1}\\
&\ge \frac{\widetilde{C}_1}{C}\|p^h\|_0 - \frac{\left|\int_\Omega p^h\nabla\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega\right|}{C\|\mathbf{w}\|_1}\,.
\end{aligned}$$

Using the fact that p^h is constant on each element Ω_e and integrating by parts gives

$$\int_\Omega p^h\nabla\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega = \sum_{\Omega_e}\int_{\Omega_e} p^h\nabla\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,d\Omega = \sum_{\Omega_e}\int_{\partial\Omega_e} p^h\mathbf{n}\cdot\left(\mathbf{w}^h-\mathbf{w}\right)\,dS\,.$$

Each interior face γ_f participates twice in this sum. Collecting the two integrals over the same face and using (2.13) we obtain

$$\sum_{\Omega_e} \int_{\partial\Omega_e} p^h \mathbf{n} \cdot \left(\mathbf{w}^h - \mathbf{w}\right)\, dS = \sum_{\gamma_f} \int_{\gamma_f} [p^h] \mathbf{n}_f \cdot \left(\mathbf{w}^h - \mathbf{w}\right)\, dS$$
$$\leq \left(\sum_{\gamma_f} \int_{\gamma_f} [p^h]^2\, dS \right)^{1/2} \left(\sum_{\gamma_f} \int_{\gamma_f} \left|\mathbf{w}^h - \mathbf{w}\right|^2\, dS \right)^{1/2} \leq C h^{1/2} \|[p^h]\|_{\Gamma_h} \|\mathbf{w}\|_1$$

which proves that

$$\frac{\left| \int_\Omega p^h \nabla \cdot \mathbf{w}^h\, d\Omega \right|}{\|\mathbf{w}^h\|_1} \geq C_1 \|p^h\|_0 - C_2 h^{1/2} \|[p^h]\|_{\Gamma_h} \qquad \forall\, p^h \in S^h\,. \tag{2.17}$$

Then, using (2.15), the lemma is proved. □

The terms

$$-h\|\nabla p^h\|_0 \quad \text{and} \quad -h^{1/2}\|[p^h]\|_0 \tag{2.18}$$

appearing in (2.11) and (2.16) quantify the inf-sup "deficiency" of the unstable pairs (2.9) and (2.10), respectively. This observation has been used implicitly in the design of stabilized methods; additional terms are introduced to counterbalance (2.18). For instance, consistently stabilized methods are based on the observation that adding a properly weighted residual of (2.1) to the continuity equation (2.2) will contribute a term that can offset $-h\|\nabla p^h\|_0$. The rest of the added terms are introduced to fulfill the consistency requirement and may actually be destabilizing. As a result, residual-based stabilization must rely on carefully selected values of parameters to keep such terms under control. Nonresidual stabilization follows the same idea but introduces balancing terms that do not involve residuals. For example, in [28], pressure jumps are added directly to the continuity equation to help offset the destabilizing effect of the $-h^{1/2}\|[p^h]\|_0$ term, while in [11] Brezzi and Pitkaranta use the first term in (2.18) to obtain a stabilized formulation for piecewise linear velocity-pressure pairs.

As a template for the design of stabilizing terms, (2.18) is insufficiently general. One is always led to consider either the gradient or the jumps of the pressure. The latter case also has the drawback of requiring face-based assembly and data structures. Below, we will derive an alternative characterization of the inf-sup "deficiency" for low-order spaces that is formulated in terms of abstract operators. This characterization does not involve gradients or jumps, does not depend on the space dimension or the type of the element shapes, and has no explicit dependence on mesh parameters. As a result, it leads to new classes of stabilized mixed methods with attractive computational properties. At this point it will suffice to specify only the ranges of the abstract operators needed to characterize the inf-sup deficiency. As we proceed to establish stability and prove convergence results, more assumptions will be added as needed. The first operator

$$\Pi_0 : L^2(\Omega) \mapsto R_0 \tag{2.19}$$

has a piecewise constant range; it will be used to stabilize (2.9). The second operator

$$\Pi_1 : L^2(\Omega) \mapsto R_1 \tag{2.20}$$

has a continuous range; it will be used to stabilize (2.10).

LEMMA 2.3. *There exists a positive constant C such that*

$$Ch\|\nabla p^h\|_0 \le \|p^h - \Pi_0 p^h\|_0 \qquad \forall\, p^h \in R_1\,. \tag{2.21}$$

There exists another positive constant C such that

$$Ch^{1/2}\|[p^h]\|_0 \le \|p^h - \Pi_1 p^h\|_0 \qquad \forall\, p^h \in R_0\,. \tag{2.22}$$

Proof. To prove (2.21), note that $\Pi_0 p^h$ is constant on each element Ω_e, and so $\nabla(\Pi_0 p^h)|_{\Omega_e} = 0$. As a result, using the inverse inequality (1.7),

$$\begin{aligned} h^2\|\nabla p^h\|_0^2 &= \sum_{\Omega_e} h^2\|\nabla p^h\|_{0,\Omega_e}^2 = \sum_{\Omega_e} h^2\|\nabla(p^h - \Pi_0 p^h)\|_{0,\Omega_e}^2 \\ &\le \sum_{\Omega_e} C_I\|p^h - \Pi_0 p^h\|_{0,\Omega_e}^2 = C_I\|p^h - \Pi_0 p^h\|_0^2\,. \end{aligned}$$

To prove (2.22), note that $\Pi_1 p^h \in R_1 \subset C^0(\Omega)$. Thus, $[(\Pi_1 p^h)|_{\gamma_f}] = 0$ on every interior element face γ_f. Using the inverse inequality (1.8),

$$\begin{aligned} h\|[p^h]\|_{\Gamma_h}^2 &= \sum_{\gamma_f} h\|[p^h]\|_{0,\gamma_f}^2 = \sum_{\gamma_f} h\|[p^h - \Pi_1 p^h]\|_{0,\gamma_f}^2 \\ &= h\|[p^h - \Pi_1 p^h]\|_{\Gamma_h}^2 \le C_I\|p^h - \Pi_1 p^h\|_0\,. \end{aligned}$$

∎

Using (2.21) and (2.22), results of Lemmas 2.1 and 2.2 can be combined into one statement. Let

$$\Pi = \begin{cases} \Pi_0 & \text{if } S^h \text{ is defined by (2.9)},\\ \Pi_1 & \text{if } S^h \text{ is defined by (2.10)}\,. \end{cases} \tag{2.23}$$

COROLLARY 2.4. *Let $\mathbf{V}^h$ and S^h be the spaces defined in* (2.9) *or* (2.10). *Then, there exist positive constants C_1 and C_2 whose values are independent of h and such that*

$$\sup_{\mathbf{v}^h \in \mathbf{V}^h} \frac{\displaystyle\int_\Omega p^h \nabla\cdot \mathbf{v}^h\, d\Omega}{\|\mathbf{v}^h\|_1} \ge C_1\|p^h\|_0 - C_2\|(I-\Pi)p^h\|_0 \qquad \forall\, p^h \in S^h\,. \tag{2.24}$$

We note that Π_0 and Π_1 are complementary in the sense that Π_0 acts on C^0 pressures and has a discontinuous range and Π_1 acts on discontinuous pressures and has a C^0 range. Note that besides the range assumption, Corollary 2.4 does not require any additional hypotheses about Π.

3. The new stabilized mixed methods. We will stabilize (2.4) by using

$$\frac{1}{2}\|(I-\Pi)p\|_0^2 \tag{3.1}$$

to compensate for the inf-sup deficiency of the low-order finite element pairs in (2.9) and (2.10). We add this term to (2.7) to obtain the following modified Lagrangian

functional:[2]

$$\widetilde{L}_m(\mathbf{v},q)=\frac{\lambda}{2}\int_\Omega |\nabla\mathbf{v}|^2\,d\Omega-\int_\Omega q\nabla\cdot\mathbf{v}\,d\Omega-\int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega-\frac{1}{2}\|(I-\Pi)q\|_0^2\,. \tag{3.2}$$

The saddle-point $(\widetilde{\mathbf{u}},\widetilde{p})$ of (3.2) satisfies the variational problem

$$A(\widetilde{\mathbf{u}},\mathbf{v})+B(\widetilde{p},\mathbf{v})=F(\mathbf{v}) \qquad \forall\,\mathbf{v}\in\mathbf{H}_0^1(\Omega), \tag{3.3}$$

$$B(q,\widetilde{\mathbf{u}})-G(\widetilde{p},q)=0 \qquad \forall\,q\in L_0^2(\Omega), \tag{3.4}$$

where

$$G(\widetilde{p},q)=\int_\Omega(\widetilde{p}-\Pi\widetilde{p})(q-\Pi q)\,d\Omega\,. \tag{3.5}$$

Equivalently, we can write (3.3)–(3.4) in the following form: seek $(\widetilde{\mathbf{u}},\widetilde{p})\in\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$ such that

$$\widetilde{Q}(\widetilde{\mathbf{u}},\widetilde{p};\mathbf{v},q)=F(\mathbf{v},q)\quad\forall\,(\mathbf{v},q)\in\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)\,, \tag{3.6}$$

where

$$\widetilde{Q}(\mathbf{u},p;\mathbf{v},q)=A(\mathbf{u},\mathbf{v})+B(p,\mathbf{v})+B(q,\mathbf{u})-G(p,q)\,. \tag{3.7}$$

The stabilized method is obtained by a restriction of (3.6) or, equivalently, of (3.3)–(3.4) to the finite element spaces (2.9) or (2.10). Thus, we seek $(\mathbf{u}^h,p^h)$ in $\mathbf{V}^h\times S^h$, such that

$$\widetilde{Q}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h)=F(\mathbf{v}^h,q^h)\quad\forall\,(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h\,. \tag{3.8}$$

The stabilization term (3.1) is not a residual of the Stokes equations. As a result, (3.8) is not a consistent finite element formulation of the Stokes equations. However, as noted earlier, for low-order elements, formally consistent stabilized methods [15, 20, 21] also lose their consistency, and so lack of consistency in our method should not be viewed as a serious flaw.

3.1. Comparison with the penalty method. The last term in (3.2) resembles the term that appears in the *penalized* Lagrangian

$$L_\epsilon(\mathbf{v},q)=\frac{\lambda}{2}\int_\Omega |\nabla\mathbf{v}|^2\,d\Omega-\int_\Omega q\nabla\cdot\mathbf{v}\,d\Omega-\int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega-\frac{\epsilon}{2}\|q\|_0^2\,. \tag{3.9}$$

However, the method (3.8) resulting from (3.2) is fundamentally different from a classical penalty approach based on (3.9). Taking first variations of (3.9) with respect to $\mathbf{v}$ and q gives the variational equation: seek $(\mathbf{u}_\epsilon,p_\epsilon)$ in $\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$ such that

$$A(\mathbf{u}_\epsilon,\mathbf{v})+B(p_\epsilon,\mathbf{v})=F(\mathbf{v}) \qquad \forall\,\mathbf{v}\in\mathbf{H}_0^1(\Omega), \tag{3.10}$$

$$B(q,\mathbf{u}_\epsilon)-\epsilon D(p_\epsilon,q)=0 \qquad \forall\,q\in L_0^2(\Omega), \tag{3.11}$$

[2]The modified Lagrangian (3.2) is in nondimensional form; its dimensional counterpart has the form

$$\widetilde{L}_m(\mathbf{v},q)=\frac{\mu}{2}\int_\Omega |\nabla\mathbf{v}|^2\,d\Omega-\int_\Omega q\nabla\cdot\mathbf{v}\,d\Omega-\int_\Omega \rho\mathbf{f}\cdot\mathbf{v}\,d\Omega-\frac{\alpha}{2}\|(I-\Pi)q\|_0^2\,,$$

where $\alpha=(\rho_{ref}u_{ref}\ell_{ref})^{-1}=\lambda/\mu$. It is important to note that α *is not a stabilization parameter*, but is merely a parameter introduced to make the dimensional form of the modified Lagrangian dimensionally correct; this observation is made obvious by examining the nondimensional form (3.2) in which the stabilization term $-\frac{1}{2}\|(I-\Pi)p\|_0^2$ is parameter free.

where $D(\cdot,\cdot)$ is the L^2 inner product. The second equation can be used to eliminate the pressure and to obtain an equation in terms of $\mathbf{u}_\epsilon$ only:

$$A(\mathbf{u}_\epsilon,\mathbf{v}) + \frac{1}{\epsilon}\int_\Omega (\nabla\cdot\mathbf{u}_\epsilon)(\nabla\cdot\mathbf{v})\,d\Omega = F(\mathbf{v}) \quad \forall\,\mathbf{v}\in\mathbf{H}^1_0(\Omega)\,. \tag{3.12}$$

Restriction of (3.12) to a discrete velocity space $\mathbf{V}^h$ leads to the classical penalty method. Alternatively, one can discretize (3.10)–(3.11), eliminate the discrete pressure from the linear system, and obtain another problem in terms of the discrete velocity only. Regardless of which version of the penalty method is used, i.e., *eliminate* and then *discretize*, or *discretize* and then *eliminate*, the ensuing penalty problem continues to require a discrete inf-sup compatibility condition. For instance, well-posedness of the *eliminate* and *discretize* method is subject to an inf-sup condition between $\mathbf{V}^h$ and an implicit pressure space defined by $\epsilon p_\epsilon = -\nabla\cdot\mathbf{u}_\epsilon$; see [24, 25]. A classical example of a failure in this method is the locking phenomena that occurs for linear velocities. In this case the implicit pair $(\mathbf{V}^h, S^h)$ is equivalent to the unstable P_1-P_0 element.

Because the penalty method still requires compatible finite element spaces, it is not a stabilization procedure. Rather, it is a solution method that allows one to solve the mixed problem more easily by uncoupling the velocity and pressure. In contrast to (3.10)–(3.11), in (3.8) we seek $(\mathbf{u}^h, p^h)\in\mathbf{V}^h\times S^h$ such that

$$A(\mathbf{u}^h,\mathbf{v}^h) + B(p^h,\mathbf{v}^h) = F(\mathbf{v}^h) \quad \forall\,\mathbf{v}^h\in\mathbf{V}^h \tag{3.13}$$

$$B(q^h,\mathbf{u}^h) - G(p^h,q^h) = 0 \qquad \forall\, q\in S^h\,. \tag{3.14}$$

In addition to the absence of a penalty parameter, another difference between (3.13)–(3.14) and the penalized problem (3.10)–(3.11) is that $G(\cdot,\cdot)$ vanishes for all pressures in the range of Π. As a result, this variable cannot be eliminated from (3.14). Of course, the main difference is that, as we shall see in the next section, (3.13)–(3.14) is stable for the low-order pairs in (2.9) and (2.10), while (3.10)–(3.11) may fail as $\epsilon\to 0$.

The penalty method can be extended to a stabilization procedure by using the stronger H^1-seminorm penalty $\epsilon/2\|\nabla q\|_0^2$ instead of the classical L^2 penalty $\epsilon/2\|q\|_0^2$. This leads to a stabilized finite element method proposed by Brezzi and Pitkaranta [11]. The bound (2.21) in Lemma 2.3 implies that for R^1 pressures their method and (3.8) have similar stability properties. However, (3.8) can be extended to constant pressures, while the method of [11] cannot.

4. Stability. To show that (3.8) is a stable variational problem, we have to additionally assume that Π is continuous as an operator $L^2(\Omega)\mapsto L^2(\Omega)$:

$$\|\Pi p\|_0 \le C\|p\|_0 \quad \forall\, p\in L^2(\Omega)\,. \tag{4.1}$$

Using (4.1), it is easy to show that $\widetilde{Q}$ is continuous, i.e.,

$$\widetilde{Q}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) \le C\left(\|\mathbf{u}^h\|_1 + \|p^h\|_0\right)\left(\|\mathbf{v}^h\|_1 + \|q^h\|_0\right) \tag{4.2}$$

for all $(\mathbf{u}^h,p^h)$ and $(\mathbf{v}^h,q^h)$ in $\mathbf{V}^h\times S^h$. We now prove the stability of the variational problem (3.8).

THEOREM 4.1. *Let $(\mathbf{V}^h,S^h)$ be one of the pairs* (2.9) *or* (2.10). *Then, there exists a positive constant C whose value is independent of h such that*

$$\sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{\widetilde{Q}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} \ge C\left(\|\mathbf{u}^h\|_1 + \|p^h\|_0\right) \ \forall\,(\mathbf{u}^h,p^h)\in\mathbf{V}^h\times S^h\,. \tag{4.3}$$

Proof. We will construct a pair $(\widehat{\mathbf{v}}^h, \widehat{q}^h)$, such that

$$\widetilde{Q}(\mathbf{u}^h, p^h; \widehat{\mathbf{v}}^h, \widehat{q}^h) \geq C \left(\|\mathbf{u}^h\|_1 + \|p^h\|_0 \right) \left(\|\widehat{\mathbf{v}}^h\|_1 + \|\widehat{q}^h\|_0 \right) .$$

Setting $(\mathbf{v}^h, q^h) = (\mathbf{u}^h, -p^h)$ yields

$$\widetilde{Q}(\mathbf{u}^h, p^h; \mathbf{u}^h, -p^h) = \lambda \|\nabla \mathbf{u}^h\|_0^2 + \| \left(I - \Pi \right) p^h \|_0^2 .$$

For a given arbitrary but fixed pressure $p^h \in S^h$, let $\mathbf{w}$ and $\mathbf{w}^h$ be the functions that satisfy (2.12) and (2.13). Assume that $\mathbf{w}^h$ is normalized so that

$$\|\nabla \mathbf{w}^h\|_0 = \sqrt{\lambda} \, \|p^h\|_0 . \tag{4.4}$$

From (2.14) and (2.21) if $\Pi = \Pi_0$ and (2.17) and (2.22) if $\Pi = \Pi_1$, we have that

$$\int_\Omega p^h \nabla \cdot \mathbf{w}^h \, d\Omega \geq C_1 \|p^h\|_0^2 - C_2 \| \left(I - \Pi \right) p^h \|_0 \|p^h\|_0 .$$

Setting $(\mathbf{v}^h, q^h) = (-\alpha \mathbf{w}^h, 0)$, where α is a real, positive parameter, together with the last inequality and (4.4), yields

$$\begin{aligned}
\widetilde{Q}(\mathbf{u}^h, p^h; -\alpha \mathbf{w}^h, 0) &= -\alpha \int_\Omega \nabla \mathbf{u}^h \cdot \nabla \mathbf{w}^h \, d\Omega + \alpha \int_\Omega p^h \nabla \cdot \mathbf{w}^h \, d\Omega \\
&\geq -\alpha \|\nabla \mathbf{u}^h\|_0 \|\nabla \mathbf{w}^h\|_0 + \alpha \big(C_1 \|p^h\|_0^2 - C_2 \|(I - \Pi) p^h\|_0 \|p^h\|_0 \big) \\
&\geq -\alpha \sqrt{\lambda} \|\nabla \mathbf{u}^h\|_0 \|p^h\|_0 + \alpha \big(C_1 \|p^h\|_0^2 - C_2 \|(I - \Pi) p^h\|_0 \|p^h\|_0 \big) .
\end{aligned}$$

As a result, for $(\mathbf{v}^h, q^h) = (\mathbf{u}^h - \alpha \mathbf{w}^h, -p^h)$, we have the bound

$$\begin{aligned}
\widetilde{Q}(\mathbf{u}^h, p^h; \mathbf{u}^h - \alpha \mathbf{w}^h, -p^h) \geq \|\nabla \mathbf{u}^h\|_0^2 + \| \left(I - \Pi \right) p^h \|_0^2 + \alpha C_1 \|p^h\|_0^2 \\
- \alpha \sqrt{\lambda} \|\nabla \mathbf{u}^h\|_0 \|p^h\|_0 - \alpha C_2 \| \left(I - \Pi \right) p^h \|_0 \|p^h\|_0 .
\end{aligned}$$

Using the ε-inequality with $\varepsilon = C_1/2$, we have that

$$\sqrt{\lambda} \|\nabla \mathbf{u}^h\|_0 \|p^h\|_0 \leq \frac{\lambda}{C_1} \|\nabla \mathbf{u}^h\|_0^2 + \frac{C_1}{4} \|p^h\|_0^2$$

and

$$C_2 \| \left(I - \Pi \right) p^h \|_0 \|p^h\|_0 \leq \frac{C_2^2}{C_1} \| \left(I - \Pi \right) p^h \|_0^2 + \frac{C_1}{4} \|p^h\|_0^2 .$$

In combination with the earlier lower bounds, these inequalities lead to

$$\begin{aligned}
&\widetilde{Q}(\mathbf{u}^h, p^h; \mathbf{u}^h - \alpha \mathbf{w}^h, -p^h) \\
&\qquad \geq \lambda \left(1 - \frac{\alpha}{C_1} \right) \|\nabla \mathbf{u}^h\|_0^2 + \frac{\alpha C_1}{2} \|p^h\|_0^2 + \left(1 - \frac{\alpha C_2^2}{C_1} \right) \| \left(I - \Pi \right) p^h \|_0^2 .
\end{aligned}$$

Choosing

$$\widehat{\alpha} = \min \left\{ \frac{C_1}{2}, \frac{C_1}{2C_2^2} \right\}$$

guarantees that

$$\left(1-\frac{\widehat{\alpha}}{C_1}\right)\geq\frac{1}{2}\quad\text{and}\quad\left(1-\frac{\widehat{\alpha}C_2^2}{C_1}\right)\geq\frac{1}{2}\,.$$

We now set

$$\widehat{\mathbf{v}}^h=\mathbf{u}^h-\widehat{\alpha}\mathbf{w}^h\qquad\text{and}\qquad\widehat{q}^h=-p^h\,.$$

It is then easy to see that

$$\begin{aligned}\widetilde{Q}(\mathbf{u}^h,p^h;\widehat{\mathbf{v}}^h,\widehat{q}^h)&\geq\frac{1}{2}\left(\lambda\|\nabla\mathbf{u}^h\|_0^2+\widehat{\alpha}C_1\|p^h\|_0^2+\|\left(I-\Pi\right)p^h\|_0^2\right)\\&\geq\frac{1}{6}\left(\sqrt{\lambda}\|\nabla\mathbf{u}^h\|_0+\sqrt{\widehat{\alpha}C_1}\|p^h\|_0+\|\left(I-\Pi\right)p^h\|_0\right)^2\\&\geq C\left(\|\nabla\mathbf{u}^h\|_0+\|p^h\|_0\right)^2,\end{aligned}$$

where the last bound follows from $(a+b+c)^2\leq 3(a^2+b^2+c^2)$. Finally, (4.4) implies that

$$\begin{aligned}\|\nabla\widehat{\mathbf{v}}^h\|_0+\|\widehat{q}^h\|_0&=\|\nabla(\mathbf{u}^h-\widehat{\alpha}\mathbf{w}^h)\|_0+\|p^h\|_0\leq\|\nabla\mathbf{u}^h\|_0+\widehat{\alpha}\|\nabla\mathbf{w}^h\|_0+\|p^h\|_0\\&\leq\|\nabla\mathbf{u}^h\|_0+\widehat{\alpha}\sqrt{\lambda}\|p^h\|_0+\|p^h\|_0\leq C\left(\|\nabla\mathbf{u}^h\|_0+\|p^h\|_0\right),\end{aligned}$$

i.e., $(\widehat{\mathbf{v}}^h,\widehat{q}^h)$ is bounded by $(\mathbf{u}^h,p^h)$ in the norm of $\mathbf{H}_0^1(\Omega)\times L^2(\Omega)$. This proves the theorem. □

Together, (4.2) and (4.3) imply that (3.8) is a stable variational problem.

Remark 1. Because $\widetilde{Q}$ is symmetric, (4.3) is sufficient to establish weak coercivity of this form.

Remark 2. The stabilized problem (3.8) is well-posed if (4.2) and (4.3) hold, i.e., if the bilinear form $\widetilde{Q}$ is continuous and weakly coercive. From the proof of Theorem 4.1, it is clear that weak coercivity depends only on Π having the appropriate range. The continuity of $\widetilde{Q}$, on the other hand, is impossible without assuming that Π itself is continuous.

5. Error estimates. To prove convergence of stabilized solutions, the properties of Π must be augmented by an approximation hypothesis. We will assume that

$$\|(I-\Pi)p\|_0\leq Ch\|p\|_1 \tag{5.1}$$

for every $p\in H^1(\Omega)$.

THEOREM 5.1. *Let* $(\mathbf{V}^h,S^h)$ *denote one of the spaces* (2.9) *or* (2.10), *let* $(\mathbf{u},p)$ *be the solution of the Stokes problem* (2.4), *and let* $(\mathbf{u}^h,p^h)\in\mathbf{V}^h\times S^h$ *solve the stabilized mixed problem* (3.8), *where the operator* Π *defined in* (2.23) *satisfies* (4.1). *Then,*

$$\begin{aligned}&\|\mathbf{u}-\mathbf{u}^h\|_1+\|p-p^h\|_0\\&\qquad\leq C\left(\inf_{q^h\in S^h}\|p-q^h\|_0+\inf_{\mathbf{v}^h\in\mathbf{V}^h}\|\mathbf{u}-\mathbf{v}^h\|_1+\|(I-\Pi)p\|_0\right).\end{aligned}\tag{5.2}$$

Proof. Since $(\mathbf{V}^h, S^h)$ is a subspace of $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$, we have from (2.4) that

$$A(\mathbf{u}, \mathbf{v}^h) + B(p, \mathbf{v}^h) = F(\mathbf{v}^h) \qquad \forall\, \mathbf{v}^h \in \mathbf{V}^h,$$
$$B(q^h, \mathbf{u}) = 0 \qquad \forall\, q^h \in S^h.$$

Subtracting these equations from (3.13)–(3.14) yields

$$A(\mathbf{u}^h - \mathbf{u}, \mathbf{v}^h) + B(p^h - p, \mathbf{v}^h) = 0 \qquad \forall\, \mathbf{v}^h \in \mathbf{V}^h,$$
$$B(q^h, \mathbf{u}^h - \mathbf{u}) = G(p^h, q) \quad \forall\, q^h \in S^h,$$

or, equivalently,

$$\widetilde{Q}(\mathbf{u}^h - \mathbf{u}, p^h - p; \mathbf{v}^h, q^h) = G(p, q^h) \quad \forall\, (\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h\,. \tag{5.3}$$

Let $(\mathbf{w}^h, r^h)$ be an arbitrary pair in $\mathbf{V}^h \times S^h$. We estimate the discrete error

$$\|\mathbf{u}^h - \mathbf{w}^h\|_1 + \|p^h - r^h\|_0$$

using the weak coercivity bound (4.3) and the error "orthogonality" (5.3):

$$\begin{aligned}
&C\left(\|\mathbf{u}^h - \mathbf{w}^h\|_1 + \|p^h - r^h\|_0\right) \\
&\quad\le \sup_{(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h} \frac{\widetilde{Q}(\mathbf{u}^h - \mathbf{w}^h, p^h - r^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} \\
&\quad= \sup_{(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h} \frac{\widetilde{Q}(\mathbf{u}^h - \mathbf{u}, p^h - p; \mathbf{v}^h, q^h) + \widetilde{Q}(\mathbf{u} - \mathbf{w}^h, p - r^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} \\
&\quad= \sup_{(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h} \frac{G(p, q^h) + \widetilde{Q}(\mathbf{u} - \mathbf{w}^h, p - r^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0}\,.
\end{aligned}$$

From (4.2) we have that

$$\widetilde{Q}(\mathbf{u} - \mathbf{w}^h, p - r^h; \mathbf{v}^h, q^h) \le C\left(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0\right)\left(\|\mathbf{v}^h\|_1 + \|q^h\|_0\right)$$

and from (4.1) we have that

$$G(p, q^h) \le C G(p, p)^{1/2} \|q^h\|_0\,.$$

As a result, there exists a positive constant C such that

$$\begin{aligned}
&C\left(\|\mathbf{u}^h - \mathbf{w}^h\|_1 + \|p^h - r^h\|_0\right) \\
&\le \sup_{(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h} \frac{G(p, p)^{1/2}\|q^h\|_0 + \left(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0\right)\left(\|\mathbf{v}^h\|_1 + \|q^h\|_0\right)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} \\
&\le G(p, p)^{1/2} + \left(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0\right) = \left(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0\right) + \|(I - \Pi)p\|_0\,.
\end{aligned}$$

To complete the proof, we use the triangle inequality to obtain

$$\begin{aligned}
&\|\mathbf{u} - \mathbf{u}^h\|_1 + \|p - p^h\|_0 \\
&\qquad\le \left(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0\right) + \left(\|\mathbf{u}^h - \mathbf{w}^h\|_1 + \|p^h - r^h\|_0\right) \\
&\qquad\le C\Big(\|\mathbf{u} - \mathbf{w}^h\|_1 + \|p - r^h\|_0 + \|(I - \Pi)p\|_0\Big)\,,
\end{aligned}$$

and then take the infimum over $\mathbf{w}^h \in \mathbf{V}^h$ and $r^h \in S^h$. ∎

Together with the assumption (5.1) this theorem can be used to show that solutions of (3.8) converge optimally with respect to the solution regularity.

COROLLARY 5.2. *Assume that* $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega) \times L_0^2(\Omega) \cap H^1(\Omega)$ *solves the Stokes problem* (2.1)–(2.2) *and that* $(\mathbf{u}^h, p^h)$ *is the solution of the stabilized mixed problem* (3.8), *where the operator* Π *defined in* (2.23) *satisfies* (4.1) *and* (5.1). *Then,*

$$\|\mathbf{u} - \mathbf{u}^h\|_1 + \|p - p^h\|_0 \leq Ch\left(\|u\|_2 + \|p\|_1\right) . \tag{5.4}$$

Proof. The assertion follows immediately from (5.2) using (1.4), (1.6), and (5.1). ∎

6. Implementation. Among the attractive features of our stabilization approach is the great flexibility in the definition of the stabilization term (3.1). The main prerequisite to achieve stabilization of the mixed method with the low-order finite element pairs (2.9) and (2.10) is to choose a Π with the appropriate range. The simplest way to accomplish this is to use standard finite element projection or interpolation operators. Then, the remaining assumptions about Π are easily verified.

From a practical viewpoint the main factors in the choice of Π are simplicity and locality, i.e., computation of its action must be done at the element level using only standard nodal data structures. With this in mind, a suitable choice of Π_0 to stabilize the lowest equal order pair (2.9) is a local L^2 projection operator. Given a function $q \in L^2(\Omega)$ we define $\Pi_0 : L^2(\Omega) \mapsto R_0$ by $\Pi_0 q = q^h \in R_0$ if and only if

$$\int_{\Omega_e} (\Pi_0 q - q)\, d\Omega_e = 0 \qquad \forall\, \Omega_e \in \mathcal{T}_h\,. \tag{6.1}$$

It is easy to see that

$$\Pi_0 q|_{\Omega_e} = \frac{1}{V(\Omega_e)} \int_{\Omega_e} q\, d\Omega_e$$

is the element average of q and that Π_0 satisfies both assumptions (4.1) and (5.1); see [18, p. 102].

A suitable choice of Π_1 to stabilize the lowest order conforming pair (2.10) is a Clement-like interpolant; see [18, p. 110]. Instead of using a projection onto a patch of elements that share the same node we choose to define our interpolant by using a projection onto the dual (or complementary) volume associated with each node. For piecewise constant pressures this choice leads to a particularly simple formula for the action of Π_1 that does not require explicit construction of a dual cell. Specifically, we define $\Pi_1 : L^2(\Omega) \mapsto R_1$ as follows. For a given node $\mathcal{N}_i$ in $\mathcal{T}_h$, let $\widehat{\Omega}_i$ denote its dual volume. Given a function $q \in L^2(\Omega)$, let q_i be the constant function on $\widehat{\Omega}_i$ that minimizes the functional

$$J_i(q) = \frac{1}{2} \int_{\widehat{\Omega}_i} (q_i - q)^2 \, d\Omega\,; \tag{6.2}$$

then set

$$\Pi_1 q = \sum_{i=1}^{N_{nodes}} q_i N_i(\mathbf{x}) \in R_1\,, \tag{6.3}$$

where N_i denotes the nodal basis of R_1 and N_{nodes} is the number of nodes in $\mathcal{T}_h$. The action of the operator defined in (6.3) can be computed locally at the element level and has the same properties as the usual Clement interpolant, i.e., (4.1) and (5.1) are satisfied. For $q = q^h \in R_0$, the functional in (6.2) further simplifies to

$$J_i(q^h) = \sum_{\Omega_e \cap \Omega_i \neq 0} V_i(\Omega_e)(q_i - q_e^h)^2,$$

where q_e^h is the constant value of q^h on Ω_e and $V_i(\Omega_e)$ is the volume fraction of the element Ω_e that belongs to the dual cell $\widehat{\Omega}_i$ associated with node $\mathcal{N}_i$. For constant pressures, we can choose

$$V_i(\Omega_e) = V(\Omega_e)/n_e\,,$$

where n_e is the number of nodes in Ω_e. Minimization of J_i then yields the formula

$$q_i = \frac{\sum_{\Omega_e \in \Omega_i} V_i(\Omega_e) q_e^h}{\sum_{\Omega_e \in \Omega_i} V_i(\Omega_e)} = \sum_{\Omega_e \in \Omega_i} d_{ie} q_e^h\,,$$

i.e., the nodal values of $\Pi_1 q^h$ are area weighted averages of the surrounding constant pressure values of q^h.

The stabilized mixed problem gives rise to a linear system of algebraic equations with a matrix that has the form

$$\begin{bmatrix} \mathbb{A} & \mathbb{B}^T \\ \mathbb{B} & -\mathbb{G} \end{bmatrix}. \tag{6.4}$$

The matrices $\mathbb{A}$ and $\mathbb{B}$ are assembled in the usual manner from the bilinear forms $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$, respectively, and $\mathbb{G}$ is a symmetric, semidefinite matrix generated at the element level from $G(\cdot,\cdot)$. The form of this matrix depends on the particular operator Π employed in the stabilization. However, computation of $\mathbb{G}$ is completely local and can be accomplished by augmenting the standard nodal assembly process by a few simple calculations. For example, the only information needed to compute $\mathbb{G}$ in the case of Π_1 is the area of each element. This information should be readily available during the standard assembly process; moreover, calculation of $\mathbb{G}$ is simple in comparison to the calculations required to determine $\mathbb{A}$ and $\mathbb{B}$.

It is also easy to see that $\mathbb{G}$ is a sparse matrix. In the case of $\Pi = \Pi_0$, its sparsity pattern is the same as for the standard nodal R_1 pressure mass matrix. In the case of $\Pi = \Pi_1$, the original mass matrix associated with piecewise constant pressures is diagonal while $\mathbb{G}$ is not. Nevertheless, the important point is that the action of $(I - \Pi)$ in both cases is obtained by multiplication of pressure degrees of freedom by a sparse matrix rather than by an inversion of a mass matrix as occurs in determining L^2 projections. As a result, in the context of iterative solution methods, our stabilization method is very efficient as it requires only one sparse matrix-vector multiply per iteration.

7. Numerical examples. In this section, we report on some numerical results obtained using the stabilized method (3.8). The main goal of these experiments is to verify the convergence rates of (5.2) for the low-order pairs (2.9) and (2.10) in two and three space dimensions. For each pair of spaces, we consider both simplicial and nonsimplicial partitions $\mathcal{T}_h$ of the computational domain into finite elements. Figure

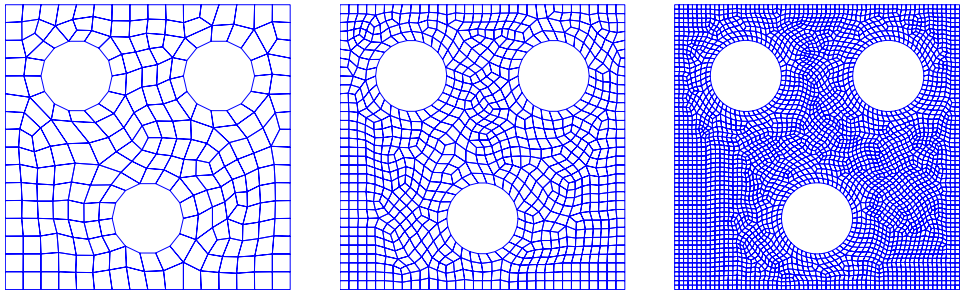


FIG. 7.1. *A sequence of refined nonuniform quadrilateral grids.*

7.1 shows an example of a nonsimplicial sequence of grids used for a convergence study in two dimensions. The following error norms are used for the investigation of convergence rates:

$$e^h_{uL_2} = \|\mathbf{u}^h - \mathbf{u}\|_0 = \sqrt{\sum_{i=1}^{d} \int_\Omega (u_i^h - u_i)^2 d\Omega}, \tag{7.1}$$

$$e^h_{uH_1} = \|\mathbf{u}^h - \mathbf{u}\|_1 = \sqrt{\sum_{i=1}^{d} \int_\Omega \nabla(u_i^h - u_i) \cdot \nabla(u_i^h - u_i) d\Omega}, \tag{7.2}$$

$$e^h_{pL_2} = \|p^h - p\|_0 = \sqrt{\int_\Omega (p^h - p)^2 d\Omega}\,, \tag{7.3}$$

where d denotes the spatial dimension, u_i, $i = 1, \ldots, d$, denote the components of the vector $\mathbf{u}$, and $(\mathbf{u}^h, p^h)$ denotes the stabilized finite element approximation of the exact solution $(\mathbf{u}, p)$. To estimate convergence rates, we select a pair of smooth functions $(\mathbf{u}, p)$, with $\mathbf{u}$ solenoidal and p having zero mean, and evaluate the Stokes equations to generate the source term $\mathbf{f}$ and the boundary data. This synthetic data is then used by (3.8) to approximate the smooth exact solution on a sequence of grids.

The first example is for a unit square with[3] $\lambda = 1$ and the smooth exact solution

$$u_1 = x + x^2 - 2xy + x^3 - 3xy^2 + x^2 y, \tag{7.4}$$

$$u_2 = -y - 2xy + y^2 - 3x^2 y + y^3 - xy^2, \tag{7.5}$$

$$p = xy + x + y + x^3 y^2 - 4/3\,. \tag{7.6}$$

The values of $\mathbf{u}$ on the boundary of the square are constrained to those given by (7.4) and (7.5). To remove the constant pressure mode from the numerical solution, the constraint

$$\int_\Omega p(\mathbf{x}) d\Omega = 0 \tag{7.7}$$

is also imposed. Results for stabilized triangular elements P_1-P_1 and P_1-P_0 and stabilized quadrilateral elements Q_1-Q_1 and Q_1-P_0 are shown in Figure 7.2. The $e^h_{uH_1}$ errors for the continuous pressure elements (P_1-P_1 and Q_1-Q_1) and the discontinuous

[3]The implementation of the new stabilized method for nonconstant viscosity is straightforward by using, e.g., viscosity values at quadrature points.

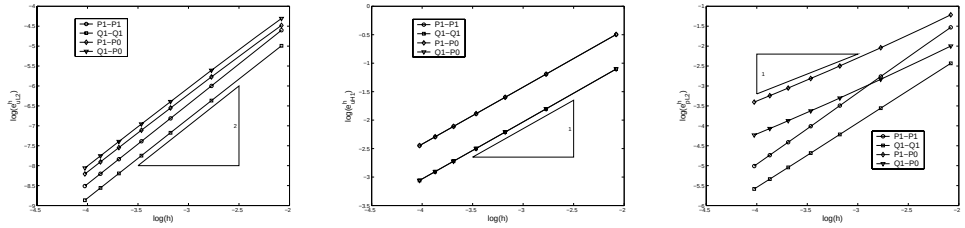


FIG. 7.2. *Errors for the first two-dimensional example (structured meshes).*

TABLE 7.1
Solution errors for triangular stabilized elements normalized with respect to results for the stable MINI element.

	P_1-P_1				P_1-P_0			
$1/h$	$e^h_{uL_2}$	$e^h_{uH_1}$	$e^h_{pL_2}$	e_{div}	$e^h_{uL_2}$	$e^h_{uH_1}$	$e^h_{pL_2}$	e_{div}
8	0.892	0.985	0.588	0.976	1.009	0.986	0.807	0.823
16	0.890	0.996	0.583	0.976	1.114	0.997	1.201	0.826
24	0.890	0.999	0.574	0.976	1.155	1.000	1.552	0.827
32	0.889	1.000	0.565	0.976	1.176	1.001	1.872	0.827
40	0.889	1.001	0.556	0.976	1.189	1.001	2.167	0.828
48	0.889	1.001	0.549	0.976	1.198	1.002	2.442	0.828
56	0.889	1.001	0.542	0.976	1.204	1.002	2.698	0.828

pressure elements (P_1-P_0 and Q_1-P_0) are nearly identical. Although not predicted by theory, the $e^h_{pL_2}$ line segment slopes for the continuous pressure elements exceed those of the discontinuous pressure elements. In all cases, the theoretical convergence rates are confirmed.

For purposes of comparison, we show in Table 7.1 the P_1-P_1 and P_1-P_0 results of Figures 7.2 normalized with respect to those of the stable MINI element. Also shown in the table are the normalized values of the maximum error in the divergence within an element as defined by

$$e_{div} = \max_e \left| \int_{\Gamma_e} \mathbf{u} \cdot \mathbf{n} \, d\Gamma_e \right| , \tag{7.8}$$

where Γ_e is the boundary of element e and $\mathbf{n}$ is the unit outward normal of Γ_e. The normalized maximum errors in the divergence are close to the stable MINI element for both the P_1-P_1 and P_1-P_0 elements. The higher normalized values of $e^h_{pL_2}$ for the P_1-P_0 elements are consistent with previous comments regarding continuous and discontinuous pressure elements.

The second example uses the same exact solution, but now the square domain has three circular cutouts as shown in Figure 7.1. Note that it is necessary to adjust the constant value of 4/3 in (7.6) to satisfy (7.7). Meshes of triangles were obtained from the quadrilateral meshes by splitting each quadrilateral into two triangles. Plots of the error norms for the different element types are shown in Figure 7.3. In this figure, $h_e = 1/\sqrt{N_e}$ where N_e is the number of quadrilateral elements in the mesh. As expected, the error norms become smaller as the meshes are refined.

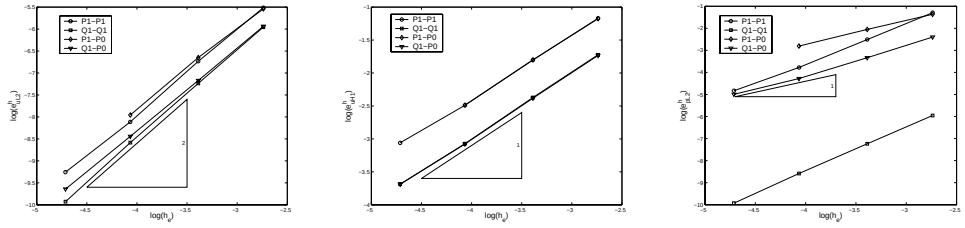


FIG. 7.3. *Errors for the second two-dimensional example (unstructured meshes).*

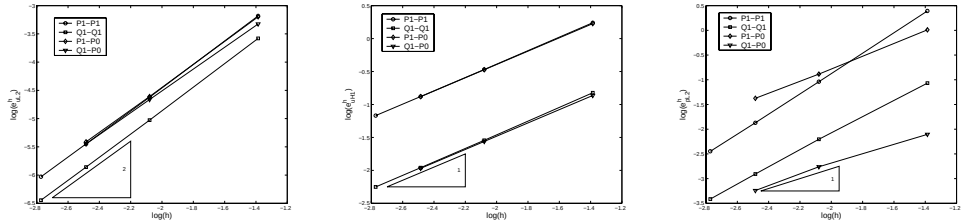


FIG. 7.4. *Errors for the three-dimensional example.*

The final example is for a unit cube with $\lambda = 1$ and the smooth exact solution

$$u_1 = x + x^2 + xy + x^3 y, \tag{7.9}$$

$$u_2 = y + xy + y^2 + x^2 y^2, \tag{7.10}$$

$$u_3 = -2z - 3xz - 3yz - 5x^2 yz, \tag{7.11}$$

$$p = xyz + x^3 y^3 z - 5/32\,. \tag{7.12}$$

The hexahedral meshes have $(1/h)^3$ elements whereas the tetrahedral meshes have $6(1/h)^3$ elements. Plots of the error norms versus element length are shown in Figure 7.4. As was the case for the two-dimensional examples, the theoretical convergence rates are confirmed. Again, the $e^h_{pL_2}$ line segment slopes for the continuous pressure elements are larger than those for the discontinuous pressure elements.

For further examples and numerical studies, we refer to [14].

8. Conclusions. We have formulated a new approach to stabilization of low-order velocity-pressure pairs for the incompressible Stokes equations. Central to our approach is the characterization of the LBB deficiency of the unstable pairs in terms of suitable operators, and their subsequent application in the formulation of a stabilized mixed variational equation. This characterization remains valid for a broad range of operators which makes our stabilization technique extremely flexible and leads to stabilized mixed methods with attractive computational properties. Most notably, our methods do not require selection of mesh-dependent stabilization parameters, retain the symmetry of the original equations, and can be implemented at the element level with minimal additional cost. Numerical examples presented in this paper demonstrate the excellent stability and accuracy properties of the new methods.

Acknowledgments. We thank the anonymous referees for their careful reading of the manuscript and suggestions that helped to improve the paper.

REFERENCES

[1] C. Baiocchi and F. Brezzi, *Stabilization of unstable numerical methods,* in Current Problems of Analysis and Mathematical Physics, (Taormina, 1992), Univ. Roma "La Sapienza," Rome, 1993, pp. 59–63.

[2] T. Barth, P. Bochev, M. Gunzburger, and J. Shadid, *A taxonomy of consistently stabilized finite element methods for the Stokes problem*, SIAM J. Sci. Comput., 25 (2004), pp. 1585–1607.

[3] R. Becker and M. Braack, *A Modification of the Least-Squares Stabilization for the Stokes Equations*, Report 03/00, University of Heidelberg, Heidelberg, Germany, 2000. Available online at http://numerik.iwr.uni-heidelberg.de.

[4] R. Becker and M. Braack, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2000), pp. 173–199.

[5] M. Behr, L. Franca, and T. Tezduyar, *Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows,* Comput. Methods Appl. Mech. Engrg., 104 (1993), pp. 31–48.

[6] J. Blasco and R. Codina, *Stabilized finite element method for the transient Navier–Stokes equations based on a pressure gradient projection*, Comput. Methods Appl. Mech. Engrg., 182 (2000), pp. 277–300.

[7] J. Blasco and R. Codina, *Space and time error estimates for a first-order, pressure stabilized finite element method for the incompressible Navier–Stokes equations*, Appl. Numer. Math., 38 (2001), pp. 475–497.

[8] P. Bochev and M. Gunzburger, *An absolutely stable pressure-Poisson stabilized finite element method for the Stokes equations*, SIAM J. Numer. Anal., 42 (2004), pp. 1189–1207.

[9] J. M. Boland and R. A. Nicolaides, *Stable and semistable low order finite elements for viscous flows*, SIAM J. Numer. Anal., 22 (1985), pp. 474–492.

[10] F. Brezzi, *On existence, uniqueness, and approximation of saddle-point problems arising from Lagrangian multipliers,* RAIRO Model. Math. Anal. Numer., 21 (1974), pp. 129–151.

[11] F. Brezzi and J. Pitkaranta, *On the stabilization of finite element approximations of the Stokes equations*, in Efficient Solutions of Elliptic Systems, Notes Numer. Fluid Mech. 10, W. Hackbusch, ed., Viewig, Braunschweig, 1984, pp. 11–19.

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, SIAM Classics in Appl. Math. 40, SIAM, Philadelphia, 2002.

[13] R. Codina and J. Blasco, *Analysis of a pressure stabilized finite element approximation of the stationary Navier–Stokes equations*, Numer. Math., 87 (2000), pp. 59–81.

[14] C. Dohrmann and P. Bochev, *A stabilized finite element method for the Stokes problem based on polynomial pressure projections*, Internat. J. Numer. Methods Fluids, 46 (2004), pp. 183–201.

[15] J. Douglas and J. Wang, *An absolutely stabilized finite element method for the Stokes problem,* Math. Comp., 52 (1989), pp. 495–508.

[16] R. Falk, *An analysis of the penalty method and extrapolation for the stationary Stokes equations*, in Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky, ed., AICA, 1975, pp. 66–69.

[17] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least-squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

[18] V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations,* Springer, Berlin, 1986.

[19] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows,* Academic Press, Boston, 1989.

[20] T. J. R. Hughes and L. P. Franca, *A new finite element formulation for computational fluid dynamics*: VII. *The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65 (1987), pp. 85–96.

[21] T. J. R. Hughes, L. P. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics*: V. *Circumventing the Babuska–Brezzi condition: A stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations,* Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[22] T. J. R. Hughes, W. Liu, and A. Brooks, *Finite element analysis of incompressible viscous flows by the penalty function formulation*, J. Comput. Phys., 30 (1979), pp. 1–60.

[23] K. Jansen, S. Collis, C. Whiting, and F. Shakib, *A better consistency for low-order stabilized finite element methods*, Comput. Methods Appl. Mech. Engrg., 174 (1999), pp. 153–170.

[24] C. Johnson and J. Pitkaranta, *Analysis of mixed finite element methods related to reduced integration*, Math. Comp., 42 (1984), pp. 9–23.

[25] T. J. Oden, *RIP-methods for Stokesian flow*, in Finite Elements in Fluids, Vol. 4, R. H. Gallagher, D. H. Norrie, J. T. Oden, and O. C. Zienkiewicz, eds., John Wiley, New York, 1982, pp. 305–318.

[26] J. Pitkaranta and T. Saarinen, *A multigrid version of a simple finite element method for the Stokes problem*, Math. Comp., 45 (1985), pp. 1–14.

[27] P. R. Schunk, M. Heroux, R. Rao, T. Baer, S. Subia, and A. Sun, *Iterative solvers and preconditioners for fully-coupled finite element formulations of incompressible fluid mechanics and related transport problems*, Tech. report, SAND 2001-3512J, Sandia National Laboratories, Albuquerque, NM, 2001.

[28] D. J. Silvester and N. Kechkar, *Stabilized bilinear-constant velocity-pressure finite elements for the conjugate gradient solution of the Stokes Problem.*, Comput. Methods Appl. Mech. Engrg., 79 (1990), pp. 71–86.

[29] D. J. Silvester, *Optimal low order finite element methods for incompressible flow*, Comput. Methods Appl. Mech. Engrg., 111 (1994), pp. 357–368.

[30] R. Stenberg, *Error analysis of some finite element methods for the Stokes problem*, Math. Comp., 54 (1990), pp. 495–508.

[31] R. Verfurth, *Error estimates for a mixed finite element approximation of the Stokes problem*, RAIRO Anal. Numer., 18 (1984), pp. 175–182.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 102–119

CONVERGENCE OF THE LLOYD ALGORITHM FOR COMPUTING CENTROIDAL VORONOI TESSELLATIONS*

QIANG DU†, MARIA EMELIANENKO‡, AND LILI JU§

Abstract. Centroidal Voronoi tessellations (CVTs) are Voronoi tessellations of a bounded geometric domain such that the generating points of the tessellations are also the centroids (mass centers) of the corresponding Voronoi regions with respect to a given density function. Centroidal Voronoi tessellations may also be defined in more abstract and more general settings. Due to the natural optimization properties enjoyed by CVTs, they have many applications in diverse fields. The Lloyd algorithm is one of the most popular iterative schemes for computing the CVTs but its theoretical analysis is far from complete. In this paper, some new analytical results on the local and global convergence of the Lloyd algorithm are presented. These results are derived through careful utilization of the optimization properties shared by CVTs. Numerical experiments are also provided to substantiate the theoretical analysis.

Key words. centroidal Voronoi tessellations, k-means, optimal vector quantizer, Lloyd algorithm, global convergence, convergence rate

AMS subject classifications. 65D99, 65C20

DOI. 10.1137/040617364

1. Introduction.

A centroidal Voronoi tessellation (CVT) is a special Voronoi tessellation of a given set such that the associated generating points are the centroids (centers of mass) of the corresponding Voronoi regions with respect to a predefined density function [7]. CVTs are indeed special as they enjoy very natural optimization properties which make them very popular in diverse scientific and engineering applications that include art design, astronomy, clustering, geometric modeling, image and data analysis, resource optimization, quadrature design, sensor networks, and numerical solution of partial differential equations [1, 2, 3, 4, 7, 8, 9, 10, 11, 13, 14, 17, 15, 26, 29, 30, 31, 39, 44, 45]. In particular, CVTs have been widely used in the design of optimal vector quantizers in electrical engineering [25, 28, 40, 43]. They are also related to the so-called method of k-means [27] in clustering analysis. CVTs can also be defined in more general cases such as those constrained to a manifold [12, 11] or those corresponding to anisotropic metrics [16, 18], and other abstract settings [7, 9].

For modern applications of the CVT concept in large-scale scientific and engineering problems, it is important to develop robust and efficient algorithms for constructing CVTs in various settings. Historically, a number of algorithms have been studied and widely used [7, 19, 25, 27, 38]. A seminal work is the algorithm first developed in the 1960s at Bell Laboratories by S. Lloyd which remains to this day one of the most popular methods due to its effectiveness and simplicity. The algorithm was later officially published in [35]. It is now commonly referred to as the Lloyd algorithm and is the main focus of this paper.

*Received by the editors October 20, 2004; accepted for publication (in revised form) August 15, 2005; published electronically February 8, 2006. This work was supported in part by NSF grants DMS-0409297, CCF-0430349, and ITR DMR-0205232.

http://www.siam.org/journals/sinum/44-1/61736.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu).

‡Department of Mathematical Sciences, Carnegie Mellon University, PA 15213 (masha@cmu.edu).

§Department of Mathematics, University of South Carolina, Columbia, SC 29208 (ju@math.sc.edu).

The Lloyd algorithm has many elegant and simple interpretations [7], but to present it more rigorously, we begin with a more detailed description of the CVT. First of all, we recall the concept of the Voronoi tessellation (or Voronoi diagram). A Voronoi tessellation refers to a tessellation of a given domain $\Omega \in \mathbb{R}^N$ by the Voronoi regions $\{V_i\}_{i=1}^k$ associated with a set of given *generating points* or *generators* $\{\mathbf{z}_i\}_{i=1}^k \subset \Omega$ [22, 33, 41]. For each i, $\{V_i\}_{i=1}^k$ consists of all points in the domain Ω that are closer to $\mathbf{z}_i$ than to all the other generating points. For a given density function ρ defined on Ω, we may define the centroids, or mass centers, of regions $\{V_i\}_{i=1}^k$ by

$$\mathbf{z}_i^* = \frac{\displaystyle\int_{V_i} \mathbf{y}\rho(\mathbf{y})\, d\mathbf{y}}{\displaystyle\int_{V_i} \rho(\mathbf{y})\, d\mathbf{y}}\,. \tag{1.1}$$

Then, a CVT refers to a Voronoi tessellation for which the generators themselves are the centroids of their respective Voronoi regions, that is, $\mathbf{z}_i = \mathbf{z}_i^*$ for all i. We refer to [7] for a more comprehensive review of the mathematical theory and diverse applications of CVTs.

In the seminal work of Lloyd on the least square quantization [35], one of the algorithms proposed for computing the CVTs (referred to as the optimal quantizers in the particular setting) is an iterative algorithm consisting of the following simple steps: starting from an initial Voronoi tessellation corresponding to an old set of generators, a new set of generators is defined by the mass centers of the Voronoi regions. This process is continued until a certain stopping criterion is met. With the notation given above, the Lloyd algorithm for constructing CVTs can be described more precisely by the following procedure.

ALGORITHM 1.1 (Lloyd algorithm for computing CVTs).

Input:

Ω, *the domain of interest;* ρ, *a density function defined on* Ω;
k, *number of generators;* $\{\mathbf{z}_i\}_{i=1}^k$, *the initial set of generators.*

Output:

$\{V_i\}_{i=1}^k$, *a CVT with* k *generators* $\{\mathbf{z}_i\}_{i=1}^k$ *in* Ω.

Iteration:

1. *Construct the Voronoi tessellation* $\{V_i\}_{i=1}^k$ *of* Ω *with generators* $\{\mathbf{z}_i\}_{i=1}^k$.
2. *Take the mass centroids of* $\{V_i\}_{i=1}^k$ *as the new set of generators* $\{\mathbf{z}_i\}_{i=1}^k$.
3. *Repeat procedures* 1 *and* 2 *until some stopping criterion is met.*

Given a set of points $\{\mathbf{z}_i\}_{i=1}^k$ and a tessellation $\{V_i\}_{i=1}^k$ of the domain, we may define the *energy functional* or the *distortion value* for the pair $(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k)$ by

$$\mathcal{H}\Big(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k\Big) = \sum_{i=1}^k \int_{V_i} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_i|^2\, d\mathbf{y}\,.$$

The minimizer of $\mathcal{H}$ necessarily forms a CVT which illustrates the optimization property of the CVT [7]. Meanwhile, it is easy to see that the Lloyd algorithm is an energy descent iteration, which gives strong indications of its practical convergence.

The Lloyd algorithm sparked enormous research efforts in later years and its variants have been proposed and studied in many contexts for different applications [25, 28, 40, 43, 35, 24, 23, 32, 34, 36]. A particular extension was made in [30] to combine the deterministic features of the Lloyd algorithm with some random sampling techniques. Despite its great success in applications and a large number of studies over

the last few decades, only limited theoretical results on the Lloyd algorithm have been obtained [7] and many fundamental issues remain open concerning its convergence.

In this paper, we present a systematic study on both the local and the global convergence properties of the Lloyd algorithm. A number of new global convergence theorems are rigorously proved, including the global convergence of subsequences for any density functions, the global convergence of the whole sequence in one-dimensional space, and the global convergence under some nondegeneracy conditions. We also present some theoretical studies on the local convergence properties of the Lloyd algorithm including estimates on the convergence rates. Some numerical results are also presented to substantiate our theoretical investigation. Many of the techniques employed in this paper, in fact, work for more general settings. As an illustration, we analyze the application of the Lloyd algorithm to the construction of the constrained CVTs on a manifold and present some similar convergence theorems.

The rest of the paper is organized as follows. We present our main convergence theorems and some detailed discussions in section 2, followed by the extensions to more general settings that are considered in section 3 and numerical results that are given in section 4. Conclusions are drawn in section 5.

2. Convergence. Since Lloyd's pioneering work, many studies have been made on the convergence of the iteration [21, 24, 32, 36]. For example, the local convergence has been proved for strictly *logarithmically concave* density functions in the one-dimensional space [32]. An extension to CVTs defined on a circle is given in [12]. The convergence analysis in multidimensional space for general density functions is far from complete. There are very few known conditions that guarantee the global convergence. We now present some new results that have not been previously explored in the literature.

For clarity, since a Voronoi tessellation is defined using a point set with k points $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^k$ as the respective generators, let us redefine the *energy functional*, or the *distortion value*, as a functional for a pair $(\mathbf{Y}, \mathbf{Z})$ with $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k) \in \mathbb{R}^{kN}$:

$$\mathcal{H}\Big(\mathbf{Y}, \mathbf{Z}\Big) = \sum_{i=1}^{k} \int_{V_i(\mathbf{Y})} \rho(\mathbf{y}) |\mathbf{y} - \mathbf{z}_i|^2 \, d\mathbf{y} \,,$$

where $\{V_i(\mathbf{Y})\}_{i=1}^k$ are the Voronoi regions with respect to $\{\mathbf{y}_i\}_{i=1}^k$. The Lloyd algorithm may be viewed as a fixed point iteration of the so-called Lloyd map [7], a mapping from a set of distinct generators $\{\mathbf{z}_i\}_{i=1}^k \subset \Omega \subset \mathbb{R}^N$ to the corresponding mass centers, defined by $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_k)^T \colon \mathbb{R}^{kN} \to \mathbb{R}^{kN}$ with

$$\mathbf{T}_i(\mathbf{Z}) = \frac{\displaystyle\int_{V_i(\mathbf{Z})} \mathbf{y} \rho(\mathbf{y}) \, d\mathbf{y}}{\displaystyle\int_{V_i(\mathbf{Z})} \rho(\mathbf{y}) \, d\mathbf{y}} \,.$$

A set of generators of a centroidal Voronoi tessellation is obviously a fixed point of $\mathbf{T}$. Moreover, the Lloyd algorithm is equivalent to a fixed point iteration of $\mathbf{T}$:

$$\mathbf{Z}_n = \mathbf{T}(\mathbf{Z}_{n-1}) \qquad \text{for } n \geq 1 \,.$$

Notice that in general, the map $\mathbf{T}$ can be defined only on an open subset of $\Omega^k \subset \mathbb{R}^{kN}$ as we need to ensure that the denominators are nonzero, that is, the corresponding Voronoi regions are nonempty. This, in particular, implies that the

generating points must be distinct. With this being noted, one needs to be cautious in applying general optimization theory concerning the convergence of energy descent algorithms [37] as such abstract theory often requires the compactness of the domain and the closedness of the associated map.

We now first quote some elementary facts for which one may find more detailed discussions in [7] and [41].

LEMMA 2.1. *Let ρ be a positive and smooth density function defined on a smooth bounded domain Ω. Then*

(1) *$\mathcal{H}$ is continuous and differentiable in $\bar{\Omega}^k \times \bar{\Omega}^k$;*

(2) $\mathcal{H}(\mathbf{Z}, \mathbf{T}(\mathbf{Z})) = \min_{\mathbf{Y} \in \bar{\Omega}^k} \mathcal{H}\Big(\mathbf{Z}, \mathbf{Y}\Big)$;

(3) $\mathcal{H}(\mathbf{Z}, \mathbf{Z}) = \min_{\mathbf{Y} \in \bar{\Omega}^k} \mathcal{H}(\mathbf{Y}, \mathbf{Z})$.

Next, we restate the strong connections between the map $\mathbf{T}$, the CVTs, and the Lloyd algorithm that we alluded to earlier.

LEMMA 2.2. *Let $\{\mathbf{Z}_n\}_1^\infty$ be the sequence of generating sets produced by the Lloyd algorithm. Then*

(1) $\mathbf{Z}_n = \mathbf{T}(\mathbf{Z}_{n-1})$;

(2) $\mathcal{H}(\mathbf{Z}_n, \mathbf{Z}_n) \leq \mathcal{H}(\mathbf{Z}_{n-1}, \mathbf{Z}_{n-1})$.

The first conclusion of the above lemma is obvious while the second one follows from properties (2) and (3) of Lemma 2.1 (for more details, see [7]). The results of Lemma 2.2 imply that the distortion (energy) values decrease when they are evaluated at consecutive iterations of the Lloyd algorithm; thus, the energy functional may be viewed as a descent function of the map $\mathbf{T}$, a fact that has been explored in [42], though the notion of a closed algorithm does not readily apply here due to the possible degeneracy of the Lloyd map $\mathbf{T}$ when some of the generating points either coincide or become arbitrarily close.

It is perhaps also interesting to note that the Lloyd algorithm may be viewed as an alternating variable algorithm for minimizing the energy functional, that is, in which one alternates between minimizing $\mathcal{H}(\mathbf{Y}, \mathbf{Z})$ with respect to $\mathbf{Y}$ and $\mathbf{Z}$. It is well known that there are examples of simple optimization problems with special objective functions for which such an alternating variable algorithm does not always converge. It is thus interesting to see whether the special features of the functional $\mathcal{H}$ can help us to establish the convergence of the Lloyd algorithm.

2.1. Existence of convergent subsequence. We now present some new convergence theorems concerning the Lloyd algorithm. It has been shown in [7] that if the density function is positive, except on a measure zero set, stationary points of the energy $\mathcal{H}$ are given by fixed points of the Lloyd map $\mathbf{T}$. The result below justifies that fixed points are attainable as a limit of Lloyd iterations.

THEOREM 2.3. *Any limit point $\mathbf{Z}$ of the Lloyd algorithm is a fixed point of the Lloyd map, and thus, $(\mathbf{Z}, \mathbf{Z})$ is a critical point of $\mathcal{H}$. Moreover, for an iteration started with a given initial guess, all elements in the set of its limit points share the same distortion value.*

Proof. The Lloyd algorithm produces a sequence $\{\mathbf{Z}_n\}$, which is bounded in $\bar{\Omega}^k$, and thus it has a convergent subsequence. Let $\mathbf{Z}$ be a limit point; then there exists a subsequence $\{\mathbf{Z}_{n_j}\}$ such that $\mathbf{Z}_{n_j} \to \mathbf{Z}$ as $n_j \to \infty$. Since the distortion values are monotonically decreasing, it follows that all limiting points must share the same distortion value.

Now, by properties of the iteration, $\mathcal{H}(\mathbf{Z}_n, \mathbf{Z}_n)$ is monotonically decreasing, so

$$\mathcal{H}\Big(\mathbf{Z}, \mathbf{Z}\Big) = \lim \mathcal{H}\Big(\mathbf{Z}_{n_j}, \mathbf{Z}_{n_j}\Big) = \inf \mathcal{H}\Big(\mathbf{Z}_n, \mathbf{Z}_n\Big) \, .$$

On the other hand, we know from Lemma 2.1 that

$$\mathcal{H}_1\Big(\mathbf{U}, \mathbf{Z}_n\Big) \mid_{\mathbf{U}=\mathbf{Z}_n} = 0 \, .$$

Here we use the notation $\mathcal{H}_1$ to denote the partial derivatives with respect to all the components of the first argument (gradient with respect to the first argument $\mathbf{U}$) and $\mathcal{H}_2$ (the gradient) with respect to the second argument.

By continuity, we get

$$\mathcal{H}_1\Big(\mathbf{Z}, \mathbf{Z}\Big) = 0 \, .$$

Now, if $\mathcal{H}_2(\mathbf{Z}, \mathbf{U}) \mid_{\mathbf{U}=\mathbf{Z}} = 0$, $(\mathbf{Z}, \mathbf{Z})$ is a critical point of $\mathcal{H}$ and we are done. Otherwise, there exists some $\mathbf{Y}$ such that

$$\mathcal{H}\Big(\mathbf{Z}, \mathbf{Y}\Big) < \mathcal{H}\Big(\mathbf{Z}, \mathbf{Z}\Big) \, .$$

Thus, for small enough δ, we have for large enough n_j that

$$\begin{aligned}
\mathcal{H}\Big(\mathbf{Z}_{n_j}, \mathbf{Y}\Big) &< \mathcal{H}\Big(\mathbf{Z}, \mathbf{Y}\Big) + \delta \\
&< \mathcal{H}\Big(\mathbf{Z}, \mathbf{Z}\Big) \\
&\leq \mathcal{H}\Big(\mathbf{Z}_{n_j+1}, \mathbf{Z}_{n_j+1}\Big) \\
&\leq \mathcal{H}\Big(\mathbf{Z}_{n_j}, \mathbf{Z}_{n_j+1}\Big) \, .
\end{aligned}$$

This contradicts the fact that

$$\mathcal{H}\Big(\mathbf{Z}_{n_j}, \mathbf{Z}_{n_j+1}\Big) = \min_{\mathbf{Y}} \mathcal{H}\Big(\mathbf{Z}_{n_j}, \mathbf{Y}\Big) \, .$$

Thus, the theorem is proved. □

The above theorem may be simply classified as a theorem for the global convergence of subsequences of the Lloyd algorithm. It leads to a more precise characterization of the algorithm and a hint on why it rarely fails, while also motivating the global convergence theorems for the whole sequence with some additional assumptions that we are going to present next.

2.2. Global convergence. As an immediate consequence of Theorem 2.3, we easily get the following result.

COROLLARY 2.4. *If the fixed point is unique, the Lloyd algorithm converges globally.*

The uniqueness of the fixed point has been established in some special cases in the literature. We will come back to this point later in the section. The uniqueness is obviously not a necessary condition, but we may in fact derive the following convergence theorem.

THEOREM 2.5. *If the set of fixed points with any particular distortion value is finite, the Lloyd algorithm converges globally.*

Proof. Convergence may fail only if the generated sequence possesses infinitely many jumps from a neighborhood of one fixed point to another. Suppose $\mathbf{U}$ and $\mathbf{V}$ are two fixed points with $||\mathbf{U} - \mathbf{V}|| = \delta > 0$. Denote the generated sequence of the Lloyd algorithm as $\mathbf{Z}_n$, i.e., $\mathbf{Z}_{n+1} = \mathbf{T}(\mathbf{Z}_n)$.

Suppose $\mathbf{Z}_{n_r} \to \mathbf{U}$ and $\mathbf{Z}_{n_l} \to \mathbf{V}$. Then for any $\delta > 0$, there exists $M > 0$ such that for all $n_r, n_l > M$ we have $||\mathbf{Z}_{n_r} - \mathbf{U}|| < \delta/3$ and $||\mathbf{Z}_{n_l} - \mathbf{V}|| < \delta/3$. The Lloyd map is continuous near the fixed points (see Proposition 3.5 in [7]), so M can be chosen to be suitably large to assure

$$||\mathbf{T}(\mathbf{Z}_{n_r}) - \mathbf{Z}_{n_r}|| < \delta/3.$$

Now suppose the sequence makes infinitely many jumps from subsequence $\{n_r\}$ to $\{n_l\}$; i.e., there are infinitely many μ, ν s.t. $n_{l_\mu} = n_{r_\nu} + 1$. Then $||\mathbf{T}(\mathbf{Z}_{n_{r_\nu}}) - \mathbf{V}|| = ||\mathbf{Z}_{n_{r_\nu}+1} - \mathbf{V}|| = ||\mathbf{Z}_{n_{l_\mu}} - \mathbf{V}||$. Hence

$$\delta = ||\mathbf{U} - \mathbf{V}|| \le ||\mathbf{U} - \mathbf{Z}_{n_{r_\nu}}|| + ||\mathbf{Z}_{n_{r_\nu}} - \mathbf{T}(\mathbf{Z}_{n_{r_\nu}})|| + ||\mathbf{T}(\mathbf{Z}_{n_{r_\nu}}) - \mathbf{V}|| < \delta.$$

We get a contradiction. □

To this end, we have proved the global convergence of the Lloyd method in case the set of fixed points, Γ, does not have an accumulation point. Note that there are situations where Γ contains accumulation points and all points in Γ share the same distortion value. For example, consider the CVTs formed with two generators in a unit disc centered at the origin for the constant density function. Simple calculation shows that the critical points fill a circle of radius $4/(3\pi)$. That is, due to the rotation symmetry, any pair of points in the opposite ends of such a circle determines a CVT, and all the critical points share the same energy values. Of course, cases like this are very rare, so this fact does not present any difficulties for the convergence of the Lloyd algorithm in most practical applications.

We now present another result which further substantiates the global convergence of Lloyd algorithm in general.

THEOREM 2.6. *If the iterations in the Lloyd algorithm stay in a compact set, where the Lloyd map* $\mathbf{T}$ *is continuous, then the algorithm is globally convergent to a critical point of* $\mathcal{H}$.

Proof. The proposition follows from the global convergence theorem (GCT), [37] and similar arguments have been presented in [42]. Indeed, the Lloyd algorithm can be regarded as a descent method with the descent function given by $\mathcal{H}(\cdot, \mathbf{T}(\cdot))$. Let $\{\mathbf{Z}_n\}_{n=1}^{\infty}$ be a sequence generated by $\mathbf{Z}_{n+1} = \mathbf{T}(\mathbf{Z}_n)$. All $\mathbf{Z}_n$'s are contained in a compact set. If Γ is the set of solutions, $\mathcal{H}(\mathbf{Y}, \mathbf{T}(\mathbf{Y})) < \mathcal{H}(\mathbf{Z}, \mathbf{T}(\mathbf{Z}))$ for all $\mathbf{Z} \notin \Gamma$, $\mathbf{Y} \in \mathbf{T}(\mathbf{Z})$ and $\mathcal{H}(\mathbf{Y}, \mathbf{T}(\mathbf{Y})) = \mathcal{H}(\mathbf{Z}, \mathbf{T}(\mathbf{Z}))$ for all $\mathbf{Z} \in \Gamma$, $\mathbf{Y} \in \mathbf{T}(\mathbf{Z})$. The continuity implies the closedness of $\mathbf{T}$ in a compact set. Applying the GCT, we get the convergence of the sequence $\mathbf{Z}_n$, and the limit $\mathbf{Z}$ is a fixed point of $\mathbf{T}$; thus, the algorithm converges to a critical point of $\mathcal{H}$. □

We note that the compactness of the iteration seems to be intuitively true but it has not been rigorously justified in the literature. The difficulty is related to showing that during the iteration, the generators of the Voronoi regions do not get arbitrarily close as the Lloyd map is not well defined at degenerating points, where some of the generators may coincide.

2.3. The compactness in the one-dimensional case. Here, we take $\Omega = [a, b]$, a compact interval, let ρ be smooth and positive, and assume that $0 < M_1 \le ||\rho||_{\infty,\Omega} \le M_2 < \infty$. Let $M_c = M_2/M_1$; obviously, $M_c \ge 1$. We verify that throughout the Lloyd algorithm, the Voronoi regions remain nondegenerate (i.e., the generating points remain distinct); thus, it will lead to the global convergence.

First, we have the following simple fact.

LEMMA 2.7. *Given an interval $V = [z_l, z_r] \in \Omega$, let z^* be the mass centroid of V with respect to the density function ρ. Then we have*

$$L(V) \le 2M_c \min(z^* - z_l,\ z_r - z^*), \tag{2.1}$$

where $L(V)$ denotes the length of V.

Proof. Without loss of generality, we suppose that $z^* - z_l \le z_r - z^*$. By the definition of mass centroid, we have

$$z^* - z_l = \frac{\displaystyle\int_{z_l}^{z_r} (x - z_l)\rho(x)\, dx}{\displaystyle\int_{z_l}^{z_r} \rho(x)\, dx} \ge \frac{M_1}{2M_2}(z_r - z_l),$$

so we get

$$z_r - z_l \le 2M_c(z^* - z_l).$$

With $z^* - z_l \le z_r - z^*$, we get the inequality (2.1). ∎

Denote by $\{z_i^{(n)}\}_{i=1}^k$ ($z_1^{(0)} < z_2^{(0)} < \cdots < z_k^{(0)}$, $n \ge 0$) the positions of the generators after n iterations in the Lloyd method and by $\{V_i^{(n)} = (y_{i-1}^{(n)}, y_i^{(n)})\}_{i=1}^k$ the corresponding Voronoi regions. Clearly, $y_0^{(n)} = a$ and $y_k^{(n)} = b$. We now present a nondegeneracy result.

LEMMA 2.8. *For any $1 < i < k$, we have*

$$L(V_i^{(n+1)}) < \min\left(\frac{L(V_i^{(n)}) + L(V_{i+1}^{(n)})}{2} + L(V_{i-1}^{(n+1)}),\ \frac{L(V_i^{(n)}) + L(V_{i-1}^{(n)})}{2} + L(V_{i+1}^{(n+1)})\right).$$

Proof. First we have

$$L(V_i^{(n+1)}) = \frac{z_{i+1}^{(n+1)} - z_i^{(n+1)}}{2} + \frac{z_i^{(n+1)} - z_{i-1}^{(n+1)}}{2}.$$

Since $z_i^{(n+1)} \in V_i^{(n)}$, $z_{i+1}^{(n+1)} \in V_{i+1}^{(n)}$, we know

$$\frac{z_{i+1}^{(n+1)} - z_i^{(n+1)}}{2} < \frac{L(V_i^{(n)}) + L(V_{i+1}^{(n)})}{2}.$$

With $L(V_{i-1}^{(n+1)}) > (z_i^{(n+1)} - z_{i-1}^{(n+1)})/2$, we get

$$L(V_i^{(n+1)}) < \frac{L(V_i^{(n)}) + L(V_{i+1}^{(n)})}{2} + L(V_{i-1}^{(n+1)}). \tag{2.2}$$

Similarly, we can prove that

$$L(V_i^{(n+1)}) < \frac{L(V_i^{(n)}) + L(V_{i-1}^{(n)})}{2} + L(V_{i+1}^{(n+1)}). \tag{2.3}$$

Combining (2.2) and (2.3), we complete the proof. ∎

This leads to the following uniform lower bound between the adjacent generators throughout the Lloyd algorithm.

PROPOSITION 2.9. *Let* $d_i^{(n)} = z_{i+1}^{(n)} - z_i^{(n)}$ *for* $i = 1, 2, \ldots, k-1$. *Then we have*

$$d_i^{(n)} > \frac{b-a}{k4^{2k-1}M_c^k}, \qquad n > k\,, \tag{2.4}$$

and consequently,

$$L(V_i^{(n)}) > \frac{b-a}{k4^{2k-1}M_c^k}, \quad 1 < i < k, \quad n > k\,, \tag{2.5}$$

and

$$L(V_i^{(n)}) > \frac{b-a}{2k4^{2k-1}M_c^k}, \quad i = 1 \text{ or } k, \quad n > k\,. \tag{2.6}$$

Proof. Let us consider any $d_i^{(n)}$ for $1 \le i \le k-1$ and $n > k$. Since $d_i^{(n)} = z_{i+1}^{(n)} - z_i^{(n)}$ and $y_i^{(n-1)} < z_{i+1}^{(n)}$, we have

$$y_i^{(n-1)} - z_i^{(n)} < d_i^{(n)}\,.$$

Then from Lemma 2.7, we have

$$L(V_i^{(n-1)}) < 2M_c d_i^{(n)}\,. \tag{2.7}$$

On the other hand, we know that $L(V_i^{(n-1)}) > (z_{i+1}^{(n-1)} - z_i^{(n-1)})/2$, which means

$$d_i^{(n-1)} < 2L(V_i^{(n-1)}) < 4M_c d_i^{(n)}\,.$$

Again by Lemma 2.7, we know that

$$L(V_{i-1}^{(n-2)}) < 8M_c^2 d_i^{(n)}\,.$$

Repeating this process, we have for $j = 1, \ldots, i$,

$$L(V_{i-j+1}^{(n-j)}) < 2^{2j-1}M_c^j d_i^{(n)}\,.$$

Now let us consider $j = i$. Clearly, $V_1^{(n-i)} = (a, y_1^{(n-i)})$, and we have

$$\begin{aligned} L(V_1^{(n-i+1)}) &< L(V_1^{(n-i)}) + L(V_2^{(n-i+1)}) \\ &< 2^{2i-1}M_c^i d_i^{(n)} + 2^{2i-3}M_c^{i-1} d_i^{(n)} \\ &< 4^i M_c^i d_i^{(n)}\,. \end{aligned}$$

Furthermore, by Lemma 2.8, we get

$$\begin{aligned} L(V_2^{(n-i+2)}) &< \frac{L(V_2^{(n-i+1)}) + L(V_1^{(n-i+1)})}{2} + L(V_3^{(n-i+2)}) \\ &< \frac{2^{2i-3}M_c^{i-1}d_i^{(n)} + 4^i M_c^i d_i^{(n)}}{2} + 2^{2i-5}M_c^{i-2}d_i^{(n)} \\ &< 4^i M_c^i d_i^{(n)}\,. \end{aligned}$$

Repeating this process, we have for $j = 1, \ldots, i-1$,

$$L(V_j^{(n-i+j)}) < 4^i M_c^i d_i^{(n)},$$

which means

$$L(V_{i-1}^{(n-1)}) < 4^i M_c^i d_i^{(n)} .$$

Using the same trick again and again, we finally arrive at

$$L(V_{i-j}^{(n-1)}) < 4^{i+j-1} M_c^i d_i^{(n)}, \qquad j = 1, \ldots, i-1.$$

Combining (2.7) and the above equation with $i, j \le k$, we get

$$L(V_j^{(n-1)}) < 4^{2k-1} M_c^k d_i^{(n)}, \qquad j = 1, \ldots, i\,. \tag{2.8}$$

By symmetry, we also have

$$L(V_j^{(n-1)}) < 4^{2k-1} M_c^k d_i^{(n)}, \qquad j = i+1, \ldots, k\,.$$

Then, we get

$$b - a = L(\Omega) = \sum_{j=1}^{k} L(V_j^{(n-1)}) < k 4^{2k-1} M_c^k d_i^{(n)},$$

which implies (2.4), (2.5), and (2.6). ∎

We then have the following theorem.

THEOREM 2.10. *For any positive and smooth density function in one dimension and a given set of k distinct generators as a starting point, the Lloyd map is continuous at any of the iteration points.*

Proof. In order to show the continuity it is enough to justify the fact that Voronoi cells do not collapse. Indeed, after a sufficient number of steps, the latter is the direct consequence of Proposition 2.9. For the initial finite number of iterations, the continuity is obvious. ∎

Finally, using Theorems 2.6 and 2.10, we get Theorem 2.11.

THEOREM 2.11. *The Lloyd algorithm is globally convergent in one dimension for any positive and smooth density function.*

Proof. Using the result of Theorem 2.10, we see that we can define a compact set (away from the degenerating points) such that for any initial condition, the Lloyd iteration (the images of the Lloyd maps) will stay in such a compact set after sufficiently many steps. Thus, we may apply Theorem 2.6 to deduce the convergence of the algorithm. ∎

The above theorem provides an affirmative answer to the question of global convergence of the Lloyd algorithm for the one-dimensional interval case without any restrictive assumptions on the density functions. It remains an open problem to verify the same conclusion in the multidimensional case.

2.4. The logarithmic concave density for the one-dimensional case. Beyond the study on the global convergence, the characterization of the convergence rate is often also important in practice. For instance, one may inquire if a geometric convergence rate can be established. This is indeed verified in [7] for the constant

density function corresponding to the unit interval $[0, 1]$, where, via the spectral analysis of $\mathbf{dT}$ at the minimizer, the established geometric convergence rate r is shown to satisfy

$$\sin^2\left(\frac{\pi}{2(k+1)}\right) \le r \le \sin^2\left(\frac{\pi}{2(k-1)}\right), \tag{2.9}$$

so that asymptotically for large k (the total number of generators) the convergence rate is on the order of $1 - \pi^2/(4k^2)$, as verified by the numerical experiments in the next section.

In general, finding the convergence rate exactly is not possible, but estimates may be obtained from the analytical bounds of the $\|\mathbf{dT}\|$.

First, it follows from Theorem 2.10 that $\mathbf{T} : \Omega^k \to \Omega^k$ is a continuously differentiable mapping away from the degenerate points, where the generating points collapse. If this mapping $\mathbf{T}$ is a contraction, i.e., $||\mathbf{dT}|| < 1$ at all nondegenerate points, the contraction mapping theorem can be used to get a good estimate of the local convergence rate for the corresponding fixed point iteration, which in our case is the Lloyd algorithm. Moreover, the contraction mapping properties also imply that $\mathbf{T}$ has a unique fixed point $\mathbf{z}^*$ in the set of nondegenerate points upon a consistent ordering. Indeed, if there existed two fixed points $\mathbf{x} = \{x_i\}_{i=1}^k$ and $\mathbf{y} = \{y_i\}_{i=1}^k$, with components corresponding to generating points whose coordinates are ordered from small to large, that is, $x_i < x_{i+1}$ and $y_i < y_{i+1}$ for all indices i, then any point along the line segment $(1-t)\mathbf{x} + t\mathbf{y}$ would remain nondegenerate and thus, by uniform continuity, we may assume that

$$\sup_{0\le t\le 1} ||\mathbf{dT}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))|| \le \alpha(\mathbf{x}, \mathbf{y}) < 1$$

for some constant $\alpha(\mathbf{x}, \mathbf{y})$ independent of t. From the multidimensional form of the mean value theorem, we then get

$$||\mathbf{x} - \mathbf{y}|| = ||\mathbf{Tx} - \mathbf{Ty}|| \le \sup_{0\le t\le 1} ||\mathbf{dT}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))||\, ||\mathbf{x} - \mathbf{y}|| \le \alpha ||\mathbf{x} - \mathbf{y}||,$$

which is possible only if $\mathbf{x} = \mathbf{y}$; thus, we have the uniqueness. We refer to [32] for similar discussions.

The concept of logarithmic concavity has played an important role in the classification of one-dimensional density functions since it is a class of density functions for which the Lloyd maps can be shown to be contractions [7].

Let us take a closer look at the structure of the Jacobian $\mathbf{dT}$. By the notation of the previous section, for the one-dimensional case (i.e., $\Omega = [a, b]$), we have

$$\frac{\partial T_i}{\partial z_i} = \frac{\partial T_i}{\partial z_{i-1}} + \frac{\partial T_i}{\partial z_{i+1}},$$

$$\frac{\partial T_i}{\partial z_{i-1}} = \frac{\rho(z_i^-)(T_i - z_i^-)}{2R_i}, \qquad \text{and} \qquad \frac{\partial T_i}{\partial z_{i+1}} = \frac{\rho(z_i^+)(z_i^+ - T_i)}{2R_i}, \tag{2.10}$$

where $R_i = \int_{V_i} \rho(y)dy$ and $V_i = [z_i^-, z_i^+]$.

The following useful relation may be found in [7, 24]:

$$R_i^2\left(1 - \sum_j \frac{\partial T_i}{\partial z_j}\right) = \frac{1}{2}\int_{V_i}\int_{V_i} \rho(t)\rho(s)\left(\frac{\rho'(s)}{\rho(s)} - \frac{\rho'(t)}{\rho(t)}\right)(t-s)dt\, ds \tag{2.11}$$

at a fixed point $\mathbf{z} = \mathbf{T}(\mathbf{z})$.

Based on this, it can be shown that for the class of logarithmically concave functions (i.e., $(\log\rho)'' < 0$), the spectral radius of the Jacobi map is less than 1 in the neighborhood of a fixed point. In fact, it is easy to show that the same estimate holds for all points as the identity (2.11) remains universally true. Hence the fixed point of the Lloyd map is unique when the generators are ordered in an increasing manner. The following convergence of the Lloyd algorithm for the logarithmically concave case is easily one of the most popular results studied in the literature.

PROPOSITION 2.12. *In one dimension, in case of logarithmically concave density, the Lloyd algorithm converges globally to the unique fixed point.*

The class of logarithmically concave functions covers many densities used in practice, for instance, linear densities and normal distributions. Notice that the result quoted in Proposition 2.12 does not provide the estimate of the actual distance of the spectral radius from 1. We now focus on getting estimates on $\theta = 1 - ||\mathbf{dT}||$ more accurately. For this, we use a more precise measure of the logarithmic concavity for the density, that is, we assume that

$$\rho(t)\rho(s)\left(\frac{\rho'(s)}{\rho(s)} - \frac{\rho'(t)}{\rho(t)}\right)(t-s) \geq c_0^2(t-s)^2 \tag{2.12}$$

for some constant $c_0 > 0$ and any (t, s) except for a set of measure zero. Upon availability of an estimate of this type, the following conclusion can be reached:

$$1 - ||\mathbf{dT}|| \geq c_0^2 \min_i \left\{ R_i^{-2} \int_{V_i}\int_{V_i} (t-s)^2 dtds \right\} \sim \frac{c_0^2}{12} \min\left\{ \frac{h_i^2}{\rho(\zeta_i)^2} \right\}$$

for some $\zeta_i \in V_i$ and $h_i = z_i^+ - z_i^-$. Let $h = \min_i h_i$, the smallest Voronoi cell size, and $M = \sup_{x\in[0,1]} \rho(x)$; then we can rewrite the above result as follows.

LEMMA 2.13. *For any smooth density ρ satisfying (2.12) on the unit interval, the Lloyd algorithm is globally convergent with a geometric convergence rate no larger than*

$$||\mathbf{dT}|| \leq 1 - \frac{c_0^2}{12}\frac{h^2}{M^2}\ . \tag{2.13}$$

The convergence estimate obtained here essentially depends on characteristics c_0 and the relative size of a Voronoi cell in comparison with the density distribution. Since the minimizer of the energy gives a nondegenerate Voronoi diagram (Proposition 3.5 in [7]), there is a positive lower bound for the distance h in the neighborhood of the solution in terms of the density and the number of generators. Moreover, for large k, due to the asymptotic equipartition of energy property in one dimension [7], after sufficiently many iterations, one can roughly estimate each cell size as

$$h_i \sim k^{-1}\rho(\zeta_i)^{-1/3}\int_0^1 \rho^{1/3}(x)dx\ .$$

Thus, we have effectively $\theta = 1 - ||\mathbf{dT}|| \geq \left(\frac{c_1}{k}\right)^2$, where for large k,

$$c_1 \sim \frac{c_0}{\sqrt{12}M^{4/3}}\int_0^1 \rho^{1/3}(x)dx\ . \tag{2.14}$$

The estimate (2.14) in general tends to be rather pessimistic; for instance, for a linear perturbation of the constant density $\rho(x) = 1 - \epsilon x$ for a small ϵ, we have $c_1 \sim \frac{3}{4\sqrt{12}}(1-(1-\epsilon)^{4/3})$, which is significantly different from $\pi/2$ in the limit as $\epsilon \to 0$ (for the constant density case, c_1 can be estimated more accurately from the estimate (2.9) as $\pi/2$). This is due to the fact that the class of constant densities shares zero value of the parameter c_0. Nevertheless, it allows us to reach the conclusion that the geometric convergence rate for all densities satisfying (2.12) is comparable with that of the constant density in the sense that θ remains of the order k^{-2} for large values of k.

We expect that such a conclusion holds for even more general density functions, but the rigorous analysis is still not available.

3. Extensions to constrained CVTs. We now briefly illustrate how much of our earlier analysis can be extended to more general settings, where the concept of CVTs can be defined. The example to be used is of constrained CVTs on general surfaces as defined in [12].

Consider a compact and smooth surface $\mathbf{S} \subset \mathbb{R}^N$. Similar to the definition of conventional CVTs, for a given set of points $\{\mathbf{z}_i\}_{i=1}^k \in \mathbf{S}$, one may define their corresponding Voronoi regions on $\mathbf{S}$ by

$$V_i = \{\, \mathbf{x} \in \mathbf{S} \;:\; |\mathbf{x}-\mathbf{z}_i| < |\mathbf{x}-\mathbf{z}_j| \quad \text{for } j = 1,\dots,k,\ j \neq i \,\}. \tag{3.1}$$

For a density function ρ defined on the surface $\mathbf{S}$ and positive almost everywhere, one may encounter a problem with the original definition when one defines centroidal Voronoi tessellations $\{(\mathbf{z}_i, V_i)\}_{i=1}^k$ of $\mathbf{S}$: the mass centroids $\{\mathbf{z}_i^*\}_{i=1}^k$ of $\{V_i\}_{i=1}^k$ as defined by (1.1) do not in general belong to $\mathbf{S}$. For example, the mass centroid of any region on the surface of a sphere is always located in the interior of the sphere. Therefore, a generalized definition of a mass centroid on surfaces is needed. For each Voronoi region $V_i \subset \mathbf{S}$, we call $\mathbf{z}_i^c$ the *constrained mass centroid* of V_i on $\mathbf{S}$ if $\mathbf{z}_i^c$ is a solution of the following problem:

$$\min_{\mathbf{z}\in\mathbf{S}} F_i(\mathbf{z})\,, \qquad \text{where} \qquad F_i(\mathbf{z}) = \int_{V_i} \rho(\mathbf{x})|\mathbf{x}-\mathbf{z}|^2\,d\mathbf{x}\,. \tag{3.2}$$

The integral over $\{V_i\}$ is understood as a standard surface integration on $\mathbf{S}$. Note that the constrained mass centroid coincides with the conventional mass center if $\mathbf{S}$ is replaced by $\mathbb{R}^N$ and V_i is a convex subset of $\mathbb{R}^N$. Clearly, for each $i = 1,\dots,k$, $F_i(\cdot)$ is convex. Since $\mathbf{S}$ is compact and $\rho(\cdot)$ is continuous almost everywhere, there exists a constant C such that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbf{S}$, we have

$$|F_i(\mathbf{z}_1) - F_i(\mathbf{z}_2)| = \left| \int_{V_i} \rho(\mathbf{x})(|\mathbf{x}-\mathbf{z}_1|^2 - |\mathbf{x}-\mathbf{z}_2|^2)\,d\mathbf{x} \right| \le C|\mathbf{z}_1 - \mathbf{z}_2|\,.$$

Thus, F_i is continuous and compact, and consequently we have the existence of solutions of (3.2), although the solution may not be unique.

We call the tessellation defined by (3.1) a *constrained centroidal Voronoi tessellation* (CCVT) if and only if the points $\{\mathbf{z}_i\}_{i=1}^k$ which serve as the generators associated with the Voronoi regions $\{V_i\}_{i=1}^k$ are the constrained mass centroids of those regions [12]. This definition of CCVT conforms with that of CVT for general spaces and clearly the energy $\mathcal{H}$ defined in (3.2) for CVTs is still valid for CCVTs. In Figure 1, we give two examples of CCVTs, one with six generators constrained to a circle

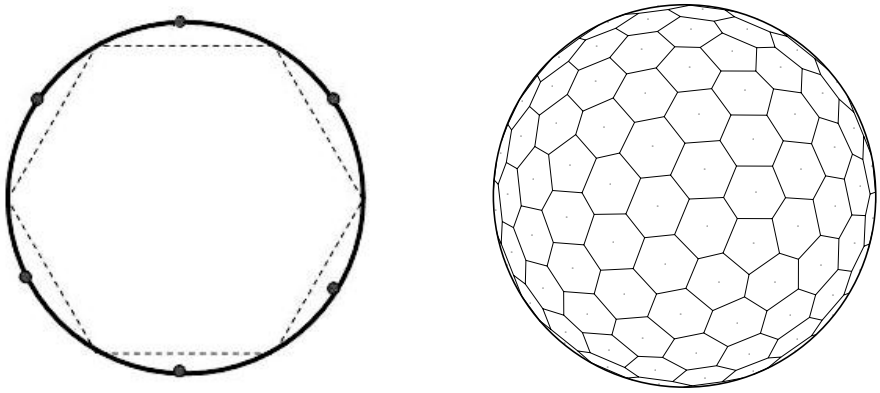


FIG. 1. *Examples of CCVTs for a circle (dots are for generators and dashes show the partition of the constrained Voronoi regions) and for a sphere (dots are generators and lines are planar projections of Voronoi edges). Only portion in one hemisphere is shown.*

(one-dimensional curve) and the other with 162 generators constrained to a sphere (two-dimensional surface). Both correspond to the constant density.

The following generalized Lloyd algorithm for computing CCVTs was proposed in [12].

ALGORITHM 3.1 (Lloyd algorithm for computing CCVTs).

Input:

$\mathbf{S}$, *the surface of interest;* ρ, *a density function defined on* $\mathbf{S}$*;*
k, *number of generators;* $\{\mathbf{z}_i\}_{i=1}^k$, *the initial set of generators.*

Output:

$\{V_i\}_{i=1}^k$, *a CCVT with* k *generators* $\{\mathbf{z}_i\}_{i=1}^k$ *in* $\mathbf{S}$.

Iteration:

1. *Construct the Voronoi tessellation* $\{V_i\}_{i=1}^k$ *of* $\mathbf{S}$ *with generators* $\{\mathbf{z}_i\}_{i=1}^k$.
2. *Take the constrained mass centroids of* $\{V_i\}_{i=1}^k$ *as the new set of generators* $\{\mathbf{z}_i\}_{i=1}^k$.
3. *Repeat the procedures* 1 *and* 2 *until some stopping criterion is met.*

It is clear that Algorithm 3.1 is almost identical to Algorithm 1.1 except the constrained mass centroids are used instead of standard mass centroids in step 2 of each iteration. So Algorithm 3.1 again can be regarded as a fixed point iteration of $\mathbf{T}$, the Lloyd map for CCVTs which now is defined to map the current generators to the constrained mass centroids of the corresponding Voronoi regions. It is transparent that the analysis done in sections 2.1 and 2.2 can be applied here, so we obtain the following general results similar to Theorems 2.3 and 2.5.

THEOREM 3.1. *Any limit point* $\mathbf{Z}$ *of the Lloyd algorithm for computing CCVTs is a fixed point of the Lloyd map for CCVTs, and thus,* $(\mathbf{Z}, \mathbf{Z})$ *is a stationary point of* $\mathcal{H}$*. Moreover, for an iteration started with a given initial guess, all elements in the set of its limit points share the same distortion value. Furthermore, if the set of fixed points with the same distortion value is finite, the Lloyd iteration for CCVTs converges globally.*

Now suppose that $\mathbf{S}$ is a smooth curve without self-intersection such as $\mathbf{S} = f(\Omega)$, where $\Omega = [a, b]$ for some smooth function f; then using the analysis similar to that provided in section 2.3, we obtain the following result.

THEOREM 3.2. *The Lloyd algorithm for computing CCVTs of* $\mathbf{S}$ *is globally convergent for any positive and smooth density function when* $\mathbf{S}$ *is a bounded smooth curve.*

Note that, unlike the one-dimensional conventional CVT in $\mathbb{R}^1$, we have not given any general estimate here on the convergence rate of the Lloyd algorithm for CCVTs. Even for the case where $\mathbf{S}$ is a bounded smooth curve, the geometric convergence rate has not been carefully derived, though the notion of contraction for the Lloyd map has been studied for density functions which share similar logarithmic concave properties with respect to the angular variable in the case of a perfect disc [12]. There are also natural generalizations of the Lloyd algorithm to the anisotropic CVTs as defined in [16] and also [18]. The details are omitted here.

4. Numerical examples. To further substantiate some of our earlier analysis, we now present a few numerical examples. All examples given below correspond to the Lloyd iteration on the interval $[0, 1]$.

4.1. Constant density. In Figure 2, we show a log-log plot of both the numerical estimates and the analytical estimate $1 - ||\mathbf{dT}|| \sim \pi^2/(4k^2)$ with respect to the constant density for various values of k, the number of generating points. The two estimates match very well and the results verify that the analytical estimates are very sharp.

4.2. Nonconstant density. Consider the case of $\rho(x) = e^{-x^2}$. Figure 3 compares the analytical estimate with the computed norms of the Jacobian for different system sizes. Here, the analytical estimate is based on $c_1^2 k^{-2}$ with the constant c_1 estimated by (2.14) with $c_0 = \sqrt{2/e}$, $M = 1$, and $\int_0^1 \rho^{1/3}(x)dx = \sqrt{3\pi} \cdot \mathrm{Erf}(1/\sqrt{3})/2$, which leads to $c_1 = \sqrt{\pi} \cdot \mathrm{Erf}(1/\sqrt{3})/2e \sim 0.19$. The plot is again given in log-log scale, and we see that although we underestimated the exact value of c_1, the slope was equal to -2 for both estimates, which indicates good agreement of the asymptotic rates on the order of $1 - O(1/k^2)$.

Figure 4 gives a similar comparison for $\rho(x) = 1 + x^4 \cos(\pi x)$. The numerical data in this case were compared to the asymptotic rate of $1 - \pi^2/4k^2$.

Figures 5–7 provide some insight into the dependence of the actual convergence factor on the number of generators and on the density function. The convergence factor in the plot is defined as the ratio of the 2-norm defects between two consecutive iterations after sufficiently many steps. A density function of the form $\rho(x) = 1 + \epsilon \cos^2(\pi x)$ is chosen. In Figure 5, we fix the number of generators to be $k = 16$, while letting ϵ vary in the range $[10^{-10}, 10^{10}]$. It is seen that the actual convergence factor and the theoretical estimate given by $||\mathbf{dT}||$ agree well in general.

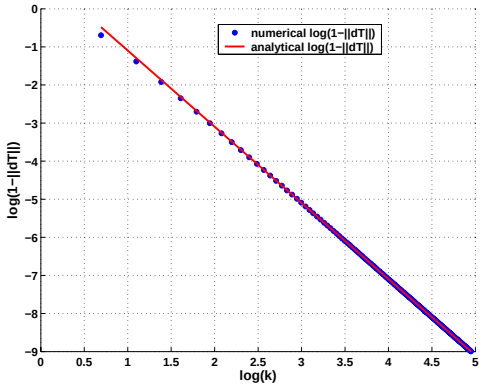


FIG. 2. *Convergence of Lloyd method for constant density.*

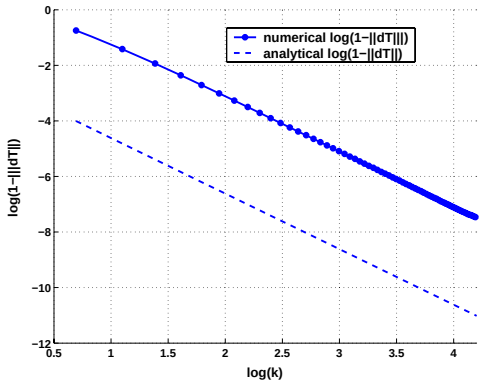


FIG. 3. *Convergence factor of Lloyd method for* $\rho(x) = e^{-x^2}$.

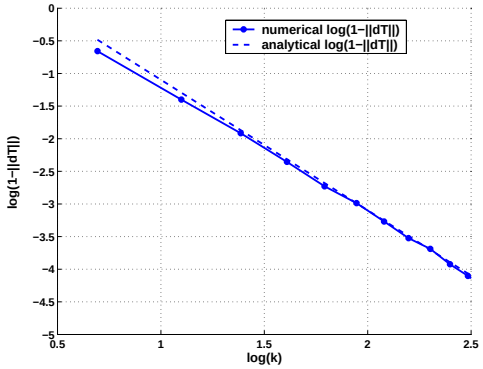


FIG. 4. *Convergence factor of Lloyd method for* $\rho(x) = 1 + x^4 \cos(\pi x)$.

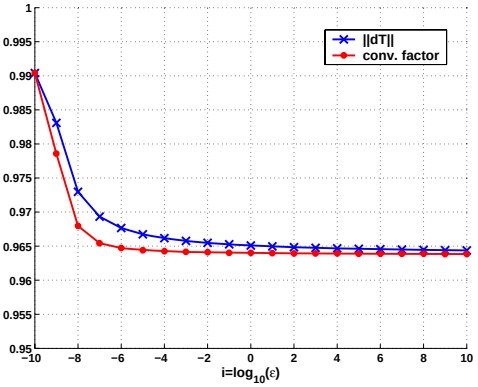


FIG. 5. *Convergence factor for* $k = 16$ *and* $\rho(x) = 1 + \epsilon \cos^2(\pi x)$ *with* $\epsilon = 10^{-10} : 10^{10}$.

To see the effect of the increasing k, in Figure 6 we fix ϵ and let the number of generators vary. The two estimates again compare well with each other.

To see more clearly the dependence of convergence rates on k, we again plot the data in a log-log scale for the density $\rho(x) = 1 + 10^3 \cos^2(\pi x)$ against the number of generators. The slope value of -2 is very evident from Figure 7, which is consistent with our earlier analysis.

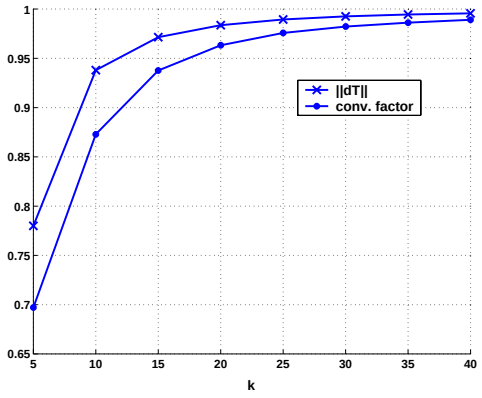


FIG. 6. *Convergence factor for* $\rho(x) = 1 + 10^3 \cos^2(\pi x)$ *and* $k = 2 : 40$.

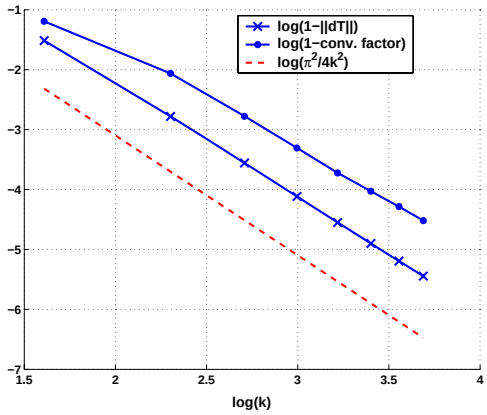


FIG. 7. *Asymptotic behavior of the convergence factor for* $\rho(x) = 1 + 10^3 \cos^2(\pi x)$.

5. Conclusions. In many practical applications of the centroidal Voronoi tessellations, it is very important to find their reliable and efficient constructions. Lloyd algorithm has been one of the most widely used techniques for such purposes. In this paper, a systematic study of both the local and the global convergence properties of the Lloyd algorithm is presented. We established several new convergence theorems, made further characterizations on the properties of the iteration, and performed relevant numerical experiments. We also extended our discussion to more general settings such as the construction of the CCVTs on a manifold. Still, one important open problem remains, that is, the global convergence of the Lloyd algorithm in any dimensions for any smooth density. The nondegeneracy of the Lloyd map should be true in this general case, but its proof has not been produced rigorously except for the one-dimensional case discussed here. We hope that our present study generates some interest along this direction, as there are certainly many issues to be considered further—in particular, the improvement of the Lloyd method for large number of generators. Even in the one-dimensional case, both our theoretical estimates and the experiments indicate the possible slow convergence rates. Recently, we have worked on making improvements in two directions: one is to explore the coupling with Newton-like methods, and another is to introduce the ideas of multilevel schemes [5, 6, 20]. As previously studied in [30], one may also consider parallel implementation issues for

these approaches. In conclusion, there are still many interesting problems associated with the construction of CVTs that can be investigated in the future.

Acknowledgements. The authors would like to thank the referees for valuable suggestions that helped us improve our presentation.

REFERENCES

[1] J. Burkardt, M. Gunzburger, and H.-C. Lee, *Centroidal Voronoi tessellation-based reduced-order modeling of complex systems*, to appear.
[2] M. Cappellari and Y. Copin, *Adaptive spatial binning of integral-field spectroscopic data using Voronoi tessellations*, Monthly Notices Roy. Astronom. Soc., 342 (2003), pp. 345–354.
[3] D. Cohen-Steiner, P. Alliez, and M. Desbrun, *Variational shape approximation*, ACM Trans. Graphics, 23 (2004), pp. 905–914.
[4] J. Cortes, S. Martinez, T. Karatas, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Tran. Robotics and Automation, 20 (2004), pp. 243–255.
[5] Q. Du and M. Emelianenko, *Uniform Convergence of a Multilevel Energy-Based Quantization Scheme*, to appear in Proceedings of the 16th International Conference on Domain Decomposition Methods, Lecture Notes Comput. Sci. Engrg., Springer.
[6] Q. Du and M. Emelianenko, *Acceleration Schemes for Computing Centroidal Voronoi Tessellations*, to appear in Proceedings of the 12th Annual Copper Mountain Conference on Multigrid Methods, Special Issue of Numer. Linear Algebra Appl.
[7] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.
[8] Q. Du and M. Gunzburger, *Grid generation and optimization based on centroidal Voronoi tessellations*, Appl. Math. Comput., 133 (2002), pp. 591–607.
[9] Q. Du and M. Gunzburger, *Centroidal Voronoi tessellation based proper orthogonal decomposition analysis*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 143, Birkhäuser, 2002, pp. 137–150.
[10] Q. Du, M. Gunzburger, and L. Ju, *Meshfree, probabilistic determination of point sets and support regions for meshless computing*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1349–1366.
[11] Q. Du, M. Gunzburger, and L. Ju, *Voronoi-based finite volume methods, optimal Voronoi meshes, and PDEs on the sphere*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 3933–3957.
[12] Q. Du, M. Gunzburger, and L. Ju, *Constrained centroidal Voronoi tessellations on general surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.
[13] Q. Du, M. Gunzburger, L. Ju, and X. Wang, *Centroidal Voronoi tessellation algorithms for image processing*, J. Math. Imaging Vision, to appear.
[14] Q. Du and L. Ju, *Finite volume methods and spherical centroidal Voronoi tessellations*, SIAM J. Numer. Anal., 43 (2005), pp. 1673–1692.
[15] Q. Du and D. Wang, *Tetrahedral mesh generation and optimization based on centroidal Voronoi tessellations*, Internat. J. Numer. Methods Engrg., 56 (2003), pp. 1355–1373.
[16] Q. Du and D. Wang, *Anisotropic centroidal Voronoi tessellations and their applications*, SIAM J. Sci. Comput., 26 (2005), pp. 737–761.
[17] Q. Du and D. Wang, *New progress in robust and quality Delaunay mesh generation*, J. Comput. Appl. Math., to appear.
[18] Q. Du and X. Wang, *Centroidal Voronoi tessellation based algorithms for vector fields visualization and segmentation*, in Proceedings of the IEEE Conference on Visualization, Austin, TX, 2004, pp. 43–50.
[19] Q. Du and T. Wong, *Numerical studies of the MacQueen's algorithm for computing the centroidal Voronoi tessellations*, Comput. Math. Appl., 44 (2002), pp. 511–523.
[20] M. Emelianenko, *Multilevel and Adaptive Methods for Some Nonlinear Optimization Problems*, Ph.D. thesis, Penn State University, University Park, PA, 2005.
[21] P. Fleischer, *Sufficient conditions for achieving minimum distortion in a quantizer* I, IEEE Int. Convention Record, 1964, pp. 104–111.
[22] S. Fortune, *Voronoi Diagrams and Delaunay Triangulations in Computing in Euclidean Geometry*, World Scientific, River Edge, NJ, 1992.
[23] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1992.

[24] R. Gray, J. Kieffer, and Y. Linde, *Locally optimal block quantizer design*, Inform. and Control, 45 (1980), pp. 178–198.
[25] R. Gray and D. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.
[26] P. Gruber, *Optimum quantization and its applications*, Adv. in Math., 186 (2004), pp. 456–497.
[27] J. Hartigan and M. Wong, *A k-means clustering algorithm*, Appl. Stat., 28 (1979), pp. 100–108.
[28] P. Heckert, *Color image quantization frame buffer display*, ACM Trans. Comput. Graph., 16 (1982), pp. 297–304.
[29] S. Hiller, H. Hellwig, and O. Deussen, *Beyond stippling—methods for distributing objects on the plane*, Computer Graphics Forum, 22 (2003), pp. 515–522.
[30] L. Ju, Q. Du, and M. Gunzburger, *Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations*, Parallel Comput., 28 (2002), pp. 1477–1500.
[31] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, *An efficient k-means clustering algorithm: Analysis and implementation*, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), pp. 881–892.
[32] J. Kieffer, *Uniqueness of locally optimal quantizer for log-concave density and convex error function*, IEEE Trans. Inform. Theory, 29 (1983), pp. 42–47.
[33] R. Klein, *Concrete and Abstract Voronoi Diagrams*, Lecture Notes in Comput. Sci. 400, Springer, Berlin, 1989.
[34] Y. Linde, A. Buzo, and R. Gray, *An algorithm for vector quantizer design*, IEEE Trans. Comm., 28 (1980), pp. 84–95.
[35] S. Lloyd, *Least square quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.
[36] F. Lu and G. Wise, *A further investigation of the Lloyd–Max algorithm for quantizer design*, in Proceedings of the 21st Annual Allerton Conference on Communication, Control, and Computing, University of Illinois, IL, 1983, pp. 481–490.
[37] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.
[38] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Statistics, L. Le Cam and J. Neyman, ed., University of California Press, Berkeley, CA, 1967, pp. 281–297.
[39] A. Mendes and I. Themido, *Multi-outlet retail site location assessment*, Internat. Trans. Oper. Res., 11 (2004), pp. 1–18.
[40] U. Moller, M. Galicki, E. Baresova, and H. Witte, *An efficient vector quantizer providing globally optimal-solutions*, IEEE Trans. Signal Process., 46 (1998), pp. 2515–2529.
[41] A. Okabe, B. Boots, and K. Sugihara, *Spatial Tessellations; Concepts and Applications of Voronoi Diagrams*, Wiley, Chichester, 1992.
[42] J. Sabin and R. Gray, *Global convergence and empirical consistency of the generalized Lloyd algorithm*, IEEE Trans. Inform. Theory, 32 (1986), pp. 148–155.
[43] A. Trushkin, *On the design of an optimal quantizer*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1180–1194.
[44] S. Valette and J. Chassery, *Approximated centroidal Voronoi diagrams for uniform polygonal mesh coarsening*, Computer Graphics Forum, 23 (2004), pp. 381–390.
[45] C. Wager, B. Coull, and N. Lange, *Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging*, J. R. Stat. Soc. B Stat. Methodol., 66 (2004), pp. 429–446.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 120–138

PROJECTION MULTILEVEL METHODS FOR QUASILINEAR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS: NUMERICAL RESULTS*

THOMAS A. MANTEUFFEL†, STEPHEN F. MCCORMICK†, OLIVER RÖHRLE‡, AND JOHN RUGE†

Abstract. The goal of this paper is to introduce a new multilevel solver for two-dimensional elliptic systems of nonlinear partial differential equations (PDEs), where the nonlinearity is of the type $u\partial v$. The incompressible Navier–Stokes equations are an important representative of this class and are the target of this study. Using a first-order system least-squares (FOSLS) approach and introducing a new variable for ∂v, for this class of PDEs we obtain a formulation in which the nonlinearity appears as a product of two different dependent variables. The result is a system that is linear within each variable but nonlinear in the cross terms. In this paper, we introduce a new multilevel method that treats the nonlinearities directly. This approach is based on a projection multilevel (PML) method [S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, SIAM, Philadelphia, 1992] applied to the FOSLS functional. The implementation of the discretization process, relaxation, coarse-grid correction, and cycling strategies is discussed, and optimal performance is established numerically. A companion paper [T. A. Manteuffel, S. F. McCormick, and O. Röhrle, *SIAM J. Numer. Anal.*, 44 (2006), pp. 139–152] establishes a two-level convergence proof for this new multilevel method.

Key words. projection method, multigrid, least squares, finite elements, quasilinear PDEs, Navier–Stokes

AMS subject classifications. 35J60, 65N12, 65N30, 65N55

DOI. 10.1137/040617698

1. Introduction.

The goal of nonlinear solution techniques is to solve the discretized nonlinear PDEs efficiently and accurately. Many popular, efficient methods for this purpose are based on multilevel strategies and all require a linearization process somewhere in the algorithm. These methods can be grouped into two broad categories, depending on when and how they apply the linearization step: global linearization such as Newton-type methods (cf. [14, 26]) and local linearization such as Brandt's FAS (full approximation scheme; cf. [6]) or Hackbusch's similar NMGM scheme (nonlinear multigrid method; cf. [17]).

Global linearization methods usually involve the solution of large linear systems of equations. Since substantial multigrid research is directed on developing robust, fast, and efficient linear solvers, there is an extensive repertoire of algorithms and knowledge to draw upon in this category of techniques. On the other hand, it is well known that the basin of attraction for efficient global linearization methods can be relatively small. Some of these problems might be handled by a full multigrid or nested iteration process that uses coarse-level processing to provide fine-level initial guesses. But problems with very small basins of attraction might need more expensive global

*Received by the editors October 26, 2004; accepted for publication (in revised form) August 16, 2005; published electronically February 8, 2006. This work was sponsored by the Department of Energy under grants DE-FC02-01ER25479 and DE-FG02-03ER25574, Lawrence Livermore National Laboratory under contract B533502, Sandia National Laboratory under contract 15268, and the National Science Foundation under VIGRE grant DMS-9810751.

http://www.siam.org/journals/sinum/44-1/61769.html

†Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309–0526 (tmanteuf@colorado.edu, stevem@colorado.edu, jruge@colorado.edu).

‡Bioengineering Institute, The University of Auckland, Private Bag 92019, Auckland 1, New Zealand (o.rohrle@auckland.ac.nz).

search methods. Local linearization methods tend to have a bigger basin of attraction, but often rely on rediscretizing each of the coarser levels separately. This might result in some loss of robustness since, for some problems with strong nonlinearities, the discretization on coarse levels might not accurately reflect the fine-level properties.

Although most research for nonlinear multilevel methods certainly focuses on these two main categories, there also exist other nonlinear solution techniques. For example, Dardyk and Yavneh [13] propose a nonlinear multigrid method that combines global and local linearization. Apparently, it is "at least as good as the more suitable of these two approaches, and often better than both" [13].

This paper introduces a new multilevel method that does not fit into either of these two categories, since we do not appeal to a linearization process anywhere in the algorithm. To achieve this direct approach, we focus on PDEs with nonlinear terms of type $u\partial v$, especially the incompressible Navier–Stokes equations, that we reformulate as a least-squares problem. Least-squares methods are based on a minimization principle for a functional constructed by taking the residual of the governing equations in some Hilbert norm. The intent is to ensure that the minimizer is the solution of the original set of equations and that the formulation is well posed. We should emphasize that our ability to avoid linearization is due to the nature of the functional that we construct. In particular, linearization is not needed in relaxation because the system to which the least-squares principle is applied is linear in its individual unknowns. Application of our solver methodology to other nonquadratic functionals is likely to require linearization in relaxation and coarse-level correction.

Least-squares methods for the Navier–Stokes equations have been addressed, for example, by Bochev and Gunzburger [4], Jiang [19], and Bochev et al. [1, 2]. In this paper, we consider a first-order system least-squares (FOSLS) method, where the functional is constructed by taking the L^2-norm of each interior first-order equation.

Instead of using FAS, Newton, or Newton-like methods to solve the resulting algebraic equations, we want to develop a new multigrid algorithm that can treat the nonlinearity directly and, thus, potentially more effectively. To this end, we consider a *projection multilevel* (*PML*) *method* cf. [24]) that solves an optimization problem by correcting a current approximation using projections onto various subspaces. In the context of FOSLS, the solution to a PDE is the minimizer of the FOSLS functional. So, naturally, we choose the minimization of the FOSLS functional as the optimization problem for our projection method. The minimization is done by corrections from certain finite element subspaces by way of the natural embedding of these spaces into the fine-grid space. The projection of the error that this defines is orthogonal with respect to the inner product associated with the functional, because it is defined as the approximation to the error from the given subspace that is best in the sense of minimizing the functional.

The algorithm developed here is new in that it is constructed with a special discretization approach that achieves optimal complexity. However, the underlying multilevel projection methodology has its origin in the so-called unigrid method introduced in [25], the indefinite-system framework presented in [16], and the Rayleigh–Ritz PML method formalized in [24]. The basic idea is to minimize the objective functional on subspaces that include the relaxation directions from all levels used in the multigrid process. As such, this methodology has been used by Mandel and McCormick for eigenvalue problems (cf. [22]); by Gelman and Mandel for constraint optimization problems (cf. [15]); by McCormick for parameter estimation, transport equations, general eigenvalue problems, Riccati equations, finite volume element methods, and image reconstruction (cf. [24]); by Tai and Xu for general convex optimization prob-

lems (cf. [30]); and by Tai for variational inequalities (cf. [29]). In fact, under certain circumstances, these methods relate to specific forms of classical multilevel methods. Consider, for example, the standard fully variational multigrid method applied to the Poisson problem in two dimensions, as given in [28], with Gauss–Seidel as the smoother, full coarsening, bilinear interpolation, and a nine-point stencil. This classical algorithm could also be classified as a PML method. It can be interpreted, at each stage, as a Rayleigh–Ritz method applied to minimizing the energy functional, where the optimization is taken as a correction over the continuous space projected onto certain subspaces of the fine-grid finite element space. This exemplifies that there exist, under certain circumstances, similarities and relations between a standard multigrid method and a PML method. In fact, PML exhibits the same basic principles as any other multilevel algorithm. Such principles include appropriate discretizations for the fine-grid problem, relaxation, coarsening, coarse-grid solves, interpolation, and cycling strategies.

The challenge in developing such a scheme is to ensure that the cost of processing coarse levels is less expensive in total than that of the fine grid. The major task in addressing this challenge is to cast the coarse subspace projection in terms of coarse-level computation. This ability we call coarse-grid realizability. We show below how this is done for our scheme applied to the Navier–Stokes equations.

To illustrate the basic ideas and principles of this new PML method, we introduce in section 2 a projection-based discretization process. Based on this process, we derive in section 3 an abstract framework for PML. In section 4, we discuss how coarse-grid realizability can be done efficiently for quasi-linear PDEs, for which the highest-order terms are linear. Additionally, we show that this is also feasible for different relaxation types and higher-order discretizations. We conclude this paper by giving numerical results (section 5) for model problems in two dimensions and making a few general remarks. While the numerical results in section 5 show optimal convergence properties, we provide in the companion paper [23] a two-level convergence proof.

2. Embedding operators and discretization by projection. As for any numerical scheme that discretely approximates continuous problems, the discretization process plays an important role. This process is even more important for multilevel schemes since they use a sequence of coarse-grid discretizations that must in some sense be compatible with the fine-grid discretization. For our particular PML method, we want to exploit a natural discretization process by using the same approach on all levels. Even though this seems to be the most natural and straightforward way to obtain discretizations for all levels, there exist other methodologies for which it is more advantageous to use a variational type of discretization process instead. Algebraic multigrid (AMG) is just one example. On coarser levels, AMG applied to a discretized PDE obtains matrices that often differ from what one would obtain using a discretization process analogous to that used on the finest level. For further details on AMG, see [7, 9, 27].

One way to relate a continuous PDE to a discrete problem is to think of the discretization process as a projection from an infinite-dimensional space onto a discrete one, with some nodal or finite element representation. (Here we restrict ourselves to a finite element representation.) To illustrate this process, consider a partial differential operator, $\mathcal{L}$, which maps between two infinite-dimensional spaces, $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{V}}_{\mathcal{L}}$ ($\mathcal{L} : \boldsymbol{\mathcal{V}} \rightarrow \boldsymbol{\mathcal{V}}_{\mathcal{L}}$). For a specific $\mathbf{g} \in \boldsymbol{\mathcal{V}}_{\mathcal{L}}$ and domain Ω, we formally obtain a PDE, which we denote by

$$\mathcal{L}(\mathbf{x}) = \mathbf{g} \qquad \text{in } \Omega. \tag{2.1}$$

For equation (2.1) to be properly defined, it may need to be taken in the weak sense, but this would complicate the discussion. We use the strong form here for simplicity. Note the use of boldface type for unknown $\mathbf{x}$ and source term $\mathbf{g}$. We do this to allow for different types of principal variables, such as pressure, temperature, and velocity. When we want to emphasize this possibility, we write these variables in component form, such as $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_v)^t$.

Now let $\mathcal{S}^h$ be a finite-dimensional subset of $\boldsymbol{\mathcal{V}}$ (e.g., a standard finite element space associated with an approximate mesh size, h). Then denote the natural embedding operator by $\mathcal{P}^h : \mathcal{S}^h \hookrightarrow \boldsymbol{\mathcal{V}}$. This operator leads to a natural discretization of our functional minimization problem as follows. Consider the least-squares functional associated with (2.1):

$$\mathcal{F}(\mathbf{x};\mathbf{g}) = \|\mathcal{L}(\mathbf{x}) - \mathbf{g}\|_{0,\Omega}^2 \qquad \forall \mathbf{x} \in \boldsymbol{\mathcal{V}}. \tag{2.2}$$

Note that $\mathcal{F}(\,\cdot\,;\mathbf{g})$ is a mapping from $\boldsymbol{\mathcal{V}}$ to $\mathbb{R}$. A discrete functional is obtained by defining $\mathcal{F}^h(\mathbf{x}^h;\mathbf{g}) := \mathcal{F}(\mathcal{P}^h\mathbf{x}^h;\mathbf{g})$, with $\mathbf{x}^h \in \mathcal{S}^h$ and $\mathcal{P}^h\mathbf{x}^h \in \boldsymbol{\mathcal{V}}$. Discretization is thus simply a matter of restricting the functional to the discrete space. This is the essence of Rayleigh–Ritz. Note that since $\mathcal{F}^h(\,\cdot\,;\mathbf{g})$ is a mapping from $\mathcal{S}^h$ to $\mathbb{R}$, notations $\mathcal{F}(\mathbf{x}^h;\mathbf{g})$ and $\mathcal{F}^h(\mathbf{x}^h;\mathbf{g})$ are equivalent. From now on, we refer to $\mathcal{F}(\mathbf{x}^h;\mathbf{g})$ as the discrete functional.

The abstract discretization process depends only on the choice of the embedding operator, $\mathcal{P}^h$, and the associated finite element space, $\mathcal{S}^h$. Hence, for coarser levels, we can define the discrete functional in the same way. Let $\mathcal{S}^{2h}$ be a finite-dimensional space (associated with an approximate mesh size, $2h$) and let $\mathcal{P}^{2h} : \mathcal{S}^{2h} \hookrightarrow \boldsymbol{\mathcal{V}}$ be the natural embedding from $\mathcal{S}^{2h}$ into $\boldsymbol{\mathcal{V}}$. Then the coarse-grid discretization of functional (2.2) is given by $\mathcal{F}(\mathbf{x}^{2h};\mathbf{g}) := \mathcal{F}(\mathcal{P}^{2h}\mathbf{x}^{2h};\mathbf{g})$, with $\mathbf{x}^{2h} \in \mathcal{S}^{2h}$. In our framework, for consecutive coarser levels, we typically choose nested spaces, so that $\mathcal{S}^{2^L h} \subset \ldots \mathcal{S}^{2h} \subset \mathcal{S}^h \subset \boldsymbol{\mathcal{V}}$. In this way, the interlevel transfer operators are induced in a natural, straightforward, and advantageous way and are easy to implement within PML. Furthermore, the coarse-grid problems are ensured to be compatible with the procedures used to define the fine-level problem, with the difference being that the coarse-level unknown is an approximation to the fine-level error and not to its solution; that is, the coarse-level correction is of the form $\mathbf{x}^h + \mathbf{c}^{2h}$. (Since $\mathbf{x}^h = \mathcal{P}^h\mathbf{x}^h$ for $\mathbf{x}^h \in \mathcal{S}^h$ and $\mathbf{c}^{2h} = \mathcal{P}^{2h}\mathbf{c}^{2h}$ for $\mathbf{c}^{2h} \in \mathcal{S}^{2h}$, we omit the embedding operators $\mathcal{P}^h$ and $\mathcal{P}^{2h}$ here and henceforth.)

Note that relaxation also depends on the choice of subspaces (and, hence, on the embeddings). We thus have to be particularly careful in picking the underlying subspaces for relaxation and coarsening.

3. Abstract framework of PML. To describe the general framework of PML applied to a functional minimization principle, let $\mathcal{F}(\mathbf{x};\mathbf{g}) : \boldsymbol{\mathcal{V}} \to \mathbb{R}$ and assume that we have a conforming finite element structure in the sense that $\mathcal{S}^{2h} \subset \mathcal{S}^h \subset \boldsymbol{\mathcal{V}}$. The aim of this section is to develop a multilevel framework that applies directly to

$$\mathcal{F}(\mathbf{x}_*^h;\mathbf{g}) = \min_{\mathbf{x}^h \in \mathcal{S}^h} \mathcal{F}(\mathbf{x}^h;\mathbf{g}), \qquad \mathbf{x}_*^h \in \mathcal{S}^h. \tag{3.1}$$

To do this, we focus on two important ingredients of multilevel methods: relaxation and coarsening. Relaxation is a generic term for an iterative process that is typically very inexpensive to use but is effective only at reducing certain "oscillatory" error components. Coarsening refers to the process of determining a coarse-level correction that, we hope, eliminates the "smooth" errors that relaxation leaves behind.

We first provide a general framework for a point or nodal relaxation scheme on the finest level. To maintain a certain form of generality in this section, let $\{\phi_n^h\}_{n=1}^{m_0}$ be a basis for $\mathcal{S}^h$, where m_0 is the dimension of $\mathcal{S}^h$. Then write $\mathcal{S}^h$ as a direct sum of the one-dimensional subspaces, $\mathcal{S}_n^h = \text{span}\{\phi_n^h\}$, $1 \le n \le m_0$: $\mathcal{S}^h = \mathcal{S}_1^h \oplus \cdots \oplus \mathcal{S}_{m_0}^h$. (Higher-dimensional subspaces can be considered for relaxation processes that update several variables at once, e.g., line or box relaxation. However, we consider only the one-dimensional case here for simplicity.)

These definitions set the stage for an abstract definition of a relaxation scheme to approximately solve for $\mathbf{x}_*^h \in \mathcal{S}^h$ in (3.1) by PML. To do so, we want to improve an initial guess, $\mathbf{x}^h$, by corrections, $\mathbf{c}^h \in \mathcal{S}_n^h$, $1 \le n \le m_0$. Thus, one sweep of relaxation consists of performing the following correction steps for each $n = 1, 2, \ldots, m_0$ in turn:

$$\text{(3.2)} \qquad \begin{cases} \mathcal{F}(\mathbf{x}^h + \mathbf{c}_{n^*}^h; \mathrm{g}) = \min\limits_{\mathbf{c}_n^h \in \mathcal{S}_n^h} \mathcal{F}(\mathbf{x}^h + \mathbf{c}_n^h; \mathrm{g}), \\ \mathbf{x}^h \leftarrow \mathbf{x}^h + \mathbf{c}_{n^*}^h. \end{cases}$$

Next, consider the coarse-grid correction process, first in terms of an exact solve, then as an iterative process. Let $\mathbf{x}^h \in \mathcal{S}^h$ be a random initial guess or a current iterate for our PML scheme. Then, the exact coarse-grid solve is described by

$$\text{(3.3)} \qquad \mathcal{F}(\mathbf{x}^h + \mathbf{c}_*^{2h}; \mathrm{g}) = \min_{\mathbf{c}^{2h} \in \mathcal{S}^{2h}} \mathcal{F}(\mathbf{x}^h + \mathbf{c}^{2h}; \mathrm{g}), \qquad \mathbf{c}_*^{2h} \in \mathcal{S}^{2h}.$$

To develop an iterative version of (3.3), we proceed in analogy to fine-grid relaxation. Let $\{\phi_n^{2h}\}_{n=1}^{m_1}$ be a basis for $\mathcal{S}^{2h}$. Then, write $\mathcal{S}^{2h}$ as a direct sum of the one-dimensional subspaces, $\mathcal{S}_n^{2h} = \text{span}\{\phi_n^{2h}\}$, $1 \le n \le m_1$: $\mathcal{S}^{2h} = \mathcal{S}_1^{2h} \oplus \cdots \oplus \mathcal{S}_{m_1}^{2h}$. Then one coarse-grid relaxation sweep consists of performing the following correction steps for each $n = 1, 2, \ldots, m_1$ in turn:

$$\text{(3.4)} \qquad \begin{cases} \mathcal{F}(\mathbf{x}^h + \mathbf{c}_{n^*}^{2h}; \mathrm{g}) = \min\limits_{\mathbf{c}_n^{2h} \in \mathcal{S}_n^{2h}} \mathcal{F}(\mathbf{x}^h + \mathbf{c}_n^{2h}; \mathrm{g}), \\ \mathbf{x}^h \leftarrow \mathbf{x}^h + \mathbf{c}_{n^*}^{2h}. \end{cases}$$

Our notation is at the crux of our ability to make PML practical. Iterative methods are commonly formulated as processes that directly update the original approximation, $\mathbf{x}^h$. Our choice of the more complicated correction form in (3.2) was made for consistency with (3.4). We complicate this notation further below by writing the respective fine- and coarse-level iterative processes as corrections to the approximate solutions, $\mathbf{c}^h$ and $\mathbf{c}^{2h}$, of (3.2) and (3.4). (To avoid further complication, we use $\mathbf{c}^h$ and $\mathbf{c}^{2h}$ to denote either the exact solutions or their approximations, a distinction that is clear by context.) Furthermore, we use these formulations to allow the multiple corrections that come from yet coarser levels. Hopefully, the three-term correction forms that we use in what follows are enough to expose the mechanisms needed to make the process efficient. The key is to write relaxation on the level $4h$ correction as a process that involves only changes to level $4h$ vectors.

To compute the corrections in (3.2) and (3.4), we use fine- and coarse-level relaxation processes. For $\mathbf{x}^h \in \mathcal{S}^h$ fixed and $\mathbf{c}^h \in \mathcal{S}^h$, the current approximation to the exact correction defined in (3.2), the nth step of a fine-grid relaxation sweep is defined by solving

$$\text{(3.5)} \qquad s = \operatorname*{argmin}_{t \in \mathbb{R}} \mathcal{F}(\mathbf{x}^h + \mathbf{c}^h + t\mathbf{d}^h; \mathrm{g}), \qquad s \in \mathbb{R},$$

and forming the update,

$$\mathbf{c}^h \leftarrow \mathbf{c}^h + s\mathbf{d}^h, \tag{3.6}$$

where $\mathbf{d}^h = \phi_n^h$, $1 \le n \le m_0$. Note that (3.5) and (3.6) describe a basic line search method, in direction $\mathbf{d}^h$, with optimal step length s. For simplicity, we combine (3.5) and (3.6) and refer to it as a *directional iteration step*. For given $\mathbf{x}^h$, $\mathbf{c}^h$, and $\mathbf{d}^h$, we denote the operator describing (3.5) and (3.6) by

$$\mathbf{c}^h \leftarrow \mathcal{D}_{\mathbf{x}^h}(\mathbf{c}^h, \mathbf{d}^h). \tag{3.7}$$

In an analogous way, coarse-grid relaxation is defined for $\mathbf{x}^h \in \mathcal{S}^h$ and $\mathbf{c}^{2h} \in \mathcal{S}^{2h}$ by $\mathbf{c}^{2h} \leftarrow \mathcal{D}_{\mathbf{x}^h}(\mathbf{c}^{2h}, \mathbf{d}^{2h})$, where $\mathbf{d}^{2h} = \phi_n^{2h}$, $1 \le n \le m_1$. Successive application of this process yields an abstract formulation of a general PML method. Assume that there are $L+1$ distinct grid levels corresponding to mesh sizes $2^l h$, $l = 0, \ldots, L$. (We label the finest level by superscript h and the coarsest one by superscript $2^L h$.) Assume that each level is defined by a finite-dimensional subspace, $\mathcal{S}^{2^l h}$, that is nested in the sense that $\mathcal{S}^{2^{l+1}h} \subset \mathcal{S}^{2^l h}$, $l = 0, \ldots, L-1$. Suppose that these spaces are written as a direct sum of one-dimensional subspaces: $\mathcal{S}^{2^l h} = \mathcal{S}_1^{2^l h} \oplus \cdots \oplus \mathcal{S}_{m_l}^{2^l h}$, $l = 0, \ldots, L$. Then one $V(0,1)$-PML cycle is defined as follows:

$$\begin{aligned}
&\mathbf{c}^{2^l h} \leftarrow \mathbf{0}, \qquad l = 0, \ldots, L \\
&\text{For } \ l = L, \ldots, 1: \qquad \text{(coarse-grid process)} \\
&\qquad \left[\begin{array}{l} \text{For } n = 0, \ldots, m_l \\ \qquad \mathbf{c}^{2^l h} \leftarrow \mathcal{D}_{\mathbf{x}^h}(\mathbf{c}^{2^l h}, \mathbf{d}_n^{2^l h}), \qquad \mathbf{d}_n^{2^l h} \in \mathcal{S}_n^{2^l h}, \\ \mathbf{c}^{2^{l-1} h} = \mathbf{c}^{2^l h} \end{array} \right. \\
&\text{For } \ l = 0: \qquad \text{(fine-grid process)} \\
&\qquad \left[\begin{array}{l} \text{For } n = 0, \ldots, m_0 \\ \qquad \mathbf{c}^h \leftarrow \mathcal{D}_{\mathbf{x}^h}(\mathbf{c}^h, \mathbf{d}_n^h), \qquad \mathbf{d}_n^h \in \mathcal{S}_n^h, \end{array} \right. \\
&\mathbf{x}^h \leftarrow \mathbf{x}^h + \mathbf{c}^h.
\end{aligned} \tag{3.8}$$

4. Coarse-grid realizability, different relaxation types, and higher-order discretizations. The key to obtaining an efficient multigrid-optimal PML implementation from (3.8) is the capability to perform the directional iteration step efficiently on coarse levels. Since the directional iteration step is based on functional evaluations, we focus now on how to do this efficiently on coarse levels.

4.1. Coarse-grid realizability. To obtain optimality, our multigrid algorithm must achieve two key objectives. First, we must be able to compute the FOSLS functional efficiently. Thus, update $\mathbf{c}^{2h}$ and the resulting new functional value must be computed quickly. Essentially, all level $2h$ calculations should in effect be performed on grid $2h$, not on grid h. Second, it must be possible to go from level $2h$ to level $4h$ without first updating the approximations on grid h.

To show how the first objective can be achieved, we need some additional notation and definitions. For simplicity, we choose the discretization to be the space, $\mathcal{S}^h$, of continuous piecewise-linear functions and the domain, $\Omega \subset \mathbb{R}^2$, to be two-dimensional, simply connected, and polygonal so that it can be partitioned into triangles. We consider here only triangulations by standard linear Lagrange triangles (cf. [5]). We need

to maintain a certain block-structured grid in order to obtain an efficient multigrid-optimal implementation of PML in two dimensions. Each level is defined by a finite-dimensional subspace, $\mathcal{S}^{2^l h}$, nested in the sense that $\mathcal{S}^{2^{l+1}h} \subset \mathcal{S}^{2^l h}$, $l = 0, \ldots, L-1$. For this section, we consider $\mathbf{x}^h = (\mathrm{x}_1^h, \ldots, \mathrm{x}_v^h)^t \in \mathcal{S}^h$ to be an arbitrary but fixed fine-grid approximation to the solution of the PDE. The $\mathbf{x}^h$ components, $\mathrm{x}_i^h : \Omega \to \mathbb{R}$ $(i = 1, \ldots, v)$, represent the different principal PDE variables (e.g., pressure, temperature, energy, and velocity). Further, we denote with $\mathbf{c}^{2^l h}$ a correction to fine-grid approximation $\mathbf{x}^h$ on level l (with approximate mesh size $2^l h$). Each component of correction $\mathbf{c}^{2^l h} = (\mathrm{c}_1^{2^l h}, \ldots, \mathrm{c}_v^{2^l h})^t \in \mathcal{S}^{2^l h}$ is a continuous piecewise-linear function and can be written, restricted to an element $\Omega_j^{2^l h}$, as a linear function as follows:

$$\mathrm{c}_i^{2^l h}(x, y)\Big|_{\Omega_j^{2^l h}} = s_1^{(i,j,l)} + s_2^{(i,j,l)} x + s_3^{(i,j,l)} y. \tag{4.1}$$

This representation holds for all $i = 1, \ldots, v$ and $j = 1, 2, \ldots, N^{(l)}$, where $N^{(l)}$ is the total number of elements on level l. The coefficients, $s_p^{(i,j,l)}$ with $p = 1, 2, 3$, are determined uniquely on level l by solving on each element, $\Omega_j^{2^l h}$ $(j = 1, 2, \ldots, N^{(l)})$, and for each $i = 1, \ldots, v$ the corresponding linear interpolation problem. In contrast to standard finite element practice, (4.1) can be seen as an alternative way to obtain a representation of $\mathbf{c}^h$ in $\mathcal{S}^h$. Standard nodal finite element bases could be used instead to represent (4.1), but this leads to substantially increased complexities for which we could not find an efficient way to avoid.

We next show how $\mathcal{F}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$ is computed for a modifiable fine-grid correction, $\mathbf{c}^h$, and an arbitrary but fixed approximation, $\mathbf{x}^h$. This is an essential step towards a multigrid-optimal algorithm and provides the basis for computing the FOSLS functional efficiently on coarser levels. For simplicity, we focus on one fine-grid element, Ω_j^h. This can be done without any loss of generality, since the sum of all fine-grid element contributions, $\mathcal{F}_{\Omega_j^h}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$, is the functional value, $\mathcal{F}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$. Further, we represent the fine-grid correction, $\mathbf{c}^h$ (level $l = 0$), as in (4.1) and consider its coefficients, $s_p^{(i,j,0)}$, as unknowns. In the next step, we use this representations to express the functional contribution, $\mathcal{F}_{\Omega_j^h}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$, in terms of the coefficients, $s_p^{(i,j,0)}$. Due to the nature of our quasi-linear first-order system and its L^2 least-squares functional, it is possible that the expansion of $\mathcal{F}_{\Omega_j^h}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$, with respect to the coefficients of $\mathbf{c}^h$, includes product terms of the coefficients, $s_p^{(i,j,0)}$, of up to order four. In the context of our new PML method, we regard all these terms as separate unknowns and store them as a matrix, $\mathbf{C}_j^h$. In this way, we are able to write the expansion of $\mathcal{F}_{\Omega_j^h}(\mathbf{x} + \mathbf{c}^h; \mathbf{g})$ as a matrix inner product of the form $\mathbf{A}_j^h : \mathbf{C}_j^h$. In the following, we refer to $\mathbf{A}_j^h$ as the local functional matrix and to $\mathbf{C}_j^h$ as the local coefficient matrix for element Ω_j^h on grid h. Now, whenever $\mathbf{c}^h$ changes on Ω_j^h, we obtain the new functional value, $\mathcal{F}_{\Omega_j^h}(\mathbf{x}^h + \mathbf{c}^h; \mathbf{g})$, by recomputing the local coefficient matrix and by evaluating the matrix inner product.

To show that we can compute the functional on coarser levels by coarse-grid calculations (the first objective), we assume a regular-structured grid. We further assume that four fine-grid elements always form one coarse-grid element. Denote by $\mathbf{C}_k^{2h}$ the coefficient matrix for $\mathbf{c}^{2h}$ on coarse-grid element Ω_k^{2h}. Further, let $\mathbf{C}_j^h$, with $j \in \{i | \Omega_i^h \subset \Omega_k^{2h}\}$, be the coefficient matrices for $\mathbf{c}^{2h}$ restricted to the fine-grid elements, Ω_j^h. The key observation now is to recognize that $\mathbf{C}_k^{2h} = \mathbf{C}_j^h$ for all

$j \in \{i|\Omega_i^h \subset \Omega_k^{2h}\}$. Then, we obtain the functional contribution for coarse-grid element Ω_k^{2h} as follows:

$$\mathcal{F}_{\Omega_k^{2h}}(\mathbf{x}^h + \mathbf{c}^{2h}; \mathbf{g}) = \sum_j \mathbf{A}_j^h : \mathbf{C}_j^h = \left(\sum_j \mathbf{A}_j^h\right) : \mathbf{C}_k^{2h} = \mathbf{A}_k^{2h} : \mathbf{C}_k^{2h}, \tag{4.2}$$

where $j \in \{i|\Omega_i^h \subset \Omega_k^{2h}\}$. Having all local fine-grid functional matrices available, we obtain the local coarse-grid functional matrices by a simple element-by-element addition of the respective fine-grid functional matrices. In this way, we can compute the functional, $\mathcal{F}(\mathbf{x}^h + \mathbf{c}^{2h}; \mathbf{g})$, for fixed fine-grid approximation $\mathbf{x}^h \in \mathcal{S}^h$ and any $\mathbf{c}^{2h} \in \mathcal{S}^{2h}$, entirely by grid $2h$ computations. We remark that, because of the nonlinear nature of our problems, these coarse-grid functional evaluations are possible only for approximations of the form $\mathbf{x}^h + \mathbf{c}^{2^l h}$, $l = 0, \ldots, L$, and not, for example, for approximations of the form $\mathbf{x}^h + \mathbf{c}^{2h} + \mathbf{c}^{4h}$. But, to fulfill the second objective, we would have to be able to compute approximations of the second type. To circumvent this drawback, we simply restrict ourselves to $V(0, \nu)$-cycles. Such a cycle is characterized for our new PML method by the following steps: fix the current approximation/initial guess; compute all local fine-grid functional matrices, $\mathbf{A}_j^h$; calculate the corresponding local coarse-grid functional matrices for all coarse levels; start the relaxation process on the coarsest level by applying ν sweeps there to compute the correction, $\mathbf{c}^{2^L h}$; interpolate this correction to the next finer level; use the interpolated correction as an initial value for relaxation on this next finer level; repeat the steps for $l = L, \ldots, 0$; and update the current approximation, $\mathbf{x}^h$, by $\mathbf{c}^h$.

By introducing local functional matrices and restricting ourselves to structured grids and $V(0, \nu)$-cycles, we can fulfill all the objectives for an efficient and multigrid optimal algorithm. Note that the same efficiency and optimality is retained if, instead of regular-structured grids, we use block-structured grids. Such grids allow the coarsest level to be unstructured, while the subsequent finer levels exhibit a regular structure.

4.2. Relaxation. The description of a $V(0, 1)$-PML cycle defines relaxation in general as $\mathbf{c}^{2^l h} \leftarrow \mathcal{D}_{\mathbf{x}^h}(\mathbf{c}^{2^l h}, \mathbf{d}_n^{2^l h})$, with $n = 0, \ldots, m_l$, where $\mathbf{x}^h$ is the current fine-grid approximation, $\mathbf{c}^{2^l h}$ is a coarse-grid correction on level $2^l h$, $\mathbf{d}_n^{2^l h} \in \mathcal{S}_n^{2^l h}$ is a search direction, and $\mathcal{S}_n^{2^l h}$, $n = 0, \ldots, m_l$, are spaces that decompose $\mathcal{S}^{2^l h}$. We clearly see that picking $\mathcal{S}_n^{2^l h}$ characterizes the type of relaxation. Before illustrating relationships between $\mathcal{S}_n^{2^l h}$ and different relaxation types, we first comment on issues concerning the realization and implementation of directional iteration or relaxation steps in general.

The classical approach for relaxation schemes, such as Gauss–Seidel or damped Jacobi, are based on a finite number of either explicitly given linear equations, typically written in matrix form, or nonlinear equations. Since relaxation for our PML method is based on a nonlinear functional minimization principle, we cannot use them in the same way that most standard approaches present them. To implement relaxation so that we keep the overall promise of avoiding linearization while obtaining an efficient algorithm, we restrict ourselves to a FOSLS functional for quasi-linear PDEs. For this class of PDE formulations, the nonlinearity appears in the functional as a cross product of two different variables, which implies linearity of the weak form with respect to each variable.

To illustrate this linearity, consider a least-squares functional consisting of the product of two variables: $\mathcal{F}([u, v]^t; 0) = \|uv\|_{0,\Omega}^2$. (For clarity, we use u and v instead

of x_1 and x_2.) Let $\mathcal{S}^h$ be a standard finite element space with approximate mesh size h. Then choose a relaxation direction for each variable: $\mathbf{d}_1^h = [d_{u^h}^h, 0]^t \in \mathcal{S}^h$ and $\mathbf{d}_2^h = [0, d_{v^h}^h]^t \in \mathcal{S}^h$. Relaxing on each variable of $\mathcal{F}([u^h, v^h]^t; 0)$ separately, we obtain for $\mathbf{x}^h = [u^h, v^h]^t \in \mathcal{S}^h$ the following relaxation process:

(4.3)
$$\begin{cases} s_1 = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \mathcal{F}(\mathbf{x}^h + t\mathbf{d}_1^h; 0) = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \mathcal{F}([u^h + td_{u^h}^h, v^h]^t; 0) =: \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \bar{\mathcal{F}}_1(t), \\ u^h \leftarrow u^h + s_1 d_{u^h}^h \end{cases}$$

and

(4.4)
$$\begin{cases} s_2 = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \mathcal{F}(\mathbf{x}^h + t\mathbf{d}_2^h; 0) = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \mathcal{F}([u^h, v^h + td_{v^h}^h]^t; 0) =: \underset{t\in\mathbb{R}}{\operatorname{argmin}}\ \bar{\mathcal{F}}_2(t), \\ v^h \leftarrow v^h + s_2 d_{v^h}^h. \end{cases}$$

For our class of PDEs, functions $\bar{\mathcal{F}}_1(t)$ and $\bar{\mathcal{F}}_2(t)$ defined in (4.3) and (4.4) are quadratic polynomials in the scalar, t. To obtain the quadratic formulation for $\bar{\mathcal{F}}_1(t)$ (or $\bar{\mathcal{F}}_2(t)$), we evaluate $\bar{\mathcal{F}}_1(t)$ (or $\bar{\mathcal{F}}_2(t)$) at three different locations and fit the functional values quadratically. In this way, the quadratic polynomial fits $\bar{\mathcal{F}}_1(t)$ (or $\bar{\mathcal{F}}_2(t)$) exactly. Actually, we need only evaluate $\bar{\mathcal{F}}_1(t)$ (or $\bar{\mathcal{F}}_2(t)$) at two locations because the current functional value ($t = 0$) is known. Also, after computing the optimal step length, which is the minimum of the quadratic polynomial, we obtain the new current functional value by plugging s_1 (or s_2) into our quadratically fitted curve.

For nonlinear PDEs with the type of nonlinearity that is the focus of this research, an alternating-variable relaxation process leads to scalar minimization problems in which the objective function is quadratic. This property is due to the nature of our constructed FOSLS functional, but it is no longer true for a relaxation scheme that simultaneously relaxes on both (resp., all) variables. Instead of (4.3) and (4.4), block relaxation on both variables leads to the following relaxation process:

(4.5)
$$\begin{cases} (s_1, s_2) = \underset{(t_1,t_2)\in\mathbb{R}^2}{\operatorname{argmin}}\ \mathcal{F}(\mathbf{x}^h + t_1\mathbf{d}_1^h + t_2\mathbf{d}_2^h; 0) \\ \qquad\quad = \underset{(t_1,t_2)\in\mathbb{R}^2}{\operatorname{argmin}}\ \mathcal{F}(\big[u^h + t_1 d_{u^h}^h, v^h + t_2 d_{v^h}^h\big]; 0) =: \underset{(t_1,t_2)\in\mathbb{R}^2}{\operatorname{argmin}}\ \bar{\mathcal{F}}_3(t_1, t_2), \\ u^h \leftarrow u^h + s_1 d_{u^h}^h, \quad v^h \leftarrow v^h + s_2 d_{v^h}^h. \end{cases}$$

Now, the possibility of performing a relaxation step without appealing to linearization depends on the capability of computing the minimizer of $\bar{\mathcal{F}}_3(t_1, t_2)$. Similar conclusions can be made for PDEs in which derivatives appear to some integer power.

Even though we illustrated only one fine-grid relaxation step for two scalar unknowns, we can apply the same techniques for more than two variables, for unknowns that are vector functions, and on coarser levels. Moreover, at this point, we see why it is extremely important to be able to compute functional values on all levels efficiently. Relaxation is the main contributor to the overall computational cost and is almost solely based on functional evaluations.

We can relax on the unknowns in an alternating fashion, as described by (4.3) and (4.4), for almost any choice of relaxation subspaces, $\mathcal{S}_n^{2^l h}$, and discretization. These choices only affect the type of relaxation. In what follows, we give two examples for

different relaxation types, a Richardson-like scheme and a Gauss–Seidel-like scheme. Although we describe the different relaxation types as if the functional had only one unknown, we still relax on the unknowns in an alternating way.

To obtain a Richardson-like relaxation scheme, we choose $m_l = 1$ on all levels. This means that there is only one relaxation step per sweep. As the single direction, $\mathbf{d}_1^{2^l h} \in \mathcal{S}_1^{2^l h} = \mathcal{S}^{2^l h}$, we make the natural choice of "steepest" descent given by the gradient of the functional with respect to the unknown. We compute the gradient of our nonlinear functional numerically: its value at node n is determined by the forward-difference formula, $\big(\mathcal{F}(\mathbf{x}^{2^l h} + s\,\mathbf{e}_n^{2^l h}; \mathrm{g}) - \mathcal{F}(\mathbf{x}^{2^l h}; \mathrm{g})\big)/s$, where $\mathbf{e}_n^{2^l h}$ is the nth nodal finite element basis function (with value one at grid point n and zero elsewhere) and s is sufficiently small; the discrete representation of the gradient, $\mathbf{d}_1^{2^l h}$, is then just the continuous piecewise polynomial in $\mathcal{S}^{2^l h}$ that has these nodal values.

If we now choose our relaxation subspaces as the span of individual basis or nodal finite element basis functions (with a value of one at a single node and zero at all other nodes), we obtain a coordinate minimization or nonlinear Gauss–Seidel relaxation process. Hence, we choose m_l to be equal to the number of nodes on level l, $\mathbf{d}_n^{2^l h}$ as the nodal finite element basis function, and $\mathcal{S}_n^{2^l h}$ as the space spanned by the nodal basis function of node n. This means that we minimize consecutively over all nodes, n, by computing the step length, $s = \min_{t\in\mathbb{R}} \mathcal{F}(\mathbf{x}^h + \mathbf{c}^{2^l h} + t\mathbf{d}_n^{2^l h}; \mathrm{g})$, and the resultant update, $\mathbf{c}^{2^l h} \leftarrow \mathbf{c}^{2^l h} + s\mathbf{d}_n^{2^l h}$. Note that this is a local process in that the approximation, $\mathbf{x}^h$, changes only at one node per step of the sweep.

Gauß-Seidel is typically a more efficient smoother than a gradient or Richardson-type process.

4.3. Higher-order discretizations. Many engineering problems require more than just a linear finite element discretization. For example, the numerical solution to the Navier–Stokes equations obtained by using linear finite elements and a triangular discretization in a FOSLS formulation usually does not conserve mass very well. This section shows the potential of using higher-order finite elements in our framework of PML. For simplicity, however, we limit ourselves in this section to standard quadratic Lagrange triangles, which generate the space, $\mathcal{S}_Q^h$, of continuous piecewise-quadratic finite elements. It should be noted that any other higher-order discretization or other element type can be implemented in a similar way. Mimicking the representation of linear functions over elements, we describe approximations or corrections in $\mathcal{S}_Q^h$ restricted to an element (in this case, we use a less cumbersome notation as we restrict our correction to a reference element, Ω_{ref}^h) by

$$c^h(x,y)\Big|_{\Omega_{ref}^h} = s_0^h + s_1^h x + s_2^h y + s_3^h xy + s_4^h x^2 + s_5^h y^2. \tag{4.6}$$

Similar to fine-grid level h, we introduce quadratic finite element spaces on coarser levels: $\mathcal{S}_Q^{2^l h}$, $l = 1, \ldots, L$. The subscript Q indicates the use of quadratic ansatz functions to generate the space. We stress that the lack of such a subscript signifies linear ansatz functions. For the multilevel implementation with quadratic finite elements, we use an unstructured triangulation for the coarsest level. All finer levels are obtained by subdividing each coarser-grid-level triangle into four equal triangles. To this end, we consider as follows two different coarse-grid correction processes that differ by the choice of the coarse-grid correction subspaces:

1. On all levels, corrections are obtained from the quadratic finite element subspaces, $\mathcal{S}_Q^{2^l h}$ $(l = 0, \ldots, L)$.

2. Only $\mathcal{S}_Q^h$ is used for the fine-grid corrections, while the coarse-grid process uses corrections from the linear finite element subspaces, $\mathcal{S}^{2^l h}$ $(l = 1, \ldots, L)$.

To achieve for both approaches an efficient or even multigrid-optimal algorithm for quadratic finite elements, we mimic ideas and techniques from previous discussions on linear finite elements. There, we introduced local functional matrices, which led to an efficient way of handling modifications to coarse-grid corrections. These local functional matrices can be computed for quadratic finite elements in a similar way. The key observation, which led to multigrid-optimality, is to recognize the need to use $V(0,\nu)$-cycles instead of $V(\mu,\nu)$-cycles for our PML method.

For the first approach, coarse-grid functional matrices are obtained as in the case for linear finite elements by adding up the respective fine-grid functional matrices. At the end of each cycle, we update the current approximation of the solution, $\mathbf{x}_Q^h$, by $\mathbf{c}_Q^h$, compute the new local functional matrices, and repeat the cycle. For the second approach, we first alter the triangulation from quadratic Lagrange triangles to linear Lagrange triangles. Then we apply a standard $V(0,\nu)$-cycle that involves relaxation on corrections represented by continuous piecewise-linear functions. At the end of this $V(0,\nu)$-cycle, we project the current (piecewise-linear) correction onto the original triangulation with quadratic Lagrange triangles. We continue the cycle by performing ν further relaxation sweeps on the correction, now represented by continuous piecewise-quadratic functions, by updating the current approximation of the solution, $\mathbf{x}_Q^h$, by $\mathbf{c}_Q^h$, and by computing the new local functional matrices to repeat the cycle.

Compared to the first approach, the second has two advantages. First, it allows reuse of most of the code for linear finite elements. Second, the coarsening process is independent of the order of the fine-grid discretization. This property becomes increasingly important as we choose increasingly higher-order discretizations on the finest level. To illustrate this, recall that our PML method treats the nonlinearity directly without any kind of linearization process. Thus, our local functional matrices are growing rapidly in complexity for higher-order discretizations. (The complexity grows for quasi-linear problems even more than for linear ones.) With the second approach, we use a multilevel strategy to compute a piecewise-linear approximation to the fine-grid correction. Having this approximation on the finest level available, we have the advantage of no longer being constrained to use local functional matrices or the same relaxation process as on coarser levels. In principle, we could consider the fine-level correction as a separate minimization process, with the advantage of having a good coarse-grid corrected initial guess. This is very appealing in particular for high-order finite elements.

However, low-order spaces do not always provide an effective coarse-level correction for high-order spaces in the same elements. An alternative is to partition the elements defining the high-order discretization into several smaller elements that could then be used to define the linear correction space (cf. [18, 21]). In our context of standard quadratic finite elements, we could split each quadratic Lagrange triangle into four linear Lagrange triangles. Although, this would violate our assumption of nested finite-dimensional subspaces, we would still obtain a good low-order approximation to the (high-order) correction on the finest level. This is very appealing for high-order element types, in particular, since this allows us, again, to consider the (high-order) fine-level problem as a separate minimization process.

5. Numerical results. Here we report on numerical results for some test problems. First, we study performance of our PML method on a set of nonlinear test

problems by adding a simple nonlinear term to the Laplace operator, using a coefficient, α, that allows us to adjust the strength of nonlinearity. We finish this section by presenting numerical results for our target application, the incompressible Navier–Stokes equations, with particular focus on the so-called Kovasznay flow.

5.1. Measuring convergence factors. Before we provide numerical results on some test problems, we first address the issue of how to measure convergence factors for our method since they play an especially important role in analyzing and evaluating a multigrid iteration.

Let $\mathcal{F}(\mathbf{x}^{2^l h};\mathsf{g})$ be the discrete nonlinear functional for a nonlinear PDE written in FOSLS form, and let $\mathbf{x}_*^{2^l h}$ be the minimizer of $\mathcal{F}(\mathbf{x}^{2^l h};\mathsf{g})$ on level l. Superscript $2^l h$ does not play an essential role here, but we use it anyway to emphasize that the operator stems from a discretization on a certain level, l. Taking our cue from the linear case, we write the functional norm defect of our current approximation as

$$\hat{\delta}_k^{2^l h} = \sqrt{\mathcal{F}(\mathbf{x}_k^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^{2^l h};\mathsf{g})}, \tag{5.1}$$

where $\mathbf{x}_k^{2^l h}$ is the approximation to the exact solution, $\mathbf{x}_*^{2^l h}$, after the kth iteration step. Note that $\hat{\delta}_k^{2^l h}$ is a positive real number because $\mathbf{x}_*^{2^l h}$ is the minimizer of $\mathcal{F}(\mathbf{x}^{2^l h};\mathsf{g})$. In analogy to computing convergence factors for linear systems (cf. [9, 31]), we define the convergence factor for the kth iteration step on level l by

$$\widehat{CF}_k^{(l)} := \frac{\hat{\delta}_k^{2^l h}}{\hat{\delta}_{k-1}^{2^l h}} = \sqrt{\frac{\mathcal{F}(\mathbf{x}_k^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^{2^l h};\mathsf{g})}{\mathcal{F}(\mathbf{x}_{k-1}^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^{2^l h};\mathsf{g})}}. \tag{5.2}$$

Since $\mathcal{F}(\mathbf{x}_*^{2^l h};\mathsf{g})$ is unknown, (5.2) cannot be used directly to compute the convergence factor. Thus, instead of considering the defect, $\hat{\delta}_k^{2^l h}$, as in (5.1), we take the approach of defining the defect of two consecutive approximations:

$$\delta_k^{2^l h} = \sqrt{\mathcal{F}(\mathbf{x}_{k-1}^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_k^{2^l h};\mathsf{g})}. \tag{5.3}$$

The attendant convergence factor estimate is then given by

$$CF_k^{(l)} := \frac{\delta_k^{2^l h}}{\delta_{k-1}^{2^l h}} = \sqrt{\frac{\mathcal{F}(\mathbf{x}_{k-1}^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_k^{2^l h};\mathsf{g})}{\mathcal{F}(\mathbf{x}_{k-2}^{2^l h};\mathsf{g}) - \mathcal{F}(\mathbf{x}_{k-1}^{2^l h};\mathsf{g})}}. \tag{5.4}$$

Note that this measures requires care with respect to machine precision and numerical cancellation. For example, if $\mathcal{F}(\mathbf{x}_{k-1}^{2^l h};\mathsf{g})$ and $\mathcal{F}(\mathbf{x}_k^{2^l h};\mathsf{g})$ in (5.4) are the same up to near machine precision, then convergence factors can give the impression of degenerating performance.

5.2. A nonlinear model problem. As a first test for our algorithm, we choose a Poisson problem with pure Dirichlet boundary conditions on $\Omega = [0,1] \times [0,1]$ that has been modified by the addition of a nonlinear term, $\alpha p p_x$. Parameter α allows us to vary the strength of the nonlinearity. This model represents a simple nonlinear PDE with the type of nonlinearity that is the focus of this research. Its FOSLS formulation

is given as follows:

$$
\begin{aligned}
\nabla p - \mathbf{u} &= \mathbf{0} && \text{in } \Omega,\\
-\frac{1}{\alpha}\nabla\cdot\,\mathbf{u} + pu_1 &= f_\Omega && \text{in } \Omega,\\
\frac{1}{\alpha}\nabla\times\,\mathbf{u} &= 0 && \text{in } \Omega,\\
p &= f_\Gamma && \text{on } \Gamma_\Omega,\\
\mathbf{n}\times\mathbf{u} &= \mathbf{n}\times f_\Gamma && \text{on } \Gamma_\Omega.
\end{aligned}
\tag{5.5}
$$

where $\nabla p = (u_1, u_2)^t$, $\mathbf{n}$ is the unit outward normal on boundary Γ_Ω, and Ω is the unit square. For further details on FOSLS formulations of the Poisson problem, see [10] and [11]. We choose $p(x, y) = x^2 + y^2$ as the exact solution and thus obtain $f_\Omega = -4/\alpha + (2x^3 + 2xy^2)$ as the right side. For all of our experiments, we use Dirichlet boundary conditions derived from the exact solution. We enforce the boundary conditions strongly by imposing them on the finite element space. Note that (5.5) arises from a more favorable scaling of the first-order system derived from the PDE, $\Delta p + \alpha\, pp_x = \tilde{f}_\Omega$. Hence, parameter α allows us to vary the strength of nonlinear term pp_x, and $\alpha = 0$ reduces it to the linear Poisson problem. Its FOSLS functional is constructed by taking the L^2-norm of each interior equation,

$$
\mathcal{F}(p, \mathbf{u}; \mathbf{g}) = \|\nabla p - \mathbf{u}\|^2_{0,\Omega} + \| -\frac{1}{\alpha}\nabla\cdot\,\mathbf{u} + pu_1 + \frac{4}{\alpha} - (2x^3 + 2xy^2)\|^2_{0,\Omega} + \|\frac{1}{\alpha}\nabla\times\,\mathbf{u}\|^2_{0,\Omega},
\tag{5.6}
$$

where $\mathbf{g} = (\mathbf{0}, f_\Omega, 0)$. The grids are based on a regular triangulation of Ω by 16 elements and 13 grid points. This coarsest level is denoted by $l = 7$, with an approximate mesh size $2^l h$, where h is the approximate mesh size with respect to the finest level. Level 6 is formed by taking every element of level 7 and subdividing it into 4 equal triangles. The midpoints of the coarse-grid element sides are the new fine-grid points. Successively finer levels are constructed in the same way. This refinement leads to 131,585 nodes (with 3 degrees of freedom per node) and 262,144 elements on level 0. A nested iteration algorithm with 10 $V(0, 4)$-Gauss–Seidel relaxation sweeps on each level is used to minimize $\mathcal{F}(p, \mathbf{u}; \mathbf{g})$ in (5.6) over the space consisting of continuous piecewise-linear functions. Extensive experiments with several problems of this type lead us to believe that $V(0, 4)$-cycles achieve nearly the best accuracy-complexity tradeoffs over other such cycling strategies. Since our exact solution cannot be represented exactly by our finite element space, the functional cannot converge to zero, but rather stagnates as the iteration reaches the level of discretization error on each grid. Table 5.1 depicts the functional norms, $\mathcal{F}(p^{2^l h}_{10}, \mathbf{u}^{2^l h}_{10}; \mathbf{g})^{\frac{1}{2}}$, obtained on each level for the linear Poisson problem and α varying between 1 and 10,000. Table 5.2 reports on the corresponding final convergence factors, $CF^{(l)}_{10}$, computed according to (5.4). Here, we choose to report the convergence factor of the last iteration, since it tends to be the worst in our numerical tests.

Note that, on each level, we obtain accuracy close to discretization level within 10 $V(0, 4)$-cycles. We have not used any special technique (e.g., streamline relaxation) to address the changing character of the operator as α increases. Thus, as expected, the final convergence factors degrade as the nonlinearity increases in dominance, but they remain grid-independent. Though one might argue that the convergence factors in the last column of Table 5.2 (α = 10,000) do not exhibit grid-independent convergence

TABLE 5.1

Measured functional norm (5.6), $\mathcal{F}(p_{10}^{2^l h}, \mathbf{u}_{10}^{2^l h}; \mathrm{g})^{\frac{1}{2}}$, *for different* α *using a linear finite element discretization and* $10V(0,4)$*-cycles with Gauss–Seidel as smoother.*

Level	Linear Poisson	Nonlinearity parameter α				
		1	10	100	1,000	10,000
	Functional norm	Functional norm	Functional norm	Functional norm	Functional norm	Functional norm
7	2.7635e−01	2.7635e−01	2.6843e−01	2.5509e−01	2.5357e−01	2.5342e−01
6	1.4162e−01	1.4207e−01	1.4046e−01	1.3540e−01	1.3460e−01	1.3452e−01
5	7.1749e−02	7.1800e−02	7.1545e−02	7.0292e−02	7.0032e−02	7.0007e−02
4	3.6021e−02	3.6027e−02	3.5992e−02	3.5762e−02	3.5714e−02	3.5710e−02
3	1.8033e−02	1.8033e−02	1.8029e−02	1.8007e−02	1.8032e−02	1.8036e−02
2	9.0197e−03	9.0198e−03	9.0193e−03	9.0259e−03	9.0756e−03	9.0822e−03
1	4.5103e−03	4.5103e−03	4.5103e−03	4.5160e−03	4.5750e−03	4.5833e−03
0	2.2552e−03	2.2552e−03	2.2552e−03	2.2583e−03	2.3232e−03	2.3331e−03

TABLE 5.2

Convergence factors, $CF_{10}^{(l)}$*, for the same experiments as in Table* 5.1.

Level	Linear Poisson	Nonlinearity parameter α				
		1	10	100	1 000	10 000
	$CF_{10}^{(l)}$	$CF_{10}^{(l)}$	$CF_{10}^{(l)}$	$CF_{10}^{(l)}$	$CF_{10}^{(l)}$	$CF_{10}^{(l)}$
7	0.031	0.029	0.128	0.258	0.272	0.274
6	0.054	0.063	0.318	0.602	0.632	0.635
5	0.091	0.068	0.443	0.776	0.809	0.812
4	0.108	0.092	0.538	0.825	0.865	0.866
3	0.116	0.098	0.581	0.872	0.891	0.893
2	0.117	0.097	0.595	0.893	0.926	0.928
1	0.117	0.090	0.599	0.908	0.944	0.946
0	0.108	0.096	0.599	0.918	0.955	0.957

factors, it is believed that grid-independent convergence factors are obtained once a sufficiently small mesh size is reached.

The fact that we reached the level of discretization error is also supported by the functional reduction factors. For continuous piecewise-linear finite elements for our problem, standard theory (cf. [12]) establishes asymptotic $O(h)$ H^1-error bounds, so H^1 ellipticity of our functional yields an $O(h)$ functional-norm bound. We might, thus, expect about a factor of 2 in functional-norm reduction from one level to the next finer one. Let the functional reduction factors as the resolution doubles be defined by

$$\beta_n^{(l)} = \sqrt{\frac{\mathcal{F}(p_n^{2^{l+1}h}, \mathbf{u}_n^{2^{l+1}h}; \mathrm{g})}{\mathcal{F}(p_n^{2^l h}, \mathbf{u}_n^{2^l h}; \mathrm{g})}}, \qquad l = 0, \ldots, L-1. \tag{5.7}$$

Table 5.3 depicts these factors for different levels and strengths of nonlinearity. For all levels and strengths of nonlinearity, we observe a functional reduction factor of about 2, which is consistent with the use of continuous piecewise-linear finite elements.

Next, we analyze error reduction factors as we step through the different levels. Consider again the same FOSLS formulation, levels, and number of V-cycles per level used for the results in Table 5.1. We now compare the numerically obtained solution, $p^{2^l h}$, with the exact solution, $p = x^2 + y^2$, for each level and for each α ($\alpha = 1$, 10, 100, 1,000, and 10,000), measured by the H^1- and L^2-norms. In Figure 5.1, we depict the H^1-error norm versus the number of elements. The L^2-error norm versus the number of elements is shown in Figure 5.2. Since we use a regular refinement strategy to step

TABLE 5.3
Functional reduction factors, $\beta_{10}^{(l)}$, based on functional norms reported in Table 5.1.

	Linear	Nonlinearity parameter α				
	Poisson	1	10	100	1 000	10 000
Level	$\beta_{10}^{(l)}$	$\beta_{10}^{(l)}$	$\beta_{10}^{(l)}$	$\beta_{10}^{(l)}$	$\beta_{10}^{(l)}$	$\beta_{10}^{(l)}$
7	—	—	—	—	—	—
6	1.95	1.95	1.91	1.88	1.88	1.88
5	1.97	1.98	1.96	1.93	1.92	1.93
4	1.99	1.99	1.99	1.97	1.96	1.96
3	2.00	2.00	2.00	1.99	1.98	1.98
2	2.00	2.00	2.00	2.00	1.99	1.99
1	2.00	2.00	2.00	2.00	1.98	1.98
0	2.00	2.00	2.00	2.00	1.97	1.97

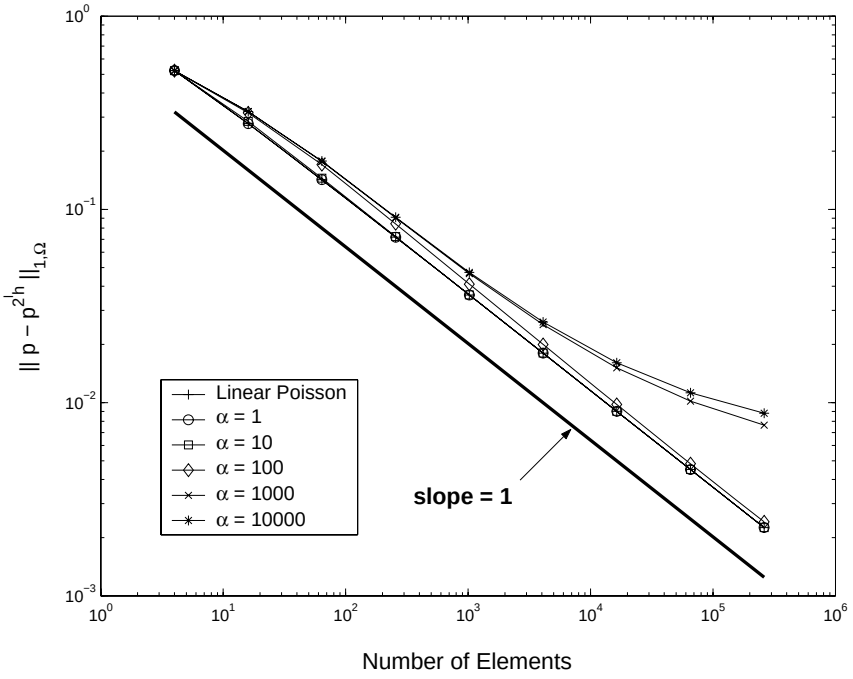


FIG. 5.1. *H^1-error, $\|p - p^{2^l h}\|_{1,\Omega}$, versus the number of elements for the linear Poisson problem and* (5.5) *with* $\alpha = 1$, 10, 100, 1,000, *and* 1,0000.

through the levels (with each refinement, we increase the number of elements by a factor of 4 and, therefore, halve our mesh size), reporting on the number of elements is the same as reporting on the mesh size. For both figures, we use a logarithmic scale for the number of elements (abscissa) and the error-norm (ordinate). For each α, the H^1-error norm (or L^2-error norm) is measured for each level and indicated with data points, which are connected in such a way that each line displays one nested iteration process for some α. Additionally, we include in Figure 5.1 a supporting line with slope 1 and in Figure 5.2 two supporting lines with slopes 1 and 2. These supporting lines should help retrieve an estimate of the error-reduction factors directly from the

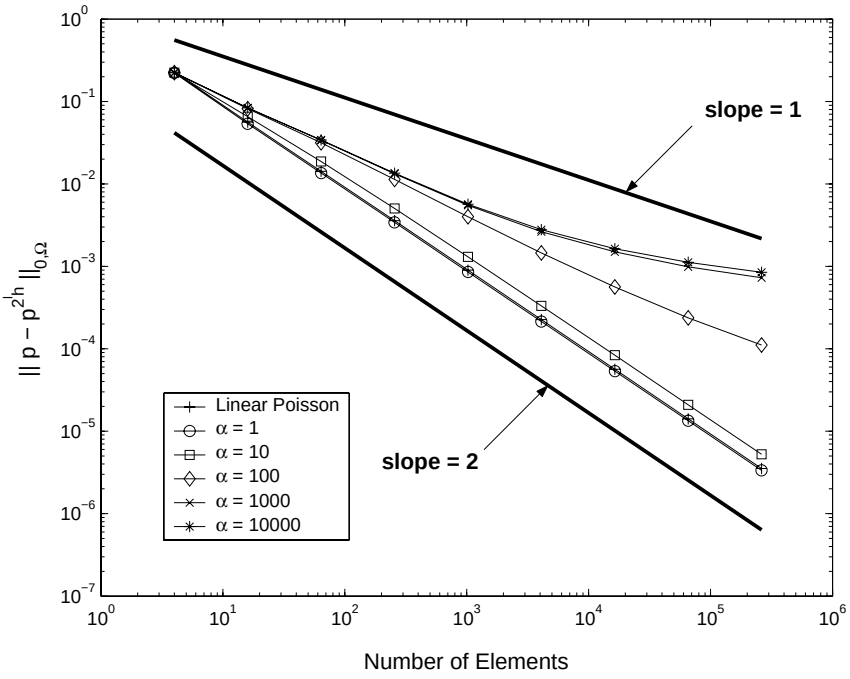


FIG. 5.2. *L^2-error, $\|p-p^{2^l h}\|_{0,\Omega}$, versus the number of elements for the linear Poisson problem and* (5.5) *with* $\alpha = 1$, 10, 100, 1,000, *and* 10,000.

graph. Note that a slope of $\mathbf{s}$ in Figures 5.1 and 5.2 means that

$$\frac{\|p - p^{2^l h}\|}{\|p - p^{2^{l+1} h}\|} \approx \Big(\frac{2^{l+1} h}{2^l h}\Big)^{\mathbf{s}} = 2^{\mathbf{s}}.$$

Hence, slope $\mathbf{s}$ translates to an error-reduction factor of $2^{\mathbf{s}}$. Analyzing Figure 5.1, the error-reduction factor from one level to the next is about 2 for every α. This coincides well with the reported functional reduction factors, $\beta_{10}^{(l)}$, in Table 5.3, and are considered to be optimal for linear finite elements. From the excellent agreement of the FOSLS functional norm and the H^1-error reduction factors, we conclude that the functional in (5.6) appears to be H^1-elliptic. This numerical observation coincides with the theoretical results of the companion paper [23], where we establish H^1 ellipticity of the FOSLS functional based on the Navier–Stokes equations and anticipate it for other quasi-linear PDEs of that class.

In Figure 5.2, we display the L^2-error norms in the same way as the H^1-error in Figure 5.1. We now observe strongly deteriorating L^2-error reduction factors with increasing strength of nonlinearity. One possible explanation for this might involve the Nitsche Trick (cf. [8]), which relates two different error norms to each other (in this case, the H^1-error norm and the L^2-error norm). Its proof is based on the assumption that the exact solution, $\mathbf{x}_*^{2^l h}$, is found on each level. With a nested iteration scheme, we compute on each level only an approximation to $\mathbf{x}_*^{2^l h}$; here, for example, we approximate $\mathbf{x}_*^{2^l h}$ by $\mathbf{x}_{10}^{2^l h}$. In separate experiments, we have been able to recover near-optimal L^2-error reduction factors by using 100 V-cycles instead of 10 on each level. This shows that better algebraic accuracy is needed on each level to control the L^2-error. This should be expected since greater L^2 accuracy is obtained

from the discretization on each level, so nested iteration should have to work harder than for H^1 accuracy to achieve it. Development of effective criteria for a nested iteration strategy that efficiently produces small H^1 and L^2 errors is still an open problem.

5.3. Kovaszany flow. While system (5.5) provides an important problem to test the behavior of the algorithm, our ultimate goal is to solve the Navier–Stokes equations. For concreteness, we focus on the steady-state incompressible Navier–Stokes equations in velocity-pressure formulation given as follows:

$$\begin{aligned} -\frac{1}{Re}\Delta\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \nabla p &= \mathbf{0} \qquad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 \qquad \text{in } \Omega. \end{aligned} \tag{5.8}$$

Velocity vector variable $\mathbf{u} = (u_1, u_2)^t$ and pressure scalar variable p are nondimensionalized. Re denotes the Reynolds number defined as $Re = (U_{ref}L)/\nu$, where L is a reference length, U_{ref} a reference velocity, and ν the kinematic viscosity (see [19]). Note that the source terms in this system are all zero. We could easily incorporate nonzero terms, but choose this simplification instead because our primary focus is on the algebraic solver and because inhomogeneities are incorporated in the boundary conditions in any case.

To obtain a first-order system from (5.8), we introduce a new velocity-flux tensor variable, $\mathbf{U} = (U_{i,j})_{2\times 2} = (\partial u_j/\partial x_i)_{2\times 2} = \nabla\mathbf{u}^t$. (See [1] for details on the FOSL-Sization of (5.8).) We thus obtain the following first-order velocity-flux form of the Navier–Stokes equations:

$$\begin{aligned} \nabla\mathbf{u}^t - \mathbf{U} &= \mathbf{0} \qquad \text{in } \Omega, \\ -\frac{1}{Re}(\nabla\cdot\mathbf{U})^t + \mathbf{U}^t\mathbf{u} + \nabla p &= \mathbf{0} \qquad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 \qquad \text{in } \Omega, \\ \frac{2}{Re}\nabla\times\mathbf{U} &= 0 \qquad \text{in } \Omega. \end{aligned} \tag{5.9}$$

The difference between this system and that proposed in [1] is the factor of 2 in the last equation and the missing trace term, $\nabla tr(\mathbf{U})$. The additional factor is a simple weighting of this equation that, by our empirical observations, results in slightly better numerical results. Concerning the trace term, because of the incompressibility condition expressed by $\partial_x u_1 + \partial_y u_2 = U_{11} + U_{22} = 0$, we are able to eliminate one of the variables by setting $U_{11} = -U_{22}$, which in turn enforces $\nabla tr(\mathbf{U}) = 0$ and therefore makes this trace equation unnecessary. Of course, system (5.9) offers but one approach to reducing the second-order problem to first order. Other choices are given, for example, in [3] and [19]. In any case, the solution of our first-order system is the minimizer of the least-squares functional given by

$$\begin{aligned} \mathcal{F}(\mathbf{u},\mathbf{U},p\,;\mathbf{g}) = \|\nabla\mathbf{u}^t - \mathbf{U}\|_{0,\Omega}^2 + \| - \frac{1}{Re}(\nabla\cdot\mathbf{U})^t + \mathbf{U}^t\mathbf{u} + \nabla p\|_{0,\Omega}^2 \\ + \|\nabla\cdot\mathbf{u}\|_{0,\Omega}^2 + \left\|\frac{2}{Re}\nabla\times\mathbf{U}\right\|_{0,\Omega}^2, \end{aligned} \tag{5.10}$$

where $\mathbf{g} = (\mathbf{0}, \mathbf{0}, 0)$ is the combined right side of the equations in (5.9).

TABLE 5.4
Convergence summary for Kovasznay flow with $Re = 40$, a nested iteration PML approach with 10 $V(0,4)$-cycles per level, Gauss–Seidel as smoother, and quadratic finite elements.

Level l	Nodes/Elements	Functional norm $\mathcal{F}(\mathbf{x}_{10}^{2^l h};\mathbf{g})^{\frac{1}{2}}$	Functional reduction factor $\beta_{10}^{(l)}$	$CF_{10}^{(l)}$
5	41/16	4.212367e+00		0.713
4	145/64	1.635538e+00	2.57	0.788
3	545/256	4.854122e−01	3.37	0.854
2	2 113/1 024	1.379767e−01	3.51	0.880
1	8 321/4 096	3.760596e−02	3.67	0.892
0	33 025/16 384	9.993762e−03	3.78	0.897

As a model problem for our algorithm applied to the Navier–Stokes equations, we turn to Kovasznay flow. This particular system is named after L. I. G. Kovasznay, who derived in [20] an analytic solution for the steady-state incompressible Navier–Stokes equations for a special laminar flow problem. We choose this problem as a test case, since it is posed on a rectangular domain, $\Omega = [-.5, 2.0] \times [-.5, 1.5]$, has a smooth solution, and exhibits no singularities. Knowledge of the analytical solution allows us to strongly impose the exact boundary conditions. Actually, for accurate error estimates, we need not appeal to an exact analytic solution, since the FOSLS functional itself naturally provides a sharp error measurement. But use of an exact solution gives a somewhat tighter estimate of any error measure we choose to use.

In Table 5.4, we give the convergence history using continuous piecewise-quadratic functions for the Kovasznay flow problem with a Reynolds number of 40. For the cycling strategy, we choose to use quadratic finite elements for the fine-grid corrections and linear finite element subspaces for the coarse-grid process (see the second approach of section 4.3). The grids are based on a regular triangulation of Ω by 16 elements and 41 nodes. Again, we use a nested iteration approach to step through the levels. On each level, we apply 10 $V(0,4)$-PML cycles, with Gauss–Seidel as smoother. For each level, we report on final functional norm values, $\mathcal{F}(\mathbf{x}_{10}^{2^l h};\mathbf{g})^{\frac{1}{2}}$, the functional reduction factor, $\beta_{10}^{(l)}$, defined as in (5.7), and the final convergence factor, $CF_{10}^{(l)}$, defined as in (5.4).

The results in Table 5.4 show that we also obtain approximate grid-independent convergence factors for the FOSLS formulation of the Navier–Stokes problem and nearly optimal finite element approximation properties. The fact that the functional reduction factor, $\beta_{10}^{(l)}$, is hovering around 3.7 instead of an optimal factor of 4 for quadratic Lagrange finite elements is probably due mostly to the approximations not yet being in the asymptotic range. Note the increase in these factors with decreasing h. (Our tests that increased the number of V-cycles showed only marginal increase in the functional reduction factors.)

Though we report here only on results for $Re = 40$ (the classical setting for the Kovasznay flow), we have done experiments for much higher Reynolds numbers. We obtained similar results, although the convergence factors naturally degraded since the PML scheme was not designed for convection-dominated problems.

REFERENCES

[1] P. BOCHEV, Z. CAI, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal, 35 (1998), pp. 990–1009.

[2] P. BOCHEV, Z. CAI, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Analysis of velocity-flux*

least-squares principles for the Navier-Stokes equations: Part II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.

[3] P. B. Bochev, *Negative norm least-squares methods for the velocity-vorticity-pressure Navier–Stokes equations*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 237–256.

[4] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[5] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, UK, 2001.

[6] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[7] A. Brandt, S. McCormick, and J. Ruge, *Algebraic multigrid (AMG) for sparse matrix equations*, in Sparsity and Its Applications (Loughborough, 1983), Cambridge University Press, Cambridge, UK, 1985, pp. 257–284.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Appl. Math. 15, Springer-Verlag, New York, 2002.

[9] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, PA, 2000.

[10] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[11] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics in Appl. Math. 40, SIAM, Philadelphia, PA, 2002.

[13] G. Dardyk and I. Yavneh, *A multilevel nonlinear method*, SIAM J. Sci. Comput., to appear.

[14] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Appl. Math. 16, SIAM, Philadelphia, PA, 1996.

[15] E. Gelman and J. Mandel, *On multilevel iterative methods for optimization problems*, Math. Programming, 48 (1990), pp. 1–17.

[16] M. Griebel, *Multilevel algorithms considered as iterative methods on semidefinite systems*, SIAM J. Sci. Comput., 15 (1994), pp. 547–565.

[17] W. Hackbusch, *Multigrid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer-Verlag, Berlin, 1985.

[18] J. J. Heys, T. A. Manteuffel, S. F. McCormick, and L. N. Olson, *Algebraic multigrid for higher-order finite elements*, J. Comput. Physics, 204 (2005), pp. 520–532.

[19] B.-N Jiang, *The Least-Squares Finite Element Method, Theory and Applications in Computational Fluid Dynamics and Electromagnetics*, Sci. Comput., Springer-Verlag, Berlin, 1998.

[20] L. I. G. Kovasznay, *Laminar flow behind two-dimensional grid*, Proc. Cambridge Philos. Soc., 44 (1948), pp. 58–62.

[21] J. Lottes and P. Fischer, *Hybrid multigrid/Schwarz algorithms for the spectral element method*, J. Sci. Comput., 24 (2005), pp. 45–78.

[22] J. Mandel and S. McCormick, *A multilevel variational method for* $A\mathbf{f}u = \lambda B\mathbf{f}u$ *on composite grids*, J. Comput. Phys., 80 (1989), pp. 442–452.

[23] T. A. Manteuffel, S. F. McCormick, and O. Röhrle, *Projection multilevel methods for quasilinear elliptic partial differential equations: Theoretical results*, SIAM J. Numer. Anal., 44 (2006), pp. 139–152.

[24] S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 62, SIAM, Philadelphia, 1992.

[25] S. F. McCormick and J. W. Ruge, *Unigrid for multigrid simulation*, Math. Comp., 41 (1983), pp. 43–62.

[26] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.

[27] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, S. F. McCormick, ed., Frontiers Appl. Math. 3, SIAM, Philadelphia, 1987, pp. 73–130.

[28] K. Stüben and U. Trottenberg, *Multigrid methods: Fundamental algorithms, Model problem analysis, and applications*, in Multigrid Methods (Cologne, 1981), Lecture Notes in Math. 960, Springer, Berlin, 1982, pp. 1–176.

[29] X.-C. Tai, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, Numer. Math., 93 (2003), pp. 755–786.

[30] X.-C. Tai and J. Xu, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124 (electronic).

[31] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, San Diego, CA, 2001.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 139–152

PROJECTION MULTILEVEL METHODS FOR QUASILINEAR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS: THEORETICAL RESULTS*

THOMAS A. MANTEUFFEL†, STEPHEN F. MCCORMICK†, AND OLIVER RÖHRLE‡

Abstract. In a companion paper [T. A. Manteuffel et al., *SIAM J. Numer. Anal.*, 44 (2006), pp. 120–138], we propose a new multilevel solver for two-dimensional elliptic systems of partial differential equations with nonlinearity of type $u\partial v$. The approach is based on a multilevel projection method (PML) [S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, SIAM, Philadelphia, 1992] applied to a first-order system least-squares functional that allows us to treat the nonlinearity directly. While the companion paper focuses on computation, here we concentrate on developing a theoretical framework that confirms optimal two-level convergence. To do so, we choose a first-order formulation of the Navier–Stokes equations as a basis of our theory. We establish continuity and coercivity bounds for the linearized Navier–Stokes equations and the full nonquadratic least-squares functional, as well as existence and uniqueness of a functional minimizer. This leads to the immediate result that one cycle of the two-level PML method reduces the functional norm by a factor that is uniformly less than 1.

Key words. projection method, multigrid, least squares, finite elements, quasilinear PDEs, Navier–Stokes

AMS subject classifications. 35J60, 65N12, 65N30, 65N55

DOI. 10.1137/040617704

1. Introduction. Our companion paper [8] introduces a new multilevel solver for two-dimensional elliptic systems of partial differential equations (PDEs) with nonlinearity of type $u\partial v$. The approach is based on a multilevel projection method (PML) [9] applied to a first-order system least-squares (FOSLS) functional, where the nonlinearity is treated directly, with no need for linearization anywhere in the algorithm. While [8] focuses on computation, the key objective of the present paper is to establish local well-posedness of our functional minimization problem. This result leads to the immediate conclusion that our two-level solver converges linearly with grid independent factors, as observed numerically in [8]. This two-grid result can be extended to W-cycles in the usual way. However, an important alternative would be to establish a V-cycle result based on the general theory developed in [11] and [12]. This alternative would naturally yield grid-dependent convergence bounds because of the weak smoothness assumptions on the problem formulation (i.e., only Lipschitz continuity on the domain boundary).

We base our theory for a two-level PML method on the first-order formulation of the Navier–Stokes formulation given in (2.1). Although we choose this formulation as a foundation for our theoretical framework, it is not limited to it: similar results can be established for other PDEs of this class.

*Received by the editors October 26, 2004; accepted for publication (in revised form) September 16, 2005; published electronically February 8, 2006. This work was sponsored by the Department of Energy under grants DE-FC02-01ER25479 and DE-FG02-03ER25574, Lawrence Livermore National Laboratory under contract B533502, Sandia National Laboratory under contract 15268, and the National Science Foundation under VIGRE grant DMS-9810751.

http://www.siam.org/journals/sinum/44-1/61770.html

†Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309–0526 (tmanteuf@colorado.edu, stevem@colorado.edu).

‡Bioengineering Institute, University of Auckland, Private Bag 92019, Auckland 1, New Zealand (o.rohrle@acukland.ac.nz).

This paper is organized in the following way. Section 2 provides the first-order system formulation, with definitions, notation, and description of one two-level PML cycle step. Section 3 establishes several continuity and coercivity bounds for the Oseen equations as well as for the full nonquadratic least-squares functional. Section 4 shows existence and uniqueness of a functional minimizer, some characteristics of coarse-grid correction and relaxation, and two-grid convergence.

2. First-order system formulation, definitions, notation, and other preliminaries. We use c and C throughout as generic constants that may change value with every occurrence but are independent of mesh size. To keep track of a specific value for a constant, subindices may be used.

FOSLS formulations for the Navier–Stokes equations are discussed in [1, 2, 3, 6]. In the framework of this paper, we consider the first-order velocity-flux formulation of the Navier–Stokes equations given in [1] and [2]:

$$\mathcal{L}(\mathbf{x}) = \mathsf{g} := \begin{cases} \nabla \mathbf{u}^t - \mathbf{U} = \mathbf{0} & \text{in } \Omega, \\ -(\nabla \cdot \mathbf{U})^t + \nabla p + Re\ \mathbf{U}^t \mathbf{u} = \mathbf{f} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega, \\ \nabla \times \mathbf{U} = \mathbf{0} & \text{in } \Omega, \\ \nabla \big(tr \mathbf{U}\big) = \mathbf{0} & \text{in } \Omega, \end{cases} \tag{2.1}$$

where Ω is a subset of $\mathbb{R}^n$ $(n = 2, 3)$ with Lipschitz continuous boundary $\partial\Omega$ and $\mathbf{f} \in L^2(\Omega)^n$. As boundary conditions, without loss of generality, we take $\mathbf{u} = \mathbf{0}$ and $\mathbf{n} \times \mathbf{U} = \mathbf{0}$ on $\partial\Omega$, where $\mathbf{n}$ is the outward unit normal on $\partial\Omega$. Writing the unknowns as $\mathbf{x} = (\mathbf{u}, \mathbf{U}, p)$, then the nonquadratic functional is constructed by taking the L^2-norm of each interior equation:

$$\mathcal{F}(\mathbf{x}; \mathsf{g}) = \|\mathcal{L}(\mathbf{x}) - \mathsf{g}\|^2_{0,\Omega}, \qquad \mathbf{x} \in \boldsymbol{\mathcal{V}}, \tag{2.2}$$

where $\mathsf{g} = (\mathbf{0}, \mathbf{f}, 0, \mathbf{0}, \mathbf{0})^T$ and the space is defined by

$$\boldsymbol{\mathcal{V}} = H^1_0(\Omega)^n \times \mathcal{V}_0 \times (H^1(\Omega)/\mathbb{R})$$

with

$$\mathcal{V}_0 = \{\mathbf{U} \in H^1(\Omega)^{n^2} : \mathbf{n} \times \mathbf{U} = \mathbf{0} \ \text{ on } \ \partial\Omega\}.$$

It is shown in [5] that the Navier–Stokes equations generally have more than one solution, unless the viscosity and the external forces satisfy very stringent requirements. However, it can also be shown that in many practical examples, these solutions are mostly isolated, i.e., there exist a neighborhood in which each solution is unique. Bifurcation phenomena are rare. We thus assume we are in a closed neighborhood, $\overline{\mathcal{B}}(\mathbf{x}_*, r)$, of an isolated solution, $\mathbf{x}_* \in \boldsymbol{\mathcal{V}}$, to (2.1), that is, a global minimum of (2.2), for which $\mathcal{F}(\mathbf{x}_*; \mathsf{g}) = 0$. The neighborhood is taken to be an H^1-ball around $\mathbf{x}_*$ with radius $r > 0$ defined as

$$\mathcal{B}(\mathbf{x}_*, r) := \big\{\mathbf{x} \in \boldsymbol{\mathcal{V}} : \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega} < r\big\},$$

where

$$\|\mathbf{x}\|^2_{1,\Omega} \equiv \|\mathbf{u}\|^2_{1,\Omega} + \|\mathbf{U}\|^2_{1,\Omega} + \|p\|^2_{1,\Omega}.$$

Its closure is $\overline{\mathcal{B}}(\mathbf{x}_*, r) = \{\mathbf{x} \in \boldsymbol{\mathcal{V}} : \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega} \leq r\}$. Several places along the way, we assume that r is so small that certain expansions we develop give us the desired bounds.

Denote by $\mathcal{L}'(\mathbf{x})[\mathbf{y}]$ the first Fréchet derivative of operator $\mathcal{L}$ at $\mathbf{x} \in \boldsymbol{\mathcal{V}}$ in direction $\mathbf{y} = (\mathbf{v}, \mathbf{V}, q) \in \boldsymbol{\mathcal{V}}$. Note that the nonlinear term, Re $\mathbf{U}^t\mathbf{u}$, in (2.1) becomes $\mathrm{Re}(\mathbf{V}^t\mathbf{u} + \mathbf{U}^t\mathbf{v})$ in $\mathcal{L}'(\mathbf{x})[\mathbf{y}]$ and that $\mathcal{L}'(\mathbf{x})[\mathbf{y}]$ is linear in $\mathbf{y}$. Also, $\mathcal{L}'(\mathbf{x})$ is the same operator as that for the Oseen equations (cf. [7]). $\mathcal{L}''(\mathbf{x})[\mathbf{y}, \mathbf{z}]$ denotes the second Fréchet derivative at $\mathbf{x}$ in directions $\mathbf{y}$ and $\mathbf{z} = (\mathbf{w}, \mathbf{W}, t) \in \boldsymbol{\mathcal{V}}$. For the linear terms of (2.1), the second Fréchet derivative is the zero operator. For the nonlinear term, we obtain $\mathrm{Re}(\mathbf{V}^t\mathbf{w} + \mathbf{W}^t\mathbf{v})$, so $\mathcal{L}''(\mathbf{x})$ is independent of $\mathbf{x}$.

Another set definition we use later is the closed line segment connecting points $\mathbf{x}, \mathbf{y} \in \boldsymbol{\mathcal{V}}$: $[\mathbf{x}, \mathbf{y}] := \{\theta\mathbf{x} + (1-\theta)\mathbf{y} : 0 \leq \theta \leq 1\}$. This notation should not be confused with the square brackets used for directional derivatives because the operator always immediately precedes the direction.

Having defined the first and second Fréchet derivatives for operator $\mathcal{L}$, we are able to express the first and second Fréchet derivatives of the nonquadratic functional in (2.2) in terms of $\mathcal{L}$ and its derivatives. For $\mathbf{x}, \mathbf{y} \in \boldsymbol{\mathcal{V}}$, the first Fréchet derivative of (2.2) in direction $\mathbf{y}$ is

$$\mathcal{F}'(\mathbf{x}; \mathbf{g})[\mathbf{y}] = 2\langle \mathcal{L}(\mathbf{x}) - \mathbf{g}, \mathcal{L}'(\mathbf{x})[\mathbf{y}] \rangle. \tag{2.3}$$

Its second Fréchet derivative in direction $[\mathbf{y}, \mathbf{y}]$ (needed later for Taylor expansions) is

$$\mathcal{F}''(\mathbf{x}; \mathbf{g})[\mathbf{y}, \mathbf{y}] = 2\,\|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 + 2\langle \mathcal{L}(\mathbf{x}) - \mathbf{g}, \mathcal{L}''(\mathbf{x})[\mathbf{y}, \mathbf{y}] \rangle. \tag{2.4}$$

Remark 1. As with all multigrid schemes, relaxation is the basis for our PML approach. One choice is steepest descent, which involves a gradient direction, $\mathbf{d}$, in $\boldsymbol{\mathcal{V}}$ and a step size, s, determined as the smallest nonnegative critical point of $\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g})$. To understand this step, it is useful to examine the polynomial

$$\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g}) = \mathcal{F}(\mathbf{x}; \mathbf{g}) - s\mathcal{F}'(\mathbf{x}; \mathbf{g})[\mathbf{d}] + \frac{s^2}{2}\mathcal{F}''(\mathbf{x}; \mathbf{g})[\mathbf{d}, \mathbf{d}] - \frac{s^3}{6}\mathcal{F}'''(\mathbf{x}; \mathbf{g})[\mathbf{d}, \mathbf{d}, \mathbf{d}] + \frac{s^4}{24}\mathcal{F}^{(4)}(\mathbf{x}; \mathbf{g})[\mathbf{d}, \mathbf{d}, \mathbf{d}, \mathbf{d}].$$

From (2.4), we see that $\mathcal{F}'''(\mathbf{x}; \mathbf{g})[\mathbf{d}, \mathbf{d}, \mathbf{d}] = 6\langle \mathcal{L}'(\mathbf{x})[\mathbf{d}], \mathcal{L}''(\mathbf{x})[\mathbf{d}, \mathbf{d}] \rangle$ and $\mathcal{F}^{(4)}(\mathbf{x}; \mathbf{g})[\mathbf{d}, \mathbf{d}, \mathbf{d}, \mathbf{d}] = 6\langle \mathcal{L}''(\mathbf{x})[\mathbf{d}, \mathbf{d}], \mathcal{L}''(\mathbf{x})[\mathbf{d}, \mathbf{d}] \rangle$. Thus, when $\mathcal{F}''(\mathbf{x}; \mathbf{g}) > 0$, an inspection of this polynomial for $s > 0$ implies that there exists a smallest nonnegative critical point, s, of $\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g})$ and that it must be a local minimum of $\mathcal{F}(\mathbf{x}, \mathbf{g})$ such that $\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g}) \leq \mathcal{F}(\mathbf{x}; \mathbf{g})$. (Note that either $\mathbf{d} = \mathbf{0}$ and $\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g}) = \mathcal{F}(\mathbf{x}; \mathbf{g})$, so $s = 0$, or $\mathcal{F}(\mathbf{x} - s\mathbf{d}; \mathbf{g})$ initially decreases but then tends to $+\infty$ as s goes from 0 to ∞.)

Similar definitions can be made for subspaces of $\boldsymbol{\mathcal{V}}$. Consider a quasi-uniform finite element partition of Ω with approximate mesh size h and let $H^h(\Omega)$ be the corresponding finite element subspace of $H^1(\Omega)$ consisting of piecewise polynomials: a function in $H^h(\Omega)$ is continuous on Ω and polynomial within each element. Let $H_0^h(\Omega)$ denote the subspace of $H^h(\Omega)$ of functions that are zero on $\partial\Omega$. Then define

$$\mathcal{S}^h = H_0^h(\Omega)^n \times \mathcal{V}_0^h \times (H^h(\Omega)/\mathbb{R}) \subset \boldsymbol{\mathcal{V}}$$

with

$$\mathcal{V}_0^h = \{\mathbf{U}^h \in H^h(\Omega)^{n^2} : \mathbf{n} \times \mathbf{U}^h = \mathbf{0} \ \text{ on } \ \partial\Omega\}.$$

Suppose also that we have a corresponding coarser $2h$ level so that the corresponding discrete space, $\mathcal{S}^{2h}$, forms a subspace of $\mathcal{S}^h$. For this paper, we assume standard nested finite element spaces, $S^{2h} \subset S^h \subset \boldsymbol{\mathcal{V}}$, that satisfy the approximation property

$$\inf_{\mathbf{x}^{2h} \in \mathcal{S}^{2h}} \|\mathbf{x}^h - \mathbf{x}^{2h}\|_{0,\Omega}^2 \leq C_1 \, h^2 \, \|\mathbf{x}^h\|_{1,\Omega}^2 \tag{2.5}$$

and the inverse estimate

$$\|\mathbf{x}^h\|_{1,\Omega}^2 \leq \frac{C_2}{h^2} \|\mathbf{x}^h\|_{0,\Omega}^2 \tag{2.6}$$

for all $\mathbf{x}^h$ in $\mathcal{S}^h$, where C_1 and C_2 are positive constants that do not depend on h (see [4]). Further, define a discrete H^1-ball by $\mathcal{B}^h(\mathbf{x}_*, r) = \{\mathbf{x}^h \in \mathcal{S}^h : \|\mathbf{x}^h - \mathbf{x}_*\|_{1,\Omega} < r\}$ and its closure by $\overline{\mathcal{B}}^h(\mathbf{x}_*, r) = \{\mathbf{x}^h \in \mathcal{S}^h : \|\mathbf{x}^h - \mathbf{x}_*\|_{1,\Omega} \leq r\}$. As we said, we choose r progressively smaller in several places in what follows. Nowhere does this requirement depend on h. However, we implicitly assume that no matter how small r becomes, h is so small that $\mathcal{B}^h(\mathbf{x}_*, r) \neq \emptyset$.

Before being able to define the relaxation scheme and two-level PML method, we introduce the discrete functional and its gradient as well as the operator norm associated with the second Fréchet derivative of functional $\mathcal{F}(\mathbf{x}; \mathsf{g})$.

DEFINITION 2.1 (discrete functional and its L^2-gradient). *Let $\mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ and define $\mathcal{F}^h(\mathbf{x}^h; \mathsf{g})$ as the restriction of $\mathcal{F}(\mathbf{x}^h; \mathsf{g})$ to space $\mathcal{S}^h$. Now let $\mathbf{y}^h \in \mathcal{S}^h$. By the definition of the first Fréchet derivative, we have*

$$\mathcal{F}^{h'}(\mathbf{x}^h; \mathsf{g})[\mathbf{y}^h] = \langle \mathcal{L}(\mathbf{x}^h) - \mathsf{g}, \mathcal{L}'(\mathbf{x}^h)[\mathbf{y}^h] \rangle,$$

and, since $\mathcal{S}^h$ is finite dimensional, the Riesz representation theorem guarantees the existence of the discrete L^2-gradient, $\nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}) \in \mathcal{S}^h$, which satisfies

$$\mathcal{F}^{h'}(\mathbf{x}^h; \mathsf{g})[\mathbf{y}^h] = \langle \mathcal{L}'^*(\mathbf{x}^h)\big(\mathcal{L}(\mathbf{x}^h) - \mathsf{g}\big), \mathbf{y}^h \rangle =: \langle \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}), \mathbf{y}^h \rangle.$$

Note that $\nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}) \in \mathcal{S}^h$ can be defined weakly by $\langle \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}), \mathbf{y}^h \rangle = \langle \mathcal{L}(\mathbf{x}^h) - \mathsf{g}, \mathcal{L}'(\mathbf{x}^h)[\mathbf{y}^h] \rangle$ for all $\mathbf{y}^h \in \mathcal{S}^h$. Note also that $\nabla^h \mathcal{F}^h(\mathbf{x}^h; \mathsf{g}) = \nabla^h \mathcal{F}^h(\mathbf{x}^h; \mathsf{g}^h)$, where g^h is the L^2-orthogonal projection of g onto space $\mathcal{L}'(\mathbf{x}^h)\mathcal{S}^h$.

Remark 2. Denote by $\mathbf{x}_*^h$ the element in $\mathcal{S}^h$ that minimizes (2.2) over $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$. Such an element exists because this set is compact and $\mathcal{F}(\mathbf{x}; \mathsf{g})$ is continuous, as we show in Theorem 3.2. Note that if $\mathbf{x}^h \in \mathcal{B}^h(\mathbf{x}_*, r)$ (i.e., the interior of the ball), then $\mathbf{x}_*^h$ is a grid h critical point in the sense that

$$\langle \nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}), \mathbf{y}^h \rangle = \langle \mathcal{L}(\mathbf{x}_*^h) - \mathsf{g}, \mathcal{L}'(\mathbf{x}_*^h)[\mathbf{y}^h] \rangle = 0 \tag{2.7}$$

for all $\mathbf{y}^h \in \mathcal{S}^h$, provided $\mathcal{F}''(\mathbf{x}; \mathsf{g})$ is bounded on $\mathcal{B}(\mathbf{x}_*, r)$, as we show in Theorem 3.1. This follows from a standard argument based on Taylor series and outlined as follows: if

$$\begin{aligned} 0 &\leq \mathcal{F}\big(\mathbf{x}_*^h - s \nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g})\big) - \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}) \\ &= -s \|\nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g})\|_{0,\Omega}^2 + s^2 \mathcal{F}''(\tilde{\mathbf{x}}^h; \mathsf{g})[\nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}), \nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g})]; \end{aligned}$$

and, for small enough but positive s, we could make the last expression negative (a contradiction) unless $\nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}) = 0$. This standard argument is referred to later in the proof of Lemma 4.3 to show that the coarse-grid correction step of PML (described next) is determined by a grid $2h$ critical point. That $\mathbf{x}_*$ is a critical point of $\mathcal{F}(\mathbf{x}; \mathsf{g})$ in $\mathcal{B}(\mathbf{x}_*, r)$ follows simply from (2.3).

DEFINITION 2.2 (discrete operator norm of the second derivative). *Let* $\mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ *and* $\mathbf{y}^h \in \mathcal{S}^h$. *Then the discrete operator norm associated with the second Fréchet derivative of functional* $\mathcal{F}(\mathbf{x}^h; \mathsf{g})$ *is defined by*

$$\left\| |\mathcal{F}''(\mathbf{x}^h; \mathsf{g})| \right\|_{0,h} = \sup_{0 \neq \mathbf{y}^h \in S^h} \frac{\left| \mathcal{F}''(\mathbf{x}^h; \mathsf{g})[\mathbf{y}^h, \mathbf{y}^h] \right|}{\langle \mathbf{y}^h, \mathbf{y}^h \rangle}.$$

Next, we define one step of relaxation. We consider two types of schemes, both of which use the discrete gradient as a descent direction. The first scheme reduces to Richardson for the linear case and the second is optimal steepest descent. The theory focuses on the Richardson-type scheme because it is simpler to analyze and it sets the stage for a simple conclusion for steepest descent.

DEFINITION 2.3 (relaxation). *One step of Richardson-type relaxation is defined by*

$$\mathbf{x}^h \leftarrow \mathbf{x}^h - \frac{\omega}{\left\| |\mathcal{F}''(\mathbf{x}^h; \mathsf{g})| \right\|_{0,h}} \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}), \tag{2.8}$$

where $\nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g})$ *is the search direction,* $1/\left\| |\mathcal{F}''(\mathbf{x}^h; \mathsf{g})| \right\|_{0,h}$ *the basic step length, and* ω *a damping parameter. One step of steepest descent is defined by*

$$\mathbf{x}^h \leftarrow \mathbf{x}^h - s \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}), \tag{2.9}$$

where s *is chosen as the smallest nonnegative root of*

$$\frac{\partial \mathcal{F}}{\partial s}\Big(\left(\mathbf{x}^h - s \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}) \right); \mathsf{g} \Big) = 0. \tag{2.10}$$

We now have all the ingredients needed to describe the two-level PML method. Its first step computes the nearest locally optimal coarse-grid correction, which Lemma 4.3 shows must exist uniquely provided we are close enough to $\mathbf{x}_*$. Its second step is one relaxation sweep given by either (2.8) or (2.9). The method in (2.8) is well defined because $\mathcal{F}''(\mathbf{x}^h; \mathsf{g})$ is nonzero, as Theorem 3.1 shows. The method in (2.9) is also well defined, as Remark 1 shows.

Step 1. For a given initial guess, $\mathbf{x}_0^h \in \mathcal{S}^h$, perform the coarse-grid correction step given by $\mathbf{x}_{\frac{1}{2}}^h \leftarrow \mathbf{x}_0^h + \mathbf{x}_*^{2h}$, where $\mathbf{x}_*^{2h}$ is the local minimizer of $\mathcal{F}(\mathbf{x}_0^h + \mathbf{x}^{2h}; \mathsf{g})$ (e.g., it is a grid $2h$ critical point) with minimal H^1-norm:

$$\mathbf{x}_*^{2h} = \operatorname*{argmin}_{\mathbf{x}^{2h} \in S^{2h}} \Big\{ \|\mathbf{x}^{2h}\|_{1,\Omega} : \nabla^{2h} \mathcal{F}(\mathbf{x}_0^h + \mathbf{x}^{2h}; \mathsf{g}) = 0,\ \mathcal{F}(\mathbf{x}_0^h + \mathbf{x}^{2h}; \mathsf{g}) \leq \mathcal{F}(\mathbf{x}_0^h; \mathsf{g}) \Big\}.$$

Step 2. Let $\mathbf{x}_1^h$ be the result of one relaxation step given by (2.8) or (2.9) applied to $\mathbf{x}_{\frac{1}{2}}^h$.

Further iterations of PML are defined in the obvious way, with $\mathbf{x}_k^h$ taking on the role of $\mathbf{x}_0^h$ and $\mathbf{x}_{k+1}^h$ being the result corresponding to $\mathbf{x}_1^h$ for $k = 1, 2, \ldots$.

3. Continuity and coercivity bounds. In this section, we first establish continuity and coercivity for the Oseen equations (Lemma 3.3). We then use Lemmas 3.1 and 3.3 to prove continuity and coercivity of $\mathcal{F}''(\mathbf{x};\mathbf{g})[\mathbf{y},\mathbf{y}]$ as a function of $\mathbf{y} \in \boldsymbol{\mathcal{V}}$, for all $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ (Theorem 3.1). The results in this section help us later to establish the key objective of our two-level method: one cycle of two-level PML reduces the functional norm by a factor that is bounded uniformly below 1 (Theorem 4.1).

LEMMA 3.1. *There exist a* γ_0*, depending only on* Re *and* Ω*, such that*

$$\|\mathcal{L}''(\mathbf{x})[\mathbf{y},\mathbf{z}]\|_{0,\Omega} \;\leq\; \gamma_0 \|\mathbf{y}\|_{1,\Omega} \|\mathbf{z}\|_{1,\Omega}$$

for all $\mathbf{x}, \mathbf{y}$*, and* $\mathbf{z}$ *in* $\boldsymbol{\mathcal{V}}$*.*

Proof. Recall for $\mathbf{x} = (\mathbf{u}, \mathbf{U}, p)$, $\mathbf{y} = (\mathbf{v}, \mathbf{W}, q)$, and $\mathbf{z} = (\mathbf{w}, \mathbf{W}, t)$ in $\boldsymbol{\mathcal{V}}$ that the second Fréchet derivative for the linear terms of (2.1) is the zero operator. For the nonlinear term, we obtain $\mathrm{Re}(\mathbf{V}^t\mathbf{w} + \mathbf{W}^t\mathbf{v})$. Then, the result follows directly from the Sobolev imbedding theorem about multiplication in Sobolev spaces (Corollary I.1.1 in [5]). ∎

LEMMA 3.2. *For all* $\mathbf{x} = (\mathbf{u}, \mathbf{U}, p) \in \boldsymbol{\mathcal{V}}$*, there exist two positive constants,* $\tilde{c}_3$ *and* $\tilde{C}_3$*, depending only on* Re*,* $\mathbf{x}$*, and* Ω*, such that*

$$\tilde{c}_3(\mathbf{x}) \; \|\mathbf{y}\|_{1,\Omega}^2 \;\leq\; \|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 \;\leq\; \tilde{C}_3(\mathbf{x}) \; \|\mathbf{y}\|_{1,\Omega}^2$$

for all $\mathbf{y} = (\mathbf{v}, \mathbf{V}, q) \in \boldsymbol{\mathcal{V}}$*.*

Proof. We use the derivation of the regularity estimate, as well as Theorems 3.2, 4.1, and 4.2 in [7], as guidelines for the proof of this lemma. Analogous to the continuity and coercivity proof for $\mathcal{L}'\big((\mathbf{u},p)\big)\big[(\mathbf{v},\mathbf{V},q)\big]$ in [7], we start from the Oseen equations in the following form:

$$\begin{aligned} -\Delta\mathbf{v} + Re\big[(\nabla\mathbf{v}^t)^t\mathbf{u} + \mathbf{U}^t\mathbf{v}\big] + \nabla q &= \mathbf{f}, \\ \nabla\cdot\mathbf{v} &= g, \end{aligned} \tag{3.1}$$

where $g \in L^2(\Omega)$. The first equation differs from that in [7] because $\mathbf{U}$ is used instead of $\nabla\mathbf{u}^t$. We also relax the smoothness assumption by only requiring $\mathbf{u}$ and $\mathbf{U}$ to be in $H_0^1(\Omega)^n$ and $\mathcal{V}_0$, respectively. ([7] requires $\mathbf{u}$ to be in $H_0^2(\Omega)^n$.)

First, we establish an a priori H^1-regularity estimate for the equations in (3.1): if Ω has Lipschitz boundary, then, for $\mathbf{f} \in H_0^{-1}(\Omega)^n$ and $g \in L_0^2(\Omega)$, the weak solution of (3.1), $(\mathbf{v}, q) \in H_0^1(\Omega)^n \times L_0^2(\Omega)$, satisfies the a priori estimate

$$\|\nabla\mathbf{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega} \leq \mathrm{const}\,\big(\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega}\big) \tag{3.2}$$

for $\delta \in (0, \frac{1}{2})$, where $\delta_\circ = 0$ for $\Omega \subset \mathbb{R}^2$ and $\delta_\circ = \frac{1}{2}$ for $\Omega \subset \mathbb{R}^3$.

To prove this estimate, we first take the pointwise dot product of the first equation of (3.1) with any $\boldsymbol{\psi} \in H_0^1(\Omega)^n$ and the dot product of the second equation of (3.1) with any $\phi \in L^2(\Omega)$, integrate it over Ω, and use integration by parts. This yields

$$\begin{aligned} \langle\nabla\mathbf{v}^t, \nabla\boldsymbol{\psi}^t\rangle + Re\langle(\nabla\mathbf{v}^t)^t\mathbf{u} + \mathbf{U}^t\mathbf{v}, \boldsymbol{\psi}\rangle - \langle q, \nabla\cdot\boldsymbol{\psi}\rangle &= \langle\mathbf{f}, \boldsymbol{\psi}\rangle, \\ \langle\nabla\cdot\mathbf{v}, \phi\rangle &= \langle g, \phi\rangle. \end{aligned} \tag{3.3}$$

Since $g \in L^2(\Omega)$, we can choose an $\mathbf{s} \in H_0^1(\Omega)^n$, according to Lemma 4.1 in [7], such that

$$\nabla\cdot\mathbf{s} = g \quad \text{and} \quad |\mathbf{s}|_{1,\Omega} \leq C\|g\|_{0,\Omega}. \tag{3.4}$$

Then, setting $\boldsymbol{v} = \mathbf{v} - \mathbf{s} \in H_0^1(\Omega)^n$ in (3.3), we have

(3.5)
$$\begin{cases} \langle \nabla \boldsymbol{v}^t, \nabla \boldsymbol{\psi}^t \rangle + \mathrm{Re}\langle (\nabla \boldsymbol{v}^t)^t \mathbf{u} + \mathbf{U}^t \boldsymbol{v}, \boldsymbol{\psi} \rangle - \langle q, \nabla \cdot \boldsymbol{\psi} \rangle &= \langle \mathbf{f}, \boldsymbol{\psi} \rangle - \langle \nabla \mathbf{s}^t, \nabla \boldsymbol{\psi}^t \rangle \\ & \quad - \mathrm{Re}\langle (\nabla \mathbf{s}^t)^t \mathbf{u} + \mathbf{U}^t \mathbf{s}, \boldsymbol{\psi} \rangle, \\ \langle \nabla \cdot \boldsymbol{v}, \phi \rangle &= \langle 0, \phi \rangle \end{cases}$$

for any $\boldsymbol{\psi} \in H_0^1(\Omega)^n$ and $\phi \in L^2(\Omega)$. For the first equation in (3.5), by taking $\boldsymbol{\psi} = \boldsymbol{v}$, we obtain

(3.6)
$$\begin{aligned} \|\nabla \boldsymbol{v}^t\|_{0,\Omega}^2 &= \langle \mathbf{f}, \boldsymbol{v} \rangle - \langle \nabla \mathbf{s}^t, \nabla \boldsymbol{v}^t \rangle - \mathrm{Re}\big(\langle (\nabla \boldsymbol{v}^t)^t \mathbf{u} + \mathbf{U}^t \boldsymbol{v}, \boldsymbol{v} \rangle + \langle (\nabla \mathbf{s}^t)^t \mathbf{u} + \mathbf{U}^t \mathbf{s}, \boldsymbol{v} \rangle\big) \\ &\le |\langle \mathbf{f}, \boldsymbol{v} \rangle| + |\langle \nabla \mathbf{s}^t, \nabla \boldsymbol{v}^t \rangle| \\ &\quad + \mathrm{Re}\ |\langle \mathbf{U}^t \boldsymbol{v}, \boldsymbol{v} \rangle + \langle (\nabla \boldsymbol{v}^t)^t \mathbf{u}, \boldsymbol{v} \rangle + \langle \mathbf{U}^t \mathbf{s}, \boldsymbol{v} \rangle + \langle (\nabla \mathbf{s}^t)^t \mathbf{u}, \boldsymbol{v} \rangle|. \end{aligned}$$

For the first term of the upper bound in (3.6), recall the definition of the $H^{-1}(\Omega)$ norm:

$$\|\mathbf{f}\|_{-1,\Omega} := \sup_{0 \neq \boldsymbol{v} \in H_0^1(\Omega)} \frac{\langle \mathbf{f}, \boldsymbol{v} \rangle}{|\boldsymbol{v}|_{1,\Omega}} \ \Rightarrow \frac{\langle \mathbf{f}, \boldsymbol{v} \rangle}{|\boldsymbol{v}|_{1,\Omega}} \le \|\mathbf{f}\|_{-1,\Omega} \forall \boldsymbol{v} \neq 0 \in H_0^1(\Omega)^n.$$

Hence for all $\boldsymbol{v} \neq 0 \in H_0^1(\Omega)^n$, we have $\langle \mathbf{f}, \boldsymbol{v} \rangle \le \|\mathbf{f}\|_{-1,\Omega} \|\nabla \boldsymbol{v}\|_{0,\Omega}$. To bound the second term, we use the Cauchy–Schwarz inequality:

$$\langle \nabla \mathbf{s}^t, \nabla \boldsymbol{v}^t \rangle \ \le \ \|\nabla \mathbf{s}^t\|_{0,\Omega} \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \ = \ |\mathbf{s}|_{1,\Omega} \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \overset{(3.4)}{\le} C \|g\|_{0,\Omega} \|\nabla \boldsymbol{v}^t\|_{0,\Omega}.$$

It remains to derive bounds for the last four terms, which are classified in [5] as trilinear. In the following, C denotes a generic constant that might depend on Re, $\Omega, \|\mathbf{u}\|_1$, and $\|\mathbf{U}\|_1$.

According to the Sobolev imbedding theorem I.1.3 in [5], the space $H^1(\Omega)$ is continuously embedded in $L^4(\Omega)$ for $n \le 4$. Then,

$$\begin{aligned} |\langle \mathbf{U}\boldsymbol{v}, \boldsymbol{v} \rangle| = \left| \sum_{i,j=1}^{n} \int_\Omega v_j U_{ij} v_i \, dx \right| &\le \sum_{i,j=1}^{n} \|v_j\|_{0,\Omega} \|U_{ij}\|_{0,4,\Omega} \|v_i\|_{0,4,\Omega} \\ &\le C \|\boldsymbol{v}\|_{0,\Omega} \|\mathbf{U}\|_{1,\Omega} \|\boldsymbol{v}\|_{1,\Omega} \le \ C \|\boldsymbol{v}\|_{\delta_\diamond + \delta, \Omega} \|\mathbf{U}\|_{1,\Omega} |\boldsymbol{v}|_{1,\Omega}. \end{aligned} \tag{3.7}$$

The last inequality is a result of the Poincaré–Friedrichs inequality ($\|\boldsymbol{v}\|_{1,\Omega} \le C|\boldsymbol{v}|_{1,\Omega}$). Applying the Sobolev imbedding theorem to the second trilinear term leads to

$$\langle (\nabla \boldsymbol{v}^t)^t \mathbf{u}, \boldsymbol{v} \rangle \le \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \|\mathbf{u}^t \boldsymbol{v}\|_{0,\Omega} \le \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \|\mathbf{u}\|_{1,\Omega} \|\boldsymbol{v}\|_{\delta_\diamond + \delta, \Omega}. \tag{3.8}$$

Similar arguments hold for the remaining two trilinear terms. Hence,

$$\langle \mathbf{U}^t \mathbf{s}, \boldsymbol{v} \rangle \ \le C \ \|\mathbf{U}\|_{1,\Omega} |\mathbf{s}|_{1,\Omega} \|\boldsymbol{v}\|_{1,\Omega} \le C \ \|\mathbf{U}\|_{1,\Omega} \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \|g\|_{0,\Omega} \tag{3.9}$$

and

$$\langle (\nabla \mathbf{s}^t)^t \mathbf{u}, \boldsymbol{v} \rangle \ \le C \ |\mathbf{s}|_{1,\Omega} \|\mathbf{u}\|_{1,\Omega} \|\boldsymbol{v}\|_{1,\Omega} \le C \ \|\mathbf{u}\|_{1,\Omega} \|\nabla \boldsymbol{v}^t\|_{0,\Omega} \|g\|_{0,\Omega}. \tag{3.10}$$

Combining the results yields

$$\|\nabla \boldsymbol{v}^t\|_{0,\Omega}^2 \le \|\mathbf{f}\|_{-1,\Omega}\|\nabla \boldsymbol{v}\|_{0,\Omega} + C\|g\|_{0,\Omega}\|\nabla \boldsymbol{v}\|_{0,\Omega} + C\|\mathbf{U}\|_{1,\Omega}\|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega}\|\nabla \boldsymbol{v}\|_{0,\Omega} + C\big(\|\mathbf{U}\|_{1,\Omega} + \|\mathbf{u}\|_{1,\Omega}\big)\|g\|_{0,\Omega}\|\nabla \boldsymbol{v}\|_{0,\Omega}.$$

Canceling $\|\nabla \boldsymbol{v}\|_{0,\Omega}$ gives

$$\|\nabla \boldsymbol{v}^t\|_{0,\Omega} \le C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega}\Big]. \tag{3.11}$$

To bound q, choose $\boldsymbol{\psi} \in H_0^1(\Omega)^n$ according to Lemma 4.1 in [7] such that

$$\nabla\cdot \boldsymbol{\psi} = q \quad \text{and} \quad |\boldsymbol{\psi}|_{1,\Omega} \le C\|q\|_{0,\Omega}. \tag{3.12}$$

Using again the first equation of (3.5), we obtain

$$\begin{aligned}\|q\|_{0,\Omega}^2 &= \langle \nabla \boldsymbol{v}^t, \nabla \boldsymbol{\psi}^t\rangle + Re\langle (\nabla \boldsymbol{v}^t)^t\mathbf{u} + \mathbf{U}^t\boldsymbol{v}, \boldsymbol{\psi}\rangle \\ &\quad - \langle \mathbf{f}, \boldsymbol{\psi}\rangle + \langle \nabla \mathbf{s}^t, \nabla \boldsymbol{\psi}^t\rangle + Re\langle (\nabla \mathbf{s}^t)^t\mathbf{u} + \mathbf{U}^t\mathbf{s}, \boldsymbol{\psi}\rangle \\ &\le \langle \nabla \boldsymbol{v}^t, \nabla \boldsymbol{\psi}^t\rangle + Re\langle (\nabla \boldsymbol{v}^t)^t\mathbf{u} + \mathbf{U}^t\boldsymbol{v}, \boldsymbol{\psi}\rangle \\ &\quad + |\langle \mathbf{f}, \boldsymbol{\psi}\rangle| + \langle \nabla \mathbf{s}^t, \nabla \boldsymbol{\psi}^t\rangle + Re\langle (\nabla \mathbf{s}^t)^t\mathbf{u} + \mathbf{U}^t\mathbf{s}, \boldsymbol{\psi}\rangle.\end{aligned}$$

We proceed similarly as with the bound $\|\nabla \boldsymbol{v}^t\|_{0,\Omega}$. For all $\boldsymbol{\psi} \ne 0 \in H_0^1(\Omega)^n$,

$$\langle \mathbf{f}, \boldsymbol{\psi}\rangle \le \|\mathbf{f}\|_{-1,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \overset{(3.12)}{\le} C\|\mathbf{f}\|_{-1,\Omega}\|q\|_{0,\Omega}, \tag{3.13}$$

$$\langle \nabla \boldsymbol{v}^t, \nabla \boldsymbol{\psi}^t\rangle \le \|\nabla \boldsymbol{v}^t\|_{0,\Omega}\|\nabla \boldsymbol{\psi}^t\|_{0,\Omega} = |\boldsymbol{v}|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \overset{(3.12)}{\le} C|\boldsymbol{v}|_{1,\Omega}\|q\|_{0,\Omega}, \tag{3.14}$$

and

$$\langle \nabla \mathbf{s}^t, \nabla \boldsymbol{\psi}^t\rangle \le \|\nabla \mathbf{s}^t\|_{0,\Omega}\|\nabla \boldsymbol{\psi}^t\|_{0,\Omega} = |\mathbf{s}|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \overset{(3.12)}{\le} C|\mathbf{s}|_{1,\Omega}\|q\|_{0,\Omega}. \tag{3.15}$$

The bounds for trilinear terms $\langle (\nabla \boldsymbol{v}^t)^t\mathbf{u}, \boldsymbol{\psi}\rangle$ and $\langle (\nabla \mathbf{s}^t)^t\mathbf{u}, \boldsymbol{\psi}\rangle$ follow directly by applying Lemma IV.2.1 in [5] and the Poincaré–Friedrichs inequality:

$$\langle (\nabla \boldsymbol{v}^t)^t\mathbf{u}, \boldsymbol{\psi}\rangle \le C\,|\boldsymbol{v}|_{1,\Omega}\|\mathbf{u}\|_{1,\Omega}\|\boldsymbol{\psi}\|_{1,\Omega} \le C\,\|\mathbf{u}\|_{1,\Omega}\|\nabla \boldsymbol{v}^t\|_{0,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \tag{3.16}$$

and

$$\langle (\nabla \mathbf{s}^t)^t\mathbf{u}, \boldsymbol{\psi}\rangle \le C\,|\mathbf{s}|_{1,\Omega}\|\mathbf{u}\|_{1,\Omega}\|\boldsymbol{\psi}\|_{1,\Omega} \le C\,\|\mathbf{u}\|_{1,\Omega}|\mathbf{s}|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega}. \tag{3.17}$$

For the remaining trilinear terms, we follow the argument in (3.7):

$$\langle \mathbf{U}^t\boldsymbol{v}, \boldsymbol{\psi}\rangle \le C\,\|\mathbf{U}\|_{1,\Omega}\|\boldsymbol{v}\|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \le C\,\|\mathbf{U}\|_{1,\Omega}\|\nabla \boldsymbol{v}^t\|_{0,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \tag{3.18}$$

and

$$\langle \mathbf{U}^t\mathbf{s}, \boldsymbol{\psi}\rangle \le C\,\|\mathbf{U}\|_{1,\Omega}\|\mathbf{s}\|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega} \le C\,\|\mathbf{U}\|_{1,\Omega}|\mathbf{s}|_{1,\Omega}|\boldsymbol{\psi}|_{1,\Omega}. \tag{3.19}$$

With (3.13)–(3.19), we have

$$\begin{aligned}\|q\|_{0,\Omega}^2 &\le C\|\mathbf{f}\|_{-1,\Omega}\|q\|_{0,\Omega} + C|\boldsymbol{v}|_{1,\Omega}\|q\|_{0,\Omega} + C|\mathbf{s}|_{1,\Omega}\|q\|_{0,\Omega} \\ &\quad + C\Big(\|\nabla \boldsymbol{v}^t\|_{0,\Omega} + |\mathbf{s}|_{1,\Omega}\Big)|\boldsymbol{\psi}|_{1,\Omega} \\ &\overset{(3.12)}{\le} C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big(\|\mathbf{f}\|_{-1,\Omega} + \|\nabla \boldsymbol{v}^t\|_{0,\Omega} + |\mathbf{s}|_{1,\Omega}\Big)\|q\|_{0,\Omega} \\ &\overset{(3.4),(3.11)}{\le} C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big(\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega}\Big)\|q\|_{0,\Omega}.\end{aligned}$$

Canceling $\|q\|_{0,\Omega}$ results in

$$\|q\|_{0,\Omega} \;\le\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big(\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega}\Big). \tag{3.20}$$

Recall that we seek an estimate for $\|\nabla \mathbf{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega}$ in terms of $\mathbf{v}$ and q and not for $\|\nabla \boldsymbol{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega}$ in terms of $\boldsymbol{v}$ and q. Earlier, we defined $\boldsymbol{v}$ to be the difference between $\mathbf{v}$ and $\mathbf{s}$. Now, adding $\mathbf{s}$ to $\boldsymbol{v}$ leads to estimates for $\|\nabla \mathbf{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega}$ in terms of $\mathbf{v}$ and q:

$$\begin{aligned}
\|\nabla \mathbf{v}^t\|_{0,\Omega} \;&\le\; \|\nabla \boldsymbol{v}^t + \nabla \mathbf{s}^t\|_{0,\Omega} \le\; \|\nabla \boldsymbol{v}^t\|_{0,\Omega} + \|\nabla \mathbf{s}^t\|_{0,\Omega} \\
&\overset{(3.11)}{\le}\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega} + \|\nabla \mathbf{s}^t\|_{0,\Omega}\Big] \\
&\le\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega} + C|\mathbf{s}|_{1,\Omega}\Big] \\
&\overset{(3.4)}{\le}\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega} + C\|g\|_{0,\Omega}\Big] \\
&\le\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega}\Big].
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|q\|_{0,\Omega} &\le C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\boldsymbol{v}\|_{\delta_\circ+\delta,\Omega}\Big] \\
&\le C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega}\Big].
\end{aligned}$$

Combining the bounds for $\|\nabla \mathbf{v}^t\|_{0,\Omega}$ and $\|q\|_{0,\Omega}$ results in the a priori estimate

$$\|\nabla \mathbf{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega} \;\le\; C(\mathrm{Re}, \|\mathbf{u}\|_{1,\Omega}, \|\mathbf{U}\|_{1,\Omega}, \Omega)\Big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{\delta_\circ+\delta,\Omega}\Big]. \tag{3.21}$$

Theorem 4.1 in [7] removes the $\|\mathbf{v}\|_{\delta_\circ+\delta,\Omega}$ term by assuming uniqueness of the solution, $(\mathbf{v}, \mathbf{V}, q) \in \boldsymbol{\mathcal{V}}$. This is a direct consequence of the standard compactness argument. Since $H_0^1(\Omega)$ is compact in $H^{\delta_\circ+\delta}(\Omega)$, where $\delta \in (0, \frac{1}{2})$ and $\delta_\circ = 0$ or $\frac{1}{2}$ depending on the spatial dimension of the domain, we can apply the standard compactness argument to (3.21) in a way similar to the estimate $\|\nabla \mathbf{v}^t\|_{0,\Omega} + \|q\|_{0,\Omega} \;\le\; C(\mathrm{Re}, \mathbf{u}, \Omega)\big[\|\mathbf{f}\|_{-1,\Omega} + \|g\|_{0,\Omega} + \|\mathbf{v}\|_{0,\Omega}\big]$ in the proof of Theorem 4.1 of [7]. We also note that the slightly different constant in (3.21) (compared to the regularity estimate in [7]) has no further implications in [7] on Theorems 3.2, 4.1, and 4.2 or their proofs. Thus, we obtain continuity and coercivity for $\mathcal{L}'(\mathbf{x})[\mathbf{y}]$ under the somewhat weaker assumptions of $\mathbf{x}$ and $\mathbf{y}$ being in $\boldsymbol{\mathcal{V}}$.

We conclude that there exist two positive constants, $\tilde{c_3}$ and $\tilde{C}_3$, which depend on Re, the H^1-norm of $\mathbf{u}$ and $\mathbf{U}$, and Ω, such that

$$\tilde{c_3}(\mathbf{x})\; \|\mathbf{y}\|_{1,\Omega}^2 \;\le\; \|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 \;\le\; \tilde{C}_3(\mathbf{x})\; \|\mathbf{y}\|_{1,\Omega}^2$$

for all $\mathbf{y} \in \boldsymbol{\mathcal{V}}$. ∎

The next lemma establishes for all $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ and r sufficiently small a uniform coercivity and continuity bound on $\|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2$.

LEMMA 3.3. *Let $\mathbf{x}_*$ be an isolated solution of (2.2) and let $\tilde{c}_3(\mathbf{x}_*)$ and $\tilde{C}_3(\mathbf{x}_*)$ be the respective coercivity and continuity constants as defined in Lemma 3.2. Then,*

$$c_3\|\mathbf{y}\|_{1,\Omega}^2 \;\le\; \|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 \;\le\; C_3\|\mathbf{y}\|_{1,\Omega}^2$$

for all $\mathbf{x} = (\mathbf{u}, \mathbf{U}, p) \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$, $\mathbf{y} = (\mathbf{v}, \mathbf{V}, q)$, *and* $\mathbf{z} = (\mathbf{w}, \mathbf{W}, t) \in \boldsymbol{\mathcal{V}}$, *where* $c_3 := \tilde{c}_3(\mathbf{x}_*) - \gamma_0 r^2 > 0$ *provided* $r < \sqrt{\tilde{c}_3(\mathbf{x}_*)/\gamma_0}$ *and* $C_3 := \tilde{C}_3(\mathbf{x}_*) + \gamma_0 r^2 > 0$.

Proof. For all $\mathbf{x}, \mathbf{y} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ and r determined later, Lemma 3.2 implies that

$$\begin{aligned}
\|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 &= \|\mathcal{L}'(\mathbf{x})[\mathbf{y}] - \mathcal{L}'(\mathbf{x}_*)[\mathbf{y}] + \mathcal{L}'(\mathbf{x}_*)[\mathbf{y}]\|_{0,\Omega}^2 \\
&\le \|\mathcal{L}''(\tilde{\mathbf{x}})[\mathbf{x} - \mathbf{x}_*, \mathbf{y}]\|_{0,\Omega}^2 + \|\mathcal{L}'(\mathbf{x}_*)[\mathbf{y}]\|_{0,\Omega}^2 \\
&\le \big(\gamma_0 \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega}^2 + \tilde{C}_3(\mathbf{x}_*)\big) \|\mathbf{y}\|_{1,\Omega}^2 \\
&\le \big(\gamma_0 r^2 + \tilde{C}_3(\mathbf{x}_*)\big) \|\mathbf{y}\|_{1,\Omega}^2 =: \; C_3 \|\mathbf{y}\|_{1,\Omega}^2
\end{aligned}$$

and

$$\begin{aligned}
\|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 &= \|\mathcal{L}'(\mathbf{x})[\mathbf{y}] - \mathcal{L}'(\mathbf{x}_*)[\mathbf{y}] + \mathcal{L}'(\mathbf{x}_*)[\mathbf{y}]\|_{0,\Omega}^2 \\
&\ge \; - \|\mathcal{L}''(\tilde{\mathbf{x}})[\mathbf{x} - \mathbf{x}_*, \mathbf{y}]\|_{0,\Omega}^2 + \|\mathcal{L}'(\mathbf{x}_*)[\mathbf{y}]\|_{0,\Omega}^2 \\
&\ge \big(- \gamma_0 \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega}^2 + \tilde{c}_3(\mathbf{x}_*)\big) \|\mathbf{y}\|_{1,\Omega}^2 \\
&\ge \big(- \gamma_0 r^2 + \tilde{c}_3(\mathbf{x}_*)\big) \|\mathbf{y}\|_{1,\Omega}^2 =: \; c_3 \|\mathbf{y}\|_{1,\Omega}^2.
\end{aligned}$$

Constant C_3 is obviously positive and $r < \sqrt{\frac{\tilde{c}_3(\mathbf{x}_*)}{\gamma_0}}$ ensures that c_3 is positive. ∎

Note that c_3 and C_3 in Lemma 3.3 depend only on Re, r, and Ω.

Next, we derive continuity and coercivity results for the full nonquadratic functional, $\mathcal{F}(\mathbf{x}; \mathbf{g})$. First, we establish these results for its second Fréchet derivative. Then, almost as a direct implication of this, we achieve continuity and coercivity for the functional norm itself. We restrict ourselves to an r that is small enough to ensure that all of these results hold uniformly for $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$.

THEOREM 3.1. *There exists an* $r > 0$ *such that for any* $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$, *the second Fréchet derivative of* $\mathcal{F}(\mathbf{x}; \mathbf{g})$ *in direction* $[\mathbf{y}, \mathbf{y}]$, $\mathbf{y} \in \boldsymbol{\mathcal{V}}$, *is* positive. *Furthermore, there exist two positive constants,* c_4 *and* C_4, *which depend only on* Re, r, *and* Ω, *such that*

$$\mathcal{F}''(\mathbf{x}; \mathbf{g})[\mathbf{y}, \mathbf{z}] \;\le\; C_4 \|\mathbf{y}\|_{1,\Omega} \|\mathbf{z}\|_{1,\Omega} \tag{3.22}$$

and

$$c_4 \|\mathbf{y}\|_{1,\Omega}^2 \;\le\; \mathcal{F}''(\mathbf{x}; \mathbf{g})[\mathbf{y}, \mathbf{y}] \tag{3.23}$$

for any $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ *and all* $\mathbf{y} \in \boldsymbol{\mathcal{V}}$.

Proof. First, we show that $\mathcal{F}''(\mathbf{x}; \mathbf{g})[\mathbf{y}, \mathbf{y}]$ is positive. Let $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$, with $r < \sqrt{\frac{\tilde{c}_3(\mathbf{x}_*)}{\gamma_0}}$, as in the proof of Lemma 3.3. Then the Cauchy–Schwarz inequality and Lemmas 3.1 and 3.3 show that

$$\begin{aligned}
\langle \mathcal{L}(\mathbf{x}) - g, \mathcal{L}''(\mathbf{x})[\mathbf{y}, \mathbf{y}] \rangle &= \; \langle \mathcal{L}'(\tilde{\mathbf{x}})[\mathbf{x} - \mathbf{x}_*], \mathcal{L}''(\mathbf{x})[\mathbf{y}, \mathbf{y}] \rangle \\
&\le \; \|\mathcal{L}'(\tilde{\mathbf{x}})[\mathbf{x} - \mathbf{x}_*]\|_{0,\Omega} \cdot \|\mathcal{L}''(\mathbf{x})[\mathbf{y}, \mathbf{y}]\|_{0,\Omega} \\
&\le \; \sqrt{C_3 \gamma_0} \; \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega} \|\mathbf{y}\|_{1,\Omega}^2,
\end{aligned} \tag{3.24}$$

where $\tilde{\mathbf{x}} \in [\mathbf{x}_*, \mathbf{x}] \subset \overline{\mathcal{B}}(\mathbf{x}_*, r)$. Then, by (2.4), (3.24), and Lemma 3.3, we have

$$\begin{aligned}
\mathcal{F}''(\mathbf{x};\mathsf{g})[\mathbf{y},\mathbf{y}] &= 2\ \|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}^2 + 2\langle \mathcal{L}(\mathbf{x}) - \mathsf{g}, \mathcal{L}''(\mathbf{x})[\mathbf{y},\mathbf{y}]\rangle \\
&\geq 2c_3\|\mathbf{y}\|_{1,\Omega}^2 - 2\sqrt{C_3}\gamma_0\ \|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}\|\mathbf{y}\|_{1,\Omega}^2 \\
&= \left(2c_3 - 2\sqrt{C_3}\gamma_0\ \|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}\right)\|\mathbf{y}\|_{1,\Omega}^2 \\
&= \left(2\tilde{c}_3(\mathbf{x}_*) - 2\gamma_0 r^2 - 2\gamma_0 r\sqrt{\tilde{C}_3(\mathbf{x}_*) + \gamma_0 r^2}\right)\ \|\mathbf{y}\|_{1,\Omega}^2 \\
&=:\ c_4(r)\ \|\mathbf{y}\|_{1,\Omega}^2.
\end{aligned}$$

Since $c_4(r)$ is continuous with respect to r and $c_4(0) = 2\tilde{c}_3(\mathbf{x}_*) > 0$, then $c_4(r)$ is positive for small enough $r > 0$. This r ensures that the second Fréchet derivative of $\mathcal{F}(\mathbf{x};\mathsf{g})$ in direction $[\mathbf{y},\mathbf{y}]$ is positive for all $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$.

The upper bound for $\mathcal{F}''(\mathbf{x};\mathsf{g})[\mathbf{y},\mathbf{z}]$ follows by Lemma 3.3, (3.24), and Lemma 3.1:

$$\begin{aligned}
\mathcal{F}''(\mathbf{x};\mathsf{g})[\mathbf{y},\mathbf{z}] &= 2\|\mathcal{L}'(\mathbf{x})[\mathbf{y}]\|_{0,\Omega}\|\mathcal{L}'(\mathbf{x})[\mathbf{z}]\|_{0,\Omega} + 2\langle \mathcal{L}(\mathbf{x}) - \mathsf{g}, \mathcal{L}''(\mathbf{x})[\mathbf{y},\mathbf{z}]\rangle \\
&\leq 2C_3\|\mathbf{y}\|_{1,\Omega}\|\mathbf{z}\|_{1,\Omega} + 2\sqrt{C_3}\gamma_0\|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}\|\mathbf{y}\|_{1,\Omega}\|\mathbf{z}\|_{1,\Omega} \\
&\leq \left(2C_3 + 2\sqrt{C_3}\gamma_0\, r\right)\|\mathbf{y}\|_{1,\Omega}\|\mathbf{z}\|_{1,\Omega} \\
&=: C_4(r)\ \|\mathbf{y}\|_{1,\Omega}\|\mathbf{z}\|_{1,\Omega}. \qquad \square
\end{aligned}$$

Remark 3. We henceforth assume that the r of Theorem 3.1 is so small that it is less than $0.4\frac{c_3}{\sqrt{C_3}\gamma_0}$. This can always be arranged by choosing r small enough.

Remark 4. The results of Lemma 3.1, Lemma 3.3, and Theorem 3.1 still hold if we restrict ourselves to a subspace of $\boldsymbol{\mathcal{V}}$ by assuming that $\mathbf{x} = \mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$, $\mathbf{y} = \mathbf{y}^h \in \mathcal{S}^h$, and $\mathbf{z} = \mathbf{z}^h \in \mathcal{S}^h$.

THEOREM 3.2. *The nonquadratic functional, $\mathcal{F}(\mathbf{x};\mathsf{g})$, is coercive and continuous for all $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$, where r, c_3, and C_3 are defined as in Theorem* 3.1:

$$\frac{1}{2}c_3\|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}^2 \leq \mathcal{F}(\mathbf{x};\mathsf{g}) \leq \frac{1}{2}C_3\|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}^2. \tag{3.25}$$

Proof. For any $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ and some $\tilde{\mathbf{x}} \in [\mathbf{x}_*, \mathbf{x}] \subset \overline{\mathcal{B}}(\mathbf{x}_*, r)$, we have

$$\mathcal{F}(\mathbf{x};\mathsf{g}) = \mathcal{F}(\mathbf{x}_*;\mathsf{g}) + \mathcal{F}'(\mathbf{x}_*;\mathsf{g})[\mathbf{x}-\mathbf{x}_*] + \frac{1}{2}\mathcal{F}''(\tilde{\mathbf{x}};\mathsf{g})[\mathbf{x}-\mathbf{x}_*, \mathbf{x}-\mathbf{x}_*].$$

The result now follows from the fact that $\mathcal{F}(\mathbf{x}_*;\mathsf{g}) = \mathcal{F}'(\mathbf{x}_*;\mathsf{g})[\mathbf{x}-\mathbf{x}_*] = 0$ (see (2.3) for the second equality) and Theorem 3.1. $\square$

A similar result can be easily established for discrete space $\mathcal{S}^h$.

THEOREM 3.3. *The functional norm, $\sqrt{\mathcal{F}(\mathbf{x}^h;\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^h;\mathsf{g})}$, is coercive and continuous for all $\mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$, where r, c_3, and C_3 are defined as in Theorem* 3.1:

$$\frac{1}{2}c_3\|\mathbf{x}^h-\mathbf{x}_*^h\|_{1,\Omega}^2 \leq \mathcal{F}(\mathbf{x}^h;\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^h;\mathsf{g}) \leq \frac{1}{2}C_3\|\mathbf{x}^h-\mathbf{x}_*^h\|_{1,\Omega}^2. \tag{3.26}$$

Proof. For any $\mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ and some $\tilde{\mathbf{x}} \in [\mathbf{x}_*, \mathbf{x}^h] \subset \overline{\mathcal{B}}^h(\mathbf{x}_*, r)$, we have

$$\mathcal{F}(\mathbf{x}^h;\mathsf{g}) - \mathcal{F}(\mathbf{x}_*^h;\mathsf{g}) = \mathcal{F}'(\mathbf{x}_*^h;\mathsf{g})[\mathbf{x}^h-\mathbf{x}_*^h] + \frac{1}{2}\mathcal{F}''(\tilde{\mathbf{x}};\mathsf{g})[\mathbf{x}^h-\mathbf{x}_*^h, \mathbf{x}^h-\mathbf{x}_*^h].$$

From Definition 2.1, we know that $\mathcal{F}'(\mathbf{x}_*^h;\mathsf{g})[\mathbf{x}^h-\mathbf{x}_*^h] = \langle \mathcal{L}(\mathbf{x}_*^h)-\mathsf{g}, \mathcal{L}'(\mathbf{x}_*^h)[\mathbf{x}^h-\mathbf{x}_*^h]\rangle = 0$ for all $\mathbf{x}^h - \mathbf{x}_*^h \in \mathcal{S}^h$ (see Remark 2). Hence, continuity and coercivity again follow directly from Theorem 3.1. $\square$

4. Convergence. This section establishes unique minimizers of the functionals we use in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$, $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$, and $\overline{\mathcal{B}}^{2h}(\mathbf{x}_*, r)$ under the assumption that r and h are sufficiently small. This is done in Lemmas 4.1, 4.2, and 4.3, respectively. Theorem 4.1 then shows that our coarse-grid correction and relaxation steps remain in a closed H^1-ball about $\mathbf{x}_*$ and it establishes uniform convergence of our two-level PML scheme.

LEMMA 4.1. *Let $\mathbf{x}_*$ be an isolated solution of (2.2) and r be defined as in Theorem 3.1. Then $\mathbf{x}_*$ is the unique minimizer in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$ of $\mathcal{F}(\mathbf{x}; \mathsf{g})$. It is characterized by $\mathcal{F}'(\mathbf{x}_*; \mathsf{g})[\mathbf{y}] = \mathbf{0}$ for all $\mathbf{y} \in \boldsymbol{\mathcal{V}}$, that is, it is the unique critical point in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$.*

Proof. The first assertion follows from (3.25). That $\mathbf{x}_*$ is a critical point follows from (2.3). We thus only need to show that it is the only critical point in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$, that is, that $\mathcal{F}'(\mathbf{x}; \mathsf{g})[\mathbf{y}] = \mathbf{0}$ for all $\mathbf{y} \in \boldsymbol{\mathcal{V}}$ and $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$ imply $\mathbf{x} = \mathbf{x}_*$. Under these assumptions, for all $\mathbf{y} \in \boldsymbol{\mathcal{V}}$, we have

$$0 = \mathcal{F}'(\mathbf{x}; \mathsf{g})[\mathbf{y}] - \mathcal{F}'(\mathbf{x}_*; \mathsf{g})[\mathbf{y}] = \mathcal{F}''(\tilde{\mathbf{x}}; \mathsf{g})[\mathbf{y}, \mathbf{x}-\mathbf{x}_*],$$

for some $\tilde{\mathbf{x}} \in [\mathbf{x}_*, \mathbf{x}] \subset \overline{\mathcal{B}}(\mathbf{x}_*, r)$. With $\mathbf{y} = \mathbf{x}-\mathbf{x}_* \in \boldsymbol{\mathcal{V}}$ and Theorem 3.1, we thus obtain

$$0 \ = \mathcal{F}''(\tilde{\mathbf{x}}; \mathsf{g})[\mathbf{x}-\mathbf{x}_*, \mathbf{x}-\mathbf{x}_*] \geq \ c_4 \|\mathbf{x}-\mathbf{x}_*\|_{1,\Omega}^2.$$

Therefore, $\mathbf{x} = \mathbf{x}_*$ and the proof is complete. ∎

Next, we prove the discrete analogue to Lemma 4.1.

LEMMA 4.2. *Let $\mathbf{x}_*$ be an isolated solution of (2.2). Let r, c_4, and C_4 be defined as in Theorem 3.1 and assume that h is sufficiently small. Then there exists a unique minimizer, $\mathbf{x}_*^h$, in $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ of $\mathcal{F}(\mathbf{x}; \mathsf{g})$. It is characterized by $\nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}) = \mathbf{0}$, that is, it is the unique grid h critical point in $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$.*

Proof. For $\mathbf{x} = \mathbf{x}^h \in \overline{\mathcal{B}}^h(\mathbf{x}_*, r) \subset \overline{\mathcal{B}}(\mathbf{x}_*, r)$, Theorem 3.2 yields

$$\frac{1}{2} c_4 \|\mathbf{x}^h - \mathbf{x}_*\|_{1,\Omega}^2 \leq \mathcal{F}(\mathbf{x}^h; \mathsf{g}) \leq \frac{1}{2} C_4 \|\mathbf{x}^h - \mathbf{x}_*\|_{1,\Omega}^2. \tag{4.1}$$

We first prove that the minimizer over $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$, which exists by compactness, is actually in $\mathcal{B}^h(\mathbf{x}_*, r)$. To this end, it suffices to show that there exists an $\mathbf{x}^h \in \mathcal{B}^h(\mathbf{x}_*, r)$ that has a smaller functional value than the minimum of $\mathcal{F}(\mathbf{x}; \mathsf{g})$ on $\partial\mathcal{B}^h(\mathbf{x}_*, r)$. To prove uniqueness of the minimizer, we then use an argument similar to that in Lemma 4.1.

Any $\mathbf{x}_\partial^h \in \partial\mathcal{B}^h(\mathbf{x}_*, r)$ must satisfy $\|\mathbf{x}_\partial^h - \mathbf{x}_*\|_{1,\Omega}^2 = r^2$. Hence, by (4.1), we have

$$\mathcal{F}(\mathbf{x}_\partial^h; \mathsf{g}) \geq \frac{1}{2} c_4 \|\mathbf{x}_\partial^h - \mathbf{x}_*\|_{1,\Omega}^2 = \frac{1}{2} c_4 r^2$$

for all $\mathbf{x}_\partial^h \in \partial\mathcal{B}^h(\mathbf{x}_*, r)$. Now let $r_1 = \sqrt{c_4/C_4}\, r$ and assume that h is so small that $\overline{\mathcal{B}}^h(\mathbf{x}_*, r_1)$ is not empty. Again by (4.1), any $\mathbf{x}^h \in \mathcal{B}^h(\mathbf{x}_*, r_1) \subset \mathcal{B}^h(\mathbf{x}_*, r)$ must satisfy

$$\mathcal{F}(\mathbf{x}^h; \mathsf{g}) \leq \frac{1}{2} C_4 \|\mathbf{x}^h - \mathbf{x}_*\|_{1,\Omega}^2 < \frac{1}{2} C_4 r_1^2 = \frac{1}{2} c_4 r^2 \leq \mathcal{F}(\mathbf{x}_\partial^h; \mathsf{g}).$$

Therefore, the minimizer, $\mathbf{x}_*^h$, of $\mathcal{F}(\mathbf{x}^h; \mathsf{g})$ over $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ must actually be in $\mathcal{B}^h(\mathbf{x}_*, r)$. Remark 2 confirms that it is a grid h critical point: $\nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}) = \mathbf{0}$.

To prove uniqueness, note that any other minimizer, $\mathbf{x}^h$, in $\mathcal{B}^h(\mathbf{x}_*, r)$ must be a grid h critical point: $\nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}) = \mathbf{0}$. It now suffices to show that $\mathbf{x}_*^h$ is the only grid

h critical point (which also proves the characterization assertion). To this end, note for all $\mathbf{y}^h \in \mathcal{S}^h$ that

$$\begin{aligned} 0 =& \langle \nabla^h \mathcal{F}(\mathbf{x}^h; \mathsf{g}), \mathbf{y}^h \rangle - \langle \nabla^h \mathcal{F}(\mathbf{x}_*^h; \mathsf{g}), \mathbf{y}^h \rangle \\ =& \mathcal{F}'(\mathbf{x}^h; \mathsf{g})[\mathbf{y}^h] - \mathcal{F}'(\mathbf{x}_*^h; \mathsf{g})[\mathbf{y}^h] = \mathcal{F}''(\tilde{\mathbf{x}}^h; \mathsf{g})[\mathbf{y}^h, \mathbf{x}^h - \mathbf{x}_*^h] \end{aligned}$$

for some $\tilde{\mathbf{x}}^h \in \mathcal{B}^h(\mathbf{x}_*, r)$. Again choosing $\mathbf{y}^h = \mathbf{x}^h - \mathbf{x}_*^h \in \mathcal{S}^h \subset \boldsymbol{\mathcal{V}}$ and using Theorem 3.1 yield

$$0 = \mathcal{F}''(\tilde{\mathbf{x}}^h; \mathsf{g})[\mathbf{x}^h - \mathbf{x}_*^h, \mathbf{x}^h - \mathbf{x}_*^h] \geq c_4 \|\mathbf{x}^h - \mathbf{x}_*^h\|_{1,\Omega}^2,$$

which establishes the result. ∎

LEMMA 4.3. *Let r, c_4, and C_4 be defined as in Theorem 3.1 and choose any $r_1 < \sqrt{c_4/C_4}\, r$. Then for $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r_1)$, there exists a unique minimizer, $\mathbf{x}_*^{2h} = \operatorname{argmin}_{\mathbf{x}^{2h} \in \mathcal{S}^{2h}, \mathbf{x}+\mathbf{x}^{2h} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)} \mathcal{F}(\mathbf{x} + \mathbf{x}^{2h}; \mathsf{g})$. If r is small enough, then this minimizer is characterized by $\nabla^{2h} \mathcal{F}(\mathbf{x} + \mathbf{x}^{2h}; \mathsf{g}) = \mathbf{0}$ with $\mathbf{x} + \mathbf{x}^{2h} \in \overline{\mathcal{B}}(\mathbf{x}_*, r)$, that is, it is the unique grid $2h$ critical point for which $\mathbf{x} + \mathbf{x}^{2h}$ stays in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$. Thus, the result, $\mathbf{x}_{\frac{1}{2}}^h$, of Step 1 of PML stays in $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ for any initial guess, $\mathbf{x}_0^h$, in $\overline{\mathcal{B}}^h(\mathbf{x}_*, r_1)$.*

Proof. The minimizer, $\mathbf{x}_*^{2h}$, clearly exists by compactness. (Note that $\{\mathbf{x} + \mathcal{S}^{2h}\} \cap \overline{\mathcal{B}}(\mathbf{x}_*, r)$ is a nonempty set because it contains $\mathbf{x} = \mathbf{x} + 0$.) To prove uniqueness and the fact that $\mathbf{x}_*^{2h}$ is a grid $2h$ critical point, first note that any $\mathbf{x}_\partial \in \partial\mathcal{B}(\mathbf{x}_*, r)$ must, by Theorem 3.2, satisfy $\mathcal{F}(\mathbf{x}_\partial; \mathsf{g}) \geq \frac{1}{2} c_4 \|\mathbf{x}_\partial - \mathbf{x}_*\|_{1,\Omega}^2 = \frac{1}{2} c_4 r^2$. Then, with $\mathbf{x} \in \overline{\mathcal{B}}(\mathbf{x}_*, r_1)$, again Theorem 3.2 implies that

$$\mathcal{F}(\mathbf{x}; \mathsf{g}) \leq \frac{1}{2} C_4 \|\mathbf{x} - \mathbf{x}_*\|_{1,\Omega}^2 \leq \frac{1}{2} C_4 r_1^2 < \mathcal{F}(\mathbf{x}_\partial; \mathsf{g}).$$

Thus, $\mathcal{F}(\mathbf{x} + \mathbf{x}_*^{2h}; \mathsf{g}) \leq \mathcal{F}(\mathbf{x}; \mathsf{g}) < \mathcal{F}(\mathbf{x}_\partial; \mathsf{g})$, which implies that $\mathbf{x} + \mathbf{x}_*^{2h} \in \mathcal{B}(\mathbf{x}_*, r)$. Then $\mathbf{x}_*^{2h}$ must satisfy the gradient condition, which follows by a similar standard argument similar to that of Remark 2.

Uniqueness and the characterization assertion can be now established as in the proofs of Lemmas 4.1 and 4.2. To this end, assume that there exists another minimizer, $\mathbf{x} + \mathbf{x}^{2h} \in \mathcal{B}^h(\mathbf{x}_*, r)$, so that $\nabla^{2h} \mathcal{F}(\mathbf{x} + \mathbf{x}^{2h}; \mathsf{g}) = \mathbf{0}$. It now suffices to show that this grid $2h$ critical point condition implies that $\mathbf{x}^{2h} = \mathbf{x}_*^{2h}$ (which also proves the characterization assertion). To this end, note for all $\mathbf{y}^{2h} \in \mathcal{S}^{2h}$, that

$$\begin{aligned} 0 =& \langle \nabla^{2h} \mathcal{F}(\mathbf{x} + \mathbf{x}^{2h}; \mathsf{g}), \mathbf{y}^{2h} \rangle - \langle \nabla^{2h} \mathcal{F}(\mathbf{x} + \mathbf{x}_*^{2h}; \mathsf{g}), \mathbf{y}^{2h} \rangle \\ =& \mathcal{F}'(\mathbf{x} + \mathbf{x}^{2h}; \mathsf{g})[\mathbf{y}^{2h}] - \mathcal{F}'(\mathbf{x} + \mathbf{x}_*^{2h}; \mathsf{g})[\mathbf{y}^{2h}] = \mathcal{F}''(\mathbf{x} + \tilde{\mathbf{x}}^{2h}; \mathsf{g})[\mathbf{y}^{2h}, \mathbf{x}^{2h} - \mathbf{x}_*^{2h}] \end{aligned}$$

for some $\mathbf{x} + \tilde{\mathbf{x}}^{2h} \in \mathcal{B}^h(\mathbf{x}_*, r)$. As before, this leads to

$$0 = \mathcal{F}''(\mathbf{x} + \tilde{\mathbf{x}}^{2h}; \mathsf{g})[\mathbf{x}^{2h} - \mathbf{x}_*^{2h}, \mathbf{x}^{2h} - \mathbf{x}_*^{2h}] \geq c_4 \|\mathbf{x}^{2h} - \mathbf{x}_*^{2h}\|_{1,\Omega}^2,$$

which proves uniqueness and the characterization assertion.

The final claim follows simply by choosing $\mathbf{x} = \mathbf{x}^h$ and noting that we can choose $r > 0$ so small that the nearest optimally corrected $\mathbf{x}^h$ must be the one in $\overline{\mathcal{B}}(\mathbf{x}_*, r)$. ∎

THEOREM 4.1. *Let r and r_1 as in Lemma 4.3, define $r_0 = \sqrt{c_4/C_4}\, r_1$, and choose h and ω sufficiently small. Then for any $\mathbf{x}_0^h \in \overline{\mathcal{B}}(\mathbf{x}_*, r_0)$, the PML iterates based on*

either (2.8) *or* (2.9) *remain in* $\overline{\mathcal{B}}^h(\mathbf{x}_*, r)$ *and converge linearly with uniformly bounded factor according to*

$$\mathcal{F}(\mathbf{x}^h_{k+1}; \mathrm{g}) - \mathcal{F}(\mathbf{x}^h_*; \mathrm{g}) \leq \kappa \left(\mathcal{F}(\mathbf{x}^h_k; \mathrm{g}) - \mathcal{F}(\mathbf{x}^h_*; \mathrm{g}) \right), \qquad k = 0, 1, 2, \ldots,$$

where $\kappa \in [0, 1)$ *depends only on* Re, r, *and* Ω.

Proof. We omit this fairly straightforward but somewhat lengthy proof and instead refer the reader to [10] for details. ∎

REFERENCES

[1] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, *Analysis of velocity-flux first-order system least-squares principles for the Navier-Stokes equations: Part* I, SIAM J. Numer. Anal, 35 (1998), pp. 990–1009.

[2] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, *Analysis of velocity-flux least-squares principles for the Navier-Stokes equations: Part* II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.

[3] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Appl. Math., Springer-Verlag, New York, 2002.

[5] V. Girault and P.-A. Raviart, *Finite element methods for Navier–Stokes equations*, in Theory and Algorithms, Springer Ser. in Comput. Math. 5, Springer-Verlag, Berlin, 1986.

[6] B.-N. Jiang, *The least-squares finite element method*, in Theory and Applications in Computational Fluid Dynamics and Electromagnetics, Scientific Comput., Springer-Verlag, Berlin, 1998.

[7] S.-D. Kim, C.-O. Lee, T. A. Manteuffel, S. F. McCormick, and Oliver Röhrle, *First-order system least-squares functionals for the Oseen equations*, J. Numer. Linear Algebra Appl., submitted.

[8] T. A. Manteuffel, S. F. McCormick, O. Röhrle, and J. Ruge, *Projection multilevel methods for quasilinear elliptic partial differential equations: Numerical results*, SIAM J. Numer. Anal., 44 (2006), pp. 120–138.

[9] S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 62, SIAM, Philadelphia, 1992.

[10] O. Röhrle, *Multilevel First-Order System Least Squares for Quasilinear Elliptic Partial Differential Equations*, Ph.D. thesis, University of Colorado, Boulder, 2004.

[11] X.-C. Tai, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, Numer. Math., 93 (2003), pp. 755–786.

[12] X.-C. Tai and J. Xu, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 153–182

REGULARIZING NEWTON–KACZMARZ METHODS FOR NONLINEAR ILL-POSED PROBLEMS*

MARTIN BURGER† AND BARBARA KALTENBACHER‡

Abstract. We introduce a class of stabilizing Newton–Kaczmarz methods for nonlinear ill-posed problems and analyze their convergence and regularization behavior. As usual for iterative methods for solving nonlinear ill-posed problems, conditions on the nonlinearity (or the derivatives) have to be imposed in order to obtain convergence. As we shall discuss in general and in some specific examples, the nonlinearity conditions obtained for the Newton–Kaczmarz methods are less restrictive than those for previously existing iteration methods and can be verified for several practical applications.

We also discuss the discretization and efficient numerical solution of the linear problems arising in each step of a Newton–Kaczmarz method, and we carry out numerical experiments for two model problems.

Key words. Newton–Kaczmarz methods, ill-posed problems, tomography

AMS subject classifications. 65J20, 65J15, 65N21, 47J06

DOI. 10.1137/040613779

1. Introduction.

The aim of this paper is to develop and analyze *Newton–Kaczmarz methods* for nonlinear inverse problems, focusing in particular on the important class of identification problems with multiple boundary data. The main idea of the Kaczmarz method is to split the inverse problem into a finite number of subproblems and to approximate its solution by performing a cyclic iteration over the subproblems.

As a *regularized Newton–Kaczmarz method* we understand the cyclic iteration where at each step one iteration of a regularized Newton method is applied to a subproblem. As we shall discuss in detail in this paper, the benefit from this approach is twofold:

1. Instead of solving one large problem in each iteration step, we can solve several smaller subproblems, which might lead to a reduction of the overall computational effort.
2. Due to the ill-posedness of the problem, conditions on the nonlinearity of the problem have to be imposed in order to ensure convergence of iterative methods (cf. [6] for an overview). These conditions are rather restrictive and cannot be verified for many practical problems, in particular for parameter identification problems using boundary data related to the solutions of partial differential equations. As we shall show below for several applications, the nonlinearity conditions for the Newton–Kaczmarz method are less restrictive and can be verified in more realistic cases.

*Received by the editors August 24, 2004; accepted for publication (in revised form) August 23, 2005; published electronically February 8, 2006.
http://www.siam.org/journals/sinum/44-1/61377.html

†Industrial Mathematics Institute, Johannes Kepler University, Altenbergerstr. 69, A 4040 Linz, Austria (martin.burger@jku.at). This author was supported by the Austrian National Science Foundation FWF through project SFB F 013/08 and the Institute of Pure and Applied Mathematics at UCLA.

‡Junior research group "Inverse Problem in Piezoelectricity," Department of Sensor Technology, Universität Erlangen-Nürnberg, Paul-Gordan-Strasse 3/5, D 91053 Erlangen, Germany (barbara.kaltenbacher@lse.eei.uni-erlangen.de). This author was supported by the German Science Foundation DFG under grant Ka 1778/1-1.

The price which one has to pay is that at least theoretically it turns out that more a priori information has to be contained in the initial values.

Another motivation for the analysis in this paper is that Kaczmarz-type methods (also called *algebraic reconstruction technique*) have been used in several applications with multiple boundary data (cf. [3, 8, 9, 24, 30]) and performed better than standard iterative methods. This paper, together with the results of Kowar and Scherzer [24] on the Landweber–Kaczmarz method, might serve to provide a theoretical basis.

Many inverse problems can be formulated as nonlinear operator equations,

$$F(x) = y, \tag{1.1}$$

or as collections of p coupled operator equations,

$$F_i(x) = y_i, \qquad i = 0, \dots p-1, \tag{1.2}$$

with nonlinear operators F_i mapping between Hilbert spaces X and Y_i. We will here assume that a solution $x^\dagger$ of (1.2) exists but need not necessarily be unique.

Note that (1.1) can be seen as a special case of (1.2) with $p = 1$; on the other hand, defining

$$F := (F_0, \dots, F_{p-1}), \qquad y := (y_0, \dots, y_{p-1}), \tag{1.3}$$

one can reduce (1.2) to (1.1). However, one potential advantage of (1.2) over (1.1) can be that it might better reflect the structure of the underlying information $(y_0, \dots, y_{p-1})$ leading to the coupled system than could a plain concatenation into one single data element y. The most important feature that we have in mind, however, is that it enables the definition of Newton-type solution methods and to prove their convergence for certain relevant problems, for which Newton-type methods applied to the single equation formulation (1.1) cannot be shown to converge.

In general we assume that we only have noisy data y_i^δ with some noise level δ bounding the noise of every measurement by

$$\|y_i^\delta - y_i\| \le \delta, \tag{1.4}$$

Note that for $p > 1$, this assumption on the noise is more restrictive than the frequently used noise bound

$$\|y^\delta - y\| \le \delta,$$

but it reflects the case of multiple measurements, where an individual noise bound is available for each. If the noise level for each measurement is different, we can make it equal by using a relative scaling between the operators F_i.

Since we are interested in the situation that (1.2) is ill-posed in the sense that small perturbations in the data can lead to large deviations in the solution, and since in practice only noisy data are available, we have to apply suitable regularization techniques (see, e.g., [11, 13, 23, 25, 28, 29, 33]). Typically, the instability in nonlinear inverse problems (1.1) corresponds to a smoothing property of the forward operator F and its linearization $F'(x)$. In particular, for an ill-posed problem, we cannot expect that $F'(x)$ is continuously invertible, and consequently a standard Newton or Gauss–Newton cannot be used. Modified Newton-type methods for solving (1.1) have been studied and analyzed in several recent publications, see; e.g., [1, 6, 14, 15, 22, 31]. Regularization is here achieved by replacing the generally unbounded inverse of $F'(x)$

in the definition of the Newton step by a bounded approximation, defined via a regularizing operator

$$G_\alpha(F'(x)) \approx F'(x)^\dagger.$$

Here, $K^\dagger$ denotes the pseudoinverse of a linear operator K, $\alpha > 0$ is a small regularization parameter, and G_α satisfies

$$G_\alpha(K)y \to K^\dagger y \quad \text{as } \alpha \to 0 \quad \forall y \in \mathcal{R}(K) \tag{1.5}$$

and

$$\|G_\alpha(K)\| \le \Phi(\alpha) \tag{1.6}$$

for any linear operator K within some uniformly bounded set. Note that, especially in view of operators K with unbounded inverses, the constant $\Phi(\alpha)$ has to tend to infinity as α goes to zero; we assume without loss of generality (w.l.o.g.) that $\Phi(\alpha)$ is strictly monotonically decreasing.

Choosing a sequence (α_n) of regularization parameters and applying the bounded operators $G_{\alpha_n}(F'(x_n))$ in place of $F'(x_n)^{-1}$ in Newton's method results in the iteration

$$x_{n+1} = x_n - G_{\alpha_n}(F'(x_n))(F(x_n) - y^\delta). \tag{1.7}$$

If G_α is defined by Tikhonov regularization

$$G_\alpha(K) = (K^*K + \alpha I)^{-1}K^*, \tag{1.8}$$

one arrives at the Levenberg–Marquardt method. (See [15]; for G_α given by a conjugate gradient iteration, see [14], and further work on this class of methods can be found in [32].)

A different class of regularized Newton methods emerged from the iteratively regularized Gauss–Newton method (IRGNM),

$$x_{n+1} = x_0 - G_{\alpha_n}(F'(x_n))(F(x_n) - y^\delta - F'(x_n)(x_n - x_0)),$$

with (1.8), which was first proposed and analyzed by Bakushinskii in [1] and later extended to regularization with general regularization operators G_{α_n} [2]; see also [19, 22]. Here, $\alpha_n \overset{n\to\infty}{\to} 0$ is an a priori chosen monotonically decreasing sequence of regularization parameters. One observes that in the limiting case $\alpha_n \to 0$ (i.e., $G_{\alpha_n}(F'(x_n)) \to F'(x)^\dagger$) also this formulation is equivalent to the usual Newton method.

In order to make these Newton-type methods applicable to multiple equations (1.2), we combine them with a Kaczmarz approach (similar to [24]). Starting from an initial guess $x_{0,i}$, we perform a Newton step for the equation $F_i(x) = y_i$, for i from 0 to $p-1$, and repeat this procedure in a cyclic manner. Incorporating the possibility of different regularization methods G^i for each equation in (1.2), and using the "overloading" notation

$$x_{0,n} := x_{0,mod(n,p)}, \quad F_n := F_{mod(n,p)}, \quad y_n := y_{mod(n,p)}, \quad G_\alpha^n := G_\alpha^{mod(n,p)}, \tag{1.9}$$

this can be written as

$$x_{n+1} = x_{0,n} - G^n_{\alpha_n}(F_n'(x_n))(F_n(x_n) - y_n^\delta - F_n'(x_n)(x_n - x_{0,n})). \tag{1.10}$$

A combination of the Levenberg–Marquardt method with a Kaczmarz approach will be discussed in section 3.

Our convergence analysis will be a local one, i.e., we will work in a neighborhood $\mathcal{B}_\rho(x^\dagger)$ of the solution, which we assume to be a subset of the domains of the operators F_i

$$\mathcal{B}_\rho(x^\dagger) \subseteq \mathcal{D}(F_i), \qquad i = 0, \dots p-1.$$

The remainder of the paper is organized as follows. In section 2 we discuss conditions on the nonlinearity of the problem and so-called source conditions, which are abstract smoothness assumptions on the solution. Section 3 contains a convergence analysis of (1.10) including the case of noisy data and convergence rates under additional regularity assumptions. In section 4, we derive some approaches for the efficient implementation of the proposed methods, and section 5 provides numerical results.

2. Nonlinearity and source conditions. In the following we shall discuss the basic conditions needed for the subsequent analysis in this paper. In particular we shall introduce conditions on the nonlinearity of the involved operators F_i and investigate their applicability to tomography-type problems.

2.1. Nonlinearity conditions. To make these methods well defined, we assume the forward operators F_i to be Fréchet differentiable with derivatives being uniformly bounded in a neighborhood of the solution. This uniform bound has to be such that applicability of the respective regularization method can be guaranteed,

$$\|F_i'(x)\| \le C_S^i \qquad \forall x \in \mathcal{B}_\rho(x^\dagger), \tag{2.1}$$

which can always be achieved by a proper scaling. In order to prove convergence of regularization methods for nonlinear ill-posed problems, one usually needs assumptions not only on the smoothness of the forward operator F but also on the type of nonlinearity it contains. Here we shall mainly consider the condition

$$F_i'(\bar{x}) = F_i'(x) R_i(\bar{x}, x) \qquad \forall \bar{x}, x \in \mathcal{B}_\rho(x^\dagger), \tag{2.2}$$

which means that the range of the Fréchet derivative of each forward operator F_i is locally invariant around the solution. The linear operators $R_i(\bar{x}, x)$ (which need not be known explicitly) should satisfy a Lipschitz type estimate

$$\|R_i(\bar{x}, x) - I\| = \|R_i(\bar{x}, x) - R_i(x, x)\| \le C_R \|\bar{x} - x\|. \tag{2.3}$$

This corresponds to an analogous assumption in the context of $p = 1$, i.e., (1.1),

$$F'(\bar{x}) = F'(x) R(\bar{x}, x) \qquad \forall \bar{x}, x \in \mathcal{B}_\rho(x^\dagger), \tag{2.4}$$

as it was used, e.g., in the convergence analysis of [22] and is closely related to the so-called affine covariant Lipschitz condition in [7]. Condition (2.2) seems to be natural especially in the context of parameter identification in PDEs from boundary measurements where the forward operator consists of a (typically invertible) solution operator for the PDE, composed with a linear operator mapping the PDE solution to the measured boundary values. In fact, by the additional freedom arising from the possibility of having different operators R_i for each i, it can be verified for important applications of parameter identification, like ultrasound tomography (see below)

and impedance tomography, for which other nonlinearity conditions used in literature cannot be proven to hold.

An alternative nonlinearity condition that can be found in the literature on regularization methods for nonlinear inverse problems (1.1) is

$$F'(\bar{x}) = R(\bar{x}, x)F'(x) \qquad \forall \bar{x}, x \in \mathcal{B}_\rho(x^\dagger) \tag{2.5}$$

with regular operators $R(\bar{x}, x)$, i.e., range invariance of the adjoints of $F'(x)$, which is closely related to the tangential cone condition used, e.g., in [14, 15, 17, 21, 22] and to the Newton–Mysovskii conditions discussed in [6].

We want to mention that the nonlinearity condition (2.2) is less restrictive than the corresponding nonlinearity condition (2.4) for the operator F defined by (1.3). If (2.4) holds, we can easily deduce (2.2) by choosing $R_i = R$ for all i. For the alternative condition

$$F_i'(\bar{x}) = R_i(\bar{x}, x)F_i'(x) \qquad \forall \bar{x}, x \in \mathcal{B}_\rho(x^\dagger) \tag{2.6}$$

and the corresponding condition (2.5) for F defined by (1.3), we obtain sufficiency in the other direction, since we can choose R to be the diagonal operator consisting of all R_i to obtain the range invariance of F'^* from (2.6).

Finally, we examine a special case of a decomposition of F_i in a linear singular and a nonlinear regular operator. As we shall see below in several examples, the operators F_i can often be written as the composition of linear trace-type operators with nonlinear parameter-to-solution maps for partial differential equations. Thus we start with a simple observation that allows us to verify the nonlinearity condition for the parameter-to-solution map only. In this context see section 5 in [18], where a class of operators satisfying the nonlinearity condition (2.5) is derived.

LEMMA 2.1. *Let X, Y, Z be Hilbert spaces, and let $L_i \in \mathcal{L}(Z, Y)$. Moreover, let $H_i : X \to Z$, $i = 0, \dots, p-1$, be continuously Fréchet differentiable operators. Then,*

$$F_i = L_i \circ H_i \tag{2.7}$$

satisfies (2.2), (2.3) *if H_i satisfies* (2.2), (2.3).

Moreover, if $H_i'(x)$ is regular for all $x \in \mathcal{B}_\rho(x^\dagger)$ with uniformly bounded inverse, and the map $x \mapsto H_i'(x)$ is Lipschitz continuous, then the condition (2.2), (2.3) *is satisfied by H_i.*

Proof. The first assertion follows from

$$F_i'(\bar{x}) = L_i \circ H_i'(\bar{x}) = L_i \circ H_i'(x) \circ R_i(\bar{x}, x) = F_i'(x) \circ R_i(\bar{x}, x).$$

Moreover, if H_i' is regular, we may define

$$R_i(\bar{x}, x) := H_i'(x)^{-1} H_i'(\bar{x}),$$

which implies (2.2). Due to the regularity of $H_i'(x)^{-1}$ and the Lipschitz-continuity of $x \mapsto H_i'(x)$, we obtain

$$\|R_i(\bar{x}, x) - I\| = \|H_i'(x)^{-1}(H_i'(\bar{x}) - H_i'(x))\| \le C_0 \|H_i'(\bar{x}) - H_i'(x)\| \le C_R \|\bar{x} - x\|,$$

i.e., (2.3) holds. ∎

2.2. Examples. In the following we discuss several examples of inverse problems satisfying the nonlinearity condition including tomography-type problems for partial differential equations in the above framework, and we show that they satisfy the nonlinearity condition (2.2).

Example 1 (reconstruction from the Dirichlet–Neumann map). We start with a rather simple model problem, namely, the estimation of the coefficient $q \geq 0$ in the partial differential equation

$$-\Delta u + qu = 0 \qquad \text{in } \Omega \subset \mathbb{R}^d$$

from measurements of the Neumann value $g = \frac{\partial u}{\partial \nu}$ on $\partial\Omega$ for several different Dirichlet values $u = f$ on $\partial\Omega$.

If we denote the different Dirichlet values by f_i, $i = 0, \ldots, p-1$, and the corresponding measurements by g_i, we may rewrite the problem as

$$F_i(q) = g_i, \qquad i = 0, \ldots, p-1,$$

where $F_i : \mathcal{D}(F_i) \subseteq L^2(\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$ is the nonlinear operator mapping q to $\frac{\partial u_i}{\partial \nu}$, where $u_i \in H^1(\Omega)$ is the weak solution of

$$\begin{aligned} -\Delta u_i + qu_i &= 0 \qquad \text{in } \Omega, \\ u_i &= f_i \qquad \text{on } \partial\Omega, \end{aligned}$$

and $\mathcal{D}(F_i)$ is to be specified below.

The decomposition (2.7) is obtained with $L : H^1(\Omega) \mapsto H^{-\frac{1}{2}}(\partial\Omega)$ being the trace operator that maps a function to its normal derivative on the boundary, and $H_i : q \mapsto u_i$ is the parameter-to-solution map.

The derivative $v_i = H_i'(q)s$ is given as the unique weak solution of

$$\begin{aligned} -\Delta v_i + qv_i + su_i &= 0 \qquad \text{in } \Omega, \\ v_i &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

Formally, we can write $H_i'(q) = -(-\Delta + q)^{-1}(u_i .)$. It can be shown easily that this operator is regular between $L^2(\Omega)$ and $H^1(\Omega)$ if $u_i > 0$. Due to a standard maximum principle for second-order elliptic differential equations, this is the case if $q \geq 0$ and $f_i > 0$. Moreover, since embedding operators are continuous and regular, the operator $H_i'(q)$ is also regular between a Sobolev space $H^\beta(\Omega)$, $\beta \geq 0$, and $H^1(\Omega)$. Thus, if $\beta > \frac{d}{2}$ (i.e., $H^\beta(\Omega) \hookrightarrow C(\bar{\Omega})$) and there exists a minimum norm solution $q^\dagger \in H^\beta(\Omega)$, which is positive in $\bar{\Omega}$, then $q \in \mathcal{B}_\rho(q^\dagger)$ is nonnegative for ρ sufficiently small and due to the above reasoning Lemma 2.1 implies that the nonlinearity condition (2.2), (2.3) is satisfied for $f_i > 0$, if we consider F_i as an operator from $\mathcal{D}(F_i) := \mathcal{B}_\rho(q^\dagger) \subseteq H^\beta(\Omega) =: X$ to $H^{-\frac{1}{2}}(\partial\Omega) =: Y_i$.

Example 2 (reconstruction from multiple sources). In some examples, one rather tries to estimate coefficients in partial differential equations from boundary measurements for different interior sources rather than from different boundary values. We consider the estimation of the coefficient $q \geq 0$ in

$$-\Delta u + qu = h \qquad \text{in } \Omega \subset \mathbb{R}^d$$

subject to a homogeneous Neumann boundary condition $\frac{\partial u}{\partial \nu} = 0$ on $\partial\Omega$, and measurements of the Dirichlet values $u = f$ on $\partial\Omega$ for different sources $h \in H^{-1}(\Omega)$. Problems

of this kind have been discussed by Lowe and Rundell [26, 27] and in an application to semiconductor devices by Fang and Ito [12].

Again, we can decompose the corresponding operators F_i into the trace operator $L : H^1(\Omega) \to L^2(\partial\Omega)$ concatenated with the parameter-to-solution maps $H_i : q \mapsto u_i$ defined by the solution of

$$\begin{aligned} -\Delta u_i + q u_i &= h_i \qquad \text{in } \Omega, \\ \frac{\partial u_i}{\partial \nu} &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

The derivative $H_i'(q)$ is almost the same as in the previous example, except for a change from Dirichlet to Neumann boundary conditions. One can verify the regularity of u_i in the same way as above for $h_i > 0$ (which allows one to apply a maximum principle for u_i) and consequently show that the nonlinearity condition (2.2), (2.3) holds.

Example 3 (SPECT). In the application of single photon emission computed tomography (SPECT), one wants to compute the source f and the coefficient $a \geq 0$ from

$$\theta_i \cdot \nabla u_i + a u_i = f \qquad \text{in } \Omega \subset \mathbb{R}^d,$$

for different values θ_i on the unit sphere, and the boundary values

$$\begin{aligned} u_i &= 0 \qquad \text{on } \partial\Omega_i^- := \{\, x \in \partial\Omega \mid \nu(x) \cdot \theta_i \leq 0 \,\}, \\ u_i &= g_i \qquad \text{on } \partial\Omega_i^+ := \{\, x \in \partial\Omega \mid \nu(x) \cdot \theta_i \geq 0 \,\}. \end{aligned}$$

Here, the condition on $\partial\Omega_i^-$ has to be understood as the boundary condition, while the values g_i on $\partial\Omega_i^+$ are the measurements. Thus, the operators F_i map (a, f) to g_i. They can be decomposed into the trace operators $L_i : L^2(\Omega) \to H^{-\frac{1}{2}}(\partial\Omega_i^+)$ and the parameter-to-solution maps $H_i : \mathcal{D}(F_i) \subseteq L^2(\Omega)^2 \to L^2(\Omega), (a, f) \mapsto u_i$.

It can be shown (cf. [30]) that the derivative $v_i = H_i'(a, f)(\hat{a}, \hat{f})$ can be determined as the unique solution of

$$\theta_i \cdot \nabla v_i + a v_i = \hat{f} - \hat{a} u_i \qquad \text{in } \Omega \subset \mathbb{R}^d,$$

subject to $v_i = 0$ on $\partial\Omega_i^-$. If $a > 0$, $f > 0$, a maximum principle applies also to the first-order equation and one may conclude $u_i > 0$, which subsequently can be used to verify the nonlinearity condition (2.2), (2.3) in the same way as for the above examples.

Example 4 (ultrasound tomography). The inverse problem in ultrasound tomography consists in finding $f \in L^2(\Omega)$ from boundary measurements $g_i = u_i$ on $\partial\Omega$ for complex-valued waves $u_i = e^{\mathbf{i}kx\cdot\theta_i} + v_i$, where v_i solves the Helmholtz equations

$$\begin{aligned} \Delta v_i + k^2(1 - f) v_i &= k^2 f e^{\mathbf{i}kx\cdot\theta_i} \qquad \text{in } \Omega, \\ \frac{\partial v_i}{\partial \nu} &= B v_i \qquad \text{on } \partial\Omega, \end{aligned}$$

with B being an appropriate operator representing the radiation condition and k a real parameter controlling the spatial resolution. Again we can decompose the operator $F_i : f \mapsto g_i$ into the trace operator $L : H^1(\Omega) \to L^2(\partial\Omega)$ and the parameter-to-solution map $H_i : \mathcal{D}(F_i) \subseteq L^2(\Omega) \to H^1(\Omega), f \mapsto u_i$.

One can show (cf. [30]) that the derivative $w_i = H_i'(f)$ is defined by the solution of

$$\begin{aligned} \Delta u_i + k^2(1-f)w_i &= k^2 f u_i \qquad \text{in } \Omega, \\ \frac{\partial w_i}{\partial \nu} &= B w_i \qquad \text{on } \partial\Omega. \end{aligned}$$

If f, k are such that the operator $\Delta + k^2(1-f)$ is regular, and if $|u_i| \neq 0$, then one can easily verify the nonlinearity condition in the same way as for the examples above.

Example 5 (nonlinear moment estimation). We finally consider a nonlinear moment estimation problem, which consists in finding $u \in L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$ a bounded domain, given

$$g_i := \int_\Omega k_i(x, u(x))\, dx \in \mathbb{R}^m$$

for given smooth kernel functions $k_i : \Omega \times \mathbb{R} \to \mathbb{R}^m$ (which could, e.g., arise from the discretization of an integral kernel, i.e., $k_i(x, u(x)) = K(x, u(x), y_i)$). Here the operator $F_i : L^2(\Omega) \to \mathbb{R}^m$ is the concatenation of the linear integration operator $L : L^2(\Omega)^m \to \mathbb{R}^m$, $w \mapsto \int_\Omega w\, dx$, and the Nemitskij-type operator $H_i : L^2(\Omega) \to L^2(\Omega), u \mapsto k_i(., u)$. The derivative of the nonlinear operator H_i is given by

$$H_i'(u)v = \frac{\partial k_i}{\partial u}(., u)v.$$

If $k_i \in C(\Omega, C_b^{1,1}(\mathbb{R}))$ and $\frac{\partial k_i}{\partial u} \neq 0$, then $H_i'(u)$ is regular and the map $u \mapsto H_i'(u)$ is Lipschitz continuous, which implies the nonlinearity condition (2.2), (2.3).

2.3. Source conditions. Convergence of regularization methods for ill-posed problems is, as a direct consequence of the instability, in general arbitrarily slow. In order to obtain convergence rates, additional regularity assumptions on the difference between an exact solution $x^\dagger$ and some initial guess x_0 used in the regularization method must be made. These have the form of so-called source wise representation conditions and in our context read as

$$x^\dagger - x_{0,i} = f(F_i'(x^\dagger)^* F_i'(x^\dagger))w_i, \quad i = 0, \dots, p-1, \tag{2.8}$$

for some w_i, where f is some real function and for the positive semidefinite operator $F_i'(x^\dagger)^* F_i'(x^\dagger)$, $f(F_i'(x^\dagger)^* F_i'(x^\dagger))$ is defined via functional calculus (cf., e.g., [11]). Condition (2.8) expresses the assumed regularity of $x^\dagger - x_{0,i}$ in terms of the smoothing property of $F'(x^\dagger)$ mentioned above. Typical functions f used here are

$$f(\lambda) := f_\nu^H(\lambda) := \lambda^\nu \tag{2.9}$$

for some Hölder exponent ν, or the weaker, but for exponentially ill-posed problems more appropriate logarithmic functions

$$f(\lambda) := f_\mu^L(\lambda) = (-\ln(\lambda))^{-\mu}. \tag{2.10}$$

Remark 1. Note that under sufficiently strong source conditions, namely, (2.8) with $f(\lambda) = O(\sqrt{\lambda})$, which corresponds to $\nu \geq \frac{1}{2}$ in (2.9), the nonlinearity assumptions on F (see the previous subsection) can be considerably relaxed in case $p = 1$. In place of (2.4) or (2.5) one only needs Lipschitz continuity of F'. This was observed by

Bakushinskii in [1] for the case $\nu \geq 1$ and IRGNM (see [21] for the case $\nu \geq \frac{1}{2}$) and later extended to several instances of a general G_α (see the monograph by Bakushinskii and Kokurin [2], as well as [22]) including those methods G_α that are considered in this paper (see section 3.3). Actually, it can be shown that the same holds true for the Newton–Kaczmarz method (1.10) for multiple equations.

3. Convergence analysis. In this section we will state a quite general convergence theorem. Its proof is closely related to convergence proofs in [6, 19, 20, 21, 22]. Therefore we shall provide the proof in a somewhat compressed form but highlight the important ideas for the convenience of the reader. We aim at giving the statements in a general and comprehensive way so that they might be of interest even for the special case $p = 1$, i.e., for (1.1). Especially, according to the authors' knowledge, the convergence result with logarithmic source conditions under nonlinearity assumptions of the type considered here is new also for $p = 1$.

3.1. Preliminaries and assumptions. To be able to carry out the estimates in the proof of Theorem 3.1, we have to make some assumptions on the regularization methods G^i, additional to (1.5), (1.6) in the introduction, i.e.,

$$G^i_\alpha(K)y \to K^\dagger y \quad \text{as } \alpha \to 0 \quad \forall y \in \mathcal{R}(K) \tag{3.1}$$

and

$$\|G^i_\alpha(K)\| \leq \Phi(\alpha) \tag{3.2}$$

for all $K \in \mathcal{L}(X, Y_i)$ with $\|K\| \leq C^i_S$. In view of the nonlinearity condition (2.2), we assume that

$$\|G^i_\alpha(KR)KR - G^i_\alpha(K)K\| \leq \bar{C}_G \|R - I\| \tag{3.3}$$

for all $K \in \mathcal{L}(X, Y_i)$, $R \in \mathcal{L}(X, X)$ with $\|K\|, \|KR\| \leq C^i_S$, $\|R - I\| \leq c < 1$, as well as

$$\|G^i_\alpha(K)K\| \leq C_G \qquad \forall K \in \mathcal{L}(X, Y_i) : \ \|K\| \leq C^i_S \tag{3.4}$$

with positive real constants $\bar{C}_G$, c, and C^i_S as in (2.1). To yield convergence rates under additional regularity conditions (2.8), the regularizing operators G^i_α have to converge to the inverse of K at some rate on the set of solutions satisfying (2.8), i.e., a condition of the form

$$\|(I - G^i_\alpha(K)K) f(K^*K)\| \leq \psi(\alpha) \qquad \forall K \in \mathcal{L}(X, Y_i) : \quad \|K\| \leq C^i_S \tag{3.5}$$

is needed, with a strictly monotone function ψ that decreases to zero as $\alpha \to 0$. Moreover, the sequence $\psi(\alpha_n)$ must not tend to zero too fast, in the sense that

$$\frac{\psi(\alpha_n)}{\psi(\alpha_{n+1})} \leq C_\psi \qquad \forall n \in \mathbb{N} \tag{3.6}$$

for some constant $C_\psi \in \mathbb{R}^+$.

In the situation of noisy data, convergence of the reconstructions as the noise level δ tends to zero is only obtained for appropriate choices of the stopping index $N = N(\delta)$ in dependence of the noise level δ. In the general case, convergence can be achieved if $N(\delta)$ is chosen such that

$$N(\delta) \to \infty \quad \text{and} \quad \Phi(\alpha_{N(\delta)}) \cdot \delta \to 0 \qquad \text{as } \delta \to 0 \tag{3.7}$$

and

$$\Phi(\alpha_n)\cdot\delta\le\tau\qquad\forall n\le N(\delta) \tag{3.8}$$

for some $\tau>0$ sufficiently small. If additional source conditions (2.8) hold, an appropriate choice is such that

$$\phi(\alpha_{N(\delta)})\le\delta<\phi(\alpha_n)\qquad\forall n\le N(\delta), \tag{3.9}$$

where

$$\phi(\alpha)=\frac{\tau\psi(\alpha)}{\Phi(\alpha)}$$

for τ defined in (3.8).

3.2. Main result. Now we shall state and prove the main convergence result of this paper, a comprehensive convergence theorem for Newton–Kaczmarz methods.

THEOREM 3.1. *Let x_n be defined by the sequence* (1.10) *with Fréchet differentiable operators F_i satisfying* (2.1), (2.2) *with* (2.3), *data y^δ satisfying* (1.4), *the regularization methods G_α^i fulfilling* (3.1), (3.2), (3.3), (3.4), *for all $K\in\mathcal{L}(X,Y_i)$, $R\in\mathcal{L}(X,X)$ with $\|K\|,\|KR\|\le C_S^i$, $\|R-I\|\le c<1$, and* (3.5), *as well as a sequence α_n tending to zero and satisfying* (3.6). *Moreover, let τ and $\|x_{0,i}-x^\dagger\|$ be sufficiently small and $x_{0,i}-x^\dagger\in\mathcal{N}(F_i'(x^\dagger)^\perp$, $i=0,\dots,p-1$.*

Then, in the noise-free case ($\delta=0$), the sequence x_n converges to $x^\dagger$ as $n\to\infty$. In case of noisy data and with the choice (3.7), (3.8), *$x_{N(\delta)}$ converges to $x^\dagger$ as $\delta\to0$.*

If the source conditions (2.8) *and* (3.5), (3.6) *hold, with $\|w^i\|$ sufficiently small, then the convergence rates*

$$\|x_n-x^\dagger\|=O(\psi(\alpha_n))$$

in the noise-free situation and, with (3.9),

$$\|x_{N(\delta)}-x^\dagger\|=O(\psi(\phi^{-1}(\delta)))$$

in the noisy situation, respectively, hold.

Proof. We will make use of the following lemma, whose proof can be found in [21].

LEMMA 3.2. *Let $\{a_n\}$ be a sequence satisfying*

$$0\le a_n\le a\quad\text{and}\quad\lim_{n\to\infty}a_n=\tilde a\le a.$$

Moreover, we assume that $\{\gamma_n\}$ is a sequence for which the estimate

$$0\le\gamma_{n+1}\le a_n+b\gamma_n+c\gamma_n^2\ ,\qquad n\in\mathbb{N}_0,\ \gamma_0\ge0,$$

holds for some $b,c\ge0$. Let $\underline{\gamma}$ and $\overline{\gamma}$ be defined as

$$\underline{\gamma}:=\frac{2a}{1-b+\sqrt{(1-b)^2-4ac}},\quad\overline{\gamma}:=\frac{1-b+\sqrt{(1-b)^2-4ac}}{2c}.$$

If $b+2\sqrt{ac}<1$ and if $\gamma_0\le\overline{\gamma}$, then

$$\gamma_n\le\max\left(\gamma_0,\underline{\gamma}\right)\ ,\quad n\in\mathbb{N}_0,$$

and if $\tilde{a} < a$, *then*

$$\limsup_{n\to\infty} \gamma_n \le \frac{2\tilde{a}}{1-b+\sqrt{(1-b)^2-4\tilde{a}c}}.$$

To derive a recursive error estimate, we assume that the current iterate x_n is in $\mathcal{B}_\rho(x^\dagger)$ and that $n < N(\delta)$ ($= \infty$ if $\delta = 0$). Then

$$\begin{aligned} x_{n+1} - x^\dagger \;=\; & \Big(I - G^n_{\alpha_n}(F_n'(x^\dagger))F_n'(x^\dagger)\Big)(x_{0,i} - x^\dagger) \\ & + \Big(G^n_{\alpha_n}(F_n'(x^\dagger))F_n'(x^\dagger) - G^n_{\alpha_n}(F_n'(x_n))F_n'(x_n)\Big)(x_{0,i} - x^\dagger) \\ & - G^n_{\alpha_n}(F_n'(x_n))(F_n(x_n) - F_n(x^\dagger) - F_n'(x_n)(x_n - x^\dagger)) \\ & - G^n_{\alpha_n}(F_n'(x_n))(y_n - y_n^\delta). \end{aligned} \tag{3.10}$$

The third term on the right-hand side can be rewritten as

$$G^n_{\alpha_n}(F_n'(x_n))F_n'(x_n)\int_0^1 \Big(R^i(x^\dagger + \theta(x_n - x^\dagger), x_n) - I\Big)d\theta(x_n - x^\dagger)$$

with $i = \mathrm{mod}(n, p)$, so

$$\begin{aligned} \|x_{n+1} - x^\dagger\| \;\le\; & \xi_n \\ & + \bar{C}_G C_R \|x_{0,i} - x^\dagger\| \;\|x_n - x^\dagger\| \\ & + \tfrac{1}{2} C_G C_R \|x_n - x^\dagger\|^2 \\ & + \Phi(\alpha_n)\delta, \end{aligned} \tag{3.11}$$

where

$$\xi_n := \|\Big(I - G^n_{\alpha_n}(F_n'(x^\dagger))F_n'(x^\dagger)\Big)(x_{0,i} - x^\dagger)\| \le \psi(\alpha_n)\|w^i\|,$$

if (2.8) holds and

$$\xi_n \to 0 \quad \text{as } n \to \infty \tag{3.12}$$

also without (2.8). The latter can be seen by (3.1) together with the following subsequence-subsequence argument.

Let $(\xi_{n_m})_{m\in\mathbb{N}}$ be an arbitrary subsequence of $(\xi_n)_{n\in\mathbb{N}}$. Then there exists an $i \in \{0, \dots, p-1\}$ such that the set $\{m \in \mathbb{N} \mid mod(n_m, p) = i\}$ has infinite cardinality. Define by $(m_l)_{l\in\mathbb{N}}$ a numbering of this set in ascending order; then for $(\xi_{n_{m_l}})_{l\in\mathbb{N}}$ we get

$$\xi_{n_{m_l}} = \|\Big(I - G^i_{\alpha_{n_{m_l}}}(F_i'(x^\dagger))F_i'(x^\dagger)\Big)(x_{0,i} - x^\dagger)\| \;\to 0 \text{ for } l \to \infty,$$

since $\alpha_{n_{m_l}} \to 0$ for $l \to \infty$.

Now we can apply induction together with Lemma 3.2 to the sequence

$$\gamma_n := \|x_n - x^\dagger\|.$$

The boundedness (3.8) in the stopping rule and our assumption on closeness of $x_{0,i}$ to $x^\dagger$ and on smallness of τ permit us to make the constants a and b sufficiently small so that the assumptions of the lemma are satisfied, and the bound $\max\{\gamma_0, \underline{\gamma}\}$ is smaller

than ρ, so that we can guarantee that the iterates remain in $\mathcal{B}_\rho(x^\dagger)$ for all $n \le N(\delta)$. Moreover, by (3.12) as well as the asymptotics (3.7) in the stopping rule, we can set $\tilde{a} = 0$ and conclude that x_n converges to $x^\dagger$ as $n \to \infty$ in the noise-free case and as $\delta \to 0$ in the noisy case, respectively.

To prove convergence rates under source conditions, we consider the sequence

$$\gamma_n := \frac{\|x_n - x^\dagger\|}{\psi(\alpha_n)},$$

which satisfies

$$\gamma_{n+1} \le C_\psi \left(\|w^i\| + \bar{C}_G C_R \|x_{0,i} - x^\dagger\| \gamma_n + \frac{1}{2} C_G C_R \psi(\alpha_n) \gamma_n^2 + \frac{\Phi(\alpha_n)}{\psi(\alpha_n)} \delta \right).$$

Hence, Lemma 3.2 together with the stopping rule (3.9) imply that x_n remains in $\mathcal{B}_\rho(x^\dagger)$ for all $n \le N(\delta)$ and that γ_n is uniformly bounded, i.e.,

$$\|x_n - x^\dagger\| \le C\psi(\alpha_n), \tag{3.13}$$

for some constant C. This immediately yields the convergence rate result in the noiseless case. To obtain the error estimate in terms of δ in the noisy case, we make use of the fact that by (3.9)

$$\delta \ge \phi(\alpha_{N(\delta)}),$$

which, since ψ and ϕ are strictly monotonically increasing, by (3.13) implies

$$\psi(\phi^{-1}(\delta)) \ge \psi(\alpha_{N(\delta)}) \ge \frac{1}{C} \|x_{N(\delta)} - x^\dagger\|. \qquad \square$$

The assumption

$$x_{0,i} - x^\dagger \in \mathcal{N}(F_i'(x^\dagger))^\perp, \quad i = 0, \dots, p-1, \tag{3.14}$$

is rather limiting, since the dimensionality of $x_0 - x^\dagger$ is related to the "smaller" space $\mathcal{N}(F_i'(x^\dagger)^\perp$. In the special case $p = 1$, the difference between x_0 and an x_0-minimum-norm-solution $x^\dagger$ will automatically lie within $\mathcal{N}(F'(x^\dagger))^\perp$ under certain nonlinearity conditions (see Proposition 2.1 in [21]). However, for general $p > 1$ one gets only $x_0 - x^\dagger \in (\bigcap_{i=0}^{p-1} \mathcal{N}(F_i'(x^\dagger)))^\perp$ and not (3.14) with $x_{0,i} := x_0$. Thus, condition (3.14) requires the choice of appropriate initial guesses $x_{0,i}$. To see the necessity of condition (3.14) for convergence, consider the linear case

$$F_i x = y_i \ , \qquad i = 0, \dots p-1, \tag{3.15}$$

with $F_i \in \mathcal{L}(X, Y_i)$, $y_i \in Y_i$, $i = 0, \dots p-1$, where the sequence x_n is defined by

$$x_{n+1} = x_{0,n} - G^n_{\alpha_n}(F_n)(F_n x_{0,n} - y_n^\delta). \tag{3.16}$$

In the case of exact data, the error can be written as

$$x_{kp+i+1} - x^\dagger = (I - G^i_{\alpha_{kp+i}}(F_i)F_i)(x_{0,i} - x^\dagger) \tag{3.17}$$

for $n = kp + i$, $k \in \mathbb{N}$, so by (3.1) and $\alpha_n \to \infty$ as $n \to \infty$,

$$x_{kp+i+1} - x^\dagger \to \mathcal{P}_{\mathcal{N}(F_i)^\perp}(x_{0,i} - x^\dagger) \text{ as } k \to \infty,$$

whence convergence of x_n to $x^\dagger$ as $n \to \infty$ implies (3.14).

In this sense, Theorem 3.1 means that the regularized Newton–Kaczmarz method is as least as good as application of Newton's method separately to each of the p equations, which might a priori not be evident due to the mixing up of the equations during the iteration (1.10). Since it takes into account more information, it should intuitively be even better, which is also reflected in our numerical tests, that showed convergence without any specific choice of the initial guesses.

Note that in the linear case, subsequent iterates completely decouple, i.e., subsequences $(x_{kp+i_1})_{k\in\mathbb{N}}$, $(x_{kp+i_2})_{k\in\mathbb{N}}$ are independent of each other for $i_1 \neq i_2$. Thus it suffices to have

$$x_{0,i} - x^\dagger \in \mathcal{N}(F_i)^\perp \tag{3.18}$$

for one $i = \bar{i} \in \{0, \dots, p-1\}$, to obtain convergence of the respective subsequence $x_{kp+\bar{i}+1}$ from standard results for linear regularization methods. The same holds true for convergence rates. Consequently, in order to get convergence (and convergence rates) with noisy data, it suffices to have (3.18) (and $x_{0,i} - x^\dagger \in \mathcal{R}(f(F_i^*F_i))$) for one $i = \bar{i} \in \{0, \dots, p-1\}$ only, and to stop the iteration at an index from the respective subsequence $kp + \bar{i} + 1$ with $k_* = k_*(\delta)$ being determined a priori from (3.7), (3.8), (3.9) or, alternatively, a posteriori from a discrepancy principle

$$\|F_i x_{k_* p+\bar{i}+1} - y_i\| \leq \tau\delta < \|F_i x_{kp+\bar{i}+1} - y_i\|, \qquad 0 \leq k < k_*.$$

Unfortunately this complete decoupling gets lost as soon as the operators F_i are nonlinear. Moreover, we have to remark that already (3.18) for one $i = \bar{i}$ is a very strong condition in case $p > 1$, since it means the other equations for $i \neq \bar{i}$ are not really required for determining $x^\dagger$.

3.3. Standard regularizing operators. Now we shall apply Theorem 3.1 to some regularization methods G^i of particular interest. Moreover, in the abstract source condition (2.8), we insert the most relevant special cases of a Hölder function f in (2.9) or a logarithmic function f in (2.10).

As important examples from a larger class of regularization methods defined by real functions $g_\alpha : \mathbb{R}^+ \mapsto \mathbb{R}^+$ approximating $\lambda \mapsto \frac{1}{\lambda}$ and

$$G_\alpha(K) := g_\alpha(K^*K)K^* \tag{3.19}$$

via functional calculus (cf., e.g., [11, 25]), we consider the following.

- Tikhonov–Philips regularization:

$$G_\alpha(K) = (K^*K + \alpha I)^{-1}K^*, \quad I - G_\alpha(K)K = \alpha(K^*K + \alpha I)^{-1}. \tag{3.20}$$

  In this case, we shall call the arising iterative method the *iteratively regularized Gauss–Newton–Kaczmarz (IRGNK) method*.

- Iterated Tikhonov regularization:

$$\begin{aligned} G_\alpha(K) &= \textstyle\sum_{l=0}^k \prod_{j=l}^k \alpha_j (K^*K + \alpha_j I)^{-1} \frac{1}{\alpha_l} K^*, \\ I - G_\alpha(K)K &= \textstyle\prod_{l=0}^k \alpha_l (K^*K + \alpha_l I)^{-1}, \end{aligned} \tag{3.21}$$

  with the effective regularization parameter

$$\alpha := \frac{1}{\sum_{l=0}^k \frac{1}{\alpha_l}}.$$

We shall call the arising iterative method the *k-iteratively regularized Gauss–Newton–Kaczmarz (IRGNK$_k$) method*. Here we distinguish between the special stationary case

$$\alpha_l :\equiv 1, \tag{3.22}$$

i.e., Lardy's method, and the (due to its faster convergence more attractive, cf. [16]) nonstationary case of, e.g., geometrically decaying α_l

$$\alpha_l := Cq^l \tag{3.23}$$

with $q \in (0,1)$.

- Landweber iteration:

$$G_\alpha(K) = \sum_{l=0}^{k} (I - K^*K)^l K^*, \quad I - G_\alpha(K)K = (I - K^*K)^{k+1}, \tag{3.24}$$
$$\alpha := \frac{1}{k+1},$$

where the scaling is assumed to be done such that $\|I - K^*K\| \leq 1$, i.e., $C_S^i = \sqrt{2}$ in (2.1). For obvious reasons, this method shall be called *Newton–Landweber–Kaczmarz (NLK) method* here and below.

These methods are well known to satisfy (3.1), (3.2) with

$$\Phi(\alpha) = C\frac{1}{\sqrt{\alpha}},$$

as well as (3.4) (cf., e.g., [11, 25, 16]). Moreover, for the Hölder-type source representation functions (2.9), they satisfy (3.5) with

$$\psi(\alpha) = C\alpha^\nu \tag{3.25}$$

(where ν is restricted to the interval $[0,1]$ in Tikhonov regularization, and to the interval $[0,k]$ in iterated Tikhonov regularization), from which one can conclude by Lemma 4 in [20] that they also satisfy (3.5) for the logarithmic functions (2.10) with

$$\psi(\alpha) = C(-\ln(\alpha))^{-\mu},$$

where w.l.o.g. both $\|K\|^2$ and α are restricted to the interval $(0, \exp(-1)]$ (i.e., $C_S^i = \exp(-1/2)$ in (2.1)) in order to avoid the singularity of f_μ^L at zero. Therewith, a decay restriction

$$\frac{\alpha_n}{\alpha_{n+1}} \leq C_\alpha \qquad \forall n \in \mathbb{N} \tag{3.26}$$

is sufficient for (3.6).

COROLLARY 3.3. *Let x_n be defined by the sequence* (1.10) *with Fréchet differentiable operators F_i satisfying* (2.1), (2.2) *with* (2.3), *data y^δ satisfying* (1.4) *and the regularization methods G_α^i defined by Tikhonov–Philips regularization, nonstationary iterated Tikhonov regularization, or Landweber iteration, as well as a sequence α_n tending to zero and satisfying* (3.26). *Moreover, let τ and $\|x_{0,i} - x^\dagger\|$ be sufficiently small and $x_{0,i} - x^\dagger \in \mathcal{N}(F_i'(x^\dagger)^\perp$, $i = 0, \ldots, p-1$.*

Then, the assertions of Theorem (3.1) *hold. In particular, under a Hölder-type source condition* (2.8) *with* (2.9)*, we obtain*

$$\|x_{N(\delta)} - x^\dagger\| = O(\delta^{\frac{2\nu}{2\nu+1}})$$

(where ν is restricted to $[0,1]$ in case of Tikhonov regularization), and under a logarithmic type source condition (2.8) *with* (2.10)

$$\|x_{N(\delta)} - x^\dagger\| = O((-\ln(\delta^2))^{-\mu}).$$

Note that the saturation of iterated Tikhonov regularization at $\nu = k$ does not take effect here, since we do not consider k but $(\sum_{l=0}^{k} \alpha_l^{-1})^{-1}$ as the regularization parameter.

Proof. It remains to show that the differences between applications of the regularization methods to two different operators can be estimated according to (3.3). Tikhonov regularization can make use of estimates presented in [21], as well as in Hohage's thesis [19], namely, for arbitrary $K \in \mathcal{L}(X, Y_i)$, $R \in \mathcal{L}(X, X)$ with $\|R - I\| \le c < 1$,

$$\begin{aligned}
&\|G_\alpha^i(KR)KR - G_\alpha^i(K)K\| \\
&= \alpha \|(K^*K + \alpha I)^{-1} - ((KR)^*KR + \alpha I)^{-1}\| \\
&= \alpha \|((KR)^*KR + \alpha I)^{-1} \Big((KR)^*KR(I - R^{-1}) + (R - I)^*K^*K \Big) (K^*K + \alpha I)^{-1}\| \\
&\le (1 + \tfrac{1}{1-c}) \|R - I\|
\end{aligned}$$

for f according to (2.9) with $\nu \le \frac{1}{2}$ or f according to (2.10).

For the iterative methods—iterated Tikhonov regularization and Landweber iteration—we make use of the identity

$$\prod_{l=0}^{k} A_l - \prod_{l=0}^{k} B_l = \sum_{l=0}^{k} \prod_{j=0}^{l-1} A_j (A_l - B_l) \prod_{j=l+1}^{k} B_j \tag{3.27}$$

for linear operators A_l, B_l, with the notation $\prod_{l=0}^{-1} A_l = I = \prod_{l=k+1}^{k} B_j$, and first consider case (a): to obtain (3.3) for Landweber iteration, we set

$$A_l := (I - (KR)^*KR), \qquad B_l := (I - K^*K), \tag{3.28}$$

and use the fact that

$$A_l - B_l = (KR)^*KR(R^{-1} - I) + (I - R^*)K^*K \tag{3.29}$$

to derive

$$\begin{aligned}
G_\alpha^i(K)K - G_\alpha^i(KR)KR = &\sum_{l=0}^{k} \prod_{j=0}^{l-1} A_j (KR)^*KR(R^{-1} - I) \prod_{j=l+1}^{k} B_j \\
&+ \sum_{l=0}^{k} \prod_{j=0}^{l-1} A_j (I - R^*)K^*K \prod_{j=l+1}^{k} B_j.
\end{aligned} \tag{3.30}$$

To estimate the sums from 0 to k we decompose them into sums from 0 to $[\frac{k}{2}]$ and from $[\frac{k}{2}] + 1$ to k and use the fact that

$$\left\| \prod_{j=0}^{l-1} A_j (KR)^*KR \right\| \le \frac{1}{l+1}, \qquad \left\| \prod_{j=l+1}^{k} K^*KB_l \right\| \le \frac{1}{k-l+1} \tag{3.31}$$

as well as

$$(KR)^*KR = I - A_l, \qquad K^*K = I - B_l \tag{3.32}$$

and a telescope sum trick to obtain, for the first sum on the right-hand side of (3.30),

$$\left\| \sum_{l=[\frac{k}{2}]+1}^{k} \prod_{j=0}^{l-1} A_j (KR)^*KR(R^{-1}-I) \prod_{j=l+1}^{k} B_j \right\| \le \sum_{l=[\frac{k}{2}]+1}^{k} \frac{1}{l+1} \|R^{-1}-I\| \le \|R^{-1}-I\|$$

and

$$\begin{aligned}
&\left\| \sum_{l=0}^{[\frac{k}{2}]} \prod_{j=0}^{l-1} A_j (KR)^*KR(R^{-1}-I) \prod_{j=l+1}^{k} B_j \right\| \\
&= \left\| \sum_{l=0}^{[\frac{k}{2}]} \prod_{j=0}^{l-1} A_j (I-A_l)(R^{-1}-I) \prod_{j=l+1}^{k} B_j \right\| \\
&= \left\| \sum_{l=1}^{[\frac{k}{2}]} \prod_{j=0}^{l-1} A_j (R^{-1}-I)(I-B_l) \prod_{j=l+1}^{k} B_j \right. \\
&\qquad \left. + (R^{-1}-I) \prod_{j=1}^{k} B_j + \prod_{j=0}^{[\frac{k}{2}]} A_j (R^{-1}-I) \prod_{j=[\frac{k}{2}]+1}^{k} B_j \right\| \\
&\le \left(\sum_{l=1}^{[\frac{k}{2}]} \frac{1}{k-l+1} + 2 \right) \|R^{-1}-I\| \le 3\|R^{-1}-I\|
\end{aligned}$$

and analogously for the second sum on the right-hand side of (3.30).

For iterated Tikhonov regularization, the estimates can be obtained analogously, this time with

$$\begin{aligned}
&A_l := \alpha_l((KR)^*KR + \alpha_l I)^{-1}, \qquad B_l := \alpha_l(K^*K + \alpha_l I)^{-1}, \\
&A_l - B_l = A_l \frac{1}{\alpha_l} \Big((KR)^*KR(R^{-1}-I) + (I-R^*)K^*K \Big) B_l, \\
&G_\alpha^i(K)K - G_\alpha^i(KR)KR = \sum_{l=0}^{k} \prod_{j=0}^{l} A_j (KR)^*KR(R^{-1}-I) \prod_{j=l}^{k} B_j \\
&\qquad\qquad + \sum_{l=0}^{k} \prod_{j=0}^{l} A_j (I-R^*)K^*K \prod_{j=l}^{k} B_j,
\end{aligned} \tag{3.33}$$

$$\left\| \prod_{j=0}^{l} A_j (KR)^*KR \right\| \le \left(\sum_{j=0}^{l} \frac{1}{\alpha_j} \right)^{-1}, \qquad \left\| \prod_{j=l}^{k} K^*KB_l \right\| \le \left(\sum_{j=l}^{k} \frac{1}{\alpha_j} \right)^{-1},$$

and

$$\frac{1}{\alpha_l}(KR)^*KR\, A_l \;=\; I - A_l, \qquad \frac{1}{\alpha_l} B_l\, K^*K \;=\; I - B_l$$

in place of (3.28), (3.29), (3.30), (3.31), (3.32), respectively. In the stationary case (3.22), again cutting of the sum at $[\frac{k}{2}]$ and the telescope trick must be used, and in the nonstationary case (3.23), we apply the telescope sum trick to the whole first sum in (3.33) and leave the second sum unchanged. ∎

We finally want to mention that these results can be extended to the situation where discretization is applied to any of the standard regularization methods.

3.4. Levenberg–Marquardt–Kaczmarz. An alternative to considering the regularized Newton–Kaczmarz approach (1.10) is the generalization of a Levenberg–Marquardt method (cf. [15], (1.7)) to the situation of multiple equations in the following form:

$$x_{n+1} = x_n - (F_n'(x_n)^* F_n'(x_n) + \alpha_n I)^{-1} F_n'(x_n)^* (F_n(x_n) - y_n^\delta).$$

Note that this formally corresponds to the (intuitively optimal) formal choice of $x_{0,n} = x_n$ in (1.10), which, however, is not admissible in view of the convergence analysis given here, that requires a cyclic repetition of the starting guesses according to $x_{0,n} = x_{0,mod(n,p)}$.

Under a nonlinearity condition of the type (2.6) and with an appropriate a posteriori choice of the sequence α_n, along the lines of the proofs in [15], and similarly to [24], one can show that the error $\|x_n - x^\dagger\|$ is monotonically decreasing up to an index $n = N(\delta)$ determined by the discrepancy principle, without having to make assumptions of the type (3.14). Moreover, the norms of the residuals are squared summable in case of exact data and therewith

$$F_n(x_n) - y_n \to 0 \text{ as } n \to \infty. \tag{3.34}$$

This implies that there exists a weakly convergent subsequence of x_n. However, the limit of a weakly convergent subsequence $(x_{n_l})_{l\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ need not necessarily be a solution to (1.2), even if the F_i are (weakly) sequentially closed, i.e., for any sequence $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{D}(F_i)$ and $f_i \in Y_i$,

$$\Big(x_k \rightharpoonup x \wedge F_i(x_k) \to f_i\Big) \quad \Rightarrow \quad \Big(x \in \mathcal{D}(F_i) \wedge F_i(x) = f_i\Big). \tag{3.35}$$

Namely, if, e.g., $(x_{n_l})_{l\in\mathbb{N}} \subseteq (x_{mp+\bar{i}})_{m\in\mathbb{N}}$ for some $\bar{i} \in \{0, \dots, p-1\}$, then (3.34) and (3.35) imply that the weak limit of $(x_{n_l})_{l\in\mathbb{N}}$ is a solution of $F_{\bar{i}}(x) = y_{\bar{i}}$ only but not necessarily of $F_i(x) = y_i$ with $i \neq \bar{i}$. Also, strong convergence to $x^\dagger$ of x_n as $n \to \infty$ in the case of exact data or of $x_{N(\delta)}$ in the noisy situation cannot be proved by methods like those used in [15], [24], even in the linear case. Still, necessary convergence conditions on the initial guess can be expected to be less restrictive for (1.7) than for (1.10) as the linear case with bounded generalized inverses indicates: setting all regularization parameters α_n to zero we arrive at the error recursion

$$x_{n+1} - x^\dagger = \mathcal{P}_{N(F_n)}(x_n - x^\dagger) = \mathcal{P}_{N(F_n)}\mathcal{P}_{N(F_{n-1})} \cdots \mathcal{P}_{N(F_0)}(x_0 - x^\dagger),$$

so that one even obtains termination of the iteration with $x_{n+1} = x^\dagger$ as soon as $\mathcal{P}_{N(F_{n-1})} \cdots \mathcal{P}_{N(F_0)}(x_0 - x^\dagger) \in \mathcal{N}(F_n)^\perp$ for some n.

4. Numerical solution methods. In the following we discuss some possible discretization strategies and methods for the solution of the arising finite-dimensional problems.

4.1. Primal method. For all the optimization approaches discussed above, one can use a standard Galerkin discretization strategy by choosing a finite-dimensional subspace $X^h \subset X$ and solving a weak form of the discretized Newton equation for x_{n+1}^h. For the IRGNK method, we have $G_{\alpha_n}^n = M_n^{-1} F_n'(x_n^h)^*$ with the positive definite operator $M_n := F_n'(x_n^h)^* F_n'(x_n^h) + \alpha_n I$. Using this special form, we can discretize a step of the IRGNK method via

$$\langle M_n(x_{n+1}^h - x_{0,n}^h), \varphi\rangle = -\langle (F_n(x_n^h) - y_n^\delta - F_n'(x_n^h)(x_n^h - x_{0,n}^h)), F_n'(x_n^h)\varphi\rangle \qquad \forall\, \varphi \in X^h.$$

By iterating this discretization procedure k times, one obtains a discrete form of the IRGNK$_k$ method. Due to the positive definiteness of M_n, one can solve this problem iteratively by a preconditioned conjugate gradient method, where all standard preconditioners for the Tikhonov regularization can be used (cf. [34] for an overview).

In the case of the Newton–Landweber iteration, we obtain the same equation for each Landweber step finally leading to x_{n+1}^h but now with $M_n = I$, which gives a quasi-explicit form for the next iteration (one only has to invert a mass matrix corresponding to the identity operator, which does not even change during the iteration).

4.2. Dual method. In the following we shall consider a dual method for the IRGNK, i.e., the Newton–Kaczmarz method with the choice $G_\alpha(K) = (K^*K + \alpha I)^{-1}K^*$. We shall now derive a dual method, which is particularly suitable for the important case that the output spaces Y_i are of lower dimensionality than the parameter space X (which is the case for the examples considered above).

A first observation is that each iteration step of the IRGNK method is equivalent to the minimization problem

$$\frac{1}{2}\|F_n(x_n) + F_n'(x_n)(x - x_n) - y_n\|^2 + \frac{\alpha_n}{2}\|x - x_{0,n}\|^2 \to \min_{x\in X}. \tag{4.1}$$

By defining the right-hand side $z := y_n - F_n(x_n) - F_n'(x_n)x_n$ and the linear operator $K := F_n'(x_n)$, this optimization problem is of the form

$$J_1(Kx) + J_2(x) \to \min_{x\in X} \tag{4.2}$$

with (omitting the index n in the regularization parameter α_n)

$$J_1(y) = \frac{1}{2}\|y - z\|^2, \qquad J_2(x) = \frac{\alpha}{2}\|x - x_{0,n}\|^2.$$

Both the functionals J_1 and J_2 are convex, and therefore standard Fenchel duality (cf. [10]) implies that the primal problem (4.2) is equivalent to the dual problem

$$J_1^*(-v) + J_2^*(K^*v) \to \min_{v\in Y_n}, \tag{4.3}$$

where J_1^* and J_2^* are the conjugate functionals, which are obtained as

$$J_1^*(v) = \sup_{y\in Y_n} \langle v, y\rangle - J_1(y) = \frac{1}{2}\|v + z\|^2 - \frac{1}{2}\|z\|^2,$$

$$J_2^*(w) = \sup_{x\in X} \langle w, x\rangle - J_2(x) = \frac{1}{2\alpha}\|w + \alpha x_{0,n}\|^2 - \frac{\alpha}{2}\|x_{0,n}\|^2.$$

Moreover, the solution v of the dual problem (4.3) and the solution x of the primal problem are connected by the optimality condition

$$K^*v = J_2'(x) = \alpha(x - x_{0,n}).$$

Thus, we may compute $x = x_{0,n} + \frac{1}{\alpha}K^*v$ once we have solved the dual problem.

By ignoring the constant terms in the conjugate functionals, we may equivalently state the dual problem as

$$\frac{1}{2}\|-v+z\|^2+\frac{1}{2\alpha}\|K^*v+\alpha x_{0,n}\|^2\to\min_{v\in Y_n}, \tag{4.4}$$

which can be discretized, e.g., by the Ritz method on a subspace of Y_n, i.e., by minimizing the functional in (4.4) on a finite-dimensional subspace $Y_n^h\subset Y_n$. This automatically yields a discretization of the update in the primal space via $x_n^h-x_{0,n}=\frac{1}{\alpha}K^*v^h$, where v^h is the discrete solution of the dual problem.

The main advantage of a dual strategy is the (possible) lower dimensionality of the spaces Y_n, which yields smaller discrete problems and consequently a faster solution. In many important cases such as the examples presented above, the spaces Y_n do not depend on the iteration index but are the same for each step, such that one does not have to change the basis over the Kaczmarz sweep.

4.3. Primal-dual methods for PDE-constrained problems. As we have seen in the examples above, the operator F_i is defined implicitly via the solution of PDEs in many applications. We formally write the partial differential equation as a nonlinear operator equation of the form

$$E_i(u_i,q)=0,$$

where $E_i:\mathcal{U}\times X\to\mathcal{V}$ is a continuously differentiable nonlinear operator such that $\frac{\partial E_i}{\partial u}$ is regular for each $u\in\mathcal{U}$. The operator F_i is typically obtained as $F_i:=L_i\circ H_i$, where $H_i(q)=u_i$. We shall derive a primal-dual solution method in this case.

One step of the IRGNK method can be rewritten as the constrained problem

$$\frac{1}{2}\|L_nv+L_nu_n-y_n\|^2+\frac{\alpha_n}{2}\|s+q_n-q_{0,n}\|^2\to\min_{(v,s)}$$

subject to the constraint that $v=H_n'(q_n)s$, which can be expressed using the implicit function theorem as

$$\frac{\partial E_{n+1}}{\partial u}(u_n,q_n)v+\frac{\partial E_{n+1}}{\partial q}(u_n,q_n)s=0,$$

where $u_n=H_n(q_n)$. Deriving the KKT conditions for this constrained problem, we obtain an indefinite system for the primal variables v, s, and a dual variable w, given by

$$\begin{pmatrix} L_n^*L_n & 0 & A_n^* \\ 0 & \alpha_nI & B_n^* \\ A_n & B_n & 0 \end{pmatrix}\begin{pmatrix} v \\ s \\ w \end{pmatrix}=\begin{pmatrix} L_n^*y_n-L_n^*L_nu_n \\ \alpha_n(q_{0,n}-q_n) \\ 0 \end{pmatrix}$$

with the linear operators $A_n:=\frac{\partial E_{n+1}}{\partial u}(u_n,q_n)$ and $B_n:=\frac{\partial E_{n+1}}{\partial q}(u_n,q_n)$.

This indefinite system can be discretized using a mixed approach, i.e., we look for a solution (v^h,s^h,w^h) in the finite-dimensional subspaces $\mathcal{U}^h\times X^h\times\mathcal{V}^h$ satisfying

$$\begin{aligned}\langle L_nv,L_n\varphi\rangle+\langle A_n\varphi,w\rangle&=\langle y_n-L_nu_n,L_n\varphi\rangle,\\ \alpha_n\langle s,\sigma\rangle+\langle B_n\sigma,w\rangle&=\alpha_n\langle q_{0,n}-q_n,\sigma\rangle,\\ \langle A_nv,\psi\rangle+\langle B_ns,\psi\rangle&=0\end{aligned}$$

for all $(\varphi,\sigma,\psi)\in\mathcal{V}^h\times X^h\times\mathcal{U}^h$.

The resulting indefinite system can be solved by a preconditioned conjugate gradient method for the Schur complement, or directly by a preconditioned Krylov subspace method for indefinite systems like GMRES, QMR, or MINRES. We refer to [4, 5] for the discussion of solution methods for indefinite systems arising from primal-dual formulations in parameter identification.

5. Numerical examples. In the following we shall present numerical results for two of the examples introduced above.

5.1. Reconstruction with multiple sources. We start with numerical results for Example 2 in the one-dimensional domain $\Omega = (0, 1)$, using $p = 20$ localized sources of the form

$$h_i(x) = 10e^{-10(x-\frac{i+1}{p+1})^2}.$$

The data correspond to the "exact solution" $q^*(x) = 5 + 5x(1 - x)$ and the initial value is $q_0 \equiv 5$. Note that in general we cannot expect the least-squares minimum norm solution $q^\dagger$ to be equal to q^*, since we use only a finite number of measurements. However, we shall see below that the resulting limit $q^\dagger$ is close to q^*, with a difference probably caused due to the limited numerical resolution only.

For the numerical solution we use the iteratively regularized Gauss–Newton–Kaczmarz method, i.e., Tikhonov regularization in $H^1(\Omega)$ as the linear regularization method. The iteration is discretized using a primal-dual method as described above, with piecewise linear finite elements on a uniform grid of size $h = 0.01$.

We first test the convergence behavior in the noise-free case. To this end, we generate the data on the same grid as we later solve the inverse problem and choose the regularization parameters as

$$\alpha_n = \alpha_0 \zeta^{-n} \tag{5.1}$$

with $\zeta = 1.1$ and $\alpha_0 = 10^{-5}$. The convergence behavior is illustrated in Figures 5.1 and 5.2 by the iterates at several different steps. The behavior during the first Kaczmarz sweep is illustrated in Figure 5.1. In iterations 1 and 4, for which we use sources localized close to the left boundary $x = 0$, the convergence is more pronounced close to the left boundary. Vice versa, for later iterations (20), with sources localized close to the right boundary $x = 1$, the reconstruction is better close to the right boundary. In the medium stage of a Kaczmarz sweep, at iterate 8 with sources localized in the middle of the interval $(0, 1)$, the iterate appears almost symmetric. In the later stage of the iteration we plot the iterates q_n at $n = 50, 60$ (i.e., those in the middle and at the end of a Kaczmarz sweep) in Figure 5.2. One observes convergence of the algorithm, which turns out to be slightly faster for the iterates in the middle of the Kaczmarz sweep. The reason for this behavior is mainly the ordering of the sources; one will of course obtain a different behavior for different ordering. We finally provide a quantitative basis for the above observations on the behavior of the iterates in Figure 5.3, where we plot the development of the error $\|q^* - q_n\|$ (dashed, on the left) and of the residual $\|F_n(q_n) - y_n\|$ (on the right). In the left plot we also plot the error at the end of each Kaczmarz sweep $\|q^* - q_{np}\|$ (solid) and in the middle of the Kaczmarz sweep $\|q^* - q_{np+n/2}\|$ (dotted). In this example it turned out that the total error is not always decreasing, but the error at the same stage of the Kaczmarz sweep $\|q^* - q_{np+j}\|$ (for $0 \le j \le p-1$) is decreasing with n. In particular, it seems that the error at the beginning and end of the sweep is always the maximum one in the sweep, while the one in the middle of the sweep is always the minimum one. Since all of them decrease toward zero, we obtain the expected worst-case convergence, but of course

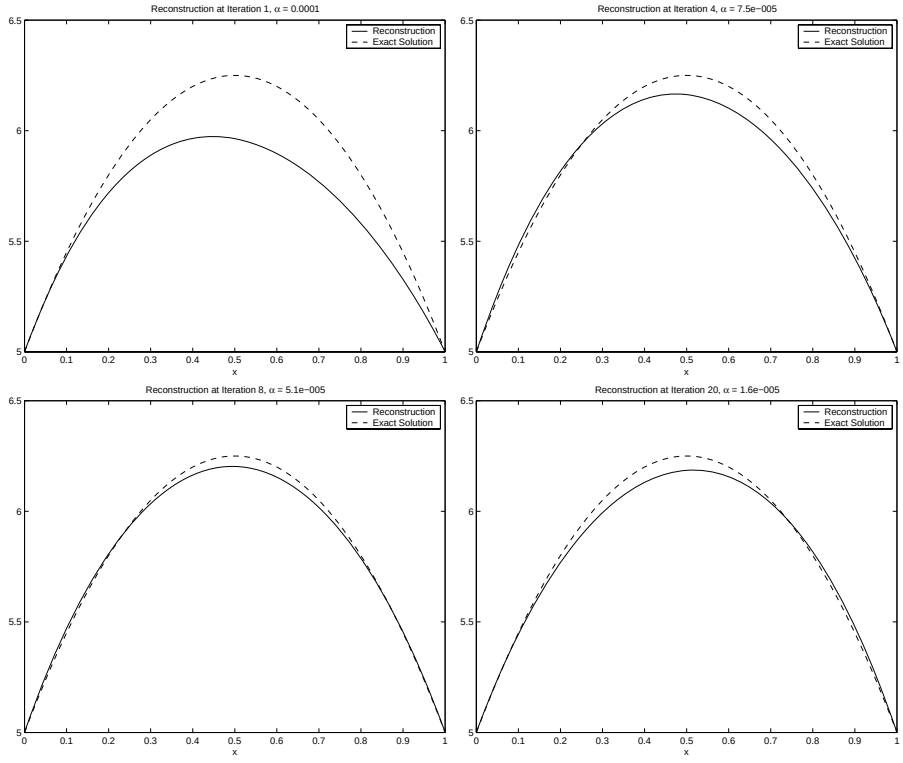


FIG. 5.1. *Reconstructions in the first example, $\delta = 0$, at iterates* 1, 4, 8, *and* 20.

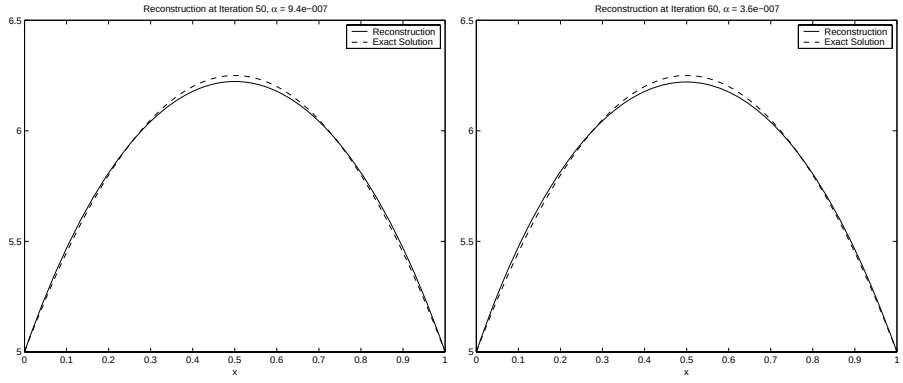


FIG. 5.2. *Reconstructions in the first example, $\delta = 0$, at iterates* 50 *and* 60.

in practice one should consider suitable orderings of the data y_i. The comparison of the residual at different iterates is even more difficult, since the operators and data are different in each step. However, we also obtain that $\|F_{np+j}(q_{np+j}) - y_{np+j}\|$ (for $0 \le j \le p-1$) is decreasing to zero with n.

For the noisy case we generated data on a finer grid of size $h = \frac{1}{347}$ in order to avoid inverse crimes. The resulting data are then perturbed using uniform random noise in the interval $[-\delta, \delta]$. The regularization parameters are chosen again via (5.1) with $\zeta = 1.1$ and $\alpha_0 = 10^{-2}$.

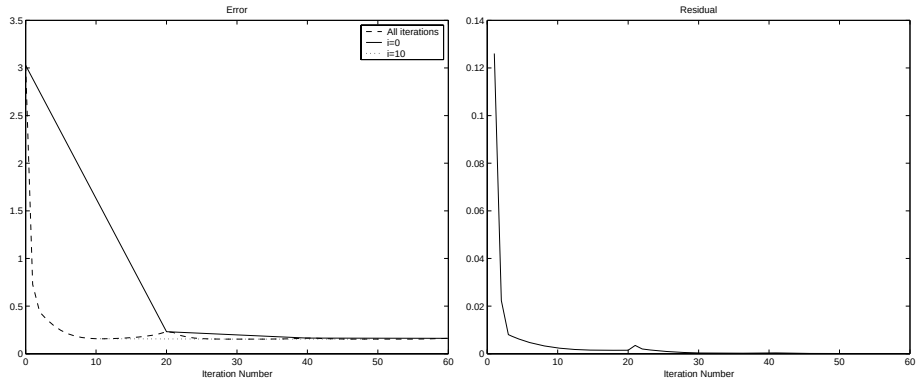


FIG. 5.3. *Plot of error (left) and residual (right) vs. iteration number in the first example,* $\delta = 0$.

We illustrate the reconstructions obtained for different noise levels (close to the minimum of the error during the iteration) in Figure 5.4. In clockwise order the plots show the reconstruction for noise level $\delta = 0.5\%$ (at iteration 90), $\delta = 1\%$ (at iteration 50), $\delta = 3\%$ (at iteration 30), and $\delta = 5\%$ (at iteration 30). One observes that the quality of the reconstruction improves with decreasing δ, i.e., the error of the iterate at the stopping index decreases with δ, thus confirming the convergence result for the noisy case. A quantitative monitoring of error and residual versus the iteration number is presented in Figure 5.5 for $\delta = 1\%$ (top), $\delta = 3\%$ (middle), and $\delta = 5\%$ (bottom). One also sees that the minimal error and residual obtained during the iteration decreases with δ as expected. As usual for ill-posed problems the error decreases only until some iteration step and then increases again although the residual is still decreasing. Note that this statement has to interpreted in a different sense, namely, for the subsequences $np + i$, $0 \le i \le p - 1$. Moreover, the variation in the error and residual during a sweep over the different sources increases with the noise level, which obviously makes the choice of the stopping index more difficult.

We finally investigate the effect of a different choice of regularization operators in each iteration step. With the localized sources we use it seems natural to localize the regularization. For this sake we compute the update from the regularized Newton equation

$$(F_n'(x_n)^* F_n'(x_n) + \alpha_n A + \beta_n B_n)(x_{n+1} - x_n) = -F_n'(x_n)^*(F(x_n) - y_n^\delta) + \alpha_n A(x_{0,n} - x_n),$$

where

$$Av := -\Delta v + v \qquad \text{and} \qquad B_n v := -\mathrm{div}(\omega_n \nabla v) + \omega_n,$$

with the weight functions $\omega_i := |x - \frac{i+1}{p+1}|$. The rationale behind the approach is that the additional part involving the operator B_n damps the update away from the point $\frac{i+1}{p+1}$, which can be considered as the location of the ith source. In this way too-large changes away from the measurement location are avoided, which usually decrease the quality of the reconstruction. The effect of the localized regularization is an improved convergence behavior, and the decay of error and residuals becomes less oscillatory, which is clearly an advantage with respect to the choice of the stopping index. Moreover, the shape of the reconstruction at different iterates depends less significantly on the localization of the last measurements used. We illustrate the

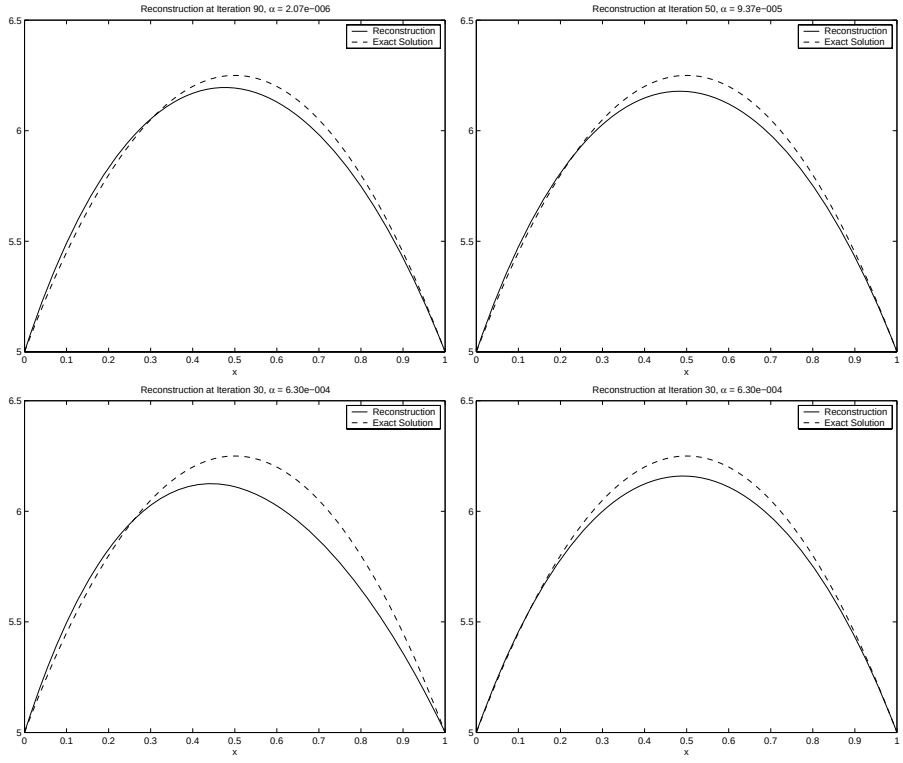


FIG. 5.4. *Reconstructions in the first example, for noise levels $\delta = 0.5\%$ (top left), $\delta = 1\%$ (top right), $\delta = 3\%$ (bottom right), $\delta = 5\%$ (bottom left).*

behavior in Figure 5.6 for 3% noise, in the same setup as above. The additional parameter β_n is chosen as $\beta_n = 0.5(1.01)^{-n}$. Compared to the same case in Figure 5.5 one observes a much smoother decay of the error and no significant differences in the behaviour at different steps of the sweep.

5.2. Reconstruction from Dirichlet–Neumann data. Our second numerical experiment is the solution of Example 1, i.e., the reconstruction of the coefficient q in

$$-\Delta u + qu = 0 \qquad \text{in } \Omega \subset \mathbb{R}^d$$

from $p = 20$ values of the Dirichlet-to-Neumann map. In our numerical example, the two-dimensional domain is $\Omega = (0,1)^2$, on which the differential equation is discretized by finite differences on a uniform grid of size $h = 0.025$.

The applied Dirichlet sources f_j are identically zero on three of the boundary segments and of the form

$$f_j(x_1,x_2) = \begin{cases} 10^3 e^{-50((x_1-j/6)^2} & \text{for } j = 1,\dots,5, x_2 = 0, \\ 10^3 e^{-50((x_1-(j-5)/6)^2} & \text{for } j = 6,\dots,10, x_2 = 1, \\ 10^3 e^{-50((x_2-(j-10)/6)^2} & \text{for } j = 11,\dots,15, x_1 = 0, \\ 10^3 e^{-50((x_2-(j-15)/6)^2} & \text{for } j = 16,\dots,20, x_1 = 1 \end{cases}$$

on the fourth segment, i.e., they approximate Dirac-delta impulses equally distributed over the boundary.

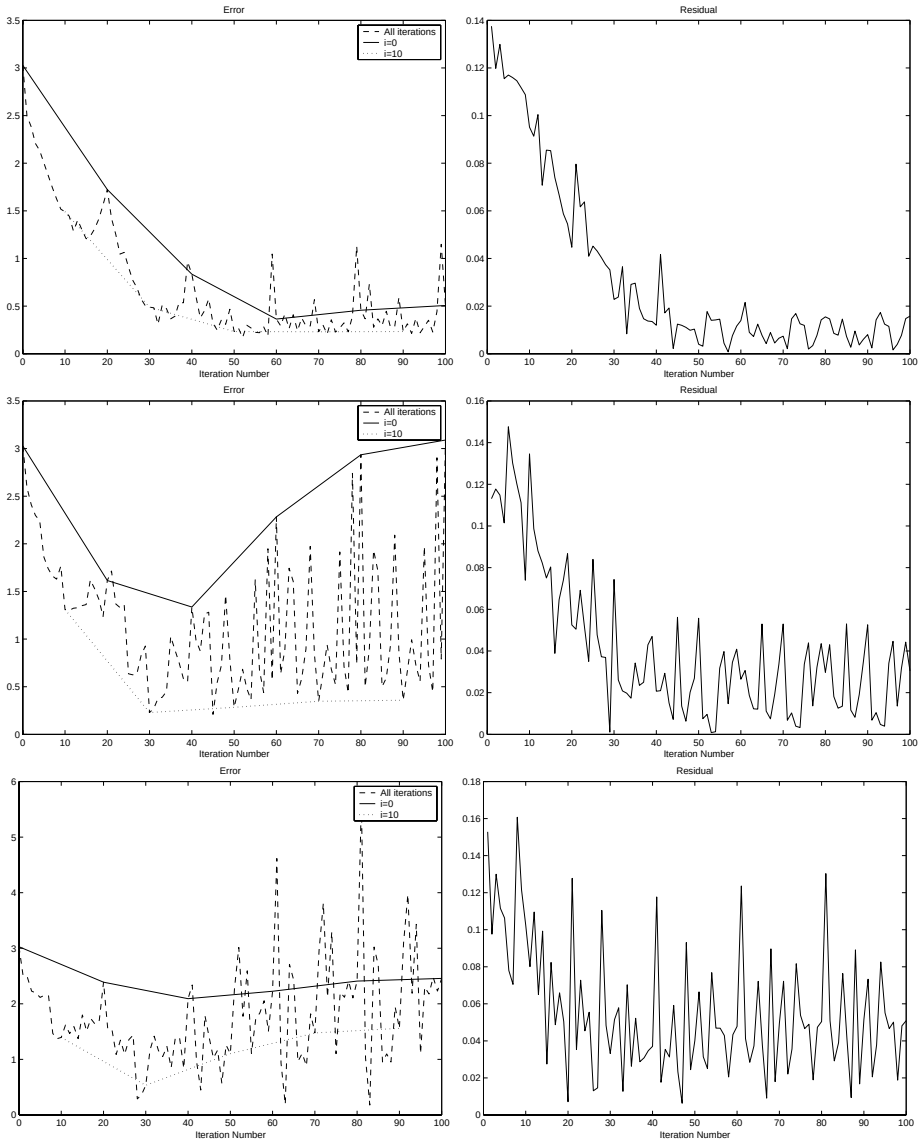


FIG. 5.5. *Plot of error (left) and residual (right) versus iteration number in the first example, for noise levels $\delta = 1\%$ (top), $\delta = 3\%$ (middle), $\delta = 5\%$ (bottom).*

In this case we use the Levenberg–Marquardt–Kaczmarz method, i.e., a Tikhonov-type stabilization in the H^1-norm in each step with prior q_n. This means that in each step of the method, the update $s = q_{n+1} - q_n$ is obtained by solving the minimization problem

$$\frac{1}{2}\left\|\frac{\partial v_n}{\partial \nu} - g_n\right\|^2_{H^{-1/2}(\partial\Omega)} + \frac{\alpha_n}{2}\|s\|^2_{H^1(\Omega)}$$

subject to the linear equation

$$-\Delta v_n + q_n v_n + s u_n = 0 \qquad \text{in } \Omega$$

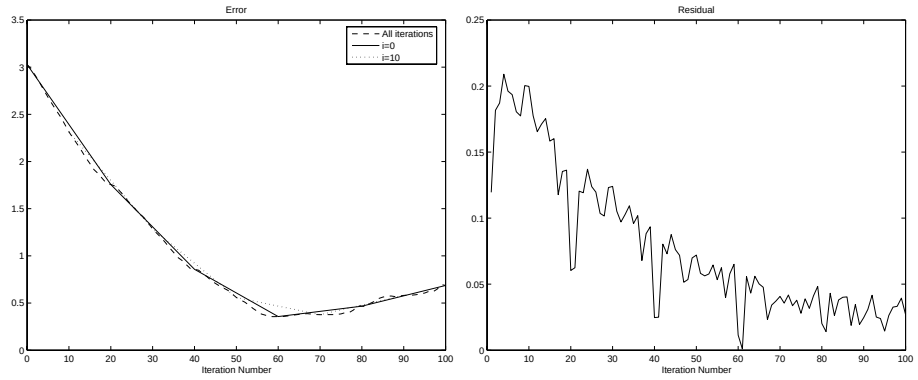


FIG. 5.6. *Plot of error (left) and residual (right) versus iteration number in the first example with localized regularization, for noise level* $\delta = 3\%$.

for v_n with homogeneous Dirichlet boundary values on $\partial\Omega$. The norm in $H^{-1/2}(\partial\Omega)$ of an element g is realized by

$$\|g\|_{H^{-1/2}(\partial\Omega)} := \|\phi_g\|_{H^1(\Omega)},$$

where $\phi_g \in H^1(\Omega)$ is the unique weak solution of

$$\int_\Omega (\nabla\phi_g \cdot \nabla\psi + \phi_g\psi)\ dx = \int_{\partial\Omega} g\psi\ d\sigma \qquad \forall\ \psi \in H^1(\Omega).$$

This means we have to solve an additional Neumann problem to evaluate the norm.

We use a primal-dual approach to discretize this problem, which means that we have to find two Lagrange multipliers corresponding to the partial differential equations for v_n and the function ϕ_g used to evaluate the norm. A careful investigation of the optimality system shows that ϕ_g can be eliminated in favor of one of the Lagrange multipliers, and the optimality system in each step becomes after straightforward transformations

$$\begin{aligned} -\Delta v_n + q v_n + s u_n &= 0, \\ -\Delta\lambda + q\lambda &= 0, \\ -\Delta\mu + \mu + (1-q)v_n - s u_n &= 0, \\ -\Delta s + s + \frac{1}{\alpha} u_n \lambda &= -\Delta(q_0 - q_n) + q_0 - q_n \end{aligned}$$

in Ω, supplemented by the boundary conditions

$$\begin{aligned} v_n &= 0, \\ \lambda - \mu &= \phi_n - \phi_{g_n}, \\ \frac{\partial\mu}{\partial\nu} &= 0, \\ s &= 0 \end{aligned}$$

on $\partial\Omega$. The functions ϕ_n and ϕ_{g_n} are the functions used to evaluate the $H^{-1/2}$-norm of $\frac{\partial u_n}{\partial\nu}$ and g_n, respectively, defined in the same way as ϕ_g above.

We start with some examples using data generated from the parameter

$$\hat{q} = 3 + 5\sin(\pi x)\sin(\pi y)$$

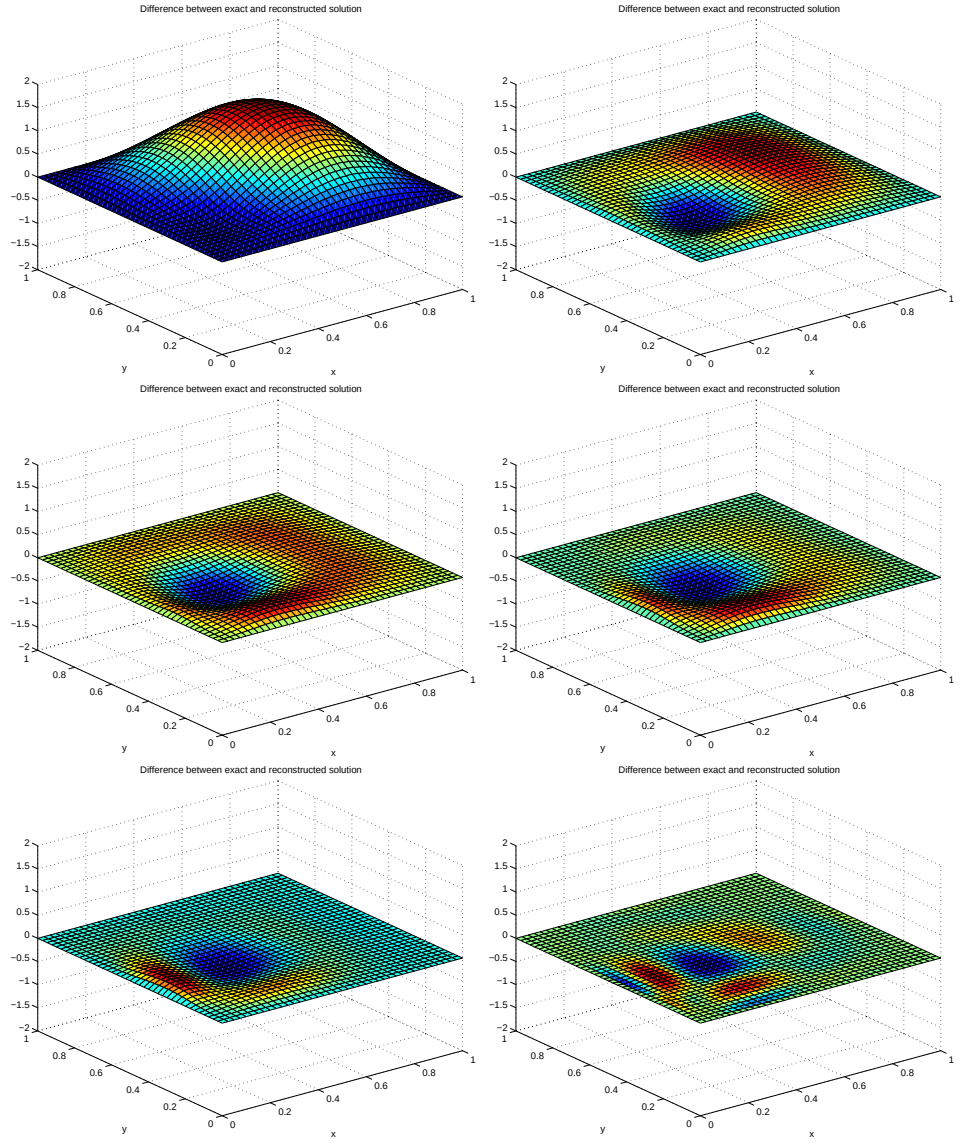


FIG. 5.7. *Difference* $\hat{q} - q_n$ *in the second example at iterates* 1, 2, 3, 5, 10, *and* 100.

and the starting value $q_0 \equiv 3$. Note that again $\hat{q}$ is not necessarily the minimum norm solution of the inverse problem with the above measurements, but since we expect that a successful reconstruction algorithm should at least approximate $\hat{q}$ and since we do not know the minimum norm solution, we measure the error as the difference between $\hat{q}$ and q_n. In order to test the convergence of exact data, we generate data on the same grid as the one used for solving the inverse problem and then perform the IRGNK algorithm with α_n chosen according to (5.1) with $\zeta = 1.05$ and $\alpha_0 = 10^{-8}$.

The difference between $\hat{q}$ and q_n is shown in Figure 5.7, at the iterates $n = 1, 2$ (top), $n = 3, 5$ (middle), and $n = 10, 100$ (bottom). One observes that the error is reduced very fast globally, but one also observes a certain local influence of the sources,

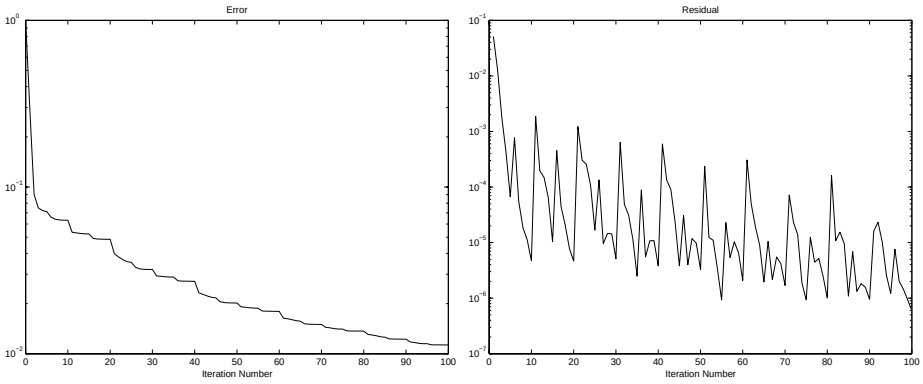


FIG. 5.8. *Semilogarithmic plot of error (left) and residual (right) versus iteration number in the second example,* $\delta = 0$.

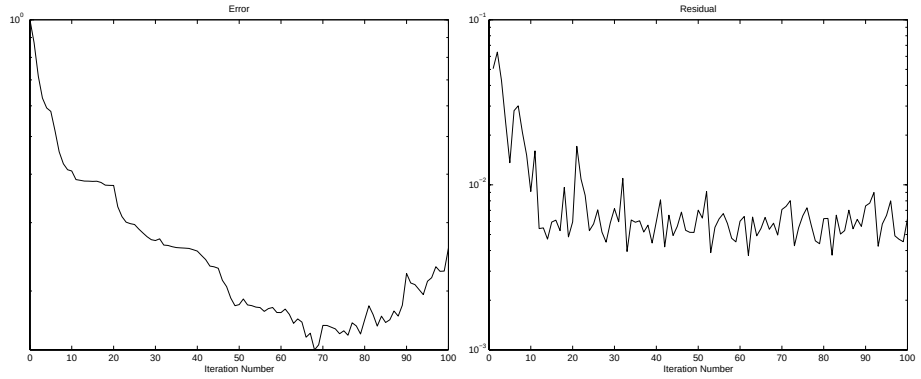


FIG. 5.9. *Semilogarithmic plot of error (left) and residual (right) versus iteration number in the second example,* $\delta = 1\%$.

i.e., the convergence seems faster closer to the support of the boundary sources. The quantitative development of the error $\|\hat{q} - q_n\|$ (left) and the residual $\|F(q_n) - g_n\|$ (right) are shown in a semilogarithmic scale in Figure 5.8.

Moreover, we test the behavior of the algorithm with respect to noise by using Gaussian random noise of variance $\delta = 1\%$ and $\delta = 0.5\%$. We plot the development of the error (left) and the residual (right) in a semilogarithmic scale in Figure 5.9 for $\delta = 1\%$ and in Figure 5.10 for $\delta = 0.5\%$. One observes the expected semiconvergence in both cases, i.e., the error reaches a minimum around which one should stop the iteration, and then starts to increase again. As expected, the minimal error appearing during the iteration decreases with the noise level, one obtains a minimal relative error 0.14 for $\delta = 1\%$ and 0.11 for $\delta = 0.5\%$.

We finally test the behavior for a more complicated exact parameter value

$$\hat{q} = 3 + 2\sin(3\pi x_1)\sin(2\pi x_2).$$

In this case we change the initial value α_0 to 10^{-12} due to the lower sensitivity of the data with respect to this parameter. The development of error and residual are

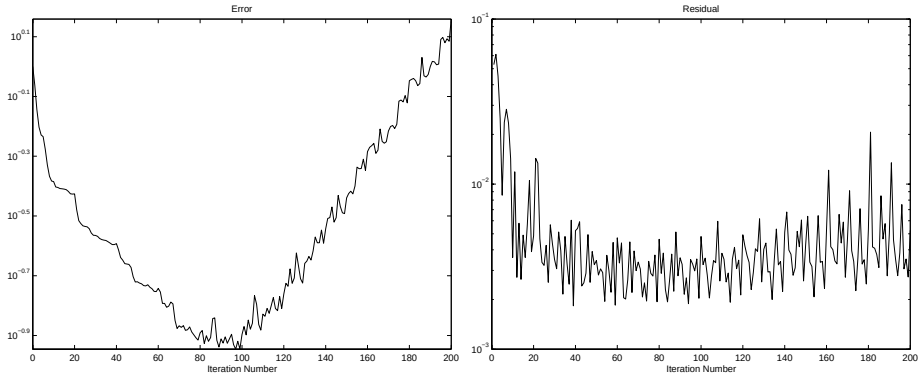


FIG. 5.10. *Semilogarithmic plot of error (left) and residual (right) versus iteration number in the second example,* $\delta = 0.5\%$.

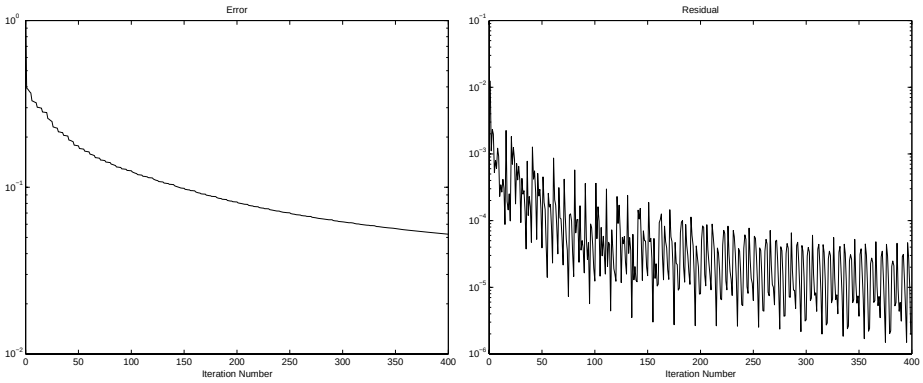


FIG. 5.11. *Semilogarithmic plot of error (left) and residual (right) versus iteration number in the second example for different exact solution,* $\delta = 0$.

shown in semilogarithmic scale in Figure 5.11. One observes that the method seems to converge in this case, too, although more slowly than in the above example, which is also caused by the lower sensitivity.

6. Conclusions and open problems. We have derived a detailed convergence analysis of regularized Newton–Kaczmarz methods for nonlinear ill-posed problems, which—as usual for ill-posed problems—can be carried out under certain conditions on the nonlinearity of the operators involved. As we have demonstrated in several examples from practice, these conditions seem not to be too restrictive in the case of Newton–Kaczmarz methods. Moreover, we have discussed the numerical solution of the linear problems arising in each step of the iteration method by three different approaches. The numerical experiments we carried out confirm the theoretical predictions.

So far, we have discussed a priori stopping rules (in the sense of [11]) only, whereas in practice it seems to be more important to have a posteriori stopping rules, which depend not only on the noise level δ but also on the actual data y^δ. In [24], a stopping rule is proposed for the Landweber–Kaczmarz method that is based on Morozov's discrepancy principle, i.e., a comparison of the residual norm with the noise level. Such

an approach would probably be appropriate also in our Newton–Kaczmarz context. However, we expect that a rigorous convergence analysis with such a residual type stopping criterion would have to be based on a nonlinearity assumption similar to (2.6) (as done in [24]) rather than the condition (2.2) that we have verified for our application examples here.

As mentioned in section 3.2, the condition (3.14) on the initial values poses a severe theoretical restriction that seems to be inevitable for Newton–Kaczmarz methods of the type (1.10) as the linear case shows. A possible way out might be to define the iteration by (1.7). Here the methods of proof considered so far for $p = 1$ (cf. [15]) rely on nonlinearity conditions of the type (2.5) but not on (2.4), in whose extension to $p > 1$, (2.2) we are interested here, however. Thus, new ideas would be necessary for proving convergence, perhaps based on a sweepwise instead of a stepwise analysis.

Acknowledgments. The major part of this work was carried out when the first author held a position in the Department of Mathematics at UCLA. The authors wish to thank Andreas Neubauer for valuable hints concerning this paper. Moreover, useful comments by the referees are gratefully acknowledged.

REFERENCES

[1] A. B. Bakushinskii, *The problem of the convergence of the iteratively regularized Gauss–Newton method*, Comput. Math. Math. Phys., 32 (1992), pp. 1353–1359.

[2] A. B. Bakushinskii and M. Yu. Kokurin, *Iterative Methods for Approximate Solution of Inverse Problems*, Springer, Dordrecht, The Netherlands, 2004.

[3] S. Biedenstein, *Numerische Verfahren zur Impedanztomographie*, Diploma thesis, University of Münster, Germany, 1997.

[4] M. Burger and W. Mühlhuber, *Iterative regularization of parameter identification problems by SQP methods*, Inverse Problems, 18 (2002), pp. 943–970.

[5] M. Burger and W. Mühlhuber, *Numerical approximation of an SQP-type method for parameter identification*, SIAM J. Numer. Anal., 40 (2002), pp. 1775–1797.

[6] P. Deuflhard, H. W. Engl, and O. Scherzer, *A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions*, Inverse Problems, 14 (1998), pp. 1081–1106.

[7] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[8] T. Dierkes, *Rekonstruktionsverfahren zur optischen Tomographie*, Ph.D. thesis, University of Münster, Germany, 2000.

[9] O. Dorn, H. Bertete-Aguirre, J. G. Berryman, and G. C. Papanicolaou, *A nonlinear inversion method for 3D-electromagnetic imaging using adjoint fields*, Inverse Problems, 15 (1999), pp. 1523–1558.

[10] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.

[11] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[12] W. Fang and K. Ito, *Reconstruction of semiconductor doping profile from LBIC image*, SIAM J. Appl. Math., 54 (1994), pp. 1067–1082.

[13] C. W. Groetsch, *Inverse Problems in Mathematical Sciences*, Vieweg, Braunschweig, Germany, 1993.

[14] M. Hanke, *Regularizing properties of a truncated Newton–CG algorithm for nonlinear inverse problems*, Numer. Funct. Anal. Optim., 18 (1997), pp. 971–993.

[15] M. Hanke, *A regularization Levenberg–Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems, 13 (1997), pp. 79–95.

[16] M. Hanke-Bourgeois and C. W. Groetsch *Nonstationary iterated Tikhonov regularization*, J. Optim. Theory Appl., 98 (1998), pp. 37–53.

[17] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[18] B. Hofmann and O. Scherzer, *Influence factors of ill-posedness for nonlinear ill-posed problems*, Inverse Problems, 10 (1994), pp. 1277–1297.

[19] T. Hohage, *Iterative Methods in Inverse Obstacle Scattering: Regularization Theory of Linear and Nonlinear Exponentially Ill-Posed Problems*, Ph.D. thesis, University of Linz, Austria, 1999.
[20] T. Hohage, *Regularization of exponentially ill-posed problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 439–464.
[21] B. Blaschke(-Kaltenbacher), A. Neubauer, and O. Scherzer, *On convergence rates for the iteratively regularized Gauss–Newton method*, IMA J. Numer. Anal., 17 (1997), pp. 421–436.
[22] B. Kaltenbacher, *Some Newton type methods for the regularization of nonlinear ill-posed problems*, Inverse Problems, 13 (1997), pp. 729–753.
[23] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, New York, 1996.
[24] R. Kowar and O. Scherzer, *Convergence analysis of a Landweber-Kaczmarz method for solving nonlinear ill-posed problems*, in Ill-Posed and Inverse Problems, S. Romanov, S. I. Kabanikhin, Y. E. Anikonov, A. L. Bukhgein, eds., VSP Publishers, Zeist, The Netherlands, 2002.
[25] A.K. Louis, *Inverse und schlecht gestellte probleme*, Teubner, Stuttgart, 1989.
[26] B. Lowe and W. Rundell, *The determination of multiple coefficients in a second order differential equation from input sources*, Inverse Problems, 9 (1993), pp. 469–482.
[27] B. Lowe and W. Rundell, *The determination of a coefficient in a parabolic equation from input sources*, IMA J. Appl. Math., 52 (1994), pp. 31–50.
[28] V.A. Morozov, *Regularization Methods for Ill-Posed Problems*, CRC Press, Boca Raton, FL, 1993.
[29] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner, Stuttgart, 1986.
[30] F. Natterer, *Numerical Solution of Bilinear Inverse Problems*, Technical Report, University of Münster, Germany, 1996.
[31] A. G. Ramm and A. B. Smirnova, *A numerical method for solving nonlinear ill-posed problems*, Nonlinear Funct. Anal. Optim., 20 (1999), pp. 317–332.
[32] A. Rieder, *On convergence rates of inexact Newton regularizations*, Numer. Math., 88 (2001), pp. 347–365.
[33] A.N. Tikhonov and V.A. Arsenin, *Methods for Solving Ill-Posed Problems*, Nauka, Moskow, 1979.
[34] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 183–198

A DISCONTINUOUS FINITE VOLUME METHOD FOR THE STOKES PROBLEMS*

XIU YE†

Abstract. We develop a new discontinuous finite volume method for solving the Stokes equations on both triangular and rectangular meshes. An optimal error estimate for the approximation of velocity is obtained in a mesh-dependent norm. First order L^2-error estimates are derived for the approximations of both velocity and pressure.

Key words. discontinuous Galerkin method, finite volume methods, Stokes problems

AMS subject classifications. Primary, 65N15, 65N30, 76D07; Secondary, 35B45, 35J50

DOI. 10.1137/040616759

1. Introduction.

Like finite element methods and finite difference methods, finite volume methods are discretization techniques for solving partial differential equations (PDEs). The integral formulation of finite volume schemes for a PDE is obtained by integrating the PDE over a control volume. In general, it represents the conservation of a quantity of interest, such as mass, momentum, or energy in fluid mechanics. Due to this natural association and its simplicity, finite volume methods are widely used in computational fluid mechanics and other applications [5, 6, 17]. Recently, Chou, Kwak, and Vassilevski [7, 8, 9, 10] applied finite volume methods involving nonconforming trial functions for diffusion, diffusion-reaction, and Stokes problems.

The discontinuous Galerkin method is a very active research field, and much literature can be found related to it [1, 2, 3, 4, 11, 12, 14, 16, 20, 19]. Discontinuous Galerkin methods use discontinuous functions as finite element approximation and enforce the connections of the approximation solutions between elements by adding some penalty terms. The flexibility of discontinuous functions gives discontinuous Galerkin methods many advantages, such as high order of accuracy, high parallelizability, localizability, and easy handling of complicated geometries.

Based on the advantages of using discontinuous functions as approximation in discontinuous Galerkin methods, it is natural to consider using discontinuous function as trial functions in the finite volume method, which we called the discontinuous finite volume method. Such method has the flexibility of the discontinuous Galerkin method and the simplicity and conservative properties of the finite volume method. In [21], a new discontinuous finite volume method was developed and analyzed for the second order elliptic problem. The local properties of the discontinuous finite volume method also reflect on the size of its control volume, which is less than half the control volume used in existing finite volume methods.

In this paper, we will extend the ideas developed in [21] to solve the Stokes equations on both triangular and rectangular meshes. In our methods, velocity is approximated by discontinuous piecewise linear functions on triangular mesh and by discontinuous piecewise rotated bilinear functions on rectangular mesh. Piecewise constant functions are used as the test functions for velocity in the discontinuous

*Received by the editors October 11, 2004; accepted for publication (in revised form) September 2, 2005; published electronically February 8, 2006.
http://www.siam.org/journals/sinum/44-1/61675.html
†Department of Mathematics, University of Arkansas, Little Rock, AR 72204 (xxye@ualr.edu).

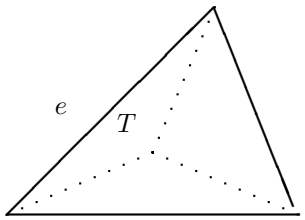


FIG. 1. *Element $T \in \mathcal{T}_h$ for triangular mesh.*

finite volume method. Therefore, after multiplying the differential equations by a test function and integrating by parts, the area integrals in the formulations will disappear, which gives the simplicity of finite volume method.

One of the advantages of using discontinuous approximation functions is it is easy to build high order elements. It is natural to consider using high order elements in the discontinuous finite volume formulations. However, use of higher order trial functions in finite volume methods may result in raising the order of the test functions. This implies that the test functions are no longer the piecewise constant that will cost the simplicity of finite volume methods. Nevertheless, it is worth a future research effort to investigate the high order discontinuous finite volume method.

Since discontinuous functions are used in the approximation, the number of unknowns is greater. However, the small support of the control volume for this method makes the method suitable for domain decomposition since the information can be updated triangle by triangle in the primary partition.

We consider the Stokes equations

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{2}$$

$$\mathbf{u} = 0 \quad \text{on } \partial\Omega, \tag{3}$$

where the symbols Δ, ∇, and $\nabla\cdot$ denote the Laplacian, gradient, and divergence operators, respectively, and $\mathbf{f}(x)$ is the external volumetric force acting on the fluid at $x \in \Omega \subset R^2$. We assume $\nu = 1$.

2. Preliminaries and notations. We will use the standard definitions for the Sobolev spaces $H^s(K)$ and their associated inner products $(\cdot,\cdot)_{s,K}$, norms $\|\cdot\|_{s,K}$, and seminorms $|\cdot|_{s,K}$, $s \geq 0$. The space $H^0(K)$ coincides with $L^2(K)$, in which case the norm and inner product are denoted by $\|\cdot\|_K$ and $(\cdot,\cdot)_K$, respectively. If $K = \Omega$, we drop K. Let $L_0^2(\Omega)$ to denote the subspace of $L^2(\Omega)$ consisting of functions with mean value zero.

Let $\mathcal{R}_h$ be a triangular or rectangular partition of Ω with $\text{diam}(\Omega) \leq h$. The triangles or rectangles in $\mathcal{R}_h$ are divided into three or four subtriangles by connecting the barycenter of the triangle or the center of the rectangles to their corner nodes, respectively, as shown in Figures 1 and 2. Then we define the dual partition $\mathcal{T}_h$ of the primal partition $\mathcal{R}_h$ to be the union of the triangles shown in Figures 1 and 2 for both rectangular and triangular mesh.

Let $P_k(T)$ consist of all the polynomials with degree less than or equal to k defined on T. We define the finite dimensional trial function space for velocity on triangular partition by

$$V_h = \{\mathbf{v} \in L^2(\Omega)^2 : \ \mathbf{v}|_K \in P_1(K)^2 \ \ \forall K \in \mathcal{R}_h\}$$

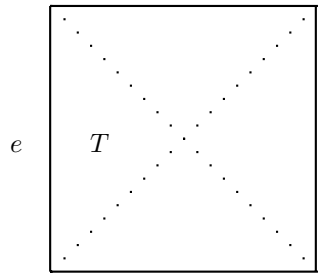


FIG. 2. *Element $T \in \mathcal{T}_h$ for rectangular mesh.*

and on rectangular partition by

$$V_h = \{\mathbf{v} \in L^2(\Omega)^2 : \ \mathbf{v}|_K \in \hat{Q}_1(K)^2 \ \ \forall K \in \mathcal{R}_h\},$$

where $\hat{Q}_1(K)$ denotes the space of functions of the form $a + bx_1 + cx_2 + d(x_1^2 - x_2^2)$ on K. Define the finite dimensional test function space W_h for velocity associated with the dual partition $\mathcal{T}_h$ as

$$W_h = \{\xi \in L^2(\Omega)^2 : \ \xi|_T \in P_0(T)^2 \ \ \forall T \in \mathcal{T}_h\}.$$

Let Q_h be the finite dimensional space for pressure

$$Q_h = \{q \in L_0^2(\Omega) : \ q|_K \in P_0(K) \ \ \forall K \in \mathcal{R}_h\}.$$

Multiplying (1) and (2) by $\xi \in W_h$ and $q \in Q_h$, respectively, we have

$$-\sum_{T\in\mathcal{T}_h} \int_{\partial T} \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \xi ds + \sum_{T\in\mathcal{T}_h} \int_{\partial T} p\xi \cdot \mathbf{n} ds = (\mathbf{f},\ \xi) \tag{4}$$

and

$$\sum_{K\in\mathcal{R}_h} \int_K \nabla \cdot \mathbf{u} q dx = 0, \tag{5}$$

where $\mathbf{n}$ is the unit outward normal vector on ∂T.

Let $T_j \in \mathcal{T}_h$ $(j = 1, \ldots, t)$ be the triangles in $K \in \mathcal{R}_h$, where $t = 3$ for triangular mesh and $t = 4$ for rectangular mesh, as shown as Figures 3 and 4. Then we have

$$\begin{aligned} \sum_{T\in\mathcal{T}_h} \int_{\partial T} \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \xi ds &= \sum_{K\in\mathcal{R}_h} \sum_{j=1}^{t} \int_{\partial T_j} \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \xi ds \\ &= \sum_{K\in\mathcal{R}_h} \sum_{j=1}^{t} \int_{A_{j+1}CA_j} \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \xi ds + \sum_{K\in\mathcal{R}_h} \int_{\partial K} \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \xi ds, \end{aligned} \tag{6}$$

where $A_{t+1} = A_1$.

For vectors $\mathbf{v}$ and $\mathbf{n}$, let $\mathbf{v} \otimes \mathbf{n}$ denote the matrix whose ijth component is $v_i n_j$ as in [13]. For two matrix valued variables σ and τ, we define $\sigma : \tau = \sum_{i,j=1}^{2} \sigma_{ij}\tau_{ij}$. Let e be an interior edge shared by two elements K_1 and K_2 in $\mathcal{T}_h$, and let $\mathbf{n}_1$ and $\mathbf{n}_2$

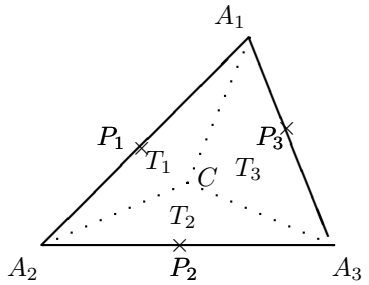


FIG. 3. *Triangular partition and its dual.*

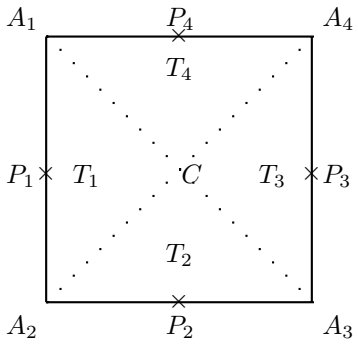


FIG. 4. *Rectangular partition and its dual.*

be unit normal vectors on e pointing exterior to K_1 and K_2, respectively. We define the average $\{\cdot\}$ and jump $[\cdot]$ on e for scalar q, vector $\mathbf{w}$ and τ, respectively.

$$\{q\} = \frac{1}{2}(q|_{\partial K_1} + q|_{\partial K_2}), \quad [q] = q|_{\partial K_1}\mathbf{n}_1 + q|_{\partial K_2}\mathbf{n}_2,$$

$$\{\mathbf{w}\} = \frac{1}{2}(\mathbf{w}|_{\partial K_1} + \mathbf{w}|_{\partial K_2}), \quad [\mathbf{w}] = \mathbf{w}|_{\partial T_K} \cdot \mathbf{n}_1 + \mathbf{w}|_{\partial K_2} \cdot \mathbf{n}_2,$$

and

$$\{\tau\} = \frac{1}{2}(\tau|_{\partial K_1} + \tau|_{\partial K_2}), \quad [\tau] = \tau|_{\partial K_1} \cdot \mathbf{n}_1 + \tau|_{\partial K_2} \cdot \mathbf{n}_2.$$

We also define a matrix valued jump $[\![\cdot]\!]$ for a vector $\mathbf{w}$ as $[\![\mathbf{w}]\!] = \mathbf{w}|_{\partial K_1} \otimes \mathbf{n}_1 + \mathbf{w}|_{\partial K_2} \otimes \mathbf{n}_2$ on e. If e is a edge on the boundary of Ω, define

$$\{q\} = q, \quad [\mathbf{w}] = \mathbf{w} \cdot \mathbf{n}, \quad \{\tau\} = \tau, \quad [\![\mathbf{w}]\!] = \mathbf{w} \otimes \mathbf{n}.$$

Let Γ denote the union of the boundaries of the rectangles K of $\mathcal{R}_h$ and $\Gamma_0 := \Gamma \backslash \partial\Omega$. A straightforward computation gives

$$\sum_{K\in\mathcal{R}_h} \int_{\partial K} q\mathbf{v} \cdot \mathbf{n} ds = \sum_{e\in\Gamma_0} \int_e [q] \cdot \{\mathbf{v}\} ds + \sum_{e\in\Gamma} \int_e \{q\}[\mathbf{v}] ds, \tag{7}$$

$$\sum_{K\in\mathcal{R}_h} \int_{\partial K} \mathbf{v} \cdot \tau\mathbf{n} ds = \sum_{e\in\Gamma_0} \int_e [\tau] \cdot \{\mathbf{v}\} ds + \sum_{e\in\Gamma} \int_e \{\tau\} : [\![\mathbf{v}]\!] ds. \tag{8}$$

Let $\int_\Gamma q ds = \sum_{e\in\Gamma}\int_e q ds$. Using (6), (8), and the fact that $[\nabla \mathbf{u}] = 0$ for $\mathbf{u} \in (H_0^1(\Omega)\cap H^2(\Omega))^2$ on Γ_0, (6) becomes

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}\frac{\partial \mathbf{u}}{\partial \mathbf{n}}\cdot\xi ds = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j}\frac{\partial \mathbf{u}}{\partial \mathbf{n}}\cdot\xi ds + \int_\Gamma [\![\xi]\!] : \{\nabla \mathbf{u}\} ds.$$

Since $[p] = 0$ for $p \in H^1(\Omega)$ on Γ_0, we also have

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T} p\xi\cdot\mathbf{n} ds = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j} p\xi\cdot\mathbf{n} ds + \int_\Gamma \{p\}[\xi] ds.$$

Let

$$a_0(\mathbf{v},\xi) = -\sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j}\frac{\partial \mathbf{v}}{\partial \mathbf{n}}\cdot\xi ds - \int_\Gamma [\![\xi]\!] : \{\nabla \mathbf{v}\} ds,$$

$$c(\xi, q) = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j} q\xi\cdot\mathbf{n} ds + \int_\Gamma \{q\}[\xi] ds,$$

and

$$b_0(\mathbf{v}, q) = \sum_{K\in\mathcal{R}_h}\int_K \nabla\cdot\mathbf{v} q dx.$$

It is clear that the solutions $(\mathbf{u}, p)$ of the Stokes equations (1)–(3) satisfy the following:

$$a_0(\mathbf{u},\xi) + c(\xi, p) = (\mathbf{f},\ \xi) \qquad \forall \xi \in W_h, \tag{9}$$

$$b_0(\mathbf{u}, q) = 0 \qquad \forall q \in Q_h. \tag{10}$$

Let $V(h) = V_h + (H^2(\Omega)\cap H_0^1(\Omega))^2$. Define a mapping $\gamma:\ V(h) \to W_h$ as shown in Figure 1 and 2.

$$\gamma\mathbf{v}|_T = \frac{1}{h_e}\int_e \mathbf{v}|_T ds, \quad T \in \mathcal{T}_h,$$

where h_e is the length of the edge e. For $\mathbf{v} = (v_1, v_2)$, γv_i (i=1, 2) is defined as

$$\gamma v_i|_T = \frac{1}{h_e}\int_e v_i|_T ds, \quad T \in \mathcal{T}_h.$$

Define the following bilinear forms:

$$\begin{aligned} A_0(\mathbf{v},\ \mathbf{w}) &= a_0(\mathbf{v},\ \gamma\mathbf{w}) \qquad \forall \mathbf{v}, \mathbf{w} \in V(h),\\ B_0(\mathbf{v}, q) &= b_0(\mathbf{v}, q) \qquad \forall \mathbf{v} \in V(h),\ \forall q \in L_0^2(\Omega),\\ C(\mathbf{v},\ q) &= c(\gamma\mathbf{v},\ q) \qquad \forall \mathbf{v} \in V(h),\ \forall q \in L_0^2(\Omega). \end{aligned}$$

Then systems (9)–(10) are equivalent to

$$A_0(\mathbf{u}, \mathbf{v}) + C(\mathbf{v}, p) = (\mathbf{f},\ \gamma\mathbf{v}) \qquad \forall \mathbf{v} \in V_h, \tag{11}$$

$$B_0(\mathbf{u}, q) = 0 \qquad \forall q \in Q_h. \tag{12}$$

3. Discontinuous finite volume formulation. In this section we propose two discontinuous finite volume formulations based on modification of the weak formulation (11)–(12) for the Stokes problem (1)–(3).

Let us introduce the bilinear forms as follows:

$$A_1(\mathbf{v},\mathbf{w}) = A_0(\mathbf{v},\mathbf{w}) + \alpha \sum_{e\in\Gamma} [\![\gamma\mathbf{v}]\!]_e : [\![\gamma\mathbf{w}]\!]_e$$

and

$$B(\mathbf{v},q) = B_0(\mathbf{v},q) - \int_\Gamma \{q\}[\gamma\mathbf{v}]ds,$$

where $\alpha > 0$ is a parameter to be determined later. For the exact solution $(\mathbf{u}, p)$ of the Stokes problem, we have

$$\begin{aligned} A_0(\mathbf{u},\mathbf{v}) &= A_1(\mathbf{u},\mathbf{v}) \qquad \forall \mathbf{v}\in V_h, \\ B_0(\mathbf{u},q) &= B(\mathbf{u},q) \qquad \forall q\in Q_h. \end{aligned}$$

Therefore, it follows from (11)–(12) that

$$A_1(\mathbf{u},\mathbf{v}) + C(\mathbf{v},p) = (\mathbf{f},\ \gamma\mathbf{v}) \qquad \forall \mathbf{v}\in V_h, \tag{13}$$

$$B(\mathbf{u},q) = 0 \qquad \forall q\in Q_h. \tag{14}$$

The corresponding discontinuous finite volume scheme for (1)–(3) seeks $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ such that

$$A_1(\mathbf{u}_h,\mathbf{v}) + C(\mathbf{v},p_h) = (\mathbf{f},\ \gamma\mathbf{v}) \qquad \forall \mathbf{v}\in V_h, \tag{15}$$

$$B(\mathbf{u}_h,q) = 0 \qquad \forall q\in Q_h. \tag{16}$$

Since the test functions are piecewise constant for velocity, the bilinear forms $A_1(\cdot,\cdot)$ and $C(\cdot,\cdot)$ in (15) do not include the area integral terms like $\sum_K \int_K \nabla\mathbf{u}_h : \nabla\mathbf{v}dx$ that normally present in the finite element formulations for the Stokes problem.

Let

$$A_*(\mathbf{v},\mathbf{w}) = -\sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j} \frac{\partial\mathbf{v}}{\partial\mathbf{n}}\cdot\gamma\mathbf{w}ds$$

and

$$C_*(\mathbf{v},q) = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j} q\gamma\mathbf{v}\cdot\mathbf{n}ds.$$

Thus

$$A_1(\mathbf{v},\mathbf{w}) = A_*(\mathbf{v},\mathbf{w}) - \int_\Gamma [\![\gamma\mathbf{w}]\!] : \{\nabla\mathbf{v}\}ds + \alpha\sum_{e\in\Gamma}[\![\gamma\mathbf{v}]\!]_e : [\![\gamma\mathbf{w}]\!]_e$$

and

$$C(\mathbf{v},q) = C_*(\mathbf{v},q) + \int_\Gamma \{q\}[\gamma\mathbf{v}]ds.$$

Let $\nabla_h \mathbf{v}$ and $\nabla_h \cdot \mathbf{v}$ be the functions whose restriction to each element $K \in \mathcal{R}_h$ are equal to $\nabla \mathbf{v}$ and $\nabla \cdot \mathbf{v}$, respectively.

LEMMA 3.1. *For any* $\mathbf{v}, \mathbf{w} \in V(h)$, *we have*

$$A_*(\mathbf{v}, \mathbf{w}) = (\nabla_h \mathbf{v}, \nabla_h \mathbf{w}) + \sum_{K \in \mathcal{R}_h} \int_{\partial K} (\gamma \mathbf{w} - \mathbf{w}) \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds + \sum_{K \in \mathcal{R}_h} (\Delta \mathbf{v}, \mathbf{w} - \gamma \mathbf{w})_K.$$

Furthermore, if $\mathbf{v}, \mathbf{w} \in V_h$, *then*

$$A_*(\mathbf{v}, \mathbf{w}) = (\nabla_h \mathbf{v}, \nabla_h \mathbf{w}).$$

Proof. Using the divergence theorem on each triangle T_j for $\mathbf{v} \in V(h)$ and the fact that $\gamma \mathbf{w}$ is a constant on each T_j, we have

$$\begin{aligned}
A_*(\mathbf{v}, \mathbf{w}) &= -\sum_{K \in \mathcal{R}_h} \sum_{j=1}^{t} \int_{A_{j+1} C A_j} \frac{\partial \mathbf{v}}{\partial \mathbf{n}} \cdot \gamma \mathbf{w} ds \\
&= \sum_{K \in \mathcal{R}_h} \sum_{j=1}^{t} \gamma \mathbf{w} \cdot \int_{A_j A_{j+1}} \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds - \sum_{K \in \mathcal{R}_h} \sum_{T_j \in K} (\Delta \mathbf{v}, \gamma \mathbf{w})_{T_j} \\
&= \sum_{K \in \mathcal{R}_h} \int_{\partial K} (\gamma \mathbf{w} - \mathbf{w}) \cdot \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds + \sum_{K \in \mathcal{R}_h} \int_{\partial K} \mathbf{w} \cdot \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds - \sum_{K \in \mathcal{R}_h} (\Delta \mathbf{v}, \gamma \mathbf{w})_K \\
&= \sum_{K \in \mathcal{R}_h} \int_{\partial K} (\gamma \mathbf{w} - \mathbf{w}) \cdot \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds + \sum_{K \in \mathcal{R}_h} (\nabla \mathbf{v}, \nabla \mathbf{w})_K + \sum_{K \in \mathcal{R}_h} (\Delta \mathbf{v}, \mathbf{w})_K - \sum_{K \in \mathcal{R}_h} (\Delta \mathbf{v}, \gamma \mathbf{w})_K \\
&= (\nabla_h \mathbf{v}, \nabla_h \mathbf{w}) + \sum_{K \in \mathcal{R}_h} \int_{\partial K} (\gamma \mathbf{w} - \mathbf{w}) \cdot \frac{\partial \mathbf{v}}{\partial \mathbf{n}} ds + \sum_{K \in \mathcal{R}_h} (\Delta \mathbf{v}, \mathbf{w} - \gamma \mathbf{w})_K.
\end{aligned}$$

If $\mathbf{v} \in V_h$, we have $\Delta \mathbf{v} = 0$ on each $K \in \mathcal{R}_h$. Then the third term in the above equation drops out. Since $\frac{\partial \mathbf{v}}{\partial \mathbf{n}}$ is constant for both triangular and rectangular mesh, by the definition of γ, the second term is zero. This completes the proof. ∎

LEMMA 3.2. *For any* $(\mathbf{v}, q) \in V(h) \times L_0^2(\Omega)$, *we have*

$$C_*(\mathbf{v}, q) = -(\nabla_h \cdot \mathbf{v}, q) + \sum_{K \in \mathcal{R}_h} \int_{\partial K} (\mathbf{v} - \gamma \mathbf{v}) \cdot \mathbf{n} q ds + \sum_{K \in \mathcal{R}_h} (\nabla q, \gamma \mathbf{v} - \mathbf{v})_K. \tag{17}$$

Furthermore, if $q \in Q_h$, *then*

$$C_*(\mathbf{v}, q) = -(\nabla_h \cdot \mathbf{v}, q) \tag{18}$$

and

$$C(\mathbf{v},q) = -B(\mathbf{v},q). \tag{19}$$

Proof. Using the divergence theorem on each triangle T_j for $\mathbf{v} \in V_h$, we have

$$\begin{aligned}
C_*(\mathbf{v},q) &= \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_{j+1}CA_j}\gamma\mathbf{v}\cdot\mathbf{n}q ds \\
&= -\sum_{K\in\mathcal{R}_h}\sum_{j=1}^{t}\int_{A_jA_{j+1}}\gamma\mathbf{v}\cdot\mathbf{n}q ds + \sum_{K\in\mathcal{R}_h}\sum_{T_j\in K}(\nabla q,\ \gamma\mathbf{v})_{T_j} \\
&= \sum_{K\in\mathcal{R}_h}\int_{\partial K}(\mathbf{v}-\gamma\mathbf{v})\cdot\mathbf{n}q ds - \sum_{K\in\mathcal{R}_h}\int_{\partial K}\mathbf{v}\cdot\mathbf{n}q ds + \sum_{K\in\mathcal{R}_h}(\nabla q,\gamma\mathbf{v})_K \\
&= \sum_{K\in\mathcal{R}_h}\int_{\partial K}(\mathbf{v}-\gamma\mathbf{v})\cdot\mathbf{n}q ds + \sum_{K\in\mathcal{R}_h}(\nabla q,\gamma\mathbf{v})_K - \sum_{K\in\mathcal{R}_h}(\nabla q,\mathbf{v})_K - \sum_{K\in\mathcal{R}_h}(\nabla\cdot\mathbf{v},q)_K \\
&= -(\nabla_h\cdot\mathbf{v},q) + \sum_{K\in\mathcal{R}_h}\int_{\partial K}(\mathbf{v}-\gamma\mathbf{v})\cdot\mathbf{n}q ds + \sum_{K\in\mathcal{R}_h}(\nabla q,\gamma\mathbf{v}-\mathbf{v})_K.
\end{aligned}$$

If $q \in Q_h$, the second and third terms in the above equation drop out. This completes the proof. ∎

We define two norms for $V(h)$ as follows:

$$\begin{aligned}
|\!|\!|\mathbf{v}|\!|\!|_1^2 &= |\mathbf{v}|_{1,h}^2 + \sum_{e\in\Gamma}[\![\gamma\mathbf{v}]\!]_e^2, \\
|\!|\!|\mathbf{v}|\!|\!|^2 &= |\!|\!|\mathbf{v}|\!|\!|_1^2 + \sum_{K\in\mathcal{R}_h} h_K^2|\mathbf{v}|_{2,K}^2,
\end{aligned}$$

where $|\mathbf{v}|_{1,h}^2 = \sum_K |\mathbf{v}|_{1,K}^2$.

The standard inverse inequality implies that there is a constant C such that

$$|\!|\!|\mathbf{v}|\!|\!| \le C|\!|\!|\mathbf{v}|\!|\!|_1 \qquad \forall \mathbf{v}\in V_h. \tag{20}$$

Let K be an element with e as an edge. It is well known (see [1]) that there exists a constant C such that for any function $g \in H^2(K)$,

$$\|g\|_e^2 \le C\left(h_K^{-1}\|g\|_K^2 + h_K|g|_{1,K}^2\right), \tag{21}$$

$$\left\|\frac{\partial g}{\partial \mathbf{n}}\right\|_e^2 \le C(h_K^{-1}|g|_{1,K}^2 + h_K|g|_{2,K}^2), \tag{22}$$

where C depends only on the minimum angle of K.

LEMMA 3.3. *For $\mathbf{v},\mathbf{w}\in V(h)$, we have*

$$A_1(\mathbf{v},\mathbf{w}) \le C|\!|\!|\mathbf{v}|\!|\!||\!|\!|\mathbf{w}|\!|\!|. \tag{23}$$

Proof. By Lemma 3.1, the inequalities (21) and (22), and the Cauchy–Schwarz inequality, we have

$$|A_*(\mathbf{v},\mathbf{w})| \le |(\nabla_h\mathbf{v},\nabla_h\mathbf{w})| + \left|\sum_{K\in\mathcal{R}_h}\int_{\partial K}(\mathbf{w}-\gamma\mathbf{w})\cdot\frac{\partial\mathbf{v}}{\partial\mathbf{n}}ds\right| + \left|\sum_{K\in\mathcal{R}_h}(\Delta\mathbf{v},\mathbf{w}-\gamma\mathbf{w})_K\right|$$
$$\le C(|\mathbf{v}|_{1,h}|\mathbf{w}|_{1,h} + \sum_{K\in\mathcal{R}_h}(h_K^{-1}\|\mathbf{w}-\gamma\mathbf{w}\|_K^2 + h_K|\mathbf{w}-\gamma\mathbf{w}|_{1,K}^2)^{\frac{1}{2}}(h_K^{-1}|\mathbf{v}|_{1,K}^2 + h_K|\mathbf{v}|_{2,K}^2)^{\frac{1}{2}}$$
$$+ \sum_{K\in\mathcal{R}_h} h|\mathbf{v}|_{2,K}|\mathbf{w}|_{1,K})$$
$$\le C\left(|\mathbf{v}|_{1,h}|\mathbf{w}|_{1,h} + \left(\sum_{K\in\mathcal{R}_h}|\mathbf{w}|_{1,K}^2\right)^{\frac{1}{2}}\left(|\mathbf{v}|_{1,h} + \left(\sum_{K\in\mathcal{R}_h}h_K^2|\mathbf{v}|_{2,K}^2\right)^{\frac{1}{2}}\right)\right.$$
$$\left. + \left(\sum_{K\in\mathcal{R}_h}h_K^2|\mathbf{v}|_{2,K}\right)^{\frac{1}{2}}|\mathbf{w}|_{1,h}\right)$$
$$\le C|||\mathbf{v}|||\,|||\mathbf{w}|||.$$

The definition of $A_1(\mathbf{v},\mathbf{w})$ and the inequality above imply that

$$A_1(\mathbf{v},\mathbf{w}) = A_*(\mathbf{v},\mathbf{w}) - \int_\Gamma [\![\gamma\mathbf{w}]\!] : \{\nabla\mathbf{v}\}ds + \alpha\sum_{e\in\Gamma}[\![\gamma\mathbf{v}]\!]_e : [\![\gamma\mathbf{w}]\!]_e$$
$$\le C\left(|||\mathbf{v}|||\,|||\mathbf{w}||| + \left(\sum_{K\in\mathcal{R}_h}(|\mathbf{v}|_{1,K}^2 + h^2|\mathbf{v}|_{2,K}^2)\right)^{\frac{1}{2}}\left(\sum_{e\in\Gamma}[\![\gamma\mathbf{w}]\!]_e^2\right)^{\frac{1}{2}}\right.$$
$$\left. + \alpha\left(\sum_{e\in\Gamma}[\![\gamma\mathbf{v}]\!]_e^2\right)^{\frac{1}{2}}\left(\sum_{e\in\Gamma}[\![\gamma\mathbf{w}]\!]_e^2\right)^{\frac{1}{2}}\right)$$
$$\le C|||\mathbf{v}|||\,|||\mathbf{w}|||. \qquad \square$$

Similarly we can prove the following lemma.

LEMMA 3.4. *For $(\mathbf{v},q)\in V(h)\times L_0^2(\Omega)$, we have*

$$C(\mathbf{v},q) \le C|||\mathbf{v}|||\left(\|q\| + \left(\sum_{K\in\mathcal{R}_h}h_K^2|q|_{1,K}^2\right)^{\frac{1}{2}}\right). \tag{24}$$

If $(\mathbf{v},q)\in V_h\times Q_h$, then

$$C(\mathbf{v},q) \le C|||\mathbf{v}|||\,\|q\|. \tag{25}$$

We will prove the coercivity of the bilinear form $A_1(\cdot,\cdot)$ in V_h in the following lemma.

LEMMA 3.5. *For any $\mathbf{v}\in V_h$, there is a constant C independent of h such that for α large enough*

$$A_1(\mathbf{v},\mathbf{v}) \ge C|||\mathbf{v}|||^2. \tag{26}$$

Proof. The trace inequality (22) and the inverse inequality give that for $\mathbf{v} \in V_h$

$$\begin{aligned}\int_\Gamma [\![\gamma \mathbf{v}]\!] : \{\nabla \mathbf{v}\} ds &\le C \left(\sum_{e\in\Gamma} \int_e h_e \{\nabla \mathbf{v}\}^2 ds \right)^{\frac{1}{2}} \left(\sum_{e\in\Gamma} \int_e h_e^{-1} [\![\mathbf{v}]\!]^2 ds \right)^{\frac{1}{2}} \\ &\le C \left(\sum_{K\in\mathcal{R}_h} (|\mathbf{v}|_{1,K}^2 + h_K^2 |\mathbf{v}|_{2,K}^2) \right)^{\frac{1}{2}} \left(\sum_{e\in\Gamma} [\![\gamma \mathbf{v}]\!]_e^2 \right)^{\frac{1}{2}} \\ &\le C |\!|\!| \mathbf{v} |\!|\!|_1 \left(\sum_{e\in\Gamma} [\![\gamma \mathbf{v}]\!]_e^2 \right)^{\frac{1}{2}}.\end{aligned}$$

Using the inequality above, (20), and Lemma 3.1, we have

$$\begin{aligned}A_1(\mathbf{v},\mathbf{v}) &= (\nabla_h \mathbf{v}, \nabla_h \mathbf{v}) + \alpha \sum_{e\in\Gamma} [\![\gamma \mathbf{v}]\!]_e^2 - \int_\Gamma [\![\gamma \mathbf{v}]\!] : \{\nabla \mathbf{v}\} ds \\ &\ge |\mathbf{v}|_{1,h}^2 + \alpha \sum_{e\in\Gamma} [\![\gamma \mathbf{v}]\!]_e^2 - C |\!|\!| \mathbf{v} |\!|\!|_1 \left(\sum_{e\in\Gamma} [\![\gamma \mathbf{v}]\!]_e^2 \right)^{\frac{1}{2}} \\ &\ge C |\!|\!| \mathbf{v} |\!|\!|_1^2 \ge C |\!|\!| \mathbf{v} |\!|\!|^2.\end{aligned}$$

The last inequality is obtained by using the arithmetic-geometric mean inequality and choosing α large enough. ∎

The proof of Lemma 3.5 indicates that the value of α depends upon the constant in the inverse inequality. Therefore, the value of α for which $A_1(\cdot,\cdot)$ is coercive is mesh dependent. Existing results for saddle-point problems indicate that it is theoretically and computationally important to have coercivity (26). Therefore, the mesh-dependence of the parameter α makes the discontinuous finite volume scheme (15)–(16) less interesting in practical computation.

To overcome the difficulty in the selection of parameter, we introduce a second discontinuous finite volume scheme which is parameter insensitive. To this end, we define a bilinear form as follows:

$$A_2(\mathbf{v},\mathbf{w}) = A_1(\mathbf{v},\mathbf{w}) + \int_\Gamma [\![\gamma \mathbf{v}]\!] : \{\nabla \mathbf{w}\} ds.$$

Similar to the bilinear form $A_1(\cdot,\cdot)$, for the exact solution $(\mathbf{u}, p)$ of the Stokes problem we have

$$A_2(\mathbf{u},\mathbf{v}) = A_0(\mathbf{u},\mathbf{v}) \qquad \forall \mathbf{v} \in V_h.$$

Consequently, the solution of the Stokes problem satisfies the following variational equations:

$$A_2(\mathbf{u},\mathbf{v}) + C(\mathbf{v},p) = (\mathbf{f}, \gamma \mathbf{v}) \qquad \forall \mathbf{v} \in V_h, \tag{27}$$

$$B(\mathbf{u},q) = 0 \qquad \forall q \in W_h. \tag{28}$$

Our second discontinuous finite volume scheme for (1)–(3) seeks $(\mathbf{u}_h, p_h) \in V_h \times W_h$ such that

$$A_2(\mathbf{u}_h,\mathbf{v}) + C(\mathbf{v},p_h) = (\mathbf{f}, \gamma \mathbf{v}) \qquad \forall \mathbf{v} \in V_h, \tag{29}$$

$$B(\mathbf{u}_h,q) = 0 \qquad \forall q \in W_h. \tag{30}$$

To see the coercivity of the bilinear form $A_2(\cdot,\cdot)$, we use its definition to obtain

$$A_2(\mathbf{v},\mathbf{v}) = (\nabla_h\mathbf{v},\nabla_h\mathbf{v}) + \alpha\sum_{e\in\Gamma}[\![\mathbf{v}]\!]_e^2 ds \tag{31}$$
$$= |\!|\!|\mathbf{v}|\!|\!|_1^2 \geq C|\!|\!|\mathbf{v}|\!|\!|^2 \qquad \forall \mathbf{v}\in V_h.$$

Thus, the coercivity (31) holds true for the bilinear form $A_2(\cdot,\cdot)$ with any value of $\alpha > 0$. Similarly, we can prove that

$$A_2(\mathbf{w},\mathbf{v}) \leq C|\!|\!|\mathbf{w}|\!|\!||\!|\!|\mathbf{v}|\!|\!| \qquad \forall \mathbf{w},\mathbf{v}\in V(h). \tag{32}$$

Let $A(\mathbf{v},\mathbf{w}) = A_1(\mathbf{v},\mathbf{w})$ or $A(\mathbf{v},\mathbf{w}) = A_2(\mathbf{v},\mathbf{w})$. In the rest of the paper, we assume that the following is true.

$$A(\mathbf{v},\mathbf{v}) \geq C|\!|\!|\mathbf{v}|\!|\!|^2. \tag{33}$$

If $A(\mathbf{v},\mathbf{w}) = A_2(\mathbf{v},\mathbf{w})$, (33) holds for any $\alpha > 0$. If $A(\mathbf{v},\mathbf{w}) = A_1(\mathbf{v},\mathbf{w})$, (33) holds for only α larger enough.

4. Error estimates. We will derive an optimal error estimates for velocity in the norm $|\!|\!|\cdot|\!|\!|$ and for pressure in the L^2-norm. A first order error estimate for velocity in L^2-norm will be obtained.

Let e be an interior edge shared by two elements K_1 and K_2 in $\mathcal{R}_h$. If $\int_e \mathbf{v}|_{K_1} ds = \int_e \mathbf{v}|_{K_2} ds$, we say that $\mathbf{v}$ is continuous on e. We say that $\mathbf{v}$ is zero at $e \in \partial\Omega$ if $\int_e \mathbf{v} ds = 0$. Define a subspace $\hat{V}_h$ of V_h by

$$\hat{V}_h = \{\mathbf{v}\in L^2(\Omega)^2 :\ \mathbf{v}|_K \in \hat{Q}_1(K)^2\ \forall K\in\mathcal{R}_h \text{ is continuous at } e\in\Gamma_0$$
$$\text{and is zero at } e\in\partial\Omega\}$$

for rectangular mesh and by

$$\hat{V}_h = \{\mathbf{v}\in L^2(\Omega)^2 :\ \mathbf{v}|_K \in P_1(K)^2\ \forall K\in\mathcal{R}_h \text{ is continuous at } e\in\Gamma_0$$
$$\text{and is zero at } e\in\partial\Omega\}$$

for triangular mesh.

It has been proved in [18] and [15] that the following discrete inf-sup condition is satisfied; i.e., there exists a positive constant β_0 such that

$$\sup_{\mathbf{v}\in\hat{V}_h} \frac{(\nabla_h\cdot\mathbf{v},\ q)}{|\mathbf{v}|_{1,h}} \geq \beta_0\|q\| \quad \forall\ q\in Q_h. \tag{34}$$

LEMMA 4.1. *The bilinear form $B(\cdot,\cdot)$ satisfies the discrete inf-sup condition*

$$\sup_{\mathbf{v}\in V_h} \frac{B(\mathbf{v},q)}{|\!|\!|\mathbf{v}|\!|\!|} \geq \beta\|q\| \quad \forall\ q\in Q_h, \tag{35}$$

where β is a positive constant independent of the mesh size h.

Proof. For $\mathbf{v}\in\hat{V}_h\subset V_h$ and $q\in Q_h$, we have $B(\mathbf{v},q) = (\nabla_h\cdot\mathbf{v},q)$ and $|\!|\!|\mathbf{v}|\!|\!|_1 = |\mathbf{v}|_{1,h}$. (34) and (20) imply that for any $q\in Q_h$

$$\beta_0\|q\| \leq \sup_{\mathbf{v}\in\hat{V}_h} \frac{(\nabla_h\cdot\mathbf{v},\ q)}{|\mathbf{v}|_{1,h}} = \sup_{\mathbf{v}\in\hat{V}_h} \frac{B(\mathbf{v},\ q)}{|\!|\!|\mathbf{v}|\!|\!|_1} \leq C \sup_{\mathbf{v}\in V_h} \frac{B(\mathbf{v},\ q)}{|\!|\!|\mathbf{v}|\!|\!|}. \tag{36}$$

With $\beta = \beta_0/C$, we have proved (35). ∎

Define an operator π_K from $H^1(K)$ to $\hat{Q}_1(K)$ or $P_1(K)$ by requiring that for any $v \in H^1(K)$,

$$\int_{e_i} \pi_K v \, ds = \int_{e_i} v \, ds \quad \text{for} \quad i = 1, \ldots, t, \tag{37}$$

where e_i, $i = 1, \ldots, t$, are the t sides of the element $K \in \mathcal{R}_h$. $t = 3$ if K is a triangle and $t = 4$ if K is a rectangle. It was proved in [18] that

$$|\pi_K v - v|_{s,K} \le Ch^{2-s}|v|_{2,K} \quad \forall\, v \in H^2(K), \quad s = 0, 1, 2. \tag{38}$$

For any $\mathbf{v} \in H_0^1(\Omega)^2$, define $\Pi_1 \mathbf{v} \in V_h$ by

$$(\Pi_1 \mathbf{v})_i|_K = \pi_K v_i \quad \forall\, K \in \mathcal{R}_h, \quad i = 1, 2. \tag{39}$$

Using the definition of Π_1 and integration by parts, we can show that

$$B(\mathbf{v} - \Pi_1 \mathbf{v}, q) = 0 \qquad \forall q \in Q_h. \tag{40}$$

The Cauchy–Schwarz inequality implies

$$\begin{aligned} [\![\gamma \mathbf{v}]\!]_e^2 &= \left(\frac{1}{h_e}\int_e [\![\mathbf{v}]\!] ds\right)^2 \le \left(\frac{1}{h_e}\right)^2 \int_e [\![\mathbf{v}]\!]^2 ds \int_e ds \\ &= \int_e \frac{1}{h_e} [\![\mathbf{v}]\!]^2 ds. \end{aligned} \tag{41}$$

(41) and (21) imply that

$$\begin{aligned} \sum_{e\in\Gamma} [\![\gamma(\mathbf{u} - \Pi_1 \mathbf{u})]\!]_e^2 &\le \int_\Gamma \frac{1}{h_e} [\![\mathbf{u} - \Pi_1 \mathbf{u}]\!]^2 ds \\ &\le C\left(|\mathbf{u} - \Pi_1 \mathbf{u}|_{1,h}^2 + \sum_{K\in\mathcal{R}_h} h^{-2}\|\mathbf{u} - \Pi_1 \mathbf{u}\|_K^2\right) \\ &\le Ch^2 \|\mathbf{u}\|_2^2. \end{aligned} \tag{42}$$

The definitions of the norm $|||\cdot|||$, (42), and (38) give

$$\begin{aligned} |||\mathbf{u} - \Pi_1 \mathbf{u}|||^2 &= |\mathbf{u} - \Pi_1 \mathbf{u}|_{1,h}^2 + \sum_{e\in\Gamma} [\![\gamma(\mathbf{u} - \Pi_1 \mathbf{u})]\!]_e^2 + \sum_{K\in\mathcal{R}_h} h^2 |\mathbf{u} - \Pi_1 \mathbf{u}|_{2,K}^2 \\ &\le Ch^2 \|\mathbf{u}\|_2^2. \end{aligned} \tag{43}$$

Let Π_2 be the L^2 projection from $L_0^2(\Omega)$ to the finite element space Q_h.

THEOREM 4.2. *Let $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ be the solution of* (15)–(16) *or* (29)–(30) *and $(\mathbf{u}, p) \in (H^2(\Omega) \cap H_0^1(\Omega))^2 \times L_0^2(\Omega) \cap H^1(\Omega)$ be the solution of* (1)–(3). *Then there exists a constant C independent of h such that*

$$|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\| \le Ch(\|\mathbf{u}\|_2 + \|p\|_1). \tag{44}$$

Proof. Let

$$\epsilon_h = \mathbf{u}_h - \Pi_1 \mathbf{u}, \qquad \eta_h = p_h - \Pi_2 p \tag{45}$$

be the error between the finite volume solution $(\mathbf{u}_h, p_h)$ and the projection $(\Pi_1\mathbf{u}, \Pi_2 p)$ of the exact solution. Denote by

$$\epsilon = \mathbf{u} - \Pi_1\mathbf{u}, \qquad \eta = p - \Pi_2 p \tag{46}$$

the error between the exact solution $(\mathbf{u}, p)$ and its projection. Subtracting (15) and (16) from (13) and (14), respectively, or subtracting (29) and (30) from (27) and (28), respectively, and using Lemma 3.2 give that with $A(\cdot,\cdot) = A_1(\cdot,\cdot)$ or $A(\cdot,\cdot) = A_2(\cdot,\cdot)$

$$A(\epsilon_h, \mathbf{v}) - B(\mathbf{v}, \eta_h) = A(\epsilon, \mathbf{v}) - B(\mathbf{v}, \eta), \tag{47}$$

$$B(\epsilon_h, q) = B(\epsilon, q) = 0 \tag{48}$$

for any $\mathbf{v} \in V_h$ and $q \in Q_h$.

By letting $\mathbf{v} = \epsilon_h$ in (47) and $q = \eta_h$ in (48), the sum of (47) and (48) gives

$$A(\epsilon_h, \epsilon_h) = A(\epsilon, \epsilon_h) - B(\epsilon_h, \eta). \tag{49}$$

Thus, it follows from the coercivity (33), the boundedness (23), (32), and (24) that

$$|||\epsilon_h|||^2 \le C \left(|||\epsilon||| |||\epsilon_h||| + \left(\|\eta\| + \left(\sum_{K\in\mathcal{R}_h} h_K^2 |\eta|_{1,K}^2 \right)^{\frac{1}{2}} \right) |||\epsilon_h||| \right),$$

which implies the following:

$$|||\epsilon_h||| \le C \left(|||\epsilon||| + \|\eta\| + \left(\sum_{K\in\mathcal{R}_h} h_K^2 |\eta|_{1,K}^2 \right)^{\frac{1}{2}} \right).$$

The above estimate can be rewritten as

$$|||\mathbf{u}_h - \Pi_1\mathbf{u}||| \le C \left(|||\mathbf{u} - \Pi_1\mathbf{u}||| + \|p - \Pi_2 p\| + \left(\sum_{K\in\mathcal{R}_h} h^2 |p - \Pi_2 p|_{1,K}^2 \right)^{\frac{1}{2}} \right).$$

Now using the triangle inequality, (38), the definition of Π_2, and the inequality above, we get

$$|||\mathbf{u} - \mathbf{u}_h||| \le C \left(|||\mathbf{u} - \Pi_1\mathbf{u}||| + \|\mathbf{u}_h - \Pi_1\mathbf{u}\| \right) \le Ch(\|\mathbf{u}\|_2 + \|p\|_1), \tag{50}$$

which completes the estimate for the velocity approximation.

Discrete inf-sup condition (35), (50), Lemma 3.2, Lemma 3.4, and inverse inequality give

$$\begin{aligned}
\|p_h - \Pi_2 p\| &\le \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{B(\mathbf{v}, p_h - \Pi_2 p)}{|||\mathbf{v}|||} \le \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{C(\mathbf{v}, \Pi_2 p - p_h)}{|||\mathbf{v}|||} \\
&= \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{C(\mathbf{v}, p - p_h) + C(\mathbf{v}, \Pi_2 p - p)}{|||\mathbf{v}|||} \\
&= \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{A(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) + C(\mathbf{v}, \Pi_2 p - p)}{|||\mathbf{v}|||} \\
&\le C \left(|||\mathbf{u} - \mathbf{u}_h||| + \|p - \Pi_2 p\| + \left(\sum_{K\in\mathcal{R}_h} h_K^2 |p - \Pi_2 p|_{1,K}^2 \right)^{\frac{1}{2}} \right) \\
&\le Ch(\|\mathbf{u}\|_2 + \|p\|_1).
\end{aligned}$$

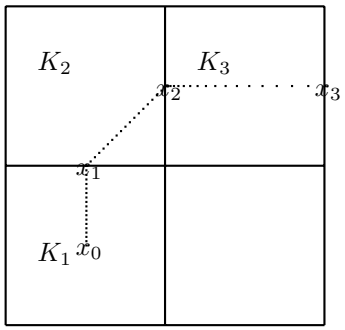


FIG. 5. *A path.*

Using the above inequality and the triangle inequality, we have completed the proof of (44). ∎

We need the following lemma for the L^2 error estimate of velocity.

LEMMA 4.3. *There exists a constant C independent of h such that the following is true:*

$$\|\mathbf{w}\| \le C|||\mathbf{w}||| \qquad \forall \mathbf{w} \in V_h.$$

Proof. The proof is similar to the proof of Lemma 3.1 in [8]. We prove this lemma only for the rectangular element. The same argument can be used to prove the lemma for the triangular element. Let $|e|$ denote the length of edge e.

Let e be an edge shared by two elements K_1 and K_2 in $\mathcal{R}_h$. Let $\mathbf{w} = (w_1, w_2)$. Define $[w_1]_* = w_1|_{\partial K_1} - w_1|_{\partial K_2}$. Interchanging K_1 and K_2 will have no effect on the procedure. Since $[w_1]_*$ is continuous on $e_i \in \Gamma$, there exists $\mathbf{x}_i \in e_i$ such that

$$[\gamma w_1]_* = \frac{1}{|e_i|}\int_{e_i} [w_1]_* ds = [w_1]_*(\mathbf{x}_i) \tag{51}$$

and

$$[\gamma w_1]_*^2 = [\gamma w_1]^2.$$

If e_i is an edge on the boundary, then $\mathbf{x}_i \in e_i$ is a point such that $\int_{e_i} w_1 ds = w_1(\mathbf{x}_i)|e_i|$. For any $\mathbf{x} = \mathbf{x}_0 \in K \in \mathcal{R}_h$, we can find a path from $\mathbf{x}_0$ to $\mathbf{x}_l$, a point on the boundary, by joining a sequence of $\mathbf{x}_i$ as shown in Figure 5, where $\mathbf{x}_i$ $(i = 1, \ldots, l)$ satisfy (51). Let C_0 be a constant such that $lh \le C_0$. Let $\{K_i\}_{i=1}^l$ be the sequence of rectangles in $\mathcal{R}_h$ containing $\mathbf{x}_i$ $(i = 0, \ldots, l)$ as shown in Figure 5.

Define $w_1(\mathbf{x}_0^1) = w_1(\mathbf{x}_0)$, $w_1(\mathbf{x}_1^2) = w_1|_{K_1}(\mathbf{x}_1)$, and $w_1(\mathbf{x}_1^1) = w_1|_{K_2}(\mathbf{x}_1)$. In general, $w_1(\mathbf{x}_i^2) = w_1|_{K_i}(\mathbf{x}_i)$ and $w_1(\mathbf{x}_i^1) = w_1|_{K_{i+1}}(\mathbf{x}_i)$.

The mean value theorem, the Cauchy–Schwarz inequality, and (51) give

$$|w_1(\mathbf{x})|^2 = |w_1(\mathbf{x}_0)|^2 = \left|\sum_{i=1}^{l}(w_1(\mathbf{x}_{i-1}^1) - w_1(\mathbf{x}_i^2)) + \sum_{i=1}^{l}[w_1]_*(\mathbf{x}_i)\right|^2 \tag{52}$$

$$\le Cl\left(\sum_{i=1}^{l}\nabla w_1(\bar{\mathbf{x}}_i)^2(\mathbf{x}_{i-1} - \mathbf{x}_i)^2 + \sum_{i=1}^{l}[\gamma w_1]_{e_i}^2\right), \tag{53}$$

where $\bar{\mathbf{x}}_i \in K_i$ is a point between $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$. As in [8], we have

$$|\nabla w_1(\bar{\mathbf{x}}_i)|^2 h^2 \le C|\nabla w_1|_{K_i}^2. \tag{54}$$

(54) implies

$$|w_1(\mathbf{x})|^2 \le Cl\left(\sum_{i=1}^{l}|\nabla w_1|_{K_i}^2+\sum_{i=1}^{l}[\gamma w_1]_{e_i}^2\right). \tag{55}$$

Integrating (55) over K gives

$$\int_K |w_1(\mathbf{x})|^2 \le Clh^2\left(\sum_{i=1}^{l}|\nabla w_1|_{K_i}^2+\sum_{i=1}^{l}[\gamma w_1]_{e_i}^2\right). \tag{56}$$

Adding over K in such way that the same K_i appears at most l times and using the fact that $lh \le C_0$, we have

$$\|w_1\|^2=\int_\Omega |w_1(\mathbf{x})|^2 \le C\left(|w_1|_{1,h}^2+\sum_{e\in\Gamma}[\gamma w_1]_e^2\right). \tag{57}$$

Similarly, we have (57) hold for w_2. Since $[\![\gamma\mathbf{w}]\!]^2=[\gamma w_1]^2+[\gamma w_2]^2$, we have completed the proof. □

THEOREM 4.4. *Let* $(\mathbf{u}_h,p_h)\in V_h\times Q_h$ *be the solution of* (15)–(16) *or* (29)–(30) *and* $(\mathbf{u},p)\in (H^2(\Omega)\cap H_0^1(\Omega))^2\times L_0^2(\Omega)\cap H^1(\Omega)$ *be the solution of* (1)–(3). *Then there exists a constant* C *independent of* h *such that*

$$\|\mathbf{u}-\mathbf{u}_h\| \le Ch(\|\mathbf{u}\|_2+\|p\|_1). \tag{58}$$

Proof. Using Lemma 4.3, (43), and (44), we have

$$\begin{aligned}\|\mathbf{u}_h-\Pi_1\mathbf{u}\| &\le C[\![\mathbf{u}_h-\Pi_1\mathbf{u}]\!] \le C([\![\mathbf{u}-\mathbf{u}_h]\!]+[\![\mathbf{u}-\Pi_1\mathbf{u}]\!])\\ &\le Ch(\|\mathbf{u}\|_2+\|p\|_1).\end{aligned} \tag{59}$$

(59), (38), and the triangle inequality imply (58). We have completed the proof. □

REFERENCES

[1] D. ARNOLD, F. BREZZI, B. COCKBURN, AND D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*.

[2] I. BABUSKA AND M. ZLAMAL, *Nonconforming elements in the finite element method with penalty*, SIAM J. Numer. Anal., 10 (1973), pp. 863–875.

[3] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[4] C. E. BAUMANN AND J. T. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[5] Z. CAI, J. MANDEL, AND S. MCCORMICK, *The finite volume element method for diffusion equations on general triangulations*, SIAM J. Numer. Anal., 28 (1991), pp. 392–403.

[6] Z. CAI AND S. MCCORMICK, *On the accuracy of the finite volume element method for diffusion equations on composite grids*, SIAM J. Numer. Anal., 27 (1990), pp. 636–655.

[7] S. H. CHOU, *Analysis and convergence of a covolume method for the generalized Stokes problem*, Math. Comput., 217 (1997), pp. 85–104.

[8] S. H. CHOU AND D. Y. KWAK, *A covolume method based on rotated bilinears for the generalized Stokes problem*, SIAM J. Numer. Anal., 2 (1998), pp. 494–507.

[9] S. H. CHOU AND D. Y. KWAK, *Analysis and convergence of a MAC scheme for the generalized Stokes problem*, Numer. Methods Partial Differential Equations, 13 (1997), pp. 147–162.

[10] S. H. Chou and P. S. Vassilevski, *A general mixed co-volume framework for constructing conservative schemes for elliptic problems*, Math. Comput., 68 (1999), pp. 991–1011.
[11] P. Castillo, B. Cockburn, I. Perugia, and D. Schoyzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.
[12] B. Cockburn, G. E. Karniadakis, and C. W. Shu, eds., *The Discontinuous Galerkin Methods: Theory, Computation and Applications*, Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, New York, 2000.
[13] B. Cockburn, G. Kanschat, D. Schotzau, and C. Schwab, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.
[14] B. Cockburn and C. W. Shu, *The local discontinuous Galerkin finite element method for convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
[15] M. Crouzeix and P. A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Anal. Numer., 7 (1973), pp. 33–76.
[16] J. Douglas, Jr., and T. Dupont, *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976.
[17] R. Lazarov, I. Michev, and P. Vassilevski, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.
[18] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.
[19] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Tech Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NV, 1973.
[20] B. Riviere, M. F. Wheeler, and V.Girault, *Improved Energy Estimates for Interior Penalty, Constrained and Discontinuous Galerkin Methods for Elliptic Problems, Part* I, Tech Report 99-09, TICAM, 1999.
[21] X. Ye, *A new discontinuous finite volume method for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 1062–1072.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 199–223

MULTIGRID ALGORITHMS FOR C^0 INTERIOR PENALTY METHODS*

SUSANNE C. BRENNER† AND LI-YENG SUNG†

Abstract. Multigrid algorithms for C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains are studied in this paper. It is shown that V-cycle, F-cycle and W-cycle algorithms are contractions if the number of smoothing steps is sufficiently large. The contraction numbers of these algorithms are bounded by $Cm^{-\alpha}$, where m is the number of presmoothing (and postsmoothing) steps, α is the index of elliptic regularity, and the positive constant C is mesh-independent. These estimates are established for a smoothing scheme that uses a Poisson solve as a preconditioner, which can be easily implemented because the C^0 finite element spaces are standard spaces for second order problems. Furthermore the variable V-cycle algorithm is also shown to be an optimal preconditioner.

Key words. multigrid methods, discontinuous Galerkin methods, fourth order problems

AMS subject classifications. 65N55, 65N30

DOI. 10.1137/040611835

1. Introduction.

C^0 interior penalty methods [29, 24] are nonconforming finite element methods for fourth order problems. Consider the following variational problem on a bounded polygonal domain in $\mathbb{R}^2$: Find $u \in H_0^2(\Omega)$ such that

$$a(u,v) = \int_\Omega f v \, dx \qquad \forall\, v \in H_0^2(\Omega), \tag{1.1}$$

where

$$a(w,v) = \sum_{i,j=1}^{2} \int_\Omega \frac{\partial^2 w}{\partial x_i \partial x_j} \frac{\partial^2 v}{\partial x_i \partial x_j} \, dx + \int_\Omega b(x) \nabla w \cdot \nabla v \, dx \tag{1.2}$$

and $f \in L_2(\Omega)$. The function $b(x)$ in (1.2) belongs to $C^1(\bar{\Omega})$ and is nonnegative on Ω. Since $\partial\Omega$ is not smooth, the solution u of (1.1) does not belong to $H^4(\Omega)$ even if $f \in C^\infty(\bar{\Omega})$ [31, 37]. In general the shift theorem [28, 4] only holds for f belonging to the Sobolev space $H^{-2+\alpha}(\Omega)$ for some $\alpha \in (\frac{1}{2}, 1]$, i.e., $u \in H^{2+\alpha}(\Omega)$ whenever $f \in H^{-2+\alpha}(\Omega)$ and

$$\|u\|_{H^{2+\alpha}(\Omega)} \le C_\Omega \|f\|_{H^{-2+\alpha}(\Omega)}. \tag{1.3}$$

(We follow the standard notation of Sobolev spaces [1, 27, 23] in this paper.)

When $b = 0$, the variational problem defined by (1.1) corresponds to the biharmonic problem. When $b > 0$, it is a scalar analog of the elliptic system that appears in strain-gradient elasticity theory [30, 41, 44]. Within the framework of finite element methods, it can be solved numerically by conforming C^1 finite elements [9, 2], nonconforming finite elements [36, 3, 40, 39] and mixed finite elements [26].

*Received by the editors July 18, 2004; accepted for publication September 28, 2005; published electronically February 8, 2006.

http://www.siam.org/journals/sinum/44-1/61183.html

†Department of Mathematics, University of South Carolina, Columbia, SC 29208 (brenner@math.sc.edu, sung@math.sc.edu). The work of the first author was supported in part by the National Science Foundation under grant DMS-03-11790.

Let $\mathcal{T}_h$ be either a simplicial triangulation or a convex quadrilateral triangulation of Ω. In the C^0 interior penalty method approach, the discrete space V_h is either a P_ℓ ($\ell \geq 2$) triangular Lagrange finite element space [27, 23] or a Q_ℓ ($\ell \geq 2$) quadrilateral Lagrange tensor product finite element space [27, 23] associated with $\mathcal{T}_h$. The discrete problem for (1.1) is then given by: Find $u_h \in V_h$ such that

$$\mathcal{A}_h(u_h, v) = \int_\Omega f v \, dx \qquad \forall\, v \in V_h, \tag{1.4}$$

where

$$\begin{aligned}\mathcal{A}_h(w,v) = &\sum_{D\in\mathcal{T}_h} \int_D \left(\sum_{i,j=1}^{2} \frac{\partial^2 w}{\partial x_i \partial x_j} \frac{\partial^2 v}{\partial x_i \partial x_j} + b(x) \nabla w \cdot \nabla v \right) dx \\ &+ \sum_{e\in\mathcal{E}_h} \int_e \left(\left\{\!\!\left\{ \frac{\partial^2 w}{\partial n^2} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] + \left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial w}{\partial n} \right]\!\!\right] \right) ds \\ &+ \sum_{e\in\mathcal{E}_h} \frac{\eta}{|e|} \int_e \left[\!\!\left[\frac{\partial w}{\partial n} \right]\!\!\right] \left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] ds,\end{aligned} \tag{1.5}$$

$\mathcal{E}_h$ is the set of all the edges of $\mathcal{T}_h$, $|e|$ is the length of the edge e, and $\eta > 0$ is a penalty parameter. The averages $\{\!\{\cdot\}\!\}$ and jumps $[\![\cdot]\!]$ in (1.5) are defined as follows.

Let e be an interior edge of $\mathcal{T}_h$ and n_e be a unit vector normal to e. Then e is shared by two elements $D_\pm \in \mathcal{T}_h$, where n_e is pointing from D_- to D_+, and we define on e

$$\left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] = \frac{\partial v_+}{\partial n_e} - \frac{\partial v_-}{\partial n_e} \quad \text{and} \quad \left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} = \frac{1}{2} \left(\frac{\partial^2 v_+}{\partial n_e^2} + \frac{\partial^2 v_-}{\partial n_e^2} \right), \tag{1.6}$$

where $v_\pm = v\big|_{D_\pm}$. For an edge e on $\partial\Omega$, we take n_e to be the outer unit normal vector and define

$$\left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] = -\frac{\partial v}{\partial n_e} \quad \text{and} \quad \left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} = \frac{\partial^2 v}{\partial n_e^2}.$$

Note that the averages and jumps are independent of the choice of n_e in (1.6).

The C^0 interior penalty method is consistent, and for η sufficiently large (which is assumed to be the case) it is also stable. Therefore the error $u - u_h$ is quasi-optimal with respect to appropriate norms [29, 24]. The C^0 interior penalty approach has certain advantages over other finite element methods:

(i) The finite element spaces are much simpler than the C^1 finite element spaces;
(ii) The lowest order C^0 interior penalty methods (i.e., those based on the P_2 or Q_2 elements) are as simple as the classical nonconforming finite element methods, but unlike such methods, the C^0 interior penalty methods come in a natural hierarchy of arbitrary orders;
(iii) Unlike mixed finite element methods, it is straightforward to construct C^0 interior penalty methods for more complicated elliptic systems;
(iv) The fact that the finite element spaces in the C^0 interior penalty approach are just standard finite element spaces for second order problems can be exploited in the design of effective smoothers for multigrid algorithms.

Remark 1.1. In the absence of hanging nodes, which is the case here, the C^0 interior penalty methods involve fewer degrees of freedom than the completely discontinuous interior penalty methods introduced in [5] because nodal values are shared on interelement boundaries.

In this paper we extend the multigrid theory for classical nonconforming finite elements (cf. [19, 22] and the references therein) to the C^0 interior penalty methods. We will prove the convergence of V-cycle, F-cycle and W-cycle algorithms when the number of smoothing steps is sufficiently large and also the optimality of the variable V-cycle algorithm as a preconditioner. In all these multigrid algorithms we use a preconditioned relaxation scheme that is much more effective than classical smoothers (such as the Richardson and the Gauss-Seidel iterations) and at the same time can be easily implemented because the finite element spaces of the C^0 interior penalty methods are the standard spaces for second order problems.

The rest of the paper is organized as follows. We set the notation and state the multigrid algorithms in section 2, and then we introduce the mesh-dependent norms and establish some basic estimates in section 3. The analysis of W-cycle and variable V-cycle algorithms is carried out in section 4. The analysis of V-cycle and F-cycle algorithms relies on the additive theory developed in [20, 22], which is recalled in section 5. The convergence results for V-cycle and F-cycle algorithms are then established in section 6. In section 7 we present the results of numerical experiments. Appendix A contains some properties of multigrid Poisson solves relevant for the convergence analysis.

For future reference we state here two elementary inequalities:

$$2ab \le \theta^2 a^2 + \theta^{-2} b^2 \qquad \text{for} \quad a, b \in \mathbb{R} \quad \text{and} \quad \theta \in (0,1), \tag{1.7}$$

$$(a+b)^2 \le (1+\theta^2) a^2 + (1+\theta^{-2}) b^2 \qquad \text{for} \quad a, b \in \mathbb{R} \quad \text{and} \quad \theta \in (0,1). \tag{1.8}$$

2. Multigrid algorithms. In this section we describe the multigrid algorithms. In view of their potential for three-dimensional (3D) problems, we will focus on C^0 interior penalty methods that are based on quadrilateral elements. Similar results can of course be obtained for triangular elements.

Let $\mathcal{T}_0$ be a triangulation of Ω by convex quadrilaterals and the triangulations of $\mathcal{T}_1, \mathcal{T}_2, \ldots$ be obtained from $\mathcal{T}_0$ through uniform subdivisions. The mesh sizes $h_k = \max_{D \in \mathcal{T}_k} \operatorname{diam} D$ thus satisfy the relation

$$h_k \approx 2^{-k} h_0. \tag{2.1}$$

Remark 2.1. In order to avoid the proliferation of constants, we will use the notation $A \lesssim B$ ($B \gtrsim A$) to represent the relation $A \le (\text{constant}) \times B$, where the positive constant is mesh-independent, i.e., it is independent of the mesh size h_k and the grid level k. The notation $A \approx B$ is equivalent to $A \lesssim B$ and $B \lesssim A$.

Let $V_k \subset H_0^1(\Omega)$ be the Q_ℓ ($\ell \ge 2$) finite element space associated with $\mathcal{T}_k$ and denote by $\mathcal{A}_k$ the symmetric bilinear form on V_k corresponding to the variational form (1.5) of the C^0 interior penalty method. The kth level discrete problem for the C^0 interior penalty method is: Find $u_k \in V_k$ such that

$$\mathcal{A}_k(u_k, v) = \int_\Omega f v \, dx \qquad \forall\, v \in V_k. \tag{2.2}$$

For η sufficiently large, the bilinear form $\mathcal{A}_k(\cdot,\cdot)$ is positive definite on V_k and we can define the discrete energy norm $\|\cdot\|_{\mathcal{A}_k}$ by

$$\|v\|_{\mathcal{A}_k} = \sqrt{\mathcal{A}_k(v,v)} \qquad \forall\, v \in V_k. \tag{2.3}$$

Note that $\mathcal{A}_k(\zeta_1,\zeta_2)$ is well-defined for $\zeta_1,\zeta_2 \in H^{2+\alpha}(\Omega)\cap H_0^2(\Omega)$, where $\alpha \in (1/2,1]$ is the index of elliptic regularity in (1.3). In fact, $\mathcal{A}_k(\zeta_1,\zeta_2) = a(\zeta_1,\zeta_2)$ because $[\![\partial\zeta_j/\partial n]\!] = 0$. In particular, in view of (1.2) and the Poincaré-Friedrichs inequality [38],

$$\mathcal{A}_k(\zeta,\zeta) = a(\zeta,\zeta) \approx |\zeta|^2_{H^2(\Omega)} \qquad \forall\, \zeta \in H^{2+\alpha}(\Omega)\cap H_0^2(\Omega). \tag{2.4}$$

However, $\mathcal{A}_k(\cdot,\cdot)$ is not positive definite on the space $V_k + [H^{2+\alpha}(\Omega)\cap H_0^2(\Omega)]$. Therefore it is necessary to introduce the following norm $\|\cdot\|_k$ for functions in $V_k + [H^{2+\alpha}(\Omega)\cap H_0^2(\Omega)]$:

$$\begin{aligned}\|w\|_k^2 = &\sum_{D\in\mathcal{T}_k}\Big(|w|^2_{H^2(D)} + |w|^2_{H^1(D)}\Big)\\ &+\sum_{e\in\mathcal{E}_k}\Big(|e|\,\|\{\!\{\partial^2 w/\partial n^2\}\!\}\|^2_{L_2(e)} + |e|^{-1}\|[\![\partial w/\partial n]\!]\|^2_{L_2(e)}\Big).\end{aligned} \tag{2.5}$$

From (2.5) it is easy to see that

$$|\mathcal{A}_k(w_1,w_2)| \lesssim \|w_1\|_k\|w_2\|_k \qquad \forall\, w_1,w_2 \in V_k + [H^{2+\alpha}(\Omega)\cap H_0^2(\Omega)]. \tag{2.6}$$

Furthermore, on V_k itself, we have (cf. (4.18), (4.20) and (4.25) of [24])

$$|v|_{H^2(\Omega,\mathcal{T}_k)} \le \|v\|_k \approx \|v\|_{\mathcal{A}_k} \lesssim |v|_{H^2(\Omega,\mathcal{T}_k)} \qquad \forall\, v \in V_k, \tag{2.7}$$

where

$$|v|^2_{H^2(\Omega,\mathcal{T}_k)} = \sum_{D\in\mathcal{T}_k}|v|^2_{H^2(D)} + \sum_{e\in\mathcal{E}_k}|e|^{-1}\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)}. \tag{2.8}$$

Let the operator $A_k : V_k \longrightarrow V_k'$ be defined by

$$\langle A_k v_1, v_2\rangle = \mathcal{A}_k(v_1,v_2) \qquad \forall\, v_1,v_2 \in V_k, \tag{2.9}$$

where $\langle\cdot,\cdot\rangle$ is the canonical bilinear form between a vector space and its dual. We can then rewrite the discrete problem (2.2) as $A_k u_k = \phi_k$, where $\phi_k \in V_k'$ is defined by $\langle\phi_k, v\rangle = \int_\Omega f v\,dx\ \forall\, v \in V_k$.

Multigrid algorithms [32, 35, 11, 17, 43] are iterative methods for the solution of equations of the form

$$A_k z = \psi, \tag{2.10}$$

where $\psi \in V_k'$ and $z \in V_k$. In the descriptions of the multigrid algorithms below we will denote the natural injection from V_{k-1} to V_k by I_{k-1}^k and its transpose from V_k' to V_{k-1}' by I_k^{k-1}, i.e.,

$$\langle\phi, I_{k-1}^k v\rangle = \langle I_k^{k-1}\phi, v\rangle \qquad \forall\, \phi \in V_k' \quad\text{and}\quad v \in V_{k-1}. \tag{2.11}$$

We also need an operator $B_k : V_k \longrightarrow V_k'$ in the preconditioned relaxation scheme used in the smoothing steps of the multigrid algorithms (cf. (2.17) and (2.19) below). Let $L_k : V_k \longrightarrow V_k'$ be the discrete Laplace operator, i.e.,

$$\langle L_k v_1, v_2\rangle = \int_\Omega \nabla v_1\cdot\nabla v_2\,dx \qquad \forall\, v_1,v_2 \in V_k. \tag{2.12}$$

Since V_k is a standard finite element space for second order problems, it is natural to consider L_k^{-1} as a preconditioner for the fourth order discrete differential operator A_k. In order to maintain the optimal complexity of multigrid algorithms, we use instead an approximation B_k of L_k with the following properties:

(i) B_k is symmetric positive definite, i.e.,

$$\langle B_k v_1, v_2\rangle = \langle B_k v_2, v_1\rangle \qquad \forall\, v_1, v_2 \in V_k, \tag{2.13}$$

$$\langle B_k v, v\rangle > 0 \qquad \forall\, v \in V_k \setminus \{0\}. \tag{2.14}$$

(ii) B_k is spectrally equivalent to the discrete Laplace operator in the sense that

$$\langle L_k v, v\rangle \le \langle B_k v, v\rangle \lesssim \langle L_k v, v\rangle = \|\nabla v\|^2_{L_2(\Omega)} \qquad \forall\, v \in V_k. \tag{2.15}$$

(iii) B_k approximates L_k in the sense that, for some $\beta \in (0, 1/2)$,

$$|v - B_k^{-1} L_k v|_{H^1(\Omega)} \lesssim h_k^{\beta} \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_k. \tag{2.16}$$

(iv) The cost for computing $B_k^{-1} v$ is of order $O(n_k)$, where n_k is the dimension of V_k.

Remark 2.2. Let $B_k^{-1} : V_k' \longrightarrow V_k$ be the Poisson solve obtained by a symmetric V-cycle algorithm, a symmetric W-cycle algorithm or a symmetric variable V-cycle algorithm. Then B_k satisfies the properties (i), (ii) and (iv). If B_k^{-1} is the Poisson solve obtained by a symmetric W-cycle algorithm with a sufficiently large number of smoothing steps or a symmetric variable V-cycle algorithm, then the operator B_k also satisfies the property (iii). Details can be found in Appendix A.

ALGORITHM 2.3 (V-cycle Algorithm). $MG_V(k, \psi, z_0, m)$ is the approximate solution of (2.10) with initial guess z_0 obtained as follows. If $k = 0$, we use a direct solve to obtain $A_0^{-1}\psi$ as the output of the V-cycle algorithm. If $k \ge 1$, we compute $MG_V(k, \psi, z_0, m)$ recursively in three steps.

Presmoothing. For $1 \le j \le m$, compute z_j recursively by

$$z_j = z_{j-1} + \gamma_k B_k^{-1}(\psi - A_k z_{j-1}), \tag{2.17}$$

where γ_k^{-1} dominates the spectral radius of the operator $B_k^{-1} A_k : V_k \longrightarrow V_k$.

Coarse Grid Correction. Compute

$$z_{m+1} = z_m + I_{k-1}^k MG_V(k-1, \varrho_{k-1}, 0, m), \tag{2.18}$$

where $\varrho_{k-1} = I_k^{k-1}(\psi - A_k z_m)$ is the transferred residual of z_m.

Postsmoothing. For $m + 2 \le j \le 2m + 1$, compute z_j recursively by

$$z_j = z_{j-1} + \gamma_k B_k^{-1}(\psi - A_k z_{j-1}). \tag{2.19}$$

The final output of the V-cycle algorithm is

$$MG_V(k, \psi, z_0, m) = z_{2m+1}. \tag{2.20}$$

ALGORITHM 2.4 (W-cycle Algorithm). If we replace the coarse grid correction step of Algorithm 2.3 by the following procedure, we have the W-cycle algorithm whose output will be denoted by $MG_W(k, \psi, z_0, m)$.

Coarse Grid Correction for the W-cycle. Compute $e_1, e_2 \in V_{k-1}$ by

$$e_j = MG_W(k-1, \varrho_{k-1}, e_{j-1}, m) \qquad \text{for} \quad 1 \le j \le 2, \tag{2.21}$$

where $e_0 = 0$, and set

$$z_{m+1} = z_m + I_{k-1}^k e_2. \tag{2.22}$$

ALGORITHM 2.5 (F-cycle Algorithm). If we replace the coarse grid correction step of Algorithm 2.3 by the following procedure, we have the F-cycle algorithm whose output will be denoted by $MG_F(k, \psi, z_0, m)$.

Coarse Grid Correction for the F-cycle. Let $e_0 = 0 \in V_{k-1}$. Compute $e_1, e_2 \in V_{k-1}$ by $e_1 = MG_F(k-1, \varrho_{k-1}, e_0, m)$, $e_2 = MG_V(k-1, \varrho_{k-1}, e_1, m)$, and set z_{m+1} by (2.22).

ALGORITHM 2.6 (Variable V-cycle Algorithm). If the numbers of smoothing steps in Algorithm 2.3 on different levels are allowed to be different, we have a variable V-cycle algorithm.

3. Mesh-dependent norms and preliminary estimates. In this section we introduce mesh-dependent norms and derive some preliminary estimates. First of all, because of (2.13) and (2.14), we can introduce a discrete inner product [6] related to the preconditioner in the smoothing steps:

$$(v_1, v_2)_k = \langle B_k v_1, v_2 \rangle \qquad \forall\, v_1, v_2 \in V_k. \tag{3.1}$$

It follows from (2.9) and (3.1) that the operator $\mathbb{A}_k = B_k^{-1} A_k : V_k \longrightarrow V_k$ satisfies

$$(\mathbb{A}_k v_1, v_2)_k = \mathcal{A}_k(v_1, v_2) \qquad \forall\, v_1, v_2 \in V_k. \tag{3.2}$$

It is clear from (1.5) and (3.2) that $\mathbb{A}_k$ is symmetric positive definite with respect to the inner product $(\cdot,\cdot)_k$. Furthermore, it follows from (2.3), (2.7), (2.8), (2.15) and standard inverse estimates [27, 23] that the spectral radius $\rho(\mathbb{A}_k)$ of $\mathbb{A}_k$ satisfies

$$\rho(\mathbb{A}_k) \lesssim h_k^{-2}. \tag{3.3}$$

Therefore we can take the parameter γ_k in (2.17) and (2.19) to be Ch_k^2 ($\leq 1/\rho(\mathbb{A}_k)$), where the positive constant C is mesh-independent.

Remark 3.1. In terms of the inner product $(\cdot,\cdot)_k$ the smoothing steps in (2.17) and (2.19) are just Richardson relaxation steps.

Remark 3.2. Using (3.3) it is not difficult to show that, with respect to the natural nodal basis of the Q_ℓ finite element space, the condition number of $\mathbb{A}_k$ (in the energy norm) is of order $O(h_k^{-2})$. On the other hand the condition number of the fourth order discrete differential operator A_k (with respect to the natural nodal basis) is of order $O(h_k^{-4})$. The reduction in the order of the condition number of $\mathbb{A}_k$ greatly improves the performance of the multigrid algorithms (cf. Remark 4.2, Tables 7.1 and 7.4 below).

For $s \in \mathbb{R}$, we define the mesh-dependent norm $|\!|\!| \cdot |\!|\!|_{s,k}$ by

$$|\!|\!| v |\!|\!|_{s,k} = (\mathbb{A}_k^s v, v)_k^{1/2} \qquad \forall\, v \in V_k. \tag{3.4}$$

It is clear from (2.3), (2.15), (3.1), (3.2) and (3.4) that

$$|\!|\!| v |\!|\!|_{0,k} = \sqrt{(v,v)_k} = \sqrt{\langle B_k v, v \rangle} \approx |v|_{H^1(\Omega)} \qquad \forall\, v \in V_k, \tag{3.5}$$

$$|\!|\!| v |\!|\!|_{1,k} = \|v\|_{\mathcal{A}_k} \qquad \forall\, v \in V_k. \tag{3.6}$$

The following well-known properties [7] of mesh-dependent norms follow immediately from (3.2)–(3.4) and the Cauchy-Schwarz inequality:

$$|\!|\!| v |\!|\!|_{s,k} \lesssim h_k^{t-s} |\!|\!| v |\!|\!|_{t,k} \qquad \forall\, v \in V_k \quad \text{and} \quad 0 \le t \le s \le 2, \tag{3.7}$$

$$|\!|\!| v |\!|\!|_{1+s,k} = \sup_{w \in V_k \setminus 0} \frac{\mathcal{A}_k(v,w)}{|\!|\!| w |\!|\!|_{1-s,k}} \qquad \forall\, v \in V_k \quad \text{and} \quad s \in \mathbb{R}. \tag{3.8}$$

Our convergence analysis in subsequent sections relies on the elliptic regularity estimate (1.3). Therefore a relation between the Sobolev norms and the mesh-dependent norms is crucial. For conforming methods such a relation is easy to derive. However, since the C^0 interior penalty methods are nonconforming (i.e., $V_k \not\subset H_0^2(\Omega)$), additional work is required here.

The key ingredient for building a link between Sobolev norms and mesh-dependent norms is the existence [24] of a C^1 finite element which is a *relative* [18, 19] of the Q_ℓ Lagrange element in the sense that (i) the shape functions of the Q_ℓ element are also shape functions of the C^1 element, and (ii) the nodal variables (degrees of freedom) of the Q_ℓ element are also nodal variables of the C^1 element. For example, we can take the C^1 elements from the generalized Bogner-Fox-Schmit family (cf. section 6 of [24]) to be the relatives of the tensor product Lagrange elements, and take the C^1 Argyris elements [2, 24] to be the relatives of the triangular Lagrange elements.

Let $\tilde{V}_k \subset H_0^2(\Omega)$ be the finite element space defined by the C^1 element. We can construct a linear map $E_k : V_k \longrightarrow \tilde{V}_k$ by averaging [24] so that the following properties hold:

$$\Pi_k E_k v = v \qquad \forall\, v \in V_k, \tag{3.9}$$

$$\|E_k v\|_{H^2(\Omega)} \lesssim \|v\|_{\mathcal{A}_k} \qquad \forall\, v \in V_k, \tag{3.10}$$

$$\|E_k v\|_{H^{1+s}(\Omega)} \approx \|v\|_{H^{1+s}(\Omega)} \qquad \forall\, v \in V_k,\ 0 \le s < \frac{1}{2}, \tag{3.11}$$

where $\Pi_k : C^0(\bar{\Omega}) \longrightarrow V_k$ is the nodal interpolation operator.

Remark 3.3. The relation (3.9) and the estimate (3.10) can be found in [24, equation (3.30) and Lemma 3]. The estimate (3.11) can be proved by the arguments in Lemma 9 of [24], where the special case $s = 1 - \alpha$ is established.

Note also that the following estimates (cf. (3.16), (3.18) and (5.3) of [24]) hold for Π_k:

$$\|\Pi_k \zeta\|_{\mathcal{A}_k} \lesssim \|\zeta\|_{H^2(\Omega)} \qquad \forall\, \zeta \in H_0^2(\Omega), \tag{3.12}$$

$$|\zeta - \Pi_k \zeta|_{H^1(\Omega)} \lesssim h_k \|\zeta\|_{H^2(\Omega)} \qquad \forall\, \zeta \in H_0^2(\Omega), \tag{3.13}$$

$$\|\zeta - \Pi_k \zeta\|_k \lesssim h_k^\alpha \|\zeta\|_{H^{2+\alpha}(\Omega)} \qquad \forall \zeta \in H^{2+\alpha}(\Omega), \tag{3.14}$$

where $\|\cdot\|_k$ is the norm defined in (2.5). Furthermore, because the finite elements are relatives, the following discrete estimate is a consequence of the equivalence of norms on finite dimensional vector spaces:

$$|\Pi_k \tilde{v}|_{H^1(\Omega)} \lesssim |\tilde{v}|_{H^1(\Omega)} \qquad \forall\, v \in \tilde{V}_k. \tag{3.15}$$

The following lemma gives useful two-level estimates for the nodal interpolation operator.

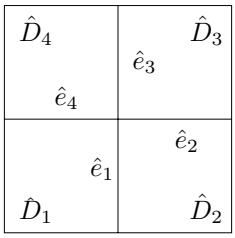


FIG. 3.1. *A subdivided referenced square* $\hat{D}$.

LEMMA 3.4. *The following estimates hold for the nodal interpolation operator*:

$$|v-\Pi_{k-1}v|_{H^1(\Omega)} \lesssim h_k\|v\|_{\mathcal{A}_k} \qquad \forall\, v\in V_k, \tag{3.16}$$

$$\|v-\Pi_{k-1}v\|_{H^{2-\alpha}(\Omega)} \lesssim h_k^{\alpha}\|v\|_{\mathcal{A}_k} \qquad \forall\, v\in V_k, \tag{3.17}$$

$$\|\Pi_{k-1}v\|_{\mathcal{A}_{k-1}} \lesssim \|v\|_{\mathcal{A}_k} \qquad \forall\, v\in V_k. \tag{3.18}$$

Proof. In view of (2.7) and (2.8), the estimate (3.16) is a consequence of

$$|v-\Pi_{k-1}v|^2_{H^1(D)} \lesssim (\operatorname{diam} D)^2\left(\sum_{\substack{D'\in\mathcal{T}_k\\ D'\subset D}}|v|^2_{H^2(D')}+\sum_{\substack{e\subset\mathcal{E}_k\\ e\subset D}}|e|^{-1}\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)}\right) \tag{3.19}$$

for all $v\in V_k$ and $D\in\mathcal{T}_{k-1}$. Since the quadrilaterals in $\mathcal{T}_k$ for $k\geq 0$ are shape regular, we can establish (3.19) by proving the following estimate on the reference square $\hat{D}$:

$$|\hat{v}-\hat{v}^I|^2_{H^1(\hat{D})} \lesssim \sum_{j=1}^4|\hat{v}|^2_{H^2(\hat{D}_j)}+\sum_{j=1}^4\|[\![\partial\hat{v}/\partial n]\!]\|^2_{L_2(\hat{e}_j)} \qquad \forall\, v\in\hat{V}, \tag{3.20}$$

where $\hat{V}\subset H^1(\hat{D})$ is the (finite dimensional) space of continuous functions whose members belong to the polynomial space $Q_\ell(\hat{D}_j)$ for each of the four subsquares $\hat{D}_j$ (cf. Figure 3.1), $\hat{v}^I\in Q_\ell(\hat{D})$ agrees with $\hat{v}$ at the nodes of the Q_ℓ element on $\hat{D}$, and $\hat{e}_j$ for $1\leq j\leq 4$ are interfaces of the subsquares. Now the estimate (3.20) follows from the observation that the square root of the right-hand side of (3.20) defines a norm on the quotient space $\hat{V}/P_1(\hat{D})$ while the square root of the left-hand side defines a seminorm on $\hat{V}/P_1(\hat{D})$.

The estimate (3.17) follows from (3.16) and the inverse estimate [8]

$$|v|_{H^{1+s}(\Omega)} \lesssim h_k^{-s}|v|_{H^1(\Omega)} \qquad \forall\, v\in V_k, \tag{3.21}$$

where $0<s<1/2$.

Finally we derive (3.18) using (2.7), (2.8), (3.18), (3.19), a trace theorem (with scaling) and a standard inverse estimate [27, 23]:

$$\begin{aligned}\|\Pi_{k-1}v\|^2_{\mathcal{A}_{k-1}} &\lesssim \sum_{D\in\mathcal{T}_{k-1}}|\Pi_{k-1}v|^2_{H^2(D)}+\sum_{e\in\mathcal{E}_k}|e|^{-1}\|[\![\partial(\Pi_{k-1}v)/\partial n]\!]\|^2_{L_2(e)}\\ &\lesssim \|\Pi_{k-1}v\|_{\mathcal{A}_{k-1}}+\sum_{e\in\mathcal{E}_k}|e|^{-1}\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)}+\sum_{e\in\mathcal{E}_k}|e|^{-1}\|[\![\partial(v-\Pi_{k-1}v)/\partial n]\!]\|^2_{L_2(e)}\end{aligned}$$

$$\lesssim \|v\|^2_{\mathcal{A}_k} + \sum_{D\in\mathcal{T}_k} (\operatorname{diam} D)^{-2}|v-\Pi_{k-1}v|^2_{H^1(D)} \lesssim \|v\|^2_{\mathcal{A}_k} \qquad \forall\, v\in V_k. \qquad \square$$

In the other direction we can also construct a map from the Sobolev spaces into V_k.

LEMMA 3.5. *There exists a linear map $J_k : L_2(\Omega) \longrightarrow V_k$ with the following properties*:

$$J_k E_k v = v \qquad \forall\, v\in V_k, \tag{3.22}$$

$$\|J_k v\|_{\mathcal{A}_k} \lesssim |v|_{H^2(\Omega)} \qquad \forall\, v\in H^2_0(\Omega), \tag{3.23}$$

$$|J_k v|_{H^1(\Omega)} \lesssim |v|_{H^1(\Omega)} \qquad \forall\, v\in H^1_0(\Omega). \tag{3.24}$$

Proof. We define J_k by

$$J_k v = \Pi_k Q_k v \qquad \forall\, v\in L_2(\Omega), \tag{3.25}$$

where $Q_k : L_2(\Omega) \longrightarrow \tilde{V}_k$ is the L_2 orthogonal projection operator. The relation (3.22) is an obvious consequence of (3.9).

Regarding Q_k we have the estimates [16]

$$\|Q_k v\|_{H^2(\Omega)} \lesssim \|v\|_{H^2(\Omega)} \qquad \forall\, v\in H^2_0(\Omega), \tag{3.26}$$

$$\|Q_k v\|_{H^1(\Omega)} \lesssim \|v\|_{H^1(\Omega)} \qquad \forall\, v\in H^1_0(\Omega). \tag{3.27}$$

The estimates (3.23) and (3.24) follow immediately from (3.12), (3.15) and (3.25)–(3.27). $\square$

LEMMA 3.6. *It holds that*

$$|\!|\!| v |\!|\!|_{s,k} \approx \|E_k v\|_{H^{1+s}(\Omega)} \qquad \forall\, v\in V_k \tag{3.28}$$

provided $0\le s\le 1$ *and* $s\ne 1/2$.

Proof. From (3.5), (3.6), (3.10) and (3.11) we have

$$\|E_k v\|_{H^2(\Omega)} \lesssim |\!|\!| v |\!|\!|_{1,k} \qquad \forall\, v\in V_k,$$
$$\|E_k v\|_{H^1(\Omega)} \lesssim |\!|\!| v |\!|\!|_{0,k} \qquad \forall\, v\in V_k,$$

which implies, by operator interpolation theory for Hilbert scales [42, 33, 11],

$$\|E_k v\|_{H^{1+s}(\Omega)} \lesssim |\!|\!| v |\!|\!|_{s,k} \qquad \forall\, v\in V_k. \tag{3.29}$$

On the other hand, from (3.5), (3.6), (3.23), (3.24) and interpolation, we have

$$|\!|\!| J_k v |\!|\!|_{s,k} \lesssim \|v\|_{H^{1+s}(\Omega)} \qquad \forall\, v\in H^{1+s}_0(\Omega), \tag{3.30}$$

which together with (3.22) implies

$$|\!|\!| v |\!|\!|_{s,k} = |\!|\!| J_k E_k v |\!|\!|_{1+s,k} \lesssim \|E_k v\|_{H^{1+s}(\Omega)} \qquad \forall\, v\in V_k. \qquad \square \tag{3.31}$$

Remark 3.7. The norm equivalence (3.28) is also valid for $s=1/2$ provided the norm on the right-hand side is replaced by the norm $\|\cdot\|_{H^{1+s}_{00}(\Omega)}$ (cf. [34, 42]).

From (3.11) and (3.28) we immediately obtain the following corollary which provides the link between mesh-dependent norms and Sobolev norms.

COROLLARY 3.8. *It holds that*

$$|\!|\!| v |\!|\!|_{s,k} \approx \|v\|_{H^{1+s}(\Omega)} \qquad \forall\, v\in V_k, \tag{3.32}$$

provided $0\le s<1/2$.

Let J_k^* be the adjoint of J_k (restricted to $H_0^2(\Omega)$) with respect to the bilinear form $a(\cdot,\cdot)$ for the continuous problem and the bilinear form $\mathcal{A}_k(\cdot,\cdot)$ for the discrete problem, i.e., $J_k^* : V_k \longrightarrow H_0^2(\Omega)$ satisfies

$$a(J_k^* v, w) = \mathcal{A}_k(v, J_k w) \qquad \forall\, v \in V_k,\ w \in H_0^2(\Omega). \tag{3.33}$$

The following lemma on J_k^* will be useful in the convergence analysis of V-cycle and F-cycle algorithms.

LEMMA 3.9. *Let $\zeta_k \in V_k$ and*

$$\phi(v) = \mathcal{A}_k(\zeta_k, J_k v) \qquad \forall\, v \in H_0^2(\Omega). \tag{3.34}$$

Then $\phi \in H^{-2+\alpha}(\Omega)$,

$$\|\phi\|_{H^{-2+\alpha}(\Omega)} \lesssim |\!|\!| \zeta_k |\!|\!|_{1+\alpha,k}, \tag{3.35}$$

and

$$\mathcal{A}_k(\zeta_k, v) = \phi(E_k v) \qquad \forall\, v \in V_k. \tag{3.36}$$

Furthermore, $\zeta = J_k^* \zeta_k \in H^{2+\alpha}(\Omega) \cap H_0^2(\Omega)$,

$$a(\zeta, w) = \phi(w) \qquad \forall\, w \in H_0^2(\Omega), \tag{3.37}$$

$$\|\zeta\|_{H^{2+\alpha}(\Omega)} \lesssim |\!|\!| \zeta_k |\!|\!|_{1+\alpha,k}, \tag{3.38}$$

and the following estimates hold:

$$\|\zeta - \zeta_k\|_k \lesssim h_k^{\alpha} |\!|\!| \zeta_k |\!|\!|_{1+\alpha,k}, \tag{3.39}$$

$$\|\zeta - \zeta_k\|_{H^{2-\alpha}(\Omega)} \lesssim h_k^{2\alpha} |\!|\!| \zeta |\!|\!|_{1+\alpha,k}. \tag{3.40}$$

Proof. From (3.8), (3.30) and (3.34) we have

$$\phi(v) \le |\!|\!| \zeta_k |\!|\!|_{1+\alpha,k} |\!|\!| J_k v |\!|\!|_{1-\alpha,k} \lesssim |\!|\!| \zeta_k |\!|\!|_{1+\alpha,k} \|v\|_{H^{2-\alpha}(\Omega)}, \tag{3.41}$$

which means that $\phi \in H^{-2+\alpha}(\Omega)$ and (3.35) is valid.

Equation (3.37) follows immediately from (3.33) and (3.34). Then $\zeta = J_k^* \zeta_k \in H^{2+\alpha}(\Omega)$ by elliptic regularity and (3.38) follows from (1.3) and (3.35).

Finally (3.22) and (3.34) imply (3.36). Therefore ζ_k is the solution of a modified C^0 interior penalty method for (3.37) studied in [24] and the error estimates (3.39) and (3.40) follow from (3.35) and Theorems 4 and 6 of [24]. ∎

4. Results for W-cycle and variable V-cycle algorithms. In this section we establish the results for W-cycle and variable V-cycle algorithms. There are two ingredients in the analysis: the *smoothing property* and the *approximation property*.

The effect of one smoothing step in (2.17) and (2.19) is measured by the operator

$$R_k = Id_k - \gamma_k \mathbb{A}_k, \tag{4.1}$$

where Id_k is the identity operator on V_k. The proof of the following result which describes the effect of the smoothing steps can be found, for example, in [32, 23].

LEMMA 4.1. *It holds that*

$$|\!|\!| R_k^m v |\!|\!|_{s,k} \lesssim h_k^{t-s} m^{(t-s)/2} |\!|\!| v |\!|\!|_{t,k} \qquad \forall\, v \in V_k \quad \textit{and} \quad 0 \le t \le s \le 2.$$

Remark 4.2. Without the preconditioner B_k^{-1}, the smoothing property becomes (for appropriately defined mesh-dependent norms)

$$\|\!| R_k^m v |\!\|_{s,k} \lesssim h_k^{t-s} m^{(t-s)/4} \|\!| v |\!\|_{t,k}.$$

In other words, the effect of m smoothing steps without preconditioning is (roughly) equivalent to the smoothing effect of $\sqrt{m}$ many smoothing steps with preconditioning. Therefore the preconditioner greatly enhances the performance of the multigrid algorithms (cf. Tables 7.1 and 7.4 below).

To measure the effect of coarse grid correction, we first recall the following well-known relation $\hat{e}_{k-1} = P_k^{k-1}(z - z_m)$ between the exact solution of the coarse grid residual equation $A_{k-1}\hat{e}_{k-1} = \varrho_{k-1}$ and the error $z - z_m$, where the operator $P_k^{k-1} : V_k \longrightarrow V_{k-1}$ is defined by

$$\mathcal{A}_{k-1}(P_k^{k-1}v, w) = \mathcal{A}_k(v, I_{k-1}^k w) \qquad \forall\, v \in V_k,\ w \in V_{k-1}. \tag{4.2}$$

The approximation property in the following result describes the effect of coarse grid correction.

LEMMA 4.3. *It holds that*

$$\|\!| (Id_k - I_{k-1}^k P_k^{k-1})v |\!\|_{1-\alpha,k} \lesssim h_k^{2\alpha} \|\!| v |\!\|_{1+\alpha,k} \qquad \forall\, v \in V_k, \tag{4.3}$$

where α is the index of elliptic regularity in (1.3).

Proof. Let $v \in V_k$ be arbitrary. We will establish (4.3) by a duality argument. Using the norm equivalence in Corollary 3.8 (with $s = 1 - \alpha$) and duality, we find

$$\begin{aligned} \|\!| (Id_k - I_{k-1}^k P_k^{k-1})v |\!\|_{1-\alpha,k} &\approx \|(Id_k - I_{k-1}^k P_k^{k-1})v\|_{H^{2-\alpha}(\Omega)} \\ &= \sup_{\phi \in H^{-2+\alpha}(\Omega)\setminus\{0\}} \frac{\phi((Id_k - I_{k-1}^k P_k^{k-1})v)}{\|\phi\|_{H^{-2+\alpha}(\Omega)}}. \end{aligned} \tag{4.4}$$

Let $\phi \in H^{-2+\alpha}(\Omega)$ be arbitrary and define $\zeta \in H_0^2(\Omega)$, $\zeta_k \in V_k$ and $\zeta_{k-1} \in V_{k-1}$ by

$$a(\zeta, v) = \phi(v) \qquad \forall\, v \in H_0^2(\Omega), \tag{4.5}$$

$$\mathcal{A}_k(\zeta_k, v) = \phi(v) \qquad \forall\, v \in V_k, \tag{4.6}$$

$$\mathcal{A}_{k-1}(\zeta_{k-1}, v) = \phi(v) \qquad \forall\, v \in V_{k-1}. \tag{4.7}$$

In other words, ζ_k and ζ_{k-1} are the approximations of ζ obtained by the C^0 interior penalty method, and the following error estimates (cf. Theorem 5 of [24]) are valid:

$$\|\zeta - \zeta_k\|_{H^{2-\alpha}(\Omega)} \lesssim h_k^{2\alpha} \|\phi\|_{H^{-2+\alpha}(\Omega)}, \tag{4.8}$$

$$\|\zeta - \zeta_{k-1}\|_{H^{2-\alpha}(\Omega)} \lesssim h_{k-1}^{2\alpha} \|\phi\|_{H^{-2+\alpha}(\Omega)}. \tag{4.9}$$

From (4.6) and (4.7) we have $\mathcal{A}_{k-1}(\zeta_{k-1}, v) = \mathcal{A}_k(\zeta_k, I_{k-1}^k v)$ for all $v \in V_{k-1}$, which implies (cf. (4.2))

$$\zeta_{k-1} = P_k^{k-1}\zeta_k. \tag{4.10}$$

We can now estimate the numerator in (4.4) by (2.1), (3.8), Corollary 3.8, (4.2), (4.6) and (4.8)–(4.10) as follows:

$$
\begin{aligned}
\phi((Id_k - I_{k-1}^k P_k^{k-1})v) &= \mathcal{A}_k(\zeta_k, v) - \mathcal{A}_k(\zeta_k, I_{k-1}^k P_k^{k-1} v) \\
&= \mathcal{A}_k(\zeta_k, v) - \mathcal{A}_{k-1}(P_k^{k-1}\zeta_k, P_k^{k-1} v) \\
&= \mathcal{A}_k(\zeta_k, v) - \mathcal{A}_{k-1}(\zeta_{k-1}, P_k^{k-1} v) \\
&= \mathcal{A}_k(\zeta_k - I_{k-1}^k \zeta_{k-1}, v) \qquad (4.11) \\
&\le |\!|\!| \zeta_k - \zeta_{k-1} |\!|\!|_{1-\alpha,k} |\!|\!| v |\!|\!|_{1+\alpha,k} \\
&\lesssim \|\zeta_k - \zeta_{k-1}\|_{H^{2-\alpha}(\Omega)} |\!|\!| v |\!|\!|_{1+\alpha,k} \\
&\le \big(\|\zeta_k - \zeta\|_{H^{2-\alpha}(\Omega)} + \|\zeta - \zeta_{k-1}\|_{H^{2-\alpha}(\Omega)}\big) |\!|\!| v |\!|\!|_{1+\alpha,k} \\
&\lesssim h_k^{2\alpha} \|\phi\|_{H^{-2+\alpha}(\Omega)} |\!|\!| v |\!|\!|_{1+\alpha,k}.
\end{aligned}
$$

The estimate (4.3) follows from (4.4) and (4.11). □

We can now apply the theory developed in [19, Theorem 4.3, Theorem 4.4, Lemma 4.7 and Theorem 4.8, where the results in [15] for the variable V-cycle is used] to derive the following results for W-cycle and variable V-cycle algorithms.

THEOREM 4.4. *The output $MG_W(k, \psi, z_0, m)$ of the W-cycle algorithm (Algorithm 2.4) applied to* (2.10) *satisfies the following estimate*:

$$
\|z - MG_W(k, \psi, z_0, m)\|_{\mathcal{A}_k} \le \frac{C}{m^\alpha} \|z - z_0\|_{\mathcal{A}_k},
$$

where the positive constant C is mesh-independent, provided that the number of smoothing steps m is greater than a positive integer m_* that is also mesh-independent.

THEOREM 4.5. *The variable V-cycle algorithm (Algorithm 2.6) is an optimal preconditioner provided the following relation is satisfied by m_k (the number of smoothing steps on level k)* :

$$
\beta_0 m_k \le m_{k-1} \le \beta_1 m_k, \qquad (4.12)
$$

where $1 < \beta_0 \le \beta_1$.

Remark 4.6. Theorems 4.4 and 4.5 have been obtained for preconditioners that satisfy (2.13)–(2.15). Therefore they are valid for B_k^{-1} obtained by a symmetric V-cycle algorithm, a symmetric W-cycle algorithm or a variable V-cycle algorithm (cf. Remark 2.2 and Appendix A).

Finally we note that (3.7) and (4.3) imply

$$
\begin{aligned}
|\!|\!| P_k^{k-1} v |\!|\!|_{1-\alpha,k} &\le |\!|\!| v |\!|\!|_{1-\alpha,k} + |\!|\!| v - I_{k-1}^k P_k^{k-1} v |\!|\!|_{1-\alpha,k} \qquad (4.13) \\
&\lesssim |\!|\!| v |\!|\!|_{1-\alpha,k} + h_k^{2\alpha} |\!|\!| v |\!|\!|_{1+\alpha,k} \lesssim |\!|\!| v |\!|\!|_{1-\alpha,k} \qquad \forall\, v \in V_k.
\end{aligned}
$$

The estimate (4.13) will be used in the convergence analysis of V-cycle and F-cycle algorithms.

5. Additive multigrid theory. In this section we briefly review the additive multigrid theory [20, 22] which will be used in the convergence analysis of V-cycle and F-cycle algorithms in section 6.

Let $\mathbb{E}_{k,m} : V_k \longrightarrow V_k$ be the error propagation operator for the kth level V-cycle algorithm, i.e., $z - MG_V(k, \psi, z_0, m) = \mathbb{E}_{k,m}(z - z_0)$, where $MG_V(k, \psi, z_0, m)$ is the

approximate solution of (2.10) obtained by the V-cycle algorithm with initial guess z_0. The operators $\mathbb{E}_k$ satisfy the well-known recurrence relation [32, 35]

$$\mathbb{E}_{k,m} = R_k^m(Id_k - I_{k-1}^k P_k^{k-1} + I_{k-1}^k \mathbb{E}_{k-1,m} P_k^{k-1}) R_k^m \tag{5.1}$$

and the initial condition $\mathbb{E}_k = 0$. Iterating (5.1) leads to the following additive expression [20, 22] for $\mathbb{E}_k$:

$$\mathbb{E}_{k,m} = \sum_{j=2}^{k} T_{k,j,m} R_j^m (Id_j - I_{j-1}^j P_j^{j-1}) R_j^m T_{j,k,m}, \tag{5.2}$$

where (for $j < k$) $T_{k,j,m}$ is the multilevel operator $R_k^m I_{k-1}^k \cdots R_{j+1}^m I_j^{j+1}$ from V_j into V_k, $T_{j,k,m} = P_{j+1}^j R_j^m \cdots P_k^{k-1} R_k^m$ is the adjoint operator of $T_{k,j,m}$ with respect to $\mathcal{A}_k(\cdot,\cdot)$, and $T_{k,k,m} = Id_k$.

A convergence theory for the V-cycle algorithm based on the additive expression (5.2) was developed in [20, 21] for second order problems. It yields the asymptotic behavior of the contraction numbers, which when combined with the results from the multiplicative theory [14, 45, 12, 13] provides a complete generalization of the classical result of Braess and Hackbusch [10] to the case of less than full elliptic regularity. This additive theory has been extended to V-cycle and F-cycle algorithms for classical nonconforming finite elements [22, 46, 47] and to interior penalty methods for second order problems [25].

Note the operator $R_j^m(Id_j - I_{j-1}^j P_j^{j-1}) R_j^m$ that appears in (5.2) is already controlled by the smoothing property (Lemma 4.1) and the approximation property (Lemma 4.3). Therefore the key in the additive approach is to control the multilevel operators $T_{k,j,m}$ and $T_{j,k,m}$. This in turn requires a careful comparison of the mesh-dependent norms on consecutive levels. In this regard the following assumptions of the additive theory [20, 22] need to be verified:

$$\|I_{k-1}^k v\|_{1,k}^2 \le (1+\theta^2)\|v\|_{1,k-1}^2 + C_1\theta^{-2}h_k^{2\mu}\|v\|_{1+\mu,k-1}^2 \qquad \forall\, v \in V_{k-1}, \tag{5.3}$$

$$\|I_{k-1}^k v\|_{1-\tau,k}^2 \le (1+\theta^2)\|v\|_{1-\tau,k-1}^2 + C_2\theta^{-2}h_k^{2\tau}\|v\|_{1,k-1}^2 \qquad \forall\, v \in V_{k-1}, \tag{5.4}$$

$$\|P_k^{k-1} v\|_{1-\tau,k}^2 \le (1+\theta^2)\|v\|_{1-\tau,k}^2 + C_3\theta^{-2}h_k^{2\tau}\|v\|_{1,k}^2 \qquad \forall\, v \in V_k, \tag{5.5}$$

where $\theta \in (0,1)$ is arbitrary, μ and τ are two parameters strictly between 0 and 1, and the positive constants C_1, C_2 and C_3 are independent of the meshes and θ.

Furthermore, we also need the following approximation property which is peculiar to nonconforming methods where the energy norm is not preserved by the coarse-to-fine intergrid transfer operator I_{k-1}^k:

$$\|(Id_{k-1} - P_k^{k-1} I_{k-1}^k) v\|_{1-\mu,k-1} \lesssim h_k^\mu \|v\|_{1,k-1} \qquad \forall\, v \in V_{k-1}. \tag{5.6}$$

Remark 5.1. The estimates (5.3) and (5.6) together imply that (cf. Lemma 4.2 of [22]), for $j \le k$,

$$\|T_{k,j,m} v\|_{1,k} \lesssim \|v\|_{1,j} \qquad \forall\, v \in V_j, \tag{5.7}$$

provided that m is sufficiently large. We can then use (5.4), (5.5) and (5.7) to derive (cf. Lemmas 4.4–4.6 of [22]), for $j \le k$, the following crucial estimate in the additive theory:

$$\|T_{j,k,m} T_{k,j,m} v\|_{1-\tau,j} \lesssim \|v\|_{1-\tau,j} \qquad \forall\, v \in V_j, \tag{5.8}$$

provided m is sufficiently large. The convergence of V-cycle algorithm for sufficiently large m follows from (5.8) and an argument based on a strengthened Cauchy-Schwarz inequality. The convergence of the F-cycle algorithm can then be established by a perturbation argument.

Therefore the heart of our convergence analysis of V-cycle and F-cycle algorithms is the derivation of the estimates (5.3)–(5.6), where there is a lot of freedom in choosing the parameters μ and τ.

We will prove the estimate (5.6) for $\mu = \alpha$ (the index of elliptic regularity in (1.3)) in this section and take up the estimates (5.3)–(5.5) in section 6. The following lemma is a stronger version of (5.6).

LEMMA 5.2. *It holds that*

$$|\!|\!|(Id_{k-1} - P_k^{k-1} I_{k-1}^k)v|\!|\!|_{1-\alpha,k-1} \lesssim h_k^{2\alpha} |\!|\!|v|\!|\!|_{1+\alpha,k-1} \qquad \forall\, v \in V_{k-1}. \tag{5.9}$$

Proof. Let $v \in V_{k-1}$ be arbitrary and define $\phi \in H_0^2(\Omega)$ by

$$\phi(w) = \mathcal{A}_{k-1}(v, J_{k-1} w) \qquad \forall\, w \in H_0^2(\Omega), \tag{5.10}$$

where $J_{k-1} : L_2(\Omega) \longrightarrow V_{k-1}$ is the map in Lemma 3.5. From Lemma 3.9 we have $\phi \in H^{-2+\alpha}(\Omega)$ and

$$\|\phi\|_{H^{-2+\alpha}(\Omega)} \lesssim |\!|\!|v|\!|\!|_{1+\alpha,k-1}. \tag{5.11}$$

Let $\zeta = J_{k-1}^* v$. Again, from Lemma 3.9 we have $\zeta \in H^{2+\alpha}(\Omega) \cap H_0^2(\Omega)$, and

$$\|\zeta - v\|_{H^{2-\alpha}(\Omega)} \lesssim h_{k-1}^{2\alpha} |\!|\!|v|\!|\!|_{1+\alpha,k-1}. \tag{5.12}$$

Finally we define $\zeta_k \in V_k$ to be the solution of the following variational problem:

$$\mathcal{A}_k(\zeta_k, w) = \phi(w) \qquad \forall\, w \in V_k, \tag{5.13}$$

i.e., ζ_k is the solution of the C^0 interior penalty method for (4.5). Therefore we have the following error estimate (cf. Theorem 5 of [24]):

$$\|\zeta - \zeta_k\|_{H^{2-\alpha}(\Omega)} \lesssim h_k^{2\alpha} \|\phi\|_{H^{-2+\alpha}(\Omega)}. \tag{5.14}$$

Moreover, from Corollary 3.8, (5.11) and (5.13) we have

$$\mathcal{A}_k(\zeta_k, w) \le \|\phi\|_{H^{-2+\alpha}(\Omega)} \|w\|_{H^{2-\alpha}(\Omega)} \lesssim |\!|\!|v|\!|\!|_{1+\alpha,k-1} |\!|\!|w|\!|\!|_{1-\alpha,k} \qquad \forall\, w \in V_k,$$

which together with (3.8) implies that

$$|\!|\!|\zeta_k|\!|\!|_{1+\alpha,k} \lesssim |\!|\!|v|\!|\!|_{1+\alpha,k-1}. \tag{5.15}$$

We can now use (2.1), Corollary 3.8, (4.3), (4.13), (5.11), (5.12), (5.14) and (5.15) to complete the proof of the lemma as follows:

$$\begin{aligned}
|\!|\!|(Id_{k-1} - P_k^{k-1} I_{k-1}^k)v|\!|\!|_{1-\alpha,k-1} &\le |\!|\!|v - P_k^{k-1}\zeta_k|\!|\!|_{1-\alpha,k-1} + |\!|\!|P_k^{k-1}(\zeta_k - v)|\!|\!|_{1-\alpha,k-1} \\
&\lesssim \|v - P_k^{k-1}\zeta_k\|_{H^{2-\alpha}(\Omega)} + |\!|\!|\zeta_k - v|\!|\!|_{1-\alpha,k} \\
&\lesssim \|v - \zeta_k\|_{H^{2-\alpha}(\Omega)} + \|\zeta_k - P_k^{k-1}\zeta_k\|_{H^{2-\alpha}(\Omega)} \\
&\lesssim \|v - \zeta\|_{H^{2-\alpha}(\Omega)} + \|\zeta - \zeta_k\|_{H^{2-\alpha}(\Omega)} + |\!|\!|(Id_k - I_{k-1}^k P_k^{k-1})\zeta_k|\!|\!|_{1-\alpha,k} \\
&\lesssim h_k^{2\alpha} |\!|\!|v|\!|\!|_{1+\alpha,k-1} + h_k^{2\alpha} |\!|\!|\zeta_k|\!|\!|_{1+\alpha,k} \lesssim h_k^{2\alpha} |\!|\!|v|\!|\!|_{1+\alpha,k-1}. \qquad \square
\end{aligned}$$

The following corollary is an immediate consequence of (3.7) and (5.9).

COROLLARY 5.3. *The estimate* (5.6) *holds for* $\mu = \alpha$.

Finally we prove a useful relation between the mesh-dependent norm $|\!|\!| \cdot |\!|\!|_{0,k}$ and Sobolev norms that will be used in the derivation of (5.3)–(5.5). We will use C in the proof of the following lemma (and others in section 6) to denote a generic mesh-independent positive constant that can take different values at different occurrences.

LEMMA 5.4. *It holds that*

$$|\!|\!| v |\!|\!|_{0,k}^2 \le (1+\theta^2)|v|_{H^1(\Omega)}^2 + C_4\theta^{-2}h_k^{2\beta}\|v\|_{H^{1+\beta}(\Omega)}^2 \qquad \forall\, v \in V_k,\ 0<\theta<1, \tag{5.16}$$

where β *is the number in* (2.16) *and the positive constant* C_4 *is mesh-independent.*

Proof. Let $\theta \in (0,1)$ and $v \in V_k$ be arbitrary. From (1.7), (2.12), (2.16), (3.5) and Corollary 3.8, we have

$$\begin{aligned}
|\!|\!| v |\!|\!|_{0,k}^2 &= \langle B_k v, v\rangle \\
&= \langle L_k v, v\rangle + \langle B_k(Id_k - B_k^{-1}L_k)v, v\rangle \\
&\le |v|_{H^1(\Omega)}^2 + |\!|\!|(Id_k - B_k^{-1}L_k)v|\!|\!|_{0,k}|\!|\!| v |\!|\!|_{0,k} \\
&\le |v|_{H^1(\Omega)}^2 + \theta^2 |\!|\!| v |\!|\!|_{0,k}^2 + C\theta^{-2}h_k^{2\beta}\|v\|_{H^{1+\beta}(\Omega)}^2 \\
&\le (1+C\theta^2)|v|_{H^1(\Omega)}^2 + C\theta^{-2}h_k^{2\beta}\|v\|_{H^{1+\beta}(\Omega)}^2,
\end{aligned}$$

which is equivalent to (5.16) because θ is arbitrary. ∎

6. Results for *V*-cycle and *F*-cycle algorithms. In this section we will complete the convergence analysis of V-cycle and F-cycle algorithms by deriving the estimates (5.3)–(5.5). We shall take the parameter μ in (5.3) to be α and the parameter τ in (5.4)–(5.5) to be the number β that appears in (2.16).

First we prove a stronger version of (5.3).

LEMMA 6.1. *There exists a positive constant* C_1 *independent of the meshes such that*

$$|\!|\!| I_{k-1}^k v |\!|\!|_{1,k}^2 \le |\!|\!| v |\!|\!|_{k-1}^2 + C_1 h_k^{2\alpha} |\!|\!| v |\!|\!|_{1+\alpha,k-1}^2 \qquad \forall\, v \in V_k,\ k \ge 1. \tag{6.1}$$

Proof. Let $v \in V_{k-1}$ and $\theta \in (0,1)$ be arbitrary, and $\zeta = J_{k-1}^* v$. From (1.5) and (3.6) we have

$$\begin{aligned}
|\!|\!| I_{k-1}^k v |\!|\!|_{1,k}^2 = \mathcal{A}_k(v,v) &= \mathcal{A}_{k-1}(v,v) + \eta \sum_{e\in\mathcal{E}_{k-1}} |e|^{-1}\|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \\
&= |\!|\!| v |\!|\!|_{1,k-1}^2 + \eta \sum_{e\in\mathcal{E}_{k-1}} |e|^{-1}\|[\![(\partial v/\partial n) - (\partial\zeta/\partial n)]\!]\|_{L_2(e)}^2.
\end{aligned} \tag{6.2}$$

Moreover, we have, from (2.5),

$$\eta \sum_{e\in\mathcal{E}_{k-1}} |e|^{-1}\|[\![(\partial v/\partial n) - (\partial\zeta/\partial n)]\!]\|_{L_2(e)}^2 \lesssim \|\zeta - v\|_{k-1}^2. \tag{6.3}$$

The estimate (6.1) follows from (6.2), (6.3) and Lemma 3.9. ∎

Since $0 < \beta < 1/2 < \alpha$, the estimates (3.7) and (6.1) imply the following corollary.

COROLLARY 6.2. *It holds that*

$$|\!|\!| I_{k-1}^k v |\!|\!|_{1,k}^2 \le |\!|\!| v |\!|\!|_{1,k-1}^2 + C_1'\theta^{-2}h_k^{2\beta} |\!|\!| v |\!|\!|_{1+\beta,k-1}^2 \qquad \forall\, v \in V_{k-1}, \tag{6.4}$$

where the positive constant C_1' *is mesh-independent.*

LEMMA 6.3. *The estimate* (5.4) *holds for* $\tau = \beta$.

Proof. Let $\theta \in (0,1)$ be arbitrary. From (2.15), (3.5), Corollary 3.8 and (5.16) we have

$$|||I_{k-1}^k v|||_{0,k}^2 \leq (1+\theta^2)|||v|||_{0,k-1}^2 + C_4'\theta^{-2}h_k^{2\beta}|||v|||_{\beta,k-1}^2 \qquad \forall\, v \in V_{k-1}, \tag{6.5}$$

where the positive constant C_4' is mesh-independent.

Let the inner product $(\!(\cdot,\cdot)\!)_{k-1,\theta}$ on V_{k-1} be defined by

$$(\!(v_1,v_2)\!)_{k-1,\theta} = (1+\theta^2)(v_1,v_2)_{k-1} + C_2\theta^{-2}h_k^{2\beta}(\mathbb{A}_{k-1}^{\beta}v_1,v_2)_{k-1} \tag{6.6}$$

for all $v_1, v_2 \in V_{k-1}$, where $C_2 = \max(C_1', C_4')$ is the maximum of the mesh-independent constants in (6.4) and (6.5). Note that $\mathbb{A}_{k-1}$ is symmetric positive definite with respect to the inner product $(\!(\cdot,\cdot)\!)_{k-1,\theta}$.

In view of (3.4) and (6.6), the estimates (6.4) and (6.5) imply

$$|||I_{k-1}^k v|||_{0,k}^2 \leq (\!(\mathbb{A}_k^0 v, v)\!)_{k-1,\theta} \qquad \forall\, v \in V_{k-1}, \tag{6.7}$$

$$|||I_{k-1}^k v|||_{1,k}^2 \leq (\!(\mathbb{A}_k^1 v, v)\!)_{k-1,\theta} \qquad \forall\, v \in V_{k-1}. \tag{6.8}$$

It follows from (3.4), (6.6)–(6.8) and interpolation between Hilbert scales that

$$\begin{aligned}|||I_{k-1}^k v|||_{1-\beta,k}^2 \leq (\!(\mathbb{A}_k^{1-\beta}v,v)\!)_{k-1,\theta}^2 &= (1+\theta^2)|||v|||_{1-\beta,k}^2 \\ &\quad + C_2\theta^{-2}h_k^{2\beta}|||v|||_{1,k-1}^2 \quad \forall\, v \in V_{k-1}. \qquad \square\end{aligned}$$

We now turn to the estimate (5.5). First we have to establish certain two-level estimates for the nodal interpolation operator with respect to the mesh-dependent norms.

LEMMA 6.4. *The following estimate holds*:

$$|||\Pi_{k-1}v|||_{1,k-1}^2 \leq (1+\theta^2)|||v|||_{1,k}^2 + C_\sharp h_k^{2\alpha}|||v|||_{1+\alpha,k}^2 \qquad \forall\, v \in V_k,\ \theta \in (0,1), \tag{6.9}$$

where the constant $C_\sharp$ *is mesh-independent.*

Proof. Let $v \in V_k$ and $\theta \in (0,1)$ be arbitrary. It follows from (1.8), (3.6), (3.18), and (4.3) that

$$\begin{aligned}|||\Pi_{k-1}v|||_{1,k-1}^2 &\leq (1+\theta^2)|||P_k^{k-1}v|||_{1,k-1}^2 + C\theta^{-2}|||\Pi_{k-1}(v-P_k^{k-1}v)|||_{1,k-1}^2 \\ &\leq (1+\theta^2)|||P_k^{k-1}v|||_{1,k-1}^2 + C\theta^{-2}|||v-P_k^{k-1}v|||_{1,k}^2 \\ &\leq (1+\theta^2)^2|||v|||_{1,k-1}^2 + C\theta^{-2}|||v-P_k^{k-1}v|||_{1,k}^2 \\ &\leq (1+\theta^2)^2|||v|||_{1,k-1}^2 + C\theta^{-2}h_k^{2\alpha}|||v|||_{1+\alpha,k}^2,\end{aligned}$$

which implies (6.9) because $\theta \in (0,1)$ is arbitrary. $\square$

Again, since $0 < \beta < 1/2 < \alpha$, the estimates (3.7) and (6.9) imply the following corollary.

COROLLARY 6.5. *It holds that*

$$|||\Pi_{k-1}v|||_{1,k-1}^2 \leq (1+\theta^2)|||v|||_{1,k}^2 + C_\sharp' h_k^{2\beta}|||v|||_{1+\beta,k}^2 \qquad \forall\, v \in V_k,\ \theta \in (0,1), \tag{6.10}$$

where the positive constant $C_\sharp'$ *is mesh-independent.*

LEMMA 6.6. *The following estimate holds*:

$$|||\Pi_{k-1}v|||_{0,k-1}^2 \leq (1+\theta^2)|||v|||_{0,k}^2 + C_\flat h_k^{2\beta}|||v|||_{\beta,k}^2 \qquad \forall\, v \in V_k,\ \theta \in (0,1), \tag{6.11}$$

where the positive constant $C_\flat$ *is mesh-independent.*

Proof. Let $\theta \in (0,1)$ and $v \in V_k$ be arbitrary. First we observe that, by (2.1), (3.6), (3.7), (3.16), (3.21) and Corollary 3.8,

$$\begin{aligned}
|\!|\!|\Pi_{k-1}v|\!|\!|_{\beta,k-1} &\lesssim \|v\|_{H^{1+\beta}(\Omega)} + \|v-\Pi_{k-1}v\|_{H^{1+\beta}(\Omega)} \\
&\lesssim |\!|\!|v|\!|\!|_{\beta,k} + h_k^{-\beta}|v-\Pi_{k-1}v|_{H^1(\Omega)} \\
&\lesssim |\!|\!|v|\!|\!|_{\beta,k} + h_k^{1-\beta}|\!|\!|v|\!|\!|_{1,k} \lesssim |\!|\!|v|\!|\!|_{\beta,k}.
\end{aligned} \tag{6.12}$$

The estimate (6.11) follows from (1.8), (2.1), (2.12), (2.15), (3.5), (3.7), (3.16), (5.16) and (6.12):

$$\begin{aligned}
|\!|\!|\Pi_{k-1}v|\!|\!|_{0,k-1}^2 &\le (1+\theta^2)|\Pi_{k-1}v|_{H^1(\Omega)}^2 + C\theta^{-2}h_{k-1}^{2\beta}|\!|\!|\Pi_{k-1}v|\!|\!|_{\beta,k-1}^2 \\
&\le (1+\theta^2)\big(|v|_{H^1(\Omega)} + |\Pi_{k-1}v - v|_{H^1(\Omega)}\big)^2 + C\theta^{-2}h_k^{2\beta}|\!|\!|v|\!|\!|_{\beta,k}^2 \\
&\le (1+\theta^2)^2|v|_{H^1(\Omega)}^2 + C\theta^{-2}h_k^2\|v\|_{\mathcal{A}_k}^2 + C\theta^{-2}h_k^{2\beta}|\!|\!|v|\!|\!|_{\beta,k}^2 \\
&\le (1+\theta^2)^2|\!|\!|v|\!|\!|_{0,k}^2 + C\theta^{-2}h_k^2|\!|\!|v|\!|\!|_{1,k}^2 + C\theta^{-2}h_k^{2\beta}|\!|\!|v|\!|\!|_{\beta,k}^2 \\
&\le (1+\theta^2)^2|\!|\!|v|\!|\!|_{0,k}^2 + C\theta^{-2}h_k^{2\beta}|\!|\!|v|\!|\!|_{\beta,k}^2,
\end{aligned}$$

which is equivalent to (6.11) because $\theta \in (0,1)$ is arbitrary. ∎

COROLLARY 6.7. *The following estimate holds*:

$$|\!|\!|\Pi_{k-1}v|\!|\!|_{1-\beta,k-1}^2 \le (1+\theta^2)|\!|\!|v|\!|\!|_{1-\beta,k}^2 + C_\natural h_k^{2\beta}|\!|\!|v|\!|\!|_{1,k}^2 \qquad \forall\, v \in V_k,\ \theta \in (0,1), \tag{6.13}$$

where the constant $C_\natural$ is mesh-independent.

Proof. We use the technique in the proof of Lemma 6.3. For any $\theta \in (0,1)$, we define the inner product $(\!(\cdot,\cdot)\!)_{k,\theta}$ on V_k by

$$(\!(v_1,v_2)\!)_{k,\theta}^2 = (1+\theta^2)(v_1,v_2)_k^2 + C_\natural\theta^{-2}h_k^{2\beta}(\mathbb{A}_k^\beta v_1, v_2)_k \tag{6.14}$$

for all $v_1, v_2 \in V_k$, where $C_\natural = \max(C_\sharp', C_\flat)$. Then $\mathbb{A}_k$ is symmetric positive definite with respect to $(\!(\cdot,\cdot)\!)_{k,\theta}$.

In view of (3.4), (6.10), (6.11) and (6.14), we have

$$|\!|\!|\Pi_{k-1}v|\!|\!|_{0,k-1}^2 \le (\!(\mathbb{A}_k^0 v, v)\!)_{k,\theta} \qquad \forall\, v \in V_k, \tag{6.15}$$

$$|\!|\!|\Pi_{k-1}v|\!|\!|_{1,k-1}^2 \le (\!(\mathbb{A}_k^1 v, v)\!)_{k,\theta} \qquad \forall\, v \in V_k. \tag{6.16}$$

The estimate (6.13) follows from (6.15), (6.16) and interpolation between Hilbert scales. ∎

We are now ready to verify (5.5).

LEMMA 6.8. *The estimate* (5.5) *holds for $\tau = \beta$.*

Proof. Let $v \in V_k$ and $\theta \in (0,1)$ be arbitrary. From (1.8), (3.6), (3.7), (3.17), Corollary 3.8, (4.3) and (6.13), we have

$$\begin{aligned}
|\!|\!|P_k^{k-1}v|\!|\!|_{1-\beta,k-1}^2 &\le \big(|\!|\!|\Pi_{k-1}v|\!|\!|_{1-\beta,k-1} + |\!|\!|P_k^{k-1}v - \Pi_{k-1}v|\!|\!|_{1-\beta,k-1}\big)^2 \\
&\le (1+\theta^2)|\!|\!|\Pi_{k-1}v|\!|\!|_{1-\beta,k-1}^2 + C\theta^{-2}h_k^{2(\beta-\alpha)}|\!|\!|P_k^{k-1}v - \Pi_{k-1}v|\!|\!|_{1-\alpha,k-1}^2 \\
&\le (1+\theta^2)|\!|\!|\Pi_{k-1}v|\!|\!|_{1-\beta,k-1}^2 + C\theta^{-2}h_k^{2(\beta-\alpha)}\|P_k^{k-1}v - \Pi_{k-1}v\|_{H^{2-\alpha}(\Omega)}^2 \\
&\le (1+\theta^2)^2|\!|\!|v|\!|\!|_{1-\beta,k}^2 + C\theta^{-2}|\!|\!|v - \Pi_{k-1}v|\!|\!|_{1-\beta,k-1}^2 \\
&\quad + C\theta^{-2}h_k^{2(\beta-\alpha)}\big(\|P_k^{k-1}v - v\|_{H^{2-\alpha}(\Omega)} + \|v-\Pi_{k-1}v\|_{H^{2-\alpha}(\Omega)}\big)^2 \\
&\le (1+\theta^2)^2|\!|\!|v|\!|\!|_{1-\beta,k}^2 + C\theta^{-2}h_k^{2\beta}|\!|\!|v|\!|\!|_{1,k}^2,
\end{aligned}$$

which is equivalent to (5.5) because $\theta \in (0,1)$ is arbitrary. ∎

We have verified the assumptions (5.3)–(5.6) for the additive theory. Therefore we can apply the results in [22] to obtain the following convergence theorems for the V-cycle and F-cycle algorithms.

THEOREM 6.9. *The output $MG_V(k,\psi,z_0,m)$ of the V-cycle algorithm (Algorithm 2.3) applied to (2.10) satisfies the following estimate*:

$$\|z - MG_V(k,\psi,z_0,m)\|_{\mathcal{A}_k} \leq \frac{C}{m^\alpha}\|z - z_0\|_{\mathcal{A}_k},$$

where the positive constant C is mesh-independent, provided that the number of smoothing steps m is greater than a positive integer m_* that is also mesh-independent.

THEOREM 6.10. *The output $MG_F(k,\psi,z_0,m)$ of the F-cycle algorithm (Algorithm 2.5) applied to (2.10) satisfies the following estimate*:

$$\|z - MG_F(k,\psi,z_0,m)\|_{\mathcal{A}_k} \leq \frac{C}{m^\alpha}\|z - z_0\|_{\mathcal{A}_k},$$

where the positive constant C is mesh-independent, provided that the number of smoothing steps m is greater than a positive integer m_* that is also mesh-independent.

Remark 6.11. Theorems 6.9 and 6.10 have been obtained for preconditioners that satisfy (2.13)–(2.16). Therefore they are valid for a Poisson solve B_k^{-1} obtained by a symmetric W-cycle algorithm with a sufficiently large number of smoothing steps or a variable V-cycle algorithm (cf. Remark 2.2 and Appendix A). However, in practice these algorithms behave equally well when the preconditioner is a symmetric V-cycle algorithm with a few smoothing steps (cf. section 7).

7. Numerical experiments. In this section we report the results of some numerical experiments for the biharmonic problem. The finite element we use is the Q_2 rectangular element and the penalty parameter η is taken to be 5.

The first set of experiments involve the biharmonic problem on the unit square, where we can take the index of elliptic regularity α to be 1. The initial triangulation $\mathcal{T}_0$ consists of one element and we compute the contraction numbers of the V-cycle, F-cycle and W-cycle algorithms on the kth level ($1 \leq k \leq 7$) with m presmoothing and m postsmoothing steps. We use the symmetric V-cycle algorithm for the Poisson problem with three presmoothing and three postsmoothing Richardson relaxation steps as the preconditioner in (2.17) and (2.19). The results are recorded in Tables 7.1–7.3. Convergence for the V-cycle, F-cycle and W-cycle algorithms is observed for $m = 5$, $m = 2$ and $m = 1$, respectively. We also observe that the performance of the F-cycle algorithm and the W-cycle algorithm are almost identical for $m \geq 6$.

Numerical experiments show that for moderate grid levels ($k \leq 7$) there is practically no difference in the performance of the multigrid algorithms whether we use a symmetric V-cycle or a symmetric W-cycle Poisson solve as the preconditioner in (2.17) and (2.19).

We also plot the contraction numbers versus the number m of smoothing steps for the 7th level V-cycle and W-cycle algorithms, where m ranges from 20 to 40. Figure 7.1 contains the resulting log-log plot. The asymptotic rate of decrease is observed to be m^{-1}, which agrees with Theorems 4.4 and 6.9.

For comparison we report in Table 7.4 the contraction numbers of the V-cycle algorithm using the Richardson relaxation scheme without a preconditioner as the smoother. Convergence is observed only for $m \geq 75$.

TABLE 7.1
Contraction numbers for the V-cycle algorithm on the unit square.

k \ m	5	6	7	8	9	10
1	0.04	0.02	0.011	0.006	0.0032	0.0017
2	0.22	0.18	0.15	0.13	0.11	0.09
3	0.32	0.29	0.26	0.23	0.21	0.19
4	0.35	0.34	0.31	0.28	0.25	0.23
5	0.42	0.37	0.34	0.31	0.29	0.27
6	0.43	0.39	0.35	0.33	0.30	0.27
7	0.44	0.39	0.36	0.34	0.31	0.29

TABLE 7.2
Contraction numbers for the F-cycle algorithm on the unit square.

k \ m	2	3	4	5	6	7	8	9	10
1	0.28	0.15	0.08	0.04	0.02	0.01	0.0060	0.0032	0.0017
2	0.50	0.35	0.27	0.22	0.18	0.15	0.13	0.11	0.09
3	0.52	0.40	0.34	0.30	0.27	0.24	0.22	0.19	0.18
4	0.53	0.42	0.37	0.34	0.31	0.28	0.26	0.24	0.22
5	0.53	0.43	0.37	0.34	0.31	0.29	0.27	0.25	0.23
6	0.53	0.44	0.38	0.34	0.32	0.29	0.27	0.25	0.23
7	0.54	0.46	0.38	0.35	0.32	0.29	0.27	0.25	0.23

TABLE 7.3
Contraction numbers for the W-cycle algorithm on the unit square.

k \ m	1	2	3	4	5	6	7	8	9	10
1	0.53	0.28	0.15	0.08	0.04	0.02	0.01	0.006	0.003	0.002
2	0.72	0.49	0.24	0.27	0.22	0.18	0.15	0.13	0.11	0.09
3	0.71	0.51	0.40	0.34	0.30	0.26	0.24	0.22	0.19	0.17
4	0.80	0.51	0.41	0.37	0.34	0.31	0.28	0.26	0.24	0.22
5	0.76	0.53	0.42	0.38	0.34	0.31	0.29	0.26	0.24	0.23
6	0.82	0.53	0.42	0.38	0.34	0.32	0.29	0.26	0.25	0.22
7	0.83	0.53	0.42	0.38	0.34	0.32	0.29	0.27	0.25	0.23

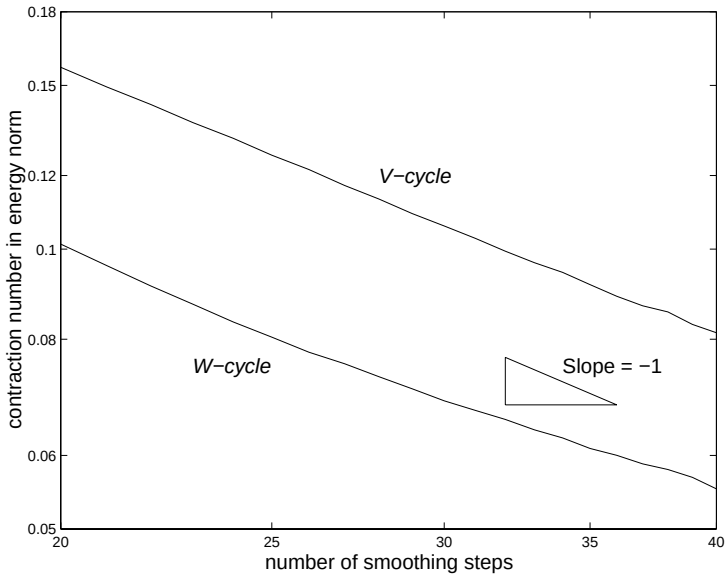


FIG. 7.1. *Asymptotic rate of decrease for the contraction numbers of the 7th level V-cycle and W-cycle algorithms.*

TABLE 7.4
Contraction numbers for the V-cycle algorithm on the unit square without a preconditioner in the smoothing steps.

k \ m	75	76	77	78	79	80	81	82	83
1	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05
2	0.47	0.47	0.46	0.46	0.46	0.46	0.46	0.45	0.45
3	0.64	0.47	0.42	0.64	0.42	0.40	0.41	0.36	0.63
4	0.60	0.60	0.58	0.57	0.54	0.52	0.50	0.51	0.49
5	0.71	0.69	0.66	0.64	0.63	0.61	0.57	0.56	0.52
6	0.76	0.74	0.72	0.70	0.68	0.65	0.62	0.60	0.56
7	0.80	0.78	0.76	0.73	0.71	0.68	0.65	0.61	0.56

In the second set of experiments we study the biharmonic problem on the L-shaped domain with vertices $(-1,-1)$, $(1,-1)$, $(1,0)$, $(0,0)$, $(0,1)$ and $(-1,1)$. In this case we can take the index of elliptic regularity α to be any number $< \alpha_* = 0.5444837368\ldots$. The initial triangulation $\mathcal{T}_0$ consists of three elements. Again we use the symmetric V-cycle algorithm for the Poisson problem with three presmoothing and three postsmoothing steps as the preconditioner. The behavior of the contraction numbers of the V-cycle, F-cycle and W-cycle algorithms are similar to those observed for the unit square.

Here we only report the results for the V-cycle algorithm in Table 7.5. From the results in Tables 7.1 and 7.5 we see that the performance on the L-shaped domain is

TABLE 7.5
Contraction numbers for the V-cycle algorithm on the L-shaped domain.

k \ m	5	6	7	8	9	10
1	0.16	0.12	0.082	0.059	0.043	0.031
2	0.27	0.21	0.19	0.16	0.14	0.13
3	0.37	0.31	0.29	0.26	0.24	0.21
4	0.42	0.37	0.34	0.31	0.28	0.26
5	0.45	0.38	0.35	0.32	0.30	0.27
6	0.46	0.40	0.36	0.34	0.31	0.28
7	0.46	0.40	0.37	0.34	0.31	0.29

slightly worse than the performance on the unit square. But it is much better than the rate $m^{-\alpha}$ predicted by Theorem 6.9. This is likely the effect of superconvergence due to the uniform grids used in our computations.

Appendix A. Some properties of multigrid Poisson solves. In this appendix we consider multigrid Poisson solves as the preconditioner in (2.17) and (2.19). We will show that properties (i)–(iv) in section 2 are satisfied by such preconditioners.

Consider the discrete Poisson problem: Find $z \in V_k$ such that

$$L_k z = \psi \qquad \forall\, v \in V_k, \tag{A.1}$$

where L_k is defined in (2.12) and $\psi \in V_k'$.

Let $S_k : V_k' \longrightarrow V_k$ be the solution operator for (A.1) generated by either a symmetric V-cycle algorithm, a symmetric W-cycle algorithm or a symmetric variable V-cycle algorithm (that satisfies (4.12)), with 0 as the initial guess and Richardson relaxation as the smoother. In terms of S_k the output $MG(k, \psi, z_0)$ of the multigrid method can be written as

$$MG(k, \psi, z_0) = z_0 + S_k(\psi - L_k z_0), \tag{A.2}$$

and $Id_k - S_k L_k$ is the error propagation operator.

The operator S_k is symmetric, or equivalently the operator $Id_k - S_k L_k$ is symmetric with respect to the bilinear form $\langle L_k \cdot, \cdot \rangle$ (cf. Lemma 7.1 of [17], where S_k is denoted by B_k). Furthermore (cf. Theorems 5.1, 7.1 and 7.2 of [17]) there exists a number $\delta \in (0, 1)$ independent of k such that

$$0 \le \langle L_k(Id_k - S_k L_k)v, v\rangle \le \delta \langle L_k v, v\rangle \qquad \forall\, v \in V_k. \tag{A.3}$$

We see from (A.3) that S_k is positive definite. Therefore we can define $B_k = S_k^{-1}$ and the operator $B_k : V_k \longrightarrow V_k'$ is symmetric positive definite. Moreover (A.3) implies that the eigenvalues of the operator $Id_k - B_k^{-1} L_k : V_k \longrightarrow V_k$ lie between 0 and δ. Since $Id_k - B_k^{-1} L_k$ is also symmetric with respect to $\langle B_k \cdot, \cdot \rangle$, we deduce that

$$0 \le \langle B_k(Id_k - B_k^{-1} L_k)v, v\rangle \le \delta \langle B_k v, v\rangle \qquad \forall\, v \in V_k. \tag{A.4}$$

The estimate (2.15) follows from (A.4) immediately.

Hence the operator B_k satisfies properties (i) and (ii) in section 2. Property (iv), which states that multigrid algorithms have optimal complexity, is also standard [32]. In particular, Theorem 4.4 and Theorem 4.5 are valid for all three types of multigrid preconditioners.

On the other hand, the proofs of Theorem 6.9 and Theorem 6.10 require property (iii). Below we will demonstrate that (2.16) is satisfied by the B_k generated by W-cycle or variable V-cycle Poisson solves.

Let $B_k^{-1} : V_k' \longrightarrow V_k$ be the preconditioner obtained by a symmetric W-cycle algorithm with m presmoothing and m postsmoothing steps. We have a well-known recurrence relation [32]:

$$\begin{aligned} Id_k - B_k^{-1}L_k =& R_k^m(Id_k - I_{k-1}^k P_k^{k-1})R_k^m \\ &+ R_k^m I_{k-1}^k (Id_{k-1} - B_{k-1}^{-1}L_{k-1})^2 P_k^{k-1} R_k^m, \end{aligned} \tag{A.5}$$

where $I_{k-1}^k : V_{k-1} \longrightarrow V_k$ is the natural injection, $P_k^{k-1} : V_k \longrightarrow V_{k-1}$ is the adjoint of I_k^{k-1} with respect to the bilinear form $\langle L_k \cdot, \cdot \rangle$ and $\langle L_{k-1} \cdot, \cdot \rangle$, and R_k is the error reduction operator of one Richardson relaxation step. Of course at the coarsest level we have $B_0^{-1} = S_0 = L_0^{-1}$ and hence

$$Id_0 - B_0^{-1}L_0 = 0. \tag{A.6}$$

Let β be any number in $(0, 1/2)$. The following estimates are valid [20] for $k \geq 1$:

$$|R_k^m(Id_k - I_{k-1}^k P_k^{k-1})R_k^m v|_{H^1(\Omega)} \lesssim h_k^\beta m^{-\beta/2} \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall v \in V_k, \tag{A.7}$$

$$\|P_k^{k-1} v\|_{H^{1+\beta}(\Omega)} \lesssim \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_k, \tag{A.8}$$

$$|(Id_k - B_k^{-1}L_k)v|_{H^1(\Omega)} \lesssim m^{-\alpha_*} |v|_{H^1(\Omega)} \qquad \forall\, v \in V_k, \tag{A.9}$$

where $\alpha_* \in (1/2, 1]$ is the index of elliptic regularity for the Poisson problem. Furthermore, we have

$$|R_k^m v|_{H^1(\Omega)} \leq |v|_{H^1(\Omega)} \qquad \forall\, v \in V_k, \text{ and } m \geq 1, \tag{A.10}$$

$$\|R_k^m v\|_{H^{1+\beta}(\Omega)} \leq C \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_k, \text{ and } m \geq 1, \tag{A.11}$$

where the positive constant C is independent of the meshes.

It follows from (2.1), (A.5), and (A.7)–(A.11) that

$$|v - B_k^{-1}L_k v|_{H^1(\Omega)} \leq C_* h_k^\beta \big[m^{-\beta/2} + \sigma m^{-\alpha_*}\big] \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_k, \tag{A.12}$$

where C_* is a mesh-independent positive constant, provided that

$$|v - B_{k-1}^{-1}L_{k-1} v|_{H^1(\Omega)} \leq \sigma h_{k-1}^\beta \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_{k-1}.$$

Hence, if m is sufficiently large, we obtain from (A.6), (A.12) and mathematical induction that

$$|v - B_k^{-1}L_k v|_{H^1(\Omega)} \leq \sigma h_k^\beta \|v\|_{H^{1+\beta}(\Omega)} \qquad \forall\, v \in V_k,\ k \geq 0, \tag{A.13}$$

if σ is the number defined by $\sigma = C_* m^{-\beta/2}/(1 - C_* m^{-\alpha_*})$. Therefore (2.16) is satisfied by the W-cycle preconditioner provided that m is sufficiently large.

Now we consider the preconditioner B_k^{-1} obtained from a variable V-cycle algorithm. Given a positive integer k, we assume that the number m_j of smoothing steps on level j satisfies

$$(1+\epsilon)m_{j+1} \le m_j \qquad \text{for} \quad 0 \le j \le k-1, \tag{A.14}$$

where ϵ is a positive number. We have an additive expression for the error propagation operator:

$$\begin{aligned} Id_k - B_k^{-1}L_k = {} & R_k^{m_k}(Id_k - I_{k-1}^k P_k^{k-1})R_k^{m_k} \\ & + R_k^{m_k} I_{k-1}^k R_{k-1}^{m_{k-1}}(Id_{k-1} - I_{k-2}^{k-1}P_{k-1}^{k-2})R_{k-1}^{m_{k-1}} P_k^{k-1} R_k^{m_k} \\ & + R_k^{m_k} I_{k-1}^k R_{k-1}^{m_{k-1}} I_{k-2}^{k-1} \\ & \times R_{k-2}^{m_{k-2}}(Id_{k-2} - I_{k-3}^{k-2}P_{k-2}^{k-3})R_{k-2}^{m_{k-2}} P_{k-1}^{k-2} R_{k-1}^{m_{k-1}} P_k^{k-1} R_k^{m_k} \\ & + \cdots . \end{aligned} \tag{A.15}$$

The following estimates are valid [20] for $k \ge 1$, $j \le k$ and $v \in V_k$:

$$\|R_k^{m_k}(Id_k - I_{k-1}^k P_k^{k-1})R_k^{m_k} v\|_{H^{1-\beta}(\Omega)} \lesssim h_k^\beta m_k^{-\alpha_* + (\beta/2)} |v|_{H^1(\Omega)}, \tag{A.16}$$

$$\|R_k^{m_k} I_{k-1}^k \cdots R_{j+1}^{m_{j+1}} I_j^{j+1} v\|_{H^{1-\beta}(\Omega)} \lesssim \|v\|_{H^{1-\beta}(\Omega)}, \tag{A.17}$$

$$|P_{j+1}^j R_{j+1}^{m_{j+1}} \cdots P_k^{k-1} R_k^{m_k} v|_{H^1(\Omega)} \lesssim |v|_{H^1(\Omega)}. \tag{A.18}$$

Combining (2.1), (A.6) and (A.15)–(A.18), we find, for any $\beta \in (0, 1/2)$,

$$\begin{aligned} |v - B_k^{-1}L_k v|_{H^{1-\beta}(\Omega)} &\le C_\beta |v|_{H^1(\Omega)} \sum_{j=2}^k h_j^\beta m_j^{-\alpha_* + (\beta/2)} \\ &\le C_\beta |v|_{H^1(\Omega)} \sum_{j=2}^k (2^{(j-k)} h_k)^\beta ((1+\epsilon)^{(j-k)} m_k)^{-\alpha_* + (\beta/2)} \\ &\le C_\beta m_k^{-\alpha_* + (\beta/2)} h_k^\beta |v|_{H^1(\Omega)} \sum_{j=2}^k [2^\beta (1+\epsilon)^{(-\alpha_* + \beta/2)}]^{j-k}, \end{aligned} \tag{A.19}$$

where C_β depends on β but not the meshes. It follows from (A.19) that there exists a positive mesh-independent constant C such that

$$|v - B_k^{-1}L_k v|_{H^{1-\beta}(\Omega)} \le C h_k^\beta |v|_{H^1(\Omega)} \qquad \forall\, v \in V_k \tag{A.20}$$

if we choose $\beta > 0$ so that

$$\beta < \max\left(\frac{1}{2}, \frac{\alpha_* \ln(1+\epsilon)}{\ln(2\sqrt{1+\epsilon})}\right).$$

The estimate (2.16) follows from (A.20) and duality. In other words, property (iii) is satisfied by the variable V-cycle preconditioner under condition (A.14).

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, 2nd ed., Academic Press, Amsterdam, 2003.

[2] J. H. Argyris, I. Fried, and D. W. Scharpf, *The TUBA family of plate elements for the matrix displacement method*, Aero. J. Roy. Aero. Soc., 72 (1968), pp. 701–709.

[3] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[4] C. Bacuta, J. H. Bramble, and J. E. Pasciak, *Shift theorems for the biharmonic Dirichlet problem*, in Recent Progress in Computational and Applied PDEs, T. Chan, et al., eds., Kluwer/Plenum, New York, 2002, pp. 1–26.

[5] G. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–89.

[6] R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal., 22 (1985), pp. 617–633.

[7] R. E. Bank and T. F. Dupont, *An optimal order process for solving finite element equations*, Math. Comp., 36 (1981), pp. 35–51.

[8] F. Ben Belgacem and S. C. Brenner, *Some nonstandard finite element estimates with applications to 3D Poisson and Signorini problems*, Electron. Trans. Numer. Anal., 12 (2001), pp. 134–148.

[9] F. K. Bogner, R. L. Fox, and L. A. Schmit, *The generation of interelement compatible stiffness and mass matrices by the use of interpolation formulas*, in Proceedings Conference on Matrix Methods in Structural Mechanics, Wright Patterson A.F.B., Dayton, OH, 1965, pp. 397–444.

[10] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967–975.

[11] J. H. Bramble, *Multigrid Methods*, Longman Scientific & Technical, Essex, 1993.

[12] J. H. Bramble and J. E. Pasciak, *New estimates for multigrid algorithms including the V-cycle*, Math. Comp., 60 (1993), pp. 447–471.

[13] J. H. Bramble and J. E. Pasciak, *Uniform convergence estimates for multigrid V-cycle algorithms with less than full elliptic regularity*, in Domain Decomposition Methods in Science and Engineering, Contemp. Math. 157, A. Quarteroni et. al., eds., Amer. Math. Soc., Providence, 1994, Contemporary Mathematics 157, pp. 17–26.

[14] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition and multigrid*, Math. Comp., 57 (1991), pp. 1–21.

[15] J. H. Bramble, J. E. Pasciak, and J. Xu, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp., 56 (1991), pp. 1–34.

[16] J. H. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.

[17] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook of Numerical Analysis, Vol. 2, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 173–415.

[18] S. C. Brenner, *Two-level additive Schwarz preconditioners for nonconforming finite element methods*, Math. Comp., 65 (1996), pp. 897–921.

[19] S. C. Brenner, *Convergence of nonconforming multigrid methods without full elliptic regularity*, Math. Comp., 68 (1999), pp. 25–53.

[20] S. C. Brenner, *Convergence of the multigrid V-cycle algorithm for second order boundary value problems without full elliptic regularity*, Math. Comp., 71 (2002), pp. 507–525.

[21] S. C. Brenner, *Smoothers, mesh dependent norms, interpolation and multigrid*, Appl. Numer. Math., 43 (2002), pp. 45–56.

[22] S. C. Brenner, *Convergence of nonconforming V-cycle and F-cycle multigrid algorithms for second order elliptic boundary value problem*, Math. Comp., 73 (2004), pp. 1041–1066.

[23] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York-Berlin-Heidelberg, 2002.

[24] S. C. Brenner and L.-Y. Sung, *C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains*, J. Sci. Comput., 22/23 (2005), pp. 83–118.

[25] S. C. Brenner and J. Zhao, *Convergence of multigrid algorithms for interior penalty methods*, Appl. Numer. Anal. Comput. Math., 2 (2004), pp. 3–18.

[26] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York-Berlin-Heidelberg, 1991.

[27] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[28] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Mathematics 1341, Springer-Verlag, Berlin-Heidelberg, 1988.
[29] G. Engel, K. Garikipati, T. J. R. Hughes, M. G. Larson, L. Mazzei, and R. L. Taylor, *Continuous/discontinuous finite element approximations of fourth order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3669–3750.
[30] N. A. Fleck and J. W. Hutchinson, *Strain gradient plasticity*, Adv. Appl. Mech., 33 (1997), pp. 295–361.
[31] P. Grisvard, *Elliptic Problems in Non Smooth Domains*, Pitman, Boston, 1985.
[32] W. Hackbusch, *Multi-grid Methods and Applications*, Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1985.
[33] S. G. Kreĭn, Ju. I. Petunin, and E. M. Semenov, *Interpolation of Linear Operators*, Translations of Mathematical Monographs 54, American Mathematical Society, Providence, 1982.
[34] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications I*, Springer-Verlag, New York, 1972.
[35] J. Mandel, S. McCormick, and R. Bank, *Variational Multigrid Theory*, in Multigrid Methods, Frontiers In Applied Mathematics 3, S. McCormick, ed., SIAM, Philadelphia, 1987, pp. 131–177.
[36] L. S. D. Morley, *The triangular equilibrium problem in the solution of plate bending problems*, Aero. Quart., 19 (1968), pp. 149–169.
[37] S. A. Nazarov and B. A. Plamenevsky, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, de Gruyter, Berlin-New York, 1994.
[38] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.
[39] T. K. Nilssen, X.-C. Tai, and R. Winther, *A robust nonconforming H^2-element*, Math. Comp., 70 (2001), pp. 489–505.
[40] Z. Shi, *On the convergence of the incomplete biquadratic nonconforming plate element*, Math. Numer. Sin., 8 (1986), pp. 53–62.
[41] J. Y. Shu, W. E. King, and N. A. Fleck, *Finite elements for materials with strain gradient effects*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 373–391.
[42] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.
[43] U. Trottenberg, C. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, San Diego, 2001.
[44] G. N. Wells, K. Garikipati, and L. Molari, *A discontinuous Galerkin formulation for a strain gradient-dependent damage model*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 3633–3645.
[45] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.
[46] J. Zhao, *Convergence of nonconforming V-cycle and F-cycle methods for the biharmonic problem using the Morley element*, Electron. Trans. Numer. Anal., 17 (2004), pp. 112–132.
[47] J. Zhao, *Multigrid Methods for Fourth Order Problems*, Ph.D. thesis, University of South Carolina, Columbia, 2004.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 224–245

hp-DISCONTINUOUS GALERKIN TIME-STEPPING FOR VOLTERRA INTEGRODIFFERENTIAL EQUATIONS*

HERMANN BRUNNER† AND DOMINIK SCHÖTZAU‡

Abstract. We present an *hp*-error analysis of the discontinuous Galerkin time-stepping method for Volterra integrodifferential equations with weakly singular kernels. We derive new error bounds that are explicit in the time-steps, the degrees of the approximating polynomials, and the regularity properties of the exact solution. It is then shown that start-up singularities can be resolved at exponential rates of convergence by using geometrically graded time-steps. Our theoretical results are confirmed in a series of numerical tests.

Key words. Volterra integrodifferential equation, discontinuous Galerkin time-stepping, geometrically refined time-steps, exponential convergence

AMS subject classifications. 65R20, 65L05, 65L60

DOI. 10.1137/040619314

1. Introduction.

We introduce and analyze the *hp*-version of the discontinuous Galerkin (DG) time-stepping method for the Volterra integrodifferential equation (VIDE):

$$u'(t) + a(t)u(t) + \int_0^t k_\alpha(t-s)b(s)u(s)\,ds = f(t), \qquad t \in [0,T], \tag{1.1}$$
$$u(0) = u_0 \in \mathbb{R}.$$

Here, a, b, and f are real functions that are continuous on $[0,T]$. Moreover, we assume that there are constants $\mu^\star \geq \mu_\star > 0$ such that

$$\mu_\star \leq a(t) \leq \mu^\star, \qquad |b(t)| \leq \mu^\star, \qquad t \in [0,T]. \tag{1.2}$$

The convolution kernel k_α is the weakly singular function given by

$$k_\alpha(s) := s^{-\alpha} \qquad \text{for } \alpha \in (0,1). \tag{1.3}$$

For any initial datum $u_0 \in \mathbb{R}$, the VIDE (1.1) has a unique solution $u : [0,T] \to \mathbb{R}$ which is continuously differentiable; see, e.g., [5, 2] and the references cited therein. More precisely, smooth (analytic) data a, b, and f in (1.1) lead to solutions u that are smooth (analytic) away from $t = 0$, but their second derivatives are unbounded at $t = 0$ and behave like

$$|u''(t)| \leq Ct^{-\alpha}, \qquad t > 0;$$

see [5, 3, 4] and [2, section 7.1]; compare also Theorem 4.1 below. This loss of regularity in u at $t = 0$ has the consequence that on uniform time-steps with length k,

*Received by the editors November 20, 2004; accepted for publication (in revised form) September 28, 2005; published electronically February 8, 2006. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).
http://www.siam.org/journals/sinum/44-1/61931.html
†Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada A1C 5S7 (hermann@math.mun.ca).
‡Mathematics Department, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (schoetzau@math.ubc.ca).

approximations U generated by standard DG or collocation methods only possess low convergence order, that is,

$$\|u - U\|_{L^\infty(0,T)} \leq Ck^{1-\alpha};$$

see [4, 2]. This problem can be overcome by using meshes that are suitably refined near $t = 0$. We will show that the *hp*-version of the DG time-stepping method with geometrically graded time-steps leads to exponential rates of convergence.

The discontinuous Galerkin method was first proposed in [11] as a nonstandard finite element method for the numerical solution of neutron transport problems. Applied to initial-value ODEs, it can be viewed as an implicit single-step scheme that allows for arbitrary variation in the time-steps and the degrees of the approximating polynomials. It has been shown in [11] that despite the underlying Galerkin structure, the discontinuous Galerkin time-stepping method corresponds to certain implicit schemes of Runge–Kutta type. Subsequently, several important issues concerning the a priori and a posteriori error analyses of these schemes have been addressed; see, e.g., [7, 9, 8, 1] and the references therein. DG time-stepping has also been applied successfully to partial differential equations, and, in the context of parabolic problems, a large body of literature exists; we refer here only to the recent monograph [18] and the references cited therein. An error analysis of the DG time-stepping method applied to a parabolic integrodifferential equation was recently presented in [10].

All the works mentioned above are concerned with the *h*-version of the DG time-stepping method where convergence is achieved on successively refined time-steps using a fixed, typically low approximation order. This is in contrast to the so-called *p*- and *hp*-versions, where approximating polynomials of high degree are employed. The *hp*-approach is particularly beneficial for piecewise analytic solutions as its judicious combination of *h*- and *p*-refinement results in exponential rates of convergence. The time discretization of linear parabolic problems by the *hp*-DG time-stepping method was recently analyzed in [15, 19]. (See also [16] for extensions to problems whose spatial operators are not self-adjoint.) In particular, it has been shown that temporal start-up singularities induced by incompatible initial data can be resolved at exponential rates of convergence. Furthermore, in [14], a complete *hp*-error analysis of the DG time-stepping method has been carried out for nonlinear initial value problems in $\mathbb{R}^d$.

In the present work, we derive new *hp*-error bounds in $L^2(0,T)$ and $L^\infty(0,T)$ for the DG time-stepping method applied to the VIDE (1.1). The L^2-framework will be particularly important in the extension of the present results to partial VIDEs. Our estimates are completely explicit in the time-steps, the polynomial degrees, and the regularity properties of the exact solution. While these estimates give optimal convergence rates in the time-steps, they also show that the DG method converges if the polynomial degrees are increased at fixed time-steps. In particular, we prove that the *p*-version DG approach gives spectral accuracy for solutions with smooth time dependence, i.e., the convergence rates are of arbitrarily high algebraic order. In order to resolve start-up singularities induced by the weakly singular kernel k_α in (1.3), we employ time-steps that are geometrically refined toward $t = 0$, combined with polynomial degrees that are linearly increasing. We show that this *hp*-version approach leads to exponential rates of convergence for analytic data a, b, and f, despite the unboundedness of the second derivative of u near $t = 0$. We present a series of numerical experiments that confirm our theoretical results.

Finally, we observe that since the main purpose of this paper is to obtain insight into the basic *hp*-error analysis of DG methods on geometrically graded time-steps

for partial VIDEs, there will be no loss of generality by using the model problem given by (1.1)–(1.3). In a sequel to this paper we shall use this insight as the key to obtain an analogous estimate for partial VIDEs; we will then also describe typical applications of such VIDEs.

The outline of the paper is as follows. In section 2, we introduce the DG time-stepping method for the VIDE (1.1) and prove existence and uniqueness of approximate solutions. In section 3 we carry out a complete *hp*-error analysis of the DG method. In section 4 we show that on the basis of precise regularity results, the solutions of (1.1) can be approximated exponentially fast on time-steps that are geometrically graded toward $t = 0$. Our theoretical results are verified in the numerical tests in section 5. Finally, we end our presentation in section 6 with concluding remarks pointing to future work and open problems.

Throughout, standard notations and conventions are used. For an interval I, we write $L^p(I)$, $1 \le p \le \infty$, for the Lebesgue space of p-integrable functions, endowed with the norm $\|\cdot\|_{L^p(I)}$. We write $W^{k,p}(I)$ for the Sobolev space of order $k \in \mathbb{N}_0$ equipped with the usual norm $\|\cdot\|_{W^{k,p}(I)}$. For a noninteger exponent $s \ge 0$, the space $W^{s,p}(I)$ is defined by the K-method of interpolation. We set $H^s(I) = W^{s,2}(I)$. We write $\mathcal{P}^r(I)$ for the space of all polynomials of degree $\le r$. We denote by C generic constants not necessarily identical at different places but always independent of the discretization parameters of interest (such as time-steps and polynomial degrees).

2. Discontinuous Galerkin time-stepping. In this section, we introduce the discontinuous Galerkin time-stepping method for the numerical approximation of the VIDE (1.1). We then show the existence and uniqueness of the approximate solutions.

2.1. Discontinuous Galerkin discretization. Let $\mathcal{M}$ be a partition of $(0,T)$ into intervals $\{I_m\}_{m=1}^M$ given by $I_m := (t_{m-1}, t_m)$ with nodes

$$0 =: t_0 < t_1 < \cdots < t_{M-1} < t_M := T.$$

The length of I_m is $k_m := t_m - t_{m-1}$. As usual, we set $k := \max_{m=1}^M k_m$. The partition $\mathcal{M}$ is called quasi-uniform if there is a constant $C > 0$ such that $k \le Ck_m$ for all $1 \le m \le M$.

We assign to each interval I_m a polynomial degree $r_m \ge 0$ and introduce the degree vector $\underline{r} = \{r_m\}_{m=1}^M$. We define $|\underline{r}| := \max_{m=1}^M r_m$. The tuple $(\mathcal{M}, \underline{r})$ is called an *hp*-discretization of $(0,T)$. If $r_m = r$ for all $1 \le m \le M$, we simply write $(\mathcal{M}, r)$.

Let $\varphi : (0,T) \to \mathbb{R}$ be a function that is piecewise continuous with respect to the partition $\mathcal{M}$. At the nodes the left- and right-sided limits of φ are defined by

$$\varphi_m^+ = \lim_{s \to 0,\ s>0} \varphi(t_m + s), \qquad 0 \le m \le M-1,$$
$$\varphi_m^- = \lim_{s \to 0,\ s>0} \varphi(t_m - s), \qquad 1 \le m \le M.$$

The jumps across interior nodes are given by $[\![\varphi]\!]_m = \varphi_m^+ - \varphi_m^-$, $1 \le m \le M-1$.

For a given *hp*-discretization $(\mathcal{M}, \underline{r})$ of $(0,T)$, we introduce the discrete space

$$\mathcal{V}(\mathcal{M}, \underline{r}) := \big\{ \varphi \in L^2(0,T) :\ \varphi|_{I_m} \in \mathcal{P}^{r_m}(I_m),\ 1 \le m \le M \big\}. \tag{2.1}$$

Note that functions in $\mathcal{V}(\mathcal{M}, \underline{r})$ can be discontinuous across the nodes $\{t_m\}$.

We consider the following discontinuous Galerkin approximation of the VIDE in (1.1): find $U \in \mathcal{V}(\mathcal{M}, \underline{r})$ such that

$$B_{DG}(U, V) = F_{DG}(V) \tag{2.2}$$

for all $V \in \mathcal{V}(\mathcal{M}, \underline{r})$.

The forms B_{DG} and F_{DG} are given by

$$\begin{aligned} B_{DG}(U,V) := & \sum_{m=1}^{M} \int_{I_m} \Big(U'(t) + a(t)U(t)\Big)V(t)\,dt \\ & + \sum_{m=1}^{M} \int_{I_m} \left(\int_0^t k_\alpha(t-s)b(s)U(s)\,ds\right) V(t)\,dt \\ & + \sum_{m=1}^{M-1} [\![U]\!]_m V_m^+ + U_0^+ V_0^+, \\ F_{DG}(V) := & \, u_0 V_0^+ + \sum_{m=1}^{M} \int_{I_m} f(t)V(t)\,dt. \end{aligned}$$

Note that the exact solution u of problem (1.1) satisfies $B_{DG}(u,V) = F_{DG}(V)$ for all $V \in \mathcal{V}(\mathcal{M}, \underline{r})$. Hence, we have the Galerkin orthogonality property

$$B_{DG}(u-U,V) = 0 \tag{2.3}$$

for all $V \in \mathcal{V}(\mathcal{M}, \underline{r})$.

Remark 2.1. The discontinuous Galerkin discretization in (2.2) is a time-stepping scheme: if U is given on I_n, $1 \le n \le m-1$, we find $U|_{I_m} \in \mathcal{P}^{r_m}(I_m)$ by solving

$$\begin{aligned} & \int_{I_m} \Big(U'(t) + a(t)U(t)\Big)V(t)\,dt + \int_{I_m} \left(\int_{t_{m-1}}^t k_\alpha(t-s)b(s)U(s)\,ds\right) V(t)\,dt + U_{m-1}^+ V_{m-1}^+ \\ & \quad = U_{m-1}^- V_{m-1}^+ + \int_{I_m} f(t)V(t)\,dt - \int_{I_m} \left(\int_0^{t_{m-1}} k_\alpha(t-s)b(s)U(s)\,ds\right) V(t)\,dt \end{aligned}$$

for all $V \in \mathcal{P}^{r_m}(I_m)$. Here, we set $U_0^- = u_0$.

2.2. Existence and uniqueness of discrete solutions. To show that the DG time-stepping method (2.2) defines a unique approximate solution $U \in \mathcal{V}(\mathcal{M}, \underline{r})$, we make use of the discrete Gronwall inequality from [10, Lemma 6.4].

LEMMA 2.2. *Let $\mathcal{M} = \{I_m\}_{m=1}^M$ be a partition of $(0,T)$ with $k = \max_{m=1}^M \{k_m\}$. Let $\{a_m\}_{m=1}^M$ and $\{b_m\}_{m=1}^M$ be sequences of numbers with $0 \le b_1 \le b_2 \le \cdots \le b_M$. Assume that there is a constant $K \ge 0$ such that*

$$a_1 \le b_1, \qquad a_m \le b_m + K \sum_{n=1}^{m} w_{m,n}(\alpha) a_n, \quad m = 2, \ldots, M,$$

where $w_{m,n}(\alpha) = \int_{I_n} (t_m - t)^{-\alpha}\,dt$. Assume further that $\delta = \frac{K k^{1-\alpha}}{1-\alpha} < 1$. Then we have

$$a_m \le C b_m, \qquad m = 1, \ldots, M,$$

with a constant $C > 0$ that solely depends on δ, K, α, and T.

Furthermore, we recall the following technical result from [10, Lemma 6.3].

LEMMA 2.3. *For $f \in L^2(0,\tau)$ and $\alpha \in (0,1)$ there holds*

$$\int_0^\tau \left(\int_0^t (t-s)^{-\alpha} f(s)\,ds\right)^2 dt \le \frac{\tau^{1-\alpha}}{(1-\alpha)} \int_0^\tau (\tau - t)^{-\alpha} \left(\int_0^t f(s)^2\,ds\right) dt.$$

We now address the existence and uniqueness of discrete solutions.

PROPOSITION 2.4. *Let* $(\mathcal{M}, \underline{r})$ *be an hp-discretization of* $(0,T)$ *with*

$$(\mu^\star/\mu_\star)^2 \frac{(Tk)^{(1-\alpha)}}{(1-\alpha)^2} < 1. \tag{2.4}$$

Then the discrete problem (2.2) *has a unique solution* $U \in \mathcal{V}(\mathcal{M}, \underline{r})$.

Remark 2.5. Note that condition (2.4) is independent of the degree vector $\underline{r}$.

Proof. We first show the uniqueness of DG solutions. To this end, let U and $\widetilde{U}$ be two solutions of (2.2). The difference $E = U - \widetilde{U}$ then satisfies

$$\int_{I_m} \Big(E' + aE\Big) V \, dt + E_{m-1}^+ V_{m-1}^+ = E_{m-1}^- V_{m-1}^+ - \int_{I_m} \left(\int_0^t k_\alpha(t-s) b(s) E(s) \, ds \right) V(t) \, dt$$

for any $V \in \mathcal{P}^{r_m}(I_m)$, $m = 1, \ldots, M$. Selecting $V = E$ yields

$$\frac{1}{2} \left(E_m^-\right)^2 + \frac{1}{2} \left(E_{m-1}^+\right)^2 + \int_{I_m} aE^2 \, dt = E_{m-1}^- E_{m-1}^+ - \int_{I_m} \left(\int_0^t k_\alpha(t-s) b(s) E(s) \, ds \right) E(t) \, dt.$$

Since

$$E_{m-1}^- E_{m-1}^+ \le \frac{1}{2} \left(E_{m-1}^-\right)^2 + \frac{1}{2} \left(E_{m-1}^+\right)^2,$$

we have

$$\frac{1}{2} \left(E_m^-\right)^2 + \int_{I_m} aE^2 \, dt \le \frac{1}{2} \left(E_{m-1}^-\right)^2 + \int_{I_m} \left(\int_0^t k_\alpha(t-s) |b(s)E(s)| \, ds \right) |E(t)| \, dt.$$

In view of $E_0^- = 0$, iterating the above estimate yields

$$\frac{1}{2} \left(E_m^-\right)^2 + \int_0^{t_m} aE^2 \, dt \le \int_0^{t_m} \left(\int_0^t k_\alpha(t-s) |b(s)E(s)| \, ds \right) |E(t)| \, dt =: S_m \tag{2.5}$$

for $1 \le m \le M$. By invoking the bounds for a and b in (1.2), the Cauchy–Schwarz inequality, and Lemma 2.3, the integral S_m in (2.5) can be bounded by

$$\begin{aligned} S_m &\le \mu^\star \mu_\star^{-1/2} \left(\int_0^{t_m} \left(\int_0^t k_\alpha(t-s) |E(s)| \, ds \right)^2 dt \right)^{\frac{1}{2}} \left(\int_0^{t_m} aE^2 \, dt \right)^{\frac{1}{2}} \\ &\le \frac{1}{2} (\mu^\star)^2 \mu_\star^{-1} \frac{t_m^{(1-\alpha)}}{(1-\alpha)} \int_0^{t_m} (t_m - t)^{-\alpha} \left(\int_0^t E(s)^2 \, ds \right) dt + \frac{1}{2} \int_0^{t_m} aE^2 \, dt \\ &\le \frac{1}{2} (\mu^\star/\mu_\star)^2 \frac{t_m^{(1-\alpha)}}{(1-\alpha)} \int_0^{t_m} (t_m - t)^{-\alpha} \left(\int_0^t a(s) E(s)^2 \, ds \right) dt + \frac{1}{2} \int_0^{t_m} aE^2 \, dt. \end{aligned}$$

Hence, we obtain

$$\frac{1}{2} \int_0^{t_m} aE^2 \, dt \le \frac{1}{2} (\mu^\star/\mu_\star)^2 \frac{T^{(1-\alpha)}}{(1-\alpha)} \sum_{n=1}^m \left(\int_{I_n} (t_m - t)^{-\alpha} \, dt \right) \left(\int_0^{t_n} aE^2 \, ds \right).$$

Setting $a_m = \int_0^{t_m} aE^2\,dt$ and $b_m = 0$, the Gronwall inequality in Lemma 2.2 gives

$$\int_0^{t_m} aE^2\,dt = 0, \qquad m = 1, \dots, M,$$

provided that (2.4) is satisfied. The boundedness of a thus shows that $E \equiv 0$ and $U \equiv \widetilde{U}$.

As problem (2.2) is linear and finite dimensional, the existence of solutions follows from their uniqueness. This completes the proof. ∎

3. Error analysis. In this section, we derive *hp*-version error bounds for the DG time-stepping method in (2.2).

3.1. Abstract error bounds. We start be showing abstract error bounds. To this end, for a continuous function $u : [0,T] \to \mathbb{R}$, we define the interpolant $\mathcal{I}u \in \mathcal{V}(\mathcal{M}, \underline{r})$ by

$$(\mathcal{I}u)_m^- = u_m^-, \qquad 1 \le m \le M, \tag{3.1}$$

$$\int_{I_m} \mathcal{I}u(t)V'(t)\,dt = \int_{I_m} u(t)V'(t)\,dt, \qquad V \in \mathcal{P}^{r_m}(I_m),\ 1 \le m \le M. \tag{3.2}$$

Remark 3.1. The same interpolant has been used in the h-version analysis in [10]; we also refer to [18] and the references cited therein in the context of parabolic problems. The *hp*-approximation properties of $\mathcal{I}$ have been thoroughly investigated in [14, 15] and will be used in section 3.2.

Let now u be the exact solution of (1.1) and $U \in \mathcal{V}(\mathcal{M}, \underline{r})$ the DG approximation in (2.2). We split the error $e = u - U$ into $e = \eta + \theta$ with $\eta := u - \mathcal{I}u$ and $\theta := \mathcal{I}u - U$. Using Galerkin orthogonality in (2.3) and the construction of $\mathcal{I}u$, the function θ satisfies

(3.3)
$$\int_{I_m} \Big(\theta' + a\theta\Big)V\,dt + \theta_{m-1}^+ V_{m-1}^+ = \theta_{m-1}^- V_{m-1}^+ - \int_{I_m} a\eta V\,dt$$
$$- \int_{I_m} \left(\int_0^t k_\alpha(t-s)b(s)\eta(s)\,ds\right) V(t)\,dt - \int_{I_m} \left(\int_0^t k_\alpha(t-s)b(s)\theta(s)\,ds\right) V(t)\,dt$$

for any $V \in \mathcal{P}^{r_m}(I_m)$ and $m = 1, \dots, M$.

Our first result establishes an L^2-control of θ in terms of η.

LEMMA 3.2. *Let $(\mathcal{M}, \underline{r})$ be an hp-discretization of $(0,T)$ with*

$$\delta = 3(\mu^\star/\mu_\star)^2 \frac{(Tk)^{(1-\alpha)}}{(1-\alpha)^2} < 1. \tag{3.4}$$

Then we have

$$\frac{1}{2}\int_0^{t_m} a\theta^2\,dt + \frac{1}{2}\left(\theta_m^-\right)^2 \le C\int_0^{t_m} a\eta^2\,dt, \qquad m = 1, \dots M,$$

with a constant $C > 0$ that solely depends on $\mu_\star$, $\mu^\star$, α, T, and δ in (3.4).

Remark 3.3. Note that assumption (3.4) is slightly stronger than that in (2.4) and thus implies the existence and uniqueness of discrete solutions.

Proof. We select $V = \theta$ in (3.3). This yields

$$\frac{1}{2}\left(\theta_m^-\right)^2 + \frac{1}{2}\left(\theta_{m-1}^+\right)^2 + \int_{I_m} a\theta^2\,dt = \theta_{m-1}^-\theta_{m-1}^+ - \int_{I_m} a\eta\theta\,dt$$
$$- \int_{I_m}\left(\int_0^t k_\alpha(t-s)b(s)\eta(s)\,ds\right)\theta(t)\,dt - \int_{I_m}\left(\int_0^t k_\alpha(t-s)b(s)\theta(s)\,ds\right)\theta(t)\,dt.$$

Since

$$\theta_{m-1}^-\theta_{m-1}^+ \le \frac{1}{2}\left(\theta_{m-1}^-\right)^2 + \frac{1}{2}\left(\theta_{m-1}^+\right)^2,$$

we obtain

$$\begin{aligned}\frac{1}{2}\left(\theta_m^-\right)^2 + \int_{I_m} a\theta^2\,dt \le \frac{1}{2}\left(\theta_{m-1}^-\right)^2 &+ \int_{I_m} a|\eta\theta|\,dt\\ &+ \int_{I_m}\left(\int_0^t k_\alpha(t-s)|b(s)\eta(s)|\,ds\right)|\theta(t)|\,dt\\ &+ \int_{I_m}\left(\int_0^t k_\alpha(t-s)|b(s)\theta(s)|\,ds\right)|\theta(t)|\,dt.\end{aligned}$$

Iterating this estimate gives

$$\frac{1}{2}\left(\theta_m^-\right)^2 + \int_0^{t_m} a\theta^2\,dt \le T_1 + T_2 + T_3,$$

where

$$\begin{aligned}T_1 &= \int_0^{t_m} a|\eta\theta|\,dt,\\ T_2 &= \int_0^{t_m}\left(\int_0^t k_\alpha(t-s)|b(s)\eta(s)|\,ds\right)|\theta(t)|\,dt,\\ T_3 &= \int_0^{t_m}\left(\int_0^t k_\alpha(t-s)|b(s)\theta(s)|\,ds\right)|\theta(t)|\,dt.\end{aligned}$$

We estimate each of the above terms separately.

First, we note that

$$T_1 \le \frac{3}{2}\int_0^{t_m} a\eta^2\,dt + \frac{1}{6}\int_0^{t_m} a\theta^2\,dt.$$

Next, using the bounds for a and b in (1.2), the Cauchy–Schwarz inequality, and Lemma 2.3, we have

$$\begin{aligned}T_2 &\le \mu^\star\mu_\star^{-1/2}\left(\int_0^{t_m}\left(\int_0^t k_\alpha(t-s)|\eta(s)|\,ds\right)^2 dt\right)^{\frac12}\left(\int_0^{t_m} a\theta^2\,dt\right)^{\frac12}\\ &\le \frac{3}{2}(\mu^\star/\mu_\star)^2\frac{T^{1-\alpha}}{(1-\alpha)}\int_0^{t_m}(t_m-t)^{-\alpha}\left(\int_0^t a(s)\eta(s)^2\,ds\right)dt + \frac{1}{6}\int_0^{t_m} a\theta^2\,dt\\ &\le \frac{3}{2}(\mu^\star/\mu_\star)^2\frac{T^{2(1-\alpha)}}{(1-\alpha)^2}\int_0^{t_m} a\eta^2\,ds + \frac{1}{6}\int_0^{t_m} a\theta^2\,dt.\end{aligned}$$

Analogously, we obtain

$$T_3 \le \frac{3}{2}(\mu^\star/\mu_\star)^2 \frac{T^{1-\alpha}}{(1-\alpha)} \int_0^{t_m} (t_m - t)^{-\alpha} \left(\int_0^t a(s)\theta^2(s) \right) ds\, dt + \frac{1}{6} \int_0^{t_m} a\theta^2\, dt$$
$$\le \frac{3}{2}(\mu^\star/\mu_\star)^2 \frac{T^{1-\alpha}}{(1-\alpha)} \sum_{n=1}^{m} \left(\int_{I_n} (t_m - t)^{-\alpha}\, dt \right) \left(\int_0^{t_n} a\theta^2\, ds \right) + \frac{1}{6} \int_0^{t_m} a\theta^2\, dt.$$

Combining the above estimates results in

$$\frac{1}{2} \left(\theta_m^-\right)^2 + \frac{1}{2} \int_0^{t_m} a\theta^2\, dt \le \max\left\{ \frac{3}{2}, \frac{3}{2}(\mu^\star/\mu_\star)^2 \frac{T^{2(1-\alpha)}}{(1-\alpha)^2} \right\} \int_0^{t_m} a\eta^2\, dt$$
$$+ \frac{3}{2}(\mu^\star/\mu_\star)^2 \frac{T^{1-\alpha}}{(1-\alpha)} \sum_{n=1}^{m} \left(\int_{I_n} (t_m - t)^{-\alpha}\, dt \right) \left(\int_0^{t_n} a\theta^2\, ds \right).$$

Setting

$$a_m = \int_0^{t_m} a\theta^2\, dt,$$
$$b_m = \max\left\{ 3, 3(\mu^\star/\mu_\star)^2 \frac{T^{2(1-\alpha)}}{(1-\alpha)^2} \right\} \int_0^{t_m} a\eta^2\, dt,$$

the assertion follows from Lemma 2.2. ∎

Next, we bound the derivative of θ as follows.

LEMMA 3.4. *We have*

$$\int_{I_m} |\theta'|^2 (t - t_{m-1})\, dt \le C k_m \int_0^{t_m} a(\theta^2 + \eta^2)\, dt, \qquad m = 1, \dots M,$$

with a constant $C > 0$ that solely depends on $\mu_\star$, $\mu^\star$, α, and T.

Proof. We choose $V(t) = \theta'(t)(t - t_{m-1})$ in (3.3) and obtain

$$\int_{I_m} |\theta'|^2 (t - t_{m-1})\, dt \le T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \int_{I_m} a|\theta\theta'(t - t_{m-1})|\, dt,$$
$$T_2 = \int_{I_m} a|\eta\theta'(t - t_{m-1})|\, dt,$$
$$T_3 = \int_{I_m} \left(\int_0^t k_\alpha(t-s)|b(s)\eta(s)|\, ds \right) |\theta'(t - t_{m-1})|\, dt,$$
$$T_4 = \int_{I_m} \left(\int_0^t k_\alpha(t-s)|b(s)\theta(s)|\, ds \right) |\theta'(t - t_{m-1})|\, dt.$$

Clearly, using the bounds for a in (1.2),

$$T_1 \le (\mu^\star)^{1/2} \left(\int_{I_m} a\theta^2\, dt \right)^{\frac{1}{2}} k_m^{\frac{1}{2}} \left(\int_{I_m} |\theta'|^2 (t - t_{m-1})\, ds \right)^{\frac{1}{2}},$$
$$T_2 \le (\mu^\star)^{1/2} \left(\int_{I_m} a\eta^2\, dt \right)^{\frac{1}{2}} k_m^{\frac{1}{2}} \left(\int_{I_m} |\theta'|^2 (t - t_{m-1})\, ds \right)^{\frac{1}{2}}.$$

Furthermore, by Lemma 2.3,

$$T_3 \le \mu^\star \left(\int_{I_m} \left(\int_0^t (t-s)^\alpha |\eta(s)|\, ds \right)^2 dt \right)^{\frac{1}{2}} k_m^{\frac{1}{2}} \left(\int_{I_m} |\theta'|^2 (t-t_{m-1})\, ds \right)^{\frac{1}{2}}$$
$$\le \mu^\star \mu_\star^{-1/2} \frac{T^{1-\alpha}}{1-\alpha} \left(\int_0^{t_m} a\eta^2\, dt \right)^{\frac{1}{2}} k_m^{\frac{1}{2}} \left(\int_{I_m} |\theta'|^2 (t-t_{m-1})\, ds \right)^{\frac{1}{2}}.$$

Analogously, we obtain

$$T_4 \le \mu^\star \mu_\star^{-1/2} \frac{T^{1-\alpha}}{1-\alpha} \left(\int_0^{t_m} a\theta^2\, dt \right)^{\frac{1}{2}} k_m^{\frac{1}{2}} \left(\int_{I_m} |\theta'|^2 (t-t_{m-1})\, ds \right)^{\frac{1}{2}}.$$

Combining these estimates results in

$$\int_{I_m} |\theta'|^2 (t-t_{m-1})\, dt \le C k_m \int_0^{t_m} a(\theta^2+\eta^2)\, dt.$$

This completes the proof. ∎

To control the L^∞-norm of θ in terms of the interpolation error η, we make use of the following inverse inequality from [14, Lemma 3.1].

LEMMA 3.5. *On each interval I_m there holds*

$$\|\varphi\|^2_{L^\infty(I_m)} \le C \left(\log(\max\{r_m, 2\}) \int_{I_m} |\varphi'(t)|^2 (t-t_{m-1})\, dt + \left(\varphi_m^-\right)^2 \right),$$

for any $\varphi \in \mathcal{P}^{r_m}(I_m)$, $r_m \ge 0$. The constant $C > 0$ is independent of k_m and r_m. Furthermore, the estimate cannot be improved asymptotically as $r_m \to \infty$.

The following result states an abstract error bound.

THEOREM 3.6. *Let $(\mathcal{M}, \underline{r})$ be an hp-discretization of $(0,T)$ satisfying (3.4). Then the error $u - U$ between the exact solution u and the DG approximation U satisfies*

$$\|u-U\|_{L^2(0,T)} \le C \|u - \mathcal{I}u\|_{L^2(0,T)}$$

and

$$\|u-U\|_{L^\infty(0,T)} \le C \log^{\frac{1}{2}}(\max\{|\underline{r}|, 2\}) \|u - \mathcal{I}u\|_{L^\infty(0,T)}$$

with a constant $C > 0$ that solely depends on $\mu_\star$, $\mu^\star$, α, T, and δ in (3.4).

Proof. As before, we split the error into $u - U = \eta + \theta$. Lemmas 3.2 and 3.4 yield

$$\left(\theta_m^-\right)^2 + \int_0^{t_m} a\theta^2\, dt + \int_0^{t_m} |\theta'|^2 (t-t_{m-1})\, dt \le C \int_0^{t_m} a\eta^2\, dt.$$

In view of the boundedness of a in (1.2), we obtain $\|\theta\|_{L^2(0,T)} \le C\|\eta\|_{L^2(0,T)}$. Furthermore, by Lemma 3.5,

$$\|\theta\|^2_{L^\infty(I_m)} \le C \log(\max\{|\underline{r}|, 2\}) \|\eta\|^2_{L^2(0,T)} \le C \log(\max\{|\underline{r}|, 2\}) \|\eta\|^2_{L^\infty(0,T)}$$

for $1 \le m \le M$. The error bounds follow from the triangle inequality. ∎

3.2. Error bounds. In this section, we employ the *hp*-version approximation properties of the interpolant $\mathcal{I}$ to make explicit the error bounds in Theorem 3.6.

We first recall the following results from [15, Theorem 3.10] and [14, Corollary 3.10]. We denote by Γ the Gamma function.

THEOREM 3.7. *Let $u|_{I_m} \in H^{s_m+1}(I_m)$ for $s_m \geq 0$. Then*

$$\|u - \mathcal{I}u\|^2_{L^2(I_m)} \leq C \left(\frac{k_m}{2}\right)^{2t_m+2} \frac{1}{\max\{1, r_m^2\}} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \|u\|^2_{H^{t_m+1}(I_m)}$$

for any real $0 \leq t_m \leq \min\{r_m, s_m\}$. The constant $C > 0$ is independent of k_m, r_m, t_m, and s_m. Moreover, if $u|_{I_m} \in W^{s_m+1,\infty}(I_m)$ for $s_m \geq 0$, then

$$\|u - \mathcal{I}u\|^2_{L^\infty(I_m)} \leq C \left(\frac{k_m}{2}\right)^{2t_m+2} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \|u\|^2_{W^{t_m+1,\infty}(I_m)}$$

for any real $0 \leq t_m \leq \min\{r_m, s_m\}$.

From Theorems 3.6 and 3.7 we obtain the following *hp*-error estimates.

THEOREM 3.8. *Let $(\mathcal{M}, \underline{r})$ be an hp-discretization of $(0,T)$ satisfying (3.4), and let $U \in \mathcal{V}(\mathcal{M}, \underline{r})$ be the DG approximation (2.2). Let the exact solution u of (1.1) satisfy*

$$u|_{I_m} \in H^{s_m+1}(I_m), \qquad s_m \geq 0, \qquad m = 1, \ldots, M.$$

Then we have the L^2-error bound

$$\|u - U\|^2_{L^2(0,T)} \leq C \sum_{m=1}^{M} \left(\left(\frac{k_m}{2}\right)^{2t_m+2} \frac{1}{\max\{1, r_m^2\}} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \|u\|^2_{H^{t_m+1}(I_m)} \right)$$

for any real $0 \leq t_m \leq \min\{s_m, r_m\}$, $1 \leq m \leq M$. Moreover, if

$$u|_{I_m} \in W^{s_m+1,\infty}(I_m), \qquad s_m \geq 0, \qquad m = 1, \ldots, M,$$

then we have the L^∞-error bound

$$\|u - U\|^2_{L^\infty(0,T)} \leq C \log\left(\max\{|\underline{r}|, 2\}\right) \cdot \max_{m=1}^{M} \left\{ \left(\frac{k_m}{2}\right)^{2t_m+2} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \|u\|^2_{W^{t_m+1,\infty}(I_m)} \right\}$$

for any real $0 \leq t_m \leq \min\{s_m, r_m\}$, $1 \leq m \leq M$.

The constants $C > 0$ solely depend on $\mu_\star$, $\mu^\star$, T, α, and δ in (3.4).

We remark that the estimates in Theorem 3.8 are explicit in the time-steps k_m, the polynomial degrees r_m, and the regularity exponents s_m of the exact solution. From the bounds in Theorem 3.8, the following convergence rates can be deduced for the *h*- and *p*-version of the DG time-stepping method.

COROLLARY 3.9. *Let $(\mathcal{M}, r)$ be an hp-discretization of $(0,T)$ satisfying (3.4) with uniform polynomial degree $r \geq 0$. Let u be the exact solution of (1.1) and U the discontinuous Galerkin approximation (2.2). If $u \in H^{s+1}(0,T)$ for $s \geq 0$, we have the L^2-error bound*

$$\|u - U\|_{L^2(0,T)} \leq C \frac{k^{\min(s,r)+1}}{r^{s+1}} \|u\|_{H^{s+1}(0,T)}.$$

Additionally, if $u \in W^{s+1,\infty}(0,T)$ *for* $s \geq 0$*, we have the* L^∞*-error bound*

$$\|u-U\|_{L^\infty(0,T)} \leq C \log\left(\max\{r,2\}\right) \frac{k^{\min(s,r)+1}}{r^s} \|u\|_{W^{s+1,\infty}(0,T)}.$$

The constants $C > 0$ *solely depend on* $\mu_\star$*,* $\mu^\star$*,* T*,* α*,* δ*, in* (3.4) *and the regularity exponent* s*.*

Proof. The proof follows from Theorem 3.8 and Stirling's formula; cf. [17]. ∎

The estimates in Corollary 3.9 show that the DG time-stepping method converges either as the time-steps are decreased ($k \to 0$) or as r is increased ($r \to \infty$). Both estimates are optimal in k. However, while the L^2-estimate is also optimal in the polynomial degree r, the L^∞-estimate is one power of r short from being optimal; this is due to the slightly suboptimal L^∞-approximation properties of the interpolant $\mathcal{I}$ in Theorem 3.7; see also [14].

It can be seen from Corollary 3.9 that for solutions u for which s is large it is more advantageous to increase r rather than to reduce k at fixed, low r. Indeed, if u is smooth on $[0,T]$, arbitrarily high algebraic convergence rates are possible if the polynomial degree r is raised. This is referred to as spectral convergence. Moreover, the p-version of the DG time-stepping method converges exponentially if the solution u is analytic on $[0,T]$. To see this, we first recall the following result.

LEMMA 3.10. *On each interval* I_m *there holds*

$$\|u-\mathcal{I}u\|_{L^2(I_m)} \leq C \inf_{q\in\mathcal{P}^{r_m}(I_m)} \|u-q\|_{H^1(I_m)},$$
$$\|u-\mathcal{I}u\|_{L^\infty(I_m)} \leq C r_m \inf_{q\in\mathcal{P}^{r_m}(I_m)} \|u-q\|_{W^{1,\infty}(I_m)}$$

with a constant $C > 0$ *independent of* I_m*,* r_m*, and* u*.*

Proof. The first estimate follows from [16, Lemma 3.6] and a scaling argument. The second estimate follows similarly from [14, Lemma 3.8]. ∎

THEOREM 3.11. *Let* $(\mathcal{M},r)$ *be an hp-discretization of* $(0,T)$ *satisfying* (3.4)*, with polynomial degree* $r \geq 0$*. Let the exact solution* u *of* (1.1) *be analytic on* $[0,T]$*. For the DG approximation* (2.2)*, we then have the error bound*

$$\|u-U\|_{L^p(0,T)} \leq C\exp(-br), \qquad p=2 \text{ or } p=\infty,$$

with constants $C, b > 0$ *that are independent of* r*.*

Proof. The assertion follows from Theorem 3.6, the results in Lemma 3.10, and standard approximation theory for analytic functions. ∎

4. Exponential convergence for analytic data. The exponential convergence result in Theorem 3.11 is valid for solutions that are analytic in $[0,T]$. However, this regularity assumption is unrealistic since, as discussed previously, solutions of (1.1) with analytic data have strong start-up singularities, due to the presence of the weakly singular kernel k_α, and are analytic only away from $t = 0$. In this section we show that despite this singular behavior, the *hp*-version of the DG method with geometrically graded time-steps near $t = 0$ yields exponential rates of convergence.

4.1. Analyticity of solutions. Let $\mathcal{A}(0,T)$ denote the space of the functions which are analytic on $[0,T]$. A function g in $\mathcal{A}(0,T)$ can be characterized by analyticity constants $C_g, d_g > 0$ and the growth conditions

$$|g^{(s)}(t)| \leq C_g d_g^s \Gamma(s+1), \qquad t\in[0,T],\ s \geq 0.$$

(See [17, pp. 78–79] for details.)

We assume the data a, b, and f to satisfy

$$a,\ b \in \mathcal{A}(0,T), \tag{4.1}$$

$$f(t) = f_1(t) + t^\beta f_2(t), \qquad f_i \in \mathcal{A}(0,T),\ i = 1,2,\ \beta > 0,\ \beta \notin \mathbb{N}. \tag{4.2}$$

The following result describes the analyticity properties of the exact solution u.

THEOREM 4.1. *Assume* (4.1)–(4.2) *and let* $\theta = \min\{2-\alpha, 1+\beta\}$. *Then there exist constants* $C, d > 0$ *depending only on the analyticity constants of* a, b, f_1, *and* f_2, *such that the solution* u *of* (1.1) *satisfies*

$$|u^{(s)}(t)| \le C d^s \Gamma(s+1) t^{\theta - s}, \qquad t \in (0,T],\ s \in \mathbb{N}.$$

Proof. This regularity result slightly generalizes earlier results in [4]; see also [12, 3] and [2]. We give a brief sketch of the proof; additional details can be found in [2, section 7.1].

The initial-value problem for the given VIDE (1.1) is equivalent to the second-kind Volterra integral equation

$$u(t) = g(t) + \int_0^t h_\alpha(t,s) b(s) u(s)\, ds, \quad t \in [0,T], \tag{4.3}$$

with

$$g(t) := u_0 + \int_0^t (f_1(s) + s^\beta f_2(s))\, ds,$$

$$h_\alpha(t,s) \ = \ -a(s) - \int_s^t k_\alpha(v-s) dv.$$

In particular, if $a(t) = a > 0,\ b(t) = \lambda > 0,\ f_i(t) = f_i = \text{const}$ for $t \in [0,T]$, then we have

$$g(t) = u_0 + f_1 t + \frac{f_2}{1+\beta} t^{1+\beta},$$

$$h_\alpha(t,s) = -a - \frac{\lambda}{1-\alpha}(t-s)^{1-\alpha}.$$

The resolvent kernel $R_\alpha(t,s)$ associated with the kernel

$$K_\alpha(t,s) := h_\alpha(t-s) b(s) \quad (t,s) \in D := \{(t,s) :\ 0 \le s \le t \le T\}$$

has the form

$$R_\alpha(t,s) = (t-s)^{1-\alpha} Q_\alpha(t,s), \quad (t,s) \in D.$$

Here,

$$Q_\alpha(t,s) := \sum_{n=1}^{\infty} (t-s)^{(n-1)(2-\alpha)} \Phi_n(t,s;\alpha),$$

where the series is uniformly convergent on D for all $\alpha \in (0,1)$. If the given data a and b are in $\mathcal{A}(0,T)$, then we have $\Phi_n(\cdot,\cdot;\alpha) \in \mathcal{A}(D)$ $(n \ge 1)$ for all $\alpha \in (0,1)$. Here, $\mathcal{A}(D)$ denotes the space of the functions that are analytic on D.

Since the (unique) solution of the VIDE (4.3) is given by

$$u(t) = g(t) + \int_0^t R_\alpha(t,s) g(s)\, ds, \quad t \in [0,T], \tag{4.4}$$

the regularity properties of the nonhomogeneous term g imply the asserted bounds for $u^{(s)}(t)$ on $(0,T]$. ∎

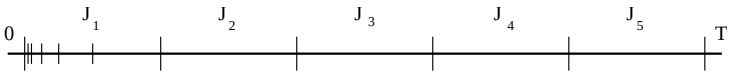


FIG. 4.1. *Example of a geometric partition $\mathcal{M}_{n,\sigma}$ of $(0,T)$. The intervals $\{J_k\}_{k=1}^5$ form the coarse partition while J_1 is geometrically refined toward $t=0$. Here, $n=5$ and $\sigma=0.5$.*

4.2. Exponential convergence for analytic data. In this section, we show that under the analyticity assumption in (4.1)–(4.2), the *hp*-version of the DG time-stepping method leads to exponential rates of convergence.

We start with the following definition.

DEFINITION 4.2. *The basic geometric partition $\widehat{\mathcal{M}}_{n,\sigma}=\{I_m\}_{m=1}^{n+1}$ of $\widehat{J}=(0,1)$ with grading factor $\sigma\in(0,1)$ and n levels of refinement is given by*

$$t_0=0, \qquad t_m=\sigma^{n-m+1}, \qquad 1\le m\le n+1.$$

Away from $t=0$, i.e., for $2\le m\le n+1$, the intervals $I_m\in\widehat{\mathcal{M}}_{n,\sigma}$ satisfy

$$k_m=t_m-t_{m-1}=\lambda t_{m-1}, \qquad \lambda:=\sigma^{-1}(1-\sigma). \tag{4.5}$$

DEFINITION 4.3. *A geometric partition $\mathcal{M}_{n,\sigma}$ of $(0,T)$ with grading factor $\sigma\in(0,1)$ and n levels of refinement is obtained by first quasi-uniformly partitioning $(0,T)$ into intervals $\{J_k\}_{k=1}^K$. The first interval $J_1=(0,t_1)$ near $t=0$ is then further subdivided into $n+1$ subintervals $\{I_m\}_{m=1}^{n+1}$ by linearly mapping to basic geometric mesh $\widehat{\mathcal{M}}_{n,\sigma}$ in Definition 4.2 onto J_1.*

An illustration of a geometric partition $\mathcal{M}_{n,\sigma}$ is given in Figure 4.1. We point out that the coarse intervals $\{J_k\}_{k=2}^K$ will be kept fixed; convergence will be achieved there by increasing the polynomial degrees.

LEMMA 4.4. *Assume (4.1)–(4.2) and set $\theta=\min\{2-\alpha,1+\beta\}$. Let $\mathcal{M}_{n,\sigma}$ be a geometric mesh of $(0,T)$ with $\{J_k\}_{k=1}^K$ denoting the underlying quasi-uniform partition of $(0,T)$ and $\{I_m\}_{m=1}^{n+1}$ the geometric refinement of J_1. Then the solution u of (1.1) satisfies*

$$\|u\|^2_{W^{1,\infty}(I_1)}\le C$$

and

$$\begin{aligned}
\|u\|^2_{W^{s+1,\infty}(I_m)} &\le Cd^{2s}\Gamma(2s+1)\sigma^{2(n-m+2)(\theta-s-1)}, && 2\le m\le n+1,\\
\|u\|^2_{W^{s+1,\infty}(I_k)} &\le Cd^{2s}\Gamma(2s+1), && 2\le k\le K,
\end{aligned}$$

for $s\ge 0$. The constants $C,d>0$ are independent of m, n, and s.

Remark 4.5. We point out that the constants C and d in Lemma 4.4 depend on the underlying quasi-uniform partition $\{J_k\}_{k=1}^K$ of $\mathcal{M}_{n,\sigma}$.

Proof. This is a simple consequence of Theorem 4.1, Definition 4.2, Definition 4.3, and properties of the Gamma function. ∎

DEFINITION 4.6. *Let $\mathcal{M}_{n,\sigma}$ be a geometric mesh of $(0,T)$ with $\{J_k\}_{k=1}^K$ denoting the underlying quasi-uniform partition of $(0,T)$ and $\{I_m\}_{m=1}^{n+1}$ the geometric refinement of J_1. A degree vector $\underline{r}$ on $\mathcal{M}_{n,\sigma}$ is called linear with slope $\mu>0$ if $r_m=\lfloor\mu m\rfloor$ on the geometrically refined elements $\{I_m\}_{m=1}^{n+1}$ and if $r_k=\lfloor\mu(n+1)\rfloor$ on the coarse elements J_k, $2\le k\le K$, away from $t=0$.*

Our next result establishes exponential rates of convergence under the analyticity assumptions in (4.1) and (4.2).

THEOREM 4.7. *Assume* (4.1)–(4.2). *Let* $\mathcal{M}_{n,\sigma}$ *be a geometric partition of* $(0,T)$ *satisfying* (3.4). *Then there exists a slope* $\mu_0 > 0$ *solely depending on* σ, α, β, *and the constants* C *and* d *in Lemma* 4.4 *such that for all linear polynomial degree vectors* $\underline{r}$ *with slope* $\mu \geq \mu_0$ *the DG approximation* $U \in \mathcal{V}(\mathcal{M}_{n,\sigma}, \underline{r})$ *satisfies the error estimate*

$$\|u - U\|_{L^p(0,T)} \leq C \exp(-bN^{\frac{1}{2}}), \qquad p = 2 \text{ or } p = \infty,$$

with constants $C, b > 0$ *that are independent of* $N = \dim(\mathcal{V}(\mathcal{M}_{n,\sigma}, \underline{r}))$.

Proof. We first note that

$$\|u - U\|_{L^2(0,T)} \leq \sqrt{T}\|u - U\|_{L^\infty(0,T)}.$$

In view of this inequality, we only need to prove the bound for the L^∞-error. To do so, we denote by $\{J_k\}_{k=1}^K$ underlying quasi-uniform partition of $\mathcal{M}_{n,\sigma}$ and by $\{I_m\}_{m=1}^{n+1}$ the geometric refinement of the first time-step J_1 near $t = 0$. From Theorem 3.8 and Lemma 3.10, we find

$$\|u - U\|^2_{L^\infty(0,T)} \leq C \log\Big(\max\{\lfloor \mu(n+1)\rfloor, 2\}\Big) \max\left\{\max_{m=1}^{n+1} e_m, \max_{k=2}^{K} e_k\right\}$$

with

$$\begin{aligned} e_m &= \left(\frac{k_m}{2}\right)^{2t_m+2} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \|u\|^2_{W^{t_m+1,\infty}(I_m)}, && 1 \leq m \leq n+1, \\ e_k &= \inf_{q \in \mathcal{P}^{r_k}(I_k)} \|u - q\|^2_{W^{1,\infty}(I_k)}, && 2 \leq k \leq K, \end{aligned}$$

and $0 \leq t_m \leq \min(s_m, r_m)$. Due to Theorem 4.1, u is analytic away from $t = 0$ and, hence, the regularity exponents s_m can be chosen arbitrarily large for $m = 2, \ldots, n+1$.

We first bound the errors $\{e_m\}$ on the geometrically refined intervals $\{I_m\}_{m=1}^{n+1}$. On the first element I_1 near $t = 0$, we select $s_1 = t_1 = 0$ and have from Lemma 4.4

$$e_1 \leq Ck_1^2 = C\sigma^{2n}.$$

Next, fix an element I_m, $2 \leq m \leq n+1$, away from $t = 0$. From Lemma 4.4 and the definition of λ in (4.5), we obtain

$$\begin{aligned} e_m &\leq C \left(\frac{\lambda\sigma^{n-m+2}}{2}\right)^{2t_m+2} \\ &\quad \cdot \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)} \left(\sigma^{n-m+2}\right)^{2(\theta-t_m-1)} d^{2t_m}\Gamma(2t_m+1) \\ &= C\,\sigma^{(n-m+2)2\theta} \left((\lambda d)^{2t_m} \frac{\Gamma(r_m+1-t_m)}{\Gamma(r_m+1+t_m)}\Gamma(2t_m+1)\right). \end{aligned}$$

Taking $t_m = \gamma_m r_m$ with $\gamma_m \in (0,1)$, Stirling's formula leads to

$$e_m \leq C\sigma^{(n-m+2)2\theta} r_m^{1/2} \left((\lambda d)^{2\gamma_m} \left(\frac{(1-\gamma_m)^{1-\gamma_m}}{(1+\gamma_m)^{1+\gamma_m}}\right)\right)^{r_m}.$$

The function $f_{\lambda,d}(\gamma) = (\lambda d)^{2\gamma} \frac{(1-\gamma)^{1-\gamma}}{(1+\gamma)^{1+\gamma}}$ satisfies

$$0 < \inf_{0<\gamma<1} f_{\lambda,d}(\gamma) =: f_{\lambda,d}(\gamma_{\min}) < 1 \qquad \text{with } \gamma_{\min} = \frac{1}{\sqrt{1+\lambda^2 d^2}}.$$

Set $f_{\min} = f_{\min}(\lambda, d) =: f_{\lambda,d}(\gamma_{\min})$ and select $\gamma_m = \gamma_{\min}$ for $2 \le m \le n+1$. Hence, for $r_m = \lfloor \mu m \rfloor$, we have

$$\begin{aligned} e_m &\le C\sigma^{(n-m+2)2\theta} r_m^{\frac{1}{2}} f_{\min}^{r_m} \le C\sigma^{(n-m+2)2\theta} (\mu m)^{\frac{1}{2}} f_{\min}^{\mu m} \\ &\le C\sigma^{2\theta n} \Big(\mu(n+1)\Big)^{\frac{1}{2}} \Big(\sigma^{(-m+2)2\theta} f_{\min}^{\mu m}\Big). \end{aligned}$$

Let

$$\mu \ge \max\left\{ \frac{2\theta \log(\sigma)}{\log(f_{\min})}, 1 \right\}. \tag{4.6}$$

Then, $f_{\min}^{\mu m} \le \sigma^{2\theta m}$ and, consequently,

$$e_m \le C\sigma^{2\theta n} \Big(\mu(n+1)\Big)^{\frac{1}{2}} (\sigma^{4\theta}) \le C\sigma^{2\theta n} \Big(\mu(n+1)\Big)^{\frac{1}{2}}, \qquad m \ge 2.$$

Thus, we obtain for $1 \le m \le n+1$ the bound

$$e_m \le C \max\left\{ \sigma^{2n}, \sigma^{2\theta n} \Big(\mu(n+1)\Big)^{\frac{1}{2}} \right\}. \tag{4.7}$$

Further, from standard approximation properties for analytic functions, we can bound the errors $\{e_k\}$ on the elements $\{J_k\}_{k=2}^K$ away from $t = 0$ as

$$e_k \le Ce^{-br_k} = Ce^{-b\lfloor \mu(n+1) \rfloor}, \qquad 2 \le k \le K, \tag{4.8}$$

with constants C and b that solely depend on the constants C and d in Lemma 4.4. Combining the estimates in (4.7) and (4.8) yields

$$\|u - U\|^2_{L^\infty(0,T)} \le C \log\big(\max\{\mu(n+1), 2\}\big) \max\left\{ \sigma^{2n}, \sigma^{2n\theta} \big(\mu(n+1)\big)^{\frac{1}{2}}, e^{-b\lfloor \mu(n+1) \rfloor} \right\}.$$

Since we have

$$\log\big(\max\{\mu(n+1), 2\}\big) \max\left\{ \sigma^{2n}, \sigma^{2n\theta} \big(\mu(n+1)\big)^{\frac{1}{2}}, (e)^{-b\lfloor \mu(n+1) \rfloor} \right\} \le C \exp(-bn),$$

as $n \to \infty$, and $N = \dim(\mathcal{V}(\mathcal{M}_{n,\sigma}, \underline{r}) \le Cn^2$, the L^∞-error bound follows. ∎

Remark 4.8. From a practical point of view, it may be more convenient to use a fixed polynomial degree r on a geometric partition $\mathcal{M}_{n,\sigma}$. In this case, exponential convergence results for all $\sigma \in (0,1)$ provided that r is proportional to the number of refinements, i.e., $r = \lfloor \mu(n+1) \rfloor$ with the slope parameter μ. Indeed, we see from the proof of Theorem 4.7 that

$$\|u - U\|_{L^\infty(0,T)} \le C \max(\sigma^{2n}, r^{\frac{1}{2}} f_{\min}^r) \le C \exp(-br) \le C \exp(-bN^{1/2}).$$

Note that condition (4.6) on the slope is not necessary in this case.

5. Numerical experiments. In this section, we present a set of numerical experiments that confirm our theoretical error bounds. Throughout, we consider problem (1.1)–(1.3) with $T = 1$ and

$$a(t) = 1, \qquad b(t) = \exp(t), \qquad u_0 = 0.$$

We choose the right-hand side f such that the solution u of (1.1) is given by

$$u(t) = t^{2-\alpha} \exp(-t). \tag{5.1}$$

Notice that this solution is analytic away from $t = 0$ and that for $\alpha \in (0,1)$, the second derivative u'' is unbounded near $t = 0$. Thus, the solution (5.1) is ideally suited to test the performance of the *hp*-version DG method.

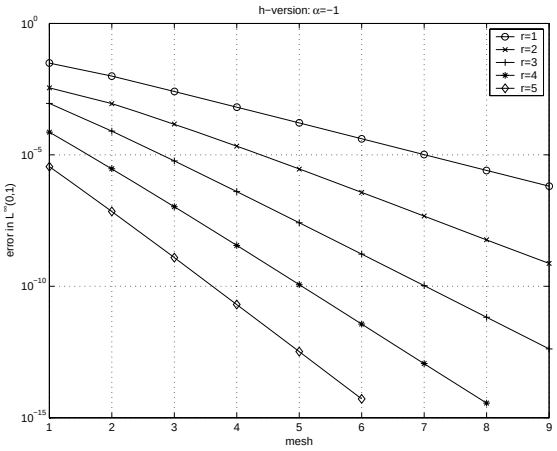


FIG. 5.1. *h-version: solution with* $\alpha = -1$.

TABLE 5.1
h-version: solution with $\alpha = -1$.

degree r	$\mathcal{M}_i$	error	order κ_i
1	7	1.03e-05	1.9982
	8	2.57e-06	1.9992
	9	6.41e-07	1.9996
2	7	4.69e-08	2.9762
	8	5.91e-09	2.9881
	9	7.42e-10	2.9941
3	7	1.05e-10	3.9852
	8	6.62e-12	3.9925
	9	4.15e-13	3.9963
4	6	3.64e-12	4.9761
	7	1.15e-13	4.9882
	8	3.59e-15	4.9940
5	4	2.03e-11	5.9170
	5	3.27e-13	5.9585
	6	5.17e-15	5.9793

5.1. Smooth solution. We start by considering the case $\alpha = -1$ so that u in (5.1) is analytic on $[0, 1]$.

In Figure 5.1, we show the errors in $L^\infty(0, 1)$ that have been obtained for the h-version DG method on a sequence $\{\mathcal{M}_i\}_{i=1}^9$ of equidistant time partitions with fixed polynomial degree $r = 1, \ldots, 5$. The partition $\mathcal{M}_i$ consists of 2^i intervals of length 2^{-i}. Hence, the straight error curves correspond to algebraic convergence in the time-step k, for each polynomial degree. To illustrate this, we compute in Table 5.1 the numerical rates of convergence $\{\kappa_i\}$ given by

$$\kappa_i = \log\left(\frac{e(\mathcal{M}_i)}{e(\mathcal{M}_{i-1})}\right) / \log(0.5)$$

with $e(\mathcal{M}_i)$ denoting the error on the partition $\mathcal{M}_i$ measured in the L^∞-norm. The convergence rates of order $r+1$ are clearly visible, which confirms the h-version result in Corollary 3.9 for a smooth solution.

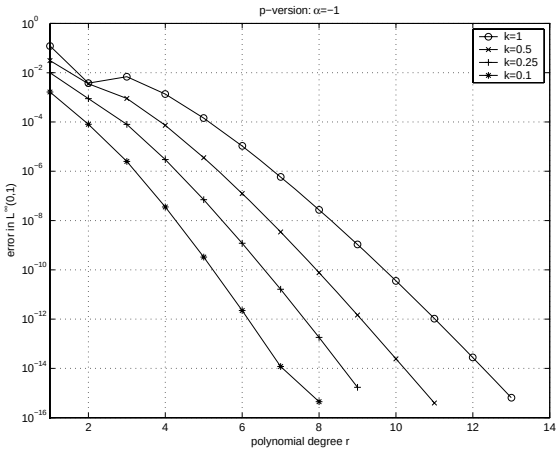


FIG. 5.2. *p-version: solution with* $\alpha = -1$.

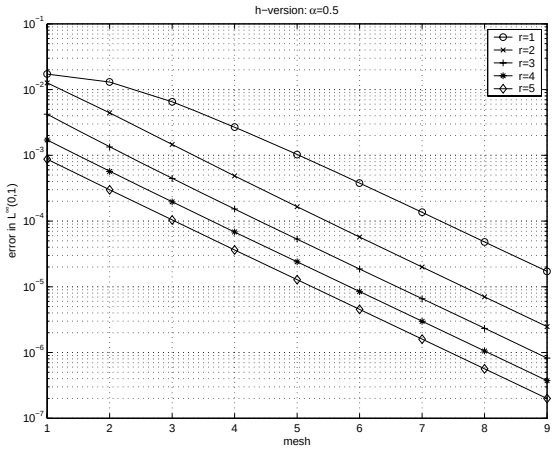


FIG. 5.3. *h-version: solution with* $\alpha = 0.5$.

Next, let us consider the p-version of the DG time-stepping method. To that end, we increase the polynomial degree from $r = 1$ to $r = 50$ for fixed partitions with time-step length $k = 1$, $k = 0.5$, $k = 0.25$, and $k = 0.1$, respectively. The performance of the p-version method is displayed in Figure 5.2. For each of the fixed time partitions the results show that exponential rates of convergence are achieved, in agreement with the theoretical findings in Theorem 3.11. (Remember that for $\alpha = -1$ the solution u is analytic in $[0, 1]$.) As expected, the smaller the underlying fixed time-step the smaller the errors that are actually obtained.

5.2. Nonsmooth solution. Next, we consider the case where $\alpha = 0.5$ so that the solution u in (5.1) has a singularity at $t = 0$. In fact, we have that $u \in W^{1.5,\infty}(0, 1)$ while the second derivative of u is unbounded near $t = 0$. In Figure 5.3, we show the performance of the h-version DG method on the uniform partitions $\mathcal{M}_i$ from section 5.1. The optimal order $r + 1$ is not obtained anymore, due to the loss of smoothness of u near the origin. Instead, the same asymptotic rate of convergence is

TABLE 5.2
h-version: solution with $\alpha = 0.5$.

degree r	i	error	order κ_i
1	7	1.3563e-04	1.4738
	8	4.8388e-05	1.4870
	9	1.7185e-05	1.4935
2	7	1.9960e-05	1.5169
	8	7.0142e-06	1.5088
	9	2.4723e-06	1.5044
3	7	6.5853e-06	1.5048
	8	2.3244e-06	1.5024
	9	8.2115e-07	1.5011
4	7	2.9812e-06	1.5023
	8	1.0532e-06	1.5011
	9	3.7221e-07	1.5006
5	7	1.5981e-06	1.5014
	8	5.6473e-07	1.5007
	9	1.9961e-07	1.5004

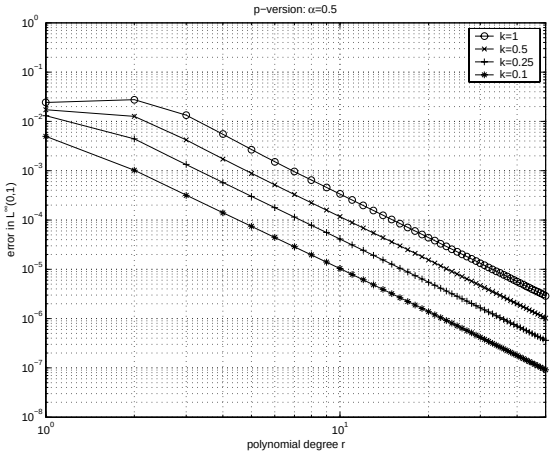


FIG. 5.4. *p-version: solution with* $\alpha = 0.5$.

observed for all polynomial degrees $r \geq 1$. This rate is computed in Table 5.2. It is of the order of 1.5 for all $r \geq 1$, thereby confirming the sharpness of the h-version result in Corollary 3.9.

Since for $\alpha = 0.5$ the solution u in (5.1) has a singularity at $t = 0$, the p-version of the DG method can only be expected to yield algebraic rates of convergence, in contrast to the test in section 5.1. Algebraic convergence behavior is indeed observed in Figure 5.4, where we increase the polynomial degree r on the same time partitions as above. The numerical convergence rates are shown in Table 5.3. In the context of the p-version DG methods, these rates are defined as

$$\kappa_r = -\log\left(\frac{e(r)}{e(r-1)}\right) / \log\left(\frac{r}{r-1}\right),$$

where $e(r)$ denotes the L^∞-error that is obtained for order r (on a fixed partition of

TABLE 5.3
p-version: solution with $\alpha = 0.5$.

r	$k = 1$		$k = 0.5$		$k = 0.25$		$k = 0.1$	
	error	κ_r	error	κ_r	error	κ_r	error	κ_r
41	5.2e-06	2.98	1.9e-06	2.98	6.5e-07	2.98	1.e-07	2.98
42	4.9e-06	2.99	1.7e-06	2.98	6.1e-07	2.98	1.5e-07	2.98
43	4.6e-06	2.98	1.6e-06	2.98	5.7e-07	2.98	1.4e-07	2.98
44	4.2e-06	2.98	1.5e-06	2.98	5.4e-07	2.98	1.4e-07	2.98
45	3.9e-06	2.99	1.4e-06	2.99	5.0e-07	2.98	1.3e-07	2.98
46	3.7e-06	2.99	1.3e-06	2.99	4.6e-07	2.98	1.2e-07	2.99
47	3.5e-06	2.99	1.2e-06	2.99	4.4e-07	2.99	1.1e-07	2.99
48	3.3e-06	2.99	1.2e-06	2.99	4.1e-07	2.99	1.0e-07	2.99
49	3.1e-06	2.99	1.1e-06	2.99	3.8e-07	2.99	9.7e-08	2.99
50	2.9e-06	2.99	1.0e-06	2.99	3.6e-07	2.99	9.2e-08	2.99

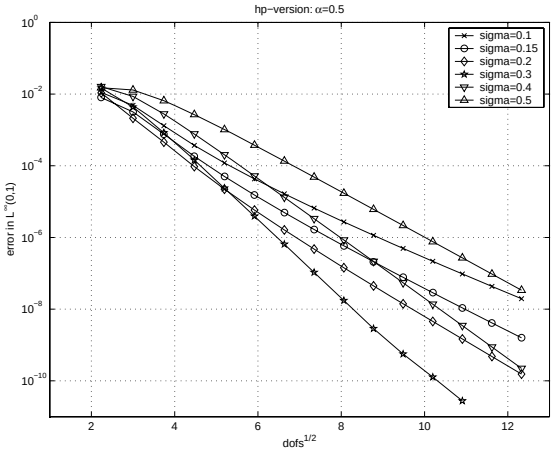


FIG. 5.5. *hp-version: solution with* $\alpha = 0.5$.

$(0, 1)$). We note that Corollary 3.9 ensures at least the order 0.5. However, rates of order 3 are observed in Table 5.3. This indicates that the estimate in Corollary 3.9 is slightly suboptimal in the polynomial degree, as remarked in the discussion after Corollary 3.9. In fact, we observe twice the rate that would correspond to the regularity exponent 1.5 of the exact solution. This doubling phenomena is well-known in p-version finite element methods for second-order boundary-value problems; see [17] and the references therein. In our context, a theoretical explanation of this observation remains an open problem.

Next, we consider the performance of the hp-version time-stepping method on the basic geometric partitions $\widehat{\mathcal{M}}_{n,\sigma} = \{I_m\}_{m=1}^{n+1}$ of $(0, 1)$ introduced in Definition 4.2. In addition, we use linearly increasing polynomial degrees as described in Definition 4.6: on time-step I_m we set $r_m = \lfloor \mu m \rfloor$ with a slope $\mu > 0$. In Figure 5.5, we display the errors against the square root of the number of degrees of freedom in the underlying discretization space, for $\mu = 1$ and various values of the grading factor σ. The straight curves indicate exponential convergence for each grading factor σ, as predicted by Theorem 4.7. It can further be seen that the grading $\sigma = 0.3$ gives the best results; for example, they are several orders of magnitude better than those for $\sigma = 0.5$. This is in contrast to the case of elliptic boundary-value problems, where the optimal

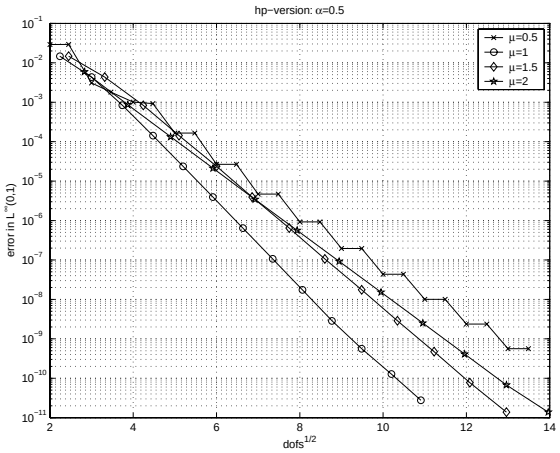


FIG. 5.6. *hp-version: solution with* $\alpha = 0.5$.

TABLE 5.4
h-version: solution with $\alpha = 0.99$.

dofs	error in $L^\infty(0,1)$
5	1.2685e-02
9	1.1820e-03
14	6.9907e-05
20	1.7007e-05
27	9.3099e-06
35	3.1848e-06
44	9.8316e-07

choice of the grading is known to be given by $\sigma \approx 0.15$, independently of the strength of the singularity; see [17] and the references therein. In Figure 5.6, we show the convergence curves for $\sigma = 0.3$ and several values of the slope parameter μ. The exponential convergence rates are less sensitive to variations in this parameter and good results are obtained for $\mu = 1$.

Finally, we test the performance of the *hp*-version DG method for the problem (5.1) with $\alpha = 0.99$. In view of the above discussions, we set $\sigma = 0.3$ and $\mu = 1$. In Table 5.4 it can be seen that with this particular choice, the *hp*-version gives an L^∞-error smaller than $1e-6$ with less than 44 degrees of freedom. To obtain the same error with the *h*-version approach on the meshes $\mathcal{M}_i$ from above and with $r = 2$, more than $10,000$ degrees of freedom are needed. This clearly underlines the suitability of *hp*-version approaches for the numerical approximation of the VIDE (1.1).

6. Concluding remarks. We conclude the paper by pointing out some extensions and future work.

In applications it often happens that at least one of the functions f_1 and f_2 in (4.2) is only *piecewise analytic* on $[0,T]$. According to the proof of Theorem 4.1 (cf. (4.4)) the corresponding solution u of (1.1) inherits this property: it is piecewise analytic on $[0,T]$, with its second derivative unbounded at $t = 0$. If the points in $[0,T]$ at which analyticity is lost are denoted by $\tau_1, \ldots, \tau_l$, it will be necessary to geometrically grade the time-steps individually near each point τ_i, $1 \le i \le l$, in order to obtain exponential convergence.

We mention in passing a related VIDE for which the above observation is relevant. Let (1.1) be replaced by

$$u'(t) + a(t)u(t) + \int_{t-\tau}^{t} k_\alpha(t-s)b(s)u(s)\,ds = f(t), \qquad t \in [0,T], \tag{6.1}$$
$$u(t) = \phi(t), \qquad t \le 0,$$

with delay $\tau > 0$. It is well known (see, e.g., [2, section 7.1]) that, regardless of the smoothness of the given functions, the solution u of (6.1) exhibits lower regularity at the so-called primary discontinuity points $\{\kappa\tau\}_{\kappa\in\mathbb{N}_0}$ induced by the delay τ. If ϕ, a, b, f_1, f_2 are analytic on $[0,T]$, then u will be analytic on each interval $(\kappa\tau,(\kappa+1)\tau]$ but only piecewise analytic on $[0,T]$.

As we mentioned in section 1, we shall study the exponential convergence of the *hp*-version of the DG method for time-stepping in a (spatially semidiscretized) parabolic *partial* VIDE (see [6]) in a forthcoming paper. Assume that such a partial VIDE has the form

$$u_t + Lu + \int_0^t k_\alpha(t-s)Bu(s)\,ds = f, \qquad t \in [0,T],\ x \in \Omega \subset \mathbb{R}^d, \tag{6.2}$$

where $-L$ denotes a strongly elliptic (spatial) partial differential operator and where B is given, for example, by $B = \Delta$ or by the scalar factor $b(s,x)$. If L_h $(= L_h(t))$ and B_h $(= B_h(s))$ denote discrete representations of L and B corresponding to a spatial discretization of (6.2) with respect to a mesh Ω_h of Ω, then (6.2) is approximated by a *system* of ordinary VIDEs analogous to (1.1) in which the roles of $a(t)$ and $b(s)$ are now assumed by the matrices $L_h(t)$ and $B_h(s)$. This suggests that our "scalar" convergence analysis can in principle be extended to these systems of VIDEs. The analysis hinges of course on appropriate regularity results for the solution of (6.2).

The situation becomes rather more difficult if we have $L = 0$ in (6.2) (see, e.g., [13]): we note that the convergence properties of the *hp*-DG method for (1.1) with $a(t) \equiv 0$ are not covered by our analysis and remain open.

Acknowledgments. The authors gratefully acknowledge the suggestion by one of the referees that led to a more general version of Theorem 4.7.

REFERENCES

[1] K. Böttcher and R. Rannacher, *Adaptive Error Control in Solving Ordinary Differential Equations by the Discontinuous Galerkin Method*, Tech. Report 96-53, IWR, Universität Heidelberg, 1996.

[2] H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Differential Equations*, Cambridge University Press, Cambridge, UK, 2004.

[3] H. Brunner, A. Pedas, and G. Vainikko, *The piecewise polynomial collocation method for nonlinear weakly singular Volterra equations*, Math. Comp., 68 (1999), pp. 1079–1095.

[4] H. Brunner, A. Pedas, and G. Vainikko, *Piecewise polynomial collocation methods for linear Volterra integrodifferential equations with weakly singular kernels*, SIAM J. Numer. Anal., 39 (2001), pp. 957–982.

[5] H. Brunner and P. van der Houwen, *The Numerical Solution of Volterra Equations*, CWI Monogr. 3, North–Holland, Amsterdam, 1986.

[6] C. Chen and T. Shih, *Finite Element Methods for Integro-Differential Equations*, World Scientific, Singapore, 1998.

[7] M. Delfour, W. Hager, and F. Trochu, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comp., 31 (1981), pp. 455–473.

[8] D. Estep, *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.

[9] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.

[10] S. Larsson, V. Thomée, and L. Wahlbin, *Numerical solution of parabolic integrodifferential equations by the discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 45–71.

[11] P. LeSaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–145.

[12] C. Lubich, *Runge-Kutta theory for Volterra and Abel integral equations of the second kind*, Math. Comp., 41 (1983), pp. 87–102.

[13] C. Lubich, I. H. Sloan, and V. Thomée, *Nonsmooth data error estimates for approximations of an evolution equation with a positive-type memory term*, Math. Comp., 65 (1996), pp. 1–17.

[14] D. Schötzau and C. Schwab, *An hp a-priori error analysis of the DG time-stepping method for initial value problems*, Calcolo, 37 (2000), pp. 207–232.

[15] D. Schötzau and C. Schwab, *Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method*, SIAM J. Numer. Anal., 38 (2000), pp. 837–875.

[16] D. Schötzau and C. Schwab, *hp-Discontinuous Galerkin time-stepping for parabolic problems*, C. R. Acad. Sci. Paris Ser. I, 333 (2001), pp. 1121–1126.

[17] C. Schwab, *p- and hp-FEM — Theory and Application to Solid and Fluid Mechanics*, Oxford University Press, Oxford, UK, 1998.

[18] V. Thomée, *Galerkin Finite Element Methods for Parabolic Equations*, Springer–Verlag, New York, 1997.

[19] T. Werder, K. Gerdes, D. Schötzau, and C. Schwab, *hp-Discontinuous Galerkin time stepping for parabolic problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6685–6708.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 246–269

APPROXIMATION THEORY FOR THE p-VERSION OF THE FINITE ELEMENT METHOD IN THREE DIMENSIONS. PART 1: APPROXIMABILITIES OF SINGULAR FUNCTIONS IN THE FRAMEWORK OF THE JACOBI-WEIGHTED BESOV AND SOBOLEV SPACES*

BENQI GUO†

Dedicated to Professor Ivo Babuška on the occasion of his 80th birthday

Abstract. This paper is the first in a series devoted to the approximation theory of the p-version of the finite element method in three dimensions. In this paper, we introduce the Jacobi-weighted Besov and Sobolev spaces in a three-dimensional setting and analyze the approximability of functions in the framework of these spaces. In particular, the Jacobi-weighted Besov and Sobolev spaces with three different weights are defined to precisely characterize the natures of the vertex singularity, the edge singularity, and the vertex-edge singularity, and to explore their best approximabilities in terms of these spaces. In the forthcoming part 2, we will apply the approximabilities of these singular functions to prove the optimal convergence of the p-version of the finite element method for elliptic problems in polyhedral domains, where the singularities of three different types occur and substantially govern the convergence of the finite element solutions.

Key words. p-version, finite element method, Jacobi-weighted Besov and Sobolev spaces, Jacobi projection, vertex singularity, edge singularity, vertex-edge singularity

AMS subject classifications. 65N30, 65N25, 35D10

DOI. 10.1137/040614803

1. Introduction.

Since the late 1970s, the p-version of the finite element method (FEM), which increases the degree of polynomials on a fixed mesh to obtain higher accuracy, has been widely used in engineering computations. There are several commercial and research codes based on the p- and hp-versions of the finite element method, for example, MSC/PROBE (MacNeal Schwendler), Poly FEM (IBM), MECHANICA (Rasna Corp.), PHLEX (Computational Mechanics), STRESSCHECK (Engineering Software Research & Development), and STRIPE (Aeronautical Research Institute of Sweden).

In 1980 it was shown that the p-version of FEM in two dimensions converges at least as fast as the traditional h-version with quasi-uniform meshes, and that it converges twice as fast as the h-version of FEM if the solution has a singularity of r^γ-type. Since then significant progress for the p-version in one and two dimensions has been made in the past two decades. The estimation of the upper bound of the approximation error in finite element solutions of the p-version in two dimensions was analyzed in [5, 6], and a detailed analysis of the p-version in one dimension is

*Received by the editors September 8, 2004; accepted for publication (in revised form) May 2, 2005; published electronically February 21, 2006. This work was partially supported by NSERC of Canada under grant OGP0046726 and by the Division of Computational Science of E-Institutes of Shanghai Municipal Education Commission under project E03004. The author was also partially supported by EPSRC of UK while participating in the special program "Computational Challenges to PDEs" in the Isaac Newton Institute for Mathematical Sciences, Cambridge University, in April–June of 2003.

http://www.siam.org/journals/sinum/44-1/61480.html

†Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China, and Department of Mathematics, University of Manitoba, Winnipeg MB R3T 2N2, Canada (guo@cc.umanitoba.ca).

available in [10]. Very recently, the author and his collaborators further developed the approximation theory of the p-version of the FEM and the boundary element method (BEM) in the framework of the Jacobi-weighted Besov and Sobolev spaces [1, 2, 3, 4, 11, 12]. In this mathematical framework, the lower and upper bounds of approximation error in FEM solutions of the p-version and in BEM solutions of p- and hp-versions for problems in polygonal domains were proved, and the optimal rate of convergence was mathematically established. The spectral method in the framework of the Jacobi-weighted Sobolev spaces has been studied and was successfully applied to singular differential equations [8, 13, 14].

In contrast to the p-version in one and two dimensions, the p-version of FEM in three dimensions is much less developed due to the complexity of three-dimensional problems. Because of a lack of effective mathematical tools and theory to deal with the complexities of three-dimensional singularities in the 1980s and 1990s, only a few results and a little analysis are available in the literature. The upper bounds in approximation error of the p-version in three dimensions were discussed for problems with singularities as a conjecture in [9] without proof and were analyzed in [15] for problems with smooth solutions belonging to $H^k(\Omega), k > 2$.

In this series of papers, we shall precisely characterize singularities and analyze the approximation to singular functions as well as smooth functions in $H^k(\Omega), k > 1$, in the framework of the Jacobi-weighted Besov and Sobolev spaces, and we prove the optimal convergence of the p-version of FEM for problems on polyhedral domains. In the first paper of the series, we shall introduce the Jacobi-weighted Besov and Sobolev spaces in three dimensions and derive the approximation results for functions in these spaces; then we verify that singular functions of different types, which arise from problems in polyhedral domains, belong to the corresponding Jacobi-weighted Besov spaces and prove their approximability by high-order polynomials. Since the approximation to functions in the Jacobi-weighted Besov and Sobolev spaces in one and two dimensions can be generalized to three dimensions without substantial difficulty and the approximability of singular functions follows from the general approximation properties for functions in the Jacobi-weighted Besov spaces and verification of the singularities in appropriate Jacobi-weighted Besov spaces, the crucial part of the paper is to prove that these singular functions belong to different Jacobi-weighted Besov spaces which are precisely designed according to the nature of these singular functions. It is well known that there are singularities of three different types in solutions of problems with piecewise analytic data and on polyhedral domains which severely govern the convergence of the FEM solution, namely, vertex singularity, edge singularity, and vertex-edge singularity. Since the vertex-edge singularity occurs in two directions and is anisotropic, the characterization of the vertex-edge singularity in the Jacobi-weighted Besov spaces is very different from those for the two-dimensional setting [1, 2, 3, 4] and for the vertex singularity and the edge singularity, which reflects the major difficulty as well as significance of the paper. The main theorems of the paper are Theorems 5.2 and 5.3, i.e., $u(x) = \rho^\gamma \sin^\sigma \theta\, \chi(\rho)\, \Psi(\theta)\, \Phi(\phi) \in B^{s,\beta}_\kappa(Q)$ with $s = 2 + 2\min\{\sigma, \gamma + (1+\beta_3)/2\} + \beta_1 + \beta_2$, the Jacobi weight $\beta = (\beta_1, \beta_2, \beta_3)$ with $\beta_i > -1$ arbitrary, and

$$\kappa = \begin{cases} 0 & \text{if } \sigma \neq \gamma + (1+\beta_3)/2, \\ 1/2 & \text{if } \sigma = \gamma + (1+\beta_3)/2, \end{cases}$$

where $Q = (-1,1)^3$ and (ρ, θ, ϕ) are the spherical coordinates with respect to the vertex $(-1,-1,-1)$ and the vertical line $L = \{x = (x_1, x_2, x_3) \mid x_1 = x_2 = -1, x_3 \in$

$(-\infty,\infty)\}$, $\chi(\rho)$, $\Psi(\theta)$, and $\Phi(\phi)$ are the usual C^∞ cutoff functions. It follows immediately from the approximability of functions in the space $B^{s,\beta}_\kappa(Q)$ that

$$\|u-\psi\|_{L^2(Q)} \le Cp^{-(2+2\min\{\sigma,\gamma+1/2\})}(1+\log p)^\kappa$$

and

$$\|u-\varphi\|_{H^1(R_0)} \le Cp^{-2\min\{\sigma,\gamma+1/2\}}\ (1+\log p)^\kappa$$

with

$$\kappa = \begin{cases} 0 & \text{if } \sigma \ne \gamma+1/2, \\ 1/2 & \text{if } \sigma = \gamma+1/2, \end{cases}$$

where R_0 denotes a conic subregion of Q which is the support of u, and ψ and φ are the Jacobi projections of u on the space $\mathcal{P}_p(Q)$ of polynomials of degree $\le p$ associated with the Legendre weight $\beta=(0,0,0)$ and the Chebyshev–Legendre weight $\beta=(-1/2,-1/2,0)$, respectively. It is worth indicating that a logarithmic term appears in the error estimation if $\sigma=\gamma+1/2$, although the function has no logarithmic singularity. This unique feature in three dimensions is precisely explored by the Jacobi-weighted space $B^{s,\beta}_\kappa(Q)$, which is an interpolation space introduced by the modified K-method. The results of this paper and forthcoming ones will significantly improve the approximation theory of the p-version of FEM in three dimensions.

The scope of the paper is as follows. In section 2 we introduce the Jacobi-weighted Besov spaces $B^{s,\beta}_\nu(Q)$ and Sobolev spaces $H^{s,\beta}_\nu(Q)$ with $Q=(-1,1)^3$, $s>0$, and integer $\nu\ge 0$, and derive error estimation of the Jacobi projections in the Jacobi-weighted Sobolev norms. In section 3 we characterize the singularity and analyze the approximability for singular functions of $\rho^\gamma\log^\nu\rho$-type with $\gamma>0,\nu\ge 0$ in terms of the space $B^{s,\beta}_{\nu^*}(Q)$. The singularity and approximability of singular functions of $r^\sigma\log^\mu r$-type with $\sigma>0$ and $\mu\ge 0$ in terms of the space $B^{s,\beta}_{\mu^*}(Q)$ are analyzed in section 4. Section 5 focuses on the characterization of singularities and the best approximation in L^2- and H^1-norms for singular functions of the $\rho^\gamma\sin^\sigma\theta$-type and $\rho^\gamma\sin^\sigma\theta\,\log^\nu\rho\,\log^\mu\sin\theta$-type with $\gamma,\sigma>0$ and integers $\nu,\mu\ge 0$ in terms of the space $B^{s,\beta}_\kappa(Q)$. Some concluding remarks are given in the last section on the effectiveness of the Jacobi-weighted Sobolev and Besov spaces by comparing the error estimations of the h- and p-versions of FEM in terms of Besov and Sobolev spaces with and without the Jacobi weights.

2. Jacobi-weighted Besov and Sobolev spaces. Let $Q=I^3=(-1,1)^3$, and let $\alpha=(\alpha_1,\alpha_2,\alpha_3)$ and $\beta=(\beta_1,\beta_2,\beta_3)$ with integer $\alpha_i\ge 0$ and real number $\beta_i>-1, 1\le i\le 3$. We introduce a weight function

$$w_{\alpha,\beta}(x)=\prod_{i=1}^3 \left(1-x_i^2\right)^{\alpha_i+\beta_i}, \tag{2.1}$$

which is referred to as the Jacobi weight. Obviously, the Jacobi polynomials and their derivatives are orthogonal with the weight $w_{\alpha,\beta}(x)$.

The Jacobi-weighted Sobolev space $H^{k,\beta}(Q)$ with integer k is defined as a closure of C^∞ functions in the norm with the Jacobi weight

$$\|u\|^2_{H^{k,\beta}(Q)}=\sum_{|\alpha|=0}^k \int_Q |D^\alpha u|^2\ w_{\alpha,\beta}(x)\,dx, \tag{2.2}$$

where $D^\alpha u = u_{x_1^{\alpha_1} x_2^{\alpha_2}, x_3^{\alpha_3}}$ and $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3$. By $|u|_{H^{k,\beta}(Q)}$ we denote the seminorm

$$|u|^2_{H^{k,\beta}(Q)} = \sum_{|\alpha|=k} \int_Q |D^\alpha u|^2 \, w_{\alpha,\beta}(x)\, dx.$$

Let $\mathcal{B}^{s,\beta}_{2,q}(Q)$ be the interpolation spaces defined by the K-method

$$(H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,q},$$

where $0 < \theta < 1, 1 \le q \le \infty, s = (1-\theta)\ell + \theta k$, ℓ and k are integers, $\ell < k$, and

$$\|u\|_{\mathcal{B}^{s,\beta}_{2,q}(Q)} = \left(\int_0^\infty t^{-q\theta} \, |K(t,u)|^q \frac{dt}{t} \right)^{1/q}, \ 1 \le q < \infty, \tag{2.3a}$$

$$\|u\|_{\mathcal{B}^{s,\beta}_{2,\infty}(Q)} = \sup_{t>0} t^{-\theta} \, K(t,u), \tag{2.3b}$$

where

$$K(t,u) = \inf_{u=v+w} \left(\|v\|_{H^{\ell,\beta}(Q)} + t\|w\|_{H^{k,\beta}(Q)} \right). \tag{2.4}$$

In particular, we are interested in the cases $q = 2$ and $q = \infty$. We shall write for $s \ge 0$ and $q = 2$

$$H^{s,\beta}(Q) = \mathcal{B}^{s,\beta}_{2,2}(Q) = (H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,2}$$

with $0 < \theta < 1$ and $s = (1-\theta)\ell + \theta k$. This space is called the Jacobi-weighted Sobolev space with fractional order if s is not an integer. It has been proved that $\mathcal{B}^{s,\beta}_{2,2}(Q) = H^{m,\beta}(Q)$ if s is an integer m in two dimensions [1]; it can be proved analogously in three dimensions.

For $q = \infty$, we shall write

$$B^{s,\beta}(Q) = \mathcal{B}^{s,\beta}_{2,\infty}(Q) = (H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,\infty},$$

which is referred to as the Jacobi-weighted Besov space. It is an exact interpolation space of θ-exponent according to [7].

For the best approximation of the singular functions such as $\rho^\gamma \log^\nu \rho, \ \nu > 0$, we need to introduce an interpolation space

$$B^{s,\beta}_\nu(Q) = (H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,\infty,\nu}$$

with integer $\nu > 0$ by a modified K-method,

$$\|u\|_{B^{s,\beta}_\nu(Q)} = \sup_{t>0} \frac{t^{-\theta} K(t,u)}{(1 + |\log t|)^\nu}. \tag{2.5}$$

Remark 2.1. The space $B^{s,\beta}_0(Q) = B^{s,\beta}(Q)$ is a standard exact interpolation space of θ-exponent; all important properties of exact interpolation spaces such as the reiteration theorem stand for $B^{s,\beta}(Q)$. It has been shown [1] that the space $B^{s,\beta}_\nu(Q)$ with $\nu > 0$ is a uniform interpolation space, but not an exact one. Hence many important properties of exact interpolation spaces do not hold for the space

$B_\nu^{s,\beta}(Q)$ with $\nu > 0$, for instance, the reiteration theorem. Fortunately a partial reiteration theorem was proved which guarantees

$$(H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,\infty,\nu} = (H^{\ell',\beta}(Q), H^{k',\beta}(Q))_{\theta',\infty,\nu}$$

as long as $(1-\theta)\ell + \theta k = (1-\theta')\ell' + \theta' k' = s$. Hence the space $B_\nu^{s,\beta}(Q)$ is well defined and does not depend on the individual values of ℓ and k but only on their combination $(1-\theta)\ell + \theta k$.

For the definition and properties of exact interpolation spaces of exponent θ, we refer to [7]. For the partial reiteration theorem and various properties of uniform interpolation space $B_\nu^{s,\beta}(Q)$ with integer $\nu > 0$, we refer to [3].

Remark 2.2. For $\beta_1 = \beta_2 = \beta_3 = 0$, the spaces $B_\nu^{s,\beta}(Q)$ are referred to as the Legendre-weighted Besov spaces. They are referred to as the Chebyshev–Legendre-weighted Besov spaces for $\beta_1 = \beta_2 = -1/2, \beta_3 = 0$ and the Chebyshev-weighted Besov spaces for $\beta_1 = \beta_2 = -1/2, \beta_3 > -1$.

We next study the approximation properties for functions in the Jacobi-weighted Sobolev spaces. Let $\mathcal{P}_p(Q)$ be a set of all polynomials of (separate) degree $\le p$. For $u \in H^{k,\beta}(Q), k \ge 0$, we have the Jacobi–Fourier expansion in $H^{0,\beta}(Q)$:

$$u(x) = \sum_{i,j,k=0}^{\infty} C_{ijk}\, P_i(x_1,\beta_1)\, P_j(x_2,\beta_2)\, P_k(x_3,\beta_3),$$

where

$$P_n(x_i,\beta_i) = \frac{(1-x_i^2)^{-\beta}}{2^n\, n!}\, \frac{d^n\,(1-x_i^2)^{\beta+n}}{dx^n}$$

is the Jacobi polynomial of degree n in variable $x_i, 1 \le i \le 3$. Then

$$u_p(x) = \sum_{i,j,k=0}^{p} C_{ijk}\, P_i(x_1,\beta_1)\, P_j(x_2,\beta_2) P_k(x_3,\beta_3)$$

is the projection of $u(x)$ on $\mathcal{P}_p(Q)$.

PROPOSITION 2.1. *Let $u \in H^{k,\beta}(Q)$, and let $u_p(x)$ be the projection of $u(x)$ on $P_p(Q)$ in $H^{0,\beta}(Q)$. Then, $u_p(x)$ is the projection on $P_p(Q)$ in $H^{\ell,\beta}(Q)$ for all $0 \le \ell \le k$, and*

$$|u_p|^2_{H^{\ell,\beta}(Q)} + |u-u_p|^2_{H^{\ell,\beta}(Q)} = |u|^2_{H^{\ell,\beta}(Q)}.$$

Proof. The proposition was proved in [3] for two dimensions. The proof can be carried over easily for one and three dimensions. □

Due to Proposition 2.1, u_p is referred to as the Jacobi projection, for which we have the following approximation property.

THEOREM 2.2. *Let $u \in H^{k,\beta}(Q)$ with integer $k \ge 1$, $\beta_i > -1, 1 \le i \le 3$, and let u_p be its $H^{0,\beta}(Q)$-projection onto $\mathcal{P}_p(Q)$. Then we have for integer $\ell \le k \le p+1$*

$$|u-u_p|_{H^{\ell,\beta}(Q)} \le C\, p^{-(k-\ell)}\, |u|_{H^{k,\beta}(Q)}. \tag{2.6}$$

Proof. The proof for one and two dimensions can be carried here for three dimensions; we will not give the details of the proof, but instead refer to [1]. □

By a standard argument of interpolation spaces, we are able to generalize Theorem 2.2 to an approximation theorem for functions in the Jacobi-weighted Besov spaces $B^{s,\beta}(Q)$.

THEOREM 2.3. *Let $u \in B^{s,\beta}(Q)$, $s > 0$, with $\beta_i > -1, 1 \le i \le 3$, and let u_p be the Jacobi projection of u on $\mathcal{P}_p(Q)$ with $p > s-1$. Then for any real $\kappa \in [0, s)$ there holds*

$$\|u - u_p\|_{H^{\kappa,\beta}(Q)} \le C\, p^{-(s-\kappa)}\, \|u\|_{B^{s,\beta}(Q)} \tag{2.7}$$

with a constant C independent of p.

THEOREM 2.4. *Let $u \in B^{s,\beta}_{\nu}(Q)$, $s > 0, \nu > 0$, with $\beta_i > -1, 1 \le i \le 3$, and let u_p be the Jacobi projection of u on $\mathcal{P}_p(Q)$ with $p > s-1$. Then for any real $\kappa \in [0, s)$, there holds*

$$\|u - u_p\|_{H^{\kappa,\beta}(Q)} \le C\, p^{-(s-\kappa)}\, (1 + \log p)^{\nu} \|u\|_{B^{s,\beta}_{\nu}(Q)} \tag{2.8}$$

with a constant C independent of p.

The proof of Theorems 2.3 and 2.4 for integer κ can be found in [3], and the usual argument of interpolation spaces leads to the estimations for noninteger κ.

3. Approximability of vertex-singular functions. Let $Q = (-1, 1)^3$, and let (ρ, θ, ϕ) be the spherical coordinates with respect to the vertex $(-1, -1, -1)$ and the vertical line $L = \{x = (x_1, x_2, x_3) \mid x_1 = x_2 = -1, x_3 \in (-\infty, \infty)\}$ with $\rho = \{\sum_{i=1}^{3}(x_i + 1)^2\}^{1/2}$, $\theta = \arctan \frac{r}{x_3+1} = \arctan \frac{\{(x_1+1)^2+(x_2+1)^2\}^{1/2}}{x_3+1} \in (0, \pi/2)$, and $\phi = \arctan \frac{x_2+1}{x_1+1} \in (0, \pi/2)$. We now consider the singular functions with $\gamma > 0$,

$$u(x) = \rho^{\gamma}\, \chi(\rho)\, \Phi(\theta, \phi) \tag{3.1}$$

and

$$v(x) = \rho^{\gamma}\, \log^{\nu} \rho\, \chi(\rho)\, \Phi(\theta, \phi) \tag{3.2}$$

with integer $\nu \ge 0$, where $\chi(\rho)$ and $\Phi(\theta, \phi)$ are C^{∞} functions such that for $0 < \rho_0 < 1$

$$\chi(\rho) = 1 \quad \text{for } 0 < \rho < \rho_0/2, \quad \chi(\rho) = 0 \quad \text{for } \rho > \rho_0,$$

and

$$\Phi(\theta, \phi) = 0 \quad \text{for } (\theta, \phi) \notin S_{\kappa_0}.$$

Hereafter, S_{κ_0} denotes a subset of the intersection of the unit sphere and Q such that the angles between the radial $A_1 - x$ and the x_i-axis are larger than κ_0. For $0 < \kappa_0 < \pi/4$, let

$$R_0 = R_{\rho_0,\kappa_0}\{x \in Q \mid 0 < \rho < \rho_0, (\theta, \phi) \in S_{\kappa_0}\}$$

as shown in Figure 3.1. Then there hold for $x \in R_0$

$$\begin{gathered} (2 - \rho_0)(1 + x_i) \le \left(1 - x_i^2\right) \le 2(1 + x_i), \quad 1 \le i \le 3, \\ \kappa_1 \le \frac{1 + x_i}{1 + x_j} \le \kappa_2, \quad 1 \le i, j \le 3, \end{gathered} \tag{3.3}$$

where $\kappa_2 = \cot \kappa_0$ and $\kappa_1 = \tan \kappa_0$. The functions defined in (3.1) and (3.2) reflect a typical singularity, referred to as the vertex singularity, which occurs in the solutions

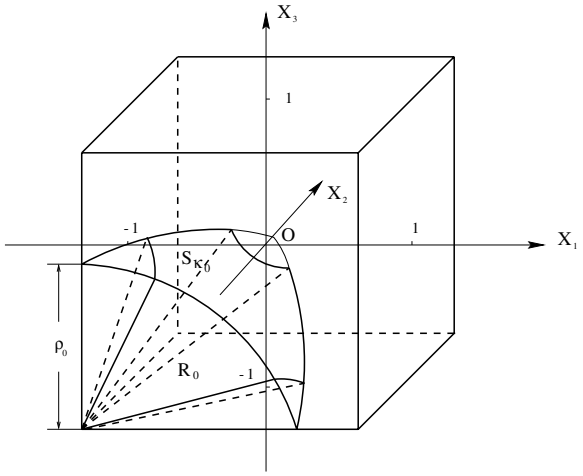


FIG. 3.1. *Cubic domain Q and subregion R_{ρ_0,κ_0}.*

of problems on polyhedral domains and severely affects the convergence of the finite element solution. Therefore, finding the best approximation to these singular functions is essential for the error estimates of the finite element solutions for problems with such singularities. It is worth indicating that the vertex singularity is isotropic, and hence the most appropriate Jacobi-weighted Besov and Sobolev spaces for their best approximation shall be isotropic as well.

3.1. Singular functions of ρ^γ-type.

THEOREM 3.1. *Let $u = \rho^\gamma \chi(\rho)\,\Phi(\theta,\phi)$ as given in* (3.1)*, and let $\beta = (\beta_1,\beta_2,\beta_3)$ with $\beta_i > -1, 1 \le i \le 3$, arbitrary. Then $u \in B^{s,\beta}(Q)$ and $u \in H^{s-\epsilon,\beta}(Q)$ with $s = 2\gamma + 3 + \sum_{i=1}^3 \beta_i$ and $\epsilon > 0$ arbitrary.*

Proof. Let $u_1 = \chi_\delta(\rho)\,u$ and $u_2 = (1-\chi_\delta(\rho))u$ with $\chi_\delta(\rho) = \chi(\frac{\rho}{\delta})$ for $\delta \in (0,\rho_0)$. Then $u = u_1 + u_2$, $u_1 \in H^{0,\beta}(Q)$ and $u_2 \in H^{k,\beta}(Q)$ for any $k > 2\gamma + 3 + \sum_{i=1}^3 \beta_i$. It is easy to see that

$$\|u_1\|^2_{H^{0,\beta}(Q)} \le C\delta^{2\gamma+3+\sum_{i=1}^3 \beta_i} \tag{3.4}$$

and

$$\|u_2\|^2_{H^{k,\beta}(Q)} \le C\delta^{2\gamma+3-k+\sum_{i=1}^3 \beta_i}. \tag{3.5}$$

Selecting $\delta = t^{\frac{2}{k}}$, we have for $t \in (0,1)$

$$\begin{aligned} K(t,u) &\le C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2}(1 + t\delta^{-k/2}) \\ &\le C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2} \le Ct^{\frac{2\gamma+3+\sum_{i=1}^3 \beta_i}{k}}, \end{aligned}$$

and for $t \ge 1$, it always holds that

$$K(t,u) \le C\|u\|_{H^{0,\beta}(Q)}.$$

Letting $\theta = \frac{2\gamma+3+\sum_{i=1}^3 \beta_i}{k}$, we have

$$\sup_{t>0} t^{-\theta}\,K(t,u) \le C,$$

which implies that $u \in (H^{0,\beta}(Q), H^{k,\beta}(Q))_{\theta,\infty} = B^{s,\beta}(Q)$ with $s = \theta k = 2\gamma + 3 + \sum_{i=1}^{3} \beta_i$.

If $\theta = \frac{2\gamma + 3 + \sum_{i=1}^{3} \beta_i - \epsilon}{k} = \frac{s-\epsilon}{k}$ with $\epsilon > 0$ arbitrary, then

$$\int_0^1 t^{-2\theta} |K(t,u)|^2 \frac{dt}{t} \le C \int_0^1 t^{-1+2\epsilon/k} dt \le C,$$

which implies that $u \in (H^{0,\beta}(Q), H^{k,\beta}(Q))_{\theta,2} = H^{s-\epsilon,\beta}(Q)$. ∎

The approximability of the singular function of ρ^γ-type is a consequence of Theorems 2.3 and 3.1.

THEOREM 3.2. *For $u = \rho^\gamma \chi(\rho) \Phi(\theta, \phi)$ given in (3.1), there exists $\psi \in \mathcal{P}_p(Q)$ with $p > 2 + 2\gamma$ such that*

$$\|u - \psi\|_{L^2(Q)} \le C p^{-(2\gamma+3)} \|u\|_{B^{2\gamma+3,\beta}(Q)} \tag{3.6}$$

with $\beta = (0,0,0)$. Also, there exists $\varphi \in \mathcal{P}_p(Q), p > 1 + 2\gamma$, such that

$$\|u - \varphi\|_{H^1(R_0)} \le C \|u - \varphi\|_{H^{1,\beta}(Q)} \le C p^{-(2\gamma+1)} \|u\|_{B^{2\gamma+2,\beta}(Q)} \tag{3.7}$$

with $\beta = (-1/3, -1/3, -1/3)$.

Proof. By Theorem 3.1 $u \in B^{s,\beta}(Q)$ with $s = 2\gamma + 3$ and $\beta = (0,0,0)$. Due to Theorem 2.2, the Jacobi projection ψ of u associated with the weight $\beta = (0,0,0)$ on $\mathcal{P}_p(Q)$ with $p > 2 + 2\gamma$ satisfies

$$\|u - \psi\|_{L^2(Q)} = \|u - \psi\|_{H^{0,\beta}(Q)} \le C\, p^{-(2\gamma+3)} \|u\|_{B^{2\gamma+3,\beta}(Q)}.$$

For $\beta = (-1/3, -1/3, -1/3)$, by Theorem 3.1, $u \in B^{s,\beta}(Q)$ with $s = 2\gamma + 2$. Owing to Theorem 2.3, it holds for the Jacobi projection φ of u associated with the weight $\beta = (-1/3, -1/3, -1/3)$ on $\mathcal{P}_p(Q)$ with $p > 1 + 2\gamma$ that

$$|u - \varphi|_{H^{\ell,\beta}(Q)} \le C\, p^{-(2\gamma+2-\ell)} \|u\|_{B^{2\gamma+2,\beta}(Q)}$$

for $\ell = 0, 1$. Note that

$$\|u - \varphi\|_{L^2(Q)} \le \|u - \varphi\|_{H^{0,\beta}(Q)} \le C p^{-(2\gamma+2)} \|u\|_{B^{2\gamma+2,\beta}(Q)}. \tag{3.8}$$

Due to (3.3), for α with $|\alpha| = 1$ and for $x \in R_0$, there exist two constants C_1 and C_2 such that

$$C_1 \le \prod_{1 \le i \le 3} \left(1 - x_i^2\right)^{\alpha_i - 1/3} \le C_2. \tag{3.9}$$

Then, we have

$$\begin{aligned} \int_{R_0} |D^\alpha (u - \varphi)|^2 \, dx &\le C \int_{R_0} |D^\alpha (u - \varphi)|^2 \prod_{1 \le i \le 3} (1 - x_i^2)^{\alpha_i - 1/3} \, dx \\ &\le C \int_Q |D^\alpha (u - \varphi)|^2 \prod_{1 \le i \le 3} (1 - x_i^2)^{\alpha_i - 1/3} \, dx \\ &\le C p^{-2(2\gamma+1)} \|u\|^2_{B^{2\gamma+2,\beta}(Q)}, \end{aligned}$$

which together with (3.8) leads to (3.7). ∎

3.2. Singular functions of $\rho^\gamma \log^\nu \rho$-type. It can be proved that the singular function $v(x) = \rho^\gamma \log^\nu \rho\, \chi(\rho)\, \Phi(\theta, \phi)$, given in (3.2), belongs to the space $B^{s-\epsilon,\beta}(Q)$ with $s = 2\gamma + 3 + \sum_{i=1}^3 \beta_i$ and $\epsilon > 0$ arbitrary. Consequently, the approximation error will lose a rate of $O(p^\epsilon)$. To avoid such a loss, the modified Jacobi-weighted Besov spaces will be the most appropriate spaces for the vertex-singular functions with logarithmic terms to describe the nature of singularity and to explore the best approximation.

THEOREM 3.3. *Let $v(x) = \rho^\gamma \log^\nu \rho\, \chi(\rho)\, \Phi(\theta, \phi)$ as given in (3.2), and let $\beta = (\beta_1, \beta_2, \beta_3)$ with $\beta_i > -1$, $1 \le i \le 3$, arbitrary. Then $v \in H^{s-\epsilon,\beta}(Q)$, and $v \in B^{s,\beta}_{\nu^*}(Q)$, with $s = 2\gamma + 3 + \sum_{i=1}^3 \beta_i$ and $\epsilon > 0$ arbitrary, and*

$$\nu^* = \begin{cases} \max\{\nu - 1, 0\} & \text{if } \gamma \text{ is an integer,} \\ \nu & \text{if } \gamma \text{ is not an integer.} \end{cases} \tag{3.10}$$

Proof. Let $v_1 = \chi_\delta(\rho)\, v$ and $v_2 = (1 - \chi_\delta(\rho))v$ with $\chi_\delta(\rho) = \chi(\frac{\rho}{\delta})$ for $\delta \in (0, \rho_0)$. Then $v = v_1 + v_2$, $v_1 \in H^{0,\beta}(Q)$ and $v_2 \in H^{k,\beta}(Q)$ for any $k > 2\gamma + 3 + \sum_{i=1}^3 \beta_i$. It is easy to see that

$$\|v_1\|^2_{H^{0,\beta}(Q)} \le C\delta^{2\gamma+3+\sum_{i=1}^3 \beta_i} |\log \delta|^{2\nu}$$

and

$$\|v_2\|^2_{H^{k,\beta}(Q)} \le C\delta^{2\gamma+3-k+\sum_{i=1}^3 \beta_i} |\log \delta|^{2\nu}.$$

Selecting $\delta = t^{\frac{2}{k}}$, we have for $t \in (0, 1)$

$$\begin{aligned} K(t, v) &\le C(\|v_1\|_{H^{0,\beta}(Q)} + t\|v_2\|_{H^{k,\beta}(Q)}) \\ &\le C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2}(1 + t\delta^{-k/2})|\log \delta|^\nu \\ &\le C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2}(1 + |\log t|)^\nu. \end{aligned}$$

For $t \ge 1$, there hold

$$K(t, v) \le C\|v\|_{H^{0,\beta}(Q)}$$

and

$$\sup_{t>1} \frac{t^{-\theta} K(t, v)}{(1 + |\log t|)^\nu} \le C\|v\|_{H^{0,\beta}(Q)}.$$

Letting $\theta = \frac{2\gamma + 3 + \sum_{i=1}^3 \beta_i}{k}$, we have

$$\sup_{0<t<1} \frac{t^{-\theta} K(t, v)}{(1 + |\log t|)^\nu} \le C,$$

which implies that $v \in (H^{0,\beta}(Q), H^{k,\beta}(Q))_{\theta,\infty,\nu} = B^{s,\beta}_\nu(Q)$ with $s = \theta k = 2\gamma + 3 + \sum_{i=1}^3 \beta_i$. Arguing similarly as in the proof of Theorem 3.1, and selecting $\theta = \frac{2\gamma + 3 + \sum_{i=1}^3 \beta_i - \epsilon}{k} = \frac{s-\epsilon}{k}$ with $\epsilon > 0$ arbitrary, we have

$$\int_0^1 t^{-2\theta} |K(t, u)|^2 \frac{dt}{t} \le C \int_0^1 t^{-1+2\epsilon/k} (1 + |\log t|)^\nu dt \le C,$$

which implies $u \in (H^{0,\beta}(Q), H^{k,\beta}(Q))_{\theta,2} = H^{s-\epsilon,\beta}(Q)$.

If γ is an integer and the integer $\nu \geq 1$, we adopt a different decomposition of $v = v_1 + v_2$ for $\delta \in (0,1)$, namely,

$$v_1 = \rho^\gamma \left(\log^\nu \rho - \log^\nu(\rho+\delta)\right) \chi(\rho)\, \Phi(\theta,\phi)$$

and

$$v_2 = \rho^\gamma \log^\nu(\rho+\delta)\chi(\rho)\, \Phi(\theta,\phi).$$

Then $v_1 \in H^{0,\beta}(Q)$ and $v_2 \in H^{k,\beta}(Q)$ for any $k > 2\gamma + 3 + \sum_{i=1}^3 \beta_i$. Using the arguments in [1, Theorem 3.9], we have

$$\|v_1\|^2_{H^{0,\beta}(Q)} \leq C\delta^{2\gamma+3+\sum_{i=1}^3 \beta_i}\, |\log\delta|^{2(\nu-1)} \tag{3.11}$$

and

$$\|v_2\|^2_{H^{k,\beta}(Q)} \leq C\delta^{2\gamma-k+3+\sum_{i=1}^3 \beta_i}\, |\log\delta|^{2(\nu-1)}. \tag{3.12}$$

Inequalities (3.11) and (3.12) lead to

$$\begin{aligned} K(t,v) &\leq C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2}(1+t\delta^{-k/2})|\log\delta|^{\nu-1} \\ &\leq C\delta^{\gamma+(3+\sum_{i=1}^3 \beta_i)/2}|\log\delta|^{\nu-1} \end{aligned}$$

and

$$\sup_{0<t<1} \frac{t^{-\theta}K(t,v)}{(1+|\log t|)^{\nu-1}} \leq C$$

with $\delta = t^{\frac{2}{k}}$ and $\theta = \frac{2\gamma+3+\sum_{i=1}^3 \beta_i}{k}$. This implies that $v \in B^{s,\beta}_{\nu-1}(Q)$ with $s = 2\gamma+3+\sum_{i=1}^3 \beta_i$. □

The precise characterization of singularity for the singular function of $\rho^\gamma \log^\nu \rho$-type given by Theorem 3.3 leads to the best approximation to the singular function of this type. The following theorem is a consequence of Theorems 2.4 and 3.3.

THEOREM 3.4. *For $v = \rho^\gamma \log^\nu \rho\, \chi(\rho)\, \Phi(\theta,\phi)$ as given in* (3.2)*, there exists $\psi(x) \in \mathcal{P}_p(Q)$ with $p > 2 + 2\gamma$ such that*

$$\|v-\psi\|_{L^2(Q)} \leq Cp^{-(2\gamma+3)}\, (1+\log p)^{\nu^*}\, \|u\|_{B^{2\gamma+3,\beta}_{\nu^*}(Q)} \tag{3.13}$$

with $\beta = (0,0,0)$. Also, there exists $\varphi(x) \in \mathcal{P}_p(Q)$ with $p > 1 + 2\gamma$ such that

$$\|u-\varphi\|_{H^1(R_0)} \leq Cp^{-(2\gamma+2)}\, (1+\log p)^{\nu^*}\, \|u\|_{B^{2\gamma+2,\beta}_{\nu^*}(Q)} \tag{3.14}$$

with $\beta = (-1/3,-1/3,-1/3)$. In both (3.13) *and* (3.14) *ν^* is given in* (3.10)*.*

Proof. By Theorem 3.3 $v \in B^{s,\beta}_{\nu^*}(Q)$ with $s = 2\gamma+3$ and $\beta = (0,0,0)$. Due to Theorem 2.2, the Jacobi projection ψ of u on $\mathcal{P}_p(Q)$ with $p > 2+2\gamma$ associated with the weight $\beta = (0,0,0)$ satisfies

$$\|u-\psi\|_{L^2(Q)} = \|u-\psi\|_{H^{0,\beta}(Q)} \leq C\, p^{-(2\gamma+3)}\, (1+\log p)^{\nu^*}\, \|u\|_{B^{2\gamma+3,\beta}_{\nu^*}(Q)}.$$

For $\beta = (-1/3, -1/3, -1/3)$, by Theorem 3.1, $v \in B^{s,\beta}_{\nu^*}(Q)$ with $s = 2 + 2\gamma$. Owing to Theorem 2.4, there holds for the Jacobi projection φ of u on $\mathcal{P}_p(Q)$ with $p > 1 + 2\gamma$ associated with the weight $\beta = (-1/3, -1/3, -1/3)$

$$|u - \varphi|_{H^{\ell,\beta}(Q)} \le C\, p^{-(2\gamma+2-\ell)}\, (1 + \log p)^{\nu^*}\, \|u\|_{B^{2\gamma+2,\beta}_{\nu^*}(Q)} \tag{3.15}$$

for $\ell = 0, 1$. Due to (3.9) and Theorem 2.4, we have for $|\alpha| = 1$

$$\begin{aligned}\int_{R_0} |D^\alpha(u-\varphi)|^2\, dx &\le C \int_Q |D^\alpha(u-\varphi)|^2 \prod_{1\le i\le 3} (1 - x_i^2)^{\alpha_i - 1/3} dx \\ &\le C p^{-2(2\gamma+2)}\, (1 + \log p)^{2\nu^*}\, \|u\|^2_{B^{2\gamma+2,\beta}_{\nu^*}(Q)},\end{aligned}$$

which together with (3.15) leads to (3.14). ∎

4. Approximability of edge-singular functions. Let $Q = (-1, 1)^3$, and let (r, ϕ, x_3) be the cylindrical coordinates with respect to the vertex (−1,−1,−1) and the vertical line $L = \{x = (x_1, x_2, x_3) \mid x_1 = x_2 = -1, x_3 \in (-\infty, \infty)\}$. Let $r = \{\sum_{i=1}^2 (x_i + 1)^2\}^{1/2}$, and let $\phi = \arctan \frac{x_2+1}{x_1+1} \in (0, \pi/2)$.

We consider the singular function with $\sigma > 0$:

$$u(x) = r^\sigma\, \chi(r)\, \Phi(\phi)\, \Psi(x_3) \tag{4.1}$$

and

$$v(x) = r^\sigma\, \log^\mu r \chi(r)\, \Phi(\phi)\, \Psi(x_3). \tag{4.2}$$

Here $\chi(r), \Psi(x_3)$, and $\Phi(\phi)$ are C^∞ functions such that for $0 < r_0 < 1$

$$\chi(r) = 1 \quad \text{for } 0 < r < r_0/2, \quad \chi(r) = 0 \quad \text{for } r > r_0,$$

and for $0 < \phi_0 < \pi/4$

$$\Phi(\phi) = 0 \quad \text{for } \phi \notin (\phi_0, \pi/2 - \phi_0),$$

and for $0 < z_0 < 1/2$

$$\Psi(x_3) = 1 \quad \text{for } x_3 \in (-1 + 2z_0, 1 - 2z_0), \quad \Psi(x_3) = 0 \quad \text{for } |x_3| \ge 1 - z_0.$$

Obviously, $u(x)$ and $v(x)$ have a support $R_{r_0,z_0} = \{x \in Q \mid 0 < r < r_0, |x_3| \le 1 - z_0\} \subset Q$. For $0 < \phi_0 < \pi/4$, let

$$R_0 = R_{r_0,\phi_0,z_0} = \{x \in Q \mid 0 < r < r_0, \phi_0 \le \phi \le \pi/2 - \phi_0, |x_3| \le 1 - z_0\},$$

as shown in Figure 4.1. Then there hold for $x \in R_0$

$$\begin{aligned} z_0(2 - z_0) &\le \left(1 - x_3^2\right) \le 1, \\ (2 - \rho_0)(1 + x_i) &\le \left(1 - x_i^2\right) \le 2(1 + x_i), \quad 1 \le i \le 2, \\ \tan\phi_0 &\le \frac{1 + x_2}{1 + x_1} \le \cot\phi_0. \end{aligned} \tag{4.3}$$

The singular functions given in (4.1) and (4.2) reflect another typical singularity in the solutions of problems in polyhedral domains and are referred to as the edge singularity. The characterization of edge singularity in appropriate functional spaces is critical to its approximability and the convergence of the finite element solutions. Although the characterization and approximability of singular functions of $r^\sigma\ \log^\mu r$-type in $Q = (-1, 1)^3$ are similar to those of vertex singular functions of $r^\gamma\ \log^\mu r$-type in two dimensions, it is worth pointing out that the edge singularity in three dimensions is anisotropic and the vertex singularity in two dimensions is isotropic.

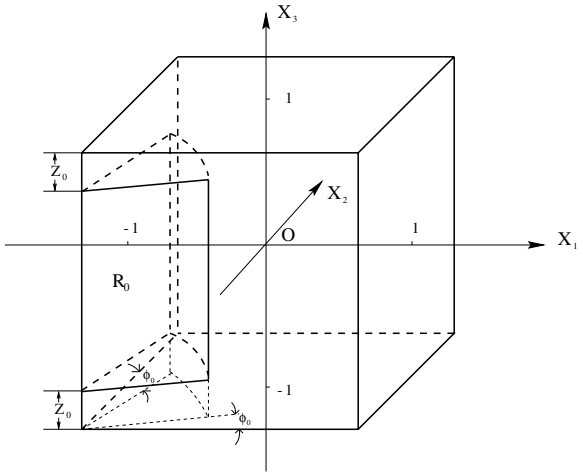


FIG. 4.1. *Cubic domain Q and subregion R_{r_0,ϕ_0,z_0}.*

4.1. Singular functions of r^σ-type.

THEOREM 4.1. *Let $u(x) = r^\sigma\,\chi(r)\,\Phi(\phi)\,\Psi(x_3)$ as given in* (4.1)*, and let $\beta = (\beta_1,\beta_2,\beta_3)$ with $\beta_i > -1$, $1 \le i \le 3$. Then $u \in B^{s,\beta}(Q)$ and $u \in H^{s-\epsilon,\beta}(Q)$ with $s = 2\sigma + 2 + \beta_1 + \beta_2$ and $\epsilon > 0$ arbitrary.*

Proof. Since u is smooth in the variable x_3 and has a support R_{r_0,z_0} in which (4.3) holds and the factor $(1-x_3^2)^{\alpha_3+\beta_3}$ is bounded from above and below, the arguments for the vertex singular functions in two dimensions can be carried over here with a minor modification. We will not repeat these here; see [1, Theorem 2.5] for details of the proof. □

Theorems 4.1 and 2.3 lead to the best approximation of the singular function u.

THEOREM 4.2. *For $u(x) = r^\sigma\,\chi(r)\,\Phi(\phi)\,\Psi(x_3)$ as given in* (4.1)*, there exists $\psi(x) \in \mathcal{P}_p(Q)$ with $p > 1 + 2\sigma$ such that*

$$\|u-\psi\|_{L^2(Q)} \le Cp^{-2(\sigma+1)}\,\|u\|_{B^{2\sigma+2,\beta}(Q)} \tag{4.4}$$

with $\beta_1 = \beta_2 = 0$ and $\beta_3 > -1$ arbitrary. Also, there exists $\varphi(x) \in \mathcal{P}_p(Q)$ with $p > 2\sigma$ such that

$$\|u-\varphi\|_{H^1(R_0)} \le C\|u-\varphi\|_{H^{1,\beta}(Q)} \le Cp^{-2\sigma}\,\|u\|_{B^{1+2\sigma,\beta}(Q)} \tag{4.5}$$

with $\beta_1 = \beta_2 = -1/2$ and $\beta_3 > -1$ arbitrary.

Proof. Due to Theorem 4.1, $u \in B^{2+2\sigma,\beta}(Q)$ with $\beta_1 = \beta_2 = 0$ and $\beta_3 > -1$ arbitrary. By Theorem 2.3, the Jacobi projection ψ of u associated with the weight $\beta = (-1/2,-1/2,\beta_3)$ on $\mathcal{P}_p(Q)$ with $p > 1 + 2\sigma$ satisfies (4.4).

For $\beta_1 = \beta_2 = -1/2$ and $\beta_3 > -1$ arbitrary, $u \in B^{1+2\sigma,\beta}(Q)$. By Theorem 2.4 the Jacobi projection φ of u associated with the weight $\beta = (-1/2,-1/2,\beta_3)$ on $\mathcal{P}_p(Q)$ with $p > 2\sigma$ satisfies

$$\|u-\varphi\|_{H^{\ell,\beta}(Q)} \le Cp^{-2\sigma-1+\ell}\,\|u\|_{B^{1+2\sigma,\beta}(Q)}$$

with $\ell = 0, 1$, which gives

$$p\|u-\varphi\|_{L^2(Q)} + \left\|\frac{\partial(u-\varphi)}{\partial x_3}\right\|_{L^2(Q)} \le Cp^{-2\sigma}\,\|u\|_{B^{1+2\sigma,\beta}(Q)}. \tag{4.6}$$

Due to (4.3), there holds for α with $\sum_{i=1}^{2}\alpha_i = 1$ and for $x \in R_0$, and there exist two constants C_1 and C_2 such that

$$C_1 \le (1-x_3^2)^{\alpha_i+\beta_3} \prod_{1\le i\le 2} (1-x_i^2)^{\alpha_i-1/2} \le C_2,$$

which implies for $|\alpha| = 1$ with $\alpha_3 = 0$

$$\|D^\alpha(u-\varphi)\|_{L^2(R_0)} \le C\|u-\varphi\|_{H^{1,\beta}(Q)} \le Cp^{-2\sigma}\ \|u\|_{B^{1+2\sigma,\beta}(Q)}.$$

This together with (4.6) leads to (4.5). ∎

4.2. Singular functions of $r^\sigma \log^\mu r$-type. For singularity with logarithmic terms we need to use the modified Jacobi-weighted Besov spaces for the best approximation.

THEOREM 4.3. *Let $v(x) = r^\sigma \log^\mu r\ \chi(r)\ \Phi(\phi)\ \Psi(x_3)$ as given in (4.2), and let $\beta = (\beta_1,\beta_2,\beta_3)$ with $\beta_i > -1$, $1 \le i \le 3$. Then $v \in B^{s,\beta}_{\mu^*}(Q)$, and $v \in H^{s-\epsilon,\beta}(Q)$ with $s = 2\sigma + 2 + \beta_1 + \beta_2$ and $\epsilon > 0$ arbitrary and*

$$\mu^* = \begin{cases} \max\{\mu-1, 0\} & \text{if } \gamma \text{ is an integer,} \\ \mu & \text{if } \gamma \text{ is not an integer.} \end{cases} \tag{4.7}$$

Proof. For the same reason mentioned in the proof of Theorem 4.1, the arguments for the vertex singular functions with the logarithmic term in two dimensions can be carried over here; we refer to [1, Theorem 3.8] for noninteger σ and [1, Theorem 3.9] for integer σ. ∎

Theorem 4.3 gives a precise characterization of the singular function of $r^\gamma \log^\mu r$-type, which avoids a loss of $O(p^\epsilon)$ in the approximation error.

THEOREM 4.4. *For $v(x) = r^\sigma \log^\mu r\ \chi(r)\,\Phi(\phi)\,\Psi(x_3)$ as given in (4.2), there exists $\psi(x) \in \mathcal{P}_p(Q), p > 1 + 2\sigma$, such that*

$$\|v-\psi\|_{L^2(Q)} \le Cp^{-(2\sigma+2)}(1+\log p)^{\mu*}\ \|v\|_{B^{2\sigma+2,\beta}_{\mu^*}(Q)} \tag{4.8}$$

with $\beta_1 = \beta_2 = 0$ and $\beta_3 > -1$ arbitrary. Also, there exists $\varphi(x) \in \mathcal{P}_p(Q), p > 2\sigma$, such that

$$\|v-\varphi\|_{H^1(R_0)} \le C\|v-\varphi\|_{H^{1,\beta}(Q)} \le Cp^{-2\sigma}(1+\log p)^{\mu^*}\ \|v\|_{B^{1+2\sigma,\beta}_{\mu^*}(Q)} \tag{4.9}$$

with $\beta_1 = \beta_2 = -1/2$ and $\beta_3 > -1$ arbitrary. In both (4.8) and (4.9) μ^* is given in (4.7).

Proof. The approximability of the singular function v is the consequence of Theorems 2.4 and 4.3. We will not elaborate details of the proof, which are similar to those of Theorems 3.4 and 4.2. ∎

5. Approximability of vertex-edge singular functions. Let $Q = (-1,1)^3$, and let (ρ,θ,ϕ) be the spherical coordinates with respect to the vertex $(-1,-1,-1)$ and the vertical line $L = \{x = (x_1,x_2,x_3) \mid x_1 = x_2 = -1, x_3 \in (-\infty,\infty)\}$ as in section 3.

We now consider the singular functions with real $\gamma,\sigma > 0$ and integers $\nu,\mu \ge 0$,

$$u(x) = \rho^\gamma\ \sin^\sigma\theta\,\chi(\rho)\,\Psi(\theta)\,\Phi(\phi) \tag{5.1}$$

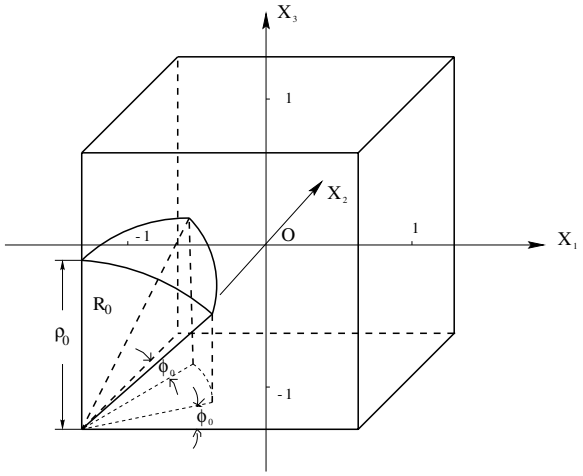


FIG. 5.1. *Cubic domain Q and subregion $R_{\rho_0,\phi_0,\theta_0}$.*

and

$$v(x) = \rho^{\gamma}\, \log^{\nu} \rho\, \sin^{\sigma} \theta\, \log^{\mu} \sin\theta\, \chi(\rho)\, \Psi(\theta)\, \Phi(\phi), \tag{5.2}$$

where $\rho = \{(x_1+1)^2+(x_2+1)^2+(x_3+1)^2\}^{1/2}$, $\chi(\rho)$ and $\Phi(\phi)$ are C^∞ cutoff functions defined in sections 3 and 4 with $0 < \rho_0 < 1$, respectively, and $\Psi(\theta)$ is a C^∞ function such that for $\theta_0 \in (0, \pi/2)$

$$\Psi(\theta) = 1 \quad \text{for } 0 \le \theta \le \theta_0/2, \quad \Psi(\theta) = 0 \quad \text{for } \theta \ge \theta_0.$$

For $0 < \phi_0 < \pi/4$, let

$$R_0 = R_{\rho_0,\theta_0,\phi_0} = \{x \in Q \mid 0 < \rho < \rho_0, \theta \in (0, \theta_0), \phi \in (\phi_0, \pi/2 - \phi_0)\}$$

as shown in Figure 5.1. Then there hold for $x \in R_0$

$$\begin{aligned} (2-\rho_0)(1+x_i) \le (1-x_i^2) &\le 2(1+x_i), \quad 1 \le i \le 3, \\ \frac{1+x_3}{1+x_i} &\ge \cot\theta_0, \quad 1 \le i \le 2, \\ \tan\phi_0 \le \frac{1+x_2}{1+x_1} &\le \cot\phi_0. \end{aligned} \tag{5.3}$$

Obviously, u has a support $R_{\rho_0,\theta_0,\phi_0} = \{x \in Q \mid 0 < \rho < \rho_0, \theta \in (0, \theta_0)\} \subset Q$.

The singularity of the functions given in (5.1) and (5.2) is the well-known vertex-edge singularity for problems on polyhedral domains, which reflect the major difficulties in characterization of the singularity and analysis of the approximability. They combine the vertex and edge singularities and are anisotropic. The combination of two types of singularities makes the analysis totally different from those in a two-dimensional setting and from those in the previous two sections for the vertex-singularity and the edge-singularity in three dimensions. Designing the Jacobi-weighted Besov spaces and proving the regularities in these spaces for the best approximation are extremely difficult and elegant.

5.1. Singular functions of $\rho^\gamma \sin^\sigma \phi$-type.

LEMMA 5.1. *Let $u(x) = \rho^\gamma \sin^\sigma \theta\, \chi(\rho)\, \Psi(\theta)\, \Phi(\phi)$ as given in (5.1). Then $u \in H^{s,\beta}(Q)$ with $\beta_i > -1$, $1 \le i \le 3$, for $s < 2 + 2\min\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$.*

Proof. Note that

$$|D^\alpha u| \le C\rho^{\gamma - |\alpha|} |\sin\theta|^{\sigma - \alpha_1 - \alpha_2}, \tag{5.4}$$

which implies that for $|\alpha| < 2 + 2\min\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$

$$\int_Q |D^\alpha u|^2 \prod_{i=1}^3 \left(1 - x_i^2\right)^{\alpha_i + \beta_i} dx \le C \int_{R_0} \rho^{2\gamma - |\alpha| + 2 + \sum_{i=1}^3 \beta_1} |\sin\theta|^{2\sigma - \alpha_1 - \alpha_2 + 1 + \sum_{i=1}^2 \beta_i} d\rho\, d\theta\, d\phi < \infty.$$

This proves the lemma for integer $s = k$. By a typical argument of interpolation spaces we are able to prove the lemma for noninteger s in general. ∎

THEOREM 5.2. *Let $u(x) = \rho^\gamma \sin^\sigma \theta\, \chi(\rho)\, \Psi(\theta)\, \Phi(\phi)$ as given in (5.1), and let $\beta = (\beta_1, \beta_2, \beta_3)$ with $\beta_i > -1, 1 \le i \le 3$. Then $u \in H^{s-\epsilon,\beta}(Q)$ and $u \in B_\kappa^{s,\beta}(Q)$ with $s = 2 + 2\min\{\sigma, \gamma + (1+\beta_3)/2\} + \beta_1 + \beta_2$, $\epsilon > 0$ arbitrary, and*

$$\kappa = \begin{cases} 0 & \text{if } \sigma \ne \gamma + (1+\beta_3)/2, \\ 1/2 & \text{if } \sigma = \gamma + (1+\beta_3)/2. \end{cases} \tag{5.5}$$

Proof. Since $r = \rho \sin\theta = \{(1+x_1)^2 + (1+x_2)^2\}^{1/2}$, we write

$$u(x) = \rho^{\gamma - \sigma}\, r^\sigma\, \chi(\rho)\, \Phi(\phi)\, \Psi(\theta),$$

and estimate (5.4) can be written as

$$|D^\alpha u(x)| \le C\rho^{\gamma - |\alpha|} |\sin\theta|^{\sigma - \alpha_1 - \alpha_2} \le C(1 + x_3)^{\gamma - \sigma - \alpha_3}\, r^{\sigma - \alpha_1 - \alpha_2}. \tag{5.6}$$

By $\varphi_\delta(r)$ we denote a C^∞ function such that $\varphi_\delta(r) = 1$ for $r < \delta$ and $\varphi_\delta(r) = 0$ for $r > 2\delta$ with $0 < \delta < \rho_0/2$. Let $u_1 = \varphi_\delta(r)u$ and $u_2 = (1 - \varphi_\delta(r))u$. Then $u_1 \in H^{0,\beta}(Q)$ due to Lemma 5.1, and $u_2 \in H^{k,\beta}(Q)$ for any $k > 2 + 2\max\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$.

Let R_{ρ_0,θ_0} be the projection of $R_{\rho_0,\theta_0,\phi_0}$ on the r-x_3 plane,

$$R_{\rho_0,\theta_0} = \{(r, x_3) \mid r\cot\theta_0 \le (1 + x_3) \le (\rho_0^2 - r^2)^{1/2}, 0 \le r \le \rho_0 \sin\theta_0\},$$

and by T_1 and T_2 we denote the triangular and rectangular regions in the r-x_3 plane, respectively,

$$T_1 = \{(r, x_3) \mid\ r\cot\theta_0 \le 1 + x_3 \le 2\delta\cot\theta_0, 0 \le r \le 2\delta\}$$

and

$$T_2 = \{(r, x_3) \mid\ 2\delta\cot\theta_0 \le 1 + x_3 \le \rho_0, 0 \le r \le 2\delta\}$$

as shown in Figure 5.2. Obviously, Supp.$u_1 \subset T_1 \cup T_2$.

Due to (5.6) there holds

$$\begin{aligned} \|u_1\|^2_{H^{0,\beta}(Q)} &= \int_{R_{\rho_0,\theta_0,\phi_0,}} |\varphi_\delta u|^2 \rho^{\sum_{i=1}^3 \beta_i} |\sin\theta|^{\beta_1 + \beta_2}\, dx \\ &\le C \int_{T_1 \cup T_2} (1 + x_3)^{2(\gamma - \sigma) + \beta_3}\, r^{2\sigma + 1 + \beta_1 + \beta_2}\, dr\, dx_3. \end{aligned} \tag{5.7}$$

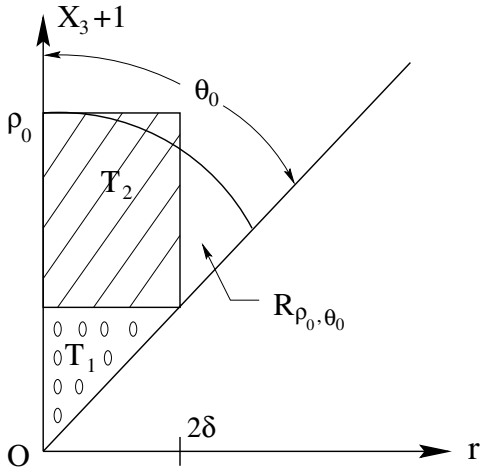


FIG. 5.2. *Regions T_1, T_2, and R_{ρ_0,θ_0}.*

Letting $\tilde{x}_3 = x_3 + 1$, we have by a simple calculation

$$\begin{aligned}
&\int_{T_1} (1+x_3)^{2(\gamma-\sigma)+\beta_3}\, r^{2\sigma+1+\beta_1+\beta_2} dr dx_3 \\
&\le C \int_0^{2\delta\cot\theta_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3}\, d\tilde{x}_3 \int_0^{\tilde{x}_3 tan\theta_0} r^{2\sigma+1+\sum_{i=1}^2 \beta_i} dr \\
&\le C\delta^{2\gamma+\sum_{i=1}^3 \beta_i+3}.
\end{aligned} \tag{5.8}$$

We also have for $\sigma \neq \gamma + (1+\beta_3)/2$

$$\begin{aligned}
&\int_{T_2} (1+x_3)^{2(\gamma-\sigma)+\beta_3}\, r^{2\sigma+1+\beta_1+\beta_2}\, dr\, dx_3 \\
&\le C \int_0^{2\delta} r^{2\sigma+1+\beta_1+\beta_2}\, dr \int_{2\delta\cot\theta_0}^{\rho_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3}\, d\tilde{x}_3 \\
&\le C(\delta^{2\gamma+3+\sum_{i=1}^3 \beta_i} + \delta^{2\sigma+2+\beta_1+\beta_2}),
\end{aligned} \tag{5.9}$$

and for $\sigma = \gamma + (1+\beta_3)/2$

$$\begin{aligned}
&\int_{T_2} (1+x_3)^{2(\gamma-\sigma)+\beta_3}\, r^{2\sigma+1+\beta_1+\beta_2}\, dr\, dx_3 \\
&\le C \int_0^{2\delta} r^{2\sigma+1+\beta_1+\beta_2} dr \int_{2\delta\cot\theta_0-1}^{\rho_0-1} (1+x_3)^{2(\gamma-\sigma)+\beta_3}\, dx_3 \\
&\le C(1+|\log\delta|)\delta^{2\sigma+2+\beta_1+\beta_2},
\end{aligned} \tag{5.10}$$

which together with (5.7)–(5.10) yields

$$\|u_1\|^2_{H^{0,\beta}(Q)} \le C(1+|\log\delta|)^{2\kappa}\delta^{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2} \tag{5.11}$$

with κ given in (5.5).

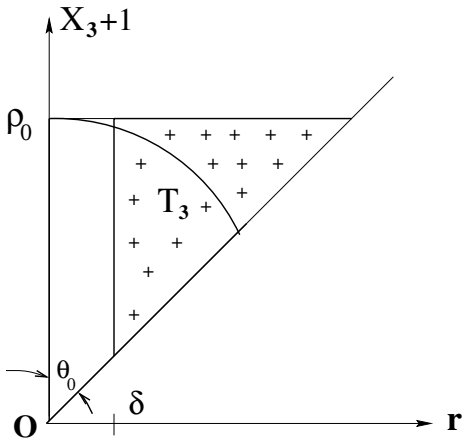


FIG. 5.3. *Region T_3.*

We next estimate $\|u_2\|_{H^{k,\beta}(Q)}$. Note that

$$\frac{\partial^k u_2}{\partial x_1^k} = (1-\varphi_\delta)\frac{\partial^k u}{\partial x_1^k} - \sum_{l=0}^{k-1}\binom{k}{l}\frac{\partial^l u}{\partial x_1^l}\frac{\partial^{k-l}\varphi_\delta}{\partial x_1^{k-l}},$$

and for $0 \le l \le k$

$$\left|\frac{\partial^{k-l}\varphi_\delta}{\partial x_1^{k-l}}\right| \le C\delta^{-(k-l)}.$$

Let

$$T_3 = \{(r, x_3) \mid\ r\cot\theta_0 \le 1 + x_3 \le \rho_0, \delta \le r \le \rho_0 \tan\theta_0\}$$

as shown in Figure 5.3. Obviously, Supp.$(1-\varphi_\delta)\frac{\partial^k u}{\partial x_1^k}$ and Supp.$\frac{\partial^{k-l}\varphi_\delta}{\partial x_1^{k-l}}$ are contained in T_3 for $0 \le l < k$. It is seen that

$$\begin{aligned}
&\int_Q \left|\frac{\partial^k u_2}{\partial x_1^k}\right|^2 (1-x_1^2)^{k+\beta_1}\prod_{i=2}^{3}(1-x_i^2)^{\beta_i}dx \\
&\le C\int_{R_0}\left(\left|\frac{\partial^k u}{\partial x_1^k}\right|^2|1-\varphi_\delta|^2 + \sum_{l=0}^{k-1}\left|\frac{\partial^l u}{\partial x_1^l}\right|^2\left|\frac{\partial^{k-l}\varphi_\delta}{\partial x_1^{k-l}}\right|^2\rho^{k+\sum_{i=1}^3\beta_i}|\sin\theta|^{k+\beta_1+\beta_2}\right)dx.
\end{aligned}$$

Due to (5.6) there hold

$$\begin{aligned}
&\int_{R_0}\left|\frac{\partial^k u}{\partial x_1^k}\right|^2|\varphi_\delta|^2\rho^{k+\sum_{i=1}^3\beta_i}|\sin\theta|^{k+\beta_1+\beta_2}dx \\
&\le C\int_{T_3}(1+x_3)^{2(\gamma-\sigma)-k+\beta_3}\, r^{2\sigma+1-k+\beta_1+\beta_2}\,dr\,dx_3 \\
&\le C\int_{\delta\cot\theta_0}^{\rho_0}\tilde{x}_3^{2(\gamma-\sigma)+\beta_3}\,d\tilde{x}_3\int_{\delta}^{\tilde{x}_3\tan\theta_0} r^{2\sigma+1-k+\beta_1+\beta_2}\,dr \\
&\le C(1+|\log\delta|)^{2\kappa}\delta^{2\gamma+3-k+\sum_{i=1}^3\beta_i}
\end{aligned} \tag{5.12}$$

and, for $l < k$,

$$\begin{aligned}
&\int_{R_0} \left|\frac{\partial^l u}{\partial x_1^l}\right|^2 \left|\frac{\partial^{k-l}\varphi_\delta}{\partial x_1^{k-l}}\right|^2 \rho^{k+\sum_{i=1}^3 \beta_i} |\sin\theta|^{k+\beta_1+\beta_2} dx \\
&\quad \le C\delta^{-2(k-l)} \int_{\delta\cot\theta_0}^{2\delta\cot\theta_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3}\, d\tilde{x}_3 \int_{\delta}^{\tilde{x}_3\tan\theta_0} r^{2(\sigma-l)+1+k+\beta_1+\beta_2} dr \\
&\quad \le C\delta^{2(\sigma+1)+\beta_1+\beta_2-k} \int_{\delta\cot\theta_0}^{2\delta\cot\theta_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3} d\tilde{x}_3 \\
&\quad \le C(1+|\log\delta|)^{2\kappa}\delta^{2\gamma+\sum_{i=1}^3\beta_i+3-k}.
\end{aligned} \tag{5.13}$$

A combination of (5.12) and (5.13) leads to

$$\int_Q \left|\frac{\partial^k u_2}{\partial x_1^k}\right|^2 \left(1-x_1^2\right)^{k+\beta_1} \prod_{i=1}^2 \left(1-x_i^2\right)^{\beta_i} dx \le C|\log\delta|^{2\kappa}\delta^{2\gamma+\sum_{i=1}^3\beta_i+3-k}.$$

The estimate on $\frac{\partial^k u_2}{\partial x_3^k}$ can be carried out similarly. Due to (5.6) there hold

$$\left|\frac{\partial^k u_2}{\partial x_3^k}\right| = \left|\varphi_\delta \frac{\partial^k u}{\partial x_3^k}\right| \le C(1+x_3)^{\gamma-\sigma-k}\, r^\sigma$$

and

$$\begin{aligned}
&\int_Q \left|\frac{\partial^k u_2}{\partial x_3^k}\right|^2 \prod_{i=1}^2 \left(1-x_i^2\right)^{\beta_i} \left(1-x_3^2\right)^{k+\beta_3} dx \\
&\le C\int_{T_3} \left|\frac{\partial^k u}{\partial x_3^k}\right|^2 \rho^{k+\sum_{i=1}^3\beta_i} |\sin\theta|^{\beta_1+\beta_2} dx \\
&\le C\int_\delta^{\rho_0\tan\theta_0} r^{2\sigma+1+\beta_1+\beta_2} dr \int_{r\cot\theta_0}^{\rho_0} \tilde{x}_3^{2(\gamma-\sigma)-k+\beta_3}\, d\tilde{x}_3 \\
&\le C\delta^{2\gamma+3-k+\sum_{i=1}^3\beta_i}.
\end{aligned}$$

We can treat all terms of $D^\alpha u_2$ with $|\alpha| \le k$ in a similar way, which gives for $k > 2\max\{\sigma, \gamma+1/2+\beta_3\}+2+\beta_1+\beta_2$

$$\|u_2\|^2_{H^{k,\beta}(Q)} \le C(1+|\log\delta|)^{2\kappa}\delta^{2\gamma+\sum_{i=1}^3\beta_i+3-k}. \tag{5.14}$$

Therefore, we have by (5.11) and (5.14)

$$\begin{aligned}
K(t,u) &= \inf_{u=v+w}\{\|v\|_{H^{0,\beta}(Q)} + t\,\|w\|_{H^{k,\beta}(Q)}\} \\
&\le C(\|u_1\|_{H^{0,\beta}(Q)} + t\,\|u_2\|_{H^{k,\beta}(Q)}) \\
&\le C(1+|\log\delta|)^\kappa \delta^{1+\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1/2+\beta_2/2}(1+t\,\delta^{-k/2}).
\end{aligned}$$

Selecting $\delta = t^{2/k}$, we have for $0 < t < 1$

$$K(t,u) \le C(1+|\log t|)^\kappa t^{\frac{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2}{k}}.$$

For $t \geq 1$, it always holds that

$$K(t,u) \leq C\|u\|_{H^{0,\beta}(Q)}.$$

Choosing $\theta = \frac{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2}{k}$, we have

$$\sup_{t>0} \frac{t^{-\theta}\, K(t,u)}{(1+|\log t|)^{\kappa}} \leq C,$$

which implies that $u \in \big(H^{0,\beta}(Q), H^{k,\beta}(Q)\big)_{\theta,\infty,\kappa} = B^{s,\beta}_{\kappa}(Q)$ with $s = \theta k = 2\min\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$ and κ given in (5.10).

Selecting $\theta = \frac{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2-\epsilon}{k} = \frac{s-\epsilon}{k}$ with $\epsilon > 0$ arbitrary gives for either $\sigma = \gamma + (1+\beta_3)/2$ or $\sigma \neq \gamma + (1+\beta_3)/2$

$$\int_0^{\infty} t^{-2\theta} |K(t,u)|^2 \frac{dt}{t} \leq C,$$

which implies that $u \in \big(H^{0,\beta}(Q), H^{k,\beta}(Q)\big)_{\theta,2} = H^{s-\epsilon,\beta}(Q)$. ∎

A combination of Theorems 5.2 and 2.3–2.4 leads to the approximability of the singular function of $\rho^{\gamma}\sin^{\sigma}\phi$-type.

THEOREM 5.3. *There exists $\psi(x) \in \mathcal{P}_p(Q)$ with $p > 1 + 2\min\{\sigma, \gamma + 1/2\}$ such that for $\beta = (0,0,0)$ and $s = 2 + 2\min\{\sigma, \gamma + 1/2\}$*

$$\|u - \psi\|_{L^2(Q)} \leq Cp^{-(2+2\min\{\sigma,\gamma+1/2\})}\, \|u\|_{B^{s,\beta}(Q)} \tag{5.15}$$

if $\sigma \neq \gamma + 1/2$, and

$$\|u - \psi\|_{L^2(Q)} \leq Cp^{-(2+2\min\{\sigma,\gamma+1/2\})}(1+\log p)^{1/2}\, \|u\|_{B^{s,\beta}_{1/2}(Q)}. \tag{5.16}$$

Also, there exists $\varphi(x) \in \mathcal{P}_p(Q)$ with $p > 2\min\{\sigma, \gamma + 1/2\}$ such that for $\beta = (-1/2, -1/2, 0)$ and $s = 1 + 2\min\{\sigma, \gamma + 1/2\}$

$$\|u - \varphi\|_{H^1(R_0)} \leq Cp^{-2\min\{\sigma,\gamma+1/2\}}\, \|u\|_{B^{s,\beta}(Q)} \tag{5.17}$$

if $\sigma \neq \gamma + 1/2$, and

$$\|u - \varphi\|_{H^1(R_0)} \leq Cp^{-2\sigma}(1+\log p)^{1/2}\, \|u\|_{B^{s,\beta}_{1/2}(Q)} \tag{5.18}$$

if $\sigma = \gamma + 1/2$.

Proof. For $\beta = (0,0,0)$, Theorem 5.2 indicates that $u \in B^{s,\beta}(Q)$ if $\sigma \neq \gamma + 1/2$ and $u \in B^{s,\beta}_{1/2}(Q)$ if $\sigma = \gamma + 1/2$ with $s = 2 + 2\min\{\sigma, \gamma + 1/2\}$. Due to Theorems 2.3–2.4, the Jacobi projection ψ of u associated with the weight $\beta = (0,0,0)$ on $\mathcal{P}_p(Q)$ with $p > 1 + 2\min\{\sigma, \gamma + 1/2\}$ satisfies (5.14) and (5.15).

Also for $\beta = (-1/2, -1/2, 0)$, Theorem 5.2 tells us that $u \in B^{s,\beta}(Q)$ if $\sigma \neq \gamma + 1/2$ and $u \in B^{s,\beta}_{1/2}(Q)$ if $\sigma = \gamma + 1/2$ with $s = 1 + 2\min\{\sigma, \gamma + 1/2\}$. Due to Theorems 2.3–2.4, the Jacobi projection φ of u associated with the weight $\beta = (-1/2, -1/2, 0)$ on $\mathcal{P}_p(Q)$ with $p > 2\min\{\sigma, \gamma + 1/2\}$ satisfies for $\ell = 0, 1$

$$|u - \varphi|_{H^{\ell,\beta}(Q)} \leq Cp^{-(2\min\{\sigma,\gamma+1/2\}+1-\ell)}\, \|u\|_{B^{s,\beta}(Q)} \tag{5.19}$$

if $\sigma \neq \gamma + 1/2$, and

$$|u - \varphi|_{H^{\ell,\beta}(Q)} \leq Cp^{-(2\min\{\sigma,\gamma+1/2\}+1-\ell)}(1+\log p)^{1/2}\, \|u\|_{B^{s,\beta}_{1/2}(Q)} \tag{5.20}$$

if $\sigma = \gamma + 1/2$. Note that

$$|u-\varphi|_{L^2(Q)} \le C|u-\varphi|_{H^{0,\beta}(Q)}. \tag{5.21}$$

Due to (5.3), there holds for $x \in R_0 = R_{\rho_0,\theta_0,\phi_0}$ and $|\alpha| = 1$

$$\prod_{i=1}^{2}(1-x_i^2)^{\alpha_i-1/2}\big(1-x_3^2\big)^{\alpha_3} \ge C_1, \tag{5.22}$$

where the positive constants C_1 is independent of x. This implies that for $|\alpha| = 1$

$$\begin{aligned}&\int_{R_0} |D^\alpha(u-\varphi)|^2\,dx\\ &\quad\le C\int_{R_0} |D^\alpha(u-\varphi)|^2 \prod_{i=1}^{2}\big(1-x_i^2\big)^{\alpha_i-1/2}\big(1-x_3^2\big)^{\alpha_3}\,dx\\ &\quad\le C|u-\varphi|^2_{H^{1,\beta}(Q)},\end{aligned}$$

which together with (5.19)–(5.21) leads to (5.17) and (5.18), completing the proof. □

5.2. Singular functions of $\rho^\gamma\ \log^\nu \rho\sin^\sigma\theta\log^\mu\sin\theta$-type. Since the function given in (5.2) can be written as

$$\begin{aligned}v(x) &= \rho^{\gamma-\sigma}\,r^\sigma\ \log^\nu\rho\,(\log\rho-\log r)^\mu\,\chi(\rho)\,\Psi(\theta)\,\Phi(\phi)\\ &= \rho^{\gamma-\sigma}\,r^\sigma\ \log^\nu\rho\,\chi(\rho)\,\Psi(\theta)\,\Phi(\phi)\sum_{l=0}^{\mu}\binom{\mu}{l}(-1)^{\mu-l}\log^l\rho\log^{\mu-l}r,\end{aligned}$$

we need to analyze the functions of this type

$$\begin{aligned}w(x) &= \rho^{\gamma-\sigma}\,r^\sigma\ \log^{\nu+l}\rho\,\log^{\mu-l}r\,\chi(\rho)\,\Psi(\theta)\,\Phi(\phi)\\ &= \rho^{\gamma-\sigma}\,r^\sigma\ \log^{\nu'}\rho\,\log^{\mu'}r\,\chi(\rho)\,\Psi(\theta)\,\Phi(\phi)\end{aligned}$$

with $\nu', \mu' \ge 0$.

THEOREM 5.4. *Let $\beta = (\beta_1,\beta_2,\beta_3)$ with $\beta_i > -1$, $1 \le i \le 3$. Then $w \in H^{s-\epsilon,\beta}(Q)$ and $w(x) \in B^{s,\beta}_{\kappa'}(Q)$ with $s = 2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2$, $\epsilon > 0$ arbitrary, and*

$$\kappa' = \begin{cases}\mu' & \text{if } \sigma < \gamma+(1+\beta_3)/2,\\ \mu'+\nu'+1/2 & \text{if } \sigma = \gamma+(1+\beta_3)/2,\\ \mu'+\nu' & \text{if } \sigma > \gamma+(1+\beta_3)/2.\end{cases} \tag{5.23}$$

Proof. We decompose the function into $w = w_1 + w_2$ with $w_1 = \varphi_\delta(r)u$ and $w_2 = (1-\varphi_\delta(r))w$, where $\varphi_\delta(r)$ is a C^∞ function defined as in the proof of Theorem 5.2. It is easy to verify that $w_1 \in H^{0,\beta}(Q)$ and $w_2 \in H^{k,\beta}(Q)$ for any $k > 2 + 2\max\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2$.

Let R_{ρ_0,θ_0}, T_i, $1 \le i \le 3$, be the regions defined as in the previous section and shown in Figure 5.2–5.3. There holds

(5.24)

$$\begin{aligned}\|w_1\|^2_{H^{0,\beta}(Q)} &= \int_{R_{\rho_0,\theta_0,\phi_0,}} |\varphi_\delta w|^2\rho^{\sum_{i=1}^3\beta_i}|\sin\theta|^{\beta_1+\beta_2}dx\\ &\le C\int_{T_1\cup T_2}(1+x_3)^{2(\gamma-\sigma)+\beta_3}\log^{2\nu'}(1+x_3)\,r^{2\sigma+1+\beta_1+\beta_2}\log^{2\mu'}r\,dr\,dx_3.\end{aligned}$$

Letting $\tilde{x}_3 = x_3 + 1$, as an analogue to estimates (5.7), we have

(5.25)
$$\begin{aligned}
&\int_{T_1} (1+x_3)^{2(\gamma-\sigma)+\beta_3} \log^{2\nu'}(1+x_3)\, r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr\, dx_3 \\
&\quad \le C \int_0^{2\delta\cot\theta_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3} \log^{2\nu'} \tilde{x}_3\, d\tilde{x}_3 \int_0^{\tilde{x}_3\tan\theta_0} r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr \\
&\quad \le C \int_0^{2\delta\cot\theta_0} \tilde{x}_3^{2\gamma+2+\sum_{i=1}^3 \beta_i} \log^{2(\nu'+\mu')} \tilde{x}_3\, d\tilde{x}_3 \\
&\quad \le C\delta^{2\gamma+\sum_{i=1}^3 \beta_i+3} (1+|\log\delta|)^{2(\nu'+\mu')}.
\end{aligned}$$

Analogously to (5.8)–(5.10) we have, for $\sigma \neq \gamma + (1+\beta_3)/2$,

(5.26)
$$\begin{aligned}
&\int_{T_2} (1+x_3)^{2(\gamma-\sigma)+\beta_3} \log^{2\nu'}(1+x_3)\, r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr\, dx_3 \\
&\quad \le C \int_0^{2\delta} r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr \int_{2\delta\cot\theta_0}^{\rho_0} \tilde{x}_3^{2(\gamma-\sigma)+\beta_3} \log^{2\nu'} \tilde{x}_3\, d\tilde{x}_3 \\
&\quad \le C(1+\delta^{2(\gamma-\sigma)+1+\beta_3} |\log\delta|^{2\nu'}) \delta^{2\sigma+2+\beta_1+\beta_2} |\log\delta|^{2\mu'} \\
&\quad \le C(\delta^{2\gamma+3+\sum_{i=1}^3 \beta_i} |\log\delta|^{2(\nu'+\mu')} + \delta^{2\sigma+2+\beta_1+\beta_2} |\log\delta|^{2\mu'}),
\end{aligned}$$

and, for $\sigma = \gamma + (1+\beta_3)/2$,

(5.27)
$$\begin{aligned}
&\int_{T_2} (1+x_3)^{2(\gamma-\sigma)+\beta_3} \log^{2\nu'}(1+x_3)\, r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr\, dx_3 \\
&\quad \le C \int_0^{2\delta} r^{2\sigma+1+\beta_1+\beta_2} \log^{2\mu'} r\, dr \int_{2\delta\cot\theta_0}^{\rho_0} \tilde{x}_3^{-1} \log^{2\nu'} \tilde{x}_3\, d\tilde{x}_3 \\
&\quad \le C(1+|\log\delta|)^{2\nu'+1} \delta^{2\sigma+2+\beta_1+\beta_2} |\log\delta|^{2\mu'} \\
&\quad \le C|\log\delta|^{2(\nu'+\mu')+1}\, \delta^{2\sigma+2+\beta_1+\beta_2}.
\end{aligned}$$

Combining (5.25)–(5.27) yields

$$\|w_1\|^2_{H^{0,\beta}(Q)} \le C|\log\delta|^{2\kappa'} \delta^{2+2\min\{\sigma,\gamma+(1+\beta_3)/2\}+\beta_1+\beta_2}. \tag{5.28}$$

Similarly we have the following estimate on $\|w_2\|^2_{H^{k,\beta}(Q)}$:

$$\|w_2\|^2_{H^{k,\beta}(Q)} \le C|\log\delta|^{2\kappa'} \delta^{2+2\min\{\sigma,\gamma+(1+\beta_3)/2\}+\beta_1+\beta_2-k}. \tag{5.29}$$

It follows from (5.28) and (5.29) that

$$K(t,w) \le C|\log\delta|^{\kappa'} \delta^{1+\min\{\sigma,\gamma+(1+\beta_3)/2\}+\beta_1/2+\beta_2/2} (1+t\delta^{-k/2}).$$

Selecting $\delta = t^{2/k}$ and $\theta = \frac{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2}{k}$, we have for $0<t<1$

$$\frac{t^{-\theta} K(t,u)}{(1+|\log t|)^{\kappa'}} \le C,$$

which implies the desired characterization of the singularity of the function $w(x)$ in the spaces $B^{s,\beta}_{\kappa'}(Q)$ with $s = 2 + 2\min\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$ and κ' given in (5.22).

Selecting $\theta = \frac{2+2\min\{\gamma+(1+\beta_3)/2,\sigma\}+\beta_1+\beta_2-\epsilon}{k} = \frac{s-\epsilon}{k}$ with $\epsilon > 0$ arbitrary, we have

$$\int_0^\infty t^{-2\theta}|K(t,u)|^2\frac{dt}{t} \le C,$$

which implies $u \in \big(H^{0,\beta}(Q), H^{k,\beta}(Q)\big)_{\theta,2} = H^{s-\epsilon,\beta}(Q)$. □

The following theorem on the characterization of singularity of the function $v(x)$ is a corollary of Theorem 5.4.

THEOREM 5.5. *Let $v(x)$ be given as in* (5.2)*, and let $\beta_i > -1, 1 \le i \le 3$. Then $w \in H^{s-\epsilon,\beta}(Q)$ and $v(x) \in B^{s,\beta}_\kappa(Q)$ with $s = 2 + 2\min\{\gamma + (1+\beta_3)/2, \sigma\} + \beta_1 + \beta_2$, $\epsilon > 0$ arbitrary, and*

$$\kappa = \begin{cases} \mu & \text{if } \sigma < \gamma + (1+\beta_3)/2, \\ \mu + \nu + 1/2 & \text{if } \sigma = \gamma + (1+\beta_3)/2, \\ \mu + \nu & \text{if } \sigma > \gamma + (1+\beta_3)/2. \end{cases} \tag{5.30}$$

Characterization of singularity of the function $v(x)$ by Theorem 5.5 *and the approximation property described in Theorem* 2.4 *give the approximability of $v(x)$.*

THEOREM 5.6. *Let $v(x)$ be given in* (5.2)*. Then there exists $\psi(x) \in \mathcal{P}_p(Q)$ with $p > 1 + 2\min\{\sigma, \gamma + 1/2\}$ such that for $\beta = (0,0,0)$ and $s = 2 + 2\min\{\sigma, \gamma + 1/2\}$*

$$\|v - \psi\|_{L^2(Q)} \le Cp^{-(2+2\min\{\sigma,\gamma+1/2\})}(1+\log p)^\kappa \, \|u\|_{B^{s,\beta}_\kappa(Q)}. \tag{5.31}$$

Also, there exists $\varphi(x) \in \mathcal{P}_p(Q)$ with $p > 2\min\{\sigma, \gamma + 1/2\}$ such that for $\beta = (-1/2, -1/2, 0)$ and $s = 1 + 2\min\{\sigma, \gamma + 1/2\}$

$$\|v - \varphi\|_{H^1(R_0)} \le Cp^{-2\min\{\sigma,\gamma+1/2\}}(1+\log p)^\kappa \, \|u\|_{B^{s,\beta}_\kappa(Q)}. \tag{5.32}$$

κ in (5.31) *and* (5.32) *is given in* (5.30).

Proof. By Theorem 5.5 $v(x) \in B^{2\min\{\gamma+1/2,\sigma\}+2,\beta}_\kappa(Q)$ with κ specified by (5.30), in particular, for $\beta = (0,0,0)$ and $\beta = (-1/2, -1/2, 0)$.

Applying Theorem 2.4 with $\beta = (0,0,0)$ leads immediately to (5.31). Applying Theorem 2.4 with $\beta = (-1/2, -1/2, 0)$ and arguing as in the proof of Theorem 5.3 we can easily obtain (5.32). Actually, ψ and φ are the Jacobi projection of v associated with the weight $\beta = (0,0,0)$ on $\mathcal{P}_p(Q)$ with $p > 1 + 2\min\{\sigma, \gamma + 1/2\}$ and with the weight $\beta = (-1/2, -1/2, 0)$ on $\mathcal{P}_p(Q)$ with $p > 2\min\{\sigma, \gamma + 1/2\}$, respectively. □

Remark 5.1. κ given in (5.30) reduces to (5.6) if $\nu = \mu = 0$. κ depends on ν and μ, but also on the relation between γ and σ. When $\sigma = \gamma + (1+\beta_3)/2$, $v(x) \in B^{s,\beta}_\kappa(Q)$ with κ increased by an extra value of $1/2$. Consequently, an extra loss of a factor $(1+\log p)^{1/2}$ occurs in the error estimate (5.31) and (5.32), which was mentioned in [16] for the p-version of BEM. Whether the extra value of $1/2$ can be removed or not is an open question for further investigation. Fortunately, the extra value of $1/2$ appears in κ, not in s.

6. Concluding remarks. The singularities of singular functions in three dimensions and their approximabilities have been analyzed in the framework of the Jacobi-weighted Besov and Sobolev spaces. To precisely characterize the singularities and investigate the approximabilities for singular functions of three different

types, Jacobi-weighted Besov and Sobolev spaces associated with three different Jacobi weights are elegantly designed. The most difficult as well as most significant work is the characterization of the functions with the singularity of $\rho^\gamma \log^\nu \rho \sin^\sigma \theta \log^\mu \sin\theta$-type in the Besov space $B^{s,\beta}_\kappa(Q)$ with κ given in (5.30). The singularity of this type is anisotropic and totally different from the singularity in two dimensions. The key for success is the decomposition of the singular function with a cutoff function $\varphi_\delta(r)$, instead of $\varphi_\delta(\rho)$ and $\varphi_\delta(\theta)$, although the singularity appears in ρ and θ. After having tried various decompositions we are convinced that only this decomposition can lead to our desired results. For the best approximation of these singular functions we select different weights, namely, $\beta = (-1/3, -1/3, -1/3)$, $\beta = (-1/2, -1/2, \beta_3)$, $\beta = (-1/2, -1/2, 0)$, respectively. We are also convinced that only this selection can give us the best error estimation in L^2- and H^1-norms. Once the weights are properly selected the approximation results follow in a natural way. Our approach for error estimation for singular functions is different from the usual approach, namely, we do not directly analyze approximation of singular functions, but verify that they belong to certain Jacobi-weighted Besov spaces.

TABLE 6.1

The value of k and s in Sobolev, Besov, and Jacobi-weighted Besov spaces for functions of $\rho^\gamma, r^\sigma, \rho^\gamma \sin^\sigma \theta$-type.

Space	$H^k(Q)$	$H^s(Q)$	$B^s(Q)$	$H^{k,\beta}(Q)$	$B^{s,\beta}(Q)$
ρ^γ	$3/2 + [\gamma]$	$3/2 + \gamma - \epsilon$	$3/2 + \gamma$	$2 + 2\gamma - \epsilon$	$2 + 2\gamma$
r^σ	$1 + [\sigma]$	$1 + \sigma - \epsilon$	$1 + \sigma$	$1 + 2\sigma - \epsilon$	$1 + 2\sigma$
$\rho^\gamma \sin^\sigma \theta$	$1 + [\lambda]$	$1 + \lambda - \epsilon$	$1 + \lambda$	$1 + 2\lambda - \epsilon$	$1 + 2\lambda$

TABLE 6.2

Accuracy of approximation in H^1-norm to singular functions of $\rho^\gamma, r^\sigma, \rho^\gamma \sin^\sigma \theta$-type by the h- and p-versions based on Sobolev, Besov, and Jacobi-weighted Besov spaces.

	h-version		p-version		
Space	$H^s(Q)$	$B^s(Q)$	$H^s(Q)$	$B^s(Q)$	$B^{s,\beta}(Q)$
ρ^γ	$h^{1/2+\gamma-\epsilon}$	$h^{1/2+\gamma+1/2}$	$p^{-(1/2+\gamma-\epsilon)}$	$p^{-(\gamma+1/2)}$	$p^{-(2\gamma+1)}$
r^σ	$h^{\sigma-\epsilon}$	h^σ	$p^{-(\sigma-\epsilon)}$	$p^{-\sigma}$	$p^{-2\sigma}$
$\rho^\gamma \sin^\sigma \theta$	$h^{\lambda-\epsilon}$	h^λ	$p^{-(\lambda-\epsilon)}$	$p^{-\lambda}$	$p^{-2\lambda}$

In Tables 6.1 and 6.2 $\lambda = \min\{\gamma + 1/2, \sigma\}$, $\sigma \neq \gamma + 1/2$, $\beta = (-1/3, -1/3, -1/3)$, $\beta = (-1/2, -1/2, \beta_3)$, $\beta = (-1/2, -1/2, 0)$ for ρ^γ, r^σ, and $\rho^\gamma \sin^\sigma \theta$, respectively.

Although the treatments for singular functions in three dimensions are quite different from and much more difficult than those in one and two dimensions, it is worth indicating that the structures of Jacobi-weighted spaces are basically the same. The difference lies only in the selection of Jacobi weights and in the way of proving that a singular function belongs to the Jacobi-weighted spaces. Hence the mathematical framework of the Jacobi-weighted Besov and Sobolev spaces is robust and uniform for problems in one, two, and three dimensions.

The singular functions with singularities of three different types are typical and appear in the solution of problems with piecewise analytical data on polyhedral domains, which govern the convergence of the finite element solutions of the h-, p-, and hp-versions (associated with quasi-uniform meshes). The function spaces used for

characterizing the singularities depend on the nature of singularities as well as the type of FEMs. Thus, the selection of function spaces is crucial to the best approximation for the finite element solutions. Tables 6.1 and 6.2 tell us how the functional spaces used for characterization of singularities and error analysis affect the estimation of approximation error measured in the H^1-norm. Hence we can conclude that the Jacobi-weighted Besov is the best theoretical tool for analyzing approximation of functions by the p- and hp-versions (associated with quasi-uniform meshes) of the FEM. Meanwhile, it can be shown that it has no substantial impact on the error estimation for the classical h-version of the FEM.

Finally, the framework we set up in three dimensions can be used for the spectral and the boundary element methods, and the analysis and results parallel to those for the finite element can be established for the spectral and the boundary element methods without substantial difficulties.

REFERENCES

[1] I. Babuška and B.Q. Guo, *Direct and inverse approximation theorems of the p-version of finite element method in the framework of weighted Besov spaces, Part* 1*: Approximability of functions in weighted Besov spaces,* SIAM J. Numer. Anal., 39 (2002), pp. 1512–1538.

[2] I. Babuška and B.Q. Guo, *Direct and inverse approximation theorems of the p-version of the finite element method in the framework of weighted Besov spaces, Part* 2*: Optimal convergence of the p-version of the finite element method*, Math. Models Methods Appl. Sci., 12 (2002), pp. 689–719.

[3] I. Babuška and B.Q. Guo, *Direct and Inverse Approximation Theorems of the p-Version of the Finite Element Method in the Framework of Weighted Besov Spaces, Part* 3*: Inverse Approximation Theorems*, TICAM report, 99-32, 1999.

[4] I. Babuška and B.Q. Guo, *Optimal estimates for lower and upper bounds of approximation errors in the p-version of the finite element method in two dimensions*, Numer. Math., 85 (2000), pp. 343–366.

[5] I. Babuška, M. Szabó, and N. Katz, *The p-version of the finite element method,* SIAM J. Numer. Anal., 18 (1981), pp. 515–545.

[6] I. Babuška and M. Suri, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1991), pp. 750–776.

[7] J. Bergh and J. Löfström, *Interpolation Spaces,* Springer-Verlag, Berlin, 1976.

[8] C. Bernardi and Y. Maday, *Spectral methods,* in Handbook of Numerical Analysis, Vol. 5, Part 2, P.G. Ciarlet and J.L. Lion, eds., Elsevier Science, Amsterdam, 1997, pp. 209–475.

[9] M.R. Dorr, *The approximation solutions of elliptic boundary value problems via the p-version of the finite element method,* SIAM J. Numer. Anal., 23 (1986), pp. 58–77.

[10] W. Gui and I. Babuška, *The h, p, and h-p versions of the finite element method in* 1 *dimension, Part* 1*: The error analysis of the p-version,* Numer. Math., 49 (1986), pp. 577–612.

[11] B.Q. Guo and N. Heuer, *The optimal convergence of the p-version of the boundary element method in two dimensions,* Numer. Math., 98 (2004), pp. 499–538.

[12] B.Q. Guo and N. Heuer, *The optimal convergence of the h-p version of the boundary element method in two dimensions,* Adv. Comput. Math., to appear.

[13] B.Y. Guo, *Jacobi approximation in certain Hilbert spaces and their applications to singular differential equations,* J. Math. Anal. Appl., 243 (2000), pp. 373–408.

[14] B.Y. Guo and L. Wang, *Jacobi approximation and Jacobi-Gauss-type interpolations in non-uniform Jacobi-weighted Sobolev spaces,* J. Approx. Theory, 128 (2004), pp. 1–41.

[15] R. Muñoz-Sola, *Polynomial liftings on a tetrahedron and applications to the h-p version of the finite element method in three dimensions,* SIAM J. Numer. Anal., 34 (1997), pp. 282–314.

[16] C. Schwab and M. Suri, *The optimal p-version approximation of singularities on polyhedra in the boundary element method,* SIAM J. Numer. Anal., 33 (1996), pp. 729–759.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 270–282

A PRIMAL-BASED PENALTY PRECONDITIONER FOR ELLIPTIC SADDLE POINT SYSTEMS*

C. R. DOHRMANN† AND R. B. LEHOUCQ‡

Abstract. A primal-based penalty preconditioner is presented for a linear set of equations arising from elliptic saddle point problems. We show that the eigenvalues of the preconditioned matrix are positive real and demonstrate that a variant of the preconditioner can be combined with the conjugate gradient algorithm. Our approach is motivated by two basic observations. First, the solution of a problem with constraints is often similar to the solution of a problem where the constraints are penalized. Second, certain methods of solution not available for a constrained problem are possible for its penalized counterpart so motivating a primal-based Schur complement approach. Numerical examples for elliptic two- and three-dimensional problems are presented that confirm theoretical results and demonstrate the effectiveness of the preconditioner.

Key words. constrained minimization, saddle point systems, mixed finite elements, Stokes problems, block preconditioners

AMS subject classifications. 65F15, 65N25, 65N30, 65N22, 65M60, 65N55, 65M55

DOI. 10.1137/040619016

1. Introduction.

We consider linear systems

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{1.1}$$

arising from finite element discretizations of saddle point problems. The matrix A is assumed to be symmetric and positive definite on the kernel of B, and C to be symmetric and positive semidefinite. We denote $u \in \mathbb{R}^n$ and $p \in \mathbb{R}^m$ the primal and dual vectors. For instance, in the case of Stokes flow and incompressible elasticity, the primal and dual variables are associated with velocity–pressure and displacement–pressure, respectively.

Several preconditioners have been investigated for specific instances of (1.1). For example, when the saddle point system represents a discrete Stokes or elasticity system, many approaches precondition the dual Schur complement $C + BA^{-1}B^T$ with a matrix spectrally equivalent to the dual mass matrix. Examples include block diagonal preconditioners [12, 25, 17], block triangular preconditioners [11, 16], and inexact Uzawa approaches [6, 9, 4]. A symmetric preconditioner for saddle point systems similar in form to ours that involves preconditioning of the dual Schur complement was recently studied in [14]. Reformulation of the saddle point problem in (1.1) as a symmetric positive definite system was considered in [5] that permits an iterative solution using the conjugate gradient algorithm. See [3, 10] for further information and

*Received by the editors November 16, 2004; accepted for publication (in revised form) August 29, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sinum/44-1/61901.html
†Sandia National Laboratories, P.O. Box 5800, MS 0847, Albuquerque, NM 87185-0847 (crdohrm@sandia.gov).
‡Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110 (rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

references on saddle point systems and their origin. Overlapping Schwarz preconditioners involving solutions of both local and coarse saddle point problems were investigated in [15, 22]. More recently, substructuring preconditioners based on balancing Neumann–Neumann methods [21, 13] and FETI-DP [18] were studied. In contrast to much previous work, we consider a preconditioner for a generic saddle point system that satisfies the minimal conditions stated after (1.1) that preconditions the primal Schur complement.

Our approach builds on the idea of preconditioning indefinite problems using a regularization approach [1] introduced by Axelsson. Preconditioning based on regularization is motivated by the idea that the solution of a penalized problem is close to that of the original constrained problem. We present theory and numerical results that extends [1] to cases where the penalized primal Schur complement $A+B^T\tilde{C}^{-1}B$ is preconditioned rather than factored directly. Here, $\tilde{C}$ is a symmetric positive definite penalty counterpart of C in (1.1).

The primal-based penalty (PBP) preconditioner and accompanying theory are presented in sections 2 and 3, respectively. The preconditioner presented is indefinite, but all the eigenvalues of the preconditioned system are real and positive provided that several mild assumptions are satisfied. A form of the preconditioner suited for conjugate gradients is presented in section 4. Numerical examples presented in section 5 confirm the theory and demonstrate the excellent performance of the preconditioner. Comparisons are also made with other preconditioners for saddle point systems. Some concluding remarks are made in the final section.

2. Preconditioner. The penalized primal Schur complement S_A is defined as

$$S_A = A + B^T\tilde{C}^{-1}B, \tag{2.1}$$

where $\tilde{C}$ is symmetric and positive definite. Since A is assumed to be positive definite on the kernel of B, it follows that S_A is positive definite. We consider a preconditioner $\mathcal{M}$ of the form

$$\mathcal{M} = \begin{bmatrix} I & B^T\tilde{C}^{-1} \\ 0 & -I \end{bmatrix} \begin{bmatrix} \hat{S}_A & 0 \\ 0 & -\tilde{C} \end{bmatrix} \begin{bmatrix} I & 0 \\ \tilde{C}^{-1}B & -I \end{bmatrix}, \tag{2.2}$$

where $\hat{S}_A$ is an approximation of S_A. The action of the preconditioner on a vector r (with primal and dual subvectors r_u and r_p) is

$$\begin{bmatrix} z_u \\ z_p \end{bmatrix} = \begin{bmatrix} I & 0 \\ \tilde{C}^{-1}B & -I \end{bmatrix} \begin{bmatrix} \hat{S}_A^{-1} & 0 \\ 0 & -\tilde{C}^{-1} \end{bmatrix} \begin{bmatrix} I & B^T\tilde{C}^{-1} \\ 0 & -I \end{bmatrix} \begin{bmatrix} r_u \\ r_p \end{bmatrix}, \tag{2.3}$$

leading to the two step application of $\mathcal{M}^{-1}r$ as

1. solve $\hat{S}_A z_u = r_u + B^T\tilde{C}^{-1}r_p$ for z_u,
2. solve $\tilde{C}z_p = Bz_u - r_p$ for z_p.

Each application of the preconditioner requires two solves with $\tilde{C}$ and one solve with $\hat{S}_A$.

3. Analysis of the preconditioner. We now investigate the eigenvalues ν of the generalized eigenproblem

$$\mathcal{A}z = \nu\mathcal{M}z, \tag{3.1}$$

where $\mathcal{A}$ is the coefficient matrix in (1.1). Using a coordinate transformation, these eigenvalues are identical to those of the generalized eigenproblem

$$\mathcal{A}\mathcal{M}^{-1}\mathcal{H}w = \nu\mathcal{H}w, \tag{3.2}$$

where $\mathcal{H}$ is defined as

$$\mathcal{H} = \begin{bmatrix} S_A - \hat{S}_A & 0 \\ 0 & \tilde{C} - C \end{bmatrix}. \tag{3.3}$$

The following lemma and corollary are needed for the main theorem of our paper.

LEMMA 3.1. *Suppose that D and E are symmetric positive definite matrices of order n. If*

$$1 < \rho_1 \leq \frac{w^T D w}{w^T E w} \leq \rho_2$$

for all $w \in \mathbb{R}^n$, then

$$\rho_1 \leq \frac{w^T(DE^{-1}D - D)w}{w^T(D-E)w} \leq \rho_2, \tag{3.4}$$

$$\rho_1 - 1 \leq \frac{w^T(D-E)E^{-1}(D-E)w}{w^T(D-E)w} \leq \rho_2 - 1, \tag{3.5}$$

$$1 - 1/\rho_1 \leq \frac{w^T(D-E)E^{-1}(D-E)w}{w^T(DE^{-1}D - D)w} \leq 1 - 1/\rho_2\,, \tag{3.6}$$

$$1 - 1/\rho_1 \leq \frac{w^T(E^{-1} - D^{-1})w}{w^T E^{-1} w} \leq 1 - 1/\rho_2. \tag{3.7}$$

Proof. If (x, λ) is the largest eigenpair of the matrix pencil (D, E), then

$$\max_w \frac{w^T D w}{w^T E w} = \frac{x^T D x}{x^T E x} \leq \rho_2.$$

Therefore, if we can show that the upper bounds of (3.4) are maximized with x, then the upper bounds are true for all $w \in \mathbb{R}^n$. Because $Dx = Ex\lambda$, then $(DE^{-1}D - D)x = Dx(\lambda - 1)$ and $(D - E)x = Ex(\lambda - 1)$ easily follow. If we premultiply these last two identities by x^T, then

$$x^T(DE^{-1}D - D)x = \frac{x^T D x}{x^T E x} x^T(D-E)x \leq \rho_2 x^T(D-E)x$$

and the upper bound of (3.4) is established. A similar argument with the minimal eigenpair (y, σ) of (D, E) establishes the lower bound.

The identity $(D - E)x = Ex(\lambda - 1)$ implies $x^T(D-E)E^{-1}(D-E)x = x^T(D - E)x(\lambda - 1)$ and so (3.5) is established using a similar argument as used for (3.4). Likewise, (3.6) follows from the identity $x^T(D-E)E^{-1}(D-E)x = x^T(DE^{-1}D - D)x(1 - 1/\lambda)$.

The identity $x^T DE^{-1}Dx = x^T DD^{-1}Dx\lambda$ follows from $Dx = Ex\lambda$ and implies

$$z^T(E^{-1} - D^{-1})z = z^T D^{-1} z(\lambda - 1) = z^T E^{-1} z \left(\frac{\lambda - 1}{\lambda}\right) \leq z^T E^{-1} z \left(\frac{\rho_2 - 1}{\rho_2}\right),$$

where $z = Dx$. The last set of inequalities (3.7) now easily follows. ∎

Using arguments similar to those for the proof of (3.4) in Lemma (3.1) we also have the following result.

COROLLARY 3.2. *Suppose that D is symmetric positive semidefinite and E is symmetric positive definite of order m. If*

$$0 \le \rho_1 \le \frac{w^T D w}{w^T E w} \le \rho_2 < 1$$

for all $w \in \mathbb{R}^m$, then

$$\rho_2 w^T(D-E)w \le w^T(DE^{-1}D - D)w \le \rho_1 w^T(D-E)w. \tag{3.8}$$

THEOREM 3.3. *If $\alpha_1 > 1,\ 0 \le \beta_1 < \beta_2 < 1,\ \gamma_1 > 0$, and*

$$\alpha_1 x^T \hat{S}_A x \le x^T S_A x \le \alpha_2 x^T \hat{S}_A x \quad \forall x \in \mathbb{R}^n, \tag{3.9}$$

$$\beta_1 y^T \tilde{C} y \ \le y^T C y \le \ \beta_2 y^T \tilde{C} y \quad \forall y \in \mathbb{R}^m, \tag{3.10}$$

$$\gamma_1 y^T B \hat{S}_A^{-1} B^T y \ \le y^T \tilde{C} y \le \ \gamma_2 y^T B \hat{S}_A^{-1} B^T y \quad \forall y \in \mathbb{R}^m, \tag{3.11}$$

and

$$0 < y^T \tilde{C} y \quad \forall y (\neq 0) \in \mathbb{R}^m, \tag{3.12}$$

then the eigenvalues of (3.2) *satisfy*

$$\delta_1 \le \nu \le \delta_2,$$

where

$$\delta_1 = \min\{\sigma_2(\alpha_1/\alpha_2), \beta_1 + \sigma_1(1-\beta_2)(\alpha_2\gamma_2)^{-1}\},$$

$$\delta_2 = \max\{2\alpha_2 - \sigma_2, \beta_2 + (1-\beta_1)(2 - \sigma_1/\alpha_2)\gamma_1^{-1}\}$$

and σ_1, σ_2 are arbitrary positive constants that satisfy $\sigma_1 + \sigma_2 = 1$.

Proof. A direct calculation gives

$$\mathcal{A}\mathcal{M}^{-1}\mathcal{H} = \begin{bmatrix} S_A \hat{S}_A^{-1} S_A - S_A & (S_A \hat{S}_A^{-1} - I) B^T P^T \\ PB(\hat{S}_A^{-1} S_A - I) & \mathcal{F} \end{bmatrix}, \tag{3.13}$$

where I is the identity matrix and

$$P = I - C\tilde{C}^{-1}, \quad \mathcal{F} = PB\hat{S}_A^{-1}B^T P^T + C - C\tilde{C}^{-1}C. \tag{3.14}$$

If $z^T = \begin{bmatrix} x^T & y^T \end{bmatrix}$, then the quadratic form $z^T \mathcal{A}\mathcal{M}^{-1}\mathcal{H} z$ contains the terms $x^T(S_A\hat{S}_A^{-1}S_A - S_A)x$, $y^T\mathcal{F}y$ and $y^T PB(\hat{S}_A^{-1}S_A - I)x$. We first estimate the absolute value of this last term.

If σ_1 and σ_2 are positive constants so that $\sigma_1 + \sigma_2 = 1$, then

$$\begin{aligned}
|y^T PB(\hat{S}_A^{-1} S_A - I)x| &= |\sigma_1 y^T PBS_A^{-1}(S_A \hat{S}_A^{-1} S_A - S_A)x + \sigma_2 y^T PB\hat{S}_A^{-1}(S_A - \hat{S}_A)x| \\
&\leq \sigma_1 |y^T PBS_A^{-1}(S_A \hat{S}_A^{-1} S_A - S_A)^{1/2}(S_A \hat{S}_A^{-1} S_A - S_A)^{1/2} x| \\
&\quad + \sigma_2 |y^T PB\hat{S}_A^{-1/2} \hat{S}_A^{-1/2}(S_A - \hat{S}_A)x| \\
&\leq \sigma_1 (y^T PB(\hat{S}_A^{-1} - S_A^{-1})B^T P^T y)^{1/2}(x^T(S_A \hat{S}_A^{-1} S_A - S_A)x)^{1/2} \\
&\quad + \sigma_2 (y^T PB\hat{S}_A^{-1} B^T P^T y)^{1/2}(x^T(S_A - \hat{S}_A)\hat{S}_A^{-1}(S_A - \hat{S}_A)x)^{1/2} \\
&\leq \frac{\sigma_1}{2} y^T PB(\hat{S}_A^{-1} - S_A^{-1})B^T P^T y + \frac{\sigma_1}{2} x^T(S_A \hat{S}_A^{-1} S_A - S_A)x \\
&\quad + \frac{\sigma_2}{2} y^T PB\hat{S}_A^{-1} B^T P^T y + \frac{\sigma_2}{2} x^T(S_A - \hat{S}_A)\hat{S}_A^{-1}(S_A - \hat{S}_A)x,
\end{aligned}$$

where we used the Cauchy–Schwarz inequality and the arithmetic mean inequality. Applications of Lemma 3.1, Corollary 3.2, and our hypothesis give

$$\begin{aligned}
2|y^T PB(\hat{S}_A^{-1} S_A - I)x| \quad &\leq \quad (1-\beta_1)(\gamma_1\alpha_2)^{-1}(\alpha_2 - \sigma_1) y^T(\tilde{C} - C)y \\
&\quad\quad + (\alpha_2 - \sigma_2) x^T(S_A - \hat{S}_A)x,
\end{aligned} \tag{3.15}$$

where we used $y^T P\tilde{C}P^T y \leq (1-\beta_1) y^T(\tilde{C} - C)y$, and

$$\begin{aligned}
(\beta_1 + (1-\beta_2)\gamma_2^{-1}) y^T(\tilde{C} - C)y \quad &\leq y^T \mathcal{F} y \leq \quad (\beta_2 + (1-\beta_1)\gamma_1^{-1}) y^T(\tilde{C} - C)y.
\end{aligned} \tag{3.16}$$

Similarly, we have

$$\begin{aligned}
-2|y^T PB(\hat{S}_A^{-1} S_A - I)x| \quad &\geq \quad -(1-\beta_2)\gamma_2^{-1}(\sigma_2 + \sigma_1(1 - 1/\alpha_2)) y^T(\tilde{C} - C)y \\
&\quad\quad - \alpha_1(\sigma_1 + \sigma_2(1 - 1/\alpha_2)) x^T(S_A - \hat{S}_A)x.
\end{aligned} \tag{3.17}$$

From (3.15)–(3.17) and Lemma 3.1 we obtain

$$\begin{aligned}
z^T \mathcal{A}\mathcal{M}^{-1}\mathcal{H} z &\leq x^T(S_A \hat{S}_A^{-1} S_A - S_A)x + 2|y^T PB(\hat{S}_A^{-1} S_A - I)x| + y^T \mathcal{F} y \\
&\leq (2\alpha_2 - \sigma_2) x^T(S_A - \hat{S}_A)x \\
&\quad + (\beta_2 + (1-\beta_1)(2 - \sigma_1/\alpha_2)\gamma_1^{-1}) y^T(\tilde{C} - C)y
\end{aligned}$$

and

$$\begin{aligned}
z^T \mathcal{A}\mathcal{M}^{-1}\mathcal{H} z &\geq x^T(S_A \hat{S}_A^{-1} S_A - S_A)x - 2|y^T PB(\hat{S}_A^{-1} S_A - I)x| + y^T \mathcal{F} y \\
&\geq \sigma_2(\alpha_1/\alpha_2) x^T(S_A - \hat{S}_A)x \\
&\quad + (\beta_1 + \sigma_1(1-\beta_2)(\alpha_2\gamma_2)^{-1}) y^T(\tilde{C} - C)y,
\end{aligned}$$

where we also used Corollary (3.2). Because $z \in \mathbb{R}^{n+m}$ is arbitrary, our theorem is established. ∎

We now briefly discuss the theorem. If the bounds (3.9)–(3.11) are independent of the discretization parameters, then bounds on the eigenvalues of the preconditioned

system are uniform with respect to the discretization. The difference of positive definite matrices that appears in the leading block of $\mathcal{H}$ can be used to define an inner product and has appeared in [5, 4, 16]. The theory developed here applies to both continuous and discontinuous interpolation of the dual variable in each element, but in practice the preconditioner is limited to the discontinuous case. The use of discontinuous interpolation in the dual variable results in a block diagonal $\tilde{C}$ so that S_A has the same sparsity structure as A.

We conclude this section with a brief discussion of how assumption (3.11) follows from more familiar bounds, e.g., for Stokes problems, if A is positive definite. Assume

$$\zeta_1 y^T B A^{-1} B^T y \leq y^T \tilde{C} y \leq \zeta_2 y^T B A^{-1} B^T y \quad \forall y \in \mathbb{R}^m,$$

where ζ_1 and ζ_2 are mesh independent positive constants. From (2.1) we obtain

$$B S_A^{-1} B^T = B A^{-1} B^T - B A^{-1} B^T (\tilde{C} + B A^{-1} B^T)^{-1} B A^{-1} B^T$$

from which follows the bounds

$$1/(1+\zeta_2) y^T \tilde{C} y \leq y^T B S_A^{-1} B^T y \leq 1/(1+\zeta_1) y^T \tilde{C} y \quad \forall y \in \mathbb{R}^m.$$

Thus, $\tilde{C}$ and $BS_A^{-1}B^T$ are spectrally equivalent. Assumption (3.11) then follows from the transitivity of spectral equivalence and (3.9). The constants γ_1 and γ_2 are independent of mesh parameters and material properties provided the same is true for α_1 and α_2.

4. Preconditioned conjugate gradient algorithm. We now consider a form of the preconditioner suitable for the conjugate gradient algorithm. The original linear system (1.1) can be expressed compactly as

$$\mathcal{A} w = d,$$

where

$$w = \begin{bmatrix} u \\ p \end{bmatrix} \quad \text{and} \quad d = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

The associated residual r is defined as

$$r = d - \mathcal{A} w.$$

We have shown in Theorem 3.3 that $\mathcal{H}\mathcal{M}^{-1}\mathcal{A}$ is symmetric and positive definite. Furthermore, if the constants in (3.9)–(3.11) are mesh independent, then $\mathcal{H}\mathcal{M}^{-1}\mathcal{A}$ is spectrally equivalent to $\mathcal{H}$. These facts motivate using the conjugate gradient algorithm to solve the equivalent linear system

$$\tilde{\mathcal{A}} w = \tilde{d},$$

where

$$\tilde{\mathcal{A}} = \mathcal{H}\mathcal{M}^{-1}\mathcal{A} \quad \text{and} \quad \tilde{d} = \mathcal{H}\mathcal{M}^{-1} d$$

using $\mathcal{H}$ as a preconditioner. The preconditioned conjugate gradient algorithm for the equivalent linear system is summarized as follows:

1. $w_0 = 0$, $r_0 = d$, $z_0 = \mathcal{M}^{-1} r_0$, $\tilde{r}_0 = \mathcal{H}\mathcal{M}^{-1} r_0$, and $k = 1$.

2. If the norm of r_{k-1} is less than a specified value, then exit. Otherwise,
3. $\beta_k = (z_{k-1}^T \tilde{r}_{k-1})/(z_{k-2}^T \tilde{r}_{k-2}) \qquad (\beta_1 = 0)$.
4. $p_k = z_{k-1} + \beta_k p_{k-1} \qquad (p_1 = z_0)$.
5. $\alpha_k = (z_{k-1} \tilde{r}_{k-1})/(p_k^T \mathcal{H}\mathcal{M}^{-1}\mathcal{A} p_k)$.
6. $w_k = w_{k-1} + \alpha_k p_k$.
7. $r_k = r_{k-1} - \alpha_k \mathcal{A} p_k$.
8. $z_k = z_{k-1} - \alpha_k \mathcal{M}^{-1}\mathcal{A} p_k$.
9. $\tilde{r}_k = \tilde{r}_{k-1} - \alpha_k \mathcal{H}\mathcal{M}^{-1}\mathcal{A} p_k$.
10. Return to Step 2.

The conjugate gradient algorithm described above is somewhat nonstandard in that two additional recurrences appear in steps 7 and 8. Application of the algorithm requires calculations of the form $\mathcal{M}^{-1}a$ and $\mathcal{H}\mathcal{M}^{-1}a$. For $a^T = \begin{bmatrix} a_u^T & a_p^T \end{bmatrix}$ we see that

$$\mathcal{M}^{-1}a = \begin{bmatrix} b_u \\ b_p \end{bmatrix} = \begin{bmatrix} \hat{S}_A^{-1}(a_u + B^T \tilde{C}^{-1} a_p) \\ \tilde{C}^{-1}(B b_u - a_p) \end{bmatrix}$$

and

$$\mathcal{H}\mathcal{M}^{-1}a = \begin{bmatrix} S_A b_u - (a_u + B^T \tilde{C}^{-1} a_p) \\ B b_u - a_p - C b_p \end{bmatrix}.$$

Notice that no calculations involving $\hat{S}_A$ are required. In addition, r_k is the residual of the original linear system at iteration k and can be used to assess convergence.

We remark that Bramble and Pasciak [5] also preconditioned an elliptic saddle point problem into a symmetric and positive definite system that could then be solved using the conjugate gradient algorithm. Klawonn [16] also generated a positive definite preconditioned system using a nonstandard inner product but did not consider a practical implementation.

5. Numerical examples. In this section, (1.1) is solved to a relative residual tolerance of 10^{-6} using both right preconditioned GMRES [24] and preconditioned conjugate gradients (PCG) for some example incompressible elasticity problems. The shear modulus G and Lamé parameter λ for an isotropic material are related to the elastic modulus E and Poisson ratio ν by

$$G = \frac{E}{2(1+\nu)}, \qquad \lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}.$$

For incompressible problems $\nu = 1/2$ and λ is infinite. All the examples in this section use $G = 1$ and $\nu = 1/2$. The preconditioner $\hat{S}_A$ used for S_A is a version of the balancing domain decomposition by constraints (BDDC) preconditioner described in [7] and analyzed in [20] that is well suited to nearly incompressible problems [8]. We note that if the original BDDC preconditioner is used rather than the modified one in [8], then the performance of the PBP preconditioner is sensitive to small changes in ν near 1/2 (see Table 5.1). The penalty matrix $\tilde{C}$ is chosen as the negative (2,2) block of the coefficient matrix in (1.1) for an identical problem with the same shear modulus but a value of ν less than 1/2.

Regarding assumption (3.9), we note that the substructuring preconditioner used for S_A has the attractive property that $\alpha_1 \geq 1$ and α_2 is mesh independent under certain additional assumptions [19]. For the conjugate gradient algorithm we scale

the preconditioned residual associated with the primal Schur complement by 1.01 as a safeguard to ensure that $\mathcal{H}$ is positive definite. Since $\tilde{C}$ is positive definite and $C = 0$, (3.10) holds with $\beta_2 < 1$. Finally, since we use stable finite elements and $\tilde{C}$ is spectrally equivalent to the pressure mass matrix, it follows from the discussion at the end of section 3 that (3.11) holds as well.

For purposes of comparison, we also present results for block diagonal and block triangular preconditioners for (1.1). Given the primal and dual residuals r_u and r_p, the preconditioned residuals z_u and z_p for the block diagonal preconditioner are given by

$$z_u = M_A^{-1} r_u \quad \text{and} \quad z_p = M_p^{-1} r_p,$$

where M_A is the BDDC preconditioner for A and M_p is the dual mass matrix. Note that the shear modulus G was chosen as 1 to obtain proper scaling of z_p. Similarly, the preconditioned residuals for the block triangular preconditioner are given by

$$z_p = -M_p^{-1} r_p \quad \text{and} \quad z_u = M_A^{-1}(r_u - B^T z_p).$$

We note that the majority of computations for the block preconditioners occur in forming and applying the BDDC preconditioner for A. Thus, the setup time and time for each iteration are nearly the same for the PBP preconditioner and the two block preconditioners.

We also make comparisons with the preconditioning technique of Bramble and Pasciak [5]. In Remark 2 of [5], Bramble and Pasciak assume that in the case of the Stokes equation the matrix

$$\begin{bmatrix} I & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}$$

is well conditioned. This is not true in our examples because the dual mass matrix M_p, which is spectrally equivalent to $BA^{-1}B^T$, is not well conditioned. The poor conditioning of M_p is a result of using discontinuous linear interpolation of the dual variable. Let the Cholesky decomposition of M_p be given by $R^T R$. By introducing the change of variables $p = R^{-1}\tilde{p}$, the matrix above becomes

$$\begin{bmatrix} I & 0 \\ 0 & R^{-T}BA^{-1}B^T R^{-1} \end{bmatrix},$$

which is well conditioned. Thus, we make use of the noted change of variables in our implementation of [5]. As is done for the PBP preconditioner, we scale the preconditioned residual associated with M_A by 1.01 to ensure that assumption (2.2) of [5] is satisfied.

The first set of examples is for a two-dimensional plane strain problem on a unit square with all displacement degrees of freedom (dofs) on the boundary constrained to zero. The entries of the right-hand side vector b were chosen as uniformly distributed random numbers in the range from 0 to 1. This choice of b ensures that the nodal forces in the problem have a significant average component without not being overly smooth. For any given mesh, the same values in b are used irrespective of the preconditioner used. For this simple geometry the finite element mesh consists of stable $Q_2 - P_1$ elements. This element uses biquadratic interpolation of displacement and discontinuous linear interpolation of pressure. In two dimensions the element has nine nodes for displacement and three element pressure dofs, while in three dimensions it

TABLE 5.1

Iterations needed to solve incompressible two-dimensional plane strain problem using the PBP preconditioner. Column ν indicates the values of ν used to define $\tilde{C}$. Results are shown for both the modified and original BDDC preconditioners for S_A. Results in parentheses are condition number estimates from PCG.

ν	Modified BDDC		Original BDDC	
	GMRES	PCG	GMRES	PCG
0.3	19	22 (16)	19	22 (16)
0.4	15	17 (7.3)	15	17 (7.3)
0.49	11	12 (3.2)	13	13 (3.7)
0.499	10	10 (3.0)	17	19 (8.6)
0.4999	9	10 (2.9)	24	31 (71)
0.49999	9	10 (2.9)	25	44 (697)

has 27 nodes for displacement and four element pressure dofs. In the final example, the stable $P_2^+ - P_1$ element is used for an unstructured mesh in two dimensions. This triangular element uses quadratic interpolation of displacement with an added bubble function and discontinuous linear interpolation of pressure. Descriptions of the $Q_2 - P_1$ and $P_2^+ - P_1$ discontinuous pressure elements can be found in [2].

Results are shown in Table 5.1 for the PBP preconditioner applied to a problem discretized by a 32×32 arrangement of square elements. Condition number estimates of the preconditioned equations are shown in parentheses for the PCG results. These eigenvalue estimates were obtained using the connection between conjugate gradients and the Lanczos method. The BDDC preconditioner is based on a regular decomposition of the mesh into 16 square substructures. Notice that the results are insensitive to changes in ν near the incompressible limit of 1/2 for the modified BDDC preconditioner used in this study.

Table 5.2 shows results for a growing number of substructures with $H/h = 4$, where H and h are the substructure and element lengths, respectively. A small growth in the number of iterations with problem size is evident in the table for all the preconditioners. Notice that the iterations required by PCG are only slightly larger than those for GMRES. The primary advantage of PCG over GMRES is the three-term recurrence that requires less storage. In contrast, GMRES uses a long recurrence and so has increased storage requirements. The PBP preconditioner is competitive with the other preconditioners. Similar results for related three-dimensional problems are shown in Tables 5.3 and 5.4.

The next example is for the unstructured mesh shown in Figure 5.1. The mesh was generated using the Matlab code described in [23]. The original triangular elements were then converted to $P_2^+ - P_1$ elements by adding midedge and center nodes. All displacements are constrained to zero on the mesh boundary for this example. The mesh has 2937 elements and 9046 nodes. Results in Table 5.5 show that the PBP preconditioner is competitive with the other approaches.

In Table 5.6 we present selected results from Tables 5.2, 5.4, and 5.5, where the linear systems involving S_A or A are solved exactly rather than approximately using the BDDC preconditioner. In addition, the results are for a relative residual tolerance reduced to 10^{-9}. Column ν reports the value of Poisson ratio used to define $\tilde{C}$ in the PBP preconditioner. Notice for values of ν away from 1/2 that the PBP preconditioner has no clear advantage over the block triangular preconditioner. This is not the case for $\nu = 0.49999$, where the PBP preconditioner with GMRES iterations

TABLE 5.2
Iterations needed to solve incompressible plane strain problems with increasing numbers of substructures (N) and $H/h = 4$. The value of ν used to define $\tilde{C}$ in the PBP preconditioner is 0.49999.

N	PBP		Block diagonal	Block triangular	Bramble–Pasciak
	GMRES	PCG	GMRES	GMRES	PCG
4	6	7 (2.0)	26	16	20
16	8	9 (2.4)	30	20	24
36	9	10 (2.9)	35	23	26
64	9	10 (3.2)	38	26	29
100	10	11 (3.3)	40	28	30
144	10	11 (3.4)	42	29	32
196	10	12 (3.4)	45	30	35
256	10	12 (3.5)	47	30	35

TABLE 5.3
Iterations needed to solve incompressible three-dimensional elasticity problem using the PBP preconditioner. The cube domain is discretized by 512 *elements and partitioned into* 64 *substructures for the BDDC preconditioner. Results for different values of ν used to define $\tilde{C}$ are shown.*

ν	GMRES	PCG
0.3	21	25 (12)
0.4	15	17 (6.1)
0.49	13	14 (3.8)
0.499	15	16 (5.0)
0.4999	17	19 (6.5)
0.49999	18	21 (6.8)

TABLE 5.4
Iterations needed to solve three-dimensional incompressible elasticity problems with increasing numbers of substructures (N) and $H/h = 2$. The value of ν used to define $\tilde{C}$ in the PBP preconditioner is 0.49.

N	PBP		Block diagonal	Block triangular	Bramble–Pasciak
	GMRES	PCG	GMRES	GMRES	PCG
8	10	11 (3.2)	36	21	23
27	13	14 (3.8)	41	22	26
64	13	14 (3.8)	46	25	30
125	13	14 (3.9)	48	25	30
216	13	14 (3.9)	51	26	32

TABLE 5.5
Iterations needed to solve two-dimensional incompressible plane strain problem shown in Figure 5.1. *The value of ν used to define $\tilde{C}$ in the PBP preconditioner is* 0.49999.

PBP		Block diagonal	Block triangular	Bramble–Pasciak
GMRES	PCG	GMRES	GMRES	PCG
11	11 (3.6)	53	35	44

exhibits a tenfold reduction in iterations.

One possible objection to the PBP preconditioner is the need for a user-specified parameter. For the examples considered here, this parameter is simply the Poisson ratio used to define $\tilde{C}$. No parameters were needed for the other three preconditioners because we considered only simple homogeneous problems with a shear modulus equal

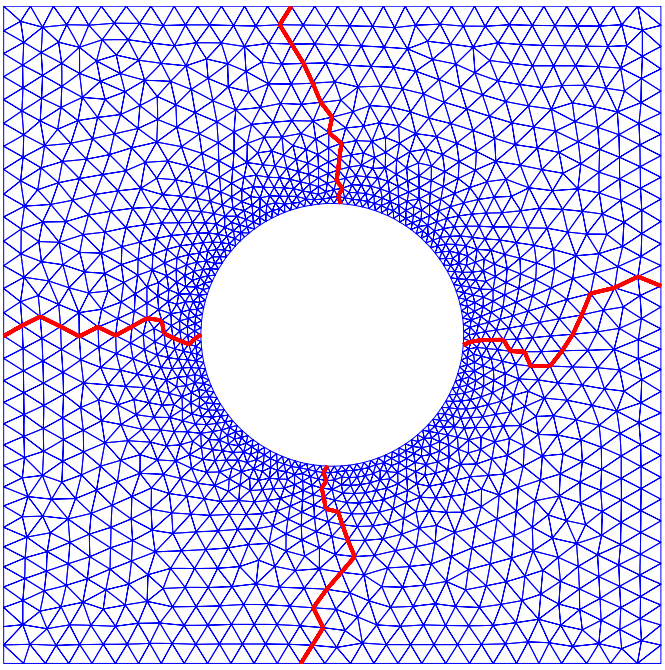


FIG. 5.1. *Unstructured mesh and decomposition into four substructures. Thick lines show boundaries shared by substructures in the preconditioner for* S_A.

TABLE 5.6

Iterations for selected problems in Tables 5.2, 5.4, *and* 5.5 *solved to a relative residual tolerance of* 10^{-9} *using direct solvers for the linear systems involving* S_A *or* A. *The column* ν *shows the value of Poisson ratio used to define* $\tilde{C}$ *in the PBP preconditioner.*

Table	ν	PBP		Block diagonal	Block triangular	Br–Pa
		GMRES	PCG	GMRES	GMRES	PCG
5.2 $N = 64$	0.3 0.4 0.49999	17 9 2	16 (4.9) 14 (2.5) 5 (1.2)	25	14	20
5.4 $N = 64$	0.3 0.4 0.49999	14 11 2	21 (5.2) 15 (2.7) 5 (1.2)	41	21	26
5.5	0.3 0.4 0.49999	16 12 3	23 (15) 17 (6.2) 5 (1.2)	37	20	31

to 1. We note that the mass matrix approximation of the dual Schur complement used by the other three preconditioners would need to be scaled appropriately to avoid degraded performance for problems with nonunit or multiple material properties. In contrast, no such scaling is needed by the present approach.

6. Conclusions. A PBP preconditioner was presented for elliptic saddle point systems. We demonstrated that the eigenvalues of the preconditioned linear system are positive and real provided certain assumptions are satisfied. A form of the precon-

ditioner suited for conjugate gradients was also presented. Numerical results in two and three dimensions were consistent with the theoretical results and demonstrated the effectiveness of the preconditioner. Excellent performance of the preconditioner was also evident when compared with other approaches.

We note that the PBP preconditioner employs a preconditioner for the penalized primal Schur complement. In contrast, the other three preconditioners considered in the previous section employ a preconditioner for the dual Schur complement. A basic difference is that the penalized primal Schur complement is known exactly, whereas the dual Schur complement is not. Consequently, it is possible to obtain much better performance from the PBP preconditioner, as shown in Table 5.6. A similar result would also likely hold for nonsymmetric saddle point systems. For example, the penalized primal Schur complement would still be known exactly for Navier–Stokes problems, but a simple mass matrix approximation of the dual Schur complement is known to degrade for increasing values of Reynolds number. Thus, one level of approximation would be removed by using the PBP preconditioner.

Acknowledgments. Theorem 3.3 was motivated in large part by related work in [16] due to Klawonn. The authors also wish to express their thanks to Michele Benzi for bringing to their attention [1] and [14].

REFERENCES

[1] O. Axelsson, *Preconditioning of indefinite problems by regularization*, SIAM J. Numer. Anal., 16 (1979), pp. 58–69.
[2] K. J. Bathe, *Finite Element Procedures*, Prentice-Hall, Upper Saddle River, NJ, 1996.
[3] M. Benzi, G. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1–137.
[4] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.
[5] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comput., 50 (1988), pp. 1–17.
[6] J. Cahouet and J.-P. Chabard, *Some fast 3d finite element solvers for the generalized Stokes problem*, Internat. J. Numer. Methods Fluids, 8 (1988), pp. 869–895.
[7] C. R. Dohrmann, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.
[8] C. R. Dohrmann, *A substructuring preconditioner for nearly incompressible elasticity problems*, Tech. Report SAND2004-5393, Sandia National Laboratories, 2004.
[9] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.
[10] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers with applications in incompressible fluid dynamics*, in Numerical Mathematics and Scientific Computation, Oxford University Press, London, 2005.
[11] H. Elman and D. Silvester, *Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations*, SIAM J. Sci. Comput., 17 (1996), pp. 33–46.
[12] M. Fortin, *Some iterative methods for incompressible flow problems*, Comput. Phys. Comm., 53 (1989), pp. 393–399.
[13] P. Goldfeld, L. F. Pavarino, and O. B. Widlund, *Balancing Neumann-Neumann preconditioners for mixed approximations of heterogeneous problems in linear elasticity*, Numer. Math., 95 (2003), pp. 283–324.
[14] M. V. Gorelova and E. V. Chizhonkov, *Preconditioning saddle point problems with the help of saddle point operators*, Comput. Math. Math. Phys., 44 (2004), pp. 1445–1455.
[15] A. Klawonn and L. F. Pavarino, *Overlapping Schwarz methods for mixed linear elasticity and Stokes problems*, Comput. Methods Appl. Mech. Engrg., 165 (1998), pp. 233–245.
[16] A. Klawonn, *Block-triangular preconditioners for saddle point problems with a penalty term*, SIAM J. Sci. Comput., 19 (1998), pp. 172–184.
[17] A. Klawonn, *An optimal preconditioner for a class of saddle point problems with a penalty term*, SIAM J. Sci. Comput., 19 (1998), pp. 540–552.

[18] J. Li, *A dual-primal feti method for incompressible Stokes equations*, Tech. Rep. 816, Courant Institute of Mathematical Sciences, 2001.
[19] J. Mandel, C. R. Dohrmann, and R. Tezaur, *An algebraic theory for primal and dual substructuring methods by constraints*, Appl. Numer. Math., (2005), pp. 167–193.
[20] J. Mandel and C. R. Dohrmann, *Convergence of a balancing domain decomposition by constraints and energy minimization*, Numer. Linear Algebra Appl., 10 (2003), pp. 639–659.
[21] L. F. Pavarino and O. B. Widlund, *Balancing Neumann-Neumann methods for incompressible Stokes equations*, Comm. Pure Appl. Math., 55 (2002), pp. 302–335.
[22] L. F. Pavarino, *Indefinite overlapping Schwarz methods for time-dependent Stokes problems*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 35–51.
[23] P. Persson and G. Strang, *A simple mesh generator in Matlab*, SIAM Rev., 46 (2004), pp. 329–345.
[24] Y. Saad and M. H. Schultz, *Gmres: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[25] D. Silvester and A. Wathen, *Fast iterative solution of stablized Stokes systems: Part* ii: *using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 283–299

A FETI-DP PRECONDITIONER WITH A SPECIAL SCALING FOR MORTAR DISCRETIZATION OF ELLIPTIC PROBLEMS WITH DISCONTINUOUS COEFFICIENTS*

N. DOKEVA†, M. DRYJA‡, AND W. PROSKUROWSKI†

Abstract. We consider two-dimensional elliptic problems with discontinuous coefficients discretized by the finite element method on geometrically conforming nonmatching triangulations across the interface using the mortar technique. The resulting discrete problem is solved by a dual-primal FETI method.

In this paper we introduce and analyze a preconditioner with a special scaling of coefficients and step parameters and establish convergence bounds. We show that the preconditioner is almost optimal with constants independent of the jumps of coefficients and step parameters. Extensive computational evidence is presented that illustrates an almost optimal convergence for a variety of situations (distribution of subregions, grid assignment, grid ratios, number of subregions) for both continuous and discontinuous problems.

Key words. domain decomposition, mortar finite element method, dual-primal FETI preconditioner, nonmatching grids, saddle-point problem, elliptic problems with discontinuous coefficients

AMS subject classifications. 65N55, 65N30, 65F10

DOI. 10.1137/040616401

1. Introduction.

In this paper we discuss a second order elliptic problem with discontinuous coefficients defined on a polygonal region $\Omega \subset \mathbb{R}^2$ which is a union of many polygons Ω_i. The problem is discretized by the finite element method (FEM) on geometrically conforming nonmatching triangulations across $\overline{\Gamma} = \cup_i \partial\Omega_i \backslash \partial\Omega$ using the mortar technique; see [1]. The resulting discrete problem is solved by a dual-primal FETI (FETI-DP) method; see [5], [6], [7] for the matching triangulation and [3], [4] for the nonmatching one. The method is discussed under the assumption of continuity of the solution at vertices of Ω_i. We prove that the method is convergent and its rate of convergence is almost optimal and independent of the jumps of coefficients, provided that a mortar side is associated with the higher coefficient. Consequently, the method is well suited for parallel processors.

The presented results are a generalization of results obtained in [4] and [3] for problems with continuous and discontinuous coefficients, respectively. In [4] a modified mortar condition at the vertices of substructures is employed using the assumption that the solution at the vertices is continuous, while in [3] a standard approximation to the mortar condition is employed. The preconditioner in [3] which does not use the scaling of the coefficients was tested for the simplest case of four subregions. In general, however, the experiments show that for discontinuous coefficients the preconditioner without proper scaling of coefficients exhibits poor convergence.

*Received by the editors October 5, 2004; accepted for publication (in revised form) June 28, 2005; published electronically March 7, 2006.

http://www.siam.org/journals/sinum/44-1/61640.html

†Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113 (dokeva@usc.edu, proskuro@math.usc.edu).

‡Department of Mathematics, Warsaw University, Banach 2, 02-097 Warsaw, Poland (dryja@mimuw.edu.pl). The work of this author was supported in part by the U.S. Department of Energy under contract DE-FG02-92ER25127 and in part by the Polish Science Foundation under grant 2P03A00524.

In this paper we introduce a preconditioner with special scaling of coefficients and step parameters. The theoretical analysis and experimental results show that the proposed preconditioner exhibits excellent properties for general cases considered here: its convergence is almost optimal with respect to the parameters of triangulations (it depends on a logarithmical factor only) and independent of the jumps of coefficients. Extensive numerical experiments on many subregions are reported.

The paper is organized as follows. In section 2, the differential and discrete problems are formulated. In section 3, a matrix form of the discrete problem is given. The preconditioner is described and analyzed in section 4. The implementation of the method and numerical experiments are presented in section 5.

2. Differential and discrete problem. We consider the following differential problems.

Find $u^* \in H_0^1(\Omega)$ such that

$$a(u^*, v) = f(v), \quad v \in H_0^1(\Omega), \tag{1}$$

where

$$a(u, v) = (\rho(x)\nabla u, \nabla u)_{L^2(\Omega)}, \quad f(v) = (f, v)_{L^2(\Omega)}.$$

We assume that Ω is a polygonal region and $\overline{\Omega} = \cup_{i=1}^N \overline{\Omega}_i$, Ω_i are disjoint polygonal subregions of diameter H_i, $\rho(x) = \rho_i$ is a positive constant on Ω_i, and $f \in L^2(\Omega)$. We solve (1) by the FEM on nonmatching triangulation across $\partial\Omega_i$. To describe a discrete problem the mortar technique is used; see [1] and [8] and the literature therein.

We impose on Ω_i a triangulation with triangular elements and parameter h_i. The resulting triangulation of Ω is nonmatching across $\partial\Omega_i$. We assume that the triangulation on each Ω_i is quasi-uniform. Let $X_i(\Omega_i)$ be a finite element space of piecewise linear continuous functions defined on the introduced triangulation. We assume that functions of $X_i(\Omega_i)$ vanish on $\partial\Omega_i \cap \partial\Omega$. Let

$$X^h(\Omega) = X_1(\Omega_1) \times \cdots \times X_N(\Omega_N).$$

Note that $X^h(\Omega) \subset L^2(\Omega)$ but $X^h(\Omega) \not\subset H_0^1(\Omega)$. To formulate a discrete problem for (1) we use the mortar technique for the geometrically conforming case. For that the following notation is used. Let Γ_{ij} be a common edge of two substructures Ω_i and Ω_j, $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$. Let $\Gamma = (\cup_i \partial\Omega_i)\backslash\partial\Omega$. We now select open edges $\gamma_m \subset \Gamma$, called *mortar*, such that $\overline{\Gamma} = \cup\overline{\gamma}_m$ and $\gamma_m \cap \gamma_n = 0$ for $m \neq n$. Let Γ_{ij} as an edge of Ω_i be denoted by $\gamma_{m(i)}$ and called *mortar* (master), and let Γ_{ij} as an edge of Ω_j be denoted by $\delta_{m(j)}$ and called *nonmortar* (slave). The criterion for choosing $\gamma_{m(i)}$ as the mortar side is that $\rho_i \geq \rho_j$, the coefficients on Ω_i and Ω_j, respectively.

Let $M(\delta_{m(j)})$ be a subspace of $W_j(\delta_{m(j)})$, the restriction of $X_j(\Omega_j)$ to $\delta_{m(j)}$, $\delta_{m(j)} \subset \partial\Omega_j$. Functions of $M(\delta_{m(j)})$ are constants on elements of the triangulation on $\delta_{m(j)}$ which touch $\partial\delta_{m(j)}$. We say that $u_i \in X_i(\Omega_i)$ and $u_j \in X_j(\Omega_j)$ on $\delta_m \equiv \delta_{m(j)} = \gamma_{m(i)} = \Gamma_{ij}$, an edge common to Ω_i and Ω_j, satisfy the mortar condition if

$$\int_{\delta_m} (u_i - u_j)\psi \, ds = 0, \quad \psi \in M(\delta_m). \tag{2}$$

Note that for the given u_i on $\gamma_{m(i)}$ and u_j on $\partial\delta_{m(j)}$, denoted by $\operatorname{Tr} u_j$, we can compute u_j at the interior nodal points of $\delta_{m(j)}$. Denoting the u_j computed in this

way by $\pi_m(u_i; \mathrm{Tr}\ u_j)$ we have

$$\int_{\delta_m} \pi_m(u_i; \mathrm{Tr}\ u_j)\psi\, ds = \int_{\delta_m} u_i\psi\, ds, \quad \psi \in M(\delta_m),$$

$$\pi_m(u_i; \mathrm{Tr}\ u_j) = \mathrm{Tr}\ u_j \ on\ \partial\delta_m.$$

Note that $\pi_m(u_i; \mathrm{Tr}\ u_j)$ is an element of X_j restricted to $\delta_{m(j)}$.

We are now in a position to introduce V^h, the space for discretization of (1). Let $V^h(\Omega)$ be a subspace of $X^h(\Omega)$ of functions which satisfy the mortar condition (2) for each $\delta_m \subset \Gamma$ and which are continuous at common vertices of the substructures. The discrete problem for (1) in V^h is defined as follows.

Find $u_h^* \in V^h$ such that

$$a_H(u_h^*, v_h) = f(v_h), \quad v_h \in V^h, \tag{3}$$

where $a_H(u,v) = \sum_{i=1}^N a_i(u,v), a_i(u,v) = \rho_i(\nabla u, \nabla v)_{L^2(\Omega_i)}$.

The problem has a unique solution and the error bound is known; see [1]. Using the basis functions of V^h, $V^h =$ span $\{\Phi_k\}$, the problem (3) is rewritten as

$$A\underline{u}_h^* = \underline{f}.$$

The form of Φ_k can be found, for example, in [4]. The matrix A is symmetric positive definite and $\mathrm{cond}(A) \le \frac{C}{\min h_i^2}$, where C here depends on the ρ_i.

3. FETI-DP equation. To derive a FETI-DP method we first rewrite the problem (3) as a saddle-point problem using Lagrange multipliers; see, for example, [8] and the literature therein. For $u = \{u_i\}_{i=1}^N \in X^h(\Omega)$ and $\psi = \{\psi_p\}_{p=1}^P \in M(\Gamma) = \prod_m M(\delta_m)$, the mortar condition (2) can be rewritten as

$$b(u,\psi) \equiv \sum_{i=1}^N \sum_{\delta_{m(i)} \subset \partial\Omega_i} \int_{\delta_{m(i)}} (u_i - u_j)\psi_k\, ds = 0,$$

where $\delta_{m(i)} = \gamma_{m(j)} = \Gamma_{ij}, \psi_k \in M(\delta_{m(i)})$. Let $\widetilde{X}^h(\Omega)$ denote a subspace of $X^h(\Omega)$ of functions which are common to the vertices of substructures.

The problem now consists of finding $(u_h^*, \lambda_h^*) \in \widetilde{X}^h(\Omega) \times M(\Gamma)$ such that

$$a(u_h^*, v_h) + b(v_h, \lambda_h^*) = f(v_h), \quad v_h \in \widetilde{X}^h(\Omega), \tag{4}$$

$$b(u_h^*, \psi_h) = 0, \quad \psi_h \in M(\Gamma). \tag{5}$$

It can be proved that u_h^*, the solution of (4)–(5), is the solution of (3) and vice versa. Therefore the problem (4)–(5) has a unique solution. This can be proved straightforwardly using the inf-sup condition, including the error bound; see [8] and the literature therein.

To derive a matrix form of (4)–(5) we first need a matrix formulation of (5). Using the nodal basis functions $\varphi_{\delta_{m(i)}}^{(l)} \in W_i(\delta_{m(i)})$, $\varphi_{\gamma_{m(j)}}^{(k)} \in W_j(\gamma_{m(j)})$, and $\psi_{\delta_{m(i)}}^{(p)} \in M_m(\delta_{m(i)})$ $(\delta_{m(i)} = \gamma_{m(j)} = \Gamma_{ij})$, (5) can be rewritten on $\bar{\delta}_{m(i)}$ as

$$B_{\delta_{m(i)}} u_{i\delta_{m(i)}} - B_{\gamma_{m(j)}} u_{j\gamma_{m(j)}} = 0, \tag{6}$$

where $u_{i\delta_{m(i)}}$ and $u_{j\gamma_{m(j)}}$ are vectors which represent $u_i|_{\delta_{m(i)}} \in W_i(\delta_{m(i)})$ and $u_j|_{\gamma_m(j)} \in W_j\ (\gamma_{m(j)})$, and ($n_{\delta_{(i)}} \equiv n_{\delta_{m(i)}}$ and $n_{\gamma_{(j)}} \equiv n_{\gamma_{m(j)}}$):

$$B_{\delta_{m(i)}} = \{(\psi^{(p)}_{\delta_{m(i)}}, \varphi^{(k)}_{\delta_{m(i)}})_{L^2(\delta_{m(i)})}\}, \quad p = 1, \dots, n_{\delta(i)}, \quad k = 0, \dots, n_{\delta(i)} + 1,$$

$$B_{\gamma_{m(j)}} = \{(\psi^{(p)}_{\delta_{m(i)}}, \varphi^{(l)}_{\gamma_{m(j)}})_{L^2(\gamma_{m(j)})}\}, \quad p = 1, \dots, n_{\delta(i)}, \quad l = 0, \dots, n_{\gamma_{(j)}} + 1.$$

Here $n_{\delta(i)}, n_{\delta(i)} + 2$, and $n_{\gamma(j)} + 2$ are the dimensions of $M_m(\delta_{m(i)})$, $W_i(\delta_{m(i)})$, and $W_j(\gamma_{m(j)})$, respectively. Note that $B_{\delta_{m(i)}}$ and $B_{\gamma_{m(j)}}$ are rectangular matrices. We split the vectors $u_{i\delta_{m(i)}}$ and $u_{j\gamma_{m(j)}}$ into vectors $u^{(r)}_{i\delta_{m(i)}}$, $u^{(c)}_{i\delta_{m(i)}}$ and $u^{(r)}_{j\gamma_{m(j)}}$, $u^{(c)}_{j\gamma_{m(j)}}$, respectively, where $u^{(c)}_{i\delta_{m(i)}}$ and $u^{(c)}_{j\gamma_{m(j)}}$ represent values of functions u_i and u_j at the end points of $\delta_{m(i)}$ and $\gamma_{m(j)}$, and $u^{(r)}_{i\delta_{m(i)}}$ and $u^{(r)}_{j\gamma_{m(j)}}$ represent values of u_i and u_j at the interior nodal points of $\delta_{m(i)}$ and $\gamma_{m(j)}$. Using this notation one can rewrite (6) as

$$(B^{(r)}_{\delta_{m(i)}} u^{(r)}_{i\delta_{m(i)}} + B^{(c)}_{\delta_{m(i)}} u^{(c)}_{i\delta_{m(i)}}) - (B^{(r)}_{\gamma_{m(j)}} u^{(r)}_{j\gamma_{m(j)}} + B^{(c)}_{\gamma_{m(j)}} u^{(c)}_{j\gamma_{m(j)}}) = 0. \tag{7}$$

Note that

$$B^{(r)}_{\delta_{m(i)}} = \{(\psi^{(p)}_{\delta_{m(i)}}, \varphi^{(k)}_{\delta_{m(i)}})_{L^2(\delta_{m(i)})}\}, \quad p,\ k = 1, \dots, n_{\delta(i)}$$

is a square tridiagonal matrix $n_{\delta(i)} \times n_{\delta(i)}$, symmetric and positive definite and $\mathrm{cond}(B^{(r)}_{\delta_{m(i)}}) \sim 1$, while the remaining matrices $B^{(c)}_{\delta_{m(i)}}$, $B^{(c)}_{\gamma_{m(j)}}$, $B^{(r)}_{\gamma_{m(j)}}$ are rectangular with dimensions $n_{\delta(i)} \times 2$, $n_{\delta(i)} \times 2$, $n_{\delta(i)} \times n_{\gamma(j)}$, respectively.

Let $K^{(l)}$ be the stiffness matrix of $a_l(\cdot\,,\cdot)$. It is represented as

$$K^{(l)} = \begin{pmatrix} K^{(l)}_{ii} & K^{(l)}_{ic} & K^{(l)}_{ir} \\ K^{(l)}_{ci} & K^{(l)}_{cc} & K^{(l)}_{cr} \\ K^{(l)}_{ri} & K^{(l)}_{rc} & K^{(l)}_{rr} \end{pmatrix}, \tag{8}$$

where the rows correspond to the interior unknowns $u^{(i)}_l$ of Ω_l, $u^{(l)}_c$ to its vertices and $u^{(r)}_l$ to its edges.

Using the above notation and the assumption of continuity of u^*_h at the vertices of $\partial\Omega_l$, (4)–(5) can be rewritten as

$$\begin{pmatrix} K_{ii} & K_{ic} & K_{ir} & 0 \\ K_{ci} & \widetilde{K}_{cc} & K_{cr} & B^T_c \\ K_{ri} & K_{rc} & K_{rr} & B^T_r \\ 0 & B_c & B_r & 0 \end{pmatrix} \begin{pmatrix} u^{(i)} \\ u^{(c)} \\ u^{(r)} \\ \widetilde{\lambda}^* \end{pmatrix} = \begin{pmatrix} f^{(i)} \\ f^{(c)} \\ f^{(r)} \\ 0 \end{pmatrix}. \tag{9}$$

Here $\widetilde{\lambda}^* = \{B^{(r)}_{\delta_{m(i)}} \lambda^*_{\delta_{m(i)}}\}, \delta_{m(i)} \subset \Gamma$, u^*_h is the solution of (4)–(5) and is represented by the vectors $u^{(i)}, u^{(c)}$, and $u^{(r)}$, which are the values of u^*_h at the interior nodal points of Ω_l, the vertices of Ω_l, and the remaining nodal points of $\partial\Omega_l \backslash \partial\Omega$, respectively; matrices K_{ii} and K_{rr} are diagonal block-matrices of $K^{(l)}_{ii}$ and $K^{(l)}_{rr}$, respectively, while matrix $\widetilde{K}_{cc}$ is built from diagonal block matrices $K^{(l)}_{cc}$ taking into account that $u^{(c)}$ are the same at the common vertices of substructures. The remaining K-matrices represent coupling between the corresponding unknowns. The mortar condition is

represented by $B = (B_c, B_r)$, where these global matrices are represented by the local ones $((B^{(r)}_{\delta_{m(i)}})^{-1}B^{(c)}_{\delta_{m(i)}},\ -(B^{(r)}_{\delta_{m(i)}})^{-1}B^{(c)}_{\gamma_{m(j)}})$, and $(I^{(r)}_{\delta_{m(i)}}, -(B^{(r)}_{\delta_{m(i)}})^{-1}B^{(r)}_{\gamma_{m(j)}})$, respectively, and $I^{(r)}_{\delta_{m(i)}}$ is an identity matrix of $n_{\delta(i)} \times n_{\delta(i)}$. The form of these matrices follows from (7) after multiplying it by $(B^{(r)}_{\delta_{m(i)}})^{-1}$.

In the system (9) we eliminate the unknowns $u^{(i)}$ and $u^{(c)}$ to obtain

$$\begin{pmatrix} \widetilde{S} & \widetilde{B}^T \\ \widetilde{B} & \widetilde{S}_{cc} \end{pmatrix} \begin{pmatrix} u^{(r)} \\ \widetilde{\lambda}^* \end{pmatrix} = \begin{pmatrix} \widetilde{f}_r \\ \widetilde{f}_c \end{pmatrix}, \tag{10}$$

where

$$\begin{cases} \widetilde{S} = K_{rr} - (K_{ri},\ K_{rc}) \begin{pmatrix} K_{ii} & K_{ic} \\ K_{ci} & \widetilde{K}_{cc} \end{pmatrix}^{-1} \begin{pmatrix} K_{ir} \\ K_{cr} \end{pmatrix}, \\ \widetilde{f}_r = f^{(r)} - (K_{ri},\ K_{rc}) \begin{pmatrix} K_{ii} & K_{ic} \\ K_{ci} & \widetilde{K}_{cc} \end{pmatrix}^{-1} \begin{pmatrix} f^{(i)} \\ f^{(c)} \end{pmatrix}, \\ \widetilde{B} = B_r - (0,\ B_c) \begin{pmatrix} K_{ii} & K_{ic} \\ K_{ci} & \widetilde{K}_{cc} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ B_c^T \end{pmatrix}, \\ \widetilde{S}_{cc} = -(0,\ B_c) \begin{pmatrix} K_{ii} & K_{ic} \\ K_{ci} & \widetilde{K}_{cc} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ B_c^T \end{pmatrix}, \\ \widetilde{f}_c = -(0,\ B_c) \begin{pmatrix} K_{ii} & K_{ic} \\ K_{ci} & \widetilde{K}_{cc} \end{pmatrix}^{-1} \begin{pmatrix} f^{(i)} \\ f^{(c)} \end{pmatrix}. \end{cases} \tag{11}$$

Note that $\widetilde{S}$ is invertible since u^*_h is continuous at the vertices of Ω_l and vanishes on $\partial\Omega$.

We next eliminate the unknown $u^{(r)}$ to get for $\widetilde{\lambda}^* \in M(\Gamma)$

$$F\widetilde{\lambda}^* = d, \tag{12}$$

where

$$F = \widetilde{B}\widetilde{S}^{-1}\widetilde{B}^T - \widetilde{S}_{cc} \quad \text{and} \quad d = \widetilde{B}\widetilde{S}^{-1}\widetilde{f}_r - \widetilde{f}_c. \tag{13}$$

This is the FETI-DP equation for the Lagrange multipliers. Since F is positive definite the problem has a unique solution. This problem can be solved by conjugate gradient iterations with a preconditioner discussed in the next section.

4. FETI-DP preconditioner. In this section we define a preconditioner for the problem (12). For that let $S^{(l)}$ denote the Schur complement of $K^{(l)}$, see (8), with respect to unknowns at the nodal points of $\partial\Omega_l$. This matrix is represented as

$$S^{(l)} = \begin{pmatrix} S^{(l)}_{rr} & S^{(l)}_{rc} \\ S^{(l)}_{cr} & S^{(l)}_{cc} \end{pmatrix}, \tag{14}$$

where the second row corresponds to unknowns at the vertices of $\partial\Omega_l$ while the first one corresponds to the remaining unknowns of $\partial\Omega_l$. Note that B_r is a matrix obtained from B defined on functions with zero values at the vertices of Ω_l and let

$$\begin{aligned} S &= \text{ diag } \{S^{(l)}\}_{l=1}^N, \qquad S_{rr} = \text{ diag } \{S^{(l)}_{rr}\}_{l=1}^N, \\ S_{cc} &= \text{ diag } \{S^{(l)}_{cc}\}_{l=1}^N, \qquad S_{cr} = (S^{(1)}_{cr}, \ldots, S^{(N)}_{cr}). \end{aligned} \tag{15}$$

The standard preconditioner developed for continuous problems in [4] is defined as

$$\bar{M}^{-1} = B_r \widehat{S}_{rr} B_r{}^T, \tag{16}$$

where $\widehat{S}_{rr} = \text{ diag } \{\widehat{S}_{rr}^{(i)}\}_{i=1}^N$, $\quad \widehat{S}_{rr}^{(i)} = S_{rr}^{(i)}$ for $\rho_i = 1$.

We employ a special scaling to generalize $\bar{M}$ to problems with discontinuous coefficients. The preconditioner M for (12) is defined as

$$M^{-1} = \widehat{B}_r \widehat{S}_{rr} \widehat{B}_r^T, \tag{17}$$

where
$\widehat{B}_r|_{\delta_{m(i)}} = (\rho_i^{1/2} I_{\delta_{m(i)}}, -\frac{h_{\delta_{m(i)}}}{h_{\gamma_{m(j)}}} \frac{\rho_i}{\rho_j} \rho_i^{1/2} B_{\delta_{m(i)}}^{-1} B_{\gamma_{m(j)}})$ for $\delta_{m(i)} \subset \partial\Omega_i$, $i = 1, \ldots, N$; $h_{\delta_{m(i)}}$ and $h_{\gamma_{m(j)}}$ are the step parameters on $\delta_{m(i)}$ and $\gamma_{m(j)}$, $\delta_{m(i)} = \gamma_{m(j)}$, respectively.

An ordering of substructures Ω_l is called mortar-nonmortar (M-N) ordering if all sides of a fixed Ω_l are mortar while all sides of the neighboring substructures of Ω_l are nonmortar.

THEOREM 4.1. *Let the mortar side be chosen where the coefficient ρ_i is larger. Then for $\lambda \in M(\Gamma)$ the following holds:*

$$c_0 \left(1 + \log \frac{H}{h}\right)^{\alpha} \langle M\lambda, \lambda\rangle \le \langle F\lambda, \lambda\rangle \le c_1 \left(1 + \log \frac{H}{h}\right)^{2} \langle M\lambda, \lambda\rangle, \tag{18}$$

where $\alpha = 0$ for M-N ordering of substructures and $\alpha = -2$ in the general case; c_0 and c_1 are positive constants independent of h_i, H_i, and the jumps of ρ_i; and $h = \min_i h_i, H = \max_i H_i$.

Proof. To prove Theorem 4.1 we need some additional facts. We first reformulate the process of reaching (12) from (9). For that we eliminate $u^{(i)}$ from the system (9). Using the notation (14) and (15) we get

$$S_{rr} u^{(r)} + S_{rc} u^{(c)} + B_r^T \widetilde{\lambda}^* = g_r, \tag{19}$$

$$S_{cr} u^{(r)} + \bar{S}_{cc} u^{(c)} + B_c^T \widetilde{\lambda}^* = g_c, \tag{20}$$

$$B_r u^{(r)} + B_c u^{(c)} = 0. \tag{21}$$

Here S_{rr} and S_{cr} ($S_{cr} = S_{cr}^T$) are defined in (15) while $\bar{S}_{cc}$ is defined by $S_{cc}^{(l)}$ (see (14)), taking into account that $u_l^{(c)}$ are the same at the common vertices of substructures.

We now eliminate $u^{(r)}$ and $u^{(c)}$ in (19)–(21). This leads to (12) with F and d of the form

$$F = F_{rr} + F_{rc} F_{cc}^{-1} F_{cr}, \quad d = d_r + F_{rc} F_{cc}^{-1} d_c. \tag{22}$$

Here

$$F_{rr} = B_r S_{rr}^{-1} B_r^T \tag{23}$$

and

$$F_{rc} = B_c - B_r S_{rr}^{-1} S_{rc}, \quad F_{cc} = \bar{S}_{cc} - S_{cr} S_{rr}^{-1} S_{rc},$$

$$d_c = g_c - S_{cr} S_{rr}^{-1} g_r, \quad d_r = B_r S_{rr}^{-1} g_r.$$

In the proof of Theorem 4.1 we will also need two lemmas.

LEMMA 4.2. *For $w \in X_1(\partial\Omega_1) \times \cdots \times X_N(\partial\Omega_N)$ with the same values at the vertices of Ω_i, the following holds:*

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}_{rr}} \leq C \left(1 + \log \frac{H}{h}\right)^2 |w|^2_S, \tag{24}$$

provided that ρ_i on the mortar side is larger than on the nonmortar side, where $z = w - I_H w$ and $I_H w$ is a linear interpolant of w on edges of $\partial\Omega_i$ with values w at the end points of the edges.

Proof. A proof of this estimate is a modification of the proof of Lemma 1 from [4]. We have

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}_{rr}} = \langle \widehat{S}_{rr} \widehat{B}_r^T B_r z, \widehat{B}_r^T B_r z \rangle.$$

Hence

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}_{rr}} = \sum_{i=1}^{N} |\widehat{B}_r^T B_r z|^2_{\widehat{S}^{(i)}}. \tag{25}$$

Note that $\widehat{B}_r^T B_r z = 0$ at the vertices. Using that, we get

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}^{(i)}} \leq C \left(\sum_{\delta_{m(i)} \subset \partial\Omega_i} |\widehat{B}_r^T B_r z|^2_{\widehat{S}_{\delta_{m(i)}}} + \sum_{\gamma_{m(i)} \subset \partial\Omega_i} |\widehat{B}_r^T B_r z|^2_{\widehat{S}_{\gamma_{m(i)}}} \right), \tag{26}$$

where $\widehat{S}_{\delta_{m(i)}}$ and $\widehat{S}_{\gamma_{m(i)}}$ are matrix representations of the $H^{1/2}_{00}$-norm on $\delta_{m(i)}$ and $\gamma_{m(i)}$, respectively. From the structure of $\widehat{B}_r$ follows

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}_{\delta_{m(i)}}} \leq 2 \left(\rho_i |z_i|^2_{\widehat{S}_{\delta_{m(i)}}} + \rho_i |B_{ij} z_j|^2_{\widehat{S}_{\delta_{m(i)}}} \right), \tag{27}$$

where here and below $z = \{z_i\}_{i=1}^N \in X^h(\Gamma)$, the restriction of $X^h(\Omega)$ to Γ, $B_{ij} \equiv (B^{(r)}_{\delta_{m(i)}})^{-1} B^{(r)}_{\gamma_{m(j)}}$, and $\delta_{m(i)} = \gamma_{m(j)}, \gamma_{m(j)} \subset \partial\Omega_j$;

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}_{\gamma_{m(i)}}} \leq 2 \left(\rho_k \left(\frac{\rho_k}{\rho_i}\right)^2 |\widehat{B}_{ki}^T z_k|^2_{\widehat{S}_{\gamma_{m(i)}}} + \rho_k \left(\frac{\rho_k}{\rho_i}\right)^2 |B_{ki}^T \widehat{B}_{ki} z_i|^2_{\widehat{S}_{\gamma_{m(i)}}} \right), \tag{28}$$

where $B_{ki} \equiv (B^{(r)}_{\delta_{m(k)}})^{-1} B^{(r)}_{\gamma_{m(i)}}$, $\widehat{B}_{ki} = \alpha_{ki} B_{ki}$, $\alpha_{ki} = \frac{h_{\delta_{m(k)}}}{h_{\gamma_{m(i)}}}$, $\gamma_{m(i)} = \delta_{m(k)}$, and $\delta_{m(k)} \subset \partial\Omega_k$. We now estimate each term of (27) and (28).

We estimate the first term of (27) as in [4]:

$$\begin{aligned} &\rho_i |z_i|^2_{\widehat{S}_{\delta_{m(i)}}} \leq C \rho_i (1 + \log \tfrac{H}{h})^2 |w_i|^2_{H^{1/2}(\partial\Omega_i)} \\ &\leq C \rho_i (1 + \log \tfrac{H}{h})^2 |w_i|^2_{\widehat{S}^{(i)}} \leq C (1 + \log \tfrac{H}{h})^2 |w_i|^2_{S^{(i)}}. \end{aligned} \tag{29}$$

To estimate the second term of (27) we use the stability of the mortar projection. Let $\pi_{\delta_{m(i)}}(z_j, 0)$ correspond to $B_{ij}(z_{j|\gamma_{m(j)}})$ for z_j restricted to $\gamma_{m(j)}$. Using that, we have

$$\begin{aligned} &\rho_i |B_{ij} z_j|^2_{\widehat{S}_{\delta_{m(i)}}} \leq C \rho_i ||\pi_{\delta_{m(i)}}(z_j, 0)||^2_{H^{1/2}_{00}(\delta_{m(i)})} \\ &\leq C \rho_i ||z_j||^2_{H^{1/2}_{00}(\gamma_{m(j)})} \leq C \rho_i (1 + \log \tfrac{H}{h})^2 |w_j|^2_{\widehat{S}^{(j)}} \leq C (1 + \log \tfrac{H}{h})^2 |w_j|^2_{S^{(j)}}. \end{aligned} \tag{30}$$

We now estimate the terms of (28). It has been shown in [4, proof of Lemma 1 and (28)] that the following holds:

$$|B_{ki}^T z_k|^2_{\widehat{S}_{\gamma_{m(i)}}} \le C|z_k|^2_{\widehat{S}_{\delta_{m(k)}}} \le C(1+\log\tfrac{H}{h})^2|w_k|^2_{\widehat{S}^{(k)}},$$

under the assumption that $h_{\delta_{m(k)}} \sim h_{\gamma_{m(i)}}$. This assumption can be removed by introducing the scaling $\alpha_{ki} = \frac{h_{\delta_{m(k)}}}{h_{\gamma_{m(i)}}}$ in B_{ki}. Thus, this estimate is valid for $\alpha_{ki}B_{ki}$ without assuming that $h_{\delta_{m(k)}} \sim h_{\gamma_{m(i)}}$; for details see the proof of Lemma 1 in [4].

Thus, the first term of (28) can be estimated as

$$\begin{aligned}&\alpha_{ki}^2\rho_k(\tfrac{\rho_k}{\rho_i})^2|B_{ki}^T z_k|^2_{\widehat{S}_{\gamma_{m(i)}}} \le \alpha_{ki}^2\rho_k|B_{ki}^T z_k|^2_{\widehat{S}_{\gamma_{m(i)}}}\\ &\le \rho_k|z_k|^2_{\widehat{S}_{\delta_{m(k)}}} \le C(1+\log\tfrac{H}{h})^2|w_k|^2_{S^{(k)}}.\end{aligned} \tag{31}$$

It remains to estimate the second term of (28). It has been shown in [4, proof of Lemma 1] that the following holds under the assumption that $h_{\delta_{m(k)}} \sim h_{\gamma_{m(i)}}$:

$$|B_{ki}^T B_{ki} z_i|^2_{\widehat{S}_{\gamma_{m(i)}}} \le C\left(1+\log\frac{H}{h}\right)^2 |w_i|^2_{\widehat{S}^{(i)}}.$$

Thus, using the scaling α_{ki} in $\widehat{B}_{ki}$ we get

$$\begin{aligned}&\alpha_{ki}^2\rho_k(\tfrac{\rho_k}{\rho_i})^2|B_{ki}^T B_{ki} z_i|^2_{\widehat{S}_{\gamma_{m(i)}}} \le \alpha_{ki}^2\rho_k|B_{ki}^T B_{ki} z_i|^2_{\widehat{S}_{\gamma_{m(i)}}}\\ &\le C\rho_k(1+\log\tfrac{H}{h})^2|w_i|^2_{\widehat{S}^{(i)}} \le C(1+\log\tfrac{H}{h})^2|w_i|^2_{S^{(i)}},\end{aligned} \tag{32}$$

without the assumption that $h_{\delta_{m(k)}} \sim h_{\gamma_{m(i)}}$.

Substituting these four estimates (29)–(32) into (27)–(28) and the resulting estimates into (26) gives

$$|\widehat{B}_r^T B_r z|^2_{\widehat{S}^{(i)}} \le C\left(1+\log\frac{H}{h}\right)^2 \left(|w_i|^2_{S^{(i)}} + \sum_j |w_j|^2_{S^{(j)}}\right),$$

where the sum is taken over $\partial\Omega_j$, which intersects $\partial\Omega_i$ by an edge. Using this in (25) provides (24). This completes the proof of Lemma 4.2. ∎

LEMMA 4.3. *For F_{rr} defined in* (23) *and $\lambda \in M(\Gamma)$,*

$$C\left(1+\log\frac{H}{h}\right)^{\alpha} \langle M\lambda,\lambda\rangle \le \langle F_{rr}\lambda,\lambda\rangle, \tag{33}$$

where $\alpha = 0$ for a M-N ordering of substructures Ω_l and $\alpha = -2$ in the general case, and C is independent of h, H, and the jumps of ρ_i.

Proof. A proof of this estimate is a modification of the proof of Theorems 2 and 3 from [4]. We first prove it for the M-N ordering of substructures. In this case $\widehat{B}_r$ can be represented as (see (17))

$$\widehat{B}_r = (\widehat{I}_N, -\widehat{B}_M), \tag{34}$$

where $\widehat{I}_N$ and $\widehat{B}_M$ are block diagonal matrices with blocks $\rho_i^{1/2} I_{\delta_{m(i)}}$ and $\alpha_{ij}\frac{\rho_i}{\rho_j}\rho_i^{1/2}(B^{(r)}_{\delta_{m(i)}})^{-1}B^{(r)}_{\gamma_{m(j)}}$, $\alpha_{ij} = \frac{h_{\delta_{m(i)}}}{h_{\gamma_{m(j)}}}$, corresponding to the N (nonmortar) and

M (mortar) substructures Ω_i, respectively. Matrix B_r is decomposed in the same way. For this ordering we can reorder matrices (15) as

$$S_{rr} = \begin{pmatrix} S_{rr}^N & 0 \\ 0 & S_{rr}^M \end{pmatrix}, \quad \widehat{S}_{rr} = \begin{pmatrix} \widehat{S}_{rr}^N & 0 \\ 0 & \widehat{S}_{rr}^M \end{pmatrix}, \tag{35}$$

where the first row corresponds to the nonmortar subregions and $S_{rr}^N = \mathrm{diag}_{i\in N}\{S_{rr}^{(i)}\}$, while the second one corresponds to mortar subregions and $S_{rr}^M = \mathrm{diag}_{i\in M}\{S_{rr}^{(i)}\}$. Then using (34) we can write preconditioner M (see (17)) in the form

$$M^{-1} = \widehat{B}_r \widehat{S}_{rr} \widehat{B}_r^T = S_{rr}^N + \widehat{B}_M \widehat{S}_{rr}^M \widehat{B}_M^T. \tag{36}$$

Note, since both terms are positive definite, that

$$\langle S_{rr}^N \lambda, \lambda\rangle \le \langle M^{-1}\lambda, \lambda\rangle,$$

and as a consequence

$$\langle M\lambda, \lambda\rangle \le \langle (S_{rr}^N)^{-1}\lambda, \lambda\rangle.$$

Using this and

$$\langle S_{rr}^{-1} B_r^T \lambda, B_r^T \lambda\rangle = \langle (S_{rr}^N)^{-1}\lambda, \lambda\rangle + \langle (S_{rr}^M)^{-1} B_M^T \lambda, B_M^T \lambda\rangle,$$

we obtain (see (23))

$$\lambda_{\min}(M^{-1/2} F_{rr} M^{-1/2}) = \min_{\lambda} \frac{\langle S_{rr}^{-1} B_r^T \lambda, B_r^T \lambda\rangle}{\langle M\lambda, \lambda\rangle} \ge 1,$$

which completes the proof for the M-N ordering.

In the case of a general ordering (non–M-N) of substructures, we have

$$\widehat{B}_r = (\widehat{I}_r^{(n)}, -\widehat{B}_r^{(m)}),$$

where $\widehat{I}_r^{(n)}$ and $\widehat{B}_r^{(m)}$ are block diagonal matrices with blocks $\rho_i^{1/2} I_{\delta_{m(i)}}$ and $\alpha_{ij}\frac{\rho_i}{\rho_j}\rho_i^{1/2}(B_{\delta_{m(i)}}^{(r)})^{-1} B_{\gamma_{m(j)}}^{(r)}$ corresponding to the nonmortar and mortar sides, respectively. In this general case matrix (15) is not block diagonal and is of the form

$$S_{rr} = \begin{pmatrix} S_{rr}^{nn} & S_{rr}^{nm} \\ S_{rr}^{mn} & S_{rr}^{mm} \end{pmatrix}, \tag{37}$$

where the first row corresponds to the nonmortar sides and the second to the mortar sides. We introduce an auxiliary matrix

$$\mathrm{diag}\{S_{rr}\} = \begin{pmatrix} S_{rr}^{nn} & 0 \\ 0 & S_{rr}^{mm} \end{pmatrix}. \tag{38}$$

Using the fact that $S_{rr} = S_{rr}^T > 0$ we get

$$\pm \begin{pmatrix} 0 & S_{rr}^{nm} \\ S_{rr}^{mn} & 0 \end{pmatrix} \le \begin{pmatrix} S_{rr}^{nn} & 0 \\ 0 & S_{rr}^{mm} \end{pmatrix},$$

from which follows that for w with zero values at the vertices of Ω_i we have

$$\langle S_{rr} w, w\rangle \le 2\langle \mathrm{diag}\{S_{rr}\} w, w\rangle. \tag{39}$$

Additionally, the following holds (see Lemma 2 in [4]):

$$\langle \mathrm{diag}\{S_{rr}\}w, w\rangle \le C\left(1+\log\frac{H}{h}\right)^2 \langle S_{rr}w, w\rangle. \tag{40}$$

The proof of Lemma 4.3 reduces to showing that

$$\lambda_{\min}(M^{-1/2}F_{rr}M^{-1/2}) = \min_{\lambda}\frac{\langle S_{rr}^{-1}B_r^T\lambda, B_r^T\lambda\rangle}{\langle(\widehat{B}_r\widehat{S}_{rr}\widehat{B}_r^T)^{-1}\lambda,\lambda\rangle} \ge \frac{C}{(1+\log\frac{H}{h})^2}.$$

(This fact has been proved in Lemma 1 of [4] for $\rho_i = 1$; the generalization for $\rho_i \neq 1$ is straightforward.) We have

$$\lambda_{\min}(M^{-1/2}F_{rr}M^{-1/2}) = \min_{\lambda}\frac{\langle F_{rr}\lambda,\lambda\rangle}{\langle M\lambda,\lambda\rangle} = \min_{\lambda}\frac{\langle (S_{rr})^{-1}B_r^T\lambda, B_r^T\lambda\rangle}{\langle(\widehat{B}_r\widehat{S}_{rr}\widehat{B}_r^T)^{-1}\lambda,\lambda\rangle}. \tag{41}$$

Using (40) we obtain the following estimate:

$$\langle S_{rr}^{nn}\lambda,\lambda\rangle = \langle \widehat{S}_{rr}^{nn}\widehat{I}_r^{(n)}\lambda, \widehat{I}_r^{(n)}\lambda\rangle \le \langle \widehat{S}_{rr}^{nn}\widehat{I}_r^{(n)}\lambda, \widehat{I}_r^{(n)}\lambda\rangle + \langle \widehat{S}_{rr}^{mm}(\widehat{B}_r^{(m)})^T\lambda, (\widehat{B}_r^{(m)})^T\lambda\rangle$$

$$= \langle \mathrm{diag}\{\widehat{S}_{rr}\}\widehat{B}_r^T\lambda, \widehat{B}_r^T\lambda\rangle \le C\left(1+\log\frac{H}{h}\right)^2 \langle \widehat{S}_{rr}\widehat{B}_r^T\lambda, \widehat{B}_r^T\lambda\rangle,$$

where $\widehat{I}_r^{(n)} = \rho_i^{1/2} I_{\delta_{m(i)}}$ on $\delta_{m(i)} \subset \partial\Omega_i$. Hence,

$$\langle(\widehat{B}_r\widehat{S}_{rr}\widehat{B}_r^T)^{-1}\lambda,\lambda\rangle \le C\left(1+\log\frac{H}{h}\right)^2 \langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle. \tag{42}$$

On the other hand, by (39)

$$\begin{aligned}\langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle &\le \langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle + \langle (S_{rr}^{mm})^{-1}\lambda,\lambda\rangle \\ &= \langle \mathrm{diag}\{S_{rr}^{-1}\}\lambda,\lambda\rangle \le 2\langle S_{rr}^{-1}\lambda,\lambda\rangle.\end{aligned} \tag{43}$$

Using (42) and (43) in (41) we get

$$\begin{aligned}\lambda_{\min}(M^{-1/2}F_{rr}M^{-1/2}) &\ge \min_{\lambda}\frac{\langle (S_{rr})^{-1}B_r^T\lambda, B_r^T\lambda\rangle}{C(1+\log\frac{H}{h})^2\langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle} \\ &\ge \min_{\lambda}\frac{\langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle}{C(1+\log\frac{H}{h})^2\langle (S_{rr}^{nn})^{-1}\lambda,\lambda\rangle} = \frac{1}{C(1+\log\frac{H}{h})^2}.\end{aligned}$$

This completes the proof of Lemma 4.3. □

Proof of Theorem 4.1. To prove the right-hand side (RHS) of Theorem 4.1 we proceed as follows. For $-\lambda \in M(\Gamma)$ we compute $w = (w^{(r)}, w^{(c)})$ by solving (19) and (20) with $g_r = 0$ and $g_c = 0$. Note that this problem has a unique solution under the assumption that $u^{(c)}$ is continuous at the cross points. Using this we get

$$\langle F\lambda,\lambda\rangle = \langle (F_{rr} + F_{rc}F_{cc}^{-1}F_{cr})\lambda,\lambda\rangle \tag{44}$$

$$= \langle (B_rS_{rr}^{-1}B_r^T + (B_c - B_rS_{rr}^{-1}S_{rc})F_{cc}^{-1}F_{cr})\lambda,\lambda\rangle = \langle B_rw^{(r)} + B_cw^{(c)},\lambda\rangle = \langle Bw,\lambda\rangle.$$

Let $I_H w$ be a linear interpolant of w on edges with values w at the end points of each edge. Note that

$$Bw = B(w - I_H w) \equiv B_r z_r$$

since $z_r \equiv w - I_H w = 0$ at the end points of the edges. Using that in (44), we get

$$\langle F\lambda, \lambda\rangle = \langle Bw, \lambda\rangle = \langle B_r z_r, \lambda\rangle. \tag{45}$$

On the other hand, using that $Sw = B^T\lambda$ (see (19) and (20)), we have

$$\begin{aligned}\langle Bw, \lambda\rangle &= \frac{\langle Bw, \lambda\rangle^2}{\langle Bw, \lambda\rangle} = \frac{\langle B_r z_r, \lambda\rangle^2}{\langle Sw, w\rangle} \\ &= \frac{\langle M^{1/2}\lambda, M^{-1/2}B_r z_r\rangle^2}{|S^{1/2}w|^2} \le \frac{|M^{1/2}\lambda|^2 |M^{-1/2}B_r z_r|^2}{|w|_S^2}.\end{aligned} \tag{46}$$

Note that by Lemma 4.2 we get

$$|M^{-1/2}B_r z_r|^2 = \langle \widehat{B}_r \widehat{S}_{rr} \widehat{B}_r^T B_r z_r, B_r z_r\rangle = |\widehat{B}_r^T B_r z_r|^2_{\widehat{S}_{rr}} \le C\left(1 + \log\frac{H}{h}\right)^2 |w|_S^2.$$

Substituting this into (46) we have

$$\langle Bw, \lambda\rangle \le C\left(1 + \log\frac{H}{h}\right)^2 |M^{1/2}\lambda|^2.$$

Using this in (45) we get the RHS estimate of (18).

To prove the left-hand side (LHS) of Theorem 4.1 we first note that

$$\langle F\lambda, \lambda\rangle \ge \langle F_{rr}\lambda, \lambda\rangle,\ \lambda \in M(\Gamma) \tag{47}$$

since $F_{cc}^{-1} > 0$. By Lemma 4.3

$$\langle F_{rr}\lambda, \lambda\rangle \ge c_0\left(1 + \log\frac{H}{h}\right)^{\alpha} \langle M\lambda, \lambda\rangle,$$

where $\alpha = 0$ for M-N ordering of substructures and $\alpha = -2$ in the general case. Using this in (47) we get the LHS of (18). ∎

5. Implementation and numerical results. The test example for all our experiments is the weak formulation, see (1), of

$$-\mathrm{div}(\rho(x)\nabla u) = f(x) \text{ in } \Omega, \tag{48}$$

with the homogenous Dirichlet boundary conditions on $\partial\Omega$, where $\Omega = (0,1)\times(0,1)$ is a union of disjoint square subregions Ω_i, $i = 1, \ldots, N$, and $\rho(x) = \rho_i$ is a positive constant in each Ω_i. The diffusion function $\rho(x)$ is chosen larger on the mortar sides of the interfaces; see Theorem 4.1.

The region Ω is cut into N regular subregions. Below we indicate the distribution of 4 coefficients ρ_i and 4 grids h_i in Ω_i, $i = 1, \ldots, 4$ with a maximum mesh ratio 8 : 1 used in our tests (for larger number of subregions, this pattern of coefficients is repeated).

For the M-N subregion ordering test case we have

$$\begin{pmatrix} 1e6 & 1 \\ 1e2 & 1e4 \end{pmatrix}, \quad \begin{pmatrix} h/8 & h \\ h/2 & h/4 \end{pmatrix}. \tag{49}$$

For the arbitrary (other than M-N) ordering of subregions test case we have

$$\begin{pmatrix} 1e6 & 1e4 \\ 1e2 & 1 \end{pmatrix}, \quad \begin{pmatrix} h/8 & h/4 \\ h/2 & h \end{pmatrix}. \tag{50}$$

Additionally, we test a 4×4 subregions case (denoted by * in the tables) that employs coefficients of the following form without a repetitive pattern:

$$\begin{pmatrix} 1e6 & 1 & 1 & 1e3 \\ 1e4 & 1e2 & 1e6 & 1 \\ 1e2 & 1e5 & 1e4 & 1e2 \\ 10 & 1e3 & 10 & 1e6 \end{pmatrix}, \quad \begin{pmatrix} h/8 & h & h & h/4 \\ h/4 & h/2 & h/8 & h \\ h/2 & h/8 & h/4 & h/2 \\ h & h/4 & h/2 & h/8 \end{pmatrix}. \tag{51}$$

5.1. Implementation. The discrete solution u_h^* of (48) is obtained as follows. Random solution at the nodal points is expressed as $(u^{(i)}, u^{(c)}, u^{(r)})$. Mortar condition on each side of the interface $\delta_{m(i)} = \gamma_{m(j)}$ is represented by (7). This gives on $\delta_{m(i)}$

$$u^{(r)}_{i\delta_{m(i)}} = (B^{(r)}_{\delta_{m(i)}})^{-1}(B^{(r)}_{\gamma_{m(j)}} u^{(r)}_{j\gamma_{m(j)}} + B^{(c)}_{\gamma_{m(j)}} u^{(c)}_{j\gamma_{m(j)}} - B^{(c)}_{\delta_{m(i)}} u^{(c)}_{i\delta_{m(i)}}). \tag{52}$$

The solution u_h^* is obtained from $(u^{(i)}, u^{(c)}, u^{(r)})$ by replacing $u^{(r)}$ on each nonmortar side by values computed by (52) and taking into account the continuity at the cross points: $u^{(c)}_{i\delta_{m(i)}} = u^{(c)}_{j\gamma_{m(j)}}$. For the given u_h^* the discrete RHS $(f^{(i)}, f^{(c)}, f^{(r)})$ is then computed.

Since $K_{ic} = 0 = K_{ci}$ in the case of triangular elements and a piecewise linear continuous finite element space, in the numerical experiments we implement somewhat simplified formulas (11):

$$\widetilde{S} = K_{rr} - K_{ri}K_{ii}^{-1}K_{ir} - K_{rc}\widetilde{K}_{cc}^{-1}K_{cr}, \quad \widetilde{f}_r = f^{(r)} - K_{ri}K_{ii}^{-1}f^{(i)} - K_{rc}\widetilde{K}_{cc}^{-1}f^{(c)},$$

$$\widetilde{B} = B_r - B_c\widetilde{K}_{cc}^{-1}K_{cr}, \quad \widetilde{S}_{cc} = -B_c\widetilde{K}_{cc}^{-1}B_c^T, \quad \text{and} \quad \widetilde{f}_c = -B_c\widetilde{K}_{cc}^{-1}f_c.$$

Computing the RHS of the Schur complement system $d = \widetilde{B}\widetilde{S}^{-1}\widetilde{f}_r - \widetilde{f}_c$ (see (13)) is equivalent to solving N coupled Neumann problems (those with Neumann boundary conditions at the interfaces and, if a subregion is adjacent to the boundary of Ω, with zero Dirichlet conditions at $\partial\Omega$) connected through the cross points and with the only nonzero values at the interfaces:

$$\begin{pmatrix} K_{ii} & 0 & K_{ir} \\ 0 & \widetilde{K}_{cc} & K_{cr} \\ K_{ri} & K_{rc} & K_{rr} \end{pmatrix} \begin{pmatrix} v_i \\ v_c \\ v_r \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \widetilde{f}_r \end{pmatrix}. \tag{53}$$

Note that this step is implemented using the capacitance matrix approach employing solvers on the subregions only. Note also that computing $\widetilde{f}_r$ requires solving N uncoupled Dirichlet problems $K_{ii}w_i = f^{(i)}$. The final result is then multiplied by $\widetilde{B}$ and corrected by $-\widetilde{f}_c$.

The preconditioned conjugate gradient (PCG) iterations to solve (12) are terminated when the norm of the residual has decreased 10^6 times in the norm generated

by the inverse of the preconditioner M^{-1}. In each PCG iteration, there are two main operations:

1. multiplication by $F = \widetilde{B}\widetilde{S}^{-1}\widetilde{B}^T - \widetilde{S}_{cc}$ (see (13)) and
2. multiplication by $M^{-1} = \widehat{B}_r\widehat{S}_{rr}\widehat{B}_r^T$ (see (17)).

Their implementation is as follows.

1. Given the search directions $p_\delta^k \in R^{n_\delta}$ at all nonmortar sides of the interfaces, we compute $r_\delta^k = Fp_\delta^k = (\widetilde{B}\widetilde{S}^{-1}\widetilde{B}^T - \widetilde{S}_{cc})p_\delta^k$ as follows: we first compute $p^k = \widetilde{B}^T p_\delta^k$; then solve for $(v_i, v_c, v_r)^T$ the N coupled Neumann problems connected through the cross points as in (53) but with the RHS $(0, 0, p^k)^T$; and finally compute $r_\delta^k = \widetilde{B}v_r - \widetilde{S}_{cc}p_\delta^k$.

2. Given the residual $r_\delta^k \in R^{n_\delta}$ at all nonmortar sides of the interfaces we compute $z_\delta^k = M^{-1}r_\delta^k = \widehat{B}_r\widehat{S}_{rr}\widehat{B}_r^T r_\delta^k$, where $\widehat{S}_{rr} = \text{ diag } \{\widehat{S}_{rr}^{(j)}\}$, $\widehat{S}_{rr}^{(j)} = S_{rr}^{(j)}$ for $\rho_i = 1$, $S_{rr}^{(j)} = K_{rr}^{(j)} - K_{ri}^{(j)}(K_{ii}^{(j)})^{-1}K_{ir}^{(j)}$ as follows: we compute $z = \widehat{B}_r^T r_\delta^k$; $v_j = (K_{ii}^{(j)})^{-1}K_{ir}^{(j)}z_j$, $z_j = z|_{\partial\Omega_j}$, which is equivalent to solving N uncoupled Dirichlet problems $K_{ii}^{(j)}v_j = K_{ir}^{(j)}z_j$ for v_j; and finally $z_\delta^k = \widehat{B}_r\widetilde{v}$, where $\widetilde{v} = \{\widetilde{v}_j\}$, $\widetilde{v}_j = K_{rr}^{(j)}z - K_{ri}^{(j)}v_j$.

After solving (12) for $\widetilde{\lambda}^*$ the final solution is obtained by solving the N coupled Neumann problems connected through the cross points (see (9))

$$\begin{pmatrix} K_{ii} & 0 & K_{ir} \\ 0 & \widetilde{K}_{cc} & K_{cr} \\ K_{ri} & K_{rc} & K_{rr} \end{pmatrix} \begin{pmatrix} u^{(i)} \\ u^{(c)} \\ u^{(r)} \end{pmatrix} = \begin{pmatrix} f^{(i)} \\ f^{(c)} - B_c^T\widetilde{\lambda}^* \\ f^{(r)} - B_r^T\widetilde{\lambda}^* \end{pmatrix}.$$

All the experiments were performed with the complete scaling of the preconditioner as in (17), including the scaling involving step parameters. In the tables, $\max \frac{H}{h_i}$ is the largest number of mesh steps on each subregion interface, "dim" is the dimension of the reduced (Schur) matrix, "# it" is the number of the PCG iterations, "$\kappa(Q)$" is the condition number estimate of the iteration matrix, and "error" is the normalized L_2 error. In all the examples the max grid ratio is $8:1$. The criterion for choosing $\gamma_{m(i)}$ as the mortar side is that $\rho_i \geq \rho_j$, the coefficients on Ω_i and Ω_j, and, if equal, where the grid is finer, $h_{\gamma_{m(i)}} \leq h_{\delta_{m(j)}}$, unless indicated otherwise.

5.2. Continuous problems. These examples serve as a comparison with the discontinuous problems investigated in further detail.

Table 1 shows that the preconditioner M of (17) employed for the continuous problem and grids (49) (with the M-N ordering of substructures) is well scalable and gives convergence logarithmically dependent on the step sizes. The exhibited dependence $\kappa(Q) = (1+\log(H/h_{\min}))^p$ with about $p = 1$ is better than the theoretical value of $p = 2$.

Table 2 shows the results for the arbitrary ordering on grids (50). Performance results for M-N ordering (Table 1) and for arbitrary ordering (Table 2) are very similar. In the latter case the computed value in the logarithmic dependence also is about $p = 1$, which is superior to the theoretical estimate of $p = 4$.

If one violates the above-mentioned recommendation and chooses $h_{\delta_{m(i)}} < h_{\gamma_{m(j)}}$, then the rate of convergence deteriorates somewhat; compare Table 3 with the upper part of Table 2.

It should be noted that results presented in Tables 1 to 3 are significantly better than those when the standard preconditioner (16) without the scaling involving step parameters is employed.

TABLE 1
Continuous coefficients. Mortar-nonmortar ordering of subregions for grids as in (49).

$\max \frac{H}{h_i}$	4×4 subregions			8×8 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	120	14	5.36	560	15	5.33
64	264	14	5.62	1232	15	5.74
128	552	14	6.27	2576	16	6.50
256	1128	15	7.17	5264	17	7.55
$\max \frac{H}{h_i}$	12×12 subregions			16×16 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	1320	15	5.31	2400	15	5.30
64	2904	15	5.76	5280	15	5.77
128	6072	16	6.54	11040	16	6.55
256	12408	17	7.62	22560	17	7.18

TABLE 2
Continuous coefficients. Arbitrary ordering of subregions for grids as in (50).

$\max \frac{H}{h_i}$	4×4 subregions			8×8 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	168	13	4.45	784	14	4.70
64	360	13	4.76	1680	14	5.06
128	744	14	5.38	3472	15	5.70
256	1512	14	6.24	7056	16	6.65
$\max \frac{H}{h_i}$	12×12 subregions			16×16 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	1848	13	4.75	3360	13	4.75
64	3960	14	5.12	7200	14	5.15
128	8184	15	5.81	14880	15	5.84
256	16632	16	6.77	30240	16	6.84

TABLE 3
The effect of choosing sides: $h_\delta < h_\gamma$. Continuous coefficients. Arbitrary ordering of subregions with grids as in (50).

$\max \frac{H}{h_i}$	4×4 subregions			8×8 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	504	15	10.50	2352	18	10.88
64	1032	15	13.97	4816	19	14.71
128	2088	16	18.03	9744	21	19.20
256	4200	17	22.74	19600	23	24.41

5.3. Discontinuous problems. For discontinuous problems the standard preconditioner (16) which does not employ scaling of coefficients exhibits poor convergence, often worse than the conjugate gradient iterations without preconditioning. Fortunately, this preconditioner allows for a multitude of scalings to be employed.

It should be pointed out that in the simplest case of M-N ordering of 2×2 subregions with only two grids and two coefficients $\rho_i : 1 = \rho_N < \rho_M$ investigated in [3], the standard preconditioner (16) displayed convergence almost independent of the ratio H/h_i (although the condition number and the number of iterations were quite high).

Several other scalings have been tried and tested. For example, for the M-N ordering as in (49) the preconditioner $M^{-1} = S_{rr}^N + B_M \widehat{S}_{rr}^M {B_M}^T$ gives convergence

TABLE 4
Discontinuous coefficients. Mortar-nonmortar ordering of subregions and grids as in (49).

$\max \frac{H}{h_i}$	4×4 subregions			8×8 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	120	3	1.03	560	3	1.03
64	264	3	1.04	1232	3	1.04
128	552	3	1.05	2576	3	1.05
256	1228	3	1.07	5264	3	1.07
$\max \frac{H}{h_i}$	12×12 subregions			16×16 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	1320	3	1.03	2400	3	1.03
64	2904	4	1.04	5280	4	1.04
128	6072	3	1.05	11040	4	1.05
256	12408	3	1.07	22560	3	1.07

TABLE 5
Discontinuous coefficients. Arbitrary ordering of subregions and grids as in (50).

$\max \frac{H}{h_i}$	4×4 subregions			8×8 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	168	8	3.27	784	9	3.40
64	360	9	4.28	1680	11	4.46
128	744	10	5.45	3472	12	5.65
256	1512	11	6.77	7056	14	7.00
$\max \frac{H}{h_i}$	12×12 subregions			16×16 subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	1848	9	3.38	3360	9	3.38
64	3960	11	4.45	7200	11	4.45
128	8184	12	5.65	14880	12	5.65
256	16632	14	7.00	30240	14	7.00

almost independent of the ratio H/h_i and the iteration count is a fraction of that obtained with preconditioner (16). However, none of these simple preconditioner scalings is satisfactory in the case of arbitrary (other than M-N) ordering, in which case a scaling that acts only on the nonmortar sides of the interfaces is required.

The preconditioner M of (17) is one of possible choices of such a scaling, and one that is exhibiting good convergence properties both in the continuous case and the discontinuous one, as we shall demonstrate.

Table 4 shows that in the case of M-N ordering of the subregions the preconditioner M gives convergence independent of the step sizes (the ratio H/h_i), the jump of coefficients, and the number of subregions.

Table 5 shows that for arbitrary ordering of subregions convergence is only logarithmically dependent of the step size, independent of the jump of coefficients, and well scalable (independent on the number of subregions). The exhibited logarithmic dependence $\kappa(Q) = (1 + \log(H/h_{\min}))^p$ with $p = 1.8$ is better than the theoretical estimate $p = 4$.

Viewing Tables 1–2 and 4–5 we can compare performances of our preconditioner for continuous and discontinuous problems. For M-N ordering we observe a much faster rate of convergence in the discontinuous case over the continuous one, while for the arbitrary ordering the rates of convergence do not differ significantly.

TABLE 6

Discontinuous coefficients. Performance comparison for arbitrary ordering with a repetitive pattern of grids as in (50) *versus that of nonrepetitive grids as in* (51) *(denoted by *).*

$\max \frac{H}{h_i}$	4×4 subregions			$4 \times 4^*$ subregions		
	dim	# it	$\kappa(Q)$	dim	# it	$\kappa(Q)$
32	168	8	3.27	160	11	4.13
64	360	9	4.28	344	12	4.44
128	744	10	5.45	712	13	4.91
256	1512	11	6.77	1448	14	5.71

TABLE 7

Discontinuous coefficients. Performance comparison for the random solution versus that of (54) *on arbitrary ordering of subregions* $4 \times 4^*$ *and grids as in* (51).

$\max \frac{H}{h_i}$	Random solution		Solution as in (54)		
	# it	$\kappa(Q)$	# it	$\kappa(Q)$	Error
32	11	4.13	10	4.16	8.57e-5
64	12	4.41	12	4.42	1.74e-5
128	13	4.91	13	5.33	4.04e-6
256	14	5.71	14	6.33	9.73e-7

Table 6 presents the comparison in performance for arbitrary ordering of 4×4 subregions between the case when the pattern of coefficients and grids is repetitive as in (50) and when it is nonrepetitive as in (51). The differences are not pronounced, which allows us to conclude that the results of experiments elsewhere in this paper with larger numbers of subregions give a reasonable representation.

We have also tested problems with extreme variations of coefficients where coefficients, ρ_i in (49) were replaced by

$$\begin{pmatrix} 1e{+}6 & 1e{+}2 \\ 1e{-}2 & 1e{-}6 \end{pmatrix}.$$

The differences in performance were only slight.

For discontinuous problems with large jump of coefficients the question of choosing sides, i.e., $h_{\gamma_{m(j)}} < h_{\delta_{m(i)}}$ versus $h_{\delta_{m(i)}} < h_{\gamma_{m(j)}}$, has virtually no effect on the convergence rate, in contrast with the continuous problems.

The variational formulation of the problem with discontinuities at the interfaces automatically imposes the continuity of the flux condition in the weak sense. The following solution (that is nonzero at the interfaces) was designed to satisfy this condition in the classical sense:

$$u(x,y) = v(x)(1-v(x))v(y)(1-v(y)), \tag{54}$$

$$v(z) = z - \frac{\sin(2m\pi z)}{2m\pi},$$

where $m = 2^k$, $k = 1$ to 4.

Choosing (54) as the exact solution allows us to test the accuracy of our solver. The results in Table 7 show that the accuracy is clearly $O(h^2)$. One needs to stress, however, that the rate of convergence remains virtually the same as with the random solution; see Table 7.

It should be mentioned that a violation of the theoretical requirement that mortar sides should be chosen where the coefficients are larger leads to a very slow convergence when preconditioner M of (17) is used.

The largest tests reported here (16×16 subregions case in Table 5) were run with the dimension of the reduced (Schur) matrix of 30,240 and about 5,500,000 grid points (degrees of freedom) in the whole domain.

6. Conclusions. In this paper we introduced and analyzed a preconditioner with special scaling involving discontinuous coefficients and step parameters, and established convergence bounds.

Extensive computational evidence presented illustrates an excellent performance of the preconditioner: its convergence is almost optimal for a variety of situations (distribution of subregions, grid assignment, grid ratios, number of subregions) and independent of the jumps of coefficients and the parameter of triangulation. This holds for both continuous and discontinuous problems (in the latter case under the theoretical assumption that a mortar side is associated with the higher coefficient).

The experiments using the proposed preconditioner also show that for discontinuous problems the choice of mortar versus nonmortar sides has little influence on convergence rate. The scaling involving step parameters removes the assumption that $h_{\delta_{m(k)}} \sim h_{\gamma_{m(i)}}$ and, for continuous problems, significantly improves the rate of convergence.

Recent experiments show that the method exhibits almost linear parallel scalability properties; see [2].

REFERENCES

[1] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.

[2] N. Dokeva and W. Proskurowski, *Parallel scalability of a FETI–DP mortar method for problems with discontinuous coefficients*, in Domain Decomposition Methods in Science and Engineering, D. Keyes and O. B. Widlund, eds., Springer, Berlin, 2006, to appear.

[3] M. Dryja and W. Proskurowski, *A FETI-DP method for the mortar discretization of elliptic problems with discontinuous coefficients*, in Domain Decomposition Methods in Science and Engineering, R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Widlund, and J. Xu, eds., Springer, Berlin, 2004, pp. 347–352.

[4] M. Dryja and O. Widlund, *A FETI-DP method for a mortar discretization of elliptic problems*, in Recent Developments in Domain Decomposition Methods, L. Pavarino and A. Toselli, eds., Springer, Berlin, 2002, pp. 41–52.

[5] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI.* I. *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[6] A. Klawonn, O. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.

[7] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.

[8] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lect. Notes Comput. Sci. Eng. 17, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Rose, T. Schlick, eds., Springer, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 300–321

NUMERICAL METHODS FOR NONCONSERVATIVE HYPERBOLIC SYSTEMS: A THEORETICAL FRAMEWORK.*

CARLOS PARÉS†

Abstract. The goal of this paper is to provide a theoretical framework allowing one to extend some general concepts related to the numerical approximation of 1-d conservation laws to the more general case of first order quasi-linear hyperbolic systems. In particular this framework is intended to be useful for the design and analysis of well-balanced numerical schemes for solving balance laws or coupled systems of conservation laws. First, the concept of path-conservative numerical schemes is introduced, which is a generalization of the concept of conservative schemes for systems of conservation laws. Then, we introduce the general definition of approximate Riemann solvers and give the general expression of some well-known families of schemes based on these solvers: Godunov, Roe, and relaxation methods. Finally, the general form of a high order scheme based on a first order path-conservative scheme and a reconstruction operator is presented.

Key words. nonconservative products, finite volume method, well-balanced schemes, approximate Riemann solvers, Godunov methods, Roe methods, relaxation methods, high order methods

AMS subject classifications. 74S10, 65M06, 35L60, 35L65, 35L67

DOI. 10.1137/050628052

1. Introduction.

The motivating question of this paper was the design of numerical schemes for P.D.E. systems that can be written under the form

$$\partial_t w + \partial_x F(w) = \mathcal{B}(w) \cdot \partial_x w + S(w)\partial_x \sigma, \tag{1}$$

where the unknown $w(x,t)$ takes values on an open convex set D of $\mathbb{R}^N$; F is a regular function from D to $\mathbb{R}^N$; $\mathcal{B}$ is a regular matrix function from D to $\mathcal{M}_{N\times N}(\mathbb{R})$; S, a function from D to $\mathbb{R}^N$; and $\sigma(x)$, a known function from $\mathbb{R}$ to $\mathbb{R}$.

System (1) includes as particular cases: systems of conservation laws ($\mathcal{B} = 0$, $S = 0$), systems of conservation laws with source term or balance laws ($\mathcal{B} = 0$), and coupled system of balance laws as defined in [7].

More precisely, the discretization of the shallow water systems that govern the flow of one shallow layer or two superposed shallow layers of immiscible homogeneous fluids was focused (see http://www.damflow.org). The corresponding systems can be written, respectively, as a balance law or a coupled system of two balance laws. Systems with similar characteristics also appear in other flow models such as two-phase flows.

It is well known that standard methods that solve correctly systems of conservation laws can fail in solving systems of balance laws, specially when approaching equilibria or near to equilibria solutions. Moreover, they can produce unstable methods when they are applied to coupled systems of conservation or balance laws. In the context of the numerical analysis of systems and coupled systems of balance laws, many authors have studied the design of well-balanced schemes, that is, schemes that

*Received by the editors March 30, 2005; accepted for publication (in revised form) July 22, 2005; published electronically March 7, 2006. This research has been partially supported by the Spanish Government Research project BFM2003-07530-C02-02.

http://www.siam.org/journals/sinum/44-1/62805.html

†Departamento de Análisis Matemático, Facultad de Ciencias, Universidad de Málaga, 29071-Málaga, Spain (pares@anamat.cie.uma.es).

preserve some equilibria: see [2], [3], [5], [7], [10], [11], [12], [17], [18], [23], [24], [27], [29], [31], [32], [36], [37], [38], [39], ...

Among the main techniques used in the derivation of well-balanced numerical schemes, one of them consists of choosing first a standard conservative scheme for the discretization of the flux terms and then discretizing the source and the coupling terms in order to obtain a consistent scheme which solves correctly a predetermined family of equilibria. If this first procedure is followed, the calculation of the correct discretization of the source and the coupling terms depends both on the specific problem and the conservative numerical scheme chosen, and it may become rather cumbersome. In [11] it was shown that the technique of modified equations can be helpful in this procedure.

Another technique consists of considering (1) as a particular case of one-dimensional quasi-linear hyperbolic system

$$\frac{\partial W}{\partial t} + \mathcal{A}(W)\frac{\partial W}{\partial x} = 0, \quad x \in \mathbb{R}, t > 0, \tag{2}$$

by adding to the system the trivial equation

$$\frac{\partial \sigma}{\partial t} = 0.$$

Once the system is rewritten under this form, piecewise constant approximations of the solutions are considered, then are updated by means of approximate Riemann solvers at the intercells.

If this second procedure is followed, the main difficulty both from the mathematical and the numerical points of view comes from the presence of nonconservative products, which makes difficult even the definition of weak solutions. Many papers have been devoted to the definition and stability of nonconservative products, and its application to the definition of weak solutions of nonconservative hyperbolic systems; see [1], [4], [6], [9], [13], [14], [26], [34], [41].

In this article we assume the definition of nonconservative products as Borel measures given by Dal Maso, LeFloch, and Murat in [14]. This definition, which depends on the choice of a family of paths in the phases space, allows one to give a rigorous definition of weak solutions of (2). Together with the definition of weak solutions, a notion of *entropy* has to be chosen as the usual Lax's concept or one related to an entropy pair. The classical theory of simple waves of hyperbolic systems of conservation laws and the results concerning the solutions of Riemann problems can then be extended to systems (2).

The choice of the family of paths may be, in general, a difficult task. The goal of this article is, once the choice is done, to provide a theoretical framework for the numerical approximation of the corresponding weak solutions of a strictly hyperbolic system (2) whose characteristic fields are either genuinely nonlinear or linearly degenerate.

The organization of the article is as follows: in section 2, a brief resume of the theory developed in [14] is presented, together with some remarks concerning the choice of paths and some properties of weak solutions.

In section 3 we introduce the concept of path-conservative numerical schemes, which is a generalization of that of conservative schemes for systems of conservation laws: a scheme will be said to be path-conservative if it conserves to some extent the Borel measure related to the nonconservative products.

Section 4 is devoted to the well-balance property: we recall the general definition of a well-balanced numerical scheme proposed in [29] and show that the well-balance property of a scheme is strongly related to its ability to approach stationary contact discontinuities.

In section 5, a general definition of approximate Reimann solvers for (2) is presented. We verify that the generalizations of the classical methods of Roe [35] and Godunov [16] presented respectively in [40] and [30] are particular cases of path-conservative methods based on approximate Riemann solvers fitting this general definition. We give also some guidelines about how to construct relaxation schemes.

Section 6 is devoted to high order methods based on reconstruction techniques. The general form of a scheme based on a first order path-conservative scheme and a reconstruction operator is presented. The schemes constructed in [8] are particular cases in which the first order method is of the Roe type. Some general results concerning the order and well-balance properties of these methods are finally presented.

2. Weak solutions. Consider the problem

$$\frac{\partial W}{\partial t} + \mathcal{A}(W)\frac{\partial W}{\partial x} = 0, \quad x \in \mathbb{R},\ t > 0, \tag{3}$$

where $W(x,t)$ belongs to Ω, an open convex subset of $\mathbb{R}^N$, and $W \in \Omega \mapsto \mathcal{A}(W) \in \mathcal{M}_{N\times N}(\mathbb{R})$ is a smooth locally bounded map. We suppose that system (3) is strictly hyperbolic, that is, for each $W \in \Omega$, $\mathcal{A}(W)$ has N real distinct eigenvalues $\lambda_1(W) < \cdots < \lambda_N(W)$, with associated eigenvectors $R_1(W),\ldots,R_N(W)$. We also suppose that for each $i = 1,\ldots,N$, the characteristic field $R_i(W)$ is either genuinely nonlinear,

$$\nabla\lambda_i(W)\cdot R_i(W) \neq 0, \quad \forall\ W \in \Omega,$$

or linearly degenerate,

$$\nabla\lambda_i(W)\cdot R_i(W) = 0, \quad \forall\ W \in \Omega.$$

The theory developed by Dal Maso, LeFloch, and Murat (see [14]) allows one to give a rigorous definition of nonconservative products associated with the choice of a family of paths in Ω.

Definition 2.1. *A* family of paths *in* $\Omega \subset \mathbb{R}^N$ *is a locally Lipschitz map*

$$\Phi\colon [0,1]\times\Omega\times\Omega \mapsto \Omega,$$

such that:

- $\Phi(0;W_L,W_R) = W_L$ *and* $\Phi(1;W_L,W_R) = W_R$*, for any* $W_L, W_R \in \Omega$*;*
- *for every arbitrary bounded set* $\mathcal{O} \subset \Omega$*, there exists a constant* k *such that*

$$\left|\frac{\partial\Phi}{\partial s}(s;W_L,W_R)\right| \leq k|W_R - W_L|,$$

for any $W_L, W_R \in \mathcal{O}$ *and almost every* $s \in [0,1]$*;*

- *for every bounded set* $\mathcal{O} \subset \Omega$*, there exists a constant* K *such that*

$$\left|\frac{\partial\Phi}{\partial s}(s;W_L^1,W_R^1) - \frac{\partial\Phi}{\partial s}(s;W_L^2,W_R^2)\right| \leq K(|W_L^1 - W_L^2| + |W_R^1 - W_R^2|),$$

for any $W_L^1, W_R^1, W_L^2, W_R^2 \in \mathcal{O}$ *and almost every* $s \in [0,1]$.

Suppose that a family of paths Φ in Ω has been chosen. Then, for $W \in (L^\infty(\mathbb{R} \times \mathbb{R}^+) \cap BV(\mathbb{R} \times \mathbb{R}^+))^N$, the nonconservative product can be interpreted as a Borel measure denoted by $[\mathcal{A}(W)W_x]_\Phi$. If the family of segments is chosen, this interpretation is equivalent to the definition of nonconservative product proposed by Volpert in [41].

Across a discontinuity with speed ξ a weak solution must satisfy the generalized Rankine–Hugoniot condition

$$\int_0^1 \big(\xi\mathcal{I} - \mathcal{A}(\Phi(s; W^-, W^+))\big)\frac{\partial \Phi}{\partial s}(s; W^-, W^+)\, ds = 0, \tag{4}$$

where $\mathcal{I}$ is the identity matrix and W^-, W^+ are the left and right limits of the solution at the discontinuity. In the particular case of a system of conservation laws (that is, if $\mathcal{A}(W)$ is the Jacobian matrix of some flux function $F(W)$), (4) is independent of the family of paths and it reduces to the usual Rankine–Hugoniot condition.

As it occurs in the conservative case, not every discontinuity is admissible. Therefore, a concept of entropic solution has to be assumed, as one of the following definitions.

DEFINITION 2.2. *A weak solution is said to be an entropic solution in the Lax sense if, at each discontinuity, there exists $i \in \{1, \dots, N\}$ such that*

$$\lambda_i(W^+) < \xi < \lambda_{i+1}(W^+) \quad \textit{and} \quad \lambda_{i-1}(W^-) < \xi < \lambda_i(W^-)$$

if the ith characteristic field is genuinely nonlinear or

$$\lambda_i(W^-) = \xi = \lambda_i(W^+)$$

if the ith characteristic field is linearly degenerate.

DEFINITION 2.3. *Given an entropy pair (η, G) for (3), i.e., a pair of regular functions from Ω to $\mathbb{R}$ such that*

$$\nabla G(W) = \nabla \eta(W) \cdot \mathcal{A}(W), \quad \forall\ W \in \Omega,$$

a weak solution is said to be entropic if it satisfies the inequality

$$\partial_t \eta(W) + \partial_x G(W) \le 0,$$

in the distributions sense.

The choice of the family of paths is important as it determines the speed of propagation of discontinuities. For scalar balance laws, rigorous justifications of the choice of the family of paths can be given using different techniques based on weak limits; see [19], [20]. In general, this choice has to be based on the physical background (see [25], [33] for instance). In any case, it is natural from the mathematical point of view to require this family to satisfy some hypotheses concerning the relation of the paths with the integral curves of the characteristic fields. Following [30], here we will assume that the family of paths satisfies the following hypotheses:

(H1) Given two states, W_L and W_R, belonging to the same integral curve γ of a linearly degenerate field, the path $\Phi(s; W_L, W_R)$ is a parameterization of the arc of γ linking W_L and W_R.

(H2) Given two states, W_L and W_R, belonging to the same integral curve γ of a genuinely nonlinear field, R_i, such that $\lambda_i(W_L) < \lambda_i(W_R)$, the path $\Phi(s; W_L, W_R)$ is a parameterization of the arc of γ linking W_L and W_R.

(H3) Let us denote by $\mathcal{RP} \subset \Omega \times \Omega$ the set of pairs (W_L, W_R) such that the Riemann problem

$$\left\{ \begin{array}{l} \dfrac{\partial W}{\partial t} + \mathcal{A}(W)\dfrac{\partial W}{\partial x} = 0, \\ \\ W(x,0) = \left\{ \begin{array}{ll} W_L & \text{if } x < 0, \\ \\ W_R & \text{if } x > 0, \end{array} \right. \end{array} \right. \tag{5}$$

has a unique self-similar solution $W(x,t) = V(x/t; W_L, W_R)$ (where the function V is piecewise regular) composed by at most N simple waves: rarefaction waves, contact discontinuities, or shocks (i.e., discontinuities satisfying the jump condition (4) and the entropy condition given by Definition 2.2 or 2.3). These simple waves connect $J+1$ intermediate states

$$W_0 = W_L; W_1, \ldots, W_{J-1}; W_J = W_R;$$

with $J \leq N$. We assume that, given two states $(W_L, W_R) \in \mathcal{RP}$, the curve described by the path $\Phi(s; W_L, W_R)$ in Ω is equal to the union of those corresponding to the paths $\Phi(s; W_j, W_{j+1})$, $j = 0, \ldots, J-1$.

If the definition of weak solutions of (3) is based on a family of paths satisfying these hypotheses, the following natural properties hold (see [30]).

PROPOSITION 2.4. *Let us suppose that the concept of weak solutions of* (3) *is defined on the basis of a family of paths satisfying hypotheses* (H1)–(H3). *Then*

(i) *Given two states W_L and W_R belonging to the same integral curve of a linearly degenerate field, the contact discontinuity given by*

$$W(x,t) = \left\{ \begin{array}{ll} W_L & \text{if } x < \sigma t, \\ W_R & \text{if } x > \sigma t, \end{array} \right.$$

where σ is the (constant) value of the corresponding eigenvalue through the integral curve, is a weak solution of (3).

(ii) *Let (W_L, W_R) be a pair belonging to $\mathcal{RP}$ and let W be the solution of the corresponding Riemann problem* (5). *The following equality holds:*

$$\left\langle \left[\mathcal{A}(W(\cdot,t))W_x(\cdot,t)\right]_\Phi, 1 \right\rangle = \int_0^1 \mathcal{A}\left(\Phi(s; W_L, W_R)\right) \frac{\partial \Phi}{\partial s}(s; W_L, W_R)\, ds.$$

Consequently, the total mass of the Borel measure $\left[\mathcal{A}(W(\cdot,t))W_x(\cdot,t)\right]_\Phi$ does not depend on t.

(iii) *Let (W_L, W_R) be a pair belonging to $\mathcal{RP}$ and let W_j be any of the intermediate states appearing in the solution of the Riemann problem* (5). *Then*

$$\begin{aligned} &\int_0^1 \mathcal{A}\left(\Phi(s; W_L, W_R)\right) \frac{\partial \Phi}{\partial s}(s; W_L, W_R)\, ds \\ &\quad = \int_0^1 \mathcal{A}\left(\Phi(s; W_L, W_j)\right) \frac{\partial \Phi}{\partial s}(s; W_L, W_j)\, ds \\ &\quad + \int_0^1 \mathcal{A}\left(\Phi(s; W_j, W_R)\right) \frac{\partial \Phi}{\partial s}(s; W_j, W_R)\, ds. \end{aligned}$$

Some general guidelines to construct a family of paths satisfying these hypotheses (at least for pairs $(W_L, W_R) \in \mathcal{RP}$) have been presented in [30].

In the following proposition we establish a property of the solution of a Riemann problem that will be of importance in the definition of generalized approximate Riemann solvers for (3).

PROPOSITION 2.5. *Given $(W_L, W_R) \in \mathcal{RP}$, the solution $W(x,t) = V(x/t; W_L, W_R)$ of the Riemann problem* (5) *satisfies the following equality:*

$$\int_0^1 \mathcal{A}\left(\Phi(s; W_L, W_R)\right) \frac{\partial \Phi}{\partial s}(s; W_L, W_R)\, ds + \int_0^\infty \left(V(v; W_L, W_R) - W_R\right)\, dv + \int_{-\infty}^0 \left(V(v; W_L, W_R) - W_L\right)\, dv = 0. \tag{6}$$

Proof. Let A, T be two positive numbers such that

$$V(x/T; W_L, W_R) = W_L, \quad \text{if } x < -A,$$
$$V(x/T; W_L, W_R) = W_R, \quad \text{if } x > A.$$

Integrating (3) in $[-A, A] \times [0, T]$, we obtain

$$\int_{-A}^{A} V(x/T; W_L, W_R)\, dx - AW_L - AW_R + \int_0^T \left\langle \left[\mathcal{A}(W(\cdot,t))W_x(\cdot,t)\right]_\Phi, 1 \right\rangle\, dt = 0.$$

Then (6) is easily obtained by taking into account (ii) of Proposition 2.4 and making the change of variables $v = x/T$ in the integral at the right-hand side. ∎

Remark 1. If the concept of entropic solution is related to an entropy pair (η, G) with convex η, the following inequality can also be proved for the solution of a Riemann problem:

$$G(W_R) + \int_0^\infty \left(\eta\big(V(v; W_L, W_R)\big) - \eta(W_R)\right)\, dv \le G(W_L) - \int_{-\infty}^0 \left(\eta\big(V(v; W_L, W_R)\big) - \eta(W_L)\right)\, dv. \tag{7}$$

The proof is identical to that corresponding to systems of conservation laws.

3. Path-conservative numerical schemes. The central concept of the theory developed in this article is that of *path-conservative* numerical scheme, which is a generalization of conservative schemes for systems of conservation laws. We recall that, given a system of conservation laws

$$\partial_t W + \partial_x F(W) = 0, \quad x \in \mathbb{R},\ t > 0, \tag{8}$$

the expression of a conservative numerical scheme is as follows:

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x}\left(G_{i-1/2} - G_{i+1/2}\right), \tag{9}$$

where Δt and Δx are the time step and the space step, which are supposed to be constant for simplicity; W_i^n represents the approximation of the average of the exact solution at the ith cell $I_i = [x_{i-1/2}, x_{i+1/2}]$ at time $t^n = n\Delta t$,

$$W_i^n \cong \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} W(x, t^n)\, dx,$$

and $G_{i+1/2} = G(W^n_{i-q}, \dots, W^n_{i+p})$ is the numerical flux at the intercell $x_{i+1/2}$,

$$G_{i+1/2} \cong \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} F(W(x_{i+1/2}, t))\, dt. \tag{10}$$

This expression is usually motivated as follows: a weak solution of (8) satisfies the equality

$$\int_a^b W(x, t_1)\, dx = \int_a^b W(x, t_0)\, dx + \int_{t_0}^{t_1} F(W(a, t))\, dt - \int_{t_0}^{t_1} F(W(b, t))\, dt, \tag{11}$$

for every rectangle $[a, b] \times [t_0, t_1]$ in $\mathbb{R} \times [0, \infty)$, and (9) is the discrete analogue of the equality (11) corresponding to the rectangle $I_i \times [t^n, t^{n+1}]$.

Let us give a reinterpretation of (9) in terms of measures in order to motivate its generalization to nonconservative problems. A weak solution can be understood as a function that satisfies the equality (8) in the sense of distributions. In the particular case of a piecewise regular weak solution, given $t > 0$ the distribution $[F(W(\cdot, t))_x]$ is defined by

$$\begin{aligned} \langle [F(W(\cdot, t))_x], \phi \rangle &= \int_{\mathbb{R}} F(W(x, t))_x \phi(x)\, dx \\ &\quad + \sum_l \left(F(W_l^+) - F(W_l^-) \right) \phi(x_l(t)), \quad \forall\ \phi \in \mathcal{D}(\mathbb{R})^N, \end{aligned} \tag{12}$$

where the derivative appearing in the integral term has to be understood in the pointwise sense; the index l of the sum runs in the number of discontinuities appearing in the solution; $x_l(t)$ is the location at time t of the lth discontinuity; W_l^- and W_l^+ the limits of the solution to the left and right of the lth discontinuity at time t; finally, $\mathcal{D}(\mathbb{R})$ represents the set of functions of class $\mathcal{C}^\infty(\mathbb{R})$ with compact support. The distribution $[F(W(\cdot, t))_x]$ can be interpreted as a Borel measure having the Lebesgue decompostion $\mu_a + \mu_s$, where μ_a is given by

$$\mu_a(E) = \int_E F(W(x, t))_x\, dx,$$

for every Borel set E, and

$$\mu_s = \sum_l \left(F(W_l^+) - F(W_l^-) \right) \delta_{x = x_l(t)}, \tag{13}$$

being $\delta_{x=a}$ the Dirac measure placed at $x = a$. Given a Borel set E, we will denote its measure by

$$\langle [F(W(\cdot, t))_x], 1_E \rangle.$$

Using this notation, (11) can be rewritten as follows:

$$\int_a^b W(x, t_1)\, dx = \int_a^b W(x, t_0)\, dx - \int_{t_0}^{t_1} \left\langle [F(W(\cdot, t))_x], 1_{[a,b]} \right\rangle\, dt. \tag{14}$$

If we now define the piecewise constant function W^n whose value at the cell I_i is the approximation W_i^n, the discrete analogue of (14) would be

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x} \langle [F(W^n)_x], 1_{I_i} \rangle, \tag{15}$$

but this equality is not equivalent to (9): notice that the measure $[F(W^n)_x]$ consists only of its singular part

$$\sum_i \left(F(W_{i+1}^n) - F(W_i^n)\right) \delta_{x=x_{i+1/2}},$$

and that the cells I_i have been defined as closed intervals. Therefore, in (15) the punctual mass placed at $x_{i+1/2}$ contribute both to cells I_i and I_{i+1}. In this sense, the conservative numerical scheme (9) can be interpreted as follows: $G_{i+1/2}$ can be considered as an *intermediate flux* that is used to split the Dirac measures placed at the intercells

$$\begin{aligned}\left(F(W_{i+1}^n) - F(W_i^n)\right) \delta_{x=x_{i+1/2}} &= \left(F(W_{i+1}^n) - G_{i+1/2}\right) \delta_{x=x_{i+1/2}} \\ &+ \left(G_{i+1/2} - F(W_i^n)\right) \delta_{x=x_{i+1/2}},\end{aligned}$$

and then, the first summand contributes to cell I_{i+1} and the second one to I_i, i.e.,

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left((F(W_i^n) - G_{i-1/2})) + (G_{i+1/2} - F(W_i^n))\right), \tag{16}$$

which is obviously equivalent to (9).

Let us now come back to nonconservative systems (3) and suppose that a family of paths Φ has been chosen to define the weak solutions. If W is again a piecewise regular weak solution, for a given time t the Borel measure related to the nonconservative product is defined as follows:

$$\begin{aligned}\langle [\mathcal{A}(W(\cdot,t))W_x(\cdot,t)]_\Phi, \phi\rangle &= \int_{\mathbb{R}} \mathcal{A}(W(x,t))W_x(x,t)\phi(x)\,dx \\ &+ \sum_l \left(\int_0^1 \mathcal{A}(\Phi(s;W_l^-,W_l^+))\frac{\partial\Phi}{\partial s}(s;W_l^-,W_l^+)\,ds\right)\phi(x_l(t)), \\ &\forall\,\phi\in\mathcal{C}_0(\mathbb{R}),\end{aligned} \tag{17}$$

which is obviously a generalization of (12). In the above equality, the expression W_x appearing in the first integral represents again the pointwise derivative of $W(\cdot,t)$; $x_l(t)$, W_l^-, W_l^+ are like in (12); and $\mathcal{C}_0(\mathbb{R})$ is the set of continuous maps with compact support.

Notice that again this measure can be decomposed as a sum $\mu_a^\Phi + \mu_s^\Phi$ where

$$\mu_a^\Phi(E) = \int_E \mathcal{A}(W(x,t))W_x(x,t)\,dx$$

for every Borel set E, and:

$$\mu_s^\Phi = \sum_l \left(\int_0^1 \mathcal{A}(\Phi(s;W_l^-,W_l^+))\frac{\partial\Phi}{\partial s}(s;W_l^-,W_l^+)\,ds\right)\delta_{x=x_l(t)}. \tag{18}$$

Given a rectangle $[a,b]\times[t_0,t_1]$ in $\mathbb{R}\times[0,\infty)$, a weak solution of (3) satisfies the equality

$$\int_a^b W(x,t_1)\,dx = \int_a^b W(x,t_0)\,dx - \int_{t_0}^{t_1} \langle [\mathcal{A}(W(\cdot,t))W_x(\cdot,t)]_\Phi, 1_{[a,b]}\rangle\,dt \tag{19}$$

that generalizes (11).

The discrete analogue of (19) is now

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\langle [\mathcal{A}(W^n)W_x^n]_\Phi \, , 1_{I_i}\rangle, \tag{20}$$

where, again, W^n is the piecewise constant function taking the value W_i^n at cell I_i. Newly, the measure $[\mathcal{A}(W^n)W_x^n]_\Phi$ consists only of its singular part,

$$\sum_i \left(\int_0^1 \mathcal{A}(\Phi(s; W_i^n, W_{i+1}^n)) \frac{\partial \Phi}{\partial s}(s; W_i^n, W_{i+1}^n)\, ds \right) \delta_{x=x_{i+1/2}}.$$

Therefore, the punctual masses placed at the intercells have to be decomposed into two terms $D_{i+1/2}^{\pm}$, one contributing to the cell I_i and the other to the cell I_{i+1}. This idea leads to the following definition.

DEFINITION 3.1. *Given a family of paths Ψ, a numerical scheme is said to be Ψ-conservative if it can be written under the form*

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\big(D_{i-1/2}^+ + D_{i+1/2}^-\big), \tag{21}$$

where

$$D_{i+1/2}^{\pm} = D^{\pm}(W_{i-q}^n, \dots, W_{i+p}^n),$$

D^- and D^+ being two continuous functions from Ω^{p+q+1} to Ω satisfying:

$$D^{\pm}(W, \dots, W) = 0, \quad \forall\ W \in \Omega, \tag{22}$$

and

$$\begin{aligned} &D^-(W_{-q}, \dots, W_p) + D^+(W_{-q}, \dots, W_p) \\ &\qquad = \int_0^1 \mathcal{A}(\Psi(s; W_0, W_1)) \frac{\partial \Psi}{\partial s}(s; W_0, W_1)\, ds, \end{aligned} \tag{23}$$

for every $W_i \in \Omega$, $i = -q, \dots, p$.

This definition generalizes the usual concept of a conservative numerical scheme for a system of conservation laws:

PROPOSITION 3.2. *Let us suppose that (3) is a system of conservation laws, i.e., $\mathcal{A}$ is the Jacobian of a flux function F. Then, every numerical scheme which is Ψ-conservative for some family of paths Ψ is consistent and conservative in the usual sense. Conversely, a consistent conservative numerical scheme is Ψ-conservative for every family of paths Ψ.*

Proof. Observe first that, in the case of a conservative system, (23) reduces to

$$D^-(W_{-q}, \dots, W_p) + D^+(W_{-q}, \dots, W_p) = F(W_1) - F(W_0).$$

Therefore, given a Ψ-conservative numerical scheme (21) we can define a numerical flux function G as follows:

$$\begin{aligned} G(W_{-q}, \dots, W_p) &= D^-(W_{-q}, \dots, W_p) + F(W_0) \\ &= -D^+(W_{-q}, \dots, W_p) + F(W_1). \end{aligned} \tag{24}$$

Then, (21) is equivalent to the conservative scheme (9) corresponding to the numerical flux G. Moreover, from (22) we easily deduce

$$G(W,\dots,W) = F(W).$$

Conversely, given a consistent conservative numerical scheme with numerical flux function G, it can be written under the form (21) by defining

$$\begin{aligned} D^-(W_{-q},\dots,W_p) &= G(W_{-q},\dots,W_p) - F(W_0),\\ D^+(W_{-q},\dots,W_p) &= -G(W_{-q},\dots,W_p) + F(W_1). \end{aligned}$$

It can be easily verified that (22) and (23) are satisfied for every family of paths Ψ. □

Remark 2. According to Proposition 3.2, a path-conservative numerical scheme applied to a conservative problem is just a conservative scheme formulated in the so-called *wave propagation form* (see [28]). It is important to notice that, in despite of its form, a path-conservative numerical scheme (21) is not a *nonconservative numerical scheme* in the usual sense: a numerical scheme for solving a *conservative problem* is said to be nonconservative if it cannot be written under the form (9).

Notice that condition (23) plays a double role. On the one hand, it is used to approximate the punctual masses associated to discontinuities. On the other hand, together with (22), it is a consistency requirement for regular solutions and smooth data. In effect, if W is a regular enough solution and $\mathcal{A}(W)$, $D^{\pm}(W_{-q},\dots,W_p)$ are also regular, from (22) and (23) it can be deduced that

$$\begin{aligned} &\frac{1}{\Delta x}\big(D^+(W(x_{i-q-1},t),\dots,W(x_{i+p-1},t)) + D^-(W(x_{i-q},t),\dots,W(x_{i+p},t))\big)\\ &\qquad = \mathcal{A}(W(x_i,t))W_x(x_i,t) + O(\Delta x). \end{aligned}$$

Path-conservative numerical schemes satisfy a certain *conservation* property. In effect, let W be a weak solution of (3) corresponding to an initial condition W_0 such that

$$W_0(x) = W_L, \quad \forall\ x < -A; \quad W_0(x) = W_R, \quad \forall\ x > A; \tag{25}$$

for some $A > 0$. Given $0 \le t_0 < t_1 < \infty$, W satisfies

$$\int_{\mathbb{R}} \big(W(x,t_1) - W(x,t_0)\big)\,dx = -\int_{t_0}^{t_1} \langle [\mathcal{A}(W(\cdot,t))W_x(\cdot,t)]_\Phi, 1\rangle\,dt. \tag{26}$$

Let us suppose now that a Ψ-conservative scheme is applied to approach this solution and let W^n be the piecewise constant function whose value at the cell I_i is W_i^n. Summing up in (21) and taking into account (17) and (23), we deduce the equality

$$\int_{\mathbb{R}} \big(W^{n+1}(x) - W^n(x)\big)\,dx = -\Delta t\,\langle [\mathcal{A}(W^n)W_x^n]_\Psi, 1\rangle, \tag{27}$$

which is clearly an approximation of (26).

As it was remarked in [29] in the context of Roe schemes, the best choice of the family of paths Ψ appearing in Definition 3.1 is the family Φ selected for the definition of weak solutions: in this case, (26) and (27) makes reference to the same

Borel measure and the jump conditions of weak solutions and numerical solutions are consistent.

In fact, a Lax–Wendroff theorem can be conjectured: if the numerical solutions obtained with a Ψ-conservative converge in an adequate sense, its limit has to be a weak solution whose definition is also related to the family of paths Ψ.

We stress that such a theorem would not be in contradiction with the negative results shown in [22] or [15]; in these works the failure of the convergence of nonconservative schemes to weak solutions of *conservative* problems was studied. But in our case, if the system is conservative, a path-conservative numerical scheme is not a nonconservative scheme (see Remark 2). Nevertheless, this kind of negative results are also expectable if a path-conservative numerical scheme based on a family Ψ is used to approach weak solutions based on a different family of paths Φ: in that case, the consistency for smooth solutions is still provided by (22) and (23) but discontinuities can be incorrectly treated. In fact, a negative result of this type was observed in [29] in the context of the approximation of shallow water systems with source term.

Unfortunately, the construction of Φ-conservative schemes can be difficult or very costly in practice. In this case, a simpler family of paths Ψ has to be chosen, as the family of segments:

$$\Psi(s; W_L, W_R) = W_L + s(W_R - W_L). \tag{28}$$

4. Well-balancing. Well-balancing is related to the numerical approximation of equilibria, i.e., steady state solutions. Notice that system (3) can only have nontrivial steady state solutions if it has some linearly degenerate fields; let $W(x)$ be a regular steady state solution

$$\mathcal{A}(W(x)) \cdot W'(x) = 0 \quad \forall\, x \in \mathbb{R}.$$

If $W'(x) \neq 0$, then 0 is an eigenvalue of $\mathcal{A}(W(x))$ and $W'(x)$ is an associated eigenvector. Therefore, $x \mapsto W(x)$ can be interpreted as a parameterization of an integral curve of a linearly degenerate characteristic field whose corresponding eigenvalue takes the value 0 through the curve. In order to define the concept of well-balancing, let us introduce the set Γ of all the integral curves γ of a linearly degenerate field of $\mathcal{A}(W)$ such that the corresponding eigenvalue vanishes on Γ. According to [29] we introduce the following definitions.

Definition 4.1. *Given a curve $\gamma \in \Gamma$, a numerical scheme for solving* (3)

$$W_j^{n+1} = W_j^n + \frac{\Delta t}{\Delta x} H(W_{j-q}^n, \dots, W_{j+p}^n) \tag{29}$$

is said to be exactly well-balanced for γ if, given any $\mathcal{C}^1$ function $x \in (\alpha, \beta) \subset \mathbb{R} \mapsto W(x) \in \Omega$ such that

$$W(x) \in \gamma, \quad \forall\, x \in (\alpha, \beta), \tag{30}$$

and $p + q + 1$ points in (α, β) $x_{-q}, \dots, x_p$ such that

$$x_{-q} < \cdots < x_p; \quad x_{i+1} - x_i = \Delta x,\ i = -q, \dots, p-1, \tag{31}$$

then

$$H(W(x_{-q}), \dots, W(x_p)) = 0. \tag{32}$$

The scheme is said to be well-balanced with order k for γ if, given any $\mathcal{C}^{k+1}$ function W and any set of points $\{x_{-q}, \dots, x_p\}$ satisfying (30), (31), *then*

$$|H(W(x_{-q}), \dots, W(x_p))| = O\left(\Delta x^{k+1}\right). \tag{33}$$

Finally, the scheme is said to be exactly well-balanced or well-balanced with order k if these properties are satisfied for any curve of Γ.

We have only considered 1-level schemes and uniform meshes in order to avoid an excess of notation, but the definition can be easily extended to more general schemes.

The well-balance property of a scheme is strongly related to its ability to approximate stationary contact discontinuities. We can state for instance the following proposition.

PROPOSITION 4.2. *Given a numerical scheme of the form* (21) *with $q = 0$ and $p = 1$ and a curve γ of Γ, the numerical scheme is exactly well-balanced for γ if and only if it solves exactly every stationary contact discontinuity linking two states belonging to γ.*

Proof. Both properties are satisfied if and only if

$$D^{\pm}(W_0, W_1) = 0, \quad \forall\, W_0, W_1 \in \gamma. \qquad \square$$

Remark 3. For numerical schemes with arbitrary values of p and q the direct implication of the proposition is also valid. To see this, observe first that a numerical scheme is exactly well-balanced for γ if and only if

$$H(W_{-q}, \dots, W_p) = 0,$$

for any given *ordered* set of states $\{W_{-q}, \dots, W_p\}$ of γ, where some of the states can be repeated. Then it can be easily shown that this property implies that the numerical scheme solves exactly stationary contact discontinuities linking two states belonging to γ.

5. Approximate Riemann solvers. This section is devoted to generalize the notion of approximate Riemann solvers introduced in [21] for conservative systems (8) and extended in [5] for balance laws. The organization of this section closely follows Bouchut's book.

DEFINITION 5.1. *Given a family of paths Ψ, a Ψ-approximate Riemann solver for* (3) *is a function $\widetilde{V} : \mathbb{R} \times \Omega \times \Omega \mapsto \Omega$ satisfying the following:*

(i) *for every $W \in \Omega$,*

$$\widetilde{V}(v; W, W) = W \quad \forall\, v \in \mathbb{R}; \tag{34}$$

(ii) *for every W_L, $W_R \in \Omega$ there exist $\lambda_{min}(W_L, W_R)$, $\lambda_{max}(W_L, W_R)$ in $\mathbb{R}$ such that,*

$$\widetilde{V}(v; W_L, W_R) = W_L, \quad \text{if } v < \lambda_{min}(W_L, W_R),$$
$$\widetilde{V}(v; W_L, W_R) = W_R, \quad \text{if } v > \lambda_{max}(W_L, W_R);$$

(iii) *for every $W_L, W_R \in \Omega$,*

$$\begin{aligned} &\int_0^1 \mathcal{A}\left(\Psi(s; W_L, W_R)\right) \frac{\partial \Psi}{\partial s}(s; W_L, W_R)\, ds \\ &\quad + \int_0^{\infty} \left(\widetilde{V}(v; W_L, W_R) - W_R\right) dv \\ &\quad + \int_{-\infty}^0 \left(\widetilde{V}(v; W_L, W_R) - W_L\right) dv = 0. \end{aligned} \tag{35}$$

Notice that (35) is a generalization of the property (6) satisfied by the exact solution of a Riemann problem (5).

Given a Ψ-approximate Riemann solver for (3) a numerical scheme can be constructed as follows:

$$W_i^{n+1} = \frac{1}{\Delta x}\left(\int_{x_{i-1/2}}^{x_i} \widetilde{V}\left(\frac{x - x_{i-1/2}}{\Delta t}; W_{i-1}^n, W_i^n\right) dx + \int_{x_i}^{x_{i+1/2}} \widetilde{V}\left(\frac{x - x_{i+1/2}}{\Delta t}; W_i^n, W_{i+1}^n\right) dx\right). \tag{36}$$

Under a CFL condition 1/2, the numerical scheme can also be written under the form (21) with

$$D_{i+1/2}^- = -\int_{-\infty}^{0} \left(\widetilde{V}(v; W_i^n, W_{i+1}^n) - W_i^n\right) dv, \tag{37}$$

$$D_{i+1/2}^+ = -\int_{0}^{\infty} \left(\widetilde{V}(v; W_i^n, W_{i+1}^n) - W_{i+1}^n\right) dv. \tag{38}$$

PROPOSITION 5.2. *A numerical scheme* (21) *based on a* Ψ*-approximate Riemann solver is* Ψ*-conservative.*

Proof. The proof is straightforward from (37), (38), and Definition 5.1. ∎

Remark 4. If the numerical scheme is intended to solve only weak solutions with *small* discontinuities, i.e., discontinuities linking pairs of states (W_L, W_R) belonging to $\mathcal{RP}$, then it is enough for the approximate Riemann solver $\widetilde{V}$ to be defined in $\mathbb{R} \times \mathcal{RP}$.

A numerical scheme (21) based on a Ψ-approximate Riemann solver is well-balanced for a curve γ of the set Γ, if and only if, given two states W_L and W_R in γ the following equalities hold:

$$\int_{-\infty}^{0} \left(\widetilde{V}(v; W_L, W_R) - W_L\right) dv = 0,$$
$$\int_{0}^{\infty} \left(\widetilde{V}(v; W_L, W_R) - W_R\right) dv = 0.$$

These equalities are trivially satisfied if

$$\widetilde{V}(v; W_L, W_R) = \begin{cases} W_L & \text{if } v < 0, \\ W_R & \text{if } v > 0, \end{cases}$$

i.e., if the approximate Riemann solver is exact for pairs of states (W_L, W_R) belonging to γ.

We recall hereafter some classical choices of approximate Riemann solvers.

5.1. Godunov methods. Godunov methods correspond to the choice of the exact Riemann solver, i.e.,

$$\widetilde{V}(v; W_L, W_R) = V(v; W_L, W_R),$$

being $V(x/t; W_L, W_R)$ the exact solution of the Riemann problem (5). This is clearly a Φ-approximate Riemann solver. Moreover, if the concept of entropic solution is

related to an entropy pair (η, G) with convex η, according to Remark 1 it is *dissipative* for this pair (see [5]).

In [30] it has been shown that if the family of paths satisfies the hypotheses (H1)–(H3) stated in section 2, Godunov methods can be written under the form (21) with

$$D^-_{i+1/2} = \int_0^1 \mathcal{A}\left(\Phi(s; W^n_i, W^n_{i+1/2})\right) \frac{\partial \Phi}{\partial s}(s; W^n_i, W^n_{i+1/2})\, ds,$$
$$D^+_{i+1/2} = \int_0^1 \mathcal{A}\left(\Phi(s, W^n_{i+1/2}, W^n_{i+1})\right) \frac{\partial \Phi}{\partial s}(s; W^n_{i+1/2}, W^n_{i+1})\, ds,$$

where $W^n_{i+1/2}$ is the (constant) value at $x = x_{i+1/2}$ of the solution of the Riemann problem related to the states W^n_i and W^n_{i+1}. If the solution is discontinuous at $x = x_{i+1/2}$ the limit to the left or the right can be chosen indifferently.

Godunov methods are exactly well-balanced (see [30]).

5.2. Roe methods. Approximate Riemann solvers are often constructed as follows: $\widetilde{V}(x/t; W_L, W_R)$ is the solution of a linear Riemann problem

$$\left\{ \begin{array}{l} \dfrac{\partial U}{\partial t} + \mathcal{A}(W_L, W_R)\dfrac{\partial U}{\partial x} = 0, \\ \\ U(x,0) = \left\{ \begin{array}{ll} W_L & \text{if } x < 0, \\ W_R & \text{if } x > 0, \end{array} \right. \end{array} \right. \tag{39}$$

where $\mathcal{A}(W_L, W_R)$ is a linearization of $\mathcal{A}(W)$. It can be easily shown that this is a Ψ-approximate Riemann solver, if and only if, $\mathcal{A}(W_L, W_R)$ is a Roe linearization in the sense defined by Toumi in [40].

DEFINITION 5.3. *Given a family of paths Ψ, a function $\mathcal{A}_\Psi : \Omega \times \Omega \mapsto \mathcal{M}_{N\times N}(\mathbb{R})$ is called a* Roe linearization *if it verifies the following properties:*

1. *for each $W_L, W_R \in \Omega$, $\mathcal{A}_\Psi(W_L, W_R)$ has N distinct real eigenvalues,*
2. *$\mathcal{A}_\Psi(W, W) = \mathcal{A}(W)$, for every $W \in \Omega$,*
3. *for any $W_L, W_R \in \Omega$,*

$$\mathcal{A}_\Psi(W_L, W_R)(W_R - W_L) = \int_0^1 \mathcal{A}(\Psi(s; W_L, W_R)) \frac{\partial \Psi}{\partial s}(s; W_L, W_R)\, ds. \tag{40}$$

Once a Roe linearization $\mathcal{A}_\Psi$ has been chosen, some straightforward calculations allow one to show that, under a CFL condition 1/2, the numerical scheme can be written under the form (21) with

$$D^-_{i+1/2} = \mathcal{A}^-_{i+1/2}(W^n_{i+1} - W^n_i),$$
$$D^+_{i+1/2} = \mathcal{A}^+_{i+1/2}(W^n_{i+1} - W^n_i),$$

where

$$\mathcal{A}_{i+1/2} = \mathcal{A}_\Psi(W^n_i, W^n_{i+1}),$$

and, as usual,

$$\mathcal{L}^\pm_{i+1/2} = \begin{bmatrix} (\lambda_1^{i+1/2})^\pm & & 0 \\ & \ddots & \\ 0 & & (\lambda_N^{i+1/2})^\pm \end{bmatrix}, \ \mathcal{A}^\pm_{i+1/2} = \mathcal{K}_{i+1/2}\mathcal{L}^\pm_{i+1/2}\mathcal{K}^{-1}_{i+1/2} \tag{41}$$

being $\mathcal{L}_{i+1/2}$ the diagonal matrix whose coefficients are the eigenvalues of $\mathcal{A}_{i+1/2}$

$$\lambda_1^{i+1/2} < \lambda_2^{i+1/2} < \cdots < \lambda_N^{i+1/2},$$

and $\mathcal{K}_{i+1/2}$ is a $N \times N$ matrix whose columns are associated eigenvectors.

As in the case of systems of conservation laws, a CFL condition 1 is used in practice, as this condition ensures the linear stability of the method. An entropy-fix technique also has to be added to the numerical scheme.

In [29] it has been shown that a Roe scheme based on a family of paths Ψ is exactly well-balanced for a curve $\gamma \in \Gamma$ if, given two states W_L and W_R in γ, the path $\Psi(s; W_L, W_R)$ is a parameterization of the arc of γ linking these states. In particular, if the family of path Ψ coincides with the family Φ used in the definition of weak solutions, the numerical scheme is exactly well-balanced. The numerical scheme is well-balanced with order k if $\Psi(s; W_L, W_R)$ approximates with order $k+1$ a regular parameterization of the arc of γ linking the states. In particular, a Roe scheme based on the family of segments (28) is always well-balanced with order 2. Moreover, it is exactly well-balanced for curves of Γ that are straight lines (see [29] for details).

The construction of Roe methods for systems of the form (1) has been studied in [29].

5.3. Relaxation methods. The goal of this paragraph is to give some guidelines about the construction of approximate Riemann solvers for nonconservative systems based on the relaxation technique. This has been done for balance laws in [5].

The idea is as follows. First of all, a new nonconservative hyperbolic system is considered,

$$\frac{\partial \widetilde{W}}{\partial t} + \mathcal{B}(\widetilde{W}) \frac{\partial \widetilde{W}}{\partial x} = 0, \quad x \in \mathbb{R},\ t > 0, \tag{42}$$

where $\widetilde{W}$ now takes values in an open convex $\widetilde{\Omega}$ of $\mathbb{R}^{\widetilde{N}}$, with $\widetilde{N} > N$. Again, $\mathcal{B}$ is a smooth locally bounded map from $\widetilde{\Omega}$ to $\mathcal{M}_{\widetilde{N}\times\widetilde{N}}(\mathbb{R})$.

Let us suppose that there exist two linear operators $\mathcal{L} : \widetilde{\Omega} \mapsto \Omega$ and $\mathcal{M} : \Omega \mapsto \widetilde{\Omega}$ such that

$$\mathcal{L}\mathcal{M}(W) = W \quad \forall\ W \in \Omega.$$

In practice, system (42) has to be chosen in such a way that it is possible to easily construct an approximate Riemann solver with good properties (this is the case, for instance, if Riemann problems related to (42) are easy to solve). Then, an approximate Riemann solver for (3) is deduced.

The main difference with the conservative case comes from the fact that, in this case, together with system (42) a family of paths in $\widetilde{\Omega}$ also has to be chosen in order to define the approximate Riemann solver for this system.

The following lemma, whose demonstration is straightforward, gives a sufficient condition to obtain a Ψ-approximate Reimann solver for (3) from a $\widetilde{\Psi}$-approximate Reimann solver for (42).

LEMMA 5.4. *Let Ψ and $\widetilde{\Psi}$ be two families of paths in Ω and $\widetilde{\Omega}$, respectively, such that*

$$\begin{aligned} &\int_0^1 \mathcal{L}\mathcal{B}\big(\widetilde{\Psi}(s; \mathcal{M}(W_L), \mathcal{M}(W_R))\big) \frac{\partial \widetilde{\Psi}}{\partial s}(s; \mathcal{M}(W_L), \mathcal{M}(W_R))\, ds \\ &\qquad = \int_0^1 \mathcal{A}\big(\Psi(s; W_L, W_R)\big) \frac{\partial \Psi}{\partial s}(s; W_L, W_R)\, ds. \end{aligned} \tag{43}$$

Then, if $\mathcal{R}(v;\widetilde{W}_L,\widetilde{W}_R)$ *is a* $\widetilde{\Psi}$*-approximate Reimann solver for* (42)*, the function*

$$\widetilde{V}(v;W_L,W_R)=\mathcal{L}\mathcal{R}(v;\mathcal{M}(W_L),\mathcal{M}(W_R)),$$

gives a Ψ*-approximate Reimann solver for* (3).

Remark 5. It can also be easily shown that, if (η,G) is an entropy pair for (3) and $(\widetilde{\eta},\widetilde{G})$ is an entropy extension to (42) (see [5]), and both η and $\widetilde{\eta}$ are convex functions, then, if $\mathcal{R}$ is dissipative for $(\widetilde{\eta},\widetilde{G})$, $\widetilde{V}$ is dissipative for (η,G).

6. High order schemes based on reconstruction of states. The goal of this section is to obtain a high order scheme for (3) based on a first order path-conservative numerical scheme (21) with $q=0$ and $p=1$, that is,

$$D^{\pm}_{i+1/2}=D^{\pm}(W^n_i,W^n_{i+1}),$$

and a reconstruction operator of order s, i.e., an operator that associates to a given sequence $\{W_i\}$ two new sequences $\{W^-_{i+1/2}\}$, $\{W^+_{i+1/2}\}$ in such a way that, whenever

$$W_i=\frac{1}{\Delta x}\int_{I_i}W(x)\,dx,\quad\forall\ i\in\mathbb{Z},$$

for some smooth function W, then

$$W^{\pm}_{i+1/2}=W(x_{i+1/2})+O\left(\Delta x^s\right),\quad\forall\ i\in\mathbb{Z}.$$

In the case of a system of conservation laws (8), high order methods based on the reconstruction of states can be built using the following procedure: a first order conservative scheme with numerical flux function $G(U,V)$ and a reconstruction operator of order s are first chosen. Next, the method of lines is used: the system is discretized only in space, leaving the problem continuous in time. Let us denote by $\overline{W}_i(t)$ the cell average of solution W of (3) over the cell I_i at time t,

$$\overline{W}_i(t)=\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}}W(x,t)\,dx.$$

The following equation can be easily obtained from (8):

$$\overline{W}'_i(t)=\frac{1}{\Delta x}\left(F(W(x_{i-1/2},t))-F(W(x_{i+1/2},t))\right). \tag{44}$$

Now, (44) is approached as follows:

$$W'_i(t)=\frac{1}{\Delta x}\big(\widetilde{G}_{i-1/2}-\widetilde{G}_{i+1/2}\big), \tag{45}$$

with

$$\widetilde{G}_{i+1/2}=G(W^-_{i+1/2}(t),W^+_{i+1/2}(t)), \tag{46}$$

$W_i(t)$ being the approximation to $\overline{W}_i(t)$, and $\{W^{\pm}_{i+1/2}(t)\}$ the reconstructions associated to the sequence $\{W_i(t)\}$. It can be shown that (45)–(46) give a semidiscrete method of order s for (8).

Notice that (45) is a system of ordinary differential equations which is solved using a standard numerical method.

Let us introduce an interpretation of (45) in terms of measures, as it was done in section 2, in order to generalize it to nonconservative systems. First, notice that (44) can also be written under the form

$$\overline{W}_i'(t) = -\frac{1}{\Delta x}\left\langle [F(W(\cdot,t))_x], 1_{I_i}\right\rangle. \tag{47}$$

Next, let us choose at every cell I_i and at every time $t>0$ a regular function P_i^t such that

$$\lim_{x\to x_{i-1/2}^+} P_i^t(x) = W_{i-1/2}^+(t), \qquad \lim_{x\to x_{i+1/2}^-} P_i^t(x) = W_{i+1/2}^-(t). \tag{48}$$

If we consider now the approximation of $W(\cdot,t)$ given by the piecewise regular function $\mathcal{W}^t$ whose restriction to I_i is P_i^t, the discrete analogue of (47) would be

$$W_i' = -\frac{1}{\Delta x}\left\langle [F(\mathcal{W}^t)_x], 1_{I_i}\right\rangle, \tag{49}$$

but, again, (49) is not equivalent to (45). In this case, $[F(\mathcal{W}^t)_x]$ is the sum of a regular measure, whose Radon–Nykodim derivative at the cell I_i is $F(P_i^t)_x$, and the singular measure

$$\sum_i \left(F(W_{i+1/2}^+(t)) - F(W_{i+1/2}^-(t))\right)\delta_{x=x_{i+1/2}}.$$

If, again, the numerical flux of the first order scheme is used to split the Dirac measures placed at the intercells

$$\begin{aligned}&\left(F(W_{i+1/2}^+(t)) - F(W_{i+1/2}^-(t))\right)\delta_{x=x_{i+1/2}}\\ &\quad = \left(F(W_{i+1/2}^+(t)) - \widetilde{G}_{i+1/2}\right)\delta_{x=x_{i+1/2}} + \left(\widetilde{G}_{i+1/2} - F(W_{i+1/2}^-(t))\right)\delta_{x=x_{i+1/2}},\end{aligned}$$

and the first and second summands are assigned, respectively, to the cells I_{i+1} and I_i, we obtain from (49),

$$\begin{aligned}W_i' = -\frac{1}{\Delta x}\Bigg(&F(W_{i-1/2}^+(t)) - \widetilde{G}_{i-1/2} + \widetilde{G}_{i+1/2} - F(W_{i+1/2}^-(t))\\ &+ \int_{x_{i-1/2}}^{x_{i+1/2}} F(P_i^t(x))_x\,dx\Bigg),\end{aligned} \tag{50}$$

which is obviously equivalent to (45).

We go now to the general case (3). In this case, the equation for the cell averages is the following:

$$\overline{W}_i' = -\frac{1}{\Delta x}\langle [\mathcal{A}(W(\cdot,t))W(\cdot,t)_x]_\Phi, 1_{I_i}\rangle. \tag{51}$$

The natural extension of (50) is then

$$W_i' = -\frac{1}{\Delta x}\left(\widetilde{D}_{i-1/2}^+ + \widetilde{D}_{i+1/2}^- + \int_{x_{i-1/2}}^{x_{i+1/2}} \mathcal{A}[P_i^t(x)]\frac{dP_i^t}{dx}(x)\,dx\right), \tag{52}$$

with

$$\widetilde{D}^{\pm}_{i+1/2} = D^{\pm}(W^{-}_{i+1/2}(t), W^{+}_{i+1/2}(t)). \tag{53}$$

In (52) the integral terms are approximations of the regular measure of the Lebesgue decomposition of $[\mathcal{A}(W(\cdot,t))W_x(\cdot,t)]_\Phi$ while the terms $\widetilde{D}^{\pm}_{i-1/2}$ are related to its singular part.

Notice that there is an important difference between the conservative and nonconservative case: while in the conservative case the numerical scheme is independent of the functions P^t_i chosen at the cells (only the property (48) is important), this is not the case for nonconservative systems. As a consequence, while the numerical scheme (45) has order s, in the case of the scheme (52) the order will depend on the choice of the functions P^t_i.

In practice, the definition of the reconstruction operator gives the natural choice of the functions P^t_i, as the usual procedure is the following: given a sequence $\{W_i\}$ of values at the cells, an approximation function is calculated at every cell I_i using the values W_j at a *stencil*,

$$P_i(x; W_{i-l}, \dots, W_{i+r}),$$

with l, r being two natural numbers. The reconstructions $W^{\pm}_{i+1/2}$ are then calculated by taking the limits of these functions at the intercells. These approximations functions are usually calculated by means of interpolation or approximation techniques. The natural choice of P^t_i is thus

$$P^t_i(x) = P_i(x; W_{i-l}(t), \dots, W_{i+r}(t)).$$

Let us now investigate the order of the numerical scheme (52). Notice first that, for regular solutions W, the differential equation (51) can be written as follows:

$$\overline{W}'_i(t) = -\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \mathcal{A}(W(x,t))W_x(x,t)\,dx. \tag{54}$$

THEOREM 6.1. *Let us suppose that $\mathcal{A}$, $D^{\pm}$ are regular and with bounded derivatives. Let us suppose also that the reconstruction operator is of order s and that, given the sequence defined by*

$$W_i = \frac{1}{\Delta x}\int_{I_i} W(x)\,dx,$$

for any smooth function W, the following approximation properties are satisfied:

$$P_i(x; W_{i-l}, \dots, W_{i+r}) = W(x) + O\left(\Delta x^{s_1}\right) \quad \forall\, x \in I_i,$$
$$\frac{d}{dx}P_i(x; W_{i-l}, \dots, W_{i+r}) = W'(x) + O\left(\Delta x^{s_2}\right) \quad \forall\, x \in I_i.$$

Then (52) *is an approximation of order at least $\bar{s} = \min(s, s_1+1, s_2+1)$ to the system* (54) *in the following sense:*

$$\begin{aligned}\widetilde{D}^{+}_{i-1/2} - \widetilde{D}^{-}_{i+1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} \mathcal{A}(P^t_i(x))\frac{dP^t_i}{dx}(x)\,dx \\ = \int_{x_{i-1/2}}^{x_{i+1/2}} \mathcal{A}(W(x,t))W_x(x,t)\,dx + O\left(\Delta x^{\bar{s}}\right),\end{aligned} \tag{55}$$

for every smooth enough solution W, being $\{W^{\pm}_{i+1/2}(t)\}$ the reconstructions corresponding to the sequence $\{\overline{W}_i(t)\}$ and P^t_i the functions defined by

$$P^t_i(x) = P_i(x; \overline{W}_{i-l}(t), \dots, \overline{W}_{i+r}(t)).$$

The proof is identical to that of the particular case studied in [8], where general high order numerical schemes based on first order Roe methods were introduced.

Remark 6. For the usual reconstruction techniques one has $s_2 \leq s_1 < s$ and the order of (52) is thus $s_2 + 1$ for nonconservative systems and s for systems of conservation laws. Therefore a loss of accuracy can be observed when a technique of reconstruction is applied to a nonconservative problem. This effect has been detected and verified numerically for WENO-Roe methods in [8].

Notice that (52) can also be written under a form similar to (21),

$$W_i' = -\frac{1}{\Delta x}\left(E^+_{i-1/2} + E^-_{i+1/2}\right), \tag{56}$$

with

$$\begin{aligned} E^+_{i+1/2} &= \widetilde{D}^+_{i+1/2} + \int_{x_{i+1/2}}^{x_{i+1}} \mathcal{A}(P^t_{i+1}(x))\frac{dP^t_{i+1}}{dx}(x)\,dx, \\ E^-_{i+1/2} &= \widetilde{D}^-_{i+1/2} + \int_{x_i}^{x_{i+1/2}} \mathcal{A}(P^t_i(x))\frac{dP^t_i}{dx}(x)\,dx. \end{aligned} \tag{57}$$

Using this notation, the following equality holds:

$$\begin{aligned} E^+_{i+1/2} + E^-_{i+1/2} = &\int_{x_i}^{x_{i+1/2}} \mathcal{A}(P^t_i(x))\frac{dP^t_i}{dx}(x)\,dx \\ &+ \int_0^1 \mathcal{A}(\Psi(s; W^-_{i+1/2}, W^+_{i+1/2}))\frac{\partial \Psi}{\partial s}(s; W^-_{i+1/2}, W^+_{i+1/2})\,ds \\ &+ \int_{x_{i+1/2}}^{x_{i+1}} \mathcal{A}(P^t_{i+1}(x))\frac{dP^t_{i+1}}{dx}(x)\,dx, \end{aligned} \tag{58}$$

with Ψ being the family of paths for which the first order numerical scheme is path-conservative.

This latter equality can be understood as a path-conservation property similar to (23), where now the path linking $W_i(t)$ and $W_{i+1}(t)$ is the composition of three paths:

$$x \in [x_i, x_{i+1/2}] \mapsto P^t_i(x), \tag{59}$$

linking $W_i(t)$ and $W^-_{i+1/2}(t)$;

$$s \in [0,1] \mapsto \Psi(s; W^-_{i+1/2}(t), W^+_{i+1/2}(t)), \tag{60}$$

linking $W^-_{i+1/2}(t)$ and $W^+_{i+1/2}(t)$; and finally,

$$x \in [x_{i+1/2}, x_{i+1}] \mapsto P^t_{i+1}(x), \tag{61}$$

linking $W^+_{i+1/2}(t)$ and $W_{i+1}(t)$. Nevertheless, this family of paths does not depend only on the states $W_i(t)$ and $W_{i+1}(t)$ (as was the case in Definition 3.1) but on the values at the stencil

$$W_{i-l}(t), \dots, W_{i+r}(t).$$

The definition of a well-balanced scheme can be easily extended for semidiscrete methods (see [8]).

DEFINITION 6.2. *Let us consider a semidiscrete method for solving* (3)*:*

$$\begin{cases} W_i' = \frac{1}{\Delta x}\mathcal{H}(\mathbf{W}(t); i), \quad i \in \mathbb{Z}, \\ \mathbf{W}(0) = \mathbf{W}_0, \end{cases} \tag{62}$$

where $\mathbf{W}(t) = \{W_i(t)\}$ *represents the vector of approximations to the cell averages of the exact solution, and* $\mathbf{W}_0 = \{W_i^0\}$ *is the vector of initial data. Let* γ *be a curve of* Γ*. The numerical method* (62) *is said to be exactly well-balanced for* γ *if, given a regular stationary solution* W*, such that*

$$W(x) \in \gamma \quad \forall\ x \in \mathbb{R},$$

the vector $\mathbf{W} = \{W(x_i)\}$*, where* x_i *denotes the center of the cell* I_i*, is a critical point for the system of differential equations* (62)*, i.e.,*

$$\mathcal{H}(\mathbf{W}; i) = 0 \quad \forall\ i,$$

and it is said to be well-balanced with order k *if:*

$$\mathcal{H}(\mathbf{W}; i) = O\left(\Delta x^k\right) \quad \forall\ i.$$

Finally, the semidiscrete method (62) *is said to be exactly well-balanced or well-balanced with order* k *if these properties are satisfied for every curve* γ *of the set* Γ*.*

We give hereafter two results concerning the well-balanced property of this scheme generalizing those presented in [8] for the particular case of Roe-based reconstruction methods, but before then we introduce a new definition.

DEFINITION 6.3. *The reconstruction operator is said to be* exactly well-balanced *for a curve* $\gamma \in \Gamma$ *if, given a sequence* $\{W_i\}$ *in* γ*, the approximation functions satisfy*

$$P_i(x; W_{i-l}, \dots, W_{i+r}) \in \gamma \quad \forall\, x \in [x_{i-1/2}, x_{i+1/2}], \tag{63}$$

for every i*.*

THEOREM 6.4. *Let* γ *belong to* Γ*. Let us suppose that both the first order scheme and the reconstruction operator are exactly well-balanced for* γ*. Then, the numerical scheme* (52) *is also exactly well-balanced for* γ*.*

THEOREM 6.5. *Under the hypothesis of Theorem* 6.1*, the scheme* (52) *is well-balanced with an order of at least* $\bar{s} = \min(s, s_1 + 1, s_2 + 1)$*.*

The proofs of these results are identical to the corresponding theorems stated in [8].

Aknowledgments. The author wishes to thank M. J. Castro, J. M. Gallardo, and M. L. Muñoz for their helpful comments; and to F. Bouchut for stimulating discussions.

REFERENCES

[1] F. Alouges and B. Merlet, *Approximate shock curves of nonconservative hyperbolic systems in one space dimension*, J. Hyperbolic Differ. Equ., 1 (2004), pp. 769–788.

[2] E. Audusse, F. Bouchut, M. O. Bristeau, R. Klein, and B. Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comp., 25 (2004), pp. 2050–2065.

[3] A. Bermúdez and M. E. Vázquez, *Upwind methods for hyperbolic conservation laws with source terms*, Comput. & Fluids, 23 (1994), pp. 1049–1071.

[4] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math., 161 (2005), pp. 223–342.

[5] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*, Birkhäuser, Basel, Switzerland, 2004.

[6] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[7] M. J. Castro, J. Macías, and C. Parés, *A Q-scheme for a class of systems of coupled conservation laws with source term. Application to a two-layer* 1*-D shallow water system*, M2AN Math. Mod. Numer. Anal., 35 (2001), pp. 107–127.

[8] M. J. Castro, J. M Gallardo, and C. Parés, *Finite volume schemes based on WENO reconstruction of states for solving nonconservative hyperbolic systems. Applications to shallow water systems*, Math. Comp., 2005 to appear.

[9] J. J. Cauret, J. F. Colombeau, and A. Y. LeRoux, *Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations*, J. Math. Anal. Appl., 139 (1989), pp. 552–573.

[10] T. Chacón, A. Domínguez, and E. D. Fernández, *A family of stable numerical solvers for shallow water equations with source terms*, Comput. Methods Appl. Mech. Eng., 192 (2003), pp. 203–225.

[11] T. Chacón, A. Domínguez, and E. D. Fernández, *Asymptotically balanced schemes for nonhomogeneous hyperbolic systems–application to the shallow water equations*, C.R. Math. Acad. Sci. Paris, 338 (2004), pp. 85–90.

[12] T. Chacón, E. D. Fernández, M. J. Castro, and C. Parés, *On well-balanced finite volume methods for nonhomogeneous nonconservative hyperbolic systems,* preprint, 2005.

[13] J. F. Colombeau and A. Heibig, *Nonconservative products in bounded variation functions*, SIAM J. Math. Anal., 23 (1992), pp. 941–949.

[14] G. Dal Maso, P. G. LeFloch, and F. Murat, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), pp. 483–548.

[15] F. De Vuyst, *Schémas Nonconservatifs et Schémas Cinétiques Oour la Simulation Numérique D'écoulements Hypersoniques Non Visqueux en Déséquilibre Thermochimique*, Thése de Doctorat de l'Université Paris VI, Paris, France, 1994.

[16] S. K. Godunov, *A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 357–393.

[17] L. Gosse, *A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms*, Comput. Math. Appl., 39 (2000), pp. 135–159.

[18] L. Gosse, *A well-balanced scheme using nonconservative products designed for hyperbolic systems of conservation laws with source terms*, Math. Models Methods Appl. Sci., 11 (2001), pp. 339–365.

[19] L. Gosse, *Localization effects and measure source terms in numerical schemes for balance laws*, Math. Comp., 71 (2002), pp. 553–582.

[20] G. Guerra, *Well-posedness for a scalar conservation law with singular nonconservative source*, J. Differential Equations, 206 (2004), pp. 438–469.

[21] A. Harten, P. D. Lax, and B. Van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.

[22] T. Hou and P. G. LeFloch, *Why nonconservative schemes converge to wrong solutions: Error analysis*, Math. Comp., 62 (1994), pp. 497–530.

[23] J. M. Greenberg and A. Y. LeRoux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[24] J. M. Greenberg, A. Y. LeRoux, R. Baraille, and A. Noussair, *Analysis and approximation of conservation laws with source terms*, SIAM J. Numer. Anal., 34 (1997), pp. 1980–2007.

[25] P. G. LeFloch, *Propagating phase boundaries. Formulation of the problem and existence via the Glimm method*, Arch. Rational Mech. Anal., 123 (1993), pp. 153–197.
[26] P. G. LeFloch and A. E. Tzavaras, *Representation of weak limits and definition of nonconservative products*, SIAM J. Math. Anal., 30 (1999), pp. 1309–1342.
[27] R. LeVeque, *Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm*, J. Comput. Phys., 146 (1998), pp. 346–365.
[28] R. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.
[29] C. Parés and M. J. Castro, *On the well-balanced property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 821–852.
[30] C. Parés, J. M. Gallardo, M. L. Muñoz, and M. J. Castro, *Godunov's method for nonconservative hyperbolic systems. Application to linear balance laws*, preprint, 2005.
[31] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint–Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.
[32] B. Perthame and C. Simeoni, *Convergence of the upwind interface source method for hyperbolic conservation laws*, in Hyperbolic Problems: Theory, Numerics, Applications, Thou and Tadmor, ed., Springer, Berlin, 2003.
[33] P. A. Raviart and L. Sainsaulieu, *A nonconservative hyperbolic system modeling spray dynamics.* I. *Solution of the Riemann problem*, Math. Models Methods Appl. Sci., 5 (1995), pp. 297–333.
[34] J. P. Raymond, *A new definition of nonconservative products and weak stability results*, Boll. Un. Mat. Ital. B, 10 (1996), pp. 681–699.
[35] P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.
[36] P. L. Roe, *Upwinding difference schemes for hyperbolic conservation laws with source terms*, in Proceedings of the Conference on Hyperbolic Problems, Carasso, Raviart, and Serre, eds., Springer, 1986, pp. 41–51.
[37] H. Tang, T. Tang, and K. Xu, *A gas-kinetic scheme for shallow-water equations with source terms*, Z. Angew. Math. Phys., 55 (2004), pp. 365–382.
[38] T. Tang and Z. H. Teng, *Error bounds for fractional step methods for conservation laws with source terms*, SIAM J. Numer. Anal., 32 (1995), pp. 110–127.
[39] E. F. Toro, *Shock-Capturing Methods for Free-Surface Shallow Flows*, Wiley, Chichester, UK, 2001.
[40] I. Toumi, *A weak formulation of Roe's approximate Riemann solver*, J. Comput. Phys., 102 (1992), pp. 360–373.
[41] A. I. Volpert, *Spaces BV and quasilinear equations*, Math. USSR Sbornik, 73 (1967), pp. 255–302.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 322–348

STABILIZED FINITE ELEMENT METHODS BASED ON MULTISCALE ENRICHMENT FOR THE STOKES PROBLEM*

RODOLFO ARAYA†, GABRIEL R. BARRENECHEA†, AND FRÉDÉRIC VALENTIN‡

Abstract. This work concerns the development of stabilized finite element methods for the Stokes problem considering nonstable different (or equal) order of velocity and pressure interpolations. The approach is based on the enrichment of the standard polynomial space for the velocity component with multiscale functions which no longer vanish on the element boundary. On the other hand, since the test function space is enriched with bubble-like functions, a Petrov–Galerkin approach is employed. We use such a strategy to propose stable variational formulations for continuous piecewise linear in velocity and pressure and for piecewise linear/piecewise constant interpolation pairs. Optimal order convergence results are derived and numerical tests validate the proposed methods.

Key words. Stokes equation, multiscale functions, SIMPLEST element, bubble function

AMS subject classification. 65N30

DOI. 10.1137/050623176

1. Introduction.

Finite element solution of the Stokes problem poses the basic problem of satisfying the discrete Babuska–Brezzi (or inf-sup) condition (see [24] and the references therein). This is indeed a restriction from the point of view of implementation since equal order velocity and pressure spaces do not satisfy this condition. On the other hand, the minimal space to imagine, namely continuous piecewise linear polynomials for the velocity and piecewise constant polynomials for the pressure, does not satisfy this condition either.

Several solutions have been proposed to overcome this restriction, starting with that in [11] and the first consistent method in [28]. Moreover, in [23, 27, 29, 34] the possibility of considering discontinuous spaces for the pressure was considered and justified. On the other hand, in [14, 13], the idea from [16] has been used to propose a new kind of stabilized finite element methods, with stabilizing terms now containing only jump terms across the interelement boundaries. For an overview of stabilized finite element methods for the Stokes problem, see [19] and [5].

On the other hand, the theoretical justification of stabilized methods has become a subject of interest in the last decade. In [2, 3, 4, 31], the connection between stabilized finite element methods and Galerkin methods enriched with bubble functions has been used to propose new stabilized finite element methods for Stokes-like and linearized Navier–Stokes problems. Also, in [22] *macro* bubbles were used to derive a method analogous to the locally stabilized method from [29] containing jump terms across the interelement boundaries of the macroelements. In the resulting method, the stabilizing

*Received by the editors January 24, 2005; accepted for publication (in revised form) August 2, 2005; published electronically March 7, 2006. A part of this work was done during the stay of the third author at the Departamento de Ingeniería Matemática of Universidad de Concepción and the stay of the first and second authors at LNCC, Petrópolis, Brazil, in the framework of the joint Chile (CONICYT)-Brazil (CNPq) project 2003-4-173 (Chile)-690221/02-9 (Brazil).

http://www.siam.org/journals/sinum/44-1/62317.html

†Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (raraya@ing-mat.udec.cl, gbarrene@ing-mat.udec.cl). The first author is partially supported by FONDECYT project 1040595. The second author is partially supported by CONICYT-Chile through FONDECYT project 1030674 and FONDAP Program on Applied Mathematics.

‡Departamento de Matemática Aplicada, Laboratório Nacional de Computação Científica, Av. Getúlio Vargas, 333, 25651-070 Petrópolis - RJ, Brazil (valentin@lncc.br).

terms are defined over the macroelements, and there is no error analysis or numerical validation of the method. All these works used the so-called bubble condensation procedure, i.e., eliminating the bubble function at the element level and writing the method as the Galerkin part, plus a term derived from the influence of the bubble functions on the formulation. A particular kind of bubble enrichment of the velocity space is the so-called residual-free bubble (RFB) method (cf. [7, 8, 12]), in which the bubble function is now the solution of a problem containing the residual of the continuous equation at the element level (see [9, 10, 32] for the a priori error analysis). This bubble part may be analytically condensed or numerically computed. In the latter case this procedure leads to the two-level finite element method.

The imposition of a zero boundary condition on the element boundary for the RFB has led to some numerical problems. Solutions for these problems have been proposed by relaxing the zero boundary condition, such as the discontinuous enrichment method [18] (for the Helmholtz equation), and, more recently, the multiscale finite element method; see [21], where the main idea may be found, and [20], where the a priori error analysis is performed (for the reaction-diffusion equation, an a posteriori error estimator based on this idea has been proposed and analyzed in [1]). A particularity of such methods is that a Petrov–Galerkin strategy is proposed, in which the test function space is enriched with bubble functions in order to have a local problem containing the residual of the momentum equation on the right-hand side. A special boundary condition (related to the one used in [25, 26, 15]) is imposed in order to solve these local problems analytically. The resulting method is of Petrov–Galerkin type, in which the trial function space is generated by a basis formed by the addition of usual polynomial basis functions and enrichment functions from the solution of the differential problem in each element (which are now, unlike the RFB, known analytically, and hence the method is not of a two-level finite element method type), and in which the test function space is the standard polynomial space.

The purpose of this work is to use the multiscale approach from [21, 20], combined with the static condensation procedure, in order to propose new stabilized finite element methods for the Stokes problem. We proceed as in [21], defining an enrichment function for the trial space for the velocity that no longer vanishes on the element boundary (and hence it is not a bubble function), and then we split it into a bubble part and a function being a harmonic extension of the boundary condition. This boundary condition comes from the solution of an elliptic ODE containing a part of the differential operator at the boundary, and a jump term as the right-hand side. Depending on the jump term chosen, this procedure will lead to different methods. Both functions are condensed, and hence we obtain a method which includes the usual Galerkin-Least-Squares (GLS) stabilizing terms at the element level, plus a positive jump term on the interelement boundaries, each one with a proper stabilization parameter. One special feature of these new methods is that the previously mentioned ODE at the element boundary may be solved analytically, and hence the stabilization parameter associated with the jump terms is known exactly.

The plan of the paper is as follows. In section 2 we present the general framework and derive a general form of the method. In sections 3 and 4 this framework is applied to derive concrete stabilized finite element methods for two families of interpolation spaces, namely $\mathbb{P}^1/\mathbb{P}^0$ and continuous $\mathbb{P}^1/\mathbb{P}^1$ elements. For both cases optimal order a priori error estimates are derived for the natural norms of the unknowns, plus some extra control on the norm of the jumps appearing in the formulation. As we already mentioned, if we change the right-hand side on the boundary condition, we can derive a new method. This is done in section 5, where we give an alternative enrichment

strategy leading to another family of methods, whose analysis is analogous to that of sections 3 and 4, and which contains a boundary term containing the residual of the Cauchy stress tensor on the internal edges of the triangulation. Numerical experiments confirming the theoretical results and comparing the performance of all the methods are presented in section 6, and some final remarks and conclusions are given in section 7.

2. The model problem and the general framework. Let Ω be an open bounded domain in $\mathbb{R}^2$ with polygonal boundary, $\boldsymbol{f} \in L^2(\Omega)^2$ and consider the following Stokes problem:

$$-\nu\,\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla\cdot \boldsymbol{u} = 0 \quad \text{in } \Omega\,, \qquad \boldsymbol{u} = \boldsymbol{0} \quad \text{on } \partial\Omega\,, \tag{1}$$

where $\nu \in \mathbb{R}^+$ is the fluid viscosity.

Now let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of Ω, built up using triangles K with boundary ∂K. Let also $\mathcal{E}_h$ be the set of internal edges of the triangulation, $h_K := diam(K)$ and $h := \max\{h_K : K \in \mathcal{T}_h\}$. Let V_h be the usual finite element space of continuous piecewise polynomials of degree k, $1 \le k \le 2$ with zero trace on $\partial\Omega$. Let also Q_h be a space of piecewise polynomials of degree l, $0 \le l \le 1$, which may be continuous or discontinuous in Ω and which belong to $L^2_0(\Omega)$. Let $H^m(\mathcal{T}_h)$ and $H^m_0(\mathcal{T}_h)$ $(m \ge 1)$ be the spaces of functions whose restriction to $K \in \mathcal{T}_h$ belongs to $H^m(K)$ and $H^m_0(K)$, respectively. Furthermore, $(\cdot\,,\cdot)_D$ stands for the inner product in $L^2(D)$ (or in $L^2(D)^2$ or $L^2(D)^{2\times 2}$, when necessary), and we denote by $\|\cdot\|_{s,D}$ $(|\cdot|_{s,D})$ the norm (seminorm) in $H^s(D)$ (or $H^s(D)^2$, if necessary). As usual, $H^0(D) = L^2(D)$, and $|\cdot|_{0,D} = \|\cdot\|_{0,D}$.

In order to propose a Petrov–Galerkin method for the Stokes problem (1), let $E_h \subset H^1_0(\Omega)$ be a finite-dimensional space, called a multiscale space, such that $V_h \cap E_h = \{0\}$. Then, we propose the following Petrov–Galerkin scheme for (1): Find $\boldsymbol{u}_1 + \boldsymbol{u}_e \in [V_h \oplus E_h]^2$ and $p \in Q_h$ such that

$$\nu(\nabla(\boldsymbol{u}_1 + \boldsymbol{u}_e), \nabla \boldsymbol{v}_h)_\Omega - (p, \nabla\cdot \boldsymbol{v}_h)_\Omega + (q, \nabla\cdot(\boldsymbol{u}_1 + \boldsymbol{u}_e))_\Omega = (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega$$

for all $\boldsymbol{v}_h \in [V_h \oplus H^1_0(\mathcal{T}_h)]^2$ and all $q \in Q_h$. Now, this Petrov–Galerkin scheme is equivalent to the following system:

$$\nu(\nabla(\boldsymbol{u}_1 + \boldsymbol{u}_e), \nabla \boldsymbol{v}_1)_\Omega - (p, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q, \nabla\cdot(\boldsymbol{u}_1 + \boldsymbol{u}_e))_\Omega = (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega \ \ \forall (\boldsymbol{v}_1, q) \in V_h^2 \times Q_h, \tag{2}$$

$$\nu(\nabla(\boldsymbol{u}_1 + \boldsymbol{u}_e), \nabla \boldsymbol{v}_b)_K - (p, \nabla\cdot \boldsymbol{v}_b)_K = (\boldsymbol{f}, \boldsymbol{v}_b)_K \quad \forall \boldsymbol{v}_b \in H^1_0(K)^2 \,\forall\, K \in \mathcal{T}_h\,. \tag{3}$$

Equation (3) above is equivalent to

$$(-\nu\Delta \boldsymbol{u}_e, \boldsymbol{v}_b)_K = (\boldsymbol{f} + \nu\Delta \boldsymbol{u}_1 - \nabla p, \boldsymbol{v}_b)_K \quad \forall \boldsymbol{v}_b \in H^1_0(K)^2,$$

which, in strong form, may be written as

$$-\nu\Delta \boldsymbol{u}_e = \boldsymbol{f} + \nu\Delta \boldsymbol{u}_1 - \nabla p \quad \text{in } K. \tag{4}$$

Now, this differential problem must be completed with boundary conditions. For reasons that will become clear in what follows, we will impose the following boundary

condition on $\boldsymbol{u}_e$:

$$\boldsymbol{u}_e = \boldsymbol{g}_e \quad \text{on each } Z \subset \partial K\,, \tag{5}$$

where $\boldsymbol{g}_e = \mathbf{0}$ if $Z \subset \partial\Omega$, and $\boldsymbol{g}_e$ is the solution of

$$\begin{aligned} -\nu\,\partial_{\boldsymbol{ss}}\boldsymbol{g}_e &= \frac{1}{h_Z}\,[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p\mathbf{I}\cdot\boldsymbol{n}]\!] \quad \text{in } Z\,,\\ \boldsymbol{g}_e &= \mathbf{0} \quad \text{at the nodes}\,, \end{aligned} \tag{6}$$

on the internal edges, where $h_Z = |Z|$, $\boldsymbol{n}$ is the normal outward vector on ∂K, $\partial_{\boldsymbol{s}}$, and $\partial_{\boldsymbol{n}}$ are the tangential and normal derivative operators, respectively, $[\![v]\!]$ stands for the jump of v across Z, and $\mathbf{I}$ is the $\mathbb{R}^{2\times 2}$ identity matrix.

REMARK 2.1. *Both the shape of the jump term and the h_Z^{-1} coefficient on the boundary condition have been suggested by the error analysis. On the other hand, if we impose as the right-hand side in* (6) *the residual of the Cauchy stress tensor on ∂K, we have another class of methods. This alternative will be analyzed in section* 5.

Now, on each $K \in \mathcal{T}_h$, we can write $\boldsymbol{u}_e|_K = \boldsymbol{u}_e^K + \boldsymbol{u}_e^{\partial K}$, where

$$\begin{aligned} -\nu\Delta\boldsymbol{u}_e^K &= \boldsymbol{f} + \nu\Delta\boldsymbol{u}_1 - \nabla p \quad \text{in } K\,,\\ \boldsymbol{u}_e^K &= \mathbf{0} \quad \text{on } \partial K\,, \end{aligned} \tag{7}$$

and

$$\begin{aligned} -\nu\Delta\boldsymbol{u}_e^{\partial K} &= \mathbf{0} \quad \text{in } K\,,\\ \boldsymbol{u}_e^{\partial K} &= \boldsymbol{g}_e \quad \text{on } \partial K\,, \end{aligned} \tag{8}$$

where $\boldsymbol{g}_e$ is the solution of (6). Such differential problems are well posed, and (3) is immediately satisfied.

In this way, we can define two operators $\mathcal{M}_K : L^2(K)^2 \to H_0^1(K)^2$ and $\mathcal{B}_K : L^2(\partial K)^2 \to H^1(K)^2$ such that

$$\boldsymbol{u}_e^K = \frac{1}{\nu}\,\mathcal{M}_K(\boldsymbol{f} + \nu\Delta\boldsymbol{u}_1 - \nabla p) \quad \forall K \in \mathcal{T}_h \tag{9}$$

and

$$\boldsymbol{u}_e^{\partial K} = \frac{1}{\nu}\,\mathcal{B}_K\left([\![\nu\,\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p\mathbf{I}\cdot\boldsymbol{n}]\!]\right) \quad \forall K \in \mathcal{T}_h\,. \tag{10}$$

Next, since the enriched part $\boldsymbol{u}_e$ is fully identified through (9)–(10) (or, equivalently, by (7)–(8)), we can perform statical condensation to derive a stabilized finite element method for our problem (1). First, integrating by parts, we have, on each $K \in \mathcal{T}_h$,

$$\begin{aligned} \nu(\nabla\boldsymbol{u}_e, \nabla\boldsymbol{v}_1)_K &= -\nu(\boldsymbol{u}_e, \Delta\boldsymbol{v}_1)_K + (\boldsymbol{u}_e, \nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1)_{\partial K},\\ (q, \nabla\cdot\boldsymbol{u}_e)_K &= -(\boldsymbol{u}_e, \nabla q)_K + (\boldsymbol{u}_e, q\mathbf{I}\cdot\boldsymbol{n})_{\partial K}. \end{aligned}$$

Using these identities we can rewrite (2) in the following way:

$$\begin{aligned} &\nu(\nabla\boldsymbol{u}_1, \nabla\boldsymbol{v}_1)_\Omega + \sum_{K\in\mathcal{T}_h}\Big[-(\boldsymbol{u}_e, \nu\Delta\boldsymbol{v}_1)_K + (\boldsymbol{u}_e, \nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1)_{\partial K}\Big] - (p, \nabla\cdot\boldsymbol{v}_1)_\Omega\\ &+(q, \nabla\cdot\boldsymbol{u}_1)_\Omega + \sum_{K\in\mathcal{T}_h}\Big[-(\boldsymbol{u}_e, \nabla q)_K + (\boldsymbol{u}_e, q\mathbf{I}\cdot\boldsymbol{n})_{\partial K}\Big] = (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega, \end{aligned} \tag{11}$$

which implies

$$\nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q, \nabla\cdot \boldsymbol{u}_1)_\Omega$$

$$+ \sum_{K\in\mathcal{T}_h} \Big[- (\boldsymbol{u}_e, \nu\Delta \boldsymbol{v}_1 + \nabla q)_K + (\boldsymbol{u}_e^{\partial K}, \nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q\mathbf{I}\cdot\boldsymbol{n})_{\partial K}\Big] = (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega, \tag{12}$$

which, applying characterizations (9)–(10), becomes

$$\nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q, \nabla\cdot \boldsymbol{u}_1)_\Omega$$

$$+ \sum_{K\in\mathcal{T}_h} \Big[\frac{1}{\nu}\left(\mathcal{M}_K(-\nu\Delta \boldsymbol{u}_1 + \nabla p) - \mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p\mathbf{I}\cdot\boldsymbol{n}]\!]\right), \nu\Delta \boldsymbol{v}_1 + \nabla q)_K \tag{13}$$

$$+\frac{1}{\nu}\left(\mathcal{B}_K\left([\![\nu\,\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p\mathbf{I}\cdot\boldsymbol{n}]\!]\right), \nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q\mathbf{I}\cdot\boldsymbol{n}\right)_{\partial K}\Big]$$

$$= (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega + \sum_{K\in\mathcal{T}_h} \frac{1}{\nu}\left(\mathcal{M}_K(\boldsymbol{f}), \nu\Delta \boldsymbol{v}_1 + \nabla q\right)_K.$$

Using this form, in the next sections we will present concrete stabilized finite element methods for both the simplest possible pair ($\mathbb{P}^1/\mathbb{P}^0$ elements) and equal order $\mathbb{P}^1/\mathbb{P}^1$ continuous finite elements.

3. The simplest element $\mathbb{P}^1/\mathbb{P}^0$.

3.1. The method. For this case, the finite element spaces are given by

$$\mathbf{V}_h := \{\boldsymbol{v} \in C^0(\overline{\Omega})^2 : \boldsymbol{v}|_K \in \mathbb{P}^1(K)^2 \,\forall\, K \in \mathcal{T}_h\} \cap H_0^1(\Omega)^2$$

for the velocity, and

$$Q_h^0 := \{q \in L_0^2(\Omega) : q|_K \in \mathbb{P}^0(K) \,\forall\, K \in \mathcal{T}_h\}$$

for the pressure. Using these spaces, we propose the following stabilized method: Find $(\boldsymbol{u}_1, p_0) \in \mathbf{V}_h \times Q_h^0$ such that

$$\mathbf{B}_0\big((\boldsymbol{u}_1, p_0), (\boldsymbol{v}_1, q_0)\big) = \mathbf{F}_0\big(\boldsymbol{v}_1, q_0\big) \quad \forall\, (\boldsymbol{v}_1, q_0) \in \mathbf{V}_h \times Q_h^0, \tag{14}$$

where

$$\mathbf{B}_0\big((\boldsymbol{u}_1, p_0), (\boldsymbol{v}_1, q_0)\big) := \nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p_0, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q_0, \nabla\cdot \boldsymbol{u}_1)_\Omega$$

$$+ \sum_{Z\in\mathcal{E}_h} \tau_Z\left([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q_0\mathbf{I}\cdot\boldsymbol{n}]\!]\right)_Z, \tag{15}$$

$$\mathbf{F}_0\big(\boldsymbol{v}_1, q_0\big) := (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega, \tag{16}$$

and τ_Z is given by

$$\tau_Z := \frac{h_Z}{12\nu}\,. \tag{17}$$

REMARK 3.1. *This method differs somewhat from other existing stabilized finite element methods with discontinuous pressure spaces (see, for example,* [23, 29, 34, 14]).

First, since τ_Z is known exactly, we have no free constants to set. To the authors' knowledge, this is the first time that the stabilization parameter corresponding to jump terms is known exactly. Furthermore, and in contrast to [22]*, the jump terms are derived without the use of a macroelement technique. Finally, another difference is the nature of the jump terms, not only containing pressure jumps, but also the jump on the normal derivative of $\boldsymbol{u}$.*

REMARK 3.2. *One of the drawbacks of the RFB method for the Stokes problem is that, due to the zero boundary condition on the element boundary, there is not a bubble-based enrichment that makes stable the $\mathbb{P}^1/\mathbb{P}^0$ element (see* [6] *for a discussion), and hence, the use of a different boundary condition makes it possible to stabilize the $\mathbb{P}^1/\mathbb{P}^0$ element.*

3.1.1. Derivation of the method. First we note that, using spaces $\mathbf{V}_h$ and Q_h^0, (13) reduces to the following: Find $(\boldsymbol{u}_1, p_0) \in \mathbf{V}_h \times Q_h^0$ such that

$$\begin{aligned} &\nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p_0, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q_0, \nabla\cdot \boldsymbol{u}_1)_\Omega \\ &+ \sum_{Z\in\mathcal{E}_h} \frac{1}{\nu} \left(\mathcal{B}_K\left([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!]\right), [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q_0\mathbf{I}\cdot\boldsymbol{n}]\!]\right)_Z = (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega \end{aligned} \tag{18}$$

for all $(\boldsymbol{v}_1, q_0) \in \mathbf{V}_h \times Q_h^0$.

REMARK 3.3. *Since $\mathcal{B}_K$ is the inverse of an elliptic operator, by denoting $\boldsymbol{v} = \mathcal{B}_K(\boldsymbol{g})$, we have, for all $\boldsymbol{g} \in L^2(\partial K)^2$,*

$$(\mathcal{B}_K(\boldsymbol{g}), \boldsymbol{g})_{\partial K} = -(\boldsymbol{v}, \partial_{\boldsymbol{ss}}\boldsymbol{v})_{\partial K} = (\partial_{\boldsymbol{s}}\boldsymbol{v}, \partial_{\boldsymbol{s}}\boldsymbol{v})_{\partial K} \geq 0\,,$$

and hence we are adding a positive term to the formulation.

Next we exploit the fact that $[\![\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!]|_Z$ is a constant function. To do so, we define the (matrix) function $\boldsymbol{b}_K^u := (\mathcal{B}_K(\boldsymbol{e}_1)|\mathcal{B}_K(\boldsymbol{e}_2))$, where $\boldsymbol{e}_1, \boldsymbol{e}_2$ are the canonical vectors in $\mathbb{R}^2$, and we remark that, from its definition, $\boldsymbol{b}_K^u = b_K^u\mathbf{I}$, where b_K^u is the solution of

$$-\Delta b_K^u = 0 \quad \text{in } K, \qquad b_K^u = g(\boldsymbol{s}) \quad \text{on each } Z \subset \partial K, \tag{19}$$

where $g = 0$ if $Z \subset \partial\Omega$, and g satisfies

$$-\partial_{\boldsymbol{ss}} g(\boldsymbol{s}) = \frac{1}{h_Z} \quad \text{in } Z\,, \qquad g = 0 \quad \text{at the nodes}\,, \tag{20}$$

in the internal edges.

REMARK 3.4. *The solution of* (20) *may be calculated explicitly and it is not difficult to realize that*

$$\frac{(b_K^u, 1)_Z}{|Z|} = \frac{h_Z}{12}\,. \tag{21}$$

Finally, since $[\![\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!]|_Z$ is a constant function we obtain

$$\begin{aligned} &(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!]), [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q_0\mathbf{I}\cdot\boldsymbol{n}]\!])_Z \\ &= \left[\int_Z \boldsymbol{b}_K^u\right] [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!]\Big|_Z \cdot [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q_0\mathbf{I}\cdot\boldsymbol{n}]\!]\Big|_Z \\ &= \frac{(b_K^u, 1)_Z}{|Z|} ([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1 + p_0\mathbf{I}\cdot\boldsymbol{n}]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1 + q_0\mathbf{I}\cdot\boldsymbol{n}]\!])_Z\,, \end{aligned}$$

and hence replacing this in (18) and using the previous remark, we obtain method (14).

3.2. Error analysis. From now on, C will denote a positive constant independent of h and ν, and that may change its value whenever it is written in two different places.

The next result states the consistency of the proposed method.

LEMMA 3.5. *Let* $(\boldsymbol{u}, p) \in [H^2(\Omega) \cap H_0^1(\Omega)]^2 \times [H^1(\Omega) \cap L_0^2(\Omega)]$ *be the weak solution of* (1) *and* $(\boldsymbol{u}_1, p_0)$ *the solution of* (14). *Then,*

$$\mathbf{B}_0\big((\boldsymbol{u} - \boldsymbol{u}_1, p - p_0), (\boldsymbol{v}_1, q_0)\big) = 0 \quad \forall\, (\boldsymbol{v}_1, q_0) \in \mathbf{V}_h \times Q_h^0. \tag{22}$$

Proof. The results follows by noting that $[\![\nu \partial_{\boldsymbol{n}} \boldsymbol{u} + p\mathbf{I}\cdot \boldsymbol{n}]\!] = \mathbf{0}$ a.e. across all the internal edges. ∎

Moreover, defining the mesh-dependent norm

$$\|(\boldsymbol{v}, q)\|_h := \left[\nu\, |\boldsymbol{v}|_{1,\Omega}^2 + \sum_{Z \in \mathcal{E}_h} \tau_Z \|[\![\nu \partial_{\boldsymbol{n}} \boldsymbol{v} + q\mathbf{I}\cdot \boldsymbol{n}]\!]\|_{0,Z}^2 \right]^{\frac{1}{2}}, \tag{23}$$

we have the following continuity and coercivity results.

LEMMA 3.6. *Let be* $(\boldsymbol{v}, q), (\boldsymbol{w}, r) \in [H^2(\mathcal{T}_h) \cap H_0^1(\Omega)]^2 \times [H^1(\mathcal{T}_h) \cap L_0^2(\Omega)]$. *Then, bilinear form* $\mathbf{B}_0$ *satisfies*

$$\mathbf{B}_0\big((\boldsymbol{v}, q), (\boldsymbol{w}, r)\big) \le \|(\boldsymbol{v}, q)\|_h \|(\boldsymbol{w}, r)\|_h + (\nabla\cdot \boldsymbol{v}, r)_\Omega - (q, \nabla\cdot \boldsymbol{w})_\Omega\,, \tag{24}$$

$$\mathbf{B}_0\big((\boldsymbol{v}, q), (\boldsymbol{v}, q)\big) = \|(\boldsymbol{v}, q)\|_h^2\,. \tag{25}$$

Proof. The result follows immediately from the definition of $\mathbf{B}_0$. ∎

In order to perform the numerical analysis of this method, we will consider the Lagrange interpolation operator $I_h : C^0(\overline{\Omega}) \to V_h$ (if $\boldsymbol{v} = (v_1, v_2) \in C^0(\overline{\Omega})^2$, we denote $I_h(\boldsymbol{v}) = (I_h(v_1), I_h(v_2))$) to approximate the velocity. Then, it is well known (cf. [17]) that

$$|v - I_h(v)|_{m,K} \le C\, h_K^{2-m}\, |v|_{2,K} \qquad \forall v \in H^2(K), \tag{26}$$

$$|v - I_h(v)|_{t,Z} \le C\, h_Z^{2-t-1/2}\, |v|_{2,\omega_Z} \qquad \forall v \in H^2(\omega_Z) \tag{27}$$

for all $K \in \mathcal{T}_h, Z \in \mathcal{E}_h$, where $\omega_Z := \cup\{K \in \mathcal{T}_h : Z \subset \partial K\}$, and $m = 0, 1, 2$, $t = 0, 1$. Let us remark that to obtain the second estimate above, we used the following local trace theorem (for a proof, see [33]): *There exists* $C > 0$, *independent of* h, *such that*

$$\|v\|_{0,\partial K}^2 \le C \left(\frac{1}{h_K} \|v\|_{0,K}^2 + h_K\, |v|_{1,K}^2 \right) \tag{28}$$

for all $v \in H^1(K)$.

In order to approximate the pressure we will consider $\Pi_h : L^2(\Omega) \to Q_h^0$ as the $L^2(\Omega)$-projection onto Q_h^0. This projection satisfies (cf. [17])

$$\|q - \Pi_h(q)\|_{0,\Omega} \le C\, h\, |q|_{1,\Omega} \tag{29}$$

if $q \in H^1(\Omega)$, and hence, using the local trace theorem (28), we obtain

$$\left[\sum_{Z\in\mathcal{E}_h} h_Z \,\|[\![q-\Pi_h(q)]\!]\|_{0,Z}^2\right]^{\frac{1}{2}} \le C\,h\,|q|_{1,\Omega} \tag{30}$$

for all $q \in H^1(\Omega)$.

LEMMA 3.7. *Suppose* $(\boldsymbol{v}, q) \in H^2(\Omega)^2 \times H^1(\Omega)$. *Then,*

$$\|(\boldsymbol{v}-I_h(\boldsymbol{v}), q-\Pi_h(q))\|_h \le Ch\left(\sqrt{\nu}\,|\boldsymbol{v}|_{2,\Omega} + \frac{1}{\sqrt{\nu}}\,|q|_{1,\Omega}\right). \tag{31}$$

Proof. The result follows immediately from the norm definition and (26), (27), (30). ∎

Using previous results we can establish the following convergence result.

THEOREM 3.8. *Let* $(\boldsymbol{u}, p) \in [H^2(\Omega)\cap H_0^1(\Omega)]^2 \times [H^1(\Omega)\cap L_0^2(\Omega)]$ *be the solution of* (1) *and* $(\boldsymbol{u}_1, p_0)$ *the solution of* (14). *Then, the following error estimate holds:*

$$\|(\boldsymbol{u}-\boldsymbol{u}_1, p-p_0)\|_h \le Ch\left(\sqrt{\nu}\,|\boldsymbol{u}|_{2,\Omega} + \frac{1}{\sqrt{\nu}}\,|p|_{1,\Omega}\right). \tag{32}$$

Proof. Let $(\tilde{\boldsymbol{u}}_h, \tilde{p}_h) := (I_h(\boldsymbol{u}), \Pi_h(p)) \in \mathbf{V}_h \times Q_h^0$. From Lemmas 3.5 and 3.6 we know that

$$\begin{aligned}
\|(\boldsymbol{u}-\boldsymbol{u}_1, p-p_0)\|_h^2 &= \mathbf{B}_0\big((\boldsymbol{u}-\boldsymbol{u}_1, p-p_0), (\boldsymbol{u}-\boldsymbol{u}_1, p-p_0)\big)\\
&= \mathbf{B}_0\big((\boldsymbol{u}-\boldsymbol{u}_1, p-p_0), (\boldsymbol{u}-\tilde{\boldsymbol{u}}_h, p-\tilde{p}_h)\big)\\
&\le C\,\|(\boldsymbol{u}-\boldsymbol{u}_1, p-p_0)\|_h\,\|(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h, p-\tilde{p}_h)\|_h\\
&\quad + (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_1), p-\tilde{p}_h)_\Omega - (\nabla\cdot(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h), p-p_0)_\Omega\,.
\end{aligned}$$

Now,

$$(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_1), p-\tilde{p}_h)_\Omega = -(\nabla\cdot\boldsymbol{u}_1, p-\tilde{p}_h)_\Omega = 0 \tag{33}$$

since $\boldsymbol{u}$ is a solenoidal field and $\nabla\cdot\boldsymbol{u}_1 \in Q_h^0$. On the other hand,

$$\begin{aligned}
(\nabla\cdot(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h), p-p_0)_\Omega &= \sum_{K\in\mathcal{T}_h}\Big[-(\nabla p, \boldsymbol{u}-\tilde{\boldsymbol{u}}_h)_K + (p-p_0, (\boldsymbol{u}-\tilde{\boldsymbol{u}}_h)\cdot\boldsymbol{n})_{\partial K}\Big]\\
&= -(\nabla p, \boldsymbol{u}-\tilde{\boldsymbol{u}}_h)_\Omega + \sum_{K\in\mathcal{T}_h}((p-p_0)\mathbf{I}\cdot\boldsymbol{n}, \boldsymbol{u}-\tilde{\boldsymbol{u}}_h)_{\partial K}\\
&\le |p|_{1,\Omega}\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_h\|_{0,\Omega} + \sum_{Z\in\mathcal{E}_h}([\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!], \boldsymbol{u}-\tilde{\boldsymbol{u}}_h)_Z\\
&\le Ch^2\,|p|_{1,\Omega}|\boldsymbol{u}|_{2,\Omega} + C\sum_{Z\in\mathcal{E}_h}\frac{h_Z^{\frac{3}{2}}}{\sqrt{\nu}}\,\|[\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}\sqrt{\nu}\,|\boldsymbol{u}|_{2,\omega_Z}\\
&\le Ch^2\,|p|_{1,\Omega}|\boldsymbol{u}|_{2,\Omega} + \frac{1}{\gamma}\sum_{Z\in\mathcal{E}_h}\frac{h_Z}{\nu}\,\|[\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 + C\gamma\sum_{Z\in\mathcal{E}_h}h^2\nu\,|\boldsymbol{u}|_{2,\omega_Z}^2\\
&\le Ch^2\left((1+\gamma)\,\nu\,|\boldsymbol{u}|_{2,\Omega}^2 + \frac{1}{\nu}|p|_{1,\Omega}^2\right) + \frac{1}{\gamma}\sum_{Z\in\mathcal{E}_h}\frac{h_Z}{\nu}\,\|[\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2\,,
\end{aligned}$$

where $\gamma > 0$. Now, using the local trace theorem (28) and the fact that $\mathbf{V}_h$ is constituted by linear polynomials we arrive at

$$\begin{aligned}
&\sum_{Z\in\mathcal{E}_h} \frac{h_Z}{\nu} \|[\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 \\
&\leq 2\sum_{Z\in\mathcal{E}_h} \frac{h_Z}{\nu} \left(\|[\![\nu\,\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 + \|[\![\nu\,\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}^2 \right) \\
&\leq C\left(\sum_{Z\in\mathcal{E}_h} \left[\frac{h_Z}{\nu} \|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 \right] + \nu\,|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}^2 + \nu\,h^2\,|\boldsymbol{u}|_{2,\Omega}^2 \right) \\
&\leq \tilde{C}\,\|(\boldsymbol{u}-\boldsymbol{u}_1,p-p_0)\|_h^2 + C\,\nu\,h^2\,|\boldsymbol{u}|_{2,\Omega}^2\,.
\end{aligned}$$

Hence, choosing $\gamma = 2\tilde{C}$ we obtain

$$\frac{1}{2}\,\|(\boldsymbol{u}-\boldsymbol{u}_1,p-p_0)\|_h^2 \leq Ch^2\left(\nu\,|\boldsymbol{u}|_{2,\Omega}^2+\frac{1}{\nu}|p|_{1,\Omega}^2\right), \tag{34}$$

and the result follows by extracting the square root. ∎

REMARK 3.9. *The last result gives a convergence result for the velocity, plus a convergence result for the jump terms. More precisely, this result implies* $|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega} \leq C\,h$ *and* $\left[\sum_{Z\in\mathcal{E}_h} h_Z\|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2\right]^{\frac{1}{2}} \leq C\,h$, *which are both optimal in order and regularity.*

3.2.1. A convergence result for the pressure. The last result of the previous section does not give convergence on the natural norm of the pressure. That is why a convergence result for the pressure in the $L^2(\Omega)$ norm is now given.

In the proof of the next result we will use the Clément interpolation operator (cf. [17, 24]), $\mathcal{C}_h : H^1(\Omega) \to V_h$. This operator satisfies

$$|v-\mathcal{C}_h(v)|_{m,\Omega} \leq C\,h^{1-m}|v|_{1,\Omega} \qquad \forall v\in H^1(\Omega) \tag{35}$$

for $m = 0, 1$, with the obvious extension to vector-valued functions.

THEOREM 3.10. *Let* $(\boldsymbol{u},p) \in [H^2(\Omega)\cap H_0^1(\Omega)]^2 \times [H^1(\Omega)\cap L_0^2(\Omega)]$ *be the solution of* (1) *and* $(\boldsymbol{u}_1,p_0)$ *the solution of* (14). *Then, the following error estimate holds:*

$$\|p-p_0\|_{0,\Omega} \leq C\,h\left[\nu\,|\boldsymbol{u}|_{2,\Omega}+\,|p|_{1,\Omega}\right]. \tag{36}$$

Proof. From the continuous inf-sup condition (see [24]), there exists $\boldsymbol{w}\in H_0^1(\Omega)^2$ such that $\nabla\cdot\boldsymbol{w} = p-p_0$ in Ω and $|\boldsymbol{w}|_{1,\Omega} \leq C\,\|p-p_0\|_{0,\Omega}$. Let $\boldsymbol{w}_h = \mathcal{C}_h(\boldsymbol{w})$. Then,

applying the consistency of the method we obtain

$$
\begin{aligned}
\|p-p_0\|_{0,\Omega}^2 \;&=\; (\nabla\cdot\boldsymbol{w}, p-p_0)_\Omega \\
&=\; (\nabla\cdot(\boldsymbol{w}-\boldsymbol{w}_h), p-p_0)_\Omega \;+\; (\nabla\cdot\boldsymbol{w}_h, p-p_0)_\Omega \\
&=\; \sum_{K\in\mathcal{T}_h} [-(\boldsymbol{w}-\boldsymbol{w}_h, \nabla p)_K \;+\; (\boldsymbol{w}-\boldsymbol{w}_h, (p-p_0)\mathbf{I}\cdot\boldsymbol{n})_{\partial K}] \\
&\quad +\; \nu\,(\nabla(\boldsymbol{u}-\boldsymbol{u}_1), \nabla\boldsymbol{w}_h)_\Omega \;+\; \sum_{Z\in\mathcal{E}_h} \tau_Z\,([\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!])_Z \\
&\leq \left[\sum_{K\in\mathcal{T}_h} \frac{h_K^2}{\nu}\,|p|_{1,K}^2 \;+\; \sum_{Z\in\mathcal{E}_h} \tau_Z \|[\![(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 \;+\; \nu|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}^2 \right. \\
&\qquad\qquad \left. +\; \sum_{Z\in\mathcal{E}_h} \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2\right]^{\frac{1}{2}} \\
&\quad \cdot \left[\sum_{K\in\mathcal{T}_h} \frac{\nu}{h_K^2}\,\|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,K}^2 \;+\; \sum_{Z\in\mathcal{E}_h} \tau_Z^{-1}\,\|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,Z}^2 \right. \\
&\qquad\qquad \left. +\,\nu\,|\boldsymbol{w}_h|_{1,\Omega}^2 \;+\; \sum_{Z\in\mathcal{E}_h} \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!]\|_{0,Z}^2\right]^{\frac{1}{2}}.
\end{aligned}
$$

Now, using the local trace theorem (28) and (35) we easily obtain

$$
\begin{aligned}
&\left[\sum_{K\in\mathcal{T}_h} \frac{\nu}{h_K^2}\|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,K}^2 + \sum_{Z\in\mathcal{E}_h} \tau_Z^{-1}\|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,Z}^2 + \nu|\boldsymbol{w}_h|_{1,\Omega}^2 + \sum_{Z\in\mathcal{E}_h} \tau_Z\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!]\|_{0,Z}^2\right]^{\frac{1}{2}} \\
&\qquad\qquad \leq C\,\sqrt{\nu}\,|\boldsymbol{w}|_{1,\Omega} \;\leq\; C\,\sqrt{\nu}\,\|p-p_0\|_{0,\Omega}\,.
\end{aligned}
$$

Hence, dividing by $\|p-p_0\|_{0,\Omega}$ and using (28) again we have

$$
\begin{aligned}
&\|p-p_0\|_{0,\Omega} \\
&\leq\; C\,\sqrt{\nu}\left[\sum_{K\in\mathcal{T}_h} \frac{h_K^2}{\nu}\,|p|_{1,K}^2 \;+\; \nu\,|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}^2 \right. \\
&\qquad \left. +\; \sum_{Z\in\mathcal{E}_h} \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\mathbf{I}\cdot\boldsymbol{n}]\!]\|_{0,Z}^2 \;+\; \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}^2\right]^{\frac{1}{2}} \\
&\leq\; C\,\sqrt{\nu}\left[\frac{h^2}{\nu}\,|p|_{1,\Omega}^2 \;+\; \|(\boldsymbol{u}-\boldsymbol{u}_1, p-p_0)\|_h^2 \;+\; \nu h^2|\boldsymbol{u}|_{2,\Omega}^2\right]^{\frac{1}{2}} \\
&\leq\; C\,\sqrt{\nu}\left[\frac{h^2}{\nu}\,|p|_{1,\Omega}^2 \;+\; \nu h^2|\boldsymbol{u}|_{2,\Omega}^2\right]^{\frac{1}{2}},
\end{aligned}
$$

and the result follows. □

3.2.2. An error estimate for $\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}$. Throughout this section we will assume that the solution of the following problem, where $(\boldsymbol{u}_1, p_0)$ is the solution of

(14), belongs to $[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$: Find $(\boldsymbol{\varphi},\pi)$ such that

$$-\nu\,\Delta\,\boldsymbol{\varphi}\,-\,\nabla\pi\,=\,\boldsymbol{u}-\boldsymbol{u}_1,\qquad \nabla\cdot\,\boldsymbol{\varphi}\,=\,0\qquad \text{in }\Omega\,,\tag{37}$$
$$\boldsymbol{\varphi}=\mathbf{0}\qquad \text{on }\partial\Omega.$$

We also assume that the following estimate holds:

$$\nu\,\|\boldsymbol{\varphi}\|_{2,\Omega}\,+\,\|\pi\|_{1,\Omega}\,\leq\,C\,\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}\,.\tag{38}$$

THEOREM 3.11. *Let $(\boldsymbol{u},p)\in[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$ be the solution of* (1) *and $(\boldsymbol{u}_1,p_0)$ the solution of* (14)*. Then, the following error estimate holds:*

$$\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}\,\leq\,C\,h^2\,\left(\,|\boldsymbol{u}|_{2,\Omega}\,+\,\frac{1}{\nu}\,|p|_{1,\Omega}\right).$$

Proof. Let $(\boldsymbol{\varphi}_h,\pi_h):=(I_h(\boldsymbol{\varphi}),\Pi_h(\pi))\in\mathbf{V}_h\times Q_h^0$. Then, multiplying the first equation in (37) by $\boldsymbol{u}-\boldsymbol{u}_1$ and the second by $-(p-p_0)$, from the definition of bilinear form $\mathbf{B}_0$, the regularity of $(\boldsymbol{\varphi},\pi)$ and the consistency of the method and Lemma 3.6, we obtain

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}^2 &= \nu\,(\nabla\boldsymbol{\varphi},\nabla(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega\,+\,(\pi,\nabla\cdot\,(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega\,-\,(p-p_0,\nabla\cdot\,\boldsymbol{\varphi})_\Omega\\
&= \mathbf{B}_0((\boldsymbol{u}-\boldsymbol{u}_1,p-p_0),(\boldsymbol{\varphi},\pi))\\
&= \mathbf{B}_0((\boldsymbol{u}-\boldsymbol{u}_1,p-p_0),(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h,\pi-\pi_h))\\
&\leq \|(\boldsymbol{u}-\boldsymbol{u}_1,p-p_0)\|_h\|(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h,\pi-\pi_h)\|_h\\
&\quad-\,(p-p_0,\nabla\cdot\,(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h))_\Omega\,+\,(\pi-\pi_h,\nabla\cdot\,(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega\,.
\end{aligned}$$

Now, using (33) we see that $(\pi-\pi_h,\nabla\cdot\,(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega=0$, and hence, using interpolation inequalities (26), Lemma 3.7 and Theorems 3.8 and 3.10, we arrive at

$$\begin{aligned}
&\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}^2\\
&\leq \|(\boldsymbol{u}-\boldsymbol{u}_1,p-p_0)\|_h\|(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h,\pi-\pi_h)\|_h\,+\,\|p-p_0\|_{0,\Omega}\|\nabla\cdot\,(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h)\|_{0,\Omega}\\
&\leq\left[\|(\boldsymbol{u}-\boldsymbol{u}_1,p-p_0)\|_h^2+\frac{1}{\nu}\|p-p_0\|_{0,\Omega}^2\right]^{\frac12}\left[\|(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h,\pi-\pi_h)\|_h^2+\nu\|\nabla\cdot\,(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h)\|_{0,\Omega}^2\right]^{\frac12}\\
&\leq Ch^2\,\left[\nu\,|\boldsymbol{u}|_{2,\Omega}^2+\frac{1}{\nu}\,|p|_{1,\Omega}^2\right]^{\frac12}\,\left[\nu\,|\boldsymbol{\varphi}|_{2,\Omega}^2+\frac{1}{\nu}\,|\pi|_{1,\Omega}^2\right]^{\frac12}\\
&\leq C\frac{1}{\sqrt{\nu}}h^2\,\left(\sqrt{\nu}\,|\boldsymbol{u}|_{2,\Omega}+\frac{1}{\sqrt{\nu}}\,|p|_{1,\Omega}\right)\,\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}\,,
\end{aligned}$$

and the result follows. ∎

4. The method using $\mathbb{P}^1/\mathbb{P}^1$ continuous elements.

4.1. The method. For this case, the finite element space for the velocity is the same as in previous section, but the pressure space is now given by

$$Q_h^1\,:=\,\{q\in C^0(\overline{\Omega})\,:\,q|_K\in\mathbb{P}^1(K)\,\forall\,K\in\mathcal{T}_h\,\}\cap L_0^2(\Omega)\,.$$

As we will see in next section, the method coming directly from (13) is given by the following: Find $(\tilde{\boldsymbol{u}}_1,\tilde{p}_1)\in\mathbf{V}_h\times Q_h^1$ such that

$$\mathbf{B}_1((\tilde{\boldsymbol{u}}_1,\tilde{p}_1),(\boldsymbol{v}_1,q_1))\,=\,\mathbf{F}(\boldsymbol{v}_1,q_1)\qquad\forall\,(\boldsymbol{v}_1,q_1)\in\mathbf{V}_h\times Q_h^1\,,\tag{39}$$

where

$$\mathbf{B}_1((\boldsymbol{u}_1,p_1),(\boldsymbol{v}_1,q_1)) := \mathbf{B}((\boldsymbol{u}_1,p_1),(\boldsymbol{v}_1,q_1)) - \sum_{K\in\mathcal{T}_h}\frac{1}{\nu}\left(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]),\nabla q_1\right)_K, \tag{40}$$

with

$$\begin{aligned}\mathbf{B}\big((\boldsymbol{u}_1,p_1),(\boldsymbol{v}_1,q_1)\big) &:= \nu(\nabla\boldsymbol{u}_1,\nabla\boldsymbol{v}_1)_\Omega - (p_1,\nabla\cdot\boldsymbol{v}_1)_\Omega + (q_1,\nabla\cdot\boldsymbol{u}_1)_\Omega\\ &+ \sum_{K\in\mathcal{T}_h}\tau_K\left(-\nu\Delta\boldsymbol{u}_1+\nabla p_1,\nu\Delta\boldsymbol{v}_1+\nabla q_1\right)_K + \sum_{Z\in\mathcal{E}_h}\tau_Z\left([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!],[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\right)_Z,\end{aligned} \tag{41}$$

$$\mathbf{F}\big(\boldsymbol{v}_1,q_1\big) := (\boldsymbol{f},\boldsymbol{v}_1)_\Omega + \sum_{K\in\mathcal{T}_h}\tau_K\left(\boldsymbol{f},\nu\Delta\boldsymbol{v}_1+\nabla q_1\right)_K, \tag{42}$$

$$\tau_K := C_1\frac{h_K^2}{\nu}, \tag{43}$$

where τ_Z is given by (17) and $C_1=\frac{1}{8}$. The value $C_1=\frac{1}{8}$ has been suggested by the error analysis of original method (39) (see Appendix A).

Now, for reasons that we will justify later (see Theorem 4.3 below), we will drop the term

$$-\sum_{K\in\mathcal{T}_h}\left(\frac{1}{\nu}\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]),\nabla q_1\right)_K$$

and analyze (and implement) the following simplified version of (39): Find $(\boldsymbol{u}_1,p_1)\in\mathbf{V}_h\times Q_h^1$ such that

$$\mathbf{B}((\boldsymbol{u}_1,p_1),(\boldsymbol{v}_1,q_1)) = \mathbf{F}(\boldsymbol{v}_1,q_1) \qquad \forall\,(\boldsymbol{v}_1,q_1)\in\mathbf{V}_h\times Q_h^1\,. \tag{44}$$

REMARK 4.1. *We see that method* (44) *has the form of a stabilized method of the GLS class, plus a nonstandard jump term formed by the residual of the Cauchy stress tensor on the edges of the triangulation. This will give us control of this residual, which is exclusive to continuous pressure spaces, since in that case pressure jumps vanish.*

REMARK 4.2. *The method is written as the restriction of a consistent method to* $\mathbb{P}^1/\mathbb{P}^1$ *elements simply to avoid some technical difficulties. A nonconsistent presentation may be given and in that case we can prove that the consistency error does not imply a loss of precision.*

As we said before, we will perform the error analysis of method (44). This is due to the fact that the error of method (39) is bounded by that of (44), as stated in the following result, whose proof may be found in Appendix A.

THEOREM 4.3. *Let* $(\boldsymbol{u},p)\in H^2(\Omega)^2\times H^1(\Omega)$ *be the solution of* (1). *Then, method* (39) *is consistent. Moreover,* (39) *has a unique solution* $(\tilde{\boldsymbol{u}}_1,\tilde{p}_1)\in\mathbf{V}_h\times Q_h^1$, *and the following error estimate holds:*

$$|\!|\!|\boldsymbol{u}-\tilde{\boldsymbol{u}}_1|\!|\!|_h^2 + \|p-\tilde{p}_1\|_h^2 \le C\left(|\!|\!|\boldsymbol{u}-\boldsymbol{u}_1|\!|\!|_h^2 + \|p-p_1\|_h^2\right),$$

where $(\boldsymbol{u}_1,p_1)\in\mathbf{V}_h\times Q_h^1$ *is the solution of* (44), *and the norms are defined as in* (49)–(50) *below.*

4.1.1. Derivation of the method. Using spaces $\mathbf{V}_h$ and Q_h^1, (13) reduces to the following: Find $(\boldsymbol{u}_1, p_1) \in \mathbf{V}_h \times Q_h^1$ such that

$$\begin{aligned}
&\nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p_1, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q_1, \nabla\cdot \boldsymbol{u}_1)_\Omega \\
&+ \sum_{K\in\mathcal{T}_h} \frac{1}{\nu} \left(\mathcal{M}_K(\nabla p_1) - \mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]), \nabla q_1\right)_K \\
&+ \sum_{Z\in\mathcal{E}_h} \frac{1}{\nu} \left(\mathcal{B}_K\left([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]\right), [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\right)_Z = (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega + \sum_{K\in\mathcal{T}_h} \frac{1}{\nu} \left(\mathcal{M}_K(\boldsymbol{f}), \nabla q_1\right)_K
\end{aligned} \tag{45}$$

for all $(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1$. Since $\nabla p_1|_K \in \mathbb{R}^2$, we have

$$\mathcal{M}_K(\nabla p_1) = \left(\mathcal{M}_K(\boldsymbol{e}_1), \mathcal{M}_K(\boldsymbol{e}_2)\right)\nabla p_1 =: \boldsymbol{b}_K^p\, \nabla p_1.$$

As in the previous section, we see that $\boldsymbol{b}_K^p = b_K^p \mathbf{I}$, where b_K^p is the solution of

$$-\Delta b_K^p = 1 \quad \text{in } K, \qquad b_K^p = 0 \quad \text{on } \partial K\,. \tag{46}$$

Hence

$$(\mathcal{M}_K(\nabla p_1), \nabla q_1)_K = \left[\int_K b_K^p\right] \nabla p_1\Big|_K \cdot \nabla q_1\Big|_K = \frac{(b_K^p, 1)_K}{|K|} (\nabla p_1, \nabla q_1)_K\,.$$

On the other hand, from the previous section we know that

$$(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]), [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!])_Z = \tau_Z\left([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\right)_Z,$$

where τ_Z has been defined in (17). Moreover, if we suppose that $\boldsymbol{f}$ is piecewise constant, we have $\mathcal{M}_K(\boldsymbol{f}) = b_K^p\, \boldsymbol{f}$, and hence, in the same way as before,

$$(\mathcal{M}_K(\boldsymbol{f}), \nabla q_1)_K = \frac{(b_K^p, 1)_K}{|K|} (\boldsymbol{f}, \nabla q_1)_K.$$

Summing all this up, we arrive at the following expression for (45): Find $(\boldsymbol{u}_1, p_1) \in \mathbf{V}_h \times Q_h^1$ such that

$$\begin{aligned}
&\nu(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}_1)_\Omega - (p_1, \nabla\cdot \boldsymbol{v}_1)_\Omega + (q_1, \nabla\cdot \boldsymbol{u}_1)_\Omega + \sum_{K\in\mathcal{T}_h} \frac{(b_K^p, 1)_K}{|K|\nu} (\nabla p_1, \nabla q_1)_K \\
&- \sum_{K\in\mathcal{T}_h} \frac{1}{\nu} (\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!]), \nabla q_1)_K + \sum_{Z\in\mathcal{E}_h} \frac{(b_K^u, 1)_Z}{|Z|\nu} ([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!])_Z \\
&= (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega + \sum_{K\in\mathcal{T}_h} \frac{(b_K^p, 1)_K}{|K|\nu} (\boldsymbol{f}, \nabla q_1)_K
\end{aligned} \tag{47}$$

for all $(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1$. Finally, since the mesh is regular by a scaling argument (cf. [31]) we have that

$$\frac{1}{|K|}(b_K^p, 1)_K \sim C_1\, h_K^2\,, \tag{48}$$

where C_1 is a positive constant independent of h and ν. Hence, replacing (48) in (47) and defining τ_K appropriately, we obtain method (39).

REMARK 4.4. *The assumption of the piecewise constant $\boldsymbol{f}$ on the right-hand side is made simply to derive the method, but it does not affect the precision of it. Indeed, if we consider a general $\boldsymbol{f} \in H^1(\Omega)^2$ and take its projection onto the space of piecewise constant functions, we keep the same order of convergence of the method* (*see Appendix* B).

4.2. Error analysis. Let us consider the mesh-dependent norms

$$|||\boldsymbol{v}|||_h^2 := \nu\,|\boldsymbol{v}|_{1,\Omega}^2 + \sum_{Z\in\mathcal{E}_h} \tau_Z \|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}]\!]\|_{0,Z}^2\,, \tag{49}$$

$$\|q\|_h^2 := \sum_{K\in\mathcal{T}_h} \tau_K |q|_{1,K}^2\,. \tag{50}$$

The first results concern the consistency and well-posedness of stabilized method (44).

LEMMA 4.5. *Let* $(\boldsymbol{u},p)\in[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$ *be the solution of* (1) *and* $(\boldsymbol{u}_1,p_1)$ *the solution of* (44). *Then,*

$$\mathbf{B}\big((\boldsymbol{u}-\boldsymbol{u}_1,p-p_1),(\boldsymbol{v}_1,q_1)\big)=0\quad\forall\,(\boldsymbol{v}_1,q_1)\in\mathbf{V}_h\times Q_h^1.$$

Proof. The result follows from the definition of $\mathbf{B}$ and the fact that $[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}]\!]=\mathbf{0}$ a.e. on the internal edges. ∎

LEMMA 4.6. *Let be* $(\boldsymbol{v}_1,q_1)\in\mathbf{V}_h\times Q_h^1$. *Then*

$$\mathbf{B}\big((\boldsymbol{v}_1,q_1),(\boldsymbol{v}_1,q_1)\big)=|||\boldsymbol{v}_1|||_h^2+\|q_1\|_h^2.$$

Proof. The result follows from the definition of $\mathbf{B}$ and the fact that $\Delta\boldsymbol{v}_1=\mathbf{0}$ in each $K\in\mathcal{T}_h$. ∎

Now, in order to approximate the velocity we will consider the Lagrange interpolation operator as in the previous section and for the pressure interpolation we will use the Clément interpolation operator $\mathcal{C}_h$ satisfying (35).

The following approximation result will be useful in what follows.

LEMMA 4.7. *Let* $(\boldsymbol{v},q)\in[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$ *and* $\tilde{q}_h:=\mathcal{C}_h(q)-\frac{(\mathcal{C}_h(q),1)_\Omega}{|\Omega|}$. *Then,*

$$|||\boldsymbol{v}-I_h(\boldsymbol{v})|||_h^2+\sum_{K\in\mathcal{T}_h}\Big[\tau_K^{-1}\|\boldsymbol{v}-I_h(\boldsymbol{v})\|_{0,K}^2+\nu h_K^2\|\Delta(\boldsymbol{v}-I_h(\boldsymbol{v}))\|_{0,K}^2\Big]\le Ch^2\,\nu\,|\boldsymbol{v}|_{2,\Omega}^2, \tag{51}$$

$$\|q-\tilde{q}_h\|_h+\frac{1}{\sqrt{\nu}}\|q-\tilde{q}_h\|_{0,\Omega}\le C\,\frac{h}{\sqrt{\nu}}\,|q|_{1,\Omega}. \tag{52}$$

Proof. The result follows from the norm definition and using $\|q-\tilde{q}_h\|_{0,\Omega}\le\|q-\mathcal{C}_h(q)\|_{0,\Omega}$ combined with (26), (27), and (35). ∎

Using Lemmas 4.5–4.7 we can establish the following convergence result.

THEOREM 4.8. *Let* $(\boldsymbol{u},p)\in[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$ *be the solution of* (1) *and* $(\boldsymbol{u}_1,p_1)$ *the solution of* (44). *Then, the following error estimate holds:*

$$|||\boldsymbol{u}-\boldsymbol{u}_1|||_h+\|p-p_1\|_h\le C\,h\,\left[\sqrt{\nu}\,|\boldsymbol{u}|_{2,\Omega}+\frac{1}{\sqrt{\nu}}\,|p|_{1,\Omega}\right]. \tag{53}$$

Proof. Let $\tilde{\boldsymbol{u}}_h:=I_h(\boldsymbol{u}),\tilde{p}_h:=\mathcal{C}_h(p)-\frac{(\mathcal{C}_h(p),1)_\Omega}{|\Omega|}$ and $(\eta^{\boldsymbol{u}},\eta^p):=(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h,p-\tilde{p}_h)$. Applying Lemma 4.6 and the consistency of the method, and integrating by parts we

have

$$
\begin{aligned}
&|||\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h|||_h^2+\|p_1-\tilde{p}_h\|_h^2 = \mathbf{B}((\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h,p_1-\tilde{p}_h),(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h,p_1-\tilde{p}_h))\\
&= \mathbf{B}((\eta^{\boldsymbol{u}},\eta^p),(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h,p_1-\tilde{p}_h))\\
&= \nu\,(\nabla\eta^{\boldsymbol{u}},\nabla(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h))_\Omega - (\eta^p,\nabla\cdot(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h))_\Omega - (\eta^{\boldsymbol{u}},\nabla(p_1-\tilde{p}_h))_\Omega\\
&\quad+\sum_{K\in\mathcal{T}_h}\tau_K\,(-\nu\Delta\eta^{\boldsymbol{u}},\nabla(p_1-\tilde{p}_h))_K+\sum_{K\in\mathcal{T}_h}\tau_K\,(\nabla\eta^p,\nabla(p_1-\tilde{p}_h))_K\\
&\quad+\sum_{Z\in\mathcal{E}_h}\tau_Z\,([\![\nu\partial_{\boldsymbol{n}}\eta^{\boldsymbol{u}}]\!],[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h)]\!])_Z\\
&\le\left[\nu\,|\eta^{\boldsymbol{u}}|_{1,\Omega}^2+\frac{1}{\nu}\,\|\eta^p\|_{0,\Omega}^2+\sum_{K\in\mathcal{T}_h}(\,\tau_K^{-1}\,\|\eta^{\boldsymbol{u}}\|_{0,K}^2+\nu^2\tau_K\,\|\Delta\eta^{\boldsymbol{u}}\|_{0,K}^2\,)\right.\\
&\quad\left.+\|\eta^p\|_h^2+\sum_{Z\in\mathcal{E}_h}\tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}\eta^{\boldsymbol{u}}]\!]\|_{0,Z}^2\right]^{\frac12}\\
&\quad\cdot\left[3\nu\,|\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h|_{1,\Omega}^2+3\sum_{K\in\mathcal{T}_h}\tau_K\,\|\nabla(p_1-\tilde{p}_h)\|_{0,K}^2+\sum_{Z\in\mathcal{E}_h}\tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h)]\!]\|_{0,Z}^2\right]^{\frac12}\\
&\quad\le\sqrt{3}\left[|||\eta^{\boldsymbol{u}}|||_h^2+\sum_{K\in\mathcal{T}_h}[\tau_K^{-1}\,\|\eta^{\boldsymbol{u}}\|_{0,K}^2+\nu h_K^2\|\Delta\eta^{\boldsymbol{u}}\|_{0,K}^2]+\|\eta^p\|_h^2+\frac{1}{\nu}\,\|\eta^p\|_{0,\Omega}^2\right]^{\frac12}\\
&\qquad\cdot\left[|||\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h|||_h^2+\|p_1-\tilde{p}_h\|_h^2\right]^{\frac12}.
\end{aligned}
$$

Hence, dividing by the last term and applying Lemma 4.7 we arrive at

$$
|||\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_h|||_h+\|p_1-\tilde{p}_h\|_h\le C\left[\nu h^2\,|\boldsymbol{u}|_{2,\Omega}^2+\frac{h^2}{\nu}\,|p|_{1,\Omega}^2\right]^{\frac12}. \tag{54}
$$

The result follows using triangular inequality and Lemma 4.7 once more. ∎

REMARK 4.9. *In particular, from the previous theorem we have an $O(h)$ convergence for $|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}$ and $\left[\sum_{Z\in\mathcal{E}_h}h_Z\,\|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}^2\right]^{\frac12}$, which are both optimal in order and regularity.*

4.2.1. A convergence result for the pressure. In the last result of the previous section we had an error estimate in the velocity, but, due to the norm definition, we did not guarantee the convergence of the pressure. The next result shows that we have an optimal error estimate in the natural norm of the pressure, which is independent of ν.

THEOREM 4.10. *Let $(\boldsymbol{u},p)\in[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$ be the solution of (1) and $(\boldsymbol{u}_1,p_1)$ the solution of (44). Then, the following error estimate holds:*

$$
\|p-p_1\|_{0,\Omega}\le C\,h\left[\nu\,|\boldsymbol{u}|_{2,\Omega}+|p|_{1,\Omega}\right]. \tag{55}
$$

Proof. From the continuous inf-sup condition (see [24]), there exists $\boldsymbol{w}\in H_0^1(\Omega)^2$ such that $\nabla\cdot\boldsymbol{w}=p-p_1$ in Ω and $|\boldsymbol{w}|_{1,\Omega}\le C\,\|p-p_1\|_{0,\Omega}$. Let $\boldsymbol{w}_h=\mathcal{C}_h(\boldsymbol{w})\in\mathbf{V}_h$.

Then, applying the consistency of the method, (35), and previous theorem, we obtain

$$\begin{aligned}
\|p-p_1\|_{0,\Omega}^2 &= (\nabla\cdot\boldsymbol{w}, p-p_1)_\Omega \,=\, (\nabla\cdot(\boldsymbol{w}-\boldsymbol{w}_h), p-p_1)_\Omega \,+\, (\nabla\cdot\boldsymbol{w}_h, p-p_1)_\Omega \\
&= -\sum_{K\in\mathcal{T}_h} (\boldsymbol{w}-\boldsymbol{w}_h, \nabla(p-p_1))_K \,+\, \nu\,(\nabla(\boldsymbol{u}-\boldsymbol{u}_1), \nabla\boldsymbol{w}_h)_\Omega \\
&\quad + \sum_{Z\in\mathcal{E}_h} \tau_Z\,([\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!], [\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!])_Z \\
&\le \sum_{K\in\mathcal{T}_h} \|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,K}|p-p_1|_{1,K} \,+\, \nu\,|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}|\boldsymbol{w}_h|_{1,\Omega} \\
&\quad + \sum_{Z\in\mathcal{E}_h} \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!]\|_{0,Z} \\
&\le \left[\sum_{K\in\mathcal{T}_h} \tau_K\,|p-p_1|_{1,K}^2 + \nu\,|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}^2 \,+\, \sum_{Z\in\mathcal{E}_h} \tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}^2\right]^{\frac12} \\
&\quad \cdot\left[\sum_{K\in\mathcal{T}_h} \tau_K^{-1}\|\boldsymbol{w}-\boldsymbol{w}_h\|_{0,K}^2 \,+\, \nu\,|\boldsymbol{w}_h|_{1,\Omega}^2 \,+\, \sum_{Z\in\mathcal{E}_h} \tau_Z\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!]\|_{0,Z}^2\right]^{\frac12} \\
&\le C\,\sqrt{\nu}\left[|\!|\!|\boldsymbol{u}-\boldsymbol{u}_1|\!|\!|_h + \|p-p_1\|_h\right]\left[|\boldsymbol{w}|_{1,\Omega}^2 \,+\, |\boldsymbol{w}_h|_{1,\Omega}^2\right]^{\frac12} \\
&\le C\,\sqrt{\nu}h\left(\sqrt{\nu}\,|\boldsymbol{u}|_{2,\Omega} \,+\, \frac{1}{\sqrt{\nu}}\,|p|_{1,\Omega}\right)\|p-p_1\|_{0,\Omega}\,,
\end{aligned}$$

where, in order to bound the term $\sum_{Z\in\mathcal{E}_h}\tau_Z\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{w}_h]\!]\|_{0,Z}^2$ we have used the local trace result (28) and $\boldsymbol{w}_h|_K\in\mathbb{P}^1(K)^2$. The result follows then by dividing by the last term. ∎

4.2.2. An error estimate for $\|\boldsymbol{u}-\boldsymbol{u_1}\|_{\boldsymbol{0,\Omega}}$. Throughout this section we will assume that the solution of the following problem, where $(\boldsymbol{u}_1, p_1)$ is the solution of (44), belongs to $[H^2(\Omega)\cap H_0^1(\Omega)]^2\times[H^1(\Omega)\cap L_0^2(\Omega)]$: Find $(\boldsymbol{\varphi},\pi)$ such that

$$\begin{aligned}
-\nu\,\Delta\boldsymbol{\varphi} \,-\, \nabla\pi \,=\, \boldsymbol{u}-\boldsymbol{u}_1, \qquad \nabla\cdot\boldsymbol{\varphi} \,=\, 0 \qquad &\text{in } \Omega\,, \\
\boldsymbol{\varphi} = \mathbf{0} \qquad &\text{on } \partial\Omega\,.
\end{aligned} \tag{56}$$

We also assume that the following estimate holds:

$$\nu\,\|\boldsymbol{\varphi}\|_{2,\Omega} \,+\, \|\pi\|_{1,\Omega} \,\le\, C\,\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}\,. \tag{57}$$

THEOREM 4.11. *Under the hypothesis of Theorem* 4.10 *the following error estimate holds:*

$$\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega} \,\le\, C\,h^2\left(|\boldsymbol{u}|_{2,\Omega} \,+\, \frac{1}{\nu}\,|p|_{1,\Omega}\right).$$

Proof. Let $(\boldsymbol{\varphi}_h,\pi_h) := (I_h(\boldsymbol{\varphi}), \mathcal{C}_h(\pi) - \frac{(\mathcal{C}_h(\pi),1)_\Omega}{|\Omega|}) \in \mathbf{V}_h\times Q_h^1$. Then, multiplying the first equation in (56) by $\boldsymbol{u}-\boldsymbol{u}_1$ and the second by $-(p-p_1)$, from the definition of bilinear form $\mathbf{B}$, the consistency of the method, the fact that $[\![\partial_{\boldsymbol{n}}\boldsymbol{\varphi}]\!] = \mathbf{0}$ a.e. on the internal edges, interpolation inequalities (26), (27), and (35), and Theorems 4.8

and 4.10, we obtain

$$
\begin{aligned}
&\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}^2 \\
&= \nu\,(\nabla\boldsymbol{\varphi},\nabla(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega \,+\, (\pi,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_1))_\Omega \,-\, (p-p_1,\nabla\cdot\boldsymbol{\varphi})_\Omega \\
&= \mathbf{B}((\boldsymbol{u}-\boldsymbol{u}_1,p-p_1),(\boldsymbol{\varphi},\pi)) \,-\, \sum_{K\in\mathcal{T}_h}\tau_K\,(-\nu\Delta(\boldsymbol{u}-\boldsymbol{u}_1)+\nabla(p-p_1),\nu\Delta\boldsymbol{\varphi}+\nabla\pi)_K \\
&= \mathbf{B}((\boldsymbol{u}-\boldsymbol{u}_1,p-p_1),(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h,\pi-\pi_h)) \\
&\quad - \sum_{K\in\mathcal{T}_h}\tau_K\,(-\nu\Delta(\boldsymbol{u}-\boldsymbol{u}_1)+\nabla(p-p_1),\nu\Delta\boldsymbol{\varphi}+\nabla\pi)_K \\
&\le \left[\nu\,|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}^2+\nu\,\|\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_1)\|_{0,\Omega}^2+\frac{1}{\nu}\,\|p-p_1\|_{0,\Omega}^2\right. \\
&\quad \left.+\sum_{Z\in\mathcal{E}_h}\tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)]\!]\|_{0,Z}^2 \,+\, 2\sum_{K\in\mathcal{T}_h}\tau_K\|-\nu\Delta\boldsymbol{u}+\nabla(p-p_1)\|_{0,K}^2\right]^{\frac12}\cdot \\
&\quad \left[\nu\,|\boldsymbol{\varphi}-\boldsymbol{\varphi}_h|_{1,\Omega}^2+\frac{1}{\nu}\,\|\pi-\pi_h\|_{0,\Omega}^2 \,+\, \nu\,\|\nabla\cdot(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h)\|_{0,\Omega}^2 \,+\, \sum_{K\in\mathcal{T}_h}\tau_K\|\nabla(\pi-\pi_h)\|_{0,K}^2\right. \\
&\quad \left.+\sum_{Z\in\mathcal{E}_h}\tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{\varphi}-\boldsymbol{\varphi}_h)]\!]\|_{0,Z}^2 \,+\, \sum_{K\in\mathcal{T}_h}\tau_K\|\nu\Delta\boldsymbol{\varphi}+\nabla\pi\|_{0,K}^2\right]^{\frac12} \\
&\le C\left[|\!|\!|\boldsymbol{u}-\boldsymbol{u}_1|\!|\!|_h^2+\nu h^2|\boldsymbol{u}|_{2,\Omega}^2+\|p-p_1\|_h^2+\frac{1}{\nu}\|p-p_1\|_{0,\Omega}^2\right]^{\frac12}\left[\nu h^2|\boldsymbol{\varphi}|_{2,\Omega}^2+\frac{h^2}{\nu}|\pi|_{1,\Omega}^2\right]^{\frac12} \\
&\le C\,\frac{1}{\sqrt{\nu}}\,h^2\,\left(\sqrt{\nu}\,|\boldsymbol{u}|_{2,\Omega}+\frac{1}{\sqrt{\nu}}\,|p|_{1,\Omega}\right)\,\|\boldsymbol{u}-\boldsymbol{u}_1\|_{0,\Omega}\,,
\end{aligned}
$$

and the result follows. ∎

REMARK 4.12. *As we claimed before, the error analysis is independent of the nature of the* $\boldsymbol{f}$ *on the right-hand side, and hence, we have actually justified method* (44) *for a general* $\boldsymbol{f}\in L^2(\Omega)^2$. *In Appendix* B *we will show that if* $\boldsymbol{f}\in H^1(\Omega)^2$, *then the difference between implementing method* (44) *and* $(\boldsymbol{f},\boldsymbol{v}_1)_\Omega+\sum_{K\in\mathcal{T}_h}\nu^{-1}(\mathcal{M}_K(\boldsymbol{f}),\nu\Delta\boldsymbol{v}_1+\nabla q)_K$ *on the right-hand side is smaller than the order of the method. On the other hand, method* (44) *has been justified for any constant* $C_1>0$, *even if it has been presented with* $C_1=\frac18$.

5. An alternative formulation including the residual on the boundary. In this section we propose another class of methods arising from a different choice of enrichment functions. We will denote by R_h the pressure space according to the choice of elements, i.e., $R_h=Q_h^1$ for $\mathbb{P}^1/\mathbb{P}^1$ elements and $R_h=Q_h^0$ for $\mathbb{P}^1/\mathbb{P}^0$ elements. The proposed method reads as follows: Find $(\boldsymbol{u}^r,p^r)\in\mathbf{V}_h\times R_h$ such that

$$
\mathbf{B}_r((\boldsymbol{u}^r,p^r),(\boldsymbol{v}_1,q))=\mathbf{F}(\boldsymbol{v}_1,q)\qquad\forall\,(\boldsymbol{v}_1,q)\in\mathbf{V}_h\times R_h\,, \tag{58}
$$

where

$$
\begin{aligned}
\mathbf{B}_r((\boldsymbol{u}_1,p),(\boldsymbol{v}_1,q)) &= \nu(\nabla\boldsymbol{u}_1,\nabla\boldsymbol{v}_1)_\Omega-(p,\nabla\cdot\boldsymbol{v}_1)_\Omega+(q,\nabla\cdot\boldsymbol{u}_1)_\Omega \\
&\quad+\sum_{K\in\mathcal{T}_h}\tau_K(-\nu\Delta\boldsymbol{u}_1+\nabla p,\nu\Delta\boldsymbol{v}_1+\nabla q)_K \\
&\quad+\sum_{Z\in\mathcal{E}_h}\tilde{\tau}_Z\,([\![-\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1+p\mathbf{I}\cdot\boldsymbol{n}]\!],[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1+q\mathbf{I}\cdot\boldsymbol{n}]\!])_Z\,,
\end{aligned} \tag{59}
$$

$\mathbf{F}$ is given by (42), τ_K by (43), and

$$\tilde{\tau}_Z := \frac{h_Z}{12\alpha\nu}\,, \tag{60}$$

where $\alpha > 0$ will be fixed in order to have a well-posed problem.

This method may be obtained in the same way as method (14) and (44) by taking the enrichment function $\boldsymbol{u}_e$ to be the solution of (4), together with the boundary conditions

$$-\nu\partial_{\boldsymbol{ss}}\boldsymbol{u}_e = \frac{1}{\alpha h_Z}[\![-\nu\partial_{\boldsymbol{n}}\boldsymbol{u}_1^r + p^r\mathbf{I}\cdot\boldsymbol{n}]\!] \text{ on each } Z\subset\partial K, \qquad \boldsymbol{u}_e = \mathbf{0} \text{ at the nodes}\,, \tag{61}$$

on the internal edges, and $\boldsymbol{u}_e = \mathbf{0}$ on $\partial K\cap\partial\Omega$. In fact, using this choice of enrichment we can perform the same derivation from sections 3 and 4, neglecting once more a cross term appearing in $\mathbb{P}^1/\mathbb{P}^1$ discretization.

REMARK 5.1. *This method is different from* (14) *and* (44) *from two viewpoints. First, the boundary term contains the residual of the Cauchy stress tensor on the trial function. This fact comes from the choice of the enriched part as being a corrector for the residual inside the element and on the boundary. The other difference is the stabilization parameter on the edges. Now, this parameter contains a constant to set.*

Now, let $|\!|\!|.|\!|\!|_h$ be the mesh-dependent norm defined by:

$$|\!|\!|(\boldsymbol{v}_1,q)|\!|\!|_h := \left[\nu\,|\boldsymbol{v}_1|^2_{1,\Omega} + \sum_{K\in\mathcal{T}_h}\tau_K\,|q|^2_{1,K} + \sum_{Z\in\mathcal{E}_h}\tilde{\tau}_Z\,\|[\![q]\!]\|^2_{0,Z}\right]^{\frac{1}{2}}. \tag{62}$$

Then, we have the following coercivity result.

LEMMA 5.2. *Let us suppose that $\alpha > C_t/3$, where $C_t > 0$ is the constant from local trace result* (28)*. Then, for all $(\boldsymbol{v}_1,q)\in\mathbf{V}_h\times R_h$ there holds*

$$\mathbf{B}_r((\boldsymbol{v}_1,q),(\boldsymbol{v}_1,q)) \geq \frac{1}{2}\,|\!|\!|(\boldsymbol{v}_1,q)|\!|\!|^2_h\,.$$

Proof. Let $(\boldsymbol{v}_1,q)\in\mathbf{V}_h\times R_h$. Then, since $\Delta\boldsymbol{v}_1 = \mathbf{0}$ on each $K\in\mathcal{T}_h$, applying local trace result (28) and the definition of $\tilde{\tau}_Z$ we obtain

$$\begin{aligned}
\mathbf{B}_r((\boldsymbol{v}_1,q),(\boldsymbol{v}_1,q)) &= \nu|\boldsymbol{v}_1|^2_{1,\Omega} \\
&\quad + \sum_{K\in\mathcal{T}_h}\tau_K\|\nabla q\|^2_{0,K} + \sum_{Z\in\mathcal{E}_h}\tilde{\tau}_Z(-\nu^2\|[\![\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\|^2_{0,Z} + \|[\![q]\!]\|^2_{0,Z}) \\
&\geq \nu\,|\boldsymbol{v}_1|^2_{1,\Omega} - \frac{C_t}{6\alpha}\nu\sum_{K\in\mathcal{T}_h}|\boldsymbol{v}_1|^2_{1,K} + \sum_{K\in\mathcal{T}_h}\tau_K\,|q|^2_{1,K} + \sum_{Z\in\mathcal{E}_h}\tilde{\tau}_Z\|[\![q]\!]\|^2_{0,Z} \\
&\geq \frac{1}{2}\,|\!|\!|(\boldsymbol{v}_1,q)|\!|\!|^2_h\,,
\end{aligned}$$

an the result follows. ∎

Once this method has been proved to be stable, following a procedure absolutely analogous to those from sections 3 and 4 we can prove the consistency of (58) and perform a complete error analysis of (58), obtaining the same results as in previous sections.

6. Numerical validations.

6.1. An analytical solution: Convergence validation. For this test case, the domain is taken as the square $\Omega = (0,1) \times (0,1)$, $\nu = 1$, and $\boldsymbol{f}$ is set such that the exact solution of our Stokes problem is given by

$$\begin{aligned} u_1(x,y) &= -256x^2(x-1)^2y(y-1)(2y-1)\,, \\ u_2(x,y) &= -u_1(y,x)\,, \\ p(x,y) &= 150(x-0.5)(y-0.5)\,. \end{aligned}$$

We perform convergence analysis for methods (14), (44), and (58) using continuous $\mathbb{P}^1/\mathbb{P}^1$ and $\mathbb{P}^1/\mathbb{P}^0$ elements.

6.1.1. The $\mathbb{P}^1/\mathbb{P}^1$ case. For this case we first depict in Figures 1–2 the convergence history for method (44). The results reproduce our theoretical results showing an $O(h)$ order of convergence for $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$,

$$|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u} - \boldsymbol{u}_1)]\!]|_h := \left[\sum_{Z \in \mathcal{E}_h} h_Z \|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u} - \boldsymbol{u}_1)]\!]\|^2_{0,Z}\right]^{\frac{1}{2}}$$

and $\|p - p_1\|_{0,\Omega}$, and an $O(h^2)$ convergence for $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$.

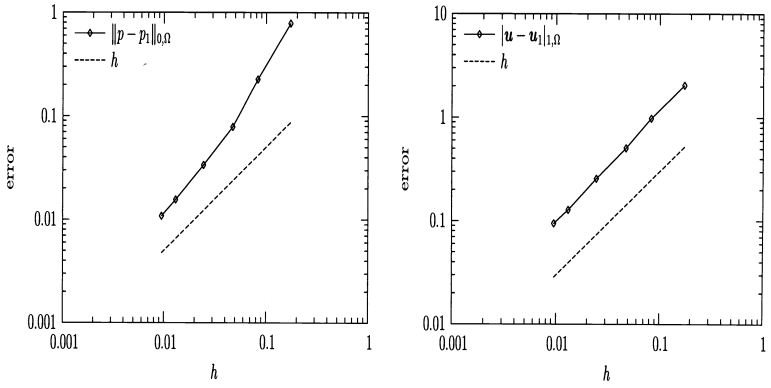


FIG. 1. *Method* (44)*: convergence history for* $\|p - p_1\|_{0,\Omega}$ *and* $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$.

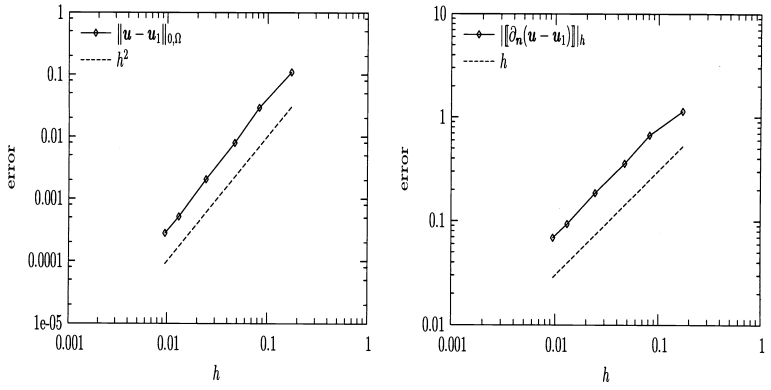


FIG. 2. *Method* (44)*: convergence history for* $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$ *and* $|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u} - \boldsymbol{u}_1)]\!]|_h$.

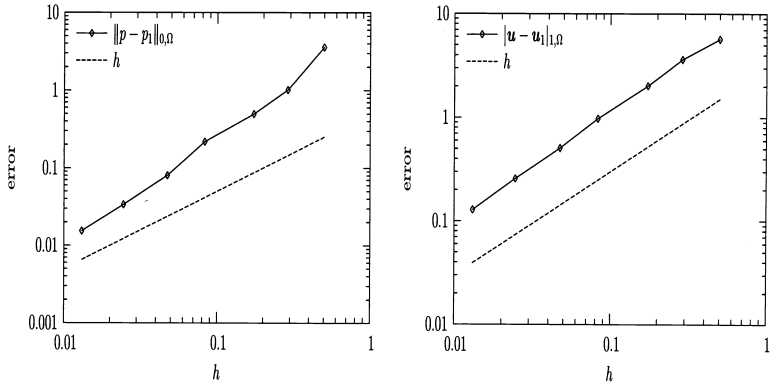


FIG. 3. *Method* (58)*: convergence history for* $\|p - p_1\|_{0,\Omega}$ *and* $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$.

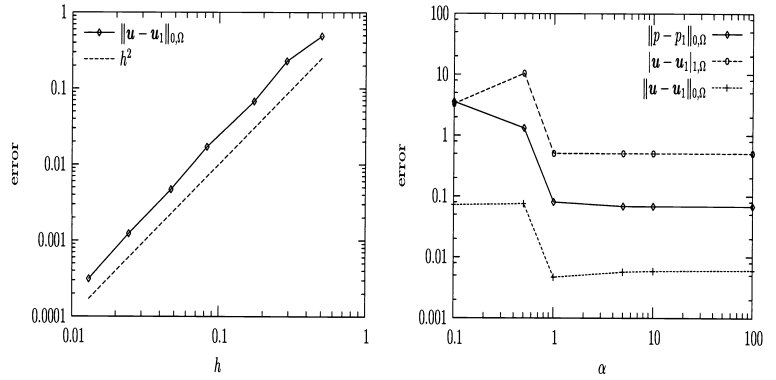


FIG. 4. *Method* (58)*: convergence history for* $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$*, and sensitivity of* (58) *with respect to* α.

Method (58) is tested next. The results are depicted in Figures 3–4 using $\alpha = 1$, where the results are in perfect accordance with the theoretical results. The justification of this choice for α may be found in Figure 4 (on the right) where we have depicted the behavior of the error in terms of α (using a mesh of around 2500 elements) and we see that for $\alpha \geq 1$ the error is almost independent of α, showing that the restriction of Lemma 5.2 is not only theoretical, but at the same time showing that, once we are inside the region predicted by the theory, the performance of the method is independent of α.

6.1.2. The $\mathbb{P}^1/\mathbb{P}^0$ case. For this case we first depict in Figures 5–6 the convergence history for method (14). The results reproduce our theoretical results showing an $O(h)$ order of convergence for $|\boldsymbol{u}-\boldsymbol{u}_1|_{1,\Omega}$, $|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u}-\boldsymbol{u}_1)+(p-p_0)\boldsymbol{n}]\!]|_h$ and $\|p-p_0\|_{0,\Omega}$, and an $O(h^2)$ convergence for $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$.

Method (58) is tested next. The results are depicted in Figures 7–8 using $\alpha = 1$, where the results are in perfect accordance with the theoretical results, giving an $O(h)$ for $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$, $|[\![p - p_0]\!]|_h$, and $\|p - p_0\|_{0,\Omega}$, and an $O(h^2)$ convergence for $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$. Concerning the choice of α, the situation now is quite different from that in the previous section. As a matter of fact, since we only control the pressure

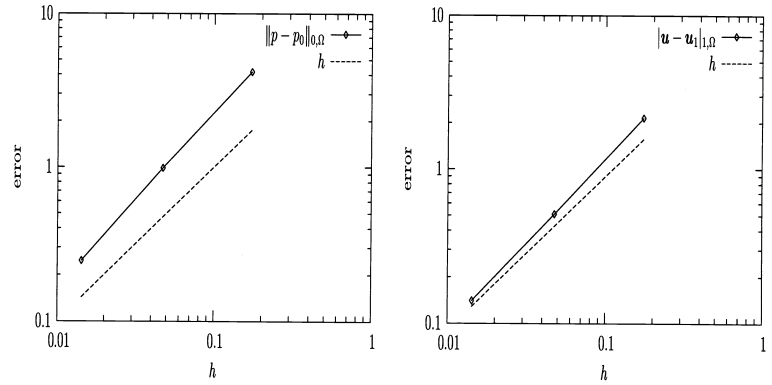


FIG. 5. *Method* (14)*: convergence history for* $\|p - p_0\|_{0,\Omega}$ *and* $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$.

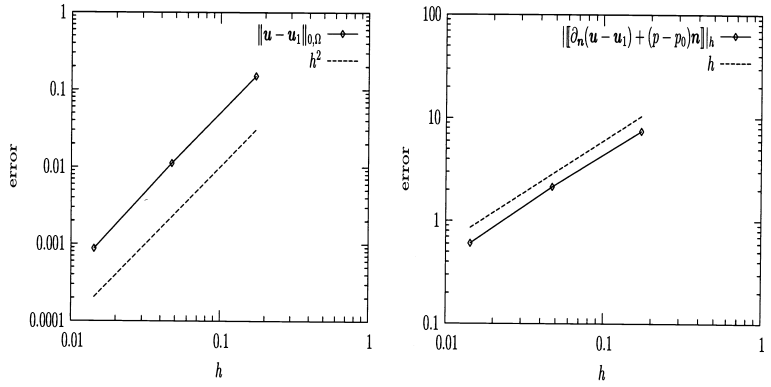


FIG. 6. *Method* (14)*: convergence history for* $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$ *and* $|[\![\partial_{\boldsymbol{n}}(\boldsymbol{u} - \boldsymbol{u}_1) + (p - p_0)\boldsymbol{n}]\!]|_h$.

via the jump terms governed by α, we can expect the error to grow as α grows, as it is shown in Figure 9 (for a mesh of 2500 elements) where we see that all the errors attain a minimum at $\alpha = 1$ (i.e., using $\tilde{\tau}_Z = \tau_Z$), and then they present a growing behavior. Values larger than 10 have been tested and the behavior is growing in all the errors. Related experiments have been performed using the GLS method (cf. [27]), obtaining similar results.

6.2. The lid-driven cavity problem. For this case we use the same domain as in the previous section, we set $\boldsymbol{f} = \boldsymbol{0}$, and the boundary conditions $\boldsymbol{u} = \boldsymbol{0}$ on $[\{0\} \times (0,1)] \cup [(0,1) \times \{0\}] \cup [\{1\} \times (0,1)]$ and $\boldsymbol{u} = (1,0)^t$ on $(0,1) \times \{1\}$. In Figure 10 we depict the pressure isovalues for both $\mathbb{P}^1/\mathbb{P}^0$ and $\mathbb{P}^1/\mathbb{P}^1$ approximations (using a mesh of around 1000 elements) showing, in both cases, the absence of oscillations.

7. Concluding remarks. In this paper we have analyzed and tested new stabilized finite element methods for the Stokes problem. These new methods arise from multiscale enrichment of the trial space for the velocity coupled with a Petrov–Galerkin strategy. This Petrov–Galerkin strategy makes it possible to perform statical condensation both at the element level and at the interelement boundary level, making the method take the form of a classical stabilized finite element method, containing jump terms on the interior edges of the triangulation, and with the corresponding

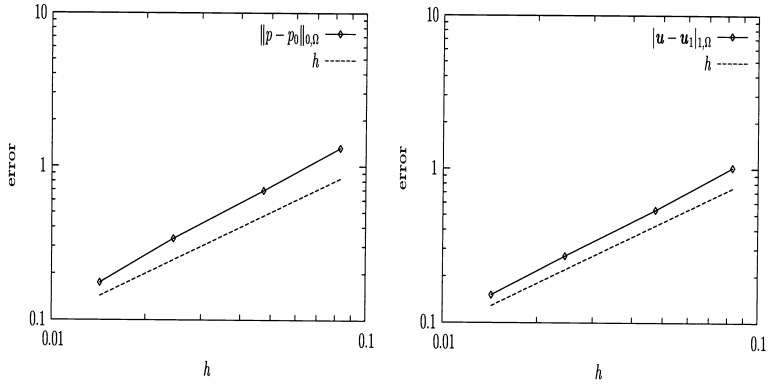


FIG. 7. *Method* (58)*: convergence history for* $\|p - p_0\|_{0,\Omega}$ *and* $|\boldsymbol{u} - \boldsymbol{u}_1|_{1,\Omega}$.

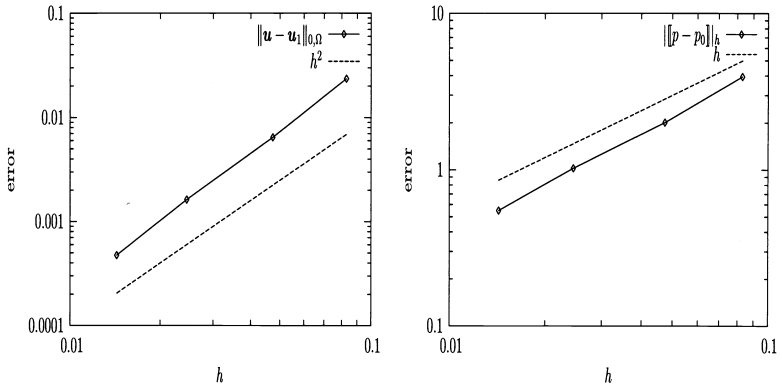


FIG. 8. *Method* (58)*: convergence history for* $\|\boldsymbol{u} - \boldsymbol{u}_1\|_{0,\Omega}$ *and* $\|p - p_0\|_{0,\Omega}$.

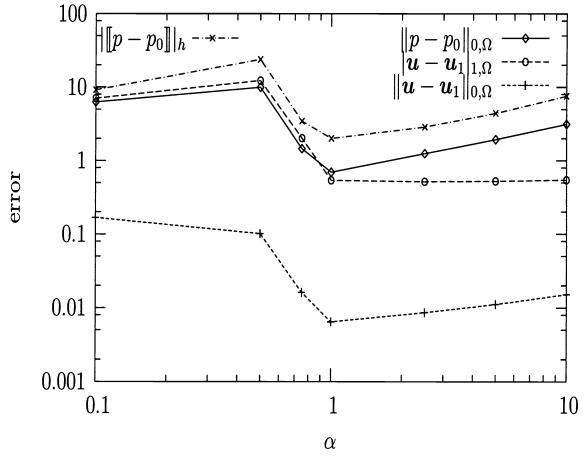


FIG. 9. *Sensitivity of method* (58) *to* α.

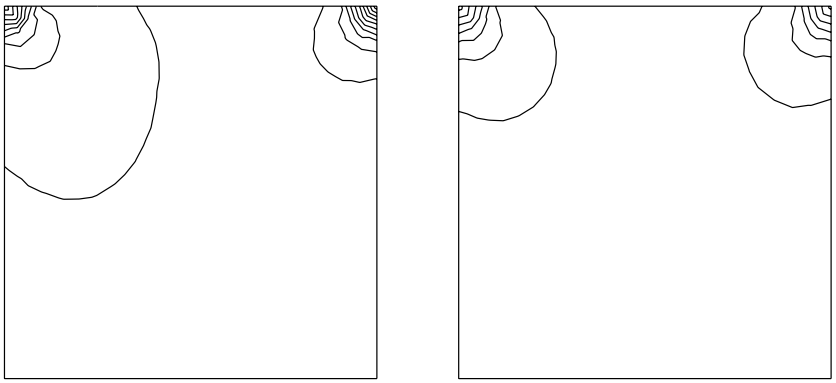


FIG. 10. *Pressure isovalues for* $\mathbb{P}^1/\mathbb{P}^0$ *(left) and* $\mathbb{P}^1/\mathbb{P}^1$ *(right) approximations.*

stabilization parameter known exactly. Optimal order error estimates were derived using the natural norms, results that were confirmed by the numerical experiments.

Our belief is that our general methodology may be applied to other mixed problems, namely the Darcy and Brinkman flow problems, and to the advection-diffusion equation. This will be the subject of future works.

Appendix A. Proof of Theorem 4.3.

The consistency of the method is immediate from the fact that $[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{u}]\!] = \mathbf{0}$ a.e. on ∂K. To prove the well-posedness of (39), we prove that $\mathbf{B}_1$ is an elliptic bilinear form. Let $(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1$; then from Lemma 4.6 we have that

$$\mathbf{B}_1((\boldsymbol{v}_1, q_1), (\boldsymbol{v}_1, q_1)) \;=\; |\!|\!|\boldsymbol{v}_1|\!|\!|_h^2 + \|q_1\|_h^2 \;-\; \sum_{K\in\mathcal{T}_h} \frac{1}{\nu}(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]), \nabla q_1)_K\,.$$

Now, in order to treat the last term above, let us denote by Z_1, Z_2, Z_3 the sides of K, and let, for $i = 1, 2, 3$, $b_K^{Z_i}$ be the solution of

$$-\Delta b_K^{Z_i} \;=\; 0 \quad \text{in } K, \qquad b_K^{Z_i} \;=\; g_i \quad \text{on each } Z \subset \partial K\,,$$

where $g_i = 0$ if $Z_i \subseteq \partial\Omega$, and g_i is the solution of

$$-\partial_{\boldsymbol{ss}}\, g_i = \frac{1}{h_{Z_i}} \quad \text{in } Z_i, \qquad g_i \;=\; 0 \quad \text{on } \partial K - Z_i\,,$$

otherwise. First, we remark that from the maximum principle, we have that $0 \le b_K^{Z_i} \le \frac{h_{Z_i}}{8}$ in K. On the other hand, it is easy to see that $\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]) \;=\; \sum_{i=1}^3 b_K^{Z_i}\,[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]|_{Z_i}$, and then, using that $|K| \le \frac{h_K^2}{2}$ and the inequality $ab \le \gamma^{-1}\frac{a^2}{4} + \gamma b^2$ $(\gamma > 0)$ (denoting

$\|\cdot\|_{\mathbb{R}^2}$ the Euclidean norm on $\mathbb{R}^2$) we arrive at

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h}\frac{1}{\nu}(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]),\nabla q_1)_K &= \sum_{K\in\mathcal{T}_h}\sum_{i=1}^{3}\frac{1}{\nu}\,(b_K^{Z_i}\,[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]|_{Z_i},\nabla q_1)_K\\
&= \sum_{Z\in\mathcal{E}_h}\sum_{K\subset\omega_Z}\frac{(b_K^Z,1)_K}{\nu}\,[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]|_Z\cdot\nabla q_1|_K\\
&\le \sum_{Z\in\mathcal{E}_h}\sum_{K\subset\omega_Z}\frac{h_Z|K|}{8\nu}\,\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]|_Z\|_{\mathbb{R}^2}\|\nabla q_1|_K\|_{\mathbb{R}^2}\\
&\le \gamma^{-1}\sum_{Z\in\mathcal{E}_h}\sum_{K\subset\omega_Z}\frac{|Z|\,\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\|_{0,Z}^2}{32\nu}+\gamma\sum_{Z\in\mathcal{E}_h}\sum_{K\subset\omega_Z}\frac{|K|\|\nabla q_1\|_{0,K}^2}{8\nu}\\
&\le \gamma^{-1}\sum_{Z\in\mathcal{E}_h}\frac{h_Z\,\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\|_{0,Z}^2}{16\nu}+\gamma\sum_{K\in\mathcal{T}_h}\frac{h_K^2\|\nabla q_1\|_{0,K}^2}{8\nu}\,. \qquad (63)
\end{aligned}$$

Hence, choosing $\gamma=\frac{14}{16}<1$ we arrive at

$$\begin{aligned}
\mathbf{B}_1((\boldsymbol{v}_1,q_1),(\boldsymbol{v}_1,q_1)) &\ge |\!|\!|\boldsymbol{v}_1|\!|\!|_h^2+\|q_1\|_h^2-\sum_{Z\in\mathcal{E}_h}\frac{h_Z}{14\nu}\|[\![\nu\partial_{\boldsymbol{n}}\boldsymbol{v}_1]\!]\|_{0,Z}^2-\gamma\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{8\nu}|q_1|_{1,K}^2\\
&\ge C_*(|\!|\!|\boldsymbol{v}_1|\!|\!|_h^2+\|q_1\|_h^2)\,,
\end{aligned}$$

where C_* is a positive constant not depending on h or ν. Now, for the error estimate, applying the coercivity result and the consistency of the method we arrive at

$$\begin{aligned}
C_*\,(|\!|\!|\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1|\!|\!|_h^2+\|p_1-\tilde{p}_1\|_h^2) &\le \mathbf{B}_1((\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1,p_1-\tilde{p}_1),(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1,p_1-\tilde{p}_1))\\
&= \mathbf{B}_1((\boldsymbol{u}_1-\boldsymbol{u},p_1-p),(\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1,p_1-\tilde{p}_1))\\
&= -\sum_{K\in\mathcal{T}_h}(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}_1-\boldsymbol{u})]\!]),\nabla(p_1-\tilde{p}_1))_K\,. \qquad (64)
\end{aligned}$$

Finally, proceeding as in (63) it is not difficult to see that

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}(\mathcal{B}_K([\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}_1-\boldsymbol{u})]\!]),\nabla(p_1-\tilde{p}_1))_K\\
&\quad\le C\sum_{Z\in\mathcal{E}_h}\tau_Z\,\|[\![\nu\partial_{\boldsymbol{n}}(\boldsymbol{u}_1-\boldsymbol{u})]\!]\|_{0,Z}^2+\frac{C_*}{2}\sum_{K\in\mathcal{T}_h}\tau_K\,|p_1-\tilde{p}_1|_{1,K}^2\\
&\quad\le C\,(|\!|\!|\boldsymbol{u}-\boldsymbol{u}_1|\!|\!|_h^2+\|p-p_1\|_h^2)+\frac{C_*}{2}\,(|\!|\!|\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1|\!|\!|_h^2+\|p_1-\tilde{p}_1\|_h^2)\,,
\end{aligned}$$

and hence, there exists $C>0$, independent of h and ν, such that

$$|\!|\!|\boldsymbol{u}_1-\tilde{\boldsymbol{u}}_1|\!|\!|_h^2+\|p_1-\tilde{p}_1\|_h^2\le C\,(|\!|\!|\boldsymbol{u}-\boldsymbol{u}_1|\!|\!|_h^2+\|p-p_1\|_h^2)\,,$$

and the result follows by triangular inequality.

REMARK A.1. *We have proved that the error of method* (39) *is bounded by the error of method* (44). *The same analysis of Theorems* 4.10 *and* 4.11 *may be carried out to prove error estimates on* $\|p-\tilde{p}_1\|_{0,\Omega}$ *and* $\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_1\|_{0,\Omega}$.

Appendix B. The error if $\boldsymbol{f}$ is not piecewise constant. As we claimed before, we have assumed that $\boldsymbol{f}$ is piecewise constant in order to derive (44), but this

assumption does not affect the convergence of the method, and hence (44) may be implemented as it is presented for a general function $\boldsymbol{f} \in L^2(\Omega)^2$. Now, if we do not suppose that $\boldsymbol{f}$ is piecewise constant in the derivation, then method (44) becomes the following: Find $(\boldsymbol{u}_h, p_h) \in \mathbf{V}_h \times Q_h^1$ such that

$$\mathbf{B}((\boldsymbol{u}_h, p_h), (\boldsymbol{v}_1, q_1)) = \mathbf{F}_h(\boldsymbol{v}_1, q_1) \tag{65}$$

for all $(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1$, where $\mathbf{B}$ is defined in (41) and $\mathbf{F}_h$ is given by

$$\mathbf{F}_h(\boldsymbol{v}_1, q_1) := (\boldsymbol{f}, \boldsymbol{v}_1)_\Omega + \sum_{K \in \mathcal{T}_h} \frac{1}{\nu} (\mathcal{M}_K(\boldsymbol{f}), \nabla q_1)_K . \tag{66}$$

Clearly, (65) has a unique solution $(\boldsymbol{u}_h, p_h) \in \mathbf{V}_h \times Q_h^1$. Moreover, the following result holds.

THEOREM B.1. *Let us suppose that $\boldsymbol{f} \in H^1(\Omega)^2$. Then, under the hypothesis of Theorems* 4.8, 4.10, *and* 4.11, *the following error estimate holds:*

$$|||\boldsymbol{u} - \boldsymbol{u}_h|||_h + \|p - p_h\|_h \leq Ch \left(\sqrt{\nu}\, |\boldsymbol{u}|_{2,\Omega} + \frac{1}{\sqrt{\nu}} |p|_{1,\Omega} + \frac{1}{\sqrt{\nu}} \|\boldsymbol{f}\|_{1,\Omega} \right) , \tag{67}$$

$$\|p - p_h\|_{0,\Omega} \leq Ch \left(\nu\, |\boldsymbol{u}|_{2,\Omega} + |p|_{1,\Omega} + \|\boldsymbol{f}\|_{1,\Omega} \right) , \tag{68}$$

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega} \leq Ch^2 \left(|\boldsymbol{u}|_{2,\Omega} + \frac{1}{\nu} |p|_{1,\Omega} + \frac{1}{\nu} \|\boldsymbol{f}\|_{1,\Omega} \right) . \tag{69}$$

Proof. Let $(\boldsymbol{u}_1, p_1)$ be the solution of (44). First, applying [30, Lem. 5.3.1], we see that

$$\begin{aligned} |||\boldsymbol{u}_1 - \boldsymbol{u}_h|||_h + \|p_1 - p_h\|_h &\leq \sup_{(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1 - \{\theta\}} \frac{\mathbf{F}(\boldsymbol{v}_1, q_1) - \mathbf{F}_h(\boldsymbol{v}_1, q_1)}{\|\boldsymbol{v}_1\|_h + \|q_1\|_h} \\ &= \sup_{(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1 - \{\theta\}} \frac{\sum_{K \in \mathcal{T}_h} (\tau_K \boldsymbol{f} - \frac{1}{\nu} \mathcal{M}_K(\boldsymbol{f}), \nabla q_1)_K}{\|\boldsymbol{v}_1\|_h + \|q_1\|_h} \\ &\leq \sup_{(\boldsymbol{v}_1, q_1) \in \mathbf{V}_h \times Q_h^1 - \{\theta\}} \frac{\sum_{K \in \mathcal{T}_h} \|\tau_K \boldsymbol{f} - \frac{1}{\nu} \mathcal{M}_K(\boldsymbol{f})\|_{0,K} |q_1|_{1,K}}{\|\boldsymbol{v}_1\|_h + \|q_1\|_h} . \end{aligned} \tag{70}$$

Now, let $\boldsymbol{f}_h$ be the piecewise constant function given by

$$\boldsymbol{f}_h\Big|_K = \frac{1}{|K|} \int_K \boldsymbol{f} .$$

This function, which is the (local) projection on the space of piecewise constant functions, satisfies (cf. [17]) $\|\boldsymbol{f} - \boldsymbol{f}_h\|_{0,K} \leq C\, h_K\, |\boldsymbol{f}|_{1,K}$. Then, applying triangular inequality we arrive at

$$\begin{aligned} \|\tau_K \boldsymbol{f} - \mathcal{M}_K(\boldsymbol{f})\|_{0,K} \leq\ & \|\tau_K (\boldsymbol{f} - \boldsymbol{f}_h)\|_{0,K} + \|\tau_K \boldsymbol{f}_h - \frac{1}{\nu} \mathcal{M}_K(\boldsymbol{f}_h)\|_{0,K} \\ & + \frac{1}{\nu} \|\mathcal{M}_K(\boldsymbol{f}_h - \boldsymbol{f})\|_{0,K} . \end{aligned} \tag{71}$$

The first term is easily bounded using the approximation properties of $\boldsymbol{f}_h$ and the definition of τ_K. Next, since $\mathcal{M}_K(\boldsymbol{f}_h) = b_K^p \boldsymbol{f}_h$ in each $K \in \mathcal{T}_h$, the second term is

bounded in the following way:

$$\begin{aligned}\|\tau_K \boldsymbol{f}_h - \frac{1}{\nu}\mathcal{M}_K(\boldsymbol{f}_h)\|_{0,K} \le\ & |\tau_K| \|\boldsymbol{f}_h\|_{0,K} + \frac{\|b_K^p\|_{0,K}}{\nu}\,\|\boldsymbol{f}_h\|_{\mathbb{R}^2} \\ \le\ & |\tau_K| \|\boldsymbol{f}_h\|_{0,K} + \frac{C_K\,|b_K^p|_{1,K}}{\nu\,|K|^{\frac{1}{2}}}\,\|\boldsymbol{f}_h\|_{0,K}\,,\end{aligned}$$

where $C_K > 0$ is the constant such that $\|v\|_{0,K} \le C_K\,|v|_{1,K}$ for all $v \in H_0^1(K)$. Furthermore, looking carefully at the behavior of the Poincaré constant C_K we can see (cf. [30, Thm. 1.2.5]) that $C_K \le h_K$. On the other hand, from the definition of b_K^p we have $|b_K^p|_{1,K}^2 = (b_K^p, 1)_K$, and then, applying (48) we arrive at

$$\frac{C_K\,|b_K^p|_{1,K}}{\nu\,|K|^{\frac{1}{2}}}\,\|\boldsymbol{f}_h\|_{0,K} \le \frac{h_K\,\sqrt{(b_K^p,1)_K}}{\nu\,|K|^{\frac{1}{2}}}\,\|\boldsymbol{f}_h\|_{0,K} \le C\,\frac{h_K^2}{\nu}\,\|\boldsymbol{f}_h\|_{0,K}\,. \tag{72}$$

To bound the third term in (71) we remark that function $\boldsymbol{e} := \mathcal{M}_K(\boldsymbol{f} - \boldsymbol{f}_h)$ satisfies $-\Delta \boldsymbol{e} = \boldsymbol{f} - \boldsymbol{f}_h$ in K, $\boldsymbol{e} = \boldsymbol{0}$ on ∂K, and hence

$$\|\boldsymbol{e}\|_{0,K} \le C_K^2\,\|\boldsymbol{f} - \boldsymbol{f}_h\|_{0,K} \le h_K^2\,\|\boldsymbol{f} - \boldsymbol{f}_h\|_{0,K} \le C\,h_K^3\,|\boldsymbol{f}|_{1,K}\,. \tag{73}$$

Hence, applying (70)–(73) (and assuming $h \le 1$), we arrive at

$$\begin{aligned}|\!|\!|\boldsymbol{u}_1 - \boldsymbol{u}_h|\!|\!|_h + \|p_1 - p_h\|_h \le\ & C \sup_{(\boldsymbol{v}_1,q_1)\in \mathbf{V}_h\times Q_h^1 - \{\theta\}} \frac{\sum_{K\in\mathcal{T}_h} \frac{h_K^2}{\nu}\|\boldsymbol{f}\|_{1,K}|q_1|_{1,K}}{|\!|\!|\boldsymbol{v}_1|\!|\!|_h + \|q_1\|_h} \\ \le\ & C\,\frac{h}{\sqrt{\nu}}\,\|\boldsymbol{f}\|_{1,\Omega}\,,\end{aligned}$$

and hence (67) follows by triangular inequality and Theorem 4.8. Estimates (68) and (69) are proved as in Theorems 4.10 and 4.11 and by using (67). ∎

REFERENCES

[1] R. Araya and F. Valentin, *A multiscale a-posteriori error estimator*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 2077–2094.

[2] C. Baiocchi, F. Brezzi, and L. P. Franca, *Virtual bubbles and Galerkin-least-squares type methods (Ga. L. S.)*, Comput. Methods Appl. Mech. Engrg., 105 (1993), pp. 125–141.

[3] G. Barrenechea, M. Fernández, and C. Vidal, *A Stabilized Finite Element Method for the Oseen Equation with Dominating Reaction*, Preprint 2004-08, Departamento de Ingeniería Matemática, Universidad de Concepción, Concepción, Chile, 2004.

[4] G. Barrenechea and F. Valentin, *An unusual stabilized finite element method for a generalized Stokes problem*, Numer. Math., 92 (2002), pp. 653–677.

[5] T. Barth, P. Bochev, M. Gunzburger, and J. Shahid, *A taxonomy of consistently stabilized finite element methods for the Stokes problem*, SIAM J. Sci. Comput., 25 (2004), pp. 1585–1607.

[6] F. Brezzi, *Recent results in the treatment of subgrid scales*, in CANUM 2000: Actes du 32e Congrès National d'Analyse Numérique, ESAIM Proc. II, Soc. Math. Indust., Paris, 2000, pp. 61–84.

[7] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–339.

[8] F. Brezzi, L. Franca, and A. Russo, *Further considerations on residual-free bubbles for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 25–33.

[9] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.

[10] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.
[11] F. Brezzi and J. Pitkaranta, *On the stabilization of finite element approximations of the Stokes problem*, in Efficient Solutions of Elliptic Systems, W. Hackbush, ed., Notes Numer. Fluid Mech. 10, Vieweg, Braunschweig, 1984, pp. 11–19.
[12] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.
[13] E. Burman, M. Fernández, and P. Hansbo, *Edge Stabilization for the Incompressible Navier–Stokes Equations: A Continuous Interior Penalty Finite Element Method*, Tech. Report RR-5349, INRIA, Le Chesnay, France, 2004.
[14] E. Burman and P. Hansbo, *A Unified Stabilized Method for Stokes' and Darcy's Equations*, Tech. Report 2002-15, Chalmers Finite Element Center, Göteborg, Sweden, 2002.
[15] A. Cangiani and E. Süli, *Enhanced RFB Method*, Tech. Report NA-03/17, Oxford University Computing Laboratory, Oxford, UK, 2003.
[16] J. Douglas and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, R. Glowinski and J.-L. Lions, eds., Springer, Berlin, 1976, pp. 207–216.
[17] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, New York, 2004.
[18] C. Farhat, I. Harari, and L. Franca, *The discontinuous enrichment method*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6455–6479.
[19] L. Franca, T. J. R. Hughes, and R. Stenberg, *Stabilized finite element methods*, in Incompressible Computational Fluid Dynamics, M. Gunzburger and R. Nicolaides, eds., Cambridge University Press, Cambridge, UK, 1993, pp. 87–107.
[20] L. Franca, A. Madureira, L. Tobiska, and F. Valentin, *Convergence analysis of a multiscale finite element method for singularly perturbed problems*, Multiscale Model. Simul., 4 (2005), pp. 839–866.
[21] L. Franca, A. Madureira, and F. Valentin, *Towards multiscale functions: Enriching finite element spaces with local but not bubble-like functions*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 3006–3021.
[22] L. P. Franca and A. Russo, *Approximation of the Stokes problem by residual-free macro bubbles*, East-West J. Numer. Math., 4 (1996), pp. 265–278.
[23] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.
[24] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[25] T. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[26] T. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly varying coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[27] T. J. R. Hughes and L. P. Franca, *A new finite element formulation for computational fluid dynamics:* VII. *The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65 (1987), pp. 85–96.
[28] T. J. R. Hughes, L. P. Franca, and M. Balestra, *A new finite element formulation for computational fluids dynamics:* V. *Circumventing the Babuska–Brezzi condition: A stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 62 (1987), pp. 85–99.
[29] N. Kechar and D. Silvester, *Analysis of a locally stabilized mixed finite element method for the Stokes problem*, Math. Comp., 58 (1992), pp. 1–10.
[30] P. A. Raviart and J.-M. Thomas, *Introduction à l'Analyse Numérique des Équations aux Dérivées Partielles*, Masson, Paris, 1983.
[31] A. Russo, *Bubble stabilization of finite element methods for the linearized incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 132 (1996), pp. 335–343.
[32] G. Sangalli, *Global and local error analysis for the residual-free bubbles method applied to advection-dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.
[33] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.
[34] L. Tobiska and R. Verfürth, *Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations*, SIAM J. Numer. Anal., 33 (1996), pp. 107–127.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 349–366

ERROR ESTIMATES FOR THE DISCONTINUOUS GALERKIN METHODS FOR PARABOLIC EQUATIONS*

K. CHRYSAFINOS‡ AND NOEL J. WALKINGTON†

Abstract. The classical discontinuous Galerkin method for a general parabolic equation is analyzed. Symmetric error estimates for schemes of arbitrary order are presented. The ideas developed below relax many assumptions required in previous work. For example, different discrete spaces may be used at each time step, and the spatial operator need not be self-adjoint or independent of time. Our error estimates are posed in terms of projections of the exact solution onto the discrete spaces and are valid under the minimal regularity guaranteed by the natural energy estimate. These projections are local and enjoy optimal approximation properties when the solution is sufficiently regular.

Key words. discontinuous Galerkin, parabolic equations, error estimates

AMS subject classification. 65M60

DOI. 10.1137/030602289

1. Introduction.

We consider the parabolic PDE of the form

$$u_t + A(t)u = F(t), \qquad u(0) = u_0. \tag{1.1}$$

The operators act on Hilbert spaces related through the standard pivot construction, $U \hookrightarrow H \simeq H' \hookrightarrow U'$, where each embedding is continuous and dense. Then, $A(.) : U \to U'$ is a linear map and $F(.) \in U'$. Our goal is to analyze the classical discontinuous Galerkin (DG) scheme and derive fully discrete error estimates under minimal regularity assumptions. The class of DG schemes considered are classical in the sense that the discrete solutions may be discontinuous in time but are conforming in space; i.e., solutions are in (a subspace of) U at each time.

Our techniques also apply to the more general implicit evolution equation [21, 22]

$$(M(t)u)_t + A(t)u = F(t), \qquad u(0) = u_0, \tag{1.2}$$

where $M(.) : H \to H$ is a self-adjoint positive definite operator. The extension of our analysis to this equation will be taken up separately. The analysis below addresses the following issues, which have not yet been adequately considered in the literature:

- The operator $A(.)$ may depend upon time and is not required to be self-adjoint. To date the sharpest estimates for DG approximations exploit classical spectral theory for self-adjoint positive definite operators, and so require A to be such an operator and to be independent of time. When $A(.)$ is not self-adjoint, multiplying (1.1) by u_t does not give an estimate for the time derivative.

*Received by the editors December 18, 2003; accepted for publication (in revised form) August 2, 2005; published electronically March 7, 2006. This work was supported in part by National Science Foundation grants DMS–0208586 and ITR 0086093. This work was also supported by the NSF through the Center for Nonlinear Analysis.

http://www.siam.org/journals/sinum/44-1/60228.html

‡Institut für Numerische Simulation, University of Bonn, Wegelerstr. 6, Bonn, Germany. Current address: (chrysafinos@ins.uni-bonn.de).

†Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213 (noelw@cmu.edu).

- The subspaces of U used for the DG approximations may be different on each time interval $(t^{n-1}, t^n]$. This adds a significant complication to the analysis, which is present even when $A = 0$. Indeed, the first step in our analysis is to consider the DG scheme for an auxiliary parabolic equation which reduces to an ODE in the Hilbert space H when the coercivity constant vanishes. Different subspaces are essential ingredients of adaptive strategies, used in conjunction with a posteriori error estimates to give guaranteed error bounds. Retriangulation is also necessary for many algorithms based upon a Lagrangian coordinate system; an example is presented below.
- DG approximations of equations of the form (1.2) have not been considered in the past. The example below shows that equations of this form arise when Lagrangian schemes are constructed for the convection diffusion equation [5, 6, 7].
- The operator $A(.)$ is not required to be strictly coercive; only semicoercivity of the form $\langle A(.)u, u\rangle \geq c|u|_U^2 - C\|u\|_H^2$ is required. Here $|.|_U$ is a seminorm such that $\|.\|_U^2 = |.|_U^2 + \|.\|_H^2$. This causes significant problems in the analysis of DG schemes since the classical Gronwall argument, used for the continuous problem, fails in the discrete setting. This failure is due to the elementary observation that functions of the form $\chi_{[0,\hat{t})}u$ are not polynomial in time unless $\hat{t}$ is a partition point, and thus they are not available as test functions in the discrete setting.[1] In the past this problem has been circumvented by bounding temporal derivatives of the solution [4, 23] so that the solution between the partition points can be controlled by the values at these points. This line of argument fails for solutions having minimal regularity. These issues are circumvented here by constructing polynomial approximations to the characteristic functions $\chi_{[0,\hat{t})}$.

As stated above, our analysis does not require any regularity beyond the natural bounds that follow from the usual energy estimate. This is essential for control problems, where solutions of the dual problem typically will not enjoy any additional regularity. Our estimates show that the error can be bounded by local projection errors of the solution projected onto the discrete subspaces. This error can also be viewed as a "local truncation error" of the ODE obtained by setting $A = 0$. Care is taken to keep track of how the various constants depend upon the coercivity constant of $A(.)$. This is important for the analysis of problems like the convection diffusion equation, where the coercivity constant is small.

The following equation can be analyzed within the general framework developed here but falls outside of the theory developed, for example, in Thomée's text [23].

Convection diffusion equation. The classical convection diffusion equation is

$$\bar{u}_t + \mathbf{V}.\nabla\bar{u} - \epsilon\Delta\bar{u} = 0,$$

and the problems that arise when ϵ is small are notorious. To address these problems this equation is sometimes considered in a Lagrangian variable. Specifically, let $\tilde{\mathbf{V}}$ be a (numerical) approximation of $\mathbf{V}$, and let $x = \chi(t, X)$ be the change of variables defined by the flow map associated with $\tilde{\mathbf{V}}$; i.e.,

$$\dot{x}(t, X) = \tilde{\mathbf{V}}(t, x(t, X)), \qquad x(0, X) = X.$$

[1] Here $\chi_{[0,\hat{t})}$ is the characteristic function equal to 1 on $[0, \hat{t})$ and zero otherwise.

If $u(t,X) = \bar{u}(t,x(t,X))$, then

$$u_t + (\mathbf{V} - \tilde{\mathbf{V}}) \cdot (F^{-T}\nabla_X u) - \epsilon \left(\frac{1}{J}\right) \operatorname{div}_X \left(JF^{-1}F^{-T}\nabla_X u\right) = 0,$$

where $F_{ij} = \partial x_i/\partial X_j$ is the Jacobian of the mapping and $J = \det(F)$. The natural weak problem for this equation is

$$\int_\Omega \Big(\big(u_t + (\mathbf{V} - \tilde{\mathbf{V}}) \cdot (F^{-T}\nabla_X u)\big)v + \epsilon (F^{-T}\nabla_X u).(F^{-T}\nabla_X v) \Big) J = 0.$$

Using the properties of determinants, we find

$$u_t J = (Ju)_t - \dot{J}u = (Ju)_t - J \operatorname{div}(\tilde{\mathbf{V}})u.$$

If $\operatorname{div}(\tilde{\mathbf{V}}) = 0$, then J is constant and the transformed problem takes the form of (1.1); otherwise, it takes the form of (1.2) with $M(.)u = Ju$, and

$$A(.)u = -\operatorname{div}(\tilde{\mathbf{V}})uJ + (\mathbf{V} - \tilde{\mathbf{V}}).(F^{-T}\nabla_X u)J - \epsilon \operatorname{div}_X \left(JF^{-1}F^{-T}\nabla_X u\right).$$

This statement of the problem generalizes the idea behind the "characteristic Galerkin" scheme introduced by Douglas and Russell in [5] and Dupont in [6].

This change of variables reduces the effective Peclet number from $|\mathbf{V}|/\epsilon$ to $|\mathbf{V} - \tilde{\mathbf{V}}|/\epsilon$, which will be $\mathcal{O}(1)$ if $\tilde{\mathbf{V}}$ is a sufficiently accurate approximation of $\mathbf{V}$. This eliminates many of the numerical difficulties encountered by algorithms based upon the classical statement; however, other problems arise. While the Jacobian of the transformation satisfies $F(0,X) = I$, its condition number grows exponentially if $\tilde{\mathbf{V}}$ is anything other than a rigid motion. In the context of a numerical scheme this problem is circumvented by reinitializing the transformation at each (or every few) time step(s). This reinitialization corresponds to changing the subspace for the numerical solution every (few) time step(s). In essence, a triangular mesh in the X coordinate system will be a distorted mesh in the x coordinate system, and reinitializing the transform corresponds to projecting the solution onto a (straight sided) triangular mesh in the x coordinates. This gives rise to different subspaces at each time step.

1.1. Related results. The discontinuous Galerkin method was first introduced to model and simulate neutron transport by Lasaint and Raviart in [13]. There is an abundant literature concerning applications of the DG scheme for hyperbolic problems; see, e.g., [3, 12, 24] and references within. The DG method for ODEs was considered by Delfour, Hager, and Trochu in [4]. They showed that the DG scheme was superconvergent at the partition points (order $2k+2$ for polynomials of degree k). The superconvergence results (and better rates in the H norm) use a duality argument, and space considerations do not permit us to develop the corresponding estimates in the current setting.

In the context of parabolic equations, DG schemes were first analyzed for linear parabolic problems by Jamet in [11], where $\mathcal{O}(\tau^k)$ results were proved, and by Eriksson, Johnson, and Thomée in [10], where $\mathcal{O}(\tau^{2k-1})$ results were proved for "smooth" initial data among others (τ being the time step size). An excellent exposition of their results and, more generally, the DG method for parabolic equations can be found in Thomée's book [23]. In [23] nodal and interior estimates are presented in various norms. One may also consult [16] for the analysis of a related formulation based on the backward Euler scheme. The relation between the DG scheme and adaptive

techniques was studied in [8] and [9]. Finally, some results concerning the analysis of parabolic integro-differential equations by the DG method are presented in [14] (see also references therein).

In [7] Dupont and Liu introduced the concept of "symmetric error estimates" for parabolic problems. They define such an error estimate to be one of the form

$$|||u - u_h||| \leq C \inf_{w_h \in \mathcal{U}_h} |||u - w_h|||,$$

where u and u_h are the exact and approximate solutions, respectively; $|||.|||$ is an appropriate norm; and $\mathcal{U}_h$ is the discrete subspace in which approximation solutions are sought. While estimates of this form are standard for elliptic problems, this is not the case for evolution problems. For example, error estimates for evolution problems approximated by the implicit Euler scheme frequently involve terms of the form $\|u_{tt}\|_{L^2(\Omega)}$. Symmetric error estimates are useful for problems where the solution u may not be very regular, such as control problems, and are used to develop a posteriori error estimates for adaptive schemes. Symmetric error estimates for moving mesh finite element methods were studied in [7, 15] (see also the references therein). Mesh modification techniques for finite elements were introduced in [17] and [18]. For some earlier work on convection-dominated problems based on the methods of characteristics and mesh modification one may consult [5] and [6], respectively.

An alternative to the symmetric error estimates are estimates of the form

$$|||u - u_h||| \leq C|||u - \mathbb{P}_h u|||, \tag{1.3}$$

where $\mathbb{P}_h : \mathcal{U} \to \mathcal{U}_h$ is a projection which exhibits optimal interpolation properties if u is sufficiently smooth. Estimates of this form enjoy the same advantages found for those proposed by Dupont and Liu. Below, estimates of the form (1.3) are developed for parabolic equations of the form (1.1), where the projection $\mathbb{P}_h u$ is nonlocal. However,

$$|||u - \mathbb{P}_h u||| \leq |||u - \mathbb{P}_h^{loc} u||| + |||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||,$$

where $\mathbb{P}_h^{loc}$ is a local projection, so the first term can be estimated using classical interpolation theory. The second term $|||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||$ vanishes if the same subspace of U is used in each partition (t^{n-1}, t^n); otherwise, it depends solely upon the jump in the interpolant of the exact solution at the partition points $\{t^n\}_{n=0}^N$. The size of the constant C in (1.3), and its dependence on various constants, plays an important role; we are careful throughout to state the dependence of the constant upon the various coercivity constants and bounds assumed for the operator A.

Error estimates for Lagrange–Galerkin approximations of convection-dominated problems for divergence-free velocity fields vanishing on the boundary are presented in [2]. Issues related to the stability of Lagrange–Galerkin approximations are also discussed in [19]. Recently there has been a lot of work on the development and analysis of DG methods for elliptic problems. A comprehensive survey and comparison of this work can be found in [1], which contains many references related to this approach.

1.2. Outline. In section 2, the DG scheme is formulated and analyzed for an auxiliary parabolic equation. This section focuses upon the difficulties that arise when different subspaces of U may be used at every time step and when the coercivity constant may be small. Error estimates are first derived at times corresponding to the partition points. Additional arguments using "discrete characteristic functions"

are developed to estimate the error at times in between the partition points. The latter arguments appear to be new.

A priori estimates for the DG approximations of (1.1) are developed in section 3. Estimates are derived in the natural norms associated with the parabolic problem; by "natural" we mean norms that arise in the natural energy estimates obtained by multiplying (1.1) by u. The results of section 2 are used in an essential fashion. Indeed, the difficulties associated with different subspaces of U at each time step are circumvented by comparing the discrete solution of the parabolic equation with an appropriate solution of the auxiliary equation. By using the "discrete characteristic functions" developed in section 2, the self-adjoint assumptions typically imposed upon $A(.)$ may be avoided.

Remark 1. When the same discrete subspace of U is used for each time step, our techniques generalize, and to some extent simplify, the classical analysis. The reader interested in this case needs to read only sections 2.3 on the construction of discrete characteristic functions, and Definition 2.1 for $\mathbb{P}_h^{loc}$, before proceeding to section 3. Remark 4 in section 3 amplifies upon this.

1.3. Notation. It is assumed throughout that the evolution of the solution to (1.1) takes place in a Hilbert space H and that the operators $A(.)$ are defined on another Hilbert space U with $U \hookrightarrow H \simeq H' \hookrightarrow U'$, where each of the embeddings are dense and continuous. The inner product on H is denoted by $(.,.)$, and the induced duality pairing between U and U' is denoted by $\langle .,. \rangle$. The norm on H is often denoted by $|.| \equiv \|.\|_H$, and it is assumed that the norm on U can be written as $\|.\|_U^2 = |.|_U^2 + \|.\|_H^2$, where $|.|_U$ is a seminorm on U (the "principle" part) and is often denoted by $\|.\|$; $\|.\|_U^2 = \|.\|^2 + |.|^2$. Standard notation of the form $L^2[0,T;U]$, $H^1[0,T;U']$, etc. is used to indicate the temporal regularity of functions with values in U, U', etc.

Approximations of (1.1) will be constructed on a partition $0 = t^0 < t^1 < \cdots < t^N = T$ of $[0,T]$; the step sizes are denoted as $\tau^n = t^n - t^{n-1}$, and $\tau = \max_n \tau^n$. On each interval of the form $(t^{n-1}, t^n]$ a subspace U_h^n of U is specified, and the approximate solutions will lie in the space

$$\mathcal{U}_h = \{u_h \in L^2[0,T;U] \mid u_h|_{(t^{n-1},t^n]} \in \mathcal{P}_k(t^{n-1}, t^n; U_h^n)\}.$$

Here $\mathcal{P}_k(t^{n-1}, t^n; U_h^n)$ is the space of polynomials of degree k or less having values in U_h^n. Notice that, by convention, functions in $\mathcal{U}_h$ are left continuous with right limits. We will write u^n for $u_h(t^n) = u_h(t_-^n)$ and let u_+^n denote $u_h(t_+^n)$. This notation is also used with functions like the error $e = u - u_h$. It is assumed that the exact solution, u, is in $C[0,T;H]$, so that the jump in the error at t^n, denoted by $[e^n]$, is $[e^n] = [u^n] = u_+^n - u^n$.

2. DG scheme for an auxiliary equation.

2.1. Background. This section addresses issues that arise when different discrete subspaces are used for each time step of the DG scheme. Schemes for the simplest (heat-type) parabolic equation are considered, which allows issues associated with different spaces at each step and the role of the coercivity constant to be isolated. In the next section the error estimates for the more general equation (1.1) will be obtained by comparing the solutions of the two equations.

Let $(.,.)_U$ be the inner product on U, and let $B : U \to U'$ be the associated Riesz map: $(u,v)_U = \langle Bu, v \rangle$. We consider the problem of recovering a function $u \in L^2[0,T;U] \cap H^1[0,T;U']$, given the initial value $u(0)$ and $f = u_t + \eta Bu$, where

$\eta \geq 0$. Specifically, consider DG finite element approximations of the initial value problem

$$u_t + \eta B u = f, \qquad u(0) = u_0. \tag{2.1}$$

In this situation there exists a unique $u \in L^2[0,T;U] \cap H^1[0,T;U'] \hookrightarrow C[0,T;H]$, which is the solution of the weak problem

$$\begin{aligned}(u(T), v(T)) - \int_0^T \langle u, v_t\rangle + \eta(u,v)_U &= (u_0, v(0)) + \int_0^T \langle f, v\rangle \\ &\forall\, v \in L^2[0,T;U] \cap H^1[0,T;U'].\end{aligned} \tag{2.2}$$

To approximate the solution of (2.2) we introduce a partition $0 = t^0 < t^1 < \cdots < t^N = T$ of $[0,T]$ and a collection $\{U_h^n\}_{n=0}^N$ of subspaces of U. The DG method constructs an approximate solution

$$u_h \in \mathcal{U}_h \equiv \{u \in L^2[0,T;U] \mid u|_{(t^{n-1},t^n]} \in \mathcal{P}_k(t^{n-1},t^n;U_h^n)\}$$

such that

$$(u^n, v^n) - \int_{t^{n-1}}^{t^n} (u_h, v_{ht}) + \eta(u_h, v_h)_U - (u^{n-1}, v_+^{n-1}) = \int_{t^{n-1}}^{t^n} \langle f, v_h\rangle \tag{2.3}$$

for all $v_h \in \mathcal{U}_h$ and each $n = 1, 2, \ldots, N$. Recall that $u^n \equiv u_h(t^n) = u_h(t_-^n)$, and use standard notation, u_+^n, u_-^n, for the traces from above and below, respectively. Integration by parts gives the following alternative form of (2.3):

$$\int_{t^{n-1}}^{t^n} (u_{ht}, v_h) + \eta(u,v)_U + (u_+^{n-1} - u^{n-1}, v_+^{n-1}) = \int_{t^{n-1}}^{t^n} \langle f, v_h\rangle. \tag{2.4}$$

2.2. Error estimate at partition points. In this elementary context it is possible to estimate the error at each partition point t^n using the ideas from [9]. We will need the following projection operators $\mathbb{P}_n^{loc}$ introduced in [10].

DEFINITION 2.1. (1) *The projection* $\mathbb{P}_n^{loc} : C[t^{n-1},t^n;H] \to \mathcal{P}_k(t^{n-1},t^n;U_h^n)$ *satisfies* $(\mathbb{P}_n^{loc}u)^n = P_n u(t^n)$, *and*

$$\int_{t^{n-1}}^{t^n} (u - \mathbb{P}_n^{loc}u, v_h) = 0 \qquad \forall\, v_h \in \mathcal{P}_{k-1}(t^{n-1},t^n;U_h^n).$$

Here we have used the convention $(\mathbb{P}_n^{loc}u)^n \equiv (\mathbb{P}_n^{loc}u)(t^n)$ *and* $P_n : H \to U_h^n$ *is the orthogonal projection operator onto* $U_h^n \subset H$.

(2) *The projection* $\mathbb{P}_h^{loc} : C[0,T;H] \to \mathcal{U}_h$ *satisfies*

$$\mathbb{P}_h^{loc}u \in \mathcal{U}_h \quad \text{and} \quad (\mathbb{P}_h^{loc}u)|_{(t^{n-1},t^n]} = \mathbb{P}_n^{loc}(u|_{[t^{n-1},t^n]}).$$

This projection satisfies the standard approximation properties [23, Theorem 12.1] and can be viewed as the one step DG approximation of $u_t = f$ on the interval $(t^{n-1},t^n]$, with exact initial data $u(t^{n-1})$ and $f = u_t$ specified. For the parabolic problem we will use the analogous *global* projection ($\mathbb{P}_h$ below), which is the DG solution of the auxiliary equation with initial data $u(0)$ and $f = u_t + Bu$ specified.

The following theorem provides a decomposition of the error into the errors due to the changing of the spaces and the projection errors. The former depend upon the size of the coercivity constant η.

THEOREM 2.2. *Let $u_h \in \mathcal{U}_h$ be the approximate solution of* (2.1) *computed using the discontinuous Galerkin scheme* (2.3)*, and write $\hat{e} = \mathbb{P}_h^{loc} u - u_h$. Then there exists a constant $C_k > 0$ depending only upon k such that*

$$|\hat{e}^n|^2 + \frac{\eta}{2}\int_0^{t^n} \|\hat{e}\|_U^2 + \frac{1}{2}\sum_{i=0}^{n-1} |\hat{e}^i - \hat{e}_+^i|^2 \le |\hat{e}^0|^2 + \int_0^{t^n} \eta \|(I - \mathbb{P}_h^{loc})u\|_U^2 + \sum_{i=0}^{n-1} 2\min\left(|(I-P_i)u(t^i)|^2, \frac{C_k^2}{\tau^{i+1}\eta}\|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2 \right). \tag{2.5}$$

Remark 2. Since $P_i(I - P_{i-1}) = 0$ when $U_h^i \subset U_h^{i-1}$, the error estimate reduces to the usual projection errors when the same discrete subspace is used at each time.

Proof. Let $e = u - u_h$ be the total error, and note that the Galerkin orthogonality gives

$$(e^n, v^n) - \int_{t^{n-1}}^{t^n} (e, v_{ht}) + \eta(e, v_h)_U - (e^{n-1}, v_+^{n-1}) = 0. \tag{2.6}$$

Letting $\hat{e} = \mathbb{P}_n^{loc} u - u_h = e - (I - \mathbb{P}_n^{loc})u$ and using the properties of $\mathbb{P}_n^{loc}$ gives

$$\begin{aligned} &(\hat{e}^n, v^n) - \int_{t^{n-1}}^{t^n} (\hat{e}, v_{ht}) + \eta(\hat{e}, v_h)_U - (\hat{e}^{n-1}, v_+^{n-1}) \\ &= ((I - P_{n-1})u(t^{n-1}), v_+^{n-1}) - \int_{t^{n-1}}^{t^n} \eta((I - \mathbb{P}_n^{loc})u, v_h)_U. \end{aligned} \tag{2.7}$$

Setting $v_h = \hat{e}$ shows

$$\begin{aligned} &\frac{1}{2}|\hat{e}^n|^2 + \int_{t^{n-1}}^{t^n} \eta\|\hat{e}\|_U^2 + \frac{1}{2}|\hat{e}^{n-1} - \hat{e}_+^{n-1}|^2 - \frac{1}{2}|\hat{e}^{n-1}|^2 \\ &= ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1}) - \int_{t^{n-1}}^{t^n} \eta((I - \mathbb{P}_n^{loc})u, \hat{e})_U. \end{aligned}$$

The second term on the right-hand side is bounded using the Cauchy–Schwarz inequality, and the first term is bounded two different ways. Since $\hat{e}^{n-1} \in U_h^{n-1}$, an estimate independent of η is computed as

$$\begin{aligned} ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1}) &= ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1} - \hat{e}^{n-1}) \\ &\le |(I - P_{n-1})u(t^{n-1})|^2 + \frac{1}{4}|\hat{e}_+^{n-1} - \hat{e}^{n-1}|^2. \end{aligned}$$

An alternative estimate is obtained upon writing

$$\begin{aligned} ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1}) &= (P_n(I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1}) \\ &\le \|P_n(I - P_{n-1})u(t^{n-1})\|_{U'}\|\hat{e}_+^{n-1}\|_U. \end{aligned}$$

We next appeal to the following "inverse" estimate for functions in $\mathcal{P}(t^{n-1}, t^n, U^n)$:

$$\|\hat{e}_+^{n-1}\|_U^2 \le \frac{C_k^2}{\tau^n}\int_{t^{n-1}}^{t^n} \|\hat{e}\|_U^2.$$

The finite dimensionality of $\mathcal{P}_k(t^{n-1}, t^n)$ shows that such an estimate holds, and a scaling argument shows that the constant takes the form C_k^2/τ^n, where C_k depends only upon k when $\tau^n = t^n - t^{n-1}$. It follows that

$$((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1}) \leq \frac{C_k^2}{\tau^n \eta} \|P_n(I - P_{n-1})u(t^{n-1})\|_{U'}^2 + \frac{\eta}{4} \int_{t^{n-1}}^{t^n} \|\hat{e}\|_U^2.$$

Substituting these estimates into (2.8) and summing completes the proof. ∎

2.3. Discrete characteristic functions. To compute the error at arbitrary times $t \in [t^{n-1}, t^n)$ we would like to substitute $v_h = \chi_{[t^{n-1},t)} u_h$ into (2.4), where $\chi_{[t^{n-1},t)}$ is the characteristic function on $[t^{n-1}, t)$. Clearly this function is not in $\mathcal{U}_h$, so in this section discrete approximations of such characteristic functions are constructed to circumvent this problem.

The construction of the discrete characteristic functions is invariant under translation, so it is convenient to work on the interval $[0, \tau)$ with $\tau = t^n - t^{n-1}$. Consider first polynomials $p \in \mathcal{P}_k(0, \tau)$. A discrete approximation of $\chi_{[0,t)}p$ is the polynomial $\tilde{p} \in \{\tilde{p} \in \mathcal{P}_k(0, \tau) | \tilde{p}(0) = p(0)\}$ satisfying

$$\int_0^\tau \tilde{p}q = \int_0^t pq \qquad \forall\, q \in \mathcal{P}_{k-1}(0, \tau).$$

The above construction is motivated by the fact that setting $q = p'$ gives $\int_0^\tau p'\tilde{p} = \int_0^t pp' = (1/2)(p^2(t) - p^2(0))$.

This elementary construction can be extended to approximations of $\chi_{[0,t)}v$ for $v \in \mathcal{P}_k(0, \tau; V)$, where V is a linear space. If $v \in \mathcal{P}_k(0, \tau; V)$, write $v = \sum_{i=0}^k p_i(t)v_i$, where $\{p_i\} \subset \mathcal{P}_k(0, \tau)$ and $\{v_i\} \subset V$. Then the discrete approximation of $\chi_{[0,t)}v$ in $\mathcal{P}_k(0, \tau; V)$ is defined to be $\tilde{v} = \sum_{i=0}^k \tilde{p}_i(t)v_i$. If V is a semi-inner product space, it is clear that

$$\tilde{v}(0) = v(0) \quad \text{and} \quad \int_0^\tau (\tilde{v}, w)_V = \int_0^t (v, w)_V \qquad \forall\, w \in \mathcal{P}_{k-1}(0, \tau; V). \tag{2.8}$$

The function $\tilde{v}$ could have been characterized directly by this equation instead of using the two stage construction given here. However, it is useful to observe that v is independent of the choice of the specific space V or the inner product $(.,.)_V$.

2.3.1. Estimates for discrete characteristic functions. The construction of the discrete characteristic functions was purely algebraic. The next two lemmas bound the map $v \mapsto \tilde{v}$.

LEMMA 2.3. *The mapping $p \mapsto \tilde{p}$ on $\mathcal{P}_k(0, \tau)$ is linear and continuous, and there exists a constant $\hat{C}_k$ depending only upon k such that $\|\tilde{p} - p\|_{L^2(0,\tau)} \leq \hat{C}_k \|p\|_{L^2(t,\tau)}$. Moreover,*

$$\|\tilde{p} - \chi_{[0,t)}p\|_{L^2(0,\tau)} \leq \|\tilde{p} - p\|_{L^2(0,\tau)} + \|p - \chi_{[0,t)}p\|_{L^2(0,\tau)} \leq (1 + \hat{C}_k)\|p\|_{L^2(t,\tau)}$$

and $\|\tilde{p}\|_{L^2(0,\tau)} \leq (1 + \hat{C}_k)\|p\|_{L^2(0,\tau)}$.

Proof. Since $\tilde{p}(0) = p(0)$, the difference may be factored as $\tilde{p} - p = t\bar{p}$ with $\bar{p} \in \mathcal{P}_{k-1}(0, \tau)$. The definition of $\tilde{p}$ shows that

$$\int_0^\tau t\bar{p}q = \int_0^\tau (\tilde{p} - p)q = -\int_t^\tau pq \qquad \forall\, q \in \mathcal{P}_{k-1}(0, \tau).$$

Setting $q = \bar{p}$ gives

$$c_k \tau \int_0^\tau \bar{p}^2 \le \int_0^\tau t\bar{p}^2 = -\int_t^\tau p\bar{p},$$

where the equivalence of norms on $\mathcal{P}_k$ was used and the scaling was chosen to make c_k independent of τ. The Cauchy–Schwarz inequality then gives

$$(c_k\tau)^2 \int_0^\tau \bar{p}^2 \le \int_t^\tau p^2,$$

which implies

$$c_k^2 \int_0^\tau (\tilde{p} - p)^2 = c_k^2 \int_0^\tau t^2\bar{p}^2 \le c_k^2\tau^2 \int_0^\tau \bar{p}^2 \le \int_t^\tau p^2. \qquad \square$$

The next lemma bounds the map $v \mapsto \tilde{v}$ on $\mathcal{P}_k(0,\tau;V)$, where V is any (semi-) inner product space.

LEMMA 2.4. *Let V be a semi-inner product space; then the mapping $\sum_{i=0}^k p_i(t)v_i \mapsto \sum_{i=0}^k \tilde{p}_i(t)v_i$ on $\mathcal{P}_k(0,\tau;V)$ is continuous in $\|\cdot\|_{L^2[0,\tau;V]}$. In particular,*

$$\|\tilde{v}\|_{L^2[0,\tau;V]} \le C_k \|v\|_{L^2[0,\tau;V]}$$

and

$$\|\tilde{v} - \chi_{[0,t)}v\|_{L^2[0,\tau;V]} \le C_k \|v\|_{L^2[0,\tau;V]},$$

where $C_k = (k+1)^{1/2}(1 + \hat{C}_k)$.

Proof. Without loss of generality write $v = \sum_{i=0}^k p_i(t)v_i$, where $\{p_i\}$ form an orthonormal basis of $\mathcal{P}_k(0,\tau)$ in $L^2(0,\tau)$, so that $\|v\|^2_{L^2[0,\tau;V]} = \sum_{i=0}^k \|v_i\|_V^2$. The lemma then follows by direct computation:

$$\begin{aligned}
\int_0^\tau \|\tilde{v}\|_V^2 &= \int_0^\tau \sum_{i,j=0}^k \tilde{p}_i(t)\tilde{p}_j(t)(v_i, v_j)_V \\
&\le \sum_{i,j=0}^k \|\tilde{p}_i\|_{L^2(0,\tau)} \|\tilde{p}_j\|_{L^2(0,\tau)} \|v_i\|_V \|v_j\|_V \\
&\le (1 + \hat{C}_k)^2 \sum_{i,j=0}^k \|v_i\|_V \|v_j\|_V \\
&\le (1 + \hat{C}_k)^2 (k+1) \left(\sum_{i=0}^k \|v_i\|_V^2 \right) \\
&\le (1 + \hat{C}_k)^2 (k+1) \int_0^\tau \|v\|_V^2.
\end{aligned}$$

The second estimate follows similarly. $\square$

2.4. Error estimates at arbitrary times. We are now ready to prove the main result of this section which shows that the error estimate of Theorem 2.2, which held at discrete times, holds for every time.

THEOREM 2.5. *Let $u_h \in \mathcal{U}_h$ be the approximate solution of* (2.1) *computed using the DG scheme* (2.3). *Let $\hat{e} = \mathbb{P}_h^{loc} u - u_h$, where $\mathbb{P}_h^{loc}$ is the projection defined in Definition* 2.1. *Then there exists a constant $\tilde{C}_k$ depending only upon k such that*

$$\sup_{t^{n-1}\le t\le t^n} |\hat{e}(t)|^2 + \eta \int_0^{t^n} \|\hat{e}\|_U^2 + \sum_{i=0}^{n-1} |\hat{e}^i - \hat{e}_+^i|^2 \le \tilde{C}_k \left(|\hat{e}^0|^2 + \eta \int_0^{t^n} \|(I - \mathbb{P}_h^{loc})u\|_U^2 + \sum_{i=0}^{n-1} \min\left(|(I-P_i)u(t^i)|^2, \frac{C_k^2}{\tau^{i+1}\eta} \|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2 \right) \right),$$

where C_k is the constant appearing in Theorem 2.2.

Proof. Given Theorem 2.2, it suffices to estimate $\sup_{t^{n-1}\le t\le t^n} |\hat{e}(t)|^2$. To bound this term, fix $t \in [t^{n-1}, t^n)$ and substitute $v_h = \tilde{e}$ into (2.7), where $\tilde{e}$ is the discrete approximation of $\chi_{[t^{n-1},t)}\hat{e}$ constructed above, to get

$$\begin{aligned}
&\frac{1}{2}|\hat{e}(t)|^2 + \frac{1}{2}|\hat{e}^{n-1} - \hat{e}_+^{n-1}|^2 - \frac{1}{2}|\hat{e}^{n-1}|^2 \\
&= ((I-P_{n-1})u(t^{n-1}), e_+^{n-1}) - \eta \int_{t^{n-1}}^{t^n} ((I-\mathbb{P}_n^{loc})u, \tilde{e})_U + (\hat{e}, \tilde{e})_U \\
&= ((I-P_{n-1})u(t^{n-1}), e_+^{n-1}) + \eta \int_{t^{n-1}}^{t^n} \frac{1}{2}\|(I-\mathbb{P}_n^{loc})u\|_U^2 + \frac{1}{2}\|\tilde{e}\|^2 + \|\hat{e}\|_U \|\tilde{e}\|_U \\
&= ((I-P_{n-1})u(t^{n-1}), e_+^{n-1}) + \eta \int_{t^{n-1}}^{t^n} \frac{1}{2}\|(I-\mathbb{P}_n^{loc})u\|_U^2 + \left(\frac{C_k}{2} + 1\right) C_k \|\hat{e}\|_U^2.
\end{aligned}$$

Here C_k is the constant in the statement of Lemma 2.4. As in the proof of Theorem 2.2, the first term on the right-hand side may be bounded by

$$|(I-P_{n-1})u(t^{n-1})|^2 + \frac{1}{4}|\hat{e}_+^{n-1} - \hat{e}^{n-1}|^2 \quad \text{or}$$
$$\frac{C_k^2}{\tau^n \eta}\|P_n(I-P_{n-1})u(t^{n-1})\|_{U'}^2 + \frac{\eta}{4}\int_{t^{n-1}}^{t^n} \|\hat{e}\|_U^2.$$

It follows that

$$|\hat{e}(t)|^2 \le |\hat{e}^{n-1}|^2 + \eta \int_{t^{n-1}}^{t^n} C(k)\|\hat{e}\|_U^2 + \|(I-\mathbb{P}_n^{loc})u\|_U^2 + 2\min\left(|(I-P_{n-1})u(t^{n-1})|^2, \frac{C_k^2}{\tau^n\eta}\|P_n(I-P_{n-1})u(t^{n-1})\|_{U'}^2 \right),$$

where $C(k)$ is a constant depending upon k. The proof follows upon using Theorem 2.2 to bound the first two terms on the right-hand side. ∎

The following definition facilitates an interpretation of the above result, which is useful for the analysis of the parabolic problem in the next section.

DEFINITION 2.6. *The projection $\mathbb{P}_h : L^2[0,T;U] \cap H^1[0,T;U'] \to \mathcal{U}_h$ is the DG approximation of the function reconstructed from $f = u' + \eta B u$ and the initial data $u(0)$. That is, $u_h = \mathbb{P}_h u$ is the solution of* (2.3), *where $f = u' + \eta B u$.*

The previous theorem can then be interpreted as an estimate of the difference $\mathbb{P}_h u - \mathbb{P}_h^{loc} u$ between the global and local projections. Bounds on $\mathbb{P}_h^{loc} u - u$ follow directly from interpolation estimates [23].

We finish this section with some comments on the optimality of these estimates in the typical situation where $H = L^2(\Omega)$, $U \subset H^1(\Omega)$, and classical finite element piecewise polynomials of degree ℓ are used to construct the subspaces $\{U_h^n\}_{n=0}^N$. Assuming $u^0 = P_0(u(0))$, $\eta = O(1)$, and $\tau \sim h$, where h is the usual mesh parameter, the error term may be estimated as

$$\eta \int_0^{t^N} \|(I - \mathbb{P}_h^{loc})u\|^2 + \sum_{i=0}^{N-1} \min\left(|(I-P_i)u(t^i)|^2, \frac{C_k^2}{\tau\eta}\|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2\right)$$
$$\sim h^{2\ell} \int_0^T \|D^{\ell+1}u\|_{L^2(\Omega)}^2 + h^{2\ell} \sup_{0\le t\le T} \|D^\ell u\|_{L^2(\Omega)}^2.$$

Here the second term in the $\min(.,.)$ is used, and the regularity assumed matches the expected regularity parabolic equations:

$$\sup_{0\le t\le T} \|D^\ell u\|_{L^2(\Omega)}^2 + \int_0^T \|D^{\ell+1}u\|_{L^2(\Omega)}^2 \le C\left(\|D^\ell u(0)\|_{L^2(\Omega)}^2 + \int_0^T \|D^{\ell-1}f\|_{L^2(\Omega)}^2\right).$$

If the solution u is smooth, the error estimate is still useful when η is small, since if $\tau \sim h$, then

$$\eta \int_0^{t^N} \|(I - \mathbb{P}_h^{loc})u\|^2 + \sum_{i=0}^{N-1} \min\left(|(I-P_i)u(t^i)|^2, \frac{C_k^2}{\tau\eta}\|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2\right)$$
$$\sim \eta h^{2\ell} \int_0^T \|D^{\ell+1}u\|_{L^2(\Omega)}^2 + \min\left(h^{2\ell+1}, \frac{h^{2(\ell+1)}}{\eta}\right) \sup_{0\le t\le T} \|D^{\ell+1}u\|_{L^2(\Omega)}^2.$$

For $\eta \ll h$ a rate of $O(h^{\ell+1/2})$ will be observed. This is typical of the rate attained by the discontinuous method for hyperbolic equations [20], and when $\eta = 0$ is sub-optimal by a factor of $h^{1/2}$.

The above rates are guaranteed to hold independently of how the mesh is chosen at each time step. If the same mesh is used at each step, then $P_{i+1}(I - P_i) = P_i(I - P_i) = 0$, so the second term in the above estimates vanishes. Adaptive mesh refinement and coarsening strategies can be employed to control this term.

3. DG scheme for parabolic PDEs.

3.1. Formulation of the DG scheme.. We now consider approximations of (1.1) using the DG scheme. Optimal error estimates are derived by extending the ideas introduced in section 2. Let $a(\cdot,\cdot)$ denote the natural bilinear form associated with $A(\cdot)$. The following continuity and coercivity conditions will be assumed for $a(.,.)$.

ASSUMPTION 1. *There exist nonnegative constants C_a, C_α, c_a, c_α with $c_a \le C_a$ such that we have*

1. *continuity of the bilinear form and data:*

$$|a(t;u,v)| \le \left(c_a\|u\|^2 + C_a|u|^2\right)^{\frac{1}{2}}\left(c_a\|v\|^2 + C_a|v|^2\right)^{\frac{1}{2}}$$

and

$$|\langle F, v\rangle| \le \|F\|_*\left(c_a\|v\|^2 + C_a|v|^2\right)^{\frac{1}{2}},$$

2. *coercivity of the bilinear form:*

$$a(t;u,u) \geq c_\alpha \|u\|^2 - C_\alpha |u|^2.$$

Remark 3.

(1) The condition $c_a \leq C_a$ is not essential and is made to simplify some of the formula below.

(2) The coercivity constant will enter the estimates through the ratio c_a/c_α. The change to Lagrangian variables in the convection diffusion example presented in the introduction enables this ratio to be bounded independently of the diffusion parameter $\epsilon > 0$.

(3) The norm $\|.\|_*$ is a dual norm equivalent to $\|.\|_{U'}$; however, the constants relating them depend upon $1/c_a$, which may be large.

In this context the natural weak formulation of (1.1) is to find $u \in \mathcal{U} \equiv L^2[0,T;U] \cap H^1[0,T;U']$ such that

$$(u(T),v(T)) + \int_0^T (\langle -u, v_t\rangle + a(t,u,v)) = (u_0, v(0)) + \int_0^T \langle F, v\rangle \tag{3.1}$$

for all $v \in \mathcal{U}$. To approximate the solution of this weak formulation let $0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of $[0,T]$, and let $\{U_h^n\}_{n=0}^N \subset U$ be closed subspaces. The DG method constructs an approximate solution satisfying $u_h|_{(t^{n-1},t^n]} \in \mathcal{P}_k(t^{n-1},t^n;U_h^n)$ and

$$\begin{aligned}(u^n,v^n) + \int_{t^{n-1}}^{t^n} \Big(-\langle u_h, v_{ht}\rangle + a(\cdot;u_h,v_h)\Big) \\ = (u^{n-1},v_+^{n-1}) + \int_{t^{n-1}}^{t^n} \langle F, v_h\rangle \qquad \forall\, v_h \in \mathcal{P}_k(t^{n-1},t^n;U_h^n).\end{aligned} \tag{3.2}$$

Integration of the temporal term by parts yields the alternative representation

$$\begin{aligned}\int_{t^{n-1}}^{t^n} \Big(\langle u_{ht}, v_h\rangle + a(\cdot;u_h,v_h)\Big) + (u_+^{n-1} - u^{n-1}, v_+^{n-1}) \\ = \int_{t^{n-1}}^{t^n} \langle F, v_h\rangle \qquad \forall\, v_h \in \mathcal{P}_k(t^{n-1},t^n;U_h^n).\end{aligned} \tag{3.3}$$

3.2. Preliminary estimates. Classical bounds for the parabolic equation (3.2) are obtained upon selecting $u = v$ in (3.1); the discrete analogue would be to set $v_h = u_h$ in (3.3). Upon observing that

$$\int_{t^{n-1}}^{t^n} \langle u_{ht}, u_h\rangle + (u_+^{n-1} - u^{n-1}, u_+^{n-1}) = \frac{1}{2}|u^n|^2 - \frac{1}{2}|u^{n-1}|^2 + \frac{1}{2}|u_+^{n-1} - u^{n-1}|^2,$$

standard energy arguments and the continuity and coercivity hypotheses in Assumption 1 lead to the inequality

$$\begin{aligned}|u^n|^2 + c_\alpha \int_{t^{n-1}}^{t^n} \|u_h\|^2 + |u^{n-1} - u_+^{n-1}|^2 \\ \leq |u^{n-1}|^2 + \int_{t^{n-1}}^{t^n} \left(\left(1 + \frac{c_a}{c_\alpha}\right) \|F\|_*^2 + (2C_\alpha + C_a)|u_h|^2 \right).\end{aligned} \tag{3.4}$$

When $u_h \in \mathcal{P}_k(t^{n-1}, t^n; U_h^n)$ with $k \geq 2$, this inequality does not control $|u_h(s)|$ for $s \in (t^{n-1}, t^n)$. When u_h is piecewise constant or linear in time ($k = 0$ or 1), the terms on the left-hand side will dominate the last term on the right-hand side for sufficiently small time steps [23, Theorem 12.4]. The case $k > 2$ has not been completely addressed previously; typically strict coercivity is assumed so that $C_\alpha \leq 0$. In this situation it is possible to write

$$\int_{t^{n-1}}^{t^n} \langle F, u_h \rangle \leq \int_{t^{n-1}}^{t^n} \frac{1}{2\epsilon} \|F\|_*^2 + \frac{\epsilon}{2}(|u_h|^2 + \|u_h\|^2)$$

and use a Poincaré inequality to control the last term. However, this requires $\epsilon \sim c_\alpha$ and thus the term $1/2\epsilon$ is large when c_α is small.

Similar difficulties are encountered with error estimates when $k \geq 2$. Letting $e = u - u_h$, the orthogonality condition becomes

$$(e^n, v^n) + \int_{t^{n-1}}^{t^n} (-\langle e, v_{ht} \rangle + a(\cdot; e, v_h)) = (e^{n-1}, v_+^{n-1})$$

for all $v_h \in \mathcal{P}_k(t^{n-1}, t^n; U_h^n)$. Writing

$$e = u - u_h = (u - \mathbb{P}_h u) + (\mathbb{P}_h u - u_h) \equiv e_p + e_h,$$

where $\mathbb{P}_h$ is the projection defined in Definition 2.6, a computation shows that

$$\begin{aligned}(e_h^n, v^n) + \int_{t^{n-1}}^{t^n} (-\langle e_h, v_{ht} \rangle + a(\cdot; e_h, v_h)) - (e_h^{n-1}, v_+^{n-1}) \\ = -(e_p^n, v^n) + \int_{t^{n-1}}^{t^n} \langle e_p, v_{ht} \rangle + (e_p^{n-1}, v_+^{n-1}) - \int_{t^{n-1}}^{t^n} a(\cdot; e_p, v_h).\end{aligned}$$

Since $\mathbb{P}_h u$ is the DG approximation of (2.1), it follows that $e_p = u - \mathbb{P}_h u$ satisfies the orthogonality condition (2.6). The first three terms on the right-hand side then simplify to $-\eta \int_{t^{n-1}}^{t^n} (e_p, v_h)_U$. When $\eta = c_a$ this gives

$$\begin{aligned}(e_h^n, v^n) + \int_{t^{n-1}}^{t^n} (\langle -e_h, v_{ht} \rangle + a(\cdot; e_h, v_h)) - (e_h^{n-1}, v_+^{n-1}) \\ = -\int_{t^{n-1}}^{t^n} a(\cdot; e_p, v_h) + c_a (e_p, v_h)_U.\end{aligned} \tag{3.5}$$

This expression is identical in form to the original scheme (3.2) for u_h with $F(.) = -a(e_p, .) - \eta(e_p, .)_U$. Setting $v_h = e_h$ gives the analogue of (3.4),

$$\begin{aligned}|e_h^n|^2 + c_\alpha \int_{t^{n-1}}^{t^n} \|e_h\|^2 + |e_h^{n-1} - e_{h+}^{n-1}|^2 - |e_h^{n-1}|^2 \\ \leq \int_{t^{n-1}}^{t^n} \left(\left(1 + \frac{4c_a}{c_\alpha} \right) \left(c_a \|e_p\|^2 + C_a |e_p|^2 \right) + (2C_\alpha + c_a)|e_h|^2 \right).\end{aligned} \tag{3.6}$$

Again the natural energy arguments for the DG scheme fail to control $e_h(t)$ for $t \in (t^{n-1}, t^n)$ when c_α is small.

Remark 4. The projection $\mathbb{P}_h$ constructed using the discrete solution of an auxiliary equation is necessary when different subspaces are used for each time step, i.e.,

$U_h^n \neq U_h^{n-1}$. Using the "standard" projection, $\mathbb{P}_h^{loc}u$ in place of $\mathbb{P}_h$ (as in [23]) gives

$$\begin{aligned}(e_h^n, v^n) + \int_{t^{n-1}}^{t^n} \Big(\langle -e_h, v_{ht}\rangle + a(\cdot; e_h, v_h)\Big) \\ = (e_h^{n-1}, v_+^{n-1}) - \int_{t^{n-1}}^{t^n} a(\cdot; e_p^{n-1}, v_h) + (e_p^{n-1}, v_+^{n-1})\end{aligned}$$

with $e_p = u - \mathbb{P}_h^{loc}u$. Note that the last term is equal to $(e_p^{n-1}, v_+^{n-1} - w_-)$ for every $w_- \in U^{n-1}$, and when $U_h^n = U_h^{n-1}$ we may select $w_- = v_+^{n-1}$ to get (3.5) (without the last term on the right, which is not present if η is taken to be 0 instead of c_a).

3.3. Stability and error estimates. In this section we show how the estimates in (3.4) and (3.6) can be augmented to provide bounds on the solution and the error for all times, particularly, for the intermediate times $t \in (t^{n-1}, t^n)$. To do this the discrete characteristic functions developed in section 2.3 will be used. Since (3.2) for u_h and (3.5) for e_h are identical in form, the same line of argument can be applied to obtain bounds or an error estimate, respectively.

THEOREM 3.1. *Let $U \hookrightarrow H \hookrightarrow U'$ be a dense embedding of Hilbert spaces and $\mathcal{U}_h$ be the subspace of $L^2[0,T;U]$ defined in section 2.1, and, let the bilinear form $a : U \times U \to \mathbb{R}$ and linear form $F : U \to \mathbb{R}$ satisfy Assumption 1. Let $u \in L^2[0,T;U] \cap H^1[0,T;U']$ be the solution of (3.1) and $u_h \in \mathcal{U}_h$ be the approximate solution computed using the DG scheme (3.2) on the partition $0 = t^0 < t^1 < \cdots < t^N = T$, and set $\tau \equiv \max_n t^n - t^{n-1}$.*

Then there exists a constant $C > 0$ depending only on k (through the constant C_k of Lemma 2.4), the constants C_a, C_α, and the ratio c_a/c_α such that

$$\begin{aligned}(1-\lambda)|u^n|^2 + \lambda \sup_{t^{n-1}\le s\le t^n} |u_h(s)|^2 + \sum_{i=0}^{n-1} e^{C(t^{n-1}-t^i)}|u^i - u_+^i|^2 \\ + (1-\lambda)\frac{c_\alpha}{2}\int_0^{t^n} \mathrm{e}^{C(t^n-s)}\|u_h(s)\|^2 ds \\ \le \big(1 + T\mathcal{O}(\tau)\big)\left(\mathrm{e}^{Ct^n}|u_h^0|^2 + C\lambda\int_0^{t^n} \mathrm{e}^{C(t^n-s)}\|F(s)\|_*^2\, ds\right)\end{aligned}$$

and

$$\begin{aligned}(1-\lambda)|e_h^n|^2 + \lambda \sup_{t^{n-1}\le s\le t^n} |e_h(s)|^2 + \sum_{i=0}^{n-1} e^{C(t^{n-1}-t^i)}|e_h^i - e_{h+}^i|^2 \\ + (1-\lambda)\frac{c_\alpha}{2}\int_0^{t^n} \mathrm{e}^{C(t^n-s)}\|e_h(s)\|^2 ds \\ \le \big(1 + T\mathcal{O}(\tau)\big)\left(\mathrm{e}^{Ct^n}|e_h^0|^2 + C\lambda\int_0^{t^n} \mathrm{e}^{C(t^n-s)}\big(c_a\|e_p(s)\|^2 + C_a|e_p(s)|^2\big)\, ds\right),\end{aligned}$$

provided $C\tau < 1$. Here $\lambda = 1/(2C_k + 4C_k c_a/c_\alpha + 1) \in (0,1)$, and $e_p = u - \mathbb{P}_h u$ and $e_h = \mathbb{P}_h u - u_h$, where $\mathbb{P}_h : L^2[0,T;U] \cap H^1[0,T;U'] \to \mathcal{U}_h$ is the projection defined in Definition 2.6 (with parameter $\eta = c_a$).

Proof. Since the line of argument to prove each inequality is identical, we prove only the second.

Fix $t \in (t^{n-1}, t^n)$, and let $\tilde{e}_h \in \mathcal{P}_k(t^{n-1}, t^n; U_h^n)$ be the discrete approximation of $\chi_{[t^n,t)} e_h$ constructed in Lemma 2.4. Setting $v_h = \tilde{e}_h$ in (3.5) and moving the term $a(.; e_h, \tilde{e}_h)$ to the right-hand side gives

$$\frac{1}{2}|e_h(t)|^2 + \frac{1}{2}|e_h^{n-1} - e_{h+}^{n-1}|^2 = \frac{1}{2}|e_h^{n-1}|^2 - \int_{t^{n-1}}^{t^n} \left(a(.; e_p, \tilde{e}_h) + c_a(e_p, \tilde{e}_h)_U + a(\cdot; e_h, \tilde{e}_h)\right).$$

The last three terms are estimated separately:

$$\begin{aligned}
\int_{t^{n-1}}^{t^n} a(.; e_h, \tilde{e}_h) &\le \int_{t^{n-1}}^{t^n} (c_a\|e_h\|^2 + C_a|e_h|^2)^{1/2} (c_a\|\tilde{e}_h\|^2 + C_a|\tilde{e}_h|^2)^{1/2} \\
&\le \left(\int_{t^{n-1}}^{t^n} c_a\|e_h\|^2 + C_a|e_h|^2\right)^{1/2} \left(\int_{t^{n-1}}^{t^n} c_a\|\tilde{e}_h\|^2 + C_a|\tilde{e}_h|^2\right)^{1/2} \\
&\le C_k \int_{t^{n-1}}^{t^n} \Big(c_a\|e_h\|^2 + C_a|e_h|^2\Big).
\end{aligned}$$

Lemma 2.4 was used to bound $\tilde{e}_h$ in terms of e_h in the last line. A similar computation shows

$$\begin{aligned}
&\int_{t^{n-1}}^{t^n} a(.; e_p, \tilde{e}_h) + c_a(e_p, \tilde{e}_h)_U \\
&\qquad \le \frac{C_k}{2} \int_{t^{n-1}}^{t^n} \left(\left(\frac{c_a}{c_\alpha} + 1\right)(c_a\|e_p\|^2 + C_a|e_p|^2) + c_\alpha\|e_h\|^2 + C_a|e_h|^2\right).
\end{aligned}$$

Combining the above gives

$$\begin{aligned}
&|e_h(t)|^2 + |e_h^{n-1} - e_{h+}^{n-1}|^2 \le |e_h^{n-1}|^2 \\
&+ C_k \int_{t^{n-1}}^{t^n} \left(\left(1 + \frac{c_a}{c_\alpha}\right)(c_a\|e_p\|^2 + C_a|e_p|^2) + c_\alpha\left(1 + \frac{2c_a}{c_\alpha}\right)\|e_h\|^2 + 3C_a|e_h|^2\right). \qquad (3.7)
\end{aligned}$$

Now form the convex combination of $(1-\lambda)$ for (3.6) and λ for (3.7). The coefficient λ is chosen so that the term involving $\|e_h\|^2$ on the right-hand side of (3.7) is dominated by the corresponding term on the left-hand side of (3.6). Specifically, let

$$\lambda C_k\left(1 + \frac{2c_a}{c_\alpha}\right) = \frac{1}{2}(1-\lambda) \qquad \text{or} \qquad \lambda = \frac{1}{(2C_k + 4C_k c_a/c_\alpha + 1)}.$$

This gives an estimate of the form

$$\begin{aligned}
(1-\lambda)|e_h^n|^2 + \lambda|e_h(t)|^2 + (1-\lambda)\frac{c_\alpha}{2}\int_{t^{n-1}}^{t^n} \|e_h\|^2 + |e_h^{n-1} - e_{h+}^{n-1}|^2 \qquad (3.8) \\
\le |e_h^{n-1}|^2 + C\lambda \int_{t^{n-1}}^{t^n} \Big(c_a\|e_p\|^2 + C_a|e_p|^2 + |e_h|^2\Big),
\end{aligned}$$

where the dependence of C upon the coercivity constants c_α and c_a is only through the ratio c_a/c_α. Bound the first and last terms on the right-hand side by

$$|e_h^{n-1}|^2 \le (1-\lambda)|e_h^{n-1}|^2 + \lambda \sup_{t^{n-2} < s \le t^{n-1}} |e_h(s)|^2$$

and

$$\int_{t^{n-1}}^{t^n} |e_h|^2 \le \tau^n \sup_{t^{n-1}<s\le t^n} |e_h(s)|^2, \qquad \tau^n \equiv t^n - t^{n-1},$$

respectively, and select the time t on the left-hand side so that $|e_h(t)| = \sup_{t^{n-1}<s\le t^n} |e_h(s)|$, to get

$$(1-\lambda)|e_h^n|^2 + \lambda(1-C\tau^n) \sup_{t^{n-1}<s\le t^n} |e_h(s)|^2 + (1-\lambda)\frac{c_\alpha}{2}\int_{t^{n-1}}^{t^n} \|e_h\|^2 + |e_h^{n-1} - e_{h+}^{n-1}|^2$$
$$\le (1-\lambda)|e_h^{n-1}|^2 + \lambda \sup_{t^{n-2}<s\le t^{n-1}} |e_h(s)|^2 + C\lambda \int_{t^{n-1}}^{t^n} (c_a\|e_p\|^2 + C_a|e_p|^2).$$

Upon introducing a factor $(1 - C\tau^n)$ in front of the first term, this inequality takes the form

$$(1-C\tau^n)\alpha^n + \beta^n \le \alpha^{n-1} + f^n,$$

and the theorem follows from the discrete Gronwall inequality. ∎

If the norms $|||.|||_\infty$, $|||.|||_2$ and jump term $J_N(e)$ are defined by

$$|||v|||_\infty^2 = \sup_{0\le s\le T} |v(s)|^2 + c_\alpha \int_0^T \mathrm{e}^{C(T-s)} \|v(s)\|^2\, ds,$$

$$|||v|||_2^2 = \int_0^T \mathrm{e}^{C(T-s)} |v(s)|^2\, ds + c_a \int_0^T e^{C(T-s)} \|v(s)\|^2\, ds,$$

and

$$J_N^2(v) = \sum_{i=0}^{N-1} \mathrm{e}^{C(T-t^i)} |v^i - v_+^i|^2,$$

then Theorem 3.1 states

$$|||\mathbb{P}_h u - u_h|||_\infty^2 + J_N^2(\mathbb{P}_h u - u_h) \le C(T)\left(|P_0 u(0) - u^0|^2 + |||u - \mathbb{P}_h u|||_2^2\right).$$

Since $\int_0^T \mathrm{e}^{C(T-s)}|e_p(s)|^2 \le \mathrm{e}^{CT} \sup_{0\le s\le T} |e_p(s)|^2$, various symmetric error estimates follow. For example, setting $u^0 = P_0 u(0)$ and using the triangle inequality gives

$$|||u - u_h|||_2 \le C(T)|||u - \mathbb{P}_h u|||_2 \quad \text{and} \quad |||u - u_h|||_\infty \le C(T)|||u - \mathbb{P}_h u|||_\infty.$$

Since $\mathbb{P}_h u$ is not a local projection, classical interpolation theory does not immediately yield rates of convergence for a specific problem. However, the results of section 2.4 may be used to estimate the right-hand sides of the above in terms of the local projection $\mathbb{P}_h^{loc} u$.

THEOREM 3.2. *Under the assumptions in Theorem* 3.1 *there exists a positive constant $C(T)$ depending only on k (through the constant C_k of Lemma* 2.4*), the constants C, λ, C_a, C_α, and the ratio c_a/c_α, and the final time T such that the following*

estimate holds:

$$\begin{aligned}|||u-u_h|||_\infty^2 &\le C(T)\big(|e_h^0|^2+|||u-\mathbb{P}_h^{loc}u|||_2^2+|||\mathbb{P}_hu-\mathbb{P}_h^{loc}u|||_\infty^2\big)\\ &\le C(T)\left(|e_h^0|^2+|||u-\mathbb{P}_h^{loc}u|||_\infty^2+c_a\int_0^{t^n}\|(I-\mathbb{P}_h^{loc})u\|_U^2\right.\\ &\qquad\left.+\sum_{i=0}^{n-1}\min\left(|(I-P_i)u(t^i)|^2,\frac{C_k^2}{\tau^{i+1}c_a}\|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2\right)\right),\end{aligned}$$

where $e_h^0=P_0u(0)-u^0$, $\mathbb{P}_h^{loc}u$ *is the local projection defined in Definition* 2.1, *and* $P_n:H\to U_h^n$ *is the orthogonal projection.*

Remark 5. Estimates for the jump terms $J_N(u-u_h)$ can also be obtained. An application of the triangle inequality gives

$$J_N(u-u_h)\le J_N(u-\mathbb{P}_h^{loc}u)+J_N(\mathbb{P}_h^{loc}u-\mathbb{P}_hu)+J_N(\mathbb{P}_hu-u_h).$$

Bounding the last term using Theorem 3.1 shows

$$\begin{aligned}J_N^2(u-u_h)&\le C(T)\Big(J_N(u-\mathbb{P}_h^{loc}u)^2+J_N^2(\mathbb{P}_h^{loc}u-\mathbb{P}_hu)+|e_n^0|^2+|||u-\mathbb{P}_hu|||_2^2\Big)\\ &\le C(T)\Big(|e_n^0|^2+J_N(u-\mathbb{P}_h^{loc}u)^2+|||u-\mathbb{P}_h^{loc}u|||_2^2\\ &\qquad+J_N^2(\mathbb{P}_h^{loc}u-\mathbb{P}_hu)+|||\mathbb{P}_hu-\mathbb{P}_n^{loc}u|||_2^2\Big).\end{aligned}$$

The last two terms may be estimated using Theorem 2.5 to give

$$\begin{aligned}J_N^2(u-u_h)\le C(T)\left(|e_n^0|^2+J_N(u-\mathbb{P}_h^{loc}u)^2+|||u-\mathbb{P}_h^{loc}u|||_2^2+c_a\int_0^{t^n}\|(I-\mathbb{P}_h^{loc})u\|_U^2\right.\\ \left.+\sum_{i=0}^{n-1}\min\left(|(I-P_i)u(t^i)|^2,\frac{C_k^2}{\tau^{i+1}c_a}\|P_{i+1}(I-P_i)u(t^i)\|_{U'}^2\right)\right).\end{aligned}$$

When the solution is smooth, the second term, $J_N(u-\mathbb{P}_h^{loc}u)$, will typically dominate the other terms.

Appendix A. Discrete Gronwall inequality. If $(1-C\tau^n)a^n+b^n\le a^{n-1}+f^n$, the discrete Gronwall inequality states that if $\max_n C\tau^n<1$, then

$$a^N+\sum_{n=1}^N\frac{b^n}{\prod_{i=n}^N(1-C\tau^i)}\le\frac{a^0}{\prod_{i=1}^N(1-C\tau^i)}+\sum_{n=1}^N\frac{f^n}{\prod_{i=n}^N(1-C\tau^i)}.$$

Since

$$\exp\left(\sum_{i=n}^NC\tau^i\right)\le\frac{1}{\prod_{i=n}^N(1-C\tau^i)}\le\left(1-\sum_{i=n}^N(C\tau^i)^2\right)^{-1}\exp\left(\sum_{i=n}^NC\tau^i\right)$$

and $(1-\sum_{i=n}^N(C\tau^i)^2)\ge1-C^2T\tau$, where $\tau=\max_n\tau^n$, we may write

$$a^N+\sum_{n=1}^Ne^{C(t^N-t^n)}b^n\le\big(1+T\mathcal{O}(\tau)\big)\left(\mathrm{e}^{Ct^N}a^0+\sum_{n=1}^N\mathrm{e}^{C(t^N-t^n)}f^n\right),$$

where $t^n=\sum_{i=1}^n\tau^n$.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] M. Bause and P. Knabner, *Uniform error analysis for Lagrange–Galerkin approximations of convection-dominated problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1954–1984.

[3] B. Cockburn, G. E. Karnadiakis, and C. W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Springer, Berlin, 2000, pp. 3–50.

[4] M. Delfour, W. Hager, and F. Trochu, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comput., 36 (1981), pp. 455–473.

[5] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[6] T. Dupont, *Mesh modification for evolutionary equations*, Math. Comp., 39 (1982), pp. 85–107.

[7] T. F. Dupont and Y. Liu, *Symmetric error estimates for moving mesh Galerkin methods for advection-diffusion equations*, SIAM J. Numer. Anal., 40 (2002), pp. 914–927.

[8] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* I. *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[9] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* II. *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.

[10] K. Eriksson, C. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modél. Math. Anal. Numér., 29 (1985), pp. 611–643.

[11] P. Jamet, *Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain*, SIAM J. Numer. Anal., 15 (1978), pp. 912–928.

[12] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[13] P. Lasaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.

[14] S. Larsson, V. Thomée, and L. B. Walhbin, *Numerical solution of parabolic integro-differential equations by the discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 45–71.

[15] Y. Liu, R. E. Bank, T. F. Dupont, S. Garcia, and R. F. Santos, *Symmetric error estimates for moving mesh mixed methods for advection-diffusion equations*, SIAM J. Numer. Anal., 40 (2003), pp. 2270–2291.

[16] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.

[17] K. Miller, *Moving finite elements* II, SIAM J. Numer. Anal., 18 (1981), pp. 1033–1057.

[18] K. Miller and R. N. Miller, *Moving finite elements* I, SIAM J. Numer. Anal., 18 (1981), pp. 1019–1032.

[19] K. W. Morton, A. Priestley, and E. Süli, *Stability of the Lagrange–Galerkin method with nonexact integration*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 625–653.

[20] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[21] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, Boston, 1979.

[22] R. E. Showlater, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, American Mathematical Society, Providence, RI, 1997.

[23] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997

[24] N. J. Walkington, *Convergence of the discontinuous Galerkin method for discontinuous solutions*, SIAM J. Numer. Anal., 42 (2005), pp. 1801–1817.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 367–384

SHAPE DESIGN IN AORTO-CORONARIC BYPASS ANASTOMOSES USING PERTURBATION THEORY*

VALERY AGOSHKOV†, ALFIO QUARTERONI‡§, AND GIANLUIGI ROZZA§

Abstract. In this paper we present a new approach in the study of aorto-coronaric bypass anastomoses configurations based on small perturbation theory. The theory of optimal control based on adjoint formulation is applied in order to optimize the shape of the zone of the incoming branch of the bypass (the toe) into the coronary (see Figure 2.1). The aim is to provide design indications in the perspective of future development for prosthetic bypasses.

Key words. optimal control, shape optimization, small perturbation theory, finite elements, haemodynamics, aorto-coronaric bypass anastomoses, design of improved medical devices

AMS subject classifications. 35Q30, 49J20, 65N30, 76Z05, 92C50

DOI. 10.1137/040613287

1. Introduction.

We consider the application of optimal control approaches to shape optimization of aorto-coronaric bypass anastomoses [22]. We analyze the "first correction" method which is derived by applying a perturbation method to the initial problem in a domain $\Omega \subset \mathbb{R}^2$ whose boundary $\partial\Omega$ is parameterized by a suitable function f. Then we propose numerical methods for its solution.

The surgical realization of a bypass to overcome a critically stenosed artery is a very common practice in an everyday cardiovascular clinic.

Improvement in the understanding of the genesis of coronary diseases is very important as it allows the reduction of surgical and postsurgical failures. It may also suggest new means in bypass surgical procedures with less invasive methods and the creation of a new shape in bypass configuration [19].

Generally speaking, mathematical modelling and numerical simulation can allow better understanding of phenomena involved in vascular diseases [6, 23, 24].

When a coronary artery is affected by a stenosis, the heart muscle can't be properly oxygenated through blood. Aorto-coronaric anastomosis restores the oxygen amount through a bypass surgery downstream of an occlusion.

At present, different kinds and shapes for aorto-coronaric bypass anastomoses are available and consequently different surgery procedures are used to set up a bypass.

*Received by the editors August 11, 2004; accepted for publication August 2, 2005; published electronically March 7, 2006. This work has been supported in part by the Swiss National Science Foundation (Project 20-65110.01) and by Italian Cofin2003-MIUR (Italian Research, University, and Education Ministry) Project "Numerical Modelling for Scientific Computing and Advanced Applications" and by Indam (Italian Institute of Advanced Mathematics).

http://www.siam.org/journals/sinum/44-1/61328.html

†Institute of Numerical Mathematics, Russian Academy of Sciences, 119991 GSP-1 Moscow, Russia (Agoshkov@inm.ras.ru). This work was prepared when this author was visiting Bernoulli Center of the Swiss Federal Institute of Technology Lausanne in the framework of the special semester on "The Mathematical Modelling of the Cardiovascular System." This author also acknowledges the support of the Russian Foundation for Basic Research (Project 04-01-00615).

‡Corresponding author. MOX, Dipartimento di Matematica "Francesco Brioschi," Politecnico di Milano, 20133 Milano, Italy (Alfio.Quarteroni@epfl.ch).

§Chair of Modelling and Scientific Computing (CMCS-IACS), École Polytechnique Fédérale de Lausanne, Station 8, CH-1015, Lausanne, Switzerland (Gianluigi.Rozza@epfl.ch). The third author acknowledges financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00270 HaeMOdel.

A bypass can be made up either by organic material (e.g., the saphena vein taken from patient's legs or the mammary artery) or by prosthetic material. The current saphenous bypass solution requires the extraction of the saphena vein with possible complications. In this respect, prosthetic bypasses are less invasive. They may feature very different shapes for bypass anastomoses, such as, e.g., cuffed arteriovenous access grafts. Different cuffed models are used such as Taylor Patch [2] and Miller Cuff Bypass [4] but also standard end-to-side anastomoses at different graft angles [3] or other shaped carbon-fiber prostheses. In the cardiovascular system, altered flow conditions such as separation, flow reversal, low and oscillatory shear stress areas, and abnormal pulse patterns are all recognized as potentially important factors in the development of arterial diseases (see [15, 18]). For all these different aspects the design of artificial arterial bypass is a very complex problem. Carbon fiber and collagen cuffed grafts instead of natural saphenous vein can be used for studying new shape design without needing "in loco" reconstruction. In this framework, optimal control (see Lions [12]) by perturbation theory (see Van Dyke [31]) provides a new approach to the problem, with the goal of improving arterial bypass graft on the basis of a better understanding of fluid dynamics aspects involved in the bypass studying.

2. Notation and problem statement. Let Ω be a bounded domain of $\mathbb{R}^2$, $\Gamma \equiv \partial\Omega$ is the boundary of Ω, $\overline{\Omega} = \Omega \cup \partial\Omega$, $\underline{x} := (x, y)$ is a point of $\overline{\Omega}$. For every scalar function ϕ and a vector function $\underline{v}$ whose components are u, v, we recall the definition of the following operators:

$$\nabla\phi = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}\right), \ \nabla \cdot \underline{v} := \mathrm{div}(\underline{v}) := \mathcal{D}(\underline{v}) = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y},$$
$$\nabla \times \underline{v} := rot(\underline{v}) = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \ \underline{rot}(\phi) = \left(\frac{\partial\phi}{\partial y}, -\frac{\partial\phi}{\partial x}\right).$$

We then recall

$$\underline{rot}(\nabla \times \underline{v}) = -\Delta\underline{v} + \nabla(\nabla \cdot \underline{v}), \ \Delta\phi = \nabla \cdot (\nabla\phi).$$

In what follows, vectors are marked with an underlined notation $\underline{v}$, aggregation of vector quantities $\underline{v}$ with scalar quantities p are indicated with $\underline{Q}$ ($\underline{Q} = (\underline{v}, p)$), $\underline{\Phi}$ or $\hat{\underline{\Phi}}$.

Consider an idealized, two-dimensional bypass bridge configuration of Figure 2.1 and the domain on Figure 2.2, where the dotted line represents the geometry of the complete anastomosis; Γ_{w_2} is the section of the original artery, Γ_{in} is the new anastomosis inflow after bypass surgery, and Γ_{out} is the anastomosis outflow. We consider the following boundary value problem for the Stokes equations [33], used to model low Reynolds blood flow in this study. For mathematical aspects related with fluid mechanics, see, for example, [14]. The problem reads: find $\underline{v}, p$ s.t.

$$\begin{cases} -\nu\Delta\underline{v} + \nabla p = \underline{F} \quad \text{in } \Omega, \\ \nabla \cdot \underline{v} = 0 \quad \text{in } \Omega, \\ \underline{v} = \underline{v}_{in} \quad \text{on } \Gamma_{in}, \ \underline{v} = 0 \quad \text{on } \Gamma_{w_1} \cup \Gamma_{w_3}, \\ -p \cdot \underline{n} + \nu\frac{\partial\underline{v}}{\partial n} = \underline{g}_{out} \quad \text{on } \Gamma_{out} \cup \Gamma_{w_2}, \end{cases} \tag{2.1}$$

where $\underline{n} = (n_1, n_2)$ is the outward unit normal vector on Γ, $\underline{F} = \underline{F}(x, y)$, $\underline{v}_{in} = \underline{v}_{in}(x, y)$, $\underline{g}_{out} = \underline{g}_{out}(x, y)$ are given vector functions, $\nu = \text{const} > 0$ and $v_f = \{\underline{v}_{in}$ on $\Gamma_{in}; \underline{0}$ on $\Gamma_{w_1} \cup \Gamma_{w_3}\}$. In the following we may need to impose some additional restriction on p (for example, $\int_\Omega p d\Omega = 0$ if $\Gamma_{in} = \Gamma$).

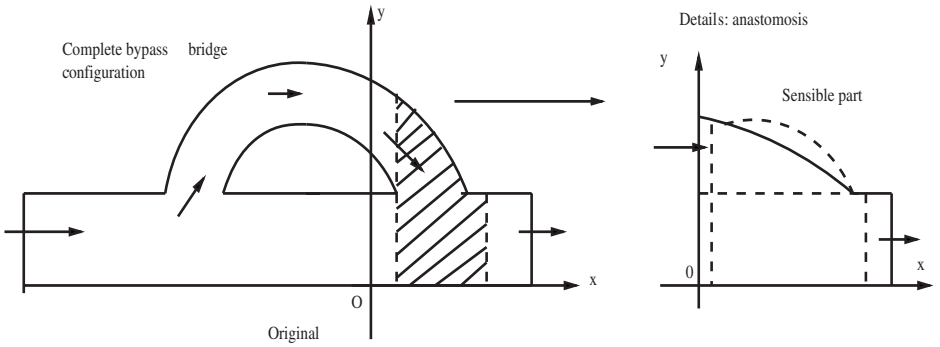


FIG. 2.1. *Idealized, 2-D bypass bridge configuration (left) and details of the sensible part for the optimization process (right). The dotted curve represents a possible shape variation.*

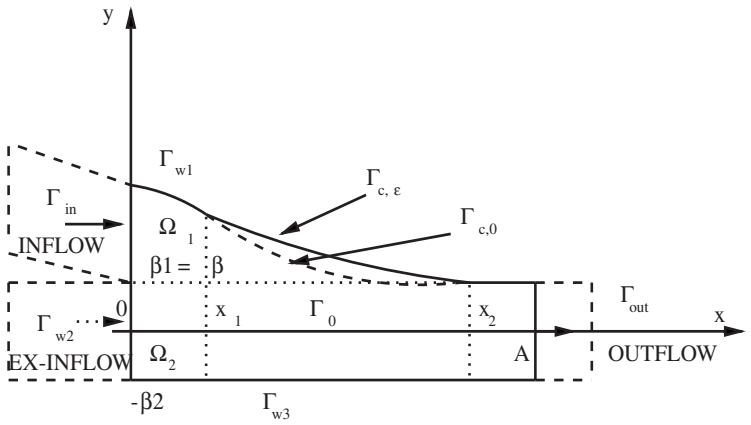


FIG. 2.2. $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2, \Gamma_w = \Gamma_{w_1} \cup \Gamma_{w_2} \cup \Gamma_{w_3}, \Gamma_0 = \partial\Omega_1 \cap \partial\Omega_2$.

The subset $\Gamma_{c,\varepsilon}$ of Γ_{w_1} is parametrized by a function $f(x,\varepsilon)$ of $\underline{x} \in [x_1, x_2]$ and of small parameter $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$, $\varepsilon_0 = \text{const}$. More precisely, we assume that $f(x,\varepsilon)$ can be developed as follows:

$$f(x,\varepsilon) = f_0(x) + \varepsilon f_1(x) + \varepsilon^2 f_2(x) + \dots, \tag{2.2}$$

where $f_k \in \mathbb{W}^{1,\infty}(x_1,x_2)$, for $k = 0$ (we recall that $\mathbb{W}^{1,\infty}(x_1,x_2)$ is the space of functions $f_k \in \mathbb{L}^\infty(x_1,x_2)$ such that all the distribution derivatives of the first order of f_k are functions of $\mathbb{L}^\infty(x_1,x_2)$), and $f_k \in \mathbb{W}_0^{1,\infty}(x_1,x_2)$, for $k \geq 1$, so that $f_k(x_1) = f_k(x_2) = 0, k \geq 1$. Here the function $f_0(x) > 0$ describes the original subset $\Gamma_{c,0}$ of the boundary of "unperturbed domain...," $\Gamma_{w0} \equiv \partial\Omega_0$ of the domain Ω_0 (see Figure 2.3 (left)), while $f_k(x), k \geq 1$ could be unknown when dealing with a control problem (see section 4).

The weak statement of (2.1) reads: find $\underline{v} \in (\mathbb{H}^1(\Omega))^2$, $p \in \mathbb{L}^2(\Omega)$ s.t.

$$\begin{cases} a(\underline{v},\underline{\hat{v}}) = b(p,\underline{\hat{v}}) + G(\underline{\hat{v}}) \quad \forall\ \underline{\hat{v}} \in \mathbb{X}, \\ b(\hat{p},\underline{v}) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\ \underline{v} = \underline{v}_f \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3}, \end{cases} \tag{2.3}$$

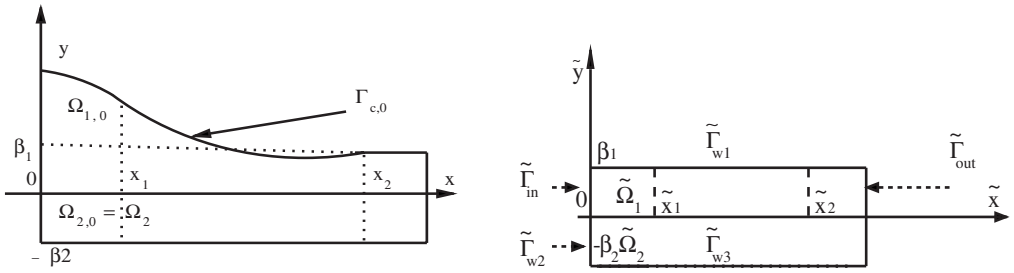


FIG. 2.3. *"Unperturbed domain"* Ω_0, $\overline{\Omega}_0 = \overline{\Omega}_{1,0} \cup \overline{\Omega}_{2,0}$ *(left). The "simple" domain* $\tilde{\Omega}$ *(right).*

where with $\underline{\hat{v}}$ we indicate test functions and

$$a(\underline{v}, \underline{\hat{v}}) = \int_\Omega \nu \nabla \underline{v} \cdot \nabla \underline{\hat{v}} d\Omega$$
$$b(p, \underline{\hat{v}}) = \int_\Omega p \nabla \cdot \underline{\hat{v}} d\Omega,\ G(\underline{\hat{v}}) = \int_\Omega \underline{F} \cdot \underline{\hat{v}} d\Omega + \int_{\Gamma_{out} \cup \Gamma_{w_2}} \underline{g}_{out} \cdot \underline{\hat{v}} d\Gamma,$$
$$\mathbb{X} := \{\underline{\hat{v}} : \underline{\hat{v}} \in (\mathbb{H}^1(\Omega))^2, \underline{\hat{v}} = 0 \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3}\}.$$

Although $a(.,.), b(.,.)$, and $G(.)$, depend on the parametrization f of the part $\Gamma_{c,\varepsilon}$, this dependence will be understood for simplicity of notations.

3. The problem for the perturbation functions. Let us introduce the reference (simple-shaded) domains $\tilde{\Omega}_1 = \{0 < \tilde{x} < A, 0 < \tilde{y} < \beta_1 \equiv \beta\}$, $\tilde{\Omega}_2 = \{0 < \tilde{x} < A, -\beta_2 < \tilde{y} < 0\}$, and $\overline{\tilde{\Omega}} = \overline{\tilde{\Omega}}_1 \cup \overline{\tilde{\Omega}}_2$ (see Figure 2.3 (right)). Then we assume that $f(x, \varepsilon) > 0$ and consider the following variables transformation:

$$\mathrm{T}_f : \overline{\Omega}_1 \cup \overline{\Omega}_2 \to \overline{\tilde{\Omega}},\ \underline{\tilde{x}} = \mathrm{T}_f(\underline{x}),$$

such as T_f is the identity in Ω_2, while $\mathrm{T}_f(x, y) = (x, \frac{\beta}{f(x,y)} y)$ in Ω_1. We set $\underline{\tilde{x}} = (\tilde{x}, \tilde{y})$ and define

$$\underline{\tilde{v}}(\underline{\tilde{x}}) := \underline{v} \circ \mathrm{T}_f^{-1}(\underline{\tilde{x}}) = \underline{v}(\tilde{x}, \tilde{y} f(\tilde{x}, \varepsilon)/\beta),$$

where $\underline{\tilde{v}} = (\tilde{u}, \tilde{v})$. Then,

$$dxdy = \frac{f(\tilde{x}, \epsilon)}{\beta} d\tilde{x} d\tilde{y}$$

and the following relations hold:

$$\text{(3.1)} \qquad \begin{cases} \frac{\partial \phi}{\partial y}(\underline{\tilde{x}}) = \frac{\beta}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{\phi}(\underline{\tilde{x}})}{\partial \tilde{y}}, \\ \frac{\partial \phi}{\partial x}(\underline{\tilde{x}}) = \frac{\partial \tilde{\phi}(\underline{\tilde{x}})}{\partial \tilde{x}} - \tilde{y} \frac{f_x(\tilde{x}, \varepsilon)}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{\phi}(\underline{\tilde{x}})}{\partial \tilde{y}} \quad \left(\text{with } f_x := \frac{df}{dx}\right), \end{cases}$$

$$\text{(3.2)} \qquad \begin{cases} \tilde{\mathcal{D}}(f)\underline{\tilde{v}}(\underline{\tilde{x}}) := ((\nabla \cdot \underline{v}) \circ \mathrm{T}_f^{-1})(\underline{\tilde{x}}) = \frac{\partial \tilde{u}}{\partial \tilde{x}} - \tilde{y} \frac{f_x(\tilde{x}, \epsilon)}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{u}}{\partial \tilde{y}} + \frac{\beta}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{v}}{\partial \tilde{y}}, \\ \tilde{\mathcal{R}}(f)\underline{\tilde{v}}(\underline{\tilde{x}}) := ((\nabla \times \underline{v}) \circ \mathrm{T}_f^{-1})(\underline{\tilde{x}}) = \frac{\partial \tilde{v}}{\partial \tilde{x}} - \tilde{y} \frac{f_x(\tilde{x}, \epsilon)}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{v}}{\partial \tilde{y}} - \frac{\beta}{f(\tilde{x}, \varepsilon)} \frac{\partial \tilde{u}}{\partial \tilde{y}}. \end{cases}$$

Then in $\tilde{\Omega}$ we have

$$\tilde{\mathcal{D}}(f)\underline{\tilde{v}} = m_2 \tilde{\nabla} \cdot \underline{\tilde{v}} + m_1 \tilde{\mathcal{D}}(f)\underline{\tilde{v}},\ \tilde{\mathcal{R}}(f)\underline{\tilde{v}} = m_2 \tilde{\nabla} \times \underline{\tilde{v}} + m_1 \tilde{\mathcal{R}}(f)\underline{\tilde{v}},$$

where $\tilde{\nabla}\phi := \left(\frac{\partial\phi}{\partial\tilde{x}}, \frac{\partial\phi}{\partial\tilde{y}}\right)$, while m_s is the characteristic function of Ω_s $(s = 1, 2)$. To simplify the notations, from now on we will set (unless otherwise specified)

$$\tilde{\underline{x}} = \underline{x},\ \tilde{\underline{v}}(\tilde{x}, \tilde{y}) := \underline{v}(x, y),\ \tilde{u} = u,\ \tilde{v} = v, \ldots, \tilde{\mathcal{D}} = \mathcal{D}, \tilde{\mathcal{R}} = \mathcal{R},\ \tilde{\Omega} \equiv \Omega,\ \tilde{\Gamma}_{w_k} \equiv \Gamma_{w_k}.$$

Then problem (2.3) in the new variables reads as follows:

$$\left\{\begin{array}{l} a(f; \underline{v}, \hat{\underline{v}}) = b(f; p, \hat{\underline{v}}) + G(f; \hat{\underline{v}}) \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\ b(f; \hat{\underline{p}}, \underline{v}) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\ \underline{v} = \underline{v}_f \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3}. \end{array}\right. \tag{3.3}$$

We have emphasized the dependence of $a(f; ., .), b(f; ., .)$, and $G(f; .)$ on f. Therefore, (with $\Omega_1 \equiv \tilde{\Omega}_1, \Omega_2 \equiv \tilde{\Omega}_2$):

$$\begin{aligned} a(f; \underline{v}, \hat{\underline{v}}) &= a_1(f; \underline{v}, \hat{\underline{v}}) + a_2(\underline{v}, \hat{\underline{v}}), \\ a_1(f; \underline{v}, \hat{\underline{v}}) &= \int_{\Omega_1} \frac{f\nu}{\beta} \left(\left(\frac{\partial \underline{v}}{\partial x} - \frac{y f_x}{f}\frac{\partial \underline{v}}{\partial y}\right) \cdot \left(\frac{\partial \hat{\underline{v}}}{\partial x} - \frac{y f_x}{f}\frac{\partial \hat{\underline{v}}}{\partial y}\right) + \frac{\beta^2}{f^2}\frac{\partial \underline{v}}{partialy} \cdot \frac{\partial \hat{\underline{v}}}{\partial y}\right) dxdy \\ a_2(\underline{v}, \hat{\underline{v}}) &= \int_{\Omega_2} \nu \left(\frac{\partial \underline{v}}{\partial x} \cdot \frac{\partial \hat{\underline{v}}}{\partial x} + \frac{\partial \underline{v}}{\partial y} \cdot \frac{\partial \hat{\underline{v}}}{\partial y}\right) dxdy, \\ b(f; p, \hat{\underline{v}}) &= b_1(f; p, \hat{\underline{v}}) + b_2(p, \hat{\underline{v}}), \\ b_1(f; p, \hat{\underline{v}}) &= \int_{\Omega_1} \frac{f}{\beta} p\mathcal{D}(f)\hat{\underline{v}} dxdy, \quad b_2(p, \hat{\underline{v}}) = \int_{\Omega_2} p\nabla \cdot \hat{\underline{v}} dxdy, \\ G(f; \hat{\underline{v}}) &= G_1(f; \hat{\underline{v}}) + G_2(\hat{\underline{v}}), \\ G_1(f; \hat{\underline{v}}) &= \int_{\Omega_1} \frac{f}{\beta} \underline{F} \cdot \hat{\underline{v}} dxdy + \int_{(\Gamma_{out} \cup \Gamma_{w2}) \cap \partial\Omega_1} \underline{g}_{out} \cdot \hat{\underline{v}} d\Gamma, \\ G_2(\hat{\underline{v}}) &= \int_{\Omega_2} \underline{F} \cdot \hat{\underline{v}} dxdy + \int_{(\Gamma_{out} \cup \Gamma_{w2}) \cap \partial\Omega_2} \underline{g}_{out} \cdot \hat{\underline{v}} d\Gamma. \end{aligned}$$

Note that the functions $\hat{\underline{v}},\ \hat{p}$ on (3.3) can be assumed to be independent of ε in what follows.

Assume that the problem (3.3) has a solution $\underline{v}, p$ that is infinitely differentiable with respect to ε:

$$\left\{\begin{array}{l} \underline{v} = \underline{v}_0 + \varepsilon\underline{v}_1 + \varepsilon^2\underline{v}_2 + \ldots \\ p = p_0 + \varepsilon p_1 + \varepsilon^2 p_2 + \ldots, \end{array}\right. \tag{3.4}$$

where $p_k \in \mathbb{L}^2, \underline{v}_k \in \mathbb{X},\ k \geq 1$. Using (2.2), (3.4), and small perturbation techniques we can derive the equations for $\underline{v}_k, p_k,\ k \geq 0$. In particular, $k = 0$ $\underline{v}_0$ and p_0 satisfy

$$\left\{\begin{array}{l} a(f_0; \underline{v}_0, \hat{\underline{v}}) = b(f_0; p_0, \hat{\underline{v}}) + G(f_0; \hat{\underline{v}}) \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\ b(f_0; \hat{\underline{p}}, \underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\ \underline{v}_0 = \underline{v}_f \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3}. \end{array}\right. \tag{3.5}$$

Correspondingly, we define

$$\mathcal{R}_{obs,0} := \mathcal{R}(f_0)\underline{v}_0. \tag{3.6}$$

For $k = 1$ the functions $\underline{v}_1$ and p_1 are the solution of the equations

$$
\left\{
\begin{array}{l}
a(f_0;\underline{v}_1,\hat{\underline{v}}) = b(f_0;p_1,\hat{\underline{v}}) + \frac{\partial}{\partial\varepsilon} b(f;p_0,\hat{\underline{v}})|_{\varepsilon=0} + \\
+ \frac{\partial}{\partial\varepsilon} G(f;\hat{\underline{v}})|_{\varepsilon=0} - \frac{\partial}{\partial\varepsilon} a(f;\underline{v}_0,\hat{\underline{v}})|_{\varepsilon=0} \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\
b(f_0;\hat{\underline{p}},\underline{v}_1) + \frac{\partial}{\partial\varepsilon} b(f;\hat{p},\underline{v}_0)|_{\varepsilon=0} = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\
\underline{v}_1 = 0 \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3},
\end{array}
\right. \tag{3.7}
$$

where

$$
\frac{\partial}{\partial\varepsilon} b(f;p_0,\hat{\underline{v}})|_{\varepsilon=0} := y b_f(f_1,p_0,\hat{\underline{v}}) = \int_{\Omega_1} \frac{f_1}{\beta} p_0 \mathcal{D}(f_0)\hat{\underline{v}}\,dxdy + \int_{\Omega_1} \frac{f_0}{\beta} p_0 \mathcal{D}_f(f_1,\hat{\underline{v}})dxdy,
$$

$$
\mathcal{D}_f(f_1,\hat{\underline{v}}) := \frac{\partial}{\partial\varepsilon}\mathcal{D}(f)\hat{\underline{v}}|_{\varepsilon=0} = -\left[y\left(\frac{f_{1,x}f_0 - f_{0,x}f_1}{f_0^2}\right)\frac{\partial\hat{u}}{\partial y} + \frac{\beta f_1}{f_0^2}\frac{\partial\hat{v}}{\partial y}\right],
$$

$$
\mathcal{D}_f(f_1,\underline{v}_0) := \frac{\partial}{\partial\varepsilon}\mathcal{D}(f)\underline{v}_0|_{\varepsilon=0} (:= \mathcal{D}_f f_1 \text{ in what follows}),
$$

$$
\frac{\partial}{\partial\varepsilon} G(f;\hat{\underline{v}})|_{\varepsilon=0} := G_1(f_1;\hat{\underline{v}}) = \int_{\Omega_1} \frac{f_1}{\beta} F\cdot\hat{\underline{v}}\,dxdy,
$$

$$
\begin{aligned}
\frac{\partial}{\partial\varepsilon} a(f;\underline{v}_0,\hat{\underline{v}})|_{\varepsilon=0} &:= a_f(f_1;\underline{v}_0,\hat{\underline{v}}) \\
&= \int_{\Omega_1} \frac{f_1\nu}{\beta}\left(\left(\frac{\partial\underline{v}_0}{\partial x} - \frac{y f_{0,x}}{f_0}\frac{\partial\underline{v}_0}{\partial y}\right)\cdot\left(\frac{\partial\hat{\underline{v}}}{\partial x} - \frac{y f_{0,x}}{f_0}\frac{\partial\hat{\underline{v}}}{\partial y}\right)\right. \\
&\qquad \left. + \frac{\beta^2}{f_0^2}\frac{\partial\underline{v}_0}{\partial y}\cdot\frac{\partial\hat{\underline{v}}}{\partial y}\right)dxdy + \\
&\quad - \int_{\Omega_1} \frac{f_0\nu}{\beta} y\frac{(f_{1,x}f_0 - f_{0,x}f_1)}{f_0^2}\left(\frac{\partial\underline{v}_0}{\partial y}\cdot\left(\frac{\partial\hat{\underline{v}}}{\partial x} - \frac{y f_{0,x}}{f_0}\frac{\partial\hat{\underline{v}}}{\partial y}\right)\right. \\
&\qquad \left. + \left(\frac{\partial\underline{v}_0}{\partial x} - \frac{y f_{0,x}}{f_0}\frac{\partial\underline{v}_0}{\partial y}\right)\cdot\frac{\partial\hat{\underline{v}}}{\partial y}\right)dxdy + \\
&\quad - \int_{\Omega_1} \frac{f_0\nu}{\beta}\left(\frac{2\beta^2 f_1}{f_0^3}\right)\frac{\partial\underline{v}_0}{\partial y}\cdot\frac{\partial\hat{\underline{v}}}{\partial y}dxdy.
\end{aligned}
$$

So the problem for $\underline{v}_1,\ p_1$ reads as follows: find $\underline{v}_1 \in \mathbb{X},\ p_1 \in \mathbb{L}^2(\Omega)$ s.t.

$$
\left\{
\begin{array}{l}
a(f_0;\underline{v}_1,\hat{\underline{v}}) - b(f_0;p_1,\hat{\underline{v}}) = b_f(f_1;p_0,\hat{\underline{v}}) + G_1(f_1;\hat{\underline{v}}) - a_f(f_1;\underline{v}_0,\hat{\underline{v}}) \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\
b(f_0;\hat{\underline{p}},\underline{v}_1) + b_f(f_1;\hat{p},\underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega),
\end{array}
\right. \tag{3.8}
$$

This is a generalized Stokes problem [7]. By a similar technique we can derive the equations for $\underline{v}_k,\ p_k$ with $k \geq 2$. However, we will not carry on this development further in this work.

4. The shape optimization problem. Suppose now that the function $f_1(x)$ in (3.7) is unknown as well as $\underline{v}_1,\ p_1$. To complete problem (3.7) we will have to formulate some "additional equations"; otherwise, we should require that f_1 be determined by minimizing a suitable "cost functional. . . ."

Problem (2.3) can be supplemented by the "additional equations":

$$
\mathcal{C}(f,\underline{v},p) = 0, \tag{4.1}
$$

where $\mathcal{C}$ is an operator (linear or nonlinear) defined on $\mathbb{H}_0^1(x_1,x_2)\times\mathbb{X}\times\mathbb{L}^2(\Omega)$. (We now consider $f\in\mathbb{H}_0^1$ for convenience.) We assume $\mathcal{C}$ to be smooth with respect to its variables $f,\underline{v},p$. Using the representations (2.2) and (3.4), we derive from (4.1) the following equation:

$$\mathcal{C}(f,\underline{v},p)=\mathcal{C}(f_0,\underline{v}_0,p_0)+\varepsilon\mathcal{C}_1(f_1,\underline{v}_1,p_1)+\mathcal{O}(\varepsilon^2)=0\quad\forall\,\varepsilon\in[-\varepsilon_0,\varepsilon_0], \tag{4.2}$$

where

$$\mathcal{C}_1(f_1,\underline{v}_1,p_1):=\frac{\partial\mathcal{C}}{\partial\varepsilon}(f,\underline{v},p)|_{\varepsilon=0}. \tag{4.3}$$

If we assume that the data of our problems are such that $\mathcal{C}(f_0,\underline{v}_0,p_0)=0$, then we can use

$$\mathcal{C}_1(f_1,\underline{v}_1,p_1)=0 \tag{4.4}$$

as an additional equation to complete (3.7). An alternative approach would consist of replacing the exact controllability equation (4.4) by the following minimization problem:

$$\inf_{f_1}\int_\Omega\frac{f_0}{\beta}|\mathcal{C}_1(f_1,\underline{v}_1,p_1)|^2dxdy, \tag{4.5}$$

where we assume that $\mathcal{C}_1$ has image in $\mathbb{L}^2(\Omega)$. Note that (4.5) is a weak statement of (4.4).

In the next sections we apply the approach described above for the completion of (3.7) and will use the following special choice of (4.1):

$$\mathcal{C}(f,\underline{v}):=((\nabla\times\underline{v})\circ\mathrm{T}_f^{-1})(x,y)-\mathcal{R}_{obs,\varepsilon}(x,y)\text{ in }\Omega_{wd}\subseteq\Omega, \tag{4.6}$$

where Ω_{wd} is a suitable subset of Ω in which we want our additional equation (or our "control") to take place. Moreover,

$$\mathcal{R}_{obs,\varepsilon}=\mathcal{R}_{obs,0}+\varepsilon\mathcal{R}_{obs,1}+\varepsilon^2\mathcal{R}_{obs,2}+\dots,\ \mathcal{R}_{obs,0}:=((\nabla\times\underline{v}_0)\circ\mathrm{T}_{f_0}^{-1}). \tag{4.7}$$

Then we have: $\mathcal{C}(f_0,\underline{v}_0)=0$, while (4.4) reads

$$\mathcal{C}(f_1,\underline{v}_1)=\mathcal{R}(f_0)\underline{v}_1+m_1\mathcal{R}_f f_1-\mathcal{R}_{obs,1}=0\text{ in }\Omega_{wd}, \tag{4.8}$$

where

$$\mathcal{R}(f_0)\underline{v}_1=(\nabla\times\underline{v}_1)\circ\mathrm{T}_{f_0}^{-1}(x,y)=\frac{\partial v_1}{\partial x}-\frac{yf_{0,x}}{f_0}\frac{\partial v_1}{\partial y}-\frac{\beta}{f_0}\frac{\partial u_1}{\partial y},$$

$$\mathcal{R}_f f_1:=\mathcal{R}_f(f_1,\underline{v}_0)=-y\frac{(f_{1,x}f_0-f_{0,x}f_1)}{f_0^2}\frac{\partial v_0}{\partial y}+\frac{\beta f_1}{f_0^2}\frac{\partial u_0}{\partial y}.$$

Therefore we have the problem: find $\underline{v}_1\in\mathbb{X},\ p_1\in\mathbb{L}^2(\Omega),\ f_1\in\mathbb{H}_0^1(x_1,x_2)$ s.t.

$$\begin{cases} a(f_0;\underline{v}_1,\hat{\underline{v}})=b(f_0;p_1,\hat{\underline{v}})+b_f(f_1;p_0,\hat{\underline{v}})+G_1(f_1;\hat{\underline{v}})-a_f(f_1;\underline{v}_0,\hat{\underline{v}})\quad\forall\,\hat{\underline{v}}\in\mathbb{X},\\ b(f_0;\underline{\hat{p}},\underline{v}_1)+b_f(f_1;\hat{p},\underline{v}_0)=0\quad\forall\,\hat{p}\in\mathbb{L}^2(\Omega),\\ \mathcal{R}(f_0)\underline{v}_1+m_1\mathcal{R}_f f_1-\mathcal{R}_{obs,1}=0\ in\ \Omega_{wd}, \end{cases} \tag{4.9}$$

where $\mathcal{R}_{obs,1}$ is a given function. Problem (4.9) is an "exact controllability problem." These problems have solutions in some particular cases only. For this reason we replace (4.9) by the following optimal control problem: find $\underline{v}_1 \in \mathbb{X},\ p_1 \in \mathbb{L}^2(\Omega)$, $f_1 \in \mathbb{H}^1_0(x_1, x_2)$ s.t.

(4.10)
$$\begin{cases} a(f_0;\underline{v}_1,\underline{\hat{v}}) - b(f_0;p_1,\underline{\hat{v}}) = b_f(f_1;p_0,\underline{\hat{v}}) + G_1(f_1;\underline{\hat{v}}) - a_f(f_1;\underline{v}_0,\underline{\hat{v}}) \quad \forall\ \underline{\hat{v}} \in \mathbb{X}, \\ b(f_0;\hat{p},\underline{v}_1) + b_f(f_1;\hat{p},\underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\ \inf_{f_1} = \frac{\alpha}{2}||f_1||^2_{\mathbb{H}^1_0(x_1,x_2)} + \gamma_1 J_1(f_1,\underline{v}_1), \end{cases}$$

where

$$J_1(f_1,\underline{v}_1) = \frac{1}{2}\int_\Omega m_{wd}\frac{f_0}{\beta}(\mathcal{R}(f_0)\underline{v}_1 + m_1\mathcal{R}_f f_1 - \mathcal{R}_{obs,1})^2 dxdy,$$

$\alpha = \text{const} \geq 0$ is a small regularization parameter, $\gamma_1 > 0$ is a weight coefficient, and m_{wd} is the characteristic function of Ω_{wd}.

Note that the third equation from (4.9) is considered in (4.10) in the least square sense; then (4.10) for $\alpha = 0$ provides the weak statement of problem (4.9). Otherwise the solution $v_1 = v_1(\alpha),\ p_1 = p_1(\alpha),\ f_1 = f_1(\alpha)$ of (4.10) represents an approximate (regularized) solution of (4.9).

We will also consider a generalized optimal control problem still given by (4.10); however, instead of J_1 we now use

$$J(f_1,\underline{v}_1,p_1) = \gamma_1 J_1(f_1,\underline{v}_1) + \gamma_2 J_2(f_1,\underline{v}_1,p_1).$$

Here $\gamma_2 = \text{const} \geq 0$ is a weight coefficient, while $J_2(f_1,\underline{v}_1,p_1)$ is an additional functional assumed to be quadratic. An example of $J_2(f_1,\underline{v}_1,p_1)$ follows.

Example 1.

(4.11)
$$J_2(f_1,\underline{v}_1,p_1) := J_2(\underline{v}_1,p_1) = \frac{1}{2}\left(||p_1 - p_{out,1}||^2_{\mathbb{L}^2(\Gamma_{out})} + \int_{\Gamma_{out}} |\underline{v}_1 - \underline{v}_{out,1}|^2 d\Gamma\right),$$

where $p_{out}, \underline{v}_{out}$ *are given.*

5. The variational equations of the optimal control problem. While considering (4.10) we can still consider the simple domain Ω of Figure 2.3(right). Another possibility consists of using the new variable transformation

$$T^{-1}_{f_0}(\underline{\tilde{x}}) = \underline{x},\ \underline{\tilde{x}} \in \Omega,\ \underline{x} \in \Omega_0, \tag{5.1}$$

which is the identity in $\tilde{\Omega}_2$, while $T^{-1}_{f_0}(\tilde{x},\tilde{y}) = (\tilde{x}, \frac{f_0(\tilde{x})}{\beta}\tilde{y})$ in $\tilde{\Omega}_1$. After applying (5.1) we will work in the "unperturbed" domain Ω_0 (see Figure 5.1) where the expressions for the bilinear forms in (4.10) become simpler. Let us use the variable transformation (5.1). Indeed problem (4.10) reads upon its reformulation in Ω_0: find $\underline{v} := \underline{v}_1,\ p := p_1, f := f_1$[1] s.t.

(5.2)
$$\begin{cases} a_0(\underline{v},\underline{\hat{v}}) - b_0(p,\underline{\hat{v}}) = b_f(f;p_0,\underline{\hat{v}}) + G_1(f;\underline{\hat{v}}) - a_f(f;\underline{v}_0,\underline{\hat{v}}) \quad \forall\ \hat{v} \in \mathbb{X} \\ b_0(\hat{p},\underline{v}) + b_f(f;\hat{p},\underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{L}^2(\Omega), \\ \inf_f = \frac{\alpha}{2}||f||^2_{\mathbb{H}^1_0(x_1,x_2)} + J(f,\underline{v},p), \end{cases}$$

[1]From now on we denote $\underline{v}_1 = \underline{v},\ p_1 = p,\ f_1 = f$; however, we should keep in mind that now $\underline{v}, p, f$ represents the "first corrections" of $\underline{v}_0,\ p_0,\ f_0$ on the unperturbed domain.

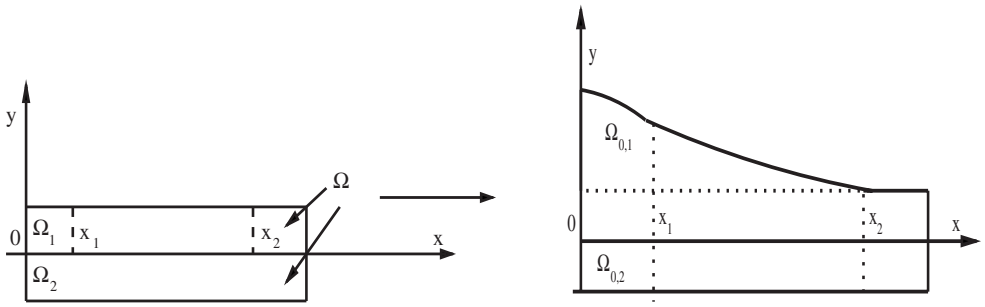


FIG. 5.1. *"Simple" domain* $\Omega \to \Omega_0$.

where

$$
\begin{aligned}
a_0(\underline{v},\underline{\hat{v}}) &= \int_{\Omega_0} \nu \left(\frac{\partial \underline{v}}{\partial x} \cdot \frac{\partial \underline{\hat{v}}}{\partial x} + \frac{\partial \underline{v}}{\partial y} \cdot \frac{\partial \underline{\hat{v}}}{\partial y} \right) dxdy, \\
b_0(p,\underline{\hat{v}}) &= \int_{\Omega_0} p \nabla \cdot \underline{\hat{v}} dxdy, \\
b_f(f,p_0,\underline{\hat{v}}) &= \int_{\Omega_{0,1}} p_0 \mathcal{D}_f(f,\underline{\hat{v}}) dxdy + \int_{\Omega_{0,1}} \frac{f}{f_0} p_0 \nabla \cdot \underline{\hat{v}} dxdy, \\
\mathcal{D}_f(f,\underline{\hat{v}}) &= -\left[y \left(\frac{f_x f_0 - f_{0,x} f}{f_0^2} \right) \frac{\partial \hat{u}}{\partial y} + \frac{f}{f_0} \frac{\partial \hat{v}}{\partial y} \right], \\
\mathcal{D}_f(f,v_0) &:= \mathcal{D}_f f, \\
G_1(f;\underline{\hat{v}}) &= \int_{\Omega_{0,1}} \frac{f}{f_0} \underline{F} \cdot \underline{\hat{v}} dxdy, \\
a_f(f;\underline{v}_0,\underline{\hat{v}}) &= \int_{\Omega_{0,1}} \frac{f\nu}{f_0} \nabla \underline{v}_0 \cdot \nabla \underline{\hat{v}} dxdy + \\
&\quad - \int_{\Omega_{0,1}} \nu y \frac{(f_x f_0 - f_{0,x} f)}{f_0^2} \left(\frac{\partial \underline{v}_0}{\partial y} \cdot \frac{\partial \underline{\hat{v}}}{\partial x} + \frac{\partial \underline{v}_0}{\partial x} \cdot \frac{\partial \underline{\hat{v}}}{\partial y} \right) dxdy + \\
&\quad - \int_{\Omega_{0,1}} \frac{2f\nu}{f_0} \frac{\partial \underline{v}_0}{\partial y} \cdot \frac{\partial \underline{\hat{v}}}{\partial y} dxdy, \\
J(f,\underline{v},p) &= \gamma_1 J_1(f,\underline{v}) + \gamma_2 J_2(f,\underline{v},p), \\
J_1(f,\underline{v}) &= \frac{1}{2} \int_{\Omega_0} m_{wd} |\nabla \times \underline{v} + m_1 \mathcal{R}_f f - \mathcal{R}_{obs,1}|^2 dxdy, \\
\mathcal{R}_f f &:= \mathcal{R}_f(f,\underline{v}_0) = -y \frac{(f_x f_0 - f_{0,x} f)}{f_0^2} \frac{\partial v_0}{\partial y} + \frac{f}{f_0} \frac{\partial u_0}{\partial y}, \\
\nabla \times \underline{v} &= \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \ \nabla \cdot \underline{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}
\end{aligned}
$$

and $J_2(f,\underline{v},p)$ are given by corresponding expressions. In order to derive the operator form of problem (5.2) we introduce the following real spaces:

$$
\begin{gathered}
\mathbb{X} \subseteq (\mathbb{L}^2(\Omega))^2 \subseteq \mathbb{X}^*, \mathbb{H}^p \subseteq \mathbb{L}^2(\Omega) \subseteq \mathbb{H}^{p*}, \\
\mathbb{H}_f \subseteq \mathbb{L}^2(x_1,x_2) \subseteq \mathbb{H}_f^*, \\
\mathbb{W} := \mathbb{X} \times \mathbb{H}^p \subseteq \mathbb{H}_0 := (\mathbb{L}^2(\Omega))^2 \times \mathbb{L}^2(\Omega) \subseteq \mathbb{W}^*.
\end{gathered}
$$

Let us reformulate (5.2) in the following form: find $\underline{\Phi} := (\underline{v}, p) \in \mathbb{W} = (\mathbb{X} \times \mathbb{H}^p)$, $f \in \mathbb{H}^f$, s.t

$$\begin{cases} \mathcal{L}(\underline{\Phi}, \hat{\underline{\Phi}}) = B(f, \hat{\underline{\Phi}}) \quad \forall\, \hat{\underline{\Phi}} = (\hat{\underline{v}}, \hat{p}) \in \mathbb{W}, \\ \inf_{f \in \mathbb{H}_f} = \frac{\alpha}{2} ||f||^2_{\mathbb{H}^1} + J(f, \underline{\Phi}), \end{cases} \tag{5.3}$$

where

$$\mathcal{L}(\underline{\Phi}, \hat{\underline{\Phi}}) := a_0(\underline{v}, \hat{\underline{v}}) - b_0(p, \hat{\underline{v}}) + b_0(\hat{p}, \underline{v}),$$

$$B(f, \hat{\underline{\Phi}}) := b_f(f, p_0, \hat{\underline{v}}) + G_1(f, \hat{\underline{v}}) - a_f(f, \underline{v}_0, \hat{\underline{v}}) - b_f(f, \hat{p}, \underline{v}_0).$$

Should $\underline{\Phi}$ be a solution of (5.3), then

$$\alpha(f, \hat{f})_{\mathbb{H}_f} + \langle J'_\Phi(f, \underline{\Phi}), \underline{\Phi}_{\hat{f}} \rangle + \langle J'_f(f, \underline{\Phi}), \hat{f} \rangle = 0, \tag{5.4}$$

for any $\hat{f} \in \mathbb{H}_f$ ($\hat{f}$ is the independent variation), where $\underline{\Phi}_{\hat{f}} \in \mathbb{W}$ satisfies the following equation:

$$\mathcal{L}(\underline{\Phi}_{\hat{f}}, \hat{\underline{\Phi}}) = B(\hat{f}, \hat{\underline{\Phi}}) \quad \forall\, \hat{\underline{\Phi}} \in \mathbb{W}. \tag{5.5}$$

In (5.4), $J'_\Phi = \frac{\partial J}{\partial \underline{\Phi}}$ and $J'_f = \frac{\partial J}{\partial f}$ are partial derivatives of J, while $\langle Q, \underline{\Phi} \rangle$ is the duality between $\mathbb{W}$ and $\mathbb{W}^*$ and $\langle g, f \rangle$ the duality between $\mathbb{H}_f$ and $\mathbb{H}^*_f$. Then we can rewrite (5.3) as a system of "optimality conditions":

$$\begin{cases} \mathcal{L}(\underline{\Phi}, \hat{\underline{\Phi}}) = B(f, \hat{\underline{\Phi}}) \quad \forall\, \hat{\underline{\Phi}} \in \mathbb{W}, \\ \alpha(f, \hat{f})_{\mathbb{H}^f} + \langle J'_\Phi(f, \underline{\Phi}), \underline{\Phi}_{\hat{f}} \rangle + \langle J'_f(f, \underline{\Phi}), \hat{f} \rangle = 0 \quad \forall\, \hat{f} \in \mathbb{H}_f. \end{cases} \tag{5.6}$$

The element $\underline{\Phi}_{\hat{f}}$ can be eliminated from (5.6) by introducing the adjoint problem: find $\underline{Q} := (\underline{q}, \sigma) \in \mathbb{W}$ s.t.

$$\mathcal{L}^*(\underline{Q}, \underline{\hat{W}}) := \mathcal{L}(\underline{\hat{W}}, \underline{Q}) = \langle J'_\Phi(f, \underline{\Phi}), \underline{\hat{W}} \rangle \quad \forall\, \underline{\hat{W}} \in \mathbb{W}. \tag{5.7}$$

Since $\underline{\Phi}_{\hat{f}} \in \mathbb{W}$ we can choose $\underline{\hat{W}} = \underline{\Phi}_{\hat{f}}$ in (5.7), yielding

$$\langle J'_\Phi(f, \underline{\Phi}), \underline{\Phi}_{\hat{f}} \rangle = \mathcal{L}(\underline{\Phi}_{\hat{f}}, \underline{Q}) = B(\hat{f}, \underline{Q}) \tag{5.8}$$

and the system of variational equations (5.6) now reads as follows:

$$\begin{cases} \mathcal{L}(\underline{\Phi}, \hat{\underline{\Phi}}) = B(f, \hat{\underline{\Phi}}) \quad \forall\, \hat{\underline{\Phi}} \in \mathbb{W}, \\ \mathcal{L}^*(\underline{Q}, \underline{\hat{W}}) = \langle J'_\Phi(f, \underline{\Phi}), \underline{\hat{W}} \rangle \quad \forall\, \underline{\hat{W}} \in \mathbb{W}, \\ \alpha(f, \hat{f})_{\mathbb{H}_f} + B(\hat{f}, \underline{Q}) + \langle J'_f(f, \underline{\Phi}), \hat{f} \rangle = 0 \quad \forall\, \hat{f} \in \mathbb{H}_f. \end{cases} \tag{5.9}$$

The first equation is the state equation. Let us define the following operators (see [1, 12, 13]):

$$L : \mathbb{W} \to \mathbb{W}^*,\ (L\underline{\Phi}, \hat{\underline{\Phi}})_{\mathbb{H}_0} := \mathcal{L}(\underline{\Phi}, \hat{\underline{\Phi}}) \quad \forall\, \underline{\Phi},\ \hat{\underline{\Phi}} \in \mathbb{W},$$

$$L^* : \mathbb{W} \to \mathbb{W}^*,\ (\underline{\hat{W}}, L^*\underline{Q})_{\mathbb{H}_0} = (L\underline{\hat{W}}, \underline{Q})_{\mathbb{H}_0} \quad \forall\, \underline{Q}, \underline{\hat{W}} \in \mathbb{W},$$

$$B : \mathbb{H}_f \to \mathbb{W}^*,\ (Bf, \underline{\Phi})_{\mathbb{H}_0} = B(f, \underline{\Phi}) \quad \forall\, f, \underline{\Phi},$$

$$\Lambda_w : \mathbb{W}^* \to \mathbb{W}^*,\ (\Lambda_w J_\Phi(f, \underline{\Phi}), \underline{\hat{W}})_{\mathbb{H}_0} := \langle J'_\Phi(f, \underline{\Phi}), \underline{\hat{W}} \rangle,$$

$$\Lambda_f : \mathbb{H}^*_f \to \mathbb{H}^*_f,\ (\Lambda_f J_f(f, \underline{\Phi}), \hat{f})_{\mathbb{L}^2(x_1, x_2)} = \langle J'_f(f, \underline{\Phi}), \hat{f} \rangle.$$

Now the system (5.9) can be written in operator form as follows:

$$\text{(5.10)}\qquad \begin{cases} L\underline{\Phi} = Bf \ \ (in\ \mathbb{W}^*), \\ L^*\underline{Q} = \Lambda_w J_\Phi(f,\underline{\Phi}) \ \ (in\ \mathbb{W}^*), \\ \alpha\Lambda_c f + B^*\underline{Q} + \Lambda_f J_f(f,\underline{\Phi}) = 0 \ \ (in\ (\mathbb{H}_f)^*), \end{cases}$$

where Λ_c is the extension to $\mathbb{H}_f$ of the following operator $\Lambda_{c,0}$:

$$\Lambda_{c,0} f := -f_{xx} + f, \quad \mathcal{D}(\Lambda_{c,0}) = \mathbb{H}^2 \cap \mathbb{H}_f.$$

Remark 1. The system (5.10) with a cost functional $J = \|C\underline{\Phi} - \underline{\Psi}\|^2_{\mathbb{H}_{ob}}$, where $C : \mathbb{W} \to \mathbb{H}_{ob}$ is a given operator and $\underline{\Psi} \in \mathbb{H}_{ob}$ a given observation function has been analyzed in [1]. In this case $J'_f = 0$ and $\Lambda_w J'_\Phi(f,\underline{\Phi}) = C^*(C\underline{\Phi} - \underline{\Psi})$.

6. Uniqueness and existence results. We analyze the particular case where the cost functional J is chosen as outlined by Example 1 of section 4.

Let J be the functional J_2 in Example 1. Then

$$\text{(6.1)}\qquad J(f,\underline{\Phi}) = J(f,\underline{v},p) = \frac{\gamma_1}{2}\int_{\Omega_0} m_{wd}|\nabla\times\underline{v} + m_1\mathcal{R}_f f - \mathcal{R}_{obs,1}|^2 d\Omega + \\ + \frac{\gamma_2}{2}\int_{\Gamma_{out}} \Big(|p - p_{out}|^2 + |\underline{v} - \underline{v}_{out}|^2\Big) d\Gamma.$$

To study the problem in this case we assume that $\Omega_{wd} = \Omega_0$ and we put here:

$$\mathbb{X} := \{\underline{v} : \underline{v} \in (\mathbb{H}^2(\Omega))^2, \underline{v} = 0 \text{ on } \Gamma_{in} \cup \Gamma_{w_1} \cup \Gamma_{w_3}\},$$

$$\mathbb{H}^p := \mathbb{H}^1(\Omega_0), \ \mathbb{H}_f := \mathbb{H}^2(x_1,x_2) \cap \mathbb{H}^1_0(x_1,x_2).$$

Here we consider $\mathbb{H}^2(\Omega_0)$ for velocity in order to use the uniqueness continuation theorem. The derivatives $J'_\Phi(f,\underline{\Phi})$ and $J'_f(f,\underline{\Phi})$ become

$$\langle J'_\Phi(f,\underline{\Phi}), \hat{\underline{\Phi}}\rangle = \gamma_1\int_{\Omega_0} m_{wd}(\nabla\times\underline{v} + m_1\mathcal{R}_f f - \mathcal{R}_{obs,1})\cdot(\nabla\times\hat{\underline{v}})d\Omega + \\ + \gamma_2\int_{\Gamma_{out}}(p - p_{out})\hat{p}d\Gamma + \gamma_2\int_{\Gamma_{out}}(\underline{v} - \underline{v}_{out})\cdot\hat{\underline{v}}d\Gamma,$$

$$\langle J'_f(f,\underline{\Phi}), \hat{f}\rangle = \gamma_1\int_{\Omega_0} m_{wd}(\nabla\times\underline{v} + m_1\mathcal{R}_f f - \mathcal{R}_{obs,1})\mathcal{R}_f\hat{f}d\Omega, \\ \forall\ \hat{\underline{\Phi}} = (\hat{\underline{v}},\hat{p})\ and\ \forall\ \hat{f}.$$

The system of variational equations (5.6) reads: find $\underline{v}_f \in \mathbb{X}, p_f \in \mathbb{H}^p$

$$\text{(6.2)}\qquad \begin{cases} a_0(\underline{v}_f,\hat{\underline{v}}) = b_0(p_f,\hat{\underline{v}}) + F(f,\hat{\underline{v}}) \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\ b_0(\hat{\underline{p}},\underline{v}_f) + b_f(f;\hat{p},\underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{H}^p(\Omega), \\ \alpha(f,\hat{f})_{\mathbb{H}_f} + \gamma_1\int_{\Omega_0} m_{wd}(\nabla\times\underline{v}_f + m_1\mathcal{R}_f f - \mathcal{R}_{obs,1})\cdot(\nabla\times\underline{v}_{\hat{f}} + m_1\mathcal{R}_f\hat{f})d\Omega + \\ +\gamma_2\int_{\Gamma_{out}}((p_f - p_{out})p_{\hat{f}} + (\underline{v}_f - \underline{v}_{out})\cdot\underline{v}_{\hat{f}})d\Gamma = 0 \quad \forall\ \hat{f} \in \mathbb{H}_f, \end{cases}$$

where

$$F(f,\hat{\underline{v}}) := b_f(f,p_0,\hat{\underline{v}}) + G_1(f,\hat{\underline{v}}) - a_f(f,\underline{v}_0,\hat{\underline{v}}),$$

and for every $\hat{f}$, $\underline{v}_{\hat{f}} = \underline{v}_f(\hat{f})$, $p_{\hat{f}} = p_f(\hat{f})$ denote the solution of the system given by the first and second equations in (6.2) corresponding to a right-hand side $f = \hat{f}$. The system (5.9) is: find $\underline{v}_f \in \mathbb{X}, p_f \in \mathbb{H}^p$

(6.3)
$$\begin{cases} a_0(\underline{v}_f, \underline{\hat{v}}) = b_0(p_f, \underline{\hat{v}}) + F(f, \underline{\hat{v}}) \quad \forall\ \underline{\hat{v}} \in \mathbb{X}, \\ b_0(\underline{\hat{p}}, \underline{v}_f) + b_f(f; \hat{p}, \underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{H}^p(\Omega), \\ a_0(\underline{\hat{q}}, \underline{q}) = -b_0(\sigma, \underline{\hat{q}}) + \gamma_1 \int_{\Omega_0} m_{wd}(\nabla \times \underline{v}_f + m_1 \mathcal{R}_f f - \mathcal{R}_{obs,1}) \cdot (\nabla \times \underline{\hat{q}}) d\Omega + \\ + \gamma_2 \int_{\Gamma_{out}} (\underline{v}_f - \underline{v}_{out}) \cdot \underline{\hat{q}} d\Gamma \quad \forall\ \underline{\hat{q}} \in \mathbb{X}, \\ -b_0(\hat{\sigma}, \underline{q}) = \gamma_2 \int_{\Gamma_{out}} (p_f - p_{out}) \hat{\sigma} d\Gamma \quad \forall\ \hat{\sigma} \in \mathbb{H}^p, \\ \alpha(f, \hat{f})_{\mathbb{H}_f} + F(\hat{f}, \underline{q}) - b_f(\hat{f}; \sigma, \underline{v}_0) + \\ + \gamma_1 \int_{\Omega_0} m_{wd}(\nabla \times \underline{v}_f + m_1 \mathcal{R}_f f - \mathcal{R}_{obs,1}) m_1 \mathcal{R}_f \hat{f} d\Omega = 0 \quad \forall\ \hat{f} \in \mathbb{H}_f. \end{cases}$$

In what follows we assume that the generalized Stokes problem (3.7) (see [7]) has a unique solution for any given $\underline{v}_0$, p_0 (the solution in the unperturbed domain Ω_0) and for each $f \in \mathbb{H}_f$ (see [8]).

Now consider the problem (6.3) for $\alpha > 0$.

PROPOSITION 6.1. *For any $\alpha > 0$, problem* (6.3) *has a unique solution for each given $\underline{\mathcal{R}}_{obs,1}$.*

Proof. Following [1], we formally invert L and L^* in the first and second equations of (5.10), then we substitute $\underline{\Phi}$, $\underline{Q}$ into the third equation and obtain the following weak problem: $f \in \mathbb{H}_f$ satisfies

$$\alpha(f, \hat{f})_{\mathbb{H}_f} + (Af, A\hat{f})_{\mathbb{L}^2(x_1,x_2)} = (G, A\hat{f})_{\mathbb{L}^2(x_1,x_2)} \quad \forall\ \hat{f} \in \mathbb{H}_f, \tag{6.4}$$

where A is a linear operator, which depends on previous operators from variational equations, while G will depend on the data more precisely from (6.2) we obtain:

$$\begin{aligned} (f, \hat{f})_{\mathbb{H}_f} &= (\Lambda_f f, \hat{f})_{\mathbb{L}^2(x_1,x_2)}, \\ (Af, A\hat{f})_{\mathbb{L}^2(x_1,x_2)} &= \gamma_1 \int_\Omega m_{wd}(\nabla \times \underline{v} + m_1 \mathcal{R}_f f) \cdot (\nabla \times \underline{v}_{\hat{f}} + m_1 \mathcal{R}_f \hat{f}) d\Omega + \\ &\quad + \gamma_2 \int_{\Gamma_{out}} (p p_{\hat{f}} + \underline{v} \cdot \underline{v}_{\hat{f}}) d\Gamma, \\ (G, A\hat{f})_{\mathbb{L}^2(x_1,x_2)} &= \gamma_1 \int_\Omega m_{wd} \mathcal{R}_{obs,1} \cdot (\nabla \times \underline{v}_{\hat{f}} + m_1 \mathcal{R}_f \hat{f}) d\Omega + \\ &\quad + \gamma_2 \int_{\Gamma_{out}} (p_{out} p_{\hat{f}} + \underline{v}_{out} \cdot \underline{v}_{\hat{f}}) d\Gamma, \end{aligned}$$

where $\underline{\Phi} = (\underline{v}, p) = L^{-1} B f$, $\underline{\Phi}_{\hat{f}} = (\underline{v}_{\hat{f}}, p_{\hat{f}}) = L^{-1} B \hat{f} \quad \forall\ \hat{f} \in \mathbb{H}_f$.

We see that if $\alpha > 0$, then the problem (6.4) has a unique solution which satisfies and: $\|f\|^2_{\mathbb{H}_f} \le \|G\|^2/(2\alpha) < \infty$. Correspondingly, we can construct $\underline{v}$, p, $\underline{q}$, σ, which jointly with f provides the unique solution of (6.3). □

Now consider the problem (6.3) with $\alpha = 0$.

PROPOSITION 6.2. *Assume that:* (i) *The solution of the generalized Stokes problem satisfies* $\left(\frac{\partial v_0}{\partial y}\right)^2 + \left(\frac{\partial u_0}{\partial y}\right)^2 > 0$ *at* $y = 0$, $x \in (x_1, x_2)$ (ii) *problem* (6.3) *has a solution. Then this solution is unique in the class* $(\mathbb{H}^2(\Omega))^2 \times \mathbb{H}^1(\Omega) \times \mathbb{W}^{1,\infty}(x_1, x_2)$.

Proof. Let $(\underline{v}_1, \dots, f_1)$ and $(\underline{v}_2, \dots, f_2)$ be two solutions of (6.3). Then for $\underline{v} = \underline{v}_1 - \underline{v}_2, \dots, f = f_1 - f_2$ from (6.2) we obtain:

$$\left\{\begin{array}{l} a_0(\underline{v}, \hat{\underline{v}}) = b_0(p, \hat{\underline{v}}) + F(f, \hat{\underline{v}}) \quad \forall\ \hat{\underline{v}} \in \mathbb{X}, \\ b_0(\hat{\underline{p}}, \underline{v}) + b_f(f; \hat{p}, \underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{H}^p(\Omega), \\ \nabla \times \underline{v} + m_1 \mathcal{R}_f f = 0 \text{ in } \Omega, \\ p = 0,\ \underline{v} = 0 \text{ on } \Gamma_{out}. \end{array}\right. \tag{6.5}$$

Consider the second and third equation from (6.5) in $\Omega_{2,0}$,

$$\nabla \cdot \underline{v} = 0,\ \nabla \times \underline{v} = 0 \text{ in } \Omega_{2,0}.$$

Then $\Delta \underline{v} = 0$ in $\Omega_{2,0}$. Considering $\hat{\underline{v}}$ with $\operatorname{supp}(\hat{\underline{v}}) \subseteq \Omega_{2,0}$ from the first equation of (6.5) we find $\nabla p = 0$, then $p = \text{const}$ in $\Omega_{2,0}$ and $-p \cdot \underline{n} + \nu \frac{\partial \underline{v}}{\partial \underline{n}} = 0$ on Γ_{out}. Since $p = 0$ on Γ_{out}, then $p = 0$ in $\Omega_{2,0}$ and $\nu \frac{\partial \underline{v}}{\partial \underline{n}} = 0$ on Γ_{out}, too. Consequently, $\underline{v}$ satisfies

$$\Delta \underline{v} = 0 \text{ in } \Omega_{2,0},\ \underline{v} = \nu \frac{\partial \underline{v}}{\partial \underline{n}} = 0 \text{ on } \Gamma_{out}.$$

This problem has only the trivial solution $\underline{v} = 0$ in $\Omega_{2,0}$. Since $\underline{v} \in (\mathbb{H}^2(\Omega))^2$, then

$$\underline{v} = \frac{\partial \underline{v}}{\partial \underline{n}} = 0 \text{ on } \Gamma_0 := \{(x, y) : y = 0, x_1 < x < x_2\}.$$

Now consider the second and third equations from (6.5) in $\Omega_{1,0}$:

$$\left\{\begin{array}{l} \nabla \cdot \underline{v} - \left[y\left(\frac{f_x f_0 - f_{0,x} f}{f_0^2}\right) \frac{\partial u_0}{\partial y} + \frac{f}{f_0} \frac{\partial v_0}{\partial y}\right] = 0 \text{ in } \Omega_{1,0}, \\ \nabla \times \underline{v} - \left[y\left(\frac{f_x f_0 - f_{0,x} f}{f_0^2}\right) \frac{\partial v_0}{\partial y} - \frac{f}{f_0} \frac{\partial u_0}{\partial y}\right] = 0 \text{ in } \Omega_{1,0}. \end{array}\right. \tag{6.6}$$

On Γ_0 we have

$$\nabla \cdot \underline{v} - \frac{f}{f_0} \frac{\partial v_0}{\partial y} = 0,\ \nabla \times \underline{v} + \frac{f}{f_0} \frac{\partial u_0}{\partial y} = 0,$$

$$|f(x)| = f_0 \frac{\left[(\nabla \cdot \underline{v})^2 + (\nabla \times \underline{v})^2\right]^{1/2}}{\left[\left(\frac{\partial v_0}{\partial y}\right)^2 + \left(\frac{\partial u_0}{\partial y}\right)^2\right]^{1/2}} \text{ on } \Gamma_0$$

(the dependence of the right-hand side on x and y is understood). Since $\underline{v} = \frac{\partial \underline{v}}{\partial \underline{n}} = \frac{\partial \underline{v}}{\partial y} = 0$ on Γ_0, then

$$\nabla \cdot \underline{v}|_{y=0} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}|_{y=0} = 0,\ \nabla \times \underline{v}|_{y=0} = \frac{\partial v}{\partial y} - \frac{\partial u}{\partial x}|_{y=0} = 0,\ x \in (x_1, x_2),$$

i.e., $f(x) = 0$. Therefore, $\underline{v} = 0,\ p = 0$, too. □

Let us once more note that if $\gamma_2 > 0$ and we introduce into considerations the cost functional J_2, then we overdeterminate the problem (4.4) for $\alpha = 0$ and the initial problem. Therefore in this case we usually have uniqueness results; however, not existence results generally. But in some physical problems the above overdeterminations (and the term $\alpha \|f\|^2_{\mathbb{H}_f}$ also) are reasonable and have a physical sense, therefore in these cases we can consider the optimal control problems like (4.5) as the problems to be independent of the initial problem (where we have only J_1). Here, we also have existence results and can name these optimal control problems as the "optimal shape design problems. . . ." Nevertheless, it is interesting to study solvability results of above variational problems as $\alpha = \gamma_2 = 0$.

7. Iterative processes. In this section we propose some iterative processes which are well suited for solving the variational equations obtained in the previous sections.

Consider the problem (5.10); if for $k = 0, 1, \ldots, f^{(k)}$ is known, then $f^{(k+1)}$ can be determinate by solving the following equations [1]:

$$\left\{\begin{array}{l} L\Phi^{(k)} = Bf^{(k)}, \\ L^*\underline{Q}^{(k)} = \Lambda_w J_\Phi(f^{(k)}, \underline{\Phi}^{(k)}), \\ \Lambda_c w^{(k)} = B^*\underline{Q}^{(k)} + \Lambda_f J_f(f^{(k)}, \underline{\Phi}^{(k)}), \\ f^{(k+1)} = f^{(k)} - \tau_k(\alpha f^{(k)} + w^{(k)}), \end{array}\right. \tag{7.1}$$

where $\{\tau_k\}$ is a family of parameters whose determination follows from the theory of extremal problems [32], the general theory of iterative processes [16, 25, 27], and the ill-posed problems theory [28, 30]. The step (7.1) would read as follows for the variational form (5.9) of problem (5.10):

$$\left\{\begin{array}{l} \mathcal{L}(\underline{\Phi}^{(k)}, \underline{\hat{\Phi}}) = B(f^{(k)}, \underline{\hat{\Phi}}) \quad \forall\ \underline{\hat{\Phi}} \in \mathbb{W}, \\ \mathcal{L}(\underline{\hat{W}}, \underline{Q}^{(k)}) = \langle J'_\Phi(f^{(k)}, \underline{\Phi}^{(k)}), \underline{\hat{W}}\rangle \quad \forall\ \underline{\hat{W}} \in \mathbb{W}, \\ (w^{(k)}, \hat{f})_{\mathbb{H}_f} = B(\hat{f}, \underline{Q}^{(k)}) + \langle J'_f(f^{(k)}, \underline{\Phi}^{(k)}), \hat{f}\rangle \quad \forall\ \hat{f} \in \mathbb{H}_f, \\ f^{(k+1)} = f^{(k)} - \tau_k(\alpha f^{(k)} + w^{(k)}). \end{array}\right. \tag{7.2}$$

Consider now problem (6.2) (with $\Omega_{wd} \subseteq \Omega$). The iterative process (7.2) for this problem reads as follows:

$$\left\{\begin{array}{l} a_0(\underline{v}^{(k)}, \underline{\hat{v}}) = b_0(p^{(k)}, \underline{\hat{v}}) + F(f^{(k)}, \underline{\hat{v}}) \quad \forall\ \underline{\hat{v}} \in \mathbb{X}, \\ b_0(\underline{\hat{p}}, \underline{v}^{(k)}) + b_f(f^{(k)}; \hat{p}, \underline{v}_0) = 0 \quad \forall\ \hat{p} \in \mathbb{H}^p(\Omega), \\ a_0(\underline{\hat{q}}, \underline{q}^{(k)}) = -b_0(\sigma^{(k)}, \underline{\hat{q}}) + \gamma_1 \int_{\Omega_0} m_{wd}(\nabla \times \underline{v}^{(k)} + m_1\mathcal{R}_f f^{(k)} - \mathcal{R}_{obs,1})\cdot \\ \cdot(\nabla \times \underline{\hat{q}})d\Omega + \gamma_2 \int_{\Gamma_{out}} (\underline{v}^{(k)} - \underline{v}_{out}) \cdot \underline{\hat{q}} d\Gamma \quad \forall\ \underline{\hat{q}} \in \mathbb{X}, \\ -b_0(\hat{\sigma}, \underline{q}^{(k)}) = \gamma_2 \int_{\Gamma_{out}} (p^{(k)} - p_{out})\hat{\sigma} d\Gamma \quad \forall\ \hat{\sigma} \in \mathbb{H}^p, \\ (w^{(k)}, \hat{f})_{\mathbb{H}_f} = F(\hat{f}, \underline{\hat{q}}) - b_f(\hat{f}; \hat{\sigma}^{(k)}, \underline{v}_0) + \\ + \gamma_1 \int_{\Omega_0} m_{wd}(\nabla \times \underline{v}^{(k)} + m_1\mathcal{R}_f f^{(k)} - \mathcal{R}_{obs,1}) m_1 \mathcal{R}_f \hat{f} d\Omega \quad \forall\ \hat{f} \in \mathbb{H}_f, \\ f^{(k+1)} = f^{(k)} - \tau_k(\alpha f^{(k)} + w^{(k)}), \quad k = 0, 1, \ldots . \end{array}\right. \tag{7.3}$$

Consider now the *finite dimensional case* in which the function $f, \{f^{(k)}\}, \hat{f}$ are all sought after in a finite-dimensional subspace $\mathbb{H}_{f,N} \subset \mathbb{H}_f$ of dimension $N < \infty$, whose basis $\varphi_i \in \mathbb{W}^{1,\infty}(x_1, x_2), i = 1, 2, \ldots, N$. Then the following theorem holds true.

THEOREM 7.1. *Assume that* $\Omega_{wd} = \Omega, (\frac{\partial v_0}{\partial y})^2 + (\frac{\partial u_0}{\partial y})^2 > 0$ *at* $y = 0, x \in (x_1, x_2)$. *Then:*

1. *the problem* (6.2) *is correctly solvable for* $\alpha \geq 0$ *and all* $N < \infty$*;*
2. *the iterative process* (7.3) *is convergent for any* $\alpha > 0$, $N < \infty$ *and provided the parameters* $\tau_k > 0$, $k = 0, 1, 2, \ldots$ *are small enough;*
3. *if* α *is sufficiently small while* k *is sufficiently large, then* $\{\underline{v}^{(k)}, p^{(k)}, f^{(k)}\}$ *can be taken as an approximate solution of problem* (6.2).

Proof.

1. The existence of the solution for $\alpha > 0$ has been proved early. Let us consider the case $\alpha = 0$. Since $f = \Sigma_{i=1}^{N} a_i \varphi_i \in \mathbb{H}_{f,N}$, then in the form (6.4) with $\alpha = 0$ we conclude that this equation is correctly solvable (because the problem (6.2) can only have a unique solution in $\mathbb{X} \times \mathbb{H}^p \times \mathbb{H}_f$; see Proposition 6.2). We assume the generalized Stokes problem to be correctly solvable for given $f \in \mathbb{H}_f$. Hence the problem (6.2) is correctly solvable too.
2. If $\alpha > 0$, then the bilinear form on the left-hand side of (6.4) is coercive and continuous with respect to the norm $\|f\|_{A,\alpha} = \sqrt{\alpha \|f\|^2_{\mathbb{H}_f} + \|Af\|^2_{\mathbb{L}^2(x_1,x_2)}}$. Then according to the general theory of iterative algorithm the process given by

$$(f^{(k+1)}, \hat{f})_{\mathbb{H}_f} = (f^{(k)}, \hat{f})_{\mathbb{H}_f} - \tau(\alpha(f^{(k)}, \hat{f})_{\mathbb{H}_f} + (Af^{(k)}, A\hat{f})_{\mathbb{L}^2(x_1,x_2)}) -$$

$$-(G, A\hat{f})_{\mathbb{L}^2(x_1,x_2)}, \ k = 0, 1, \ldots$$

is convergent for small $\tau > 0$. Hence the process (7.3) is convergent also and

$$\|\underline{v}^{(k)} - \underline{v}\|_{\mathbb{X}} + \|p^{(k)} - p\|_{\mathbb{H}^p} + \|f - f^{(k)}\|_{\mathbb{H}_f} \to 0, \ k \to \infty. \tag{7.4}$$

If $\Lambda_C^{-1} A^* A \in [C_1, C_2], C_1, C_2 = \text{const}$, and $\tau_k = 2/(2\alpha + C_1 + C_2)$, then (7.4) becomes (see [1]):

(7.5)

$$\|\underline{v}^{(k)} - \underline{v}\|_{\mathbb{X}} + \|p^{(k)} - p\|_{\mathbb{H}^p} + \|f - f^{(k)}\|_{\mathbb{H}_f} \le C\Big(\frac{C_2 - C_1}{2\alpha + C_1 + C_2}\Big)^k \to 0, \ k \to \infty.$$

3. Let $\underline{v}_0, \ p_0, \ f_0$ be a solution of (6.2) when $\alpha = 0$. According to the theory of ill-posed problem ([28] and [30]) we have: $\|f_0 - f_\alpha\|_{\mathbb{H}^p} \to 0$ as $\alpha \to +0$, where $(f_\alpha, \underline{v}_\alpha, p_\alpha)$ is the solution of (6.2) for $\alpha > 0$. Hence

$$\|\underline{v}_0 - \underline{v}_\alpha\|_{\mathbb{X}} + \|p_0 - p_\alpha\|_{\mathbb{H}^p} \to 0, \ as \ \alpha \to +0.$$

Then owing to (7.4) we conclude that the statement of Theorem 7.1 holds true also. □

The simple schemes in Figure 7.1 can be considered as examples of the above problems when $f \in \mathbb{H}_{f,N}$ for small N (the dimension of $\mathbb{H}_{f,N}$).

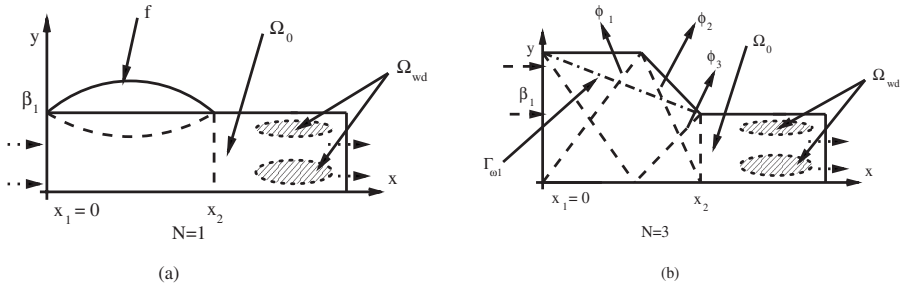


FIG. 7.1. *Domain Ω with N shape functions:* (a) $N = 1$, $f = \beta_1 + a\varphi_0(x)$, $\varphi_0 = x(x_2 - x)$; (b) $N = 3$, $f = \beta_1 + \Sigma_{i=1}^{3} a_i \varphi_i$.

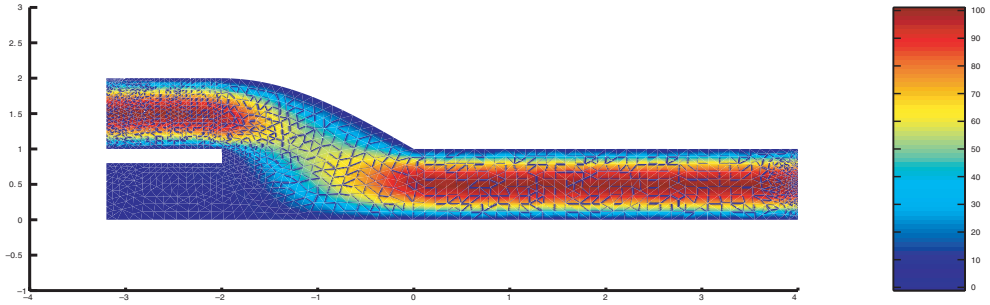


FIG. 8.1. *Idealized two-dimensional bypass configuration before optimal shape design process: Iso-velocity* $[cms^{-1}]$.

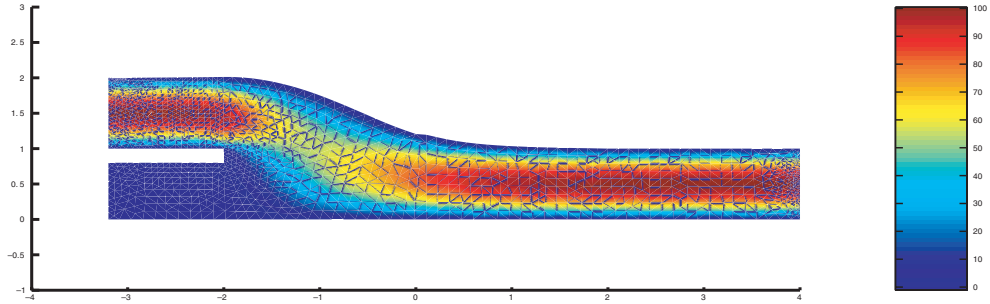


FIG. 8.2. *Bypass configuration at the end of shape optimization using first corrections: Iso-velocity* $[cms^{-1}]$.

8. Test problem and numerical results. To test our method we consider some test problems on simplified configurations. Numerical simulations have been carried out using *Bamg* [11], a Bi-dimensional Anisotropic Mesh Generator and *FreeFem*, a finite element Library developed at INRIA [10], the French National Institute for Research in Computer Science and Control, with the development of algorithms based on control theory and adjoint formulation for generalized Stokes problem. For application of finite element method to incompressible flow, see [9]. In this section we present numerical results using as cost functional the $\mathbb{L}^2$ norm of the vorticity in the downfield zone of the new incoming branch of the bypass.

Wall curvature was considered only in the zone of the incoming branch of the bypass where we set $f_0 = sin(x)$; in other parts we used piecewise constant function. The graft angle of the bypass incoming branch (which influences vorticity) is equal to zero (between the artery and the new incoming branch there isn't a relative angle).

Velocity values $\underline{v}_{in}$ at the inflow are chosen in such a way that the Reynolds number $Re = \frac{\bar{v} \cdot D}{\nu}$ has order 10^3. Blood kinematic viscosity $\nu = \frac{\mu}{\rho}$ is equal to $4 \cdot 10^{-6}\ m^2\ s^{-1}$, blood density $\rho = 1\ g\ cm^{-3}$ and dynamic viscosity $\mu = 4 \cdot 10^{-2}\ g\ cm^{-1} s^{-1}$; $\bar{v}$ is a mean inflow velocity related with $\underline{v}_{in}$, while D is the arterial diameter (3.5 mm) [23].

Figures 8.1–8.3 provide a preliminary account of numerical results and show how the shape of the bypass using generalized steady Stokes equations in an optimal control problem is smoothed out at the corner. Figure 8.1 refers to the original configuration; whereas Figure 8.2 refers to the configuration obtained after 25 iterations of the optimization algorithm (the vorticity has been reduced by about 30%).

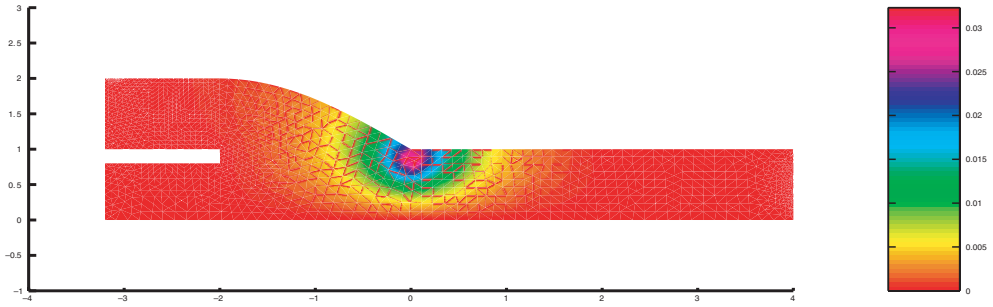


FIG. 8.3. *Adjoint solution* $\underline{q}$ *in Bypass configuration in the reference domain.*

9. Future developments. The development of tools for geometry reconstruction from medical data (medical imaging and other noninvasive means) and their integration with numerical simulation could provide improvements in disease diagnosis procedures.

In this study we have focused on the problem of determining the first corrections for the shape design of simplified two-dimensional bypass configurations.

Using the numerical method developed in this paper it is possible to realize the iterative process for solving initial nonlinear problems. For that it is sufficient to consider $f = f_0 + \varepsilon f_1$, where f_0 is the initial configuration and f_1 the computed first correction as the new f_0, then to calculate a new first correction and so on.

Optimal control and shape optimization applied to fully unsteady incompressible Stokes and Navier–Stokes equations and possibly the coupled fluid-structure problem and the setting of the problem in a three-dimensional geometry will provide more realistic design indications concerning surgical prosthesis realizations.

A further development will be devoted to build domain decomposition methods [26] based on optimal control approaches and efficient schemes for reduced-basis methodology approximations (see, for example, [20] and [21]) which could be more efficient for use in a repetitive design environment as optimal shape design methodology requires; see [29] for the state of the art of the problem.

Acknowledgments. Bernoulli Center of EPFL is acknowledged for the support of the authors during the special semester on the "Mathematical Modelling of the Cardiovascular System."

REFERENCES

[1] V. I. AGOSHKOV, *Optimal Control Approaches and Adjoint Equations in the Mathematical Physics Problems*, Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow, 2003.

[2] J. S. COLE, J. K. WATTERSON, AND M. J. G. O'REILLY, *Numerical investigation of the haemodynamics at a patched arterial bypass anastomosis*, Medical Engineering and Physics, 24 (2002), pp. 393–401.

[3] J. S. COLE, J. K. WATTERSON, AND M. J. G. O'REILLY, *Is there a haemodynamic advantage associated with cuffed arterial anastomoses?* Journal of Biomechanics, 35 (2002) pp. 1337–46.

[4] J. S. COLE, L. D. WIJESINGHE, J. K. WATTERSON, AND D. J. A. SCOTT, *Computational and experimental simulations of the haemodynamics at cuffed arterial bypass graft anastomoses*, in Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 216 (2002), pp. 135–143.

[5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Wiley, New York, 1966.

[6] Y. C. Fung, *Biodynamics: Circulation*, Springer-Verlag, New York, 1984.
[7] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations, Volume* I: *Linearized Steady Problem*, Springer-Verlag, New York, 1994.
[8] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[9] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Elements Method*, Wiley, New York, 2000.
[10] F. Hecht, O. Pironneau, and K. Ohtsuka, *Freefem++ Manual* 1.34, http://www.freefem.org, Paris, 2003.
[11] F. Hecht, *BAMG: Bidimensional Anisotropic Mesh Generator*, User Guide, INRIA, Rocquencourt, Cedex, France, 1998.
[12] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[13] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York, 1972.
[14] P. L. Lions, *Mathematical Topics in Fluid Mechanics. Volume* I: *Incompressible Models*, Oxford Science Publications, Clarendon Press, Oxford, 1996.
[15] F. Loth, S. A. Jones, D. P. Giddens, H. S. Bassiouny, C. K. Zarins, and S. Glagov, *Measurement of velocity and wall shear stress inside a PTE vascular graft model under steady flow conditions*, Journal of Biomechanical Engineering, 119 (1997), pp. 187–194.
[16] G. I. Marchuk, *Methods of Numerical Mathematics*, Nauka, Moscow, 1989.
[17] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Oxford University Press, Oxford, 2001.
[18] J. A. Moore, D. A. Steinman, S. Prakash, C. R. Ethier, and K. W. Johnston, *A numerical study of blood flow patterns in anatomically realistic and simplified end-to-side anastomoses*, ASME, J. Biomechanical Engineering, 121 (1999), pp. 265–272.
[19] K. Perktold, M. Hofer, G. Karner, W. Trubel, and H. Schima, *Computer Simulation of Vascular Fluid Dynamics and Mass Transport: Optimal Design of Arterial Bypass Anastomoses*, in Proceedings of ECCOMAS 98, K. Papailion et al., eds., Wiley, New York, 1998, pp. 484–489.
[20] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. T. Patera, and G. Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J. Fluids Engineering, 172 (2002), pp. 70–80.
[21] C. Prud'homme, D. Rovas, K. Veroy, and A. T. Patera, *A mathematical and computational framework for reliable real-time solution of parametrized partial differential equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 747–771.
[22] A. Quarteroni and G. Rozza, *Optimal control and shape optimization in aorto-coronaric bypass anastomoses*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1801–1823.
[23] A. Quarteroni and L. Formaggia, *Mathematical Modelling and Numerical Simulation of the Cardiovascular System*, in Modelling of Living Systems, Handbook of Numerical Analysis Series, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2004.
[24] A. Quarteroni, M. Tuveri, and A. Veneziani, *Computational vascular fluid dynamics: Problems, models, and methods*, Computing and Visualization in Science, 2 (2000), pp. 163–197.
[25] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.
[26] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.
[27] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*, Springer-Verlag, New York, 2000.
[28] A. N. Tikhonov and V. Ya. Arsenin, *Numerical Methods for Solving Ill-posed Problems*, Nauka, Moscow, 1974.
[29] G. Rozza, *Reduced Basis Methods for Elliptic Equations in sub-domains with A-Posteriori Error Bounds and Adaptivity*, EPFL-IACS report N.16-2004, Appl. Numer. Math., 55 (2005), pp. 403–424.
[30] G. M. Vainikko and A. Y. Veretennikov, *Iterative Procedures in Ill-posed Problems*, Nauka, Moscow, 1986.
[31] M. Van Dyke, *Perturbation Methods in Fluid Mechanics*, The Parabolic Press, Stanford, CA, 1975.
[32] F. P. Vasiliev, *Methods for Solving the Extremum Problems*, Nauka, Moscow, 1981.
[33] R. K. Zeytounian, *Theory and Applications of Viscous Fluid Flow*, Springer-Verlag, Berlin, 2004.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 385–411

CYCLIC DIGITAL NETS, HYPERPLANE NETS, AND MULTIVARIATE INTEGRATION IN SOBOLEV SPACES*

GOTTLIEB PIRSIC†, JOSEF DICK‡, AND FRIEDRICH PILLICHSHAMMER§

Abstract. Cyclic nets are a special case of digital nets and were recently introduced by Niederreiter. Here we present a construction algorithm for such nets, where we use the root mean square worst-case error of a randomly digitally shifted point set in a weighted Sobolev space as a selection criterion. This yields a feasible construction algorithm since for a cyclic net with q^m points (with fixed bijections and fixed ground field) there are q^m possible choices.

Our results here match the convergence rate and strong tractability results for polynomial lattice rules, hence providing us with an alternative construction algorithm. Further, we improve upon previous results by including constructions over arbitrary finite fields and an arbitrary choice of bijections.

Key words. cyclic digital net, weighted $\mathcal{L}_2$-discrepancy, quasi Monte Carlo algorithm, component-by-component algorithm

AMS subject classifications. 11K38, 65D30

DOI. 10.1137/050622638

1. Introduction.

In quasi Monte Carlo (QMC) one considers the approximation of an integral $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ by the average of $f(\boldsymbol{x}_h)$ for sample points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$. This approach might appear simple at first, but for high dimensions the question of how to choose a good point set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ becomes a truly challenging problem, with many questions yet to be answered (see, for example, [5, 13]). Generally speaking one wants the points $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ to be evenly spread over the unit cube. To assess the quality of a point set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, in other words, to measure the distribution properties of $P = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, one often uses a norm of the discrepancy function. The discrepancy function is given by

$$\Delta(P, \boldsymbol{z}) = \frac{A_N(P, [0, \boldsymbol{z}))}{N} - |\boldsymbol{z}|,$$

where $\boldsymbol{z} = (z_1, \ldots, z_s) \in [0,1]^s$, $A_N(P, [0, \boldsymbol{z}))$ is the number of points of P in $[0, \boldsymbol{z}) := \prod_{j=1}^{s} [0, z_j)$ and $|\boldsymbol{z}| = z_1 \cdots z_s$. By taking a norm of this function we obtain a quality measure of the point set P. In this paper we consider the so-called $\mathcal{L}_2$-discrepancy. In the classical case this corresponds to the 2-norm of the discrepancy function. Ever since the paper [20] by Sloan and Woźniakowski it has become popular to consider weighted discrepancies; specifically in our case this means we consider the *weighted*

*Received by the editors January 13, 2005; accepted for publication (in revised form) September 19, 2005; published electronically March 7, 2006.
http://www.siam.org/journals/sinum/44-1/62263.html

†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria (gottlieb.pirsic@oeaw.ac.at).

‡School of Mathematics, University of New South Wales, Sydney 2052, Australia (josi@maths.unsw.edu.au), and Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (josi@math.hkbu.edu.hk). This author is supported by the Australian Research Council under its Center of Excellence Program and by the Hong Kong Research Grants Council grant HKBU/2009/04P.

§Institut für Finanzmathematik, Universität Linz, Altenbergerstrasse 69, A-4040 Linz, Austria (friedrich.pillichshammer@jku.at). This author is supported by Austrian Research Foundation (FWF) project S9606 and project P17022-N12.

$\mathcal{L}_{2,\boldsymbol{\gamma}}$-*discrepancy*, which is given by

$$\mathcal{L}^2_{2,\boldsymbol{\gamma}}(P) = \sum_{\substack{u\subseteq\{1,\dots,s\}\\ u\neq\emptyset}} \prod_{j\in u} \gamma_j \int_{[0,1]^{|u|}} |\Delta(P,(\boldsymbol{z}_u,1))|^2 \,\mathrm{d}\boldsymbol{z}_u, \tag{1.1}$$

where $\boldsymbol{z}_u$ denotes the vector from $[0,1]^{|u|}$ containing the components of $\boldsymbol{z}$ whose indices are in u and $(\boldsymbol{z}_u,1)$ is the vector $\boldsymbol{z}$ from $[0,1]^s$ with all components whose indices are not in u replaced by 1. Here $\boldsymbol{\gamma} = (\gamma_1,\gamma_2,\dots)$ is a sequence of nonnegative real numbers γ_j, $j \geq 1$, the so-called weights. As is apparent from (1.1) the weights can be used to modify the importance of lower dimensional projections. We remark that in this paper we consider only product weights and not general weights.

In [20] it was also shown that the $\mathcal{L}_2$-discrepancy coincides with the worst-case error in certain weighted Sobolev spaces. Later on in the paper we prefer to state our results in terms of this worst-case error, as is usually done (for details see section 3). On the other hand it also seems enlightening to understand the geometrical meaning of this worst-case error, hence we also described the measure in terms of the $\mathcal{L}_2$-discrepancy rather than the worst-case error.

With this measure at hand we are able to compare the distribution properties of point sets in the unit cube. The point sets which we consider here are so-called (t,m,s)-nets and were introduced by Niederreiter [10]. A special construction scheme of such nets goes by the name of digital nets. In order to construct a digital net over some finite field, one needs to find $m \times m$ matrices $C_1,\dots,C_s$ with elements in the finite field $\mathbb{F}_q$. This provides us with a point set of q^m points. But the number of possible choices of generating matrices is q^{sm^2}, where s denotes the dimension. Hence using computer search to choose the best one is not feasible for a practically useful number of points. Niederreiter [12] also introduced a special subclass of digital nets, namely, polynomial lattices. Thereby the number of choices is reduced to q^{ms}. Moreover, it was shown in [4] that it is sufficient to search with a special algorithm (component-by-component), which further reduces the number of polynomial lattices considered to sq^m.

An alternative to polynomial lattices are cyclic nets, which were introduced by Niederreiter in 2004 [14]. In this paper we show how we can also construct cyclic nets using algorithms similar to [4]. Indeed, the construction algorithm used for cyclic nets matches the Korobov-type construction of polynomial lattice rules. Both of these have a search space of size at most q^m. Subsequently we also generalize the notion of cyclic nets whereby we succeed in introducing an analogue to the component-by-component algorithm of polynomial lattice rules. In the following we will call this construction scheme hyperplane nets.

The upper bounds presented here are comparable, although they are more general in the sense that we now also allow arbitrary finite fields. (Formerly we considered only finite fields of prime order.) In this situation one also needs bijections between the finite field and the digits $\{0,1,\dots,q-1\}$. The results presented here show that cyclic nets perform just as well as polynomial lattices, also achieving the best possible convergence rate and strong tractability results under appropriate conditions on the weights. Similar results have also been obtained for lattice rules; see [6, 18].

The paper is organized as follows. In the subsequent section we state the definition of (t,m,s)-nets, cyclic nets, hyperplane nets, and Walsh functions. Walsh functions are characters over the group of digital nets and are hence very useful for analyzing digital nets (see [3] for more information). Section 3 is concerned with construction

algorithms for cyclic nets and hyperplane nets. In that section we also prove upper bounds on the $\mathcal{L}_2$-discrepancy (or worst-case error), whereby the good performance of our construction algorithm is ensured. Finally, in an appendix we generalize the results in the appendix of [3], allowing now more general bijections which in turn allows us to obtain results for constructions of cyclic and hyperplane nets over arbitrary finite fields.

2. (t, m, s)-nets in base b. In this section we recall the definition of (digital) (t, m, s)-nets in base b and a special construction of such nets due to Niederreiter.

A detailed theory of (t, m, s)-nets was developed in [10]. (See also [11, Chapter 4] for a survey of this theory.) Those (t, m, s)-nets in a base b provide sets of b^m points in the s-dimensional unit cube $[0, 1)^s$, which are extremely well distributed if the quality parameter t is small.

DEFINITION 2.1 ((t, m, s)-nets). *Let $b \geq 2$, $s \geq 1$, and $0 \leq t \leq m$ be integers. Then a point set P consisting of b^m points in $[0, 1)^s$ forms a (t, m, s)-net in base b if every subinterval $J = \prod_{j=1}^{s} [a_j\, b^{-d_j}, (a_j + 1)\, b^{-d_j})$ of $[0, 1)^s$, with integers $d_j \geq 0$ and integers $0 \leq a_j < b^{d_j}$ for $1 \leq j \leq s$ and of volume b^{t-m}, contains exactly b^t points of P.*

In practice, all concrete constructions of (t, m, s)-nets in base b are based on the general construction scheme of digital nets. To avoid too many technical notions—and since we only deal with this case—in the following we restrict ourselves to digital nets defined over the finite field $\mathbb{F}_q$ of prime-power order q. For a more general definition (over arbitrary finite, commutative rings) see, for example, Niederreiter [11], Larcher [7], or Larcher, Niederreiter, and Schmid [8].

DEFINITION 2.2 (digital (t, m, s)-nets). *Let q be a prime-power and let $s \geq 1$ and $m \geq 1$ be integers. Let $C_1, \ldots, C_s$ be $m \times m$ matrices over $\mathbb{F}_q$. Now we construct q^m points in $[0, 1)^s$: for $0 \leq h \leq q^m - 1$ let $h = h_0 + h_1 q + \cdots + h_{m-1} q^{m-1}$ be the q-adic expansion of h. Consider an arbitrary but fixed bijection $\varphi : \{0, 1, \ldots, q - 1\} \longrightarrow \mathbb{F}_q$. Identify h with the vector $\vec{h} = (\varphi(h_0), \ldots, \varphi(h_{m-1}))^\top \in \mathbb{F}_q^m$, where $\top$ means the transpose of the vector. For $1 \leq j \leq s$ multiply the matrix C_j by $\vec{h}$, i.e.,*

$$C_j \vec{h} =: (y_{j,1}(h), \ldots, y_{j,m}(h))^\top \in \mathbb{F}_q^m,$$

and set

$$x_{h,j} := \frac{\varphi^{-1}(y_{j,1}(h))}{q} + \cdots + \frac{\varphi^{-1}(y_{j,m}(h))}{q^m}.$$

If for some integer t with $0 \leq t \leq m$ the point set consisting of the points

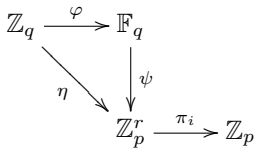
$$\boldsymbol{x}_h = (x_{h,1}, \ldots, x_{h,s})$$

for $0 \leq h < q^m$ is a (t, m, s)-net in base q, then it is called a digital (t, m, s)-net over $\mathbb{F}_q$ or, in brief, a digital net (over $\mathbb{F}_q$). The C_j are called its generating matrices.

Concerning the determination of the quality parameter t of digital nets we refer to Niederreiter [11, Theorem 4.28]; see also [17].

An essential tool for the investigation of digital nets is Walsh functions. A very general definition, corresponding to the most general construction of digital nets over finite rings, was given in [8]. There, Walsh functions over a finite abelian group G, using some bijection φ, were defined. Here we restrict ourselves to the case of $G = \mathbb{F}_{p^r}, p$ prime. We restate the definitions for this special case here for the sake of convenience.

DEFINITION 2.3 (Walsh functions). *Let $q = p^r$, p prime, $r \in \mathbb{N}_0$, and let $\mathbb{F}_q$ be the finite field with q elements. Let $\mathbb{Z}_q = \{0, 1, \dots, q-1\} \subset \mathbb{Z}$ with ring operations modulo q and let $\varphi : \mathbb{Z}_q \longrightarrow \mathbb{F}_q$ be a bijection such that $\varphi(0) = 0$, the neutral element of addition in $\mathbb{F}_q$. Moreover denote by ψ the isomorphism of additive groups $\psi : \mathbb{F}_q \longrightarrow \mathbb{Z}_p^r$ and define $\eta := \psi \circ \varphi$. For $1 \le i \le r$ denote by π_i the projection $\pi_i : \mathbb{Z}_p^r \longrightarrow \mathbb{Z}_p$, $\pi_i(x_1, \dots, x_r) = x_i$.*



Let now $k \in \mathbb{N}_0$ with base q representation $k = \kappa_0 + \kappa_1 q + \cdots + \kappa_{m-1} q^{m-1}$ where $\kappa_l \in \mathbb{Z}_q$ and let $x \in [0,1)$ with base q representation $x = x_1/q + x_2/q^2 + \cdots$. Then the kth Walsh function over the finite field $\mathbb{F}_q$ with respect to the bijection φ is defined by

$$ {}_{\mathbb{F}_q,\varphi}\mathrm{wal}_k(x) := \prod_{l=0}^{m-1} \prod_{i=1}^{r} \exp\left(2\pi \mathtt{i} \frac{(\pi_i \circ \eta)(\kappa_l)(\pi_i \circ \eta)(x_l)}{p}\right). $$

For convenience, in the rest of the paper we will omit the subscript and simply write wal_k if there is no ambiguity.

Multivariate Walsh functions are defined by multiplication of the univariate components, i.e., for $\boldsymbol{x} = (x_1, \dots, x_s) \in [0,1)^s, \boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s, s > 1$, we set

$$ \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^{s} \mathrm{wal}_{k_j}(x_j). $$

We summarize some important properties of Walsh functions over finite fields which will be used throughout the paper. The proofs of the subsequent results can be found, e.g., in [9, 15].

PROPOSITION 2.4. *Let p, q, $\mathbb{F}_q$, and φ be as in Definition 2.3. For x, y with q-adic representations $x = \sum_{i=w}^{\infty} x_i q^{-i}$ and $y = \sum_{i=w}^{\infty} y_i q^{-i}$, $w \in \mathbb{Z}$ (hence the following operations are also defined for integers), define $x \oplus_\varphi y := \sum_{i=w}^{\infty} z_i q^{-i}$ where $z_i := \varphi^{-1}(\varphi(x_i) + \varphi(y_i))$ and $\ominus_\varphi x := \sum_{i=w}^{\infty} v_i q^{-i}$ where $v_i := \varphi^{-1}(-\varphi(x_i))$. Further we set $x \ominus_\varphi y := x \oplus_\varphi (\ominus_\varphi y)$. For vectors $\boldsymbol{x}, \boldsymbol{y}$ we define the operations componentwise. Then we have the following:*

1. *For all $k, l \in \mathbb{N}_0$ and all $x, y \in [0,1)$ we have*

$$ \mathrm{wal}_k(x) \cdot \mathrm{wal}_l(x) = \mathrm{wal}_{k \oplus_\varphi l}(x), \quad \mathrm{wal}_k(x) \cdot \mathrm{wal}_k(y) = \mathrm{wal}_k(x \oplus_\varphi y) $$

and

$$ \mathrm{wal}_k(x) \cdot \overline{\mathrm{wal}_l(x)} = \mathrm{wal}_{k \ominus_\varphi l}(x), \quad \mathrm{wal}_k(x) \cdot \overline{\mathrm{wal}_k(y)} = \mathrm{wal}_k(x \ominus_\varphi y). $$

2. *We have*

$$ \sum_{k=0}^{q-1} \mathrm{wal}_l(k/q) = \begin{cases} 0 & \text{if } l \neq 0, \\ q & \text{if } l = 0. \end{cases} $$

3. *We have*

$$ \int_0^1 \mathrm{wal}_0(x)\,\mathrm{d}x = 1 \quad \text{and} \quad \int_0^1 \mathrm{wal}_k(x)\,\mathrm{d}x = 0 \text{ if } k > 0. $$

4. *For all* $\boldsymbol{k}, \boldsymbol{l} \in \mathbb{N}_0^s$ *we have the following orthogonality properties:*

$$\int_{[0,1)^s} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\overline{\mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} = \begin{cases} 1 & \text{if } \boldsymbol{k} = \boldsymbol{l}, \\ 0 & \text{otherwise.} \end{cases}$$

5. *For any* $f \in \mathcal{L}_2([0,1)^s)$ *and any* $\boldsymbol{\sigma} \in [0,1)^s$ *we have*

$$\int_{[0,1)^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^s} f(\boldsymbol{x} \oplus_\varphi \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{x}.$$

6. *For any integer* $s \geq 1$ *the system* $\{\mathrm{wal}_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{N}_0^s\}$ *is a complete orthonormal system in* $\mathcal{L}_2([0,1)^s)$.

Let $\mathbb{F}_q = \mathbb{Z}_p[\theta]$ such that $\{1, \theta, \ldots, \theta^{r-1}\}$ is a basis of $\mathbb{F}_q$ over $\mathbb{Z}_p$ as a vector space. Then the isomorphism ψ between $\mathbb{F}_q$ and $\mathbb{Z}_p^r$ shall be given by

$$\psi(x) = (x_1, \ldots, x_r)^\top \text{ for } x = \sum_{i=1}^{r} x_i \theta^{i-1}, x_i \in \mathbb{Z}_p.$$

Let ψ be extended to vectors over $\mathbb{F}_q$, i.e., such that for arbitrary m, vectors in $\mathbb{F}_q^m$ get mapped to vectors in $\mathbb{Z}_p^{rm}$.

Also let φ be extended to nonnegative integers by setting

$$\varphi(k) := (\varphi(\kappa_0), \ldots, \varphi(\kappa_{m-1}))^\top \text{ for } k = \sum_{i=0}^{m-1} \kappa_i q^i, k_i \in \{0, \ldots, q-1\}.$$

We will also use the concatenation $\eta(k) := \psi(\varphi(k))$. We have the following commutative diagram:

$$\begin{array}{ccc} \mathbb{Z}_{q^m} & \xrightarrow{\varphi} & \mathbb{F}_q^m \\ & \searrow^{\eta} & \downarrow{\scriptstyle \psi,\Psi} \\ & & \mathbb{Z}_p^{rm} \end{array}$$

We now define a map Ψ of the linear transformations over $\mathbb{F}_q$ into the linear transformations over $\mathbb{Z}_p$. Let the representation of the element θ^r in $\mathbb{F}_q$ be given by $\theta^r = \theta_0 + \theta_1\theta + \cdots + \theta_{r-1}\theta^{r-1}$, $\theta_i \in \mathbb{Z}_p, i = 0, \ldots, r-1$. By Θ we denote the matrix

$$\Theta := \begin{pmatrix} 0 & 0 & 0 & \cdots & \theta_0 \\ 1 & 0 & 0 & \cdots & \theta_1 \\ 0 & 1 & 0 & \cdots & \theta_2 \\ & \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & \theta_{r-1} \end{pmatrix}.$$

It is easy to see that Θ acts on a vector $(x_1, \ldots, x_r) \in \mathbb{Z}_p^r$ in the same way as the linear transformation $x \mapsto \theta x, x \in \mathbb{F}_q$ does on $x_1 + x_2\theta + \cdots + x_r\theta^{r-1}$, i.e., $\Theta\psi(x) = \psi(\theta x)$. If $\alpha = \sum_{i=0}^{r-1} a_i\theta^i$ is the representation of an arbitrary element, denote by $\Psi(\alpha)$ the matrix

$$\Psi(\alpha) := \sum_{i=0}^{r-1} a_i \Theta^i.$$

Clearly then $\Psi(\alpha)\psi(x) = \psi(\alpha x)$. By linearity the mapping Ψ can be extended to matrices by applying it to the matrix entries and letting the matrices run together, i.e., with some abuse of notation,

$$\Psi(A) := (\Psi(a_{i,j}))_{i,j} \in \mathbb{Z}_p^{rm_1 \times rm_2} \qquad \text{for } A = (a_{i,j})_{i,j} \in \mathbb{F}_q^{m_1 \times m_2}, a_{i,j} \in \mathbb{F}_q,$$

for arbitrary m_1, m_2. Again by linearity $\Psi(A)\psi(\boldsymbol{x}) = \psi(A\boldsymbol{x})$ holds as well (for $A \in \mathbb{F}_q^{m_1 \times m_2}, \boldsymbol{x} \in \mathbb{F}_q^{m_2}$, $m_1, m_2 \in \mathbb{N}$).

LEMMA 2.5. *Let $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{q^m-1}\}$ be a digital net over $\mathbb{F}_q$ with bijection φ, where $\varphi(0) = 0$, generated by the $m \times m$ matrices $C_1, \dots, C_s$ over $\mathbb{F}_q$, $m > 0$. Then for any vector $\boldsymbol{k} = (k_1, \dots, k_s)$ of nonnegative integers $0 \le k_1, \dots, k_s < q^m$ we have*

$$\sum_{h=0}^{q^m-1} {}_{\mathbb{F}_q,\varphi}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h) = \begin{cases} q^m & \text{if } C_1^\top \varphi(k_1) + \cdots + C_s^\top \varphi(k_s) = \boldsymbol{0}, \\ 0 & \text{else}, \end{cases}$$

where $\boldsymbol{0}$ is the zero vector in $\mathbb{F}_q^m$.

Proof. Denote by ω_p the pth root of unity, i.e., $\omega_p = \exp(2\pi\mathtt{i}/p)$. For each k_j, $1 \le j \le s$ let $\kappa_{j,l}$ denote the lth q-adic digit of k_j, i.e., $k_j = \kappa_{j,0} + \cdots + \kappa_{j,m-1}q^{m-1}$. For $0 \le h \le q^m - 1$ let $\boldsymbol{x}_h = (x_{h,1}, \dots, x_{h,s})$. Then we have

$$\begin{aligned}\Sigma &:= \sum_{h=0}^{q^m-1} {}_{\mathbb{F}_q,\varphi}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h) = \sum_{h=0}^{q^m-1} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \prod_{i=1}^{r} \omega_p^{(\pi_i \circ \eta)(\kappa_{j,l})(\pi_i \circ \eta)(x_{h,j,l})} \\ &= \sum_{h=0}^{q^m-1} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \omega_p^{\sum_{i=1}^{r}(\pi_i \circ \eta)(\kappa_{j,l})(\pi_i \circ \eta)(x_{h,j,l})} = \sum_{h=0}^{q^m-1} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \omega_p^{\langle \eta(\kappa_{j,l}), \eta(x_{h,j,l})\rangle},\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product. By the definition of digital nets we have

$$x_{h,j,l} = \varphi^{-1}(\langle \boldsymbol{c}_{j,l}^\top, \vec{h} \rangle),$$

where $\boldsymbol{c}_{j,l}$ denotes the lth row vector of the matrix C_j and where $\vec{h} = (\varphi(h_0), \dots, \varphi(h_{m-1}))^\top$ if $h = h_0 + \cdots + h_{m-1}q^{m-1}$. Therefore we obtain

$$\eta(x_{h,j,l}) = \psi \circ \varphi(\varphi^{-1}(\langle \boldsymbol{c}_{j,l}^\top, \vec{h} \rangle)) = \psi(\langle \boldsymbol{c}_{j,l}^\top, \vec{h} \rangle).$$

Since φ is a bijection we get

$$\begin{aligned}\Sigma &= \sum_{h=0}^{q^m-1} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \omega_p^{\langle \eta(\kappa_{j,l}), \psi(\langle \boldsymbol{c}_{j,l}^\top, \vec{h} \rangle)\rangle} = \sum_{\boldsymbol{h} \in \mathbb{F}_q^m} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \omega_p^{\langle \eta(\kappa_{j,l}), \psi(\langle \boldsymbol{c}_{j,l}^\top, \boldsymbol{h} \rangle)\rangle} \\ &= \sum_{\boldsymbol{h}' \in \mathbb{Z}_p^{rm}} \prod_{j=1}^{s} \prod_{l=0}^{m-1} \omega_p^{\langle \eta(\kappa_{j,l}), \Psi(\boldsymbol{c}_{j,l})\boldsymbol{h}' \rangle} = \sum_{\boldsymbol{h}' \in \mathbb{Z}_p^{rm}} \omega_p^{\langle \boldsymbol{h}', \sum_{j=1}^{s} \sum_{l=0}^{m-1} \Psi(\boldsymbol{c}_{j,l})^\top \eta(\kappa_{j,l}) \rangle}.\end{aligned}$$

We have

$$\sum_{l=0}^{m-1} \Psi(\boldsymbol{c}_{j,l})^\top \eta(\kappa_{j,l}) = \Psi(C_j^\top)\eta(k_j),$$

since, denoting by $c_{j,l,i}$ the components of $\boldsymbol{c}_{j,l}$,

$$\begin{pmatrix} \Psi(\boldsymbol{c}_{j,1})^\top \cdots \Psi(\boldsymbol{c}_{j,m})^\top \end{pmatrix} = \begin{pmatrix} \Psi(c_{j,1,1})^\top & \cdots & \Psi(c_{j,m,1})^\top \\ \vdots & \ddots & \vdots \\ \Psi(c_{j,1,m})^\top & \cdots & \Psi(c_{j,m,m})^\top \end{pmatrix} = \Psi(C_j^\top).$$

So we obtain

$$\Sigma = \sum_{\boldsymbol{h}' \in \mathbb{Z}_p^{rm}} \omega_p^{\langle \boldsymbol{h}', \sum_{j=1}^{s} \Psi(C_j^\top)\eta(k_j)\rangle}.$$

Bringing this into a form where we can evaluate the exponential sums, we get

$$\begin{aligned}\Sigma &= \prod_{l=0}^{m-1} \prod_{i=0}^{r-1} \sum_{h=0}^{p-1} \left(\omega_p^{(i\ m\ +\ l)th\ component\ of\ \left(\sum_{j=1}^{s} \Psi(C_j^\top)\eta(k_j)\right)}\right)^h \\ &= \begin{cases} q^m & \text{if } \Psi(C_1^\top)\eta(k_1) + \cdots + \Psi(C_s^\top)\eta(k_s) = \mathbf{0} \in \mathbb{Z}_p^{rm}, \\ 0 & \text{else,} \end{cases}\end{aligned}$$

and the lemma is proved by noting that

$$\Psi(C_1^\top)\eta(k_1) + \cdots + \Psi(C_s^\top)\eta(k_s) = \psi(C_1^\top\varphi(k_1) + \cdots + C_s^\top\varphi(k_s)) = \mathbf{0}$$

iff

$$C_1^\top\varphi(k_1) + \cdots + C_s^\top\varphi(k_s) = \mathbf{0},$$

since $\psi(0) = 0$. □

We next define cyclic digital nets following Niederreiter's article in [14]. (See this article of Niederreiter for more about the background of this notion and exact definitions of some terms not explained further in this paper.)

DEFINITION 2.6. *Let integers $m \geq 1, s \geq 2$ and a finite field $\mathbb{F}_q$ be given. Fix an element $\alpha \in \mathbb{F}_{q^m}$ and consider the set of polynomials*

$$\mathcal{P}_\alpha := \{f \in \mathcal{P}, f(\alpha) = 0\} \quad \subseteq \quad \mathcal{P} := \{f \in \mathbb{F}_{q^m}[x], \deg(f) < s\}.$$

For each $j = 1, \ldots, s$ choose an ordered basis $\mathcal{B}_j$ of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$ and define ϕ as the mapping

$$\phi : f(x) = \sum_{j=1}^{s} \gamma_j x^{j-1} \in \mathcal{P} \mapsto (\gamma_{1,1}, \ldots, \gamma_{1,m}, \ldots, \gamma_{s,1}, \ldots, \gamma_{s,m}) \in \mathbb{F}_q^{ms},$$

where $(\gamma_{j,1}, \ldots, \gamma_{j,m})$ is the coordinate vector of γ_j with respect to the chosen basis $\mathcal{B}_j$.

We denote by $\mathcal{C}_\alpha$ the orthogonal subspace in $\mathbb{F}_q^{ms}$ of the image $\mathcal{N}_\alpha := \phi(\mathcal{P}_\alpha)$. Let

$$C_\alpha = (C_1^\top \cdots C_s^\top) \in \mathbb{F}_q^{m \times sm}$$

be a matrix whose row space is $\mathcal{C}_\alpha$. Then the C_j are the generating matrices of a cyclic digital net with respect to $\mathcal{B}_1, \ldots, \mathcal{B}_s$ *and C_α is its overall generating matrix. We shall from now on assume a fixed choice of bases $\mathcal{B}_j$ and will therefore not explicitly mention them again.*

In the following we will again use the idea of employing linear representations (i.e., the mapping ψ), but with $\mathbb{F}_q$ in the role of $\mathbb{Z}_p$ and $\mathbb{F}_{q^m}$ in the role of $\mathbb{F}_q$. To be more precise, let $\mathbb{F}_{q^m} = \mathbb{F}_q[\omega]$, such that the powers of ω form a basis of $\mathbb{F}_{q^m}/\mathbb{F}_q$. Let $\omega^m = \beta_0 + \cdots + \beta_{m-1}\omega^{m-1}, \beta_l \in \mathbb{F}_q$, and P the matrix

$$P := \begin{pmatrix} 0 & 0 & \cdots & \beta_0 \\ 1 & 0 & \cdots & \beta_1 \\ \vdots & \ddots & 0 & \vdots \\ 0 & \cdots & 1 & \beta_{m-1} \end{pmatrix}.$$

Now, if we have the representation of α in $\mathbb{F}_{q^m}$ as $\alpha = \sum_{l=0}^{m-1} a_l \omega^l, a_l \in \mathbb{F}_q$, define

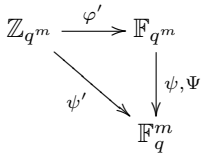
$$\psi(\alpha) := (a_0, \ldots, a_{m-1}) \in \mathbb{F}_q^m, \qquad \Psi(\alpha) := \sum_{l=0}^{m-1} a_l P^l \in \mathbb{F}_q^{m \times m}.$$

Note that for any $\alpha, x \in \mathbb{F}_{q^m} \setminus \{0\}$ we have $\Psi(\alpha)\psi(x) = \psi(\alpha x) \neq \mathbf{0} \in \mathbb{F}_q^m$ as $\alpha x \neq 0 \in \mathbb{F}_{q^m}$. Hence it follows that for any $\alpha \in \mathbb{F}_{q^m} \setminus \{0\}$ we have that the matrix $\Psi(\alpha)$ is regular.

Furthermore, for $k = \sum_{l=0}^{m-1} \kappa_l q^l$, let

$$\varphi'(k) := \sum_{l=0}^{m-1} \varphi(\kappa_l)\omega^l, \qquad \psi'(k) := \psi(\varphi'(k))$$

and define all extensions to vectors and matrices as above. We have the following commutative diagram:



Note that we have $\psi' = \varphi$.

Using similar methods as in Lemma 2.5 we can give the generating matrices for $\mathcal{C}_\alpha$ in the following form.

THEOREM 2.7. *Let $m, s, \mathbb{F}_q$ and $\alpha \in \mathbb{F}_{q^m} = \mathbb{F}_q[\omega]$ be given and define s matrices $B_j = (\psi(b_{j,1}), \ldots, \psi(b_{j,m}))^{-1}$, where the $b_{j,l}$ constitute the chosen basis $\mathcal{B}_j$. Then the generating matrices of the net are given by $C_j = (\Psi(\alpha^{j-1})B_j)^\top = (\Psi(\alpha)^{j-1}B_j)^\top$, $j = 1, \ldots, s$. Furthermore, it follows that C_j is regular for $j = 1, \ldots, s$.*

Proof. Let ϕ_1 be the (additive) isomorphism between $\mathcal{P} \subset \mathbb{F}_{q^m}[x]$ and $\mathbb{F}_{q^m}^s$. To arrive at the ϕ of Definition 2.6 we have to account for the choice of arbitrary bases $\mathcal{B}_j$. We do this by multiplying with the transformation matrix B^{-1}, where B is a square, block diagonal matrix with the matrices B_j of the statement of the theorem in its diagonal. Then $\phi(f) = B^{-1}\psi(\phi_1(f)), f \in \mathcal{P}$. We summarize these relations in the following diagrams:

Our first goal is to describe $\mathcal{N}_\alpha^\circ := \psi(\phi_1(\mathcal{P}_\alpha))$. Clearly, $\phi_1(\mathcal{P}_\alpha)$ is the space of all vectors orthogonal to $(1, \alpha, \ldots, \alpha^{s-1})^\top$. So $\boldsymbol{x} \in \phi_1(\mathcal{P}_\alpha)$ iff

$$\begin{aligned} 0 = (1, \alpha, \ldots, \alpha^{s-1})\boldsymbol{x} \iff \mathbf{0} &= \psi((1, \alpha, \ldots, \alpha^{s-1})\boldsymbol{x}) \\ &= \Psi((1, \alpha, \ldots, \alpha^{s-1}))\psi(\boldsymbol{x}), \end{aligned}$$

hence $\mathcal{N}_\alpha^\circ$ is the orthogonal space to the row space of

$$C_\alpha^\circ := \Psi((1, \alpha, \ldots, \alpha^{s-1})) = (\Psi(1), \Psi(\alpha), \ldots, \Psi(\alpha^{s-1})).$$

If the $\mathcal{B}_j$ are again taken into account, we have that $\mathcal{N}_\alpha$ is the image of $\mathcal{N}_\alpha^\circ$ under the automorphism $\boldsymbol{x} \mapsto B^{-1}\boldsymbol{x}$; accordingly its orthogonal space is the image under $\boldsymbol{x} \mapsto \boldsymbol{x}B$. Thus $C_\alpha := C_\alpha^\circ B$ is the overall generating matrix of the cyclic digital net (i.e., its row space is said to be orthogonal space) and $C_j := (\Psi(\alpha^{j-1})B_j)^\top$ are its generating matrices by the duality theory of digital nets. By the considerations before Lemma 2.5, Ψ is a ring homomorphism, so $\Psi(\alpha^j) = \Psi(\alpha)^j$.

In order to show that the matrices C_j are regular, recall that for any $\alpha \in \mathbb{F}_{q^m} \setminus \{0\}$ the matrix $\Psi(\alpha)$ is regular and as B_j is regular as well, it follows that C_j has to be regular. □

Remark 2.8. Note that every digital net with regular generating matrices C_j is cyclic with respect to some choice of bases $\mathcal{B}_j$. However, the focus in this paper lies on the class of all cyclic nets, i.e., where α runs through all elements in $\mathbb{F}_{q^m} \setminus \{0\}$, for fixed bases $\mathcal{B}_j$ and we show that there is at least one good cyclic net (i.e., good choice of $\alpha \in \mathbb{F}_{q^m} \setminus \{0\}$) for each fixed choice of $\mathcal{B}_j$. Those classes of cyclic nets which we use in our search algorithms do depend on the choice of bases, but once chosen the bases remain fixed throughout the search algorithm. In particular the search space of our algorithm will still be of size q^m for any given choice of bases.

Remark 2.9. It can be shown with a little calculation that Korobov polynomial lattice rules can be constructed (up to reordering of points) as cyclic nets. There, we have all B_j equal to the identity matrix. Note that with a suitably modified definition of cyclic nets (namely, if we consider arbitrary polynomial residue class rings) this also works for composite moduli f.

With similar little difficulty, Schmid's constacyclic shift-nets can be realized as cyclic nets, using the construction for $\mathbb{F}_{q^m} = \mathbb{F}_q[\theta]$, where $\theta^m = k$, if $f(x) = x^m - k$ is irreducible, and k is the factor for the shifted elements. (Again, we could also extend the definition of cyclic nets to include arbitrary polynomial residue class rings instead of only $\mathbb{F}_{q^m}/\mathbb{F}_q$.) Here, all B_j are constant and equal to the first, unshifted matrix C_1, and α is always chosen equal to θ.

These two relations are considered in detail in [16].

In view of Theorem 2.7 we propose the following generalization of the cyclic net construction.

DEFINITION 2.10. *Given a finite field $\mathbb{F}_q$, $\mathbb{F}_{q^m} = \mathbb{F}_q[\omega]$ as above, choose s elements $\alpha_1, \ldots, \alpha_s \in \mathbb{F}_{q^m}$ and regular matrices $B_j \in \mathbb{F}_q^{m\times m}$ and let the generating matrices of a digital net be defined by the matrices $C_j = (\Psi(\alpha_j)B_j)^\top$. A digital net constructed in this manner shall be called a* hyperplane net with respect to $\mathcal{B}_1, \ldots, \mathcal{B}_s$, *where by $\mathcal{B}_j$ we denote the ordered bases corresponding to the matrices B_j as in Theorem* 2.7. *Again, we shall from now on assume a fixed choice of bases $\mathcal{B}_j$ and will therefore no longer explicitly mention them.*

Remark 2.11. Note that the generating matrices C_j of the hyperplane net are regular provided that $\alpha_j \neq 0 \in \mathbb{F}_{q^m}$. For $\alpha_j = 0$ we obtain that $C_j = \mathbf{0} \in \mathbb{F}_q^{m\times m}$, the matrix consisting only of the neutral element with respect to addition in $\mathbb{F}_q$.

As a consequence, by Remark 2.8 a hyperplane net with regular generating matrices can also be considered as a cyclic net for some choice of bases $\mathcal{B}_j$. But the class of all hyperplane nets with fixed bases $\mathcal{B}_j$ is a proper superclass of the class of all cyclic nets with the same fixed bases $\mathcal{B}_j$. Hence the search space over all hyperplane nets is larger than the search space over all cyclic nets, as for the search we fix the bases in advance.

The name of the generalized construction is motivated by the following corollary of Lemma 2.5.

COROLLARY 2.12. *Let $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{q^m-1}\}$ be a digital net over $\mathbb{F}_q$ generated by the $m \times m$ matrices $C_1, \dots, C_s$ over $\mathbb{F}_q$, $m > 0$, as given in Definition 2.10. Then for any vector $0 \le k_1, \dots, k_s < q^m$ of nonnegative integers we have*

$$\sum_{h=0}^{q^m-1} {}_{\mathbb{F}_q,\varphi}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h) = \begin{cases} q^m & \text{if } \alpha_1\varphi'(\tau_1(k_1)) + \cdots + \alpha_s\varphi'(\tau_s(k_s)) = 0, \\ 0 & \text{else,} \end{cases}$$

with permutations $\tau_j(k) = \psi'^{-1}(B_j\psi'(k))$, and B_j as in Theorem 2.7.

Proof. By Definition 2.10 and Theorem 2.7, the generating matrices of the net are $C_j = (\Psi(\alpha_j)B_j)^\top$, so by Lemma 2.5 the sum equals q^m, iff (note that $\psi' = \varphi$)

$$\begin{aligned} \sum_{j=1}^{s} C_j^\top \psi'(k_j) &= \sum_{j=1}^{s} \Psi(\alpha_j)B_j\psi'(k_j) \\ = \sum_{j=1}^{s} \Psi(\alpha_j)\psi'(\tau_j(k_j)) = \mathbf{0} &\iff \sum_{j=1}^{s} \psi(\alpha_j\varphi'(\tau_j(k_j))) = \mathbf{0} \\ \iff \psi\left(\sum_{j=1}^{s} \alpha_j\varphi'(\tau_j(k_j))\right) = \mathbf{0} &\iff \sum_{j=1}^{s} \alpha_j\varphi'(\tau_j(k_j)) = 0, \end{aligned}$$

and vanishes otherwise, so the corollary follows. ∎

Remark 2.13. With this corollary it is not difficult to show that an equivalent definition in the spirit of Definition 2.6 exists: replace the $\mathcal{P}$ of Definition 2.6 by the space of linear forms

$$\mathcal{P} = \{f(x_1, \dots, x_s) = x_1\gamma_1 + \cdots + x_s\gamma_s, x_j \in \mathbb{F}_{q^m}\} \subset \mathbb{F}_{q^m}[x_1, \dots, x_s]$$

and $\mathcal{P}_\alpha$ by $\mathcal{P}_{\boldsymbol{\alpha}} = \{f \in \mathcal{P}, f(\alpha_1, \dots, \alpha_s) = 0\}$.

Remark 2.14. Similar to cyclic nets, polynomial lattice rules can be regarded as a special case ($B_j = I$) of hyperplane nets. Note that we shall again need to modify the definition of hyperplane nets to account for composite moduli f. Together with Korobov polynomial lattice rules and cyclic digital nets etc. we get a hierarchy of nets that is dealt with specifically and in more detail in [16].

3. Multivariate integration in weighted Sobolev spaces. In this section we consider multivariate integration in the weighted Sobolev space $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ induced by the reproducing kernel given by (see [2, 6, 18, 19])

$$K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s}(1 + \gamma_j\varrho_{w_j}(x_j, y_j)),$$

where $\boldsymbol{w} = (w_1, \dots, w_s) \in [0,1]^s$ and

$$\begin{aligned} \varrho_w(x,y) &= \frac{|x-w| + |y-w| - |x-y|}{2} \\ &= \begin{cases} \min(|x-w|, |y-w|) & \text{if } (x-w)(y-w) \ge 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The inner product in $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ is given by

$$\langle f, g\rangle_{H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}} := f(\boldsymbol{w})g(\boldsymbol{w}) + \sum_{\substack{u\subseteq\{1,\ldots,s\}\\ u\neq\emptyset}} \prod_{j\in u} \gamma_j^{-1} \int_{[0,1)^{|u|}} \frac{\partial^{|u|} f}{\partial \boldsymbol{x}_u}(\boldsymbol{x}_u, \boldsymbol{w}) \frac{\partial^{|u|} g}{\partial \boldsymbol{x}_u}(\boldsymbol{x}_u, \boldsymbol{w})\,\mathrm{d}\boldsymbol{x}_u,$$

where for $\boldsymbol{x} = (x_1, \ldots, x_s)$ and $u \subseteq \{1, \ldots, s\}$, $u \neq \emptyset$, we use the notation $\boldsymbol{x}_u = (x_j)_{j\in u}$ and $(\boldsymbol{x}_u, \boldsymbol{w})$ denotes the s-dimensional vector whose jth component is x_j if $j \in u$ and w_j if $j \notin u$. The Sobolev space $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ can also be defined as the set of all square integrable functions where the norm induced by the above inner product is finite.

Choose a prime-power base $q = p^r$ and let $x = \frac{x_1}{q} + \frac{x_2}{q^2} + \cdots$ and $\sigma = \frac{\sigma_1}{q} + \frac{\sigma_2}{q^2} + \cdots$ be the base q representation of x and σ. Further choose a bijection $\varphi : \{0, 1, \ldots, q-1\} \longrightarrow \mathbb{F}_q$ with $\varphi(0) = 0$. Then the digitally shifted point (with respect to the bijection φ) $y = x \oplus_\varphi \sigma$ is given by $y = \frac{y_1}{q} + \frac{y_2}{q^2} + \cdots$, where $y_i = \varphi^{-1}(\varphi(x_i) + \varphi(\sigma_i))$. For vectors $\boldsymbol{x}$ and $\boldsymbol{\sigma}$ we define the digitally shifted point $\boldsymbol{x} \oplus_\varphi \boldsymbol{\sigma}$ componentwise. Obviously, the shift depends on the base q as well as on the bijection φ.

For a point set $P_N = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ and a $\boldsymbol{\sigma} \in [0,1)^s$ let $P_{N,\varphi,\boldsymbol{\sigma}} = \{\boldsymbol{x}_0 \oplus_\varphi \boldsymbol{\sigma}, \ldots, \boldsymbol{x}_{N-1} \oplus_\varphi \boldsymbol{\sigma}\}$ be the digitally shifted point set.

We recall that the *worst-case error* $e(P_N, K)$ for the integration of functions f from a reproducing kernel Hilbert space H with reproducing kernel K by means of a QMC-algorithm

$$Q_{N,s}(P_N, f) = \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n)$$

using a point set $P_N = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ is defined as

$$e(P_N, K) := \sup_{f\in H, \|f\|\leq 1} \left| \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - Q_{N,s}(P_N, f) \right|.$$

(In [20, Theorem 1] it is shown that for $\boldsymbol{w} = (1, \ldots, 1)$ we have $e(P_N, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) = \mathcal{L}_{2,\boldsymbol{\gamma}}(P_N)$, the weighted $\mathcal{L}_{2,\boldsymbol{\gamma}}$-discrepancy of the point set P_N; see (1.1).)

Let the mean square worst-case error $\widehat{e}^2(P_N, K)$ be given by

$$\mathbb{E}[e^2(P_{N,\varphi,\boldsymbol{\sigma}}, K)] = \int_{[0,1)^s} e^2(P_{N,\varphi,\boldsymbol{\sigma}}, K)\,\mathrm{d}\boldsymbol{\sigma}.$$

Then we have $\widehat{e}^2(P_N, K) = e^2(P_N, K_{\mathrm{ds}})$, where

$$K_{\mathrm{ds}}(\boldsymbol{x}, \boldsymbol{y}) := \int_{[0,1)^s} K(\boldsymbol{x} \oplus_\varphi \boldsymbol{\sigma}, \boldsymbol{y} \oplus_\varphi \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{\sigma}$$

is the so-called shift invariant kernel of the kernel K. The proof of this result is similar to that of [3, Theorem 7].

In Appendix A of this paper it is shown that the shift invariant kernel $K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}$ for the reproducing kernel $K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ is given by

$$K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^s} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{k}) \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y})},$$

where $\boldsymbol{w} = (w_1, \ldots, w_s) \in [0,1]^s$ and $\widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{k}) = \prod_{j=1}^s \widehat{r}_q(w_j, \gamma_j, k_j)$, where

$$\widehat{r}_q(w,\gamma,k) := \begin{cases} 1 + \gamma(w^2 - w + \frac{1}{3}) & \text{if } k = 0, \\ -\frac{\gamma}{2}\left(\frac{1}{3q^{2a}} + \frac{2}{q^{2a}}\Re\left(\sum_{\substack{u,v=0 \\ v \ge u}}^{q-1} \frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u \ominus_\varphi v}{q}\right)}{q}\right)\right) & \text{if } k > 0. \end{cases}$$

Here for $k > 0$, κ_{a-1} denotes the most significant bit in the base q representation of k and $\Re$ is the real part function. Note that this result generalizes the result in [3, Appendix A], as we now also allow Walsh functions over arbitrary finite fields and arbitrary bijections ϕ between $\mathbb{Z}_q$ and $\mathbb{F}_q$ which satisfy $\phi(0) = 0$.

Remark 3.1. Since $K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ is a reproducing kernel, it is easy to see that the corresponding shift invariant kernel $K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}$ is a reproducing kernel as well and from this one can see (by the properties of reproducing kernels; see [1]) that $\widehat{r}_q(w,\gamma,k)$ is nonnegative for any $k \in \mathbb{N}_0$.

Further, for $x = \frac{x_1}{q} + \frac{x_2}{q^2} + \cdots$ and $y = \frac{y_1}{q} + \frac{y_2}{q^2} + \cdots$ we define $\rho_{\mathrm{ds},q,w}(x,x) := w^2 - w + \frac{1}{2}$ and if $x \neq y$,

$$\rho_{\mathrm{ds},q,w}(x,y) := w^2 - w + \frac{1}{2} - \frac{1}{2q^{i_0+1}} \times \left(\sum_{\substack{u=0 \\ u < u \oplus_\varphi x_{i_0} \ominus_\varphi y_{i_0}}}^{q-1} (u \oplus_\varphi x_{i_0} \ominus_\varphi y_{i_0} - u) + \sum_{\substack{u=0 \\ u < u \oplus_\varphi y_{i_0} \ominus_\varphi x_{i_0}}}^{q-1} (u \oplus_\varphi y_{i_0} \ominus_\varphi x_{i_0} - u) \right), \tag{3.1}$$

where i_0 is the smallest index such that the digits of x and y differ. Note that we have $\rho_{\mathrm{ds},q,w}(x,y) = \rho_{\mathrm{ds},q,w}(x \ominus_\varphi y, 0)$. Using Lemma B.2 it can easily be checked that

$$K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^s (1 + \gamma_j \rho_{\mathrm{ds},q,w_j}(x_j, y_j)).$$

Now we obtain, as in [3], that the mean square worst-case error for integration in the weighted Sobolev space $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ by using a random digital shift in base q with respect to a bijection φ on the point set $P_N = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$, with $\boldsymbol{x}_h = (x_{h,1}, \ldots, x_{h,s})$, is given by

$$\begin{aligned} &\widehat{e}^2(P_n, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) \\ &= \int_{[0,1)^{2s}} K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - \frac{2}{N}\sum_{h=0}^{N-1}\int_{[0,1)^s} K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x}_h,\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ &\quad + \frac{1}{N^2}\sum_{h,n=0}^{N-1} K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x}_h,\boldsymbol{x}_n) \\ &= -\prod_{j=1}^s \left(1 + \gamma_j\left(w_j^2 - w_j + \tfrac{1}{3}\right)\right) + \frac{1}{N^2}\sum_{h,n=0}^{N-1}\sum_{\boldsymbol{k}\in\mathbb{N}_0^s} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h)\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n)} \\ &= -\prod_{j=1}^s \left(1 + \gamma_j\left(w_j^2 - w_j + \tfrac{1}{3}\right)\right) + \frac{1}{N^2}\sum_{h,n=0}^{N-1}\prod_{j=1}^s (1 + \gamma_j \rho_{\mathrm{ds},q,w_j}(x_{h,j}, x_{n,j})). \end{aligned}$$

For the special case where the point set $P_{q^m,\varphi}$ is a digital (t,m,s)-net over $\mathbb{F}_q$ with generating matrices $C_1,\dots,C_s$ and bijection φ (the same bijection as for the digital shift) we obtain

$$\begin{aligned}\widehat{e}^2(P_{q^m,\varphi},K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) &= -\prod_{j=1}^{s}\Big(1+\gamma_j\Big(w_j^2-w_j+\tfrac{1}{3}\Big)\Big)\\ &\quad+\frac{1}{q^{2m}}\sum_{n=0}^{q^m-1}\left(\sum_{h=0}^{q^m-1}\sum_{\boldsymbol{k}\in\mathbb{N}_0^s}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h\ominus_\varphi\boldsymbol{x}_n)\right)\\ &= -\prod_{j=1}^{s}\Big(1+\gamma_j\Big(w_j^2-w_j+\tfrac{1}{3}\Big)\Big)\\ &\quad+\frac{1}{q^{2m}}\sum_{n=0}^{q^m-1}\left(\sum_{h=0}^{q^m-1}\prod_{j=1}^{s}(1+\gamma_j\rho_{\mathrm{ds},q,w_j}(x_{h,j}\ominus_\varphi x_{n,j},0))\right).\end{aligned}$$

It is easy to show that a digital net $P_{q^m,\varphi}$ over $\mathbb{F}_q$ generated by matrices $C_1,\dots,C_s$ with bijection φ together with the addition $\oplus_\varphi$ becomes a group. Hence each term in the sum over n has the same value and therefore we obtain

$$\begin{aligned}&\widehat{e}^2(P_{q^m,\varphi},K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})\\ &= -\prod_{j=1}^{s}\Big(1+\gamma_j\Big(w_j^2-w_j+\tfrac{1}{3}\Big)\Big)+\sum_{\boldsymbol{k}\in\mathbb{N}_0^s}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k})\frac{1}{q^m}\sum_{h=0}^{q^m-1}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h)\\ &= -\prod_{j=1}^{s}\Big(1+\gamma_j\Big(w_j^2-w_j+\tfrac{1}{3}\Big)\Big)+\frac{1}{q^m}\sum_{h=0}^{q^m-1}\prod_{j=1}^{s}(1+\gamma_j\rho_{\mathrm{ds},q,w_j}(x_{h,j},0)).\end{aligned}$$

For $k\in\mathbb{N}_0$, $k=\kappa_0+\kappa_1 q+\cdots$ denote an m-bit truncation by

$$\mathrm{tc}_{m,\varphi}(k)=(\varphi(\kappa_0),\dots,\varphi(\kappa_{m-1}))^\top\in\mathbb{F}_q^m.$$

For vectors $\boldsymbol{k}\in\mathbb{N}_0^s$ the mapping $\mathrm{tc}_{m,\varphi}$ is defined componentwise. Further define

$$\mathcal{N}=\{\boldsymbol{k}=(k_1,\dots,k_s)\in(\mathbb{F}_q^m)^s : C_1^\top k_1+\cdots+C_s^\top k_s=\boldsymbol{0}\},$$

where $\boldsymbol{0}$ is the zero vector in $\mathbb{F}_q^m$.

Using these definitions and Lemma 2.5 we obtain the next theorem.

THEOREM 3.2. *Let $P_{q^m,\varphi}=\{\boldsymbol{x}_0,\dots,\boldsymbol{x}_{q^m-1}\}$ be a digital (t,m,s)-net over $\mathbb{F}_q$ generated by $C_1,\dots,C_s$ and with respect to the bijection φ, where $\varphi(0)=0$.*

1. *The mean square worst-case error for integration in the weighted Sobolev space $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ by using the digital net $P_{q^m,\varphi}$ is given by*

$$\widehat{e}^2(P_{q^m,\varphi},K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})=\sum_{\substack{\boldsymbol{k}\in\mathbb{N}_0^s\setminus\{\boldsymbol{0}\}\\ \mathrm{tc}_{m,\varphi}(\boldsymbol{k})\in\mathcal{N}}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k}).$$

2. *Let $\boldsymbol{x}_h=(x_{h,1},\dots,x_{h,s})$ for $0\le h\le q^m-1$. Then we have*

$$\begin{aligned}&\widehat{e}^2(P_{q^m,\varphi},K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})\\ &= -\prod_{j=1}^{s}\Big(1+\gamma_j\Big(w_j^2-w_j+\tfrac{1}{3}\Big)\Big)+\frac{1}{q^m}\sum_{h=0}^{q^m-1}\prod_{j=1}^{s}(1+\gamma_j\rho_{\mathrm{ds},q,w_j}(x_{h,j},0)),\end{aligned}\tag{3.2}$$

where $\rho_{\mathrm{ds},q,w}$ is given by (3.1).

Note that formula (3.2) allows us to compute the mean square worst-case error for integration in the weighted Sobolev space $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ for any digital net over $\mathbb{F}_q$ in $O(q^m s)$ operations.

3.1. Integration in $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ with cyclic nets. In this subsection we consider the special case where the digital net used for the QMC rule is a cyclic net.

Let $\alpha \in \mathbb{F}_{q^m}$, $\mathcal{B}_1, \ldots, \mathcal{B}_s$ be s ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$ and let $C_1, \ldots, C_s$ be given as in Theorem 2.7. Then we have

$$\mathcal{N} = \mathcal{N}_\alpha = \phi(\mathcal{P}_\alpha);$$

see Definition 2.6. Let φ be a bijection from $\mathbb{Z}_q$ to $\mathbb{F}_q$ with $\varphi(0) = 0$. The cyclic net generated by $C_1, \ldots, C_s$ with respect to φ will be denoted by $P_{\alpha,\varphi}$.

ALGORITHM 3.3. *Given a dimension $s \geq 2$, an integer $m \geq 1$, and weights $\boldsymbol{\gamma} = (\gamma_j)_{j\geq 1}$,*

1. *Choose a prime power q, a finite field $\mathbb{F}_q$ with q elements, a bijection $\varphi : \{0, 1, \ldots, q-1\} \longrightarrow \mathbb{F}_q$ with $\varphi(0) = 0$, and s ordered bases $\mathcal{B}_1, \ldots, \mathcal{B}_s$ of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$.*
2. *Find $\alpha \in \mathbb{F}_{q^m} \setminus \{0\}$ that minimizes $\widehat{e}^2(P_{\alpha,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})$.*

In the following theorem we show that this construction yields the same upper bound as the Korobov construction of polynomial lattice rules, which is of course not surprising as Korobov polynomial lattice rules are just a special case.

THEOREM 3.4. *Let $s \geq 2$, let q be a prime power, let $\mathbb{F}_q$ be a finite field with q elements, let $\varphi : \{0, 1, \ldots, q-1\} \longrightarrow \mathbb{F}_q$ be a bijection with $\varphi(0) = 0$, and let $\mathcal{B}_1, \ldots, \mathcal{B}_s$ be s ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$. Further let $m \geq 1$. Assume that $\alpha^* \in \mathbb{F}_{q^m} \setminus \{0\}$ is constructed by Algorithm 3.3. Then we have*

$$\widehat{e}^2(P_{\alpha^*,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) \leq \Big(\frac{s}{q^m - 1}\Big)^{\frac{1}{\lambda}} \prod_{j=1}^{s} \Big(\Big(1 + \gamma_j \left[w_j^2 - w_j + \tfrac{1}{3}\right]\Big)^{\lambda} + \gamma_j^{\lambda} \zeta_q(\lambda)\Big)^{\frac{1}{\lambda}}$$

for all $\frac{1}{2} < \lambda \leq 1$. Here for $\lambda = 1$, $\zeta_q(1) = \frac{1}{6}$ and for $\frac{1}{2} < \lambda < 1$ we have

$$\zeta_2(\lambda) = \frac{1}{3^{\lambda}(2^{2\lambda} - 2)} \quad \text{and} \quad \zeta_q(\lambda) = \frac{(q-1)q^{2\lambda}}{6^{\lambda}(q^{2\lambda} - q)} \quad \text{for} \quad q \neq 2.$$

Proof. Since

$$\min_{\alpha \in \mathbb{F}_{q^m} \setminus \{0\}} \widehat{e}^2(P_{\alpha,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) \leq \left(\frac{1}{q^m - 1} \sum_{\alpha \in \mathbb{F}_{q^m} \setminus \{0\}} \widehat{e}^2(P_{\alpha,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})^{\lambda}\right)^{\frac{1}{\lambda}},$$

it is enough to show that the inequality

$$\frac{1}{q^m - 1} \sum_{\alpha \in \mathbb{F}_{q^m} \setminus \{0\}} \widehat{e}^2(P_{\alpha,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})^{\lambda} \leq \frac{s}{q^m - 1} \prod_{j=1}^{s} \Big(\Big(1 + \gamma_j \left[w_j^2 - w_j + \tfrac{1}{3}\right]\Big)^{\lambda} + \gamma_j^{\lambda} \zeta_q(\lambda)\Big)$$

holds. With Jensen's inequality, which states that for a sequence (a_k) of nonnegative reals we have

$$\Big(\sum a_k\Big)^{\lambda} \leq \sum a_k^{\lambda}$$

for any $0 < \lambda \le 1$, we obtain

$$\begin{aligned} M_{s,q} &:= \frac{1}{q^m - 1} \sum_{\alpha \in \mathbb{F}_{q^m} \setminus \{0\}} \widehat{e}^2(P_\alpha, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}})^\lambda \\ &\le \frac{1}{q^m - 1} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l})^\lambda A(\mathrm{tc}_{m,\varphi}(\boldsymbol{l})), \end{aligned}$$

where for $\boldsymbol{k} \in \mathbb{F}_q^{sm}$ we define

$$A(\boldsymbol{k}) := \#\{\alpha \in \mathbb{F}_{q^m} \setminus \{0\} : \boldsymbol{k} \in \mathcal{N}_\alpha\}.$$

Now $\boldsymbol{k} \in \mathbb{F}_q^{sm} \setminus \{\boldsymbol{0}\}$ is contained in $\mathcal{N}_\alpha$ iff α is a zero of the corresponding polynomial $\phi^{-1}(\boldsymbol{k})$. This polynomial has degree of at most $s-1$ and hence it has at most $s-1$ zeros. Thus $A(\boldsymbol{k}) \le s-1$. Further, we have $A(\boldsymbol{0}) = q^m - 1$.

For $\boldsymbol{l} \in \mathbb{N}_0^s$ we have $\mathrm{tc}_{m,\varphi}(q^m \boldsymbol{l}) = \boldsymbol{0}$ and hence

$$\begin{aligned} M_{s,q} &\le \frac{1}{q^m - 1} \Bigg(\sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, q^m \boldsymbol{l})^\lambda A(\boldsymbol{0}) \\ &\qquad + \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \sum_{\substack{\boldsymbol{l}^* \in \mathbb{N}_0^s \\ 0 < \|\boldsymbol{l}^*\|_\infty < q^m}} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l}^* + q^m \boldsymbol{l})^\lambda A(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}^*)) \Bigg) \\ &\le \sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, q^m \boldsymbol{l})^\lambda + \frac{s-1}{q^m - 1} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l})^\lambda. \end{aligned}$$

The first sum in the last line in the inequality above can be estimated by

$$\begin{aligned} &\sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, q^m \boldsymbol{l})^\lambda \\ &\quad = -\prod_{j=1}^s \widehat{r}_q(w_j, \gamma_j, 0)^\lambda + \prod_{j=1}^s \sum_{k=0}^\infty \widehat{r}_q(w_j, \gamma_j, q^m k)^\lambda \\ &\quad = -\prod_{j=1}^s \left(1 + \gamma_j \left(w_j^2 - w_j + \tfrac{1}{3}\right)\right)^\lambda \\ &\qquad + \prod_{j=1}^s \left(\left(1 + \gamma_j \left(w_j^2 - w_j + \tfrac{1}{3}\right)\right)^\lambda + \sum_{k=1}^\infty \widehat{r}_q(w_j, \gamma_j, q^m k)^\lambda \right) \\ &\quad \le \frac{1}{q^{2\lambda m}} \prod_{j=1}^s \left(\left(1 + \gamma_j \left[w_j^2 - w_j + \tfrac{1}{3}\right]\right)^\lambda + \sum_{k=1}^\infty \widehat{r}_q(w_j, \gamma_j, k)^\lambda \right) \end{aligned}$$

and the second sum is

$$\begin{aligned} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l})^\lambda &= \prod_{j=1}^s \sum_{k=0}^\infty \widehat{r}_q(w_j, \gamma_j, k)^\lambda \\ &= \prod_{j=1}^s \left(\left(1 + \gamma_j \left[w_j^2 - w_j + \tfrac{1}{3}\right]\right)^\lambda + \sum_{k=1}^\infty \widehat{r}_q(w_j, \gamma_j, k)^\lambda \right). \end{aligned}$$

We have

$$\sum_{k=1}^{\infty} \widehat{r}_q(w_j,\gamma_j,k)^\lambda =: \gamma_j^\lambda \mu_q(\lambda).$$

First note that

$$\left|\frac{1}{6}+\Re\left(\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}\right)\right| \leq \frac{1}{6}+\frac{1}{q}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}(v-u)$$
$$= \frac{1}{6}+\frac{1}{q}\frac{q(q+1)(q-1)}{6} = \frac{q^2}{6}.$$

For $\frac{1}{2}<\lambda<1$ we have

$$\mu_q(\lambda) = \sum_{k=1}^{\infty}\left(-\frac{1}{2}\left(\frac{1}{3q^{2a}}+\frac{2}{q^{2a}}\Re\left(\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}\right)\right)\right)^\lambda$$
$$\leq \sum_{a=1}^{\infty}\sum_{k=q^{a-1}}^{q^a-1}\frac{1}{q^{2\lambda a}}\left(\frac{q^2}{6}\right)^\lambda = \frac{(q-1)q^{2\lambda}}{6^\lambda(q^{2\lambda}-q)} =: \zeta_q(\lambda)$$

and for $\lambda=1$ we have

$$\mu_q(1)$$
$$= -\frac{1}{2}\sum_{a=1}^{\infty}\sum_{k=q^{a-1}}^{q^a-1}\left(\frac{1}{3q^{2a}}+\frac{2}{q^{2a}}\Re\left(\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}\right)\right)$$
$$= -\frac{1}{6}\sum_{a=1}^{\infty}\frac{1}{q^{2a}}(q^a-q^{a-1}) - \Re\left(\sum_{a=1}^{\infty}\frac{1}{q^{2a}}\sum_{k=q^{a-1}}^{q^a-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}\right).$$

Now with Lemma B.1 we obtain

$$\sum_{a=1}^{\infty}\frac{1}{q^{2a}}\sum_{k=q^{a-1}}^{q^a-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}$$
$$= \sum_{a=1}^{\infty}\frac{1}{q^{a+2}}\sum_{\kappa_{a-1}=1}^{q-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}(v-u)\mathrm{wal}_{\kappa_{a-1}}(\tfrac{u\ominus_\varphi v}{q}) = \sum_{a=1}^{\infty}\frac{1}{q^{a+2}}q\frac{1-q^2}{6} = -\frac{q+1}{6q}.$$

Therefore we have

$$\mu_q(1) = -\frac{1}{6q}+\frac{q+1}{6q} = \frac{1}{6} =: \zeta_q(1).$$

Let now $q=2$ and $\frac{1}{2}<\lambda<1$. Then

$$\mu_2(\lambda) = \sum_{a=1}^{\infty}\sum_{k=2^{a-1}}^{2^a-1}\left(-\frac{1}{2}\left(\frac{1}{3\cdot 2^{2a}}+\frac{2}{2^{2a}}\sum_{\substack{u,v=0\\ v\geq u}}^{1}\frac{(v-u)\mathrm{wal}_1\left(\frac{u\ominus_\varphi v}{2}\right)}{2}\right)\right)^\lambda$$

and by using

$$\sum_{\substack{u,v=0\\ v\ge u}}^{1} \frac{(v-u)\mathrm{wal}_1\left(\frac{u\ominus_\varphi v}{2}\right)}{2} = \frac{\mathrm{wal}_1\left(\frac{0\ominus_\varphi 1}{2}\right)}{2} = -\frac{1}{2},$$

we obtain

$$\mu_2(\lambda) = \frac{1}{3^\lambda(2^{2\lambda}-2)} =: \zeta_2(\lambda).$$

The theorem follows. □

The following corollary shows that under certain conditions on the weights we can obtain an upper bound which depends only polynomially on the dimension and thus proving tractability. (See [20] for more information on tractability.) An analogous result was shown for the Korobov construction of polynomial lattice rules; see [4].

COROLLARY 3.5. *Let $s \ge 2$, let q be a prime power, $\mathbb{F}_q$ be a finite field with q elements, $\varphi : \{0,1,\dots,q-1\} \longrightarrow \mathbb{F}_q$ be a bijection with $\varphi(0) = 0$, and $\mathcal{B}_1,\dots,\mathcal{B}_s$ be s ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$. Let $m \ge 1$ and suppose $\alpha^* \in \mathbb{F}_{q^m} \setminus \{0\}$ is constructed by Algorithm 3.3. Let $N = q^m$.*

1. *We have*

$$\widehat{e}(P_{\alpha^*,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) \le c_{s,\boldsymbol{w},\boldsymbol{\gamma},\delta} s^{1-\delta} N^{-1+\delta} \qquad \text{for all } 0 < \delta \le \tfrac{1}{2},$$

where

$$c_{s,\boldsymbol{w},\boldsymbol{\gamma},\delta} := 2^{1-\delta} \prod_{j=1}^{s} \left(1 + \gamma_j^{\frac{1}{2(1-\delta)}} \left[\left(w_j^2 - w_j + \tfrac{1}{3}\right)^{\frac{1}{2(1-\delta)}} + \zeta_q\left(\tfrac{1}{2(1-\delta)}\right)\right]\right)^{1-\delta}.$$

2. *Under the assumption*

$$A := \limsup_{s\to\infty} \frac{\sum_{j=1}^{s} \gamma_j}{\log s} < \infty$$

we obtain $c_{s,\boldsymbol{w},\boldsymbol{\gamma},1/2} \le \bar{c}_\eta s^{(A+\eta)/2}$ and therefore

$$\widehat{e}(P_{\alpha^*,\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) \le \bar{c}_\eta s^{(1+(A+\eta))/2} N^{-\frac{1}{2}} \qquad \text{for all } \eta > 0,$$

where the constant $\bar{c}_\eta$ depends only on the arbitrarily chosen parameter η. Thus the root mean square worst-case error of the cyclic net generated by α^* (with respect to the bijection φ) satisfies a bound which depends only polynomially on the dimension.

The result can be shown using the methods employed in the proof of [4, Corollary 4.8].

3.2. Integration in $H_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}$ with hyperplane nets. In this subsection we consider the special case where the digital net used for the QMC rule is a hyperplane net.

Let $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_s) \in \mathbb{F}_{q^m}^s$, $\mathcal{B}_1,\dots,\mathcal{B}_s$ be s ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$ and let $C_1,\dots,C_s$ be given as in Definition 2.10. Let

$$\mathcal{N}_{\boldsymbol{\alpha}} = \{\boldsymbol{k} = (k_1,\dots,k_s) \in (\mathbb{F}_q^m)^s : C_1^\top k_1 + \cdots + C_s^\top k_s = \boldsymbol{0}\},$$

where $\mathbf{0}$ is the zero vector in $\mathbb{F}_q^m$. Let φ be a bijection from $\mathbb{Z}_q$ to $\mathbb{F}_q$ with $\varphi(0) = 0$. The hyperplane net generated by $C_1, \dots, C_s$ with respect to φ will be denoted by $P_{\boldsymbol{\alpha},\varphi}$.

From now on we write $\widehat{e}^2(P_{\boldsymbol{\alpha},\varphi}, K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}) = \widehat{e}^2(\alpha_1, \dots, \alpha_s)$ to stress the dependence of the mean square worst-case error on $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_s)$.

ALGORITHM 3.6. *Given a dimension $s \geq 2$, an integer $m \geq 1$, and weights $\boldsymbol{\gamma} = (\gamma_j)_{j \geq 1}$,*

1. *Choose a prime power q, a finite field $\mathbb{F}_q$ with q elements, a bijection $\varphi : \{0, 1, \dots, q-1\} \longrightarrow \mathbb{F}_q$ with $\varphi(0) = 0$, and ordered bases $\mathcal{B}_1, \dots, \mathcal{B}_s$ of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$.*
2. *Choose $\alpha_1 \in \mathbb{F}_{q^m} \setminus \{0\}$.*
3. *For $d = 2, 3, \dots, s$ find $\alpha_d \in \mathbb{F}_{q^m} \setminus \{0\}$ by minimizing the mean square worst-case error $\widehat{e}^2(\alpha_1, \dots, \alpha_d)$.*

In the following theorem we show that this construction yields the same upper bound as the component-by-component construction of polynomial lattice rules.

THEOREM 3.7. *Let $s \geq 2$, let q be a prime power, let $\mathbb{F}_q$ be a finite field with q elements, let $\varphi : \{0, 1, \dots, q-1\} \longrightarrow \mathbb{F}_q$ be a bijection with $\varphi(0) = 0$, and let $\mathcal{B}_1, \dots, \mathcal{B}_s$ be ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$. Further let $m \geq 1$. Assume that $(\alpha_1^*, \dots, \alpha_s^*) \in (\mathbb{F}_{q^m} \setminus \{0\})^s$ is constructed by Algorithm 3.6. Then for all $d = 1, 2, \dots, s$ we have*

$$\widehat{e}^2(\alpha_1^*, \dots, \alpha_d^*) \leq (q^m - 1)^{-\frac{1}{\lambda}} \prod_{j=1}^{d} \left(\left(1 + \gamma_j \left[w_j^2 - w_j + \tfrac{1}{3} \right] \right)^{\lambda} + \zeta_q(\lambda) \gamma_j^{\lambda} \right)^{\frac{1}{\lambda}}$$

for all $\frac{1}{2} < \lambda \leq 1$. Here $\zeta_q(\lambda)$ is defined as in Theorem 3.4.

Proof. First we show that the inequality

$$\widehat{e}^2(\alpha_1^*, \dots, \alpha_d^*) \leq \left(\frac{1}{q^m - 1} \sum_{\boldsymbol{l} \in \mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l})^{\lambda} \right)^{\frac{1}{\lambda}} \tag{3.3}$$

holds for all $d = 1, 2, \dots, s$.

We start with $d = 1$. The generating matrix C_1 is regular for all $\alpha_1 \in \mathbb{F}_{q^m} \setminus \{0\}$ and hence $\mathcal{N}_{\alpha_1} = \{\mathbf{0}\}$. Hence we have

$$\widehat{e}^2(\alpha_1^*) = \sum_{\substack{k \in \mathbb{N}_0 \setminus \{0\} \\ q^m \mid k}} \widehat{r}_q(\omega_1, \gamma_1, k) = \sum_{k=1}^{\infty} \widehat{r}_q(\omega_1, \gamma_1, q^m k) = \frac{1}{q^{2m}} \sum_{k=1}^{\infty} \widehat{r}_q(\omega_1, \gamma_1, k).$$

The result for $d = 1$ now follows by applying Jensen's inequality to the infinite sum above.

Suppose for some $1 \leq d < s$ we have $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_d^*) \in (\mathbb{F}_{q^m} \setminus \{0\})^d$ and

$$\widehat{e}^2(\boldsymbol{\alpha}^*) \leq \left(\frac{1}{q^m - 1} \sum_{\boldsymbol{l} \in \mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{l})^{\lambda} \right)^{\frac{1}{\lambda}}.$$

Now we have

$$\begin{aligned}\widehat{e}^2(\boldsymbol{\alpha},\alpha_{d+1}) &= \sum_{\substack{(\boldsymbol{l},l_{d+1})\in\mathbb{N}_0^{d+1}\setminus\{\mathbf{0}\}\\(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}),\mathrm{tc}_{m,\varphi}(l_{d+1}))\in\mathcal{N}_{(\boldsymbol{\alpha},\alpha_{d+1})}}} \widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})\\ &= \widehat{r}_q(w_{d+1},\gamma_{d+1},0)\widehat{e}^2(\boldsymbol{\alpha})+\theta(\alpha_{d+1}),\end{aligned}$$

where

$$\theta(\alpha_{d+1}) = \sum_{l_{d+1}=1}^{\infty} \widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1}) \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}),\mathrm{tc}_{m,\varphi}(l_{d+1}))\in\mathcal{N}_{(\boldsymbol{\alpha},\alpha_{d+1})}}} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l}).$$

In Algorithm 3.6, α_{d+1}^* is chosen such that the mean square worst-case error $\widehat{e}^2(\boldsymbol{\alpha}^*,\alpha_{d+1})$ is minimized. Since the only dependence on α_{d+1} is in $\theta(\alpha_{d+1})$, we have $\theta(\alpha_{d+1}^*) \le \theta(\alpha_{d+1})$ for all $\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}$. Hence for any $\lambda\le 1$ we obtain

$$\theta(\alpha_{d+1}^*) \le \left(\frac{1}{q^m-1}\sum_{\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}}\theta(\alpha_{d+1})^\lambda\right)^{\frac{1}{\lambda}}.$$

From Jensen's inequality it follows that

$$\theta(\alpha_{d+1})^\lambda \le \sum_{l_{d+1}=1}^{\infty} \widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})^\lambda \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}),\mathrm{tc}_{m,\varphi}(l_{d+1}))\in\mathcal{N}_{(\boldsymbol{\alpha},\alpha_{d+1})}}} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda.$$

If l_{d+1} is a multiple of q^m, then $\mathrm{tc}_{m,\varphi}(l_{d+1})=0$ and the sum is independent of α_{d+1}. Otherwise $\mathrm{tc}_{m,\varphi}(l_{d+1})\ne 0$. We obtain

$$\begin{aligned}\frac{1}{q^m-1}\sum_{\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}}\theta(\alpha_{d+1})^\lambda &\le \sum_{\substack{l_{d+1}=1\\q^m\mid l_{d+1}}}^{\infty} \widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})^\lambda \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\\mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\\ &+ \frac{1}{q^m-1}\sum_{\substack{l_{d+1}=1\\q^m\nmid l_{d+1}}}^{\infty} \widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})^\lambda \sum_{\boldsymbol{l}\in\mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda \sum_{\substack{\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}\\(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}),\mathrm{tc}_{m,\varphi}(l_{d+1}))\in\mathcal{N}_{(\boldsymbol{\alpha},\alpha_{d+1})}}} 1.\end{aligned}$$

From $\mathrm{tc}_{m,\varphi}(l_{d+1})\ne 0$ we obtain

$$\begin{aligned}&\sum_{\boldsymbol{l}\in\mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda \sum_{\substack{\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}\\(\mathrm{tc}_{m,\varphi}(\boldsymbol{l}),\mathrm{tc}_{m,\varphi}(l_{d+1}))\in\mathcal{N}_{(\boldsymbol{\alpha},\alpha_{d+1})}}} 1\\ &= \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\\mathrm{tc}_{m,\varphi}(\boldsymbol{l})\notin\mathcal{N}_{\boldsymbol{\alpha}}}} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda = \sum_{\boldsymbol{l}\in\mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda - \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\\mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda.\end{aligned}$$

Therefore we can estimate

$$\begin{aligned}
&\frac{1}{q^m-1}\sum_{\alpha_{d+1}\in\mathbb{F}_{q^m}\setminus\{0\}}\theta(\alpha_{d+1})^\lambda \le \sum_{\substack{l_{d+1}=1\\ q^m\mid l_{d+1}}}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})^\lambda \sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\ \mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\\
&\quad+\frac{1}{q^m-1}\sum_{\substack{l_{d+1}=1\\ q^m\nmid l_{d+1}}}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l_{d+1})^\lambda\left(\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda-\sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\ \mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)\\
&\le\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},q^m l)^\lambda\sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\ \mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\\
&\quad+\frac{1}{q^m-1}\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\left(\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda-\sum_{\substack{\boldsymbol{l}\in\mathbb{N}_0^d\\ \mathrm{tc}_{m,\varphi}(\boldsymbol{l})\in\mathcal{N}_{\boldsymbol{\alpha}}}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)\\
&\le\frac{1}{q^m-1}\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda.
\end{aligned}$$

Here the last inequality follows from the fact that

$$\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},q^m l)^\lambda-\frac{1}{q^m-1}\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\le 0,$$

which follows from the definition of $\widehat{r}_q$ and since $\lambda>\frac{1}{2}$. Now we have

$$\theta(\alpha_{d+1}^*)\le\left(\frac{1}{q^m-1}\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)^{\frac{1}{\lambda}}.$$

From this it follows that

$$\begin{aligned}
\widehat{e}^2(\boldsymbol{\alpha}^*,\alpha_{d+1}^*)&\le\widehat{r}_q(w_{d+1},\gamma_{d+1},0)\widehat{e}^2(\boldsymbol{\alpha}^*)\\
&\quad+\left(\frac{1}{q^m-1}\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)^{\frac{1}{\lambda}}\\
&\le\left(\frac{1}{q^m-1}\sum_{\boldsymbol{l}\in\mathbb{N}_0^d}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)^{\frac{1}{\lambda}}\left(\widehat{r}_q(w_{d+1},\gamma_{d+1},0)+\left(\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\right)^{\frac{1}{\lambda}}\right).
\end{aligned}$$

Again from Jensen's inequality we obtain

$$\widehat{r}_q(w_{d+1},\gamma_{d+1},0)+\left(\sum_{l=1}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\right)^{\frac{1}{\lambda}}\le\left(\sum_{l=0}^{\infty}\widehat{r}_q(w_{d+1},\gamma_{d+1},l)^\lambda\right)^{\frac{1}{\lambda}}$$

and hence

$$\widehat{e}^2(\boldsymbol{\alpha}^*,\alpha_{d+1}^*)\le\left(\frac{1}{q^m-1}\sum_{\boldsymbol{l}\in\mathbb{N}_0^{d+1}}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda\right)^{\frac{1}{\lambda}}.$$

This finishes our induction proof of inequality (3.3).

Finally we have

$$\sum_{\boldsymbol{l}\in\mathbb{N}_0^d} \widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{l})^\lambda = \prod_{j=1}^d \sum_{l=0}^\infty \widehat{r}_q(w_j,\gamma_j,l)^\lambda$$
$$= \prod_{j=1}^d \left((1+\gamma_j[w_j^2 - w_j + \tfrac{1}{3}])^\lambda + \sum_{l=1}^\infty \widehat{r}_q(w_j,\gamma_j,l)^\lambda \right).$$

As in the proof of Theorem 3.4 we obtain

$$\sum_{l=1}^\infty \widehat{r}_q(w_j,\gamma_j,l)^\lambda \le \gamma_j^\lambda \zeta_q(\lambda)$$

and the result follows. □

The following corollary shows that under certain conditions on the weights we can obtain an upper bound which depends only polynomially on the dimension, and, with stronger conditions on the weights, we can also obtain an upper bound which is independent of the dimension, thus proving strong tractability. (See [20] for more information on (strong) tractability.) An analogous result was shown for the component-by-component construction of polynomial lattice rules; see [4].

COROLLARY 3.8. *Let $s \ge 2$, let q be prime power, $\mathbb{F}_q$ be a finite field with q elements, $\varphi : \{0,1,\dots,q-1\} \longrightarrow \mathbb{F}_q$ be a bijection with $\varphi(0) = 0$, and $\mathcal{B}_1, \dots \mathcal{B}_s$ be s ordered bases of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$. Further let $m \ge 1$. Suppose $\boldsymbol{\alpha}^* \in (\mathbb{F}_{q^m} \setminus \{0\})^s$ is constructed by Algorithm 3.6. Let $N = q^m$.*

1. *We have*

$$\widehat{e}(\boldsymbol{\alpha}^*) \le c_{s,\boldsymbol{w},\boldsymbol{\gamma},\delta} N^{-1+\delta} \qquad \text{for all } 0 < \delta \le \tfrac{1}{2},$$

where

$$c_{s,\boldsymbol{w},\boldsymbol{\gamma},\delta} := 2^{1-\delta} \prod_{j=1}^s \left(1 + \gamma_j^{\frac{1}{2(1-\delta)}} \left[\left(w_j^2 - w_j + \tfrac{1}{3}\right)^{\frac{1}{2(1-\delta)}} + \zeta_q\left(\tfrac{1}{2(1-\delta)}\right)\right]\right)^{1-\delta}.$$

2. *Suppose*

$$\sum_{j=1}^\infty \gamma_j^{\frac{1}{2(1-\delta)}} < \infty.$$

Then $c_{s,\boldsymbol{w},\boldsymbol{\gamma},\delta} \le c_{\infty,\boldsymbol{w},\boldsymbol{\gamma},\delta} < \infty$ and we have

$$\widehat{e}(\boldsymbol{\alpha}^*) \le c_{\infty,\boldsymbol{w},\boldsymbol{\gamma},\delta} N^{-1+\delta} \qquad \text{for all } 0 < \delta \le \tfrac{1}{2}.$$

Thus the root mean square worst-case error of the hyperplane net generated by $\boldsymbol{\alpha}^*$ (with respect to the bijection φ) is bounded independently of the dimension.

3. *Under the assumption*

$$A := \limsup_{s\to\infty} \frac{\sum_{j=1}^s \gamma_j}{\log s} < \infty$$

we obtain $c_{s,\boldsymbol{w},\boldsymbol{\gamma},1/2} \le \widetilde{c}_\eta s^{(A+\eta)/2}$ and therefore

$$\widehat{e}(\boldsymbol{\alpha}^*) \le \widetilde{c}_\eta s^{(A+\eta)/2} N^{-\frac{1}{2}} \qquad \text{for all } \eta > 0,$$

where the constant $\widetilde{c}_\eta$ depends only on the arbitrarily chosen parameter η. Thus the root mean square worst-case error of the hyperplane net generated by $\boldsymbol{\alpha}^*$ (with respect to the bijection φ) satisfies a bound which depends only polynomially on the dimension.

The result can be shown using the methods employed in the proof of [4, Corollary 4.5].

Appendix A. Computation of the digital shift invariant kernel. Here we compute the digital shift invariant kernel for the reproducing kernel

$$K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{s}(1+\gamma_j \varrho_{w_j}(x_j,y_j)),$$

where $\boldsymbol{w} = (w_1,\dots,w_s) \in [0,1]^s$ and

$$\begin{aligned}\varrho_w(x,y) &= \frac{|x-w|+|y-w|-|x-y|}{2}\\ &= \begin{cases}\min(|x-w|,|y-w|) & \text{if } (x-w)(y-w)\ge 0,\\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

We have

$$\begin{aligned}K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x},\boldsymbol{y}) &:= \int_{[0,1)^s} K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}(\boldsymbol{x}\oplus_\varphi \boldsymbol{\sigma}, \boldsymbol{y}\oplus_\varphi \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{\sigma}\\ &= \prod_{j=1}^{s}\int_0^1 K_{\mathrm{sob},1,w_j,\gamma_j}(x_j\oplus_\varphi \sigma, y_j\oplus_\varphi \sigma)\,\mathrm{d}\sigma,\end{aligned}$$

where $K_{\mathrm{sob},1,w,\gamma}(x,y) := 1+\gamma\rho_w(x,y)$. So it suffices to compute

$$K_{\mathrm{ds},q,\gamma,w,\varphi}(x,y) := \int_0^1 K_{\mathrm{sob},1,w,\gamma}(x\oplus_\varphi \sigma, y\oplus_\varphi \sigma)\,\mathrm{d}\sigma.$$

It can easily be seen that the function $K_{\mathrm{sob},1,w,\gamma}(x,y)$ is in $\mathcal{L}_2([0,1)^2)$ and therefore we can apply [3, Lemma 5], from which we find that

$$K_{\mathrm{ds},q,\gamma,w,\varphi}(x,y) = \sum_{k=0}^{\infty}\widehat{K}(k)\mathrm{wal}_k(x)\overline{\mathrm{wal}_k(y)},$$

where

$$\widehat{K}(k) = \int_0^1\int_0^1 K_{\mathrm{sob},1,w,\gamma}(x,y)\overline{\mathrm{wal}_k(x)}\mathrm{wal}_k(y)\,\mathrm{d}x\,\mathrm{d}y.$$

By Proposition 2.4(3) it follows easily that

$$\widehat{K}(k) = \begin{cases}1+\gamma(w^2-w+\frac{1}{3}) & \text{if } k=0,\\ -\frac{\gamma}{2}\int_0^1\int_0^1 |x-y|\mathrm{wal}_k(y\ominus_\varphi x)\,\mathrm{d}x\,\mathrm{d}y & \text{if } k>0.\end{cases}$$

Now we evaluate the last integral.

LEMMA A.1. *Let $q^{a-1} \le k < q^a, \kappa_{a-1} = \lfloor k/q^{a-1} \rfloor > 0$. Then*

$$\tau(k) := \int_{[0,1)^2} |x-y| \mathrm{wal}_k(y \ominus_\varphi x)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{3q^{2a}} + \frac{2}{q^{2a}} \Re \left(\sum_{\substack{u,v=0 \\ v \ge u}}^{q-1} \frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u \ominus_\varphi v}{q}\right)}{q} \right),$$

where $\Re$ is the real part function.

Proof. First we remark that

$$\int_{u/q^a}^{(u+1)/q^a} \int_{v/q^a}^{(v+1)/q^a} |x-y|\,\mathrm{d}x\,\mathrm{d}y = \begin{cases} 1/(3q^{3a}) & \text{if } u = v, \\ |v-u|/(q^{3a}) & \text{otherwise.} \end{cases}$$

In the same way as in [3, Appendix A], we partition the unit square into subsquares where the Walsh function is constant and get

$$\tau(k) = \sum_{u,v=0}^{q^a-1} \mathrm{wal}_k\left(\frac{u \ominus_\varphi v}{q^a}\right) \int_{u/q^a}^{(u+1)/q^a} \int_{v/q^a}^{(v+1)/q^a} |x-y|\,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{3q^{2a}} + \frac{1}{q^{3a}} \sum_{\substack{u,v=0 \\ v>u}}^{q^a-1} (v-u) \left(\mathrm{wal}_k\left(\frac{u \ominus_\varphi v}{q^a}\right) + \overline{\mathrm{wal}_k\left(\frac{u \ominus_\varphi v}{q^a}\right)} \right)$$
$$= \frac{1}{3q^{2a}} + \frac{2}{q^{3a}} \Re \left(\sum_{\substack{u,v=0 \\ v>u}}^{q^a-1} (v-u)\mathrm{wal}_k\left(\frac{u \ominus_\varphi v}{q^a}\right) \right). \tag{A.1}$$

Let

$$0 \le u = qu' + u_0 < v = qv' + v_0 < q^a, \qquad v' > u', \qquad 0 \le u_0, v_0 < q.$$

Since $|\mathrm{wal}_k((u' \ominus_\varphi v')/q^{a-1})| = 1$ and $\sum_{u_0,v_0}(v'-u')\mathrm{wal}_{\kappa_{a-1}}((u_0 \ominus_\varphi v_0)/q) = 0$, we have

$$\left| \sum_{u_0,v_0=0}^{q-1} ((qv'+v_0) - (qu'+u_0))\mathrm{wal}_k\left(\frac{(qu'+u_0) \ominus_\varphi (qv'+v_0)}{q^a}\right) \right| = \left| \sum_{u_0,v_0=0}^{q-1} T_{\kappa_{a-1}}(u_0,v_0) \right|,$$

where $T_\kappa(u,v) := (v-u)\mathrm{wal}_\kappa((u \ominus_\varphi v)/q)$. By the character properties of the Walsh function system

$$\sum_{u_0,v_0=0}^{q-1} T_{\kappa_{a-1}}(u_0,v_0) = \sum_{u_0,v_0=0}^{q-1} (v_0-u_0)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u_0 \ominus_\varphi v_0}{q}\right) \tag{A.2}$$
$$= \sum_{v_0=0}^{q-1} v_0 \overline{\mathrm{wal}}_{\kappa_{a-1}}\left(\frac{v_0}{q}\right) \sum_{u_0=0}^{q-1} \mathrm{wal}_{\kappa_{a-1}}\left(\frac{u_0}{q}\right) - \sum_{u_0=0}^{q-1} u_0 \mathrm{wal}_{\kappa_{a-1}}\left(\frac{u_0}{q}\right) \sum_{v_0=0}^{q-1} \overline{\mathrm{wal}}_{\kappa_{a-1}}\left(\frac{v_0}{q}\right) = 0.$$

So the only (u,v) left in the sum (A.1) are $(qu'+u_0, qu'+v_0)$ for $0 \le u' < q^{a-1}$. In

this case the summands in the sum of (A.1) equal $T_{\kappa_{a-1}}(u_0, v_0)$ for fixed u', we get

$$\begin{aligned}\tau(k) &= \frac{1}{3q^{2a}} + \frac{2}{q^{3a}}\Re\left(\sum_{u'=0}^{q^{a-1}-1}\sum_{\substack{u_0,v_0=0\\ v_0>u_0}}^{q-1} T_{\kappa_{a-1}}(u_0,v_0)\right)\\ &= \frac{1}{3q^{2a}} + \frac{2}{q^{2a+1}}\Re\left(\sum_{\substack{u_0,v_0=0\\ v_0\geq u_0}}^{q-1}(v_0-u_0)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u_0\ominus_\varphi v_0}{q}\right)\right),\end{aligned}$$

as claimed. (Note the difference from (A.2) in the summation range of v_0.) ∎

Now define

$$\widehat{r}_q(w,\gamma,k) := \begin{cases} 1+\gamma(w^2-w+\frac{1}{3}) & \text{if } k=0,\\ -\frac{\gamma}{2}\left(\frac{1}{3q^{2a}} + \frac{2}{q^{2a}}\Re\left(\sum\limits_{\substack{u,v=0\\ v\geq u}}^{q-1}\frac{(v-u)\mathrm{wal}_{\kappa_{a-1}}\left(\frac{u\ominus_\varphi v}{q}\right)}{q}\right)\right) & \text{if } k>0,\end{cases}$$

and for $\boldsymbol{w}=(w_1,\dots,w_s)$, $\boldsymbol{k}=(k_1,\dots,k_s)$ and $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\dots)$ define

$$\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k}) := \prod_{j=1}^{s}\widehat{r}_q(w_j,\gamma_j,k_j).$$

Then we obtain the following theorem.

THEOREM A.2. *The digital shift invariant kernel for the reproducing kernel $K_{\mathrm{sob},s,\boldsymbol{w},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y})$, where the digital shift is taken in prime-power base q and with respect to the bijection φ, is given by*

$$K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^s}\widehat{r}_q(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y})}.$$

Appendix B. Some other useful results. Here we prove two results which are used in section 3.

LEMMA B.1. *With $T_\kappa(u,v) := (v-u)\mathrm{wal}_\kappa((u\ominus_\varphi v)/q)$, for $l\in\{0,\dots,q-1\}$ we have*

$$\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\mathrm{wal}_\kappa\left(\frac{l}{q}\right)T_\kappa(u,v) = q\left(\frac{1-q^2}{6} + \sum_{\substack{u=0,\\ u<u\oplus_\varphi l}}^{q-1}(u\oplus_\varphi l) - u\right)$$

and

$$\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}\mathrm{wal}_\kappa\left(\frac{l}{q}\right)\overline{T_\kappa(u,v)} = q\left(\frac{1-q^2}{6} + \sum_{\substack{u=0,\\ u<u\ominus_\varphi l}}^{q-1}(u\ominus_\varphi l) - u\right).$$

Proof. For the first sum we have

$$\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\geq u}}^{q-1}(v-u)\mathrm{wal}_\kappa\left(\frac{l\oplus_\varphi u\ominus_\varphi v}{q}\right) = \sum_{\substack{u,v=0\\ v\geq u}}^{q-1}(v-u)\sum_{\kappa=1}^{q-1}\mathrm{wal}_\kappa\left(\frac{l\oplus_\varphi u\ominus_\varphi v}{q}\right),$$

where the right sum can be simplified such that the last line is equal to

$$\begin{aligned}
&\sum_{\substack{u,v=0\\ v\ge u}}^{q-1}(v-u)\times\begin{cases}q-1 & \text{if } v=u\oplus_\varphi l,\\ -1 & \text{else,}\end{cases}\\
=&\sum_{u=0}^{q-1}\sum_{v=u}^{q-1}-(v-u)+q\sum_{\substack{u,v=0\\ v\ge u, v=u\oplus_\varphi l}}^{q-1}(v-u)=\sum_{d=0}^{q-1}-(q-d)d+q\sum_{\substack{u=0\\ u<u\oplus_\varphi l}}^{q-1}((u\oplus_\varphi l)-u)\\
=&q\left(\frac{-(q^2-1)}{6}+\sum_{\substack{u=0\\ u<u\oplus_\varphi l}}^{q-1}((u\oplus_\varphi l)-u)\right).
\end{aligned}$$

The second part follows from the first by

$$\begin{aligned}
&\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\ge u}}^{q-1}\mathrm{wal}_\kappa\left(\frac{l}{q}\right)\overline{T_\kappa\Big(u,v\Big)}\\
=&\overline{\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\ge u}}^{q-1}\overline{\mathrm{wal}_\kappa\left(\frac{l}{q}\right)}T_\kappa\Big(u,v\Big)}=\overline{\sum_{\kappa=1}^{q-1}\sum_{\substack{u,v=0\\ v\ge u}}^{q-1}\mathrm{wal}_\kappa\left(\frac{\ominus_\varphi l}{q}\right)T_\kappa\Big(u,v\Big)},
\end{aligned}$$

which finishes the proof. □

The next lemma was used in section 3 to give an explicit computable representation of the shift invariant kernel $K_{\mathrm{ds},q,\boldsymbol{\gamma},\boldsymbol{w},\varphi}$.

LEMMA B.2. *If $x\neq y$,*

$$\begin{aligned}
&\sum_{k=1}^{\infty}\frac{-\tau(k)}{2}\mathrm{wal}_k(x)\overline{\mathrm{wal}_k(y)}=\frac{1}{6}-\frac{1}{2q^{i_0+1}}\\
&\times\left(\sum_{\substack{u=0,\\ u<u\oplus_\varphi x_{i_0}\ominus_\varphi y_{i_0}}}^{q-1}(u\oplus_\varphi x_{i_0}\ominus_\varphi y_{i_0}-u)+\sum_{\substack{u=0,\\ u<u\oplus_\varphi y_{i_0}\ominus_\varphi x_{i_0}}}^{q-1}(u\oplus_\varphi y_{i_0}\ominus_\varphi x_{i_0}-u)\right)
\end{aligned}$$

with x_{i_0}, y_{i_0} denoting the i_0th fractional q-adic digits of x and y, where i_0 is the smallest index such that the digits differ. For $x=y$ the sum is equal to $1/6$. (Note that for $k>0$ we have $\widehat{r}_q(w,\gamma,k)=-\frac{\gamma}{2}\tau(k)$.)

Proof. Let $D_{a,\kappa}$ denote $D_{a,\kappa}=\sum_{k=\kappa q^{a-1}+\dots}\mathrm{wal}_k$, where the sum ranges over all k with the leading term κq^{a-1} in the q-adic expansion. By the character properties of the Walsh function set we have

$$D_{a,\kappa}=\mathrm{wal}_{\kappa q^{a-1}}\sum_{0\le k<q^{a-1}}\mathrm{wal}_k=\mathrm{wal}_{\kappa q^{a-1}}\cdot q^{a-1}\cdot 1_{[0,q^{-(a-1)})}.$$

First, let $x\neq y$, $i_0=\max(\{i: x_j=y_j \text{ for all } j=0,\dots,i\ge 0\})+1$, where x_i,y_i are the q-adic digits of x and y. Then, with $T_\kappa(u,v)=(v-u)\mathrm{wal}_\kappa((u\ominus_\varphi v)/q)$ as above,

and since $\tau(k)$ depends only on the q-adic length a and most significant digit κ,

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{-\tau(k)}{2} \mathrm{wal}_k(x)\overline{\mathrm{wal}_k(y)} = \sum_{a=1}^{\infty}\sum_{\kappa=1}^{q-1} \frac{-\tau(\kappa q^{a-1})}{2} D_{a,\kappa}(x \ominus_\varphi y)\\
&= \sum_{a=1}^{\infty}\sum_{\kappa=1}^{q-1} \frac{1}{q^{2a}} \Re\left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} T_\kappa(u,v) - \frac{1}{6}\right) D_{a,\kappa}(x \ominus_\varphi y)\\
&= \sum_{a=1}^{i_0-1}\sum_{\kappa=1}^{q-1} \frac{1}{q^{2a}} \Re\left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} T_\kappa(u,v) - \frac{1}{6}\right) D_{a,\kappa}(x \ominus_\varphi y)\\
&\quad + \sum_{\kappa=1}^{q-1} \frac{1}{q^{2i_0}} \Re\left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} T_\kappa(u,v) - \frac{1}{6}\right) D_{i_0,\kappa}(x \ominus_\varphi y)\\
&= \sum_{a=1}^{i_0-1}\sum_{\kappa=1}^{q-1} \frac{1}{q^{a+1}} \Re\left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} T_\kappa(u,v) - \frac{1}{6}\right)\\
&\quad + \sum_{\kappa=1}^{q-1} \frac{1}{2q^{i_0+1}} \left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} \mathrm{wal}_{\kappa q^{i_0-1}}(x \ominus_\varphi y) T_\kappa(u,v) - \frac{\mathrm{wal}_{\kappa q^{i_0-1}}(x \ominus_\varphi y)}{6}\right)\\
&\quad + \sum_{\kappa=1}^{q-1} \frac{1}{2q^{i_0+1}} \left(\frac{-1}{q}\sum_{\substack{u,v=0,\\ v\geq u}}^{q-1} \mathrm{wal}_{\kappa q^{i_0-1}}(x \ominus_\varphi y) \overline{T_\kappa(u,v)} - \frac{\mathrm{wal}_{\kappa q^{i_0-1}}(x \ominus_\varphi y)}{6}\right)\\
&= \sum_{a=1}^{i_0-1} \frac{1}{q^{a+1}} \left(\frac{q^2-1}{6} - \frac{q-1}{6}\right) + \frac{1}{2q^{i_0+1}} \left(2 \cdot \frac{q^2-1}{6}\right.\\
&\quad \left. - \left(\sum_{\substack{u=0,\\ u<u\oplus_\varphi x_{i_0}\ominus_\varphi y_{i_0}}}^{q-1} (u \oplus_\varphi x_{i_0} \ominus_\varphi y_{i_0} - u) + \sum_{\substack{u=0,\\ u<u\oplus_\varphi y_{i_0}\ominus_\varphi x_{i_0}}}^{q-1} (u \oplus_\varphi y_{i_0} \ominus_\varphi x_{i_0} - u)\right) - 2 \cdot \frac{-1}{6}\right)\\
&= \frac{1}{6} - \frac{1}{2q^{i_0+1}} \left(\sum_{\substack{u=0,\\ u<u\oplus_\varphi x_{i_0}\ominus_\varphi y_{i_0}}}^{q-1} (u \oplus_\varphi x_{i_0} \ominus_\varphi y_{i_0} - u) + \sum_{\substack{u=0,\\ u<u\oplus_\varphi y_{i_0}\ominus_\varphi x_{i_0}}}^{q-1} (u \oplus_\varphi y_{i_0} \ominus_\varphi x_{i_0} - u)\right)
\end{aligned}$$

by Lemma B.1.

If $x = y$, it is easy to see (e.g., by letting $i_0 \to \infty$ in the above term) that the second term vanishes. □

Acknowledgments. The authors would like to thank the anonymous referees for their valuable comments and suggestions, which were very helpful in improving the quality and readability of the paper.

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] J. Dick, *On the convergence rate of the component-by-component construction of good lattice rules*, J. Complexity, 20 (2004), pp. 493–522.

[3] J. Dick and F. Pillichshammer, *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity, 21 (2005), pp. 149–195.

[4] J. Dick, F. Y. Kuo, F. Pillichshammer, and I. H. Sloan, *Construction algorithms for polynomial lattice rules for multivariate integration*, Math. Comp., 74 (2005), pp. 1895–1921.

[5] S. Heinrich, *Some open problems concerning the star-discrepancy*, J. Complexity, 19 (2003), pp. 416–419.

[6] F. Y. Kuo, *Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces*, J. Complexity, 19 (2003), pp. 301–320.

[7] G. Larcher, *Digital point sets: Analysis and application*, in Random and Quasi-Random Point Sets, P. Hellekalek and G. Larcher, eds., Lecture Notes in Statist. 138, Springer, New York, 1998, pp. 167–222.

[8] G. Larcher, H. Niederreiter, and W. Ch. Schmid, *Digital nets and sequences constructed over finite rings and their application to quasi-Monte Carlo integration*, Monatsh. Math., 121 (1996), pp. 231–253.

[9] G. Larcher and G. Pirsic, *Base change problems for generalized Walsh series and multivariate numerical integration*, Pacific J. Math., 189 (1999), pp. 75–105.

[10] H. Niederreiter, *Point sets and sequences with small discrepancy*, Monatsh. Math., 104 (1987), pp. 273–337.

[11] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Ser. Appl. Math. 63, SIAM, Philadelphia, 1992.

[12] H. Niederreiter, *Low-discrepancy point sets obtained by digital constructions over finite fields*, Czechoslovak Math. J., 42 (1992), pp. 143–166.

[13] H. Niederreiter, *Some current issues in quasi-Monte Carlo methods*, J. Complexity, 19 (2003), pp. 428–433.

[14] H. Niederreiter, *Digital nets and coding theory*, in Coding, Cryptography and Combinatorics, K. Q. Feng, H. Niederreiter, and C. P. Xing, eds., Birkhäuser, Basel, 2004, pp. 247–257.

[15] G. Pirsic, *Embedding Theorems and Numerical Integration of Walsh Series over Groups*, Ph.D. thesis, University of Salzburg, 1997; also available online at http://www.ricam.oeaw.ac.at/people/page/pirsic/.

[16] G. Pirsic, *A small taxonomy of integration node sets*, Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber.II, submitted.

[17] W. Ch. Schmid and G. Pirsic, *Calculation of the quality parameter of digital nets and application to their construction*, J. Complexity, 17 (2001), pp. 827–839.

[18] I. H. Sloan, F. Y. Kuo, and S. Joe, *Constructing randomly shifted lattice rules in weighted Sobolev spaces*, SIAM J. Numer. Anal., 40 (2002), pp. 1650–1655.

[19] I. H. Sloan, F. Y. Kuo, and S. Joe, *On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces*, Math. Comp., 71 (2002), pp. 1609–1640.

[20] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.

SIAM J. NUMER. ANAL.
Vol. 44, No. 1, pp. 412–448

DISCRETE INGHAM INEQUALITIES AND APPLICATIONS*

MIHAELA NEGREANU† AND ENRIQUE ZUAZUA‡

Abstract. In this paper we prove a discrete version of the classical Ingham inequality for nonharmonic Fourier series whose exponents satisfy a gap condition. Time integrals are replaced by discrete sums on a discrete mesh. We prove that, as the mesh becomes finer and finer, the limit of the discrete Ingham inequality is the classical continuous one. This analysis is partially motivated by control-theoretical issues. As an application we analyze the control/observation properties of numerical approximation schemes of the 1-d wave equation. The discrete Ingham inequality provides observability and controllability results which are uniform with respect to the mesh-size in suitable classes of numerical solutions in which the high frequency components have been filtered. We also discuss the optimality of these results in connection with the dispersion diagrams of the numerical schemes.

Key words. wave equation, numerical approximation schemes, nonharmonic analysis, discrete Fourier transform, Ingham inequalities, observability, controllability, dispersion, group velocity

AMS subject classifications. 42C99, 65T50, 65M06, 65N06

DOI. 10.1137/050630015

1. Introduction.

Families of "nonharmonic" exponentials $\left\{e^{i\lambda_k t}\right\}$ appear in various fields of mathematics and signal processing. One of the central problems arising in all of these applications is the question of the Riesz basis property.

The following inequality for nonharmonic Fourier series due to Ingham is well known (see [9] and [26, p. 162]): *Assume that the strictly increasing sequence* $\{\lambda_k\}_{k\in\mathbb{Z}}$ *of real numbers satisfies the "gap" condition*

$$\lambda_{k+1} - \lambda_k \geq \gamma \quad \text{for all } k \in \mathbb{Z}, \tag{1.1}$$

for some $\gamma > 0$. *Then, for all* $T > 2\pi/\gamma$ *there exist two positive constants* C_1, C_2 *depending only on* γ *and* T *such that*

$$C_1(T,\gamma) \sum_{k=-\infty}^{\infty} |a_k|^2 \leq \int_0^T \left| \sum_{k=-\infty}^{\infty} a_k e^{it\lambda_k} \right|^2 dt \leq C_2(T,\gamma) \sum_{k=-\infty}^{\infty} |a_k|^2, \tag{1.2}$$

for every complex sequence $(a_k)_{k\in\mathbb{Z}} \in \ell^2$, *where*

$$C_1(T,\gamma) = \frac{2T}{\pi}\left(1 - \frac{4\pi^2}{T^2\gamma^2}\right) > 0, \tag{1.3}$$

$$C_2(T,\gamma) = \frac{8T}{\pi}\left(1 + \frac{4\pi^2}{T^2\gamma^2}\right) > 0, \tag{1.4}$$

*Received by the editors April 26, 2005; accepted for publication (in revised form) September 26, 2005; published electronically March 7, 2006. This research was supported by grant BFM2002-03345 of the Spanish MCYT and the network "New materials, adaptive systems and their nonlinearities: modelling, control and numerical simulation" of the EU.

http://www.siam.org/journals/sinum/44-1/63001.html

†Departamento de Álgebra, Facultad de Matemática, Universidad Complutense de Madrid, 28040 Madrid, Spain (mihaela_negreanu@mat.ucm.es).

‡Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain (enrique.zuazua@uam.es).

and ℓ^2 is the Hilbert space of square summable sequences,

$$\ell^2 = \left\{ \{a_k\} \ : \ \|a_k\|_{\ell^2}^2 = \sum_{k\in\mathbb{N}} |a_k|^2 < \infty \right\}. \tag{1.5}$$

This result shows that the sequence of exponentials $\left\{e^{i\lambda_k t}\right\}$ forms a Riesz basis of its span for $T > 2\pi/\gamma$ (see [26, Chapter 3, p. 112]).

As we have mentioned above, one of the main applications of Ingham's inequality and its variants is the control of wave-like equations and other closely related problems like observability or inverse problems. The problem of observability for wave equations consists of analyzing whether the energy of the waves propagating in a domain with suitable boundary conditions can be estimated in terms of the energy concentrated on a given subregion of the domain (or its boundary) where propagation occurs in a given time interval. On the other hand, the goal in controllability problems is to drive the solutions of a given dynamical system (continuous or discrete) to a given state at a given final time by means of a control acting on the system on that subregion (or its boundary). It is well known that the two problems are equivalent provided one chooses an appropriate functional setting, which depends on the equation (see, for instance, [17]).

In the context of partial differential equations, using the Fourier representation of the solutions, the problem of observability can be reduced to an application of Ingham's inequality in which the sequence $\{\lambda_k\}$ is constituted by the spectrum of the generator of the underlying semigroup. However, the gap condition (1.1) that is required to apply Ingham's inequality often limits the range of applicability of this technique to 1-d problems like strings and beams. This has led to a significant number of controllability results (see [15]) and also to far reaching generalizations of the Ingham theorem under weakened gap conditions (see [2], [4], [5], [8], [12], [13]). The most complete result in this direction has been obtained independently by Baiocchi, Komornik, and Loreti in [2], [3], [4], and Avdonin and Moran in [1].

In the numerical analysis of those observability inequalities and for studying the controllability properties of numerical schemes the need of a discrete version of this inequality arises naturally (see [7], [19], [20], [21]). This paper is devoted to proving a discrete version of that Ingham inequality.

The inequality we prove is uniform with respect to the mesh-size Δt in the time-discretization and, in the limit as $\Delta t \to 0$, yields the classical Ingham inequality above.

The discrete Ingham inequality we prove is the natural tool to prove observability/controllability properties for fully discrete schemes for the approximation of the 1-d wave equation and other closely related models (vibrating beams, Schrödinger equation, etc.) and to show that the controls of the limiting continuous model are the limit of the controls of the full discrete schemes. However, it is important to recall that, as it is by now well known [28], numerical approximation schemes often introduce spurious high frequency solutions that may be an obstacle for uniform (with respect to the mesh-size) observability/controllability results. Thus, one often needs to filter or cut-off those spurious numerical solutions. Our generalization of Ingham's inequality to the discrete context explains how this filtering has to be done in order to guarantee uniform results.

As an example of application of our discrete Ingham inequality we perform the analysis of the observability/controllability properties of the most standard centered fully discrete schemes for the wave equation.

The main reason for the lack of uniform observability/controllability of the numerical high frequency spurious solutions, is that they generate high frequency wave packets for which the group velocity is of the order of the mesh-size ([28]). Thus, as the mesh-size tends to zero, since the velocity becomes smaller and smaller, the time for observability/controllability increasing in a divergent way. This fact is related to the dispersion diagram associated to the numerical approximation scheme, since, roughly, the slope of the dispersion diagram is the group velocity of propagation of wave packets and also coincides with the spectral gap. Part of this article is devoted to explaining the connections of these notions and to show how combining the qualitative information that the dispersion diagram provides with the discrete Ingham inequality, one can get precise information on how the filtering should be implemented, if needed.

As proved in the original article by Ingham (see [9, p. 368]), an L^1-version of inequality (1.2) also holds. More precisely, for every increasing sequence $\{\lambda_k\}_{k\in\mathbb{Z}}$ of real numbers satisfying the "gap" condition (1.1) we also have

$$C_1(T,\gamma)\,|a_k| \le \int_0^T \left| \sum_{k=-\infty}^{\infty} a_k e^{it\lambda_k} \right| dt \le C_2(T,\gamma)\,|a_k| \quad \text{for all } k \in \mathbb{Z} \tag{1.6}$$

for all $T > 2\pi/\gamma$.

In this paper we also prove a discrete version of this inequality.

Our proofs are strongly inspired in that by Ingham (see also [26]), which is based on the use of a suitable cut-off, nonnegative function, with compact support on the time interval $(0,T)$ and whose Fourier transform is "concentrated" around $\tau = 0$. We use the same function in the physical space, but its Fourier transform has to be replaced by the discrete one. One of the key points in the proof is a careful comparison between the continuous and discrete transforms of this weight function. This is done by using a key result by N. Trefethen [23].

This paper is organized as follows: in section 2 we state our discrete Ingham inequality (see Theorem 2.1), we analyze the necessity of its hypotheses and compare both the continuous and discrete inequalities. We also formulate a discrete version of the L^1 analogue (1.6) (see Theorem 2.2). In section 3 we discuss the application of this result to the study of the properties of the solutions of fully discrete approximations of the wave equation. In section 4 the controllability problem for the discrete system is addressed and the main results of existence, characterization, and convergence of the discrete controls are presented and proved. In section 5 we discuss these results in connection with the dispersion diagrams of the discrete equations under consideration. Finally, section 6 is devoted to proving the discrete Ingham inequality and its discrete L^1 version.

The discrete Ingham inequality we present in this paper has been announced in [21].

2. Main results. The main result of this paper is as follows.

THEOREM 2.1 (discrete Ingham inequality). *Let $\{\lambda_k\}_{k\in\mathbb{Z}}$ be an increasing sequence of real numbers satisfying for some $\gamma > 0$ the "gap" condition*

$$\lambda_{k+1} - \lambda_k \ge \gamma > 0 \quad \text{for all } k \in \mathbb{Z}. \tag{2.1}$$

Let $T > 0$ and $0 < \Delta t \le 1$. Assume that $\{\lambda_k\}_{|k|\le N}$ satisfies the additional condition

$$|\lambda_k - \lambda_l| \le \frac{2\pi - (\Delta t)^p}{\Delta t} \quad \text{for all } |k| \le N,\ |l| \le N, \text{ for some } 0 \le p < 1/2, \tag{2.2}$$

where $2N \leq M$ *and* $M = [T/\Delta t - 1]$. *Then, there exists a positive number* $\epsilon(\Delta t)$ *such that, for all* $T > T_0(\Delta t) := 2\pi/\gamma + \epsilon(\Delta t)$, *there exist two positive constants* $C_j(\Delta t, T, \gamma) > 0$, $j = 1, 2$, *such that*

(2.3)
$$C_1(\Delta t, T, \gamma) \sum_{k=-N}^{N} |a_k|^2 \leq \Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} \right|^2 \leq C_2(\Delta t, T, \gamma) \sum_{k=-N}^{N} |a_k|^2,$$

for every complex sequence $(a_k)_{k\in\mathbb{Z}} \in \ell^2$.

Moreover, if γ *and* p *in* (2.1) *and* (2.2) *are kept fixed, then* $\epsilon(\Delta t) = o(\Delta t)^{1-2p}$ *and the constants in* (2.3) *satisfy*

(2.4)
$$C_j(\Delta t, T, \gamma) = C_j(T, \gamma) + \delta_j(\Delta t), \quad j = 1, 2, \quad \text{with } \delta_1(\Delta t) \leq 0 \text{ and } \delta_2(\Delta t) \geq 0,$$

where $C_j(T, \gamma)$, $j = 1, 2$, *are the Ingham constants in* (1.3) *and* (1.4) *and* $\lim_{\Delta t\to 0} \delta_j(\Delta t) = 0$, $j = 1, 2$.

Concerning the L^1-version of Ingham inequality in (1.6), the following theorem holds.

THEOREM 2.2. *Under the hypotheses of Theorem* 2.1 *we also have the following discrete version of* (1.6):

(2.5)
$$C_1(\Delta t, T, \gamma)\,|a_k| \leq \Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} \right| \leq C_2(\Delta t, T, \gamma)\,|a_k| \quad \text{for all } |k| \leq N.$$

As in Theorem 2.1, *the time* T *and the constants in this inequality remain uniform as* $\Delta t \to 0$ *and converge to those of the continuous Ingham inequality* (1.6).

Remark 2.3. Condition $T > 2\pi/\gamma$ is optimal for the classical Ingham inequality (see [26, p. 163]). In this sense, the condition $T > 2\pi/\gamma + \epsilon(\Delta t)$ in Theorem 2.1 is asymptotically optimal since $\epsilon(\Delta t) \to 0$ as $\Delta t \to 0$.

It is important to emphasize that the time T and the constants C_j, $j = 1, 2$ in (2.3) are uniform in Δt. This is essential for the applications in numerical analysis in which $\Delta t \to 0$. The uniformity may be guaranteed because of the assumptions (2.1)–(2.2) on the sequence $\{\lambda_k\}_k$.

More precisely, when comparing the continuous and discrete inequalities, the following can be said:

- In both continuous and discrete cases, the sequence $\{\lambda_k\}_k$ is required to satisfy the so-called *gap condition* (2.1).
- The restriction (2.2) imposed on $\{\lambda_k\}_k$ in Theorem 2.1 is not needed in the classical continuous Ingham inequality (1.2).
- It is easy to see that, for every $N \in \mathbb{N}$ fixed, if we pass to the limit $\Delta t \to 0$ in (2.3), we get the classical Ingham inequality (1.2). Indeed, for (1.2) to be true for all sequences $(a_k)_{k\in\mathbb{Z}} \in \ell^2$ it is sufficient, by density, to prove it for sequences with only a finite number of nonzero components.
  In that case (1.2) is the limit of (2.3) because of the convergence of the minimal time T and the constants C_j, $j = 1, 2$, in (2.3) to those of (1.2).
  We also have a discrete Ingham inequality (2.3) for every sequence $(\lambda_k)_k$ verifying conditions (2.1) and (2.2), with $0 \leq p \leq 1$. But, if $p \geq 1/2$,

$\varepsilon(\Delta t) = o(\Delta t)^{1-2p} \to \infty$, so $T_0(\Delta t) \to \infty$, and this makes it of little use in practice because we are looking for a uniform (with respect to Δt) time T.

On the other hand, the restriction $2N \leq M$, with $M = [T/\Delta t - 1]$, is sharp. Indeed, when $2N > M$ one can find nontrivial values of the coefficients $\{a_k\}_k$ such that

$$\sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} = 0, \quad 0 \leq n \leq M \tag{2.6}$$

and

$$\sum_{k=-N}^{N} |a_k|^2 \neq 0. \tag{2.7}$$

Observe that (2.6) is a system of $M+1$ homogeneous linear equations with $2N+1$ unknown quantities a_k. If $2N > M$, this system necessarily has nontrivial solutions. This is in agreement with common sense. Indeed, in view of the fact that we only make $M+1$ measurements for $n = 0, \ldots, M$ one cannot expect to recover more than $M+1$ coefficients of the solution.

When $2N \leq M$, (2.6)–(2.7) do not hold. However if $\lambda_k - \lambda_l \in 2\pi\mathbb{Z}/\Delta t$ for certain values of k and l with $k \neq l$ the sequence $a_k = -a_l = 1$, $a_n = 0$, $n \neq k$, and l satisfies (2.6). Then, an inequality of type (2.3) is impossible. So, it is natural to impose on the sequence $\{\lambda_k\}_k$ the condition $\lambda_k - \lambda_l \notin 2\pi\mathbb{Z}/\Delta t$ for a discrete Ingham inequality (2.3) to hold.

In fact, to avoid *aliasing* one has to restrict the increasing sequence of real numbers $\{\lambda_k\}_k$ to be such that $\lambda_k - \lambda_l \in [2\pi m/\Delta t, 2\pi(m+1)/\Delta t]$, for some $m \in \mathbb{Z}$. Therefore, it is natural to impose the condition

$$|\lambda_k - \lambda_l| < \frac{2\pi}{\Delta t}.$$

In our theorem this latter condition is implied by the stronger one, (2.2), which is needed for the uniform estimates in (2.3) to hold. More precisely, the restriction $0 \leq p < 1/2$ in (2.2) is needed to guarantee the asymptotically optimal time $T > 2\pi/\gamma + \epsilon(\Delta t)$, with $\epsilon(\Delta t) \to 0$ as $\Delta t \to 0$ since $\epsilon(\Delta t) = o(\Delta t)^{1-2p}$.

Remark 2.4. The condition $T > T_0(\Delta t)$ is necessary for the proof of the first inequality in (2.3) and in (1.6) (to have $C_1(\Delta t, T, \gamma) > 0$). The second inequality in (2.3) and (1.6), respectively, holds for all $T > 0$. In this respect the situation is the same as for the continuous inequalities (1.2).

3. Application to the uniform observability of the full discretizations of the 1-d wave equation.

3.1. The wave equation. This section is motivated by the classical problem of control of waves. More precisely, it is related with the controllability of the 1-d wave equation: given $T > 0$ and $(u^0, u^1) \in L^2(0,1) \times H^{-1}(0,1)$, the problem is to find a control function $v \in L^2(0,T)$ such that the solution of the system

$$\begin{cases} u_{tt} - u_{xx} = 0, & 0 < x < 1, \quad 0 < t < T, \\ u(0,t) = 0, \quad u(1,t) = v(t), & 0 < t < T, \\ u(x,0) = u^0(x), \qquad u_t(x,0) = u^1(x), & 0 < x < 1, \end{cases} \tag{3.1}$$

satisfies

$$u(T) = u_t(T) = 0, \qquad 0 < x < 1. \tag{3.2}$$

This property is well known to be true for $T \geq 2$. This problem has been studied and solved in a much more general setting and, in particular, for multidimensional wave equations [17]. Several approaches to the problem have been developed. In particular, the Hilbert uniqueness method (HUM) introduced by Lions in [17] offers a general way of reducing the problem to the so-called *observability problem* for the adjoint (up to an inversion in time) wave equation in the absence of control:

$$\begin{cases} \phi_{tt} - \phi_{xx} = 0, & 0 < x < 1, \quad 0 < t < T, \\ \phi(0,t) = \phi(1,t) = 0, & 0 < t < T, \\ \phi(x,0) = \phi^0(x), \qquad \phi_t(x,0) = \phi^1(x), & 0 < x < 1. \end{cases} \tag{3.3}$$

It is well known that the energy

$$E(t) = \frac{1}{2} \int_0^1 \left(\mid \phi_x(x,t) \mid^2 + \mid \phi_t(x,t) \mid^2 \right) dx \tag{3.4}$$

of the solutions of (3.3) satisfies

$$\frac{dE(t)}{dt} = E'(t) = 0 \quad \text{for all } t \in [0,T]$$

and therefore is conserved in time.

The observability problem is as follows: *To find $T > 0$ such that there exists a constant $C(T) > 0$ for which*

$$E(0) \leq C(T) \int_0^T \mid \phi_x(1,t) \mid^2 dt \tag{3.5}$$

holds for every solution of (3.3).

HUM allows showing that, once the observability inequality (3.5) is satisfied for the adjoint system (3.3), system (3.1) is controllable in time T. Moreover, HUM provides a systematic method to build the control $v = \phi_x(1,t)$ of minimal $L^2(0,T)$-norm.

In the context of the 1-d wave equation (3.3), inequality (3.5) can be easily proved by several methods including *Fourier series, D'Alembert Formula*, *multiplier techniques*, and *Ingham's theorem* (1.2), provided $T \geq 2$.

In order to solve the problem (3.5) applying the classical Ingham inequality, one uses Fourier series techniques. Indeed, the solution of (3.3) admits the Fourier development

$$\phi(x,t) = \sum_{k \in \mathbb{Z} \setminus \{0\}} a_k e^{i\lambda_k t} \varphi_k(x), \tag{3.6}$$

with $\{\lambda_k\}_k$, $\lambda_k = k\pi = -\lambda_{-k}$, $k > 0$, being the sequence of eigenvalues of the system, $\varphi_k(x) = \sin(k\pi x)$, the corresponding eigenfunctions and $a_k \in \mathbb{C}$ the Fourier coefficients, which can be computed explicitly in terms of the initial data in (3.3).

By definition (3.4) of the conserved energy of the solution ϕ of (3.3) given by (3.6), we have

$$E_\phi = \frac{1}{2} \sum_{k \in \mathbb{Z} \setminus \{0\}} k^2 \pi^2 \left| a_k \right|^2. \tag{3.7}$$

On the other hand, in view of the explicit form of $\phi_x(1,t)$, inequality (3.5) may be written as:

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} k^2\pi^2 \left|a_k\right|^2 \leq C(T)\int_0^T \left|\sum_{k\in\mathbb{Z}\setminus\{0\}} (-1)^k k\pi a_k e^{i\lambda_k t}\right|^2 dt. \tag{3.8}$$

According to Ingham's inequality (1.2), (3.8) holds for $T > 2$, since the gap of the sequence $\{\lambda_k\}_k$ is constant, $\gamma = \pi$, and, consequently, the minimal observability time is $2\pi/\gamma = 2$. In this particular case the inequality holds also for the minimal time $T = 2$. This is due to the orthogonality properties of the trigonometric polynomials. But, in general, i.e., for a general sequence $(\lambda_k)_{k\in\mathbb{Z}}$ satisfying the gap condition (1.1), it is well known that the Ingham inequality (1.2) may fail for the minimal time $T = 2\pi/\gamma$ (see [26, p. 163]).

In order to obtain numerical approximations of the controls, it is natural to analyze the controllability and observability properties of numerical approximation schemes. We first recall some well-known facts about the space semi-discretization schemes to later address space-time discretizations.

3.2. Space semi-discretizations. First, we consider the semi-discrete version of the observability problem (3.5): Take $N \in \mathbb{N}$, set $h = 1/(N+1)$ and consider the finite-difference space semi-discretization of (3.3):

$$\begin{cases} \phi_j'' = \dfrac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{h^2}, & t > 0, \quad j = 1,\dots,N, \\ \phi_0 = \phi_{N+1} = 0, & t > 0, \\ \phi(0) = \phi_{0,j}, \quad \phi_j'(0) = \phi_{1,j}, & j = 1,\dots,N. \end{cases} \tag{3.9}$$

The energy of system (3.9) is given by

$$E_h(t) = \frac{h}{2}\sum_{j=1}^{N} |\ \phi_j(t)\ |^2 + \frac{h}{2}\sum_{j=0}^{N} \frac{|\ \phi_{j+1}(t) - \phi_j(t)\ |^2}{h^2} \tag{3.10}$$

and it is also conserved in time.

The semi-discrete version of (3.5) is

$$E_h(0) \leq C\int_0^T \left|\frac{\phi_N(t)}{h}\right|^2 dt. \tag{3.11}$$

More precisely, one seeks for a positive constant $C > 0$ such that (3.11) holds.
The corresponding eigenvalue problem is of the form

$$\begin{cases} -[\varphi_{k+1} + \varphi_{k-1} - 2\varphi_k]/h^2 = \lambda^2\varphi_k, \quad k = 1,\dots,N, \\ \varphi_0 = \varphi_{N+1} = 0. \end{cases} \tag{3.12}$$

The eigenvalues and eigenvectors of (3.12) may be computed explicitly (see [10, p. 456]); one then has

$$\begin{cases} \lambda_k^2(h) = \dfrac{4}{h^2}\sin^2\left(\dfrac{\pi k h}{2}\right), & k = 1,\dots,N, \\ \bar\varphi_k \equiv (\varphi_{k,1},\dots,\varphi_{k,N});\ \varphi_{k,j} = \sin(k\pi j h), & j,k = 1,\dots,N. \end{cases} \tag{3.13}$$

The solutions of (3.9) in Fourier series are

$$\bar{\phi} = \sum_{k=-N, k\neq 0}^{N} a_k e^{i\lambda_k(h)t}\bar{\varphi}_k, \tag{3.14}$$

where $\bar{\phi} = (\phi_1, \dots, \phi_N)$.

As pointed out in [11], (3.11) holds for all $T > 0$ and $h > 0$, but, the observability constant in (3.11) may not remain uniformly bounded as $h \to 0$, for any $T > 0$. More precisely,

$$\sup_{\bar{\phi}\in\mathcal{S}_h}\left[\frac{E_h(0)}{\int_0^T \mid \phi_N(t)/h \mid^2 dt}\right] \to \infty, \ as \ h \to 0, \tag{3.15}$$

where $\mathcal{S}_h$ is the set of all solutions of (3.9). This is due to the pathological behavior of the high frequency numerical solutions.

In the light of Ingham's inequality (1.2), the lack of uniform observability as h tends to zero may be explained because of the lack of gap between consecutive eigenvalues (see [11], [28]). In particular, the gap between the largest eigenvalues entering in the Fourier development of the solution of (3.9) may be bounded above as follows:

$$\lambda_N(h) - \lambda_{N-1}(h) \leq \frac{3\pi^2 h}{2} \to 0, \quad \text{as} \quad h \to 0. \tag{3.16}$$

As it was proved in [11], a suitable cut-off or filtering of the spurious numerical high frequencies may be a good cure for these pathologies. Given $0 < \alpha < 1$, we introduce the following classes of filtered solution of (3.9):

$$\mathcal{C}_\alpha(h) = \left\{\bar{\phi} \text{ sol. of (3.9)} \ : \bar{\phi} = \sum_{|k|\leq \alpha N, \ k\neq 0} a_k e^{i\lambda_k t}\bar{\varphi}_k\right\}. \tag{3.17}$$

In the class $\mathcal{C}_\alpha(h)$ the high frequencies corresponding to the indexes $j > \alpha N$ have been cut-off. This guarantees a uniform gap condition

$$\lambda_{k+1}(h) - \lambda_k(h) \geq \pi \cos\left(\frac{\pi\alpha}{2}\right), \quad \text{for} \quad k \leq \alpha/h. \tag{3.18}$$

Consequently, applying Ingham's inequality, we may deduce the uniform observability in the class $\mathcal{C}_\alpha(h)$ for

$$T > T(\alpha) = 2/\cos(\pi\alpha/2). \tag{3.19}$$

Let us explain this in more detail.

By definition (3.10) of the conserved energy and taking into account the orthogonality properties of the eigenvectors (see [11], [20]), we have

$$E_h = \frac{1}{4}\sum_{k=-\alpha N, k\neq 0}^{\alpha N} |a_k|^2 \left(1 + \lambda_k^2(h)\right). \tag{3.20}$$

Then, inequality (3.11) in the class $\mathcal{C}_\alpha(h)$ may be rewritten as

$$\sum_{k=-\alpha N, k\neq 0}^{\alpha N} |a_k|^2 \left(1 + \lambda_k^2(h)\right) \leq C(T)\int_0^T \left|\sum_{k=-\alpha N, k\neq 0}^{\alpha N} \frac{\sin(Nk\pi h)}{h} a_k e^{i\lambda_k t}\right|^2 dt. \tag{3.21}$$

Applying now Ingham's theorem (1.2) for the real sequence $(\lambda_k(h))_{|k|\leq \alpha N}$, in view of (3.18), it follows that if $T > T(\alpha)$ with $T(\alpha)$ as in (3.19), there exists a constant $C > 0$ such that

$$\sum_{k=-\alpha N, k\neq 0}^{\alpha N} \left| a_k \frac{\sin(Nk\pi h)}{h} \right|^2 \leq C(T) \int_0^T \left| \sum_{k=-\alpha N, k\neq 0}^{\alpha N} \frac{\sin(Nk\pi h)}{h} a_k e^{i\lambda_k t} \right|^2 dt, \tag{3.22}$$

holds for every solution of (3.9) in the class $\mathcal{C}_\alpha(h)$. Finally, it is sufficient to observe that

$$\sum_{k=-\alpha N, k\neq 0}^{\alpha N} \left| a_k \frac{\sin(Nk\pi h)}{h} \right|^2 \sim E_h,$$

to obtain a uniform observability inequality (3.11) in each class $\mathcal{C}_\alpha(h)$ for all $0 < \alpha < 1$. Note, however, that the minimal time $T(\alpha)$ depends on the filtering parameter α and, in particular, $T(\alpha) \to 2$ as $\alpha \to 0$ and $T(\alpha) \to \infty$ as $\alpha \to 1$ (see [28] for a rigorous proof).

As a further step towards a complete theory of numerical approximations of controls it is natural to address the same issue for full space-time discretizations. This issue is addressed in the following section.

3.3. Fully discrete approximations. The main ingredient to derive the fully discrete analogue of (3.5) for a finite-difference full discretization of a homogeneous 1-d wave equation (3.3) is the Fourier representation of solutions combined, this time, with our discrete Ingham inequality in Theorem 2.1.

Given $M, N \in \mathbb{N}$ we set $\Delta x = 1/(N+1)$ and $\Delta t = T/(M+1)$ and introduce the nets

$$0 = x_0 < x_1 = \Delta x < \cdots < x_N = N\Delta x < x_{N+1} = 1,$$

$$0 = t_0 < t_1 = \Delta t < \cdots < t_M = M\Delta t < t_{M+1} = T$$

with $x_j = j\Delta x$ and $t_n = n\Delta t$, $j = 0, 1, \ldots, N+1$, $n = 0, 1, \ldots, M+1$.

We consider the following finite-difference discretization of (3.1):

$$\begin{cases} \dfrac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} = \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}, & j = 1, 2, \ldots, N;\ n = 1, 2, \ldots, M, \\ u_0^n = 0, \qquad u_{N+1}^n = v_{\Delta x}^n, & n = 1, 2, \ldots, M, \\ u_j^0 = u_{0j}, \quad u_j^1 = \Delta t u_{1j} + u_{0j}, & j = 1, 2, \ldots, N. \end{cases} \tag{3.23}$$

We shall denote by $\bar{u}^n = (u_1^n, \ldots, u_N^n)$ the solution at the time step n. As in the context of the continuous wave equation above, we consider the uncontrolled system

$$\begin{cases} \dfrac{\phi_j^{n+1} - 2\phi_j^n + \phi_j^{n-1}}{(\Delta t)^2} = \dfrac{\phi_{j+1}^n - 2\phi_j^n + \phi_{j-1}^n}{(\Delta x)^2}, & j = 1, 2, \ldots, N; \quad n = 1, 2, \ldots, M, \\ \phi_0^n = \phi_{N+1}^n = 0, & n = 1, 2, \ldots, M, \\ \phi_j^0 = \phi_{0j}, \quad \phi_j^1 = \phi_{0j} + \Delta t \phi_{1j}, & j = 1, 2, \ldots, N, \end{cases} \tag{3.24}$$

a central finite difference discretization of (3.3).

Under the stability condition $\mu = \Delta t/\Delta x \leq 1$ (μ is the Courant number), the scheme (3.24) is convergent of order 2.

However, as observed in [14], the resulting discrete sequence of controls $v^n_{\Delta x} = -\phi^n_N/\Delta x$ obtained with a discrete HUM method may have an unstable behavior as $(\Delta t, \Delta x) \to (0,0)$. More precisely, it is possible to exhibit initial conditions such that the discrete controls $v^n_{\Delta x}$ do not converge towards the control v for (3.1) (see [28]). Once more, filtering of high frequencies is an efficient cure for these instabilities and our discrete Ingham inequality is the tool to analyze how it behaves.

The energy of (3.24) is

$$E_n = \frac{\Delta x}{2}\sum_{j=0}^{N}\left[\left(\frac{\phi_j^{n+1}-\phi_j^n}{\Delta t}\right)^2 + \left(\frac{\phi_{j+1}^{n+1}-\phi_j^{n+1}}{\Delta x}\right)\left(\frac{\phi_{j+1}^n-\phi_j^n}{\Delta x}\right)\right] \geq 0, \tag{3.25}$$

which is a discretization of the continuous energy E in (3.4), and it is conserved in all the time steps $E_n = E_0$, $n = 1, \ldots, M$, for the solutions of (3.24) (see [20]).

Solutions of (3.24) admit the Fourier development (see [20])

$$\bar{\phi}^n = \sum_{k=-N, k\neq 0}^{N} a_k e^{i\lambda_k n\Delta t}\bar{\varphi}_{|k|}, \tag{3.26}$$

with $a_k \in \mathbb{C}$, $\bar{\varphi}_k = (\varphi_{k,1}, \ldots, \varphi_{k,N}) = (\sin(k\pi\Delta x), \ldots, \sin(Nk\pi\Delta x))$ and

$$\lambda_k = \operatorname{sgn}(k)\frac{2}{\Delta t}\arcsin\left(\frac{\Delta t}{\Delta x}\sin\frac{k\pi\Delta x}{2}\right). \tag{3.27}$$

Our goal is to analyze the discrete version of the observability inequality (3.5)

$$E_0 \leq C\left[\Delta t\sum_{n=0}^{M}\left|\frac{\phi_N^n}{\Delta x}\right|^2\right], \tag{3.28}$$

where E_0 is the conserved energy of the solutions of the discrete system (3.24). This inequality implies by HUM a controllability property of the discrete analogue (3.23) of the control system (3.1). Of course, we seek for a positive constant $C > 0$ in (3.28), independent on Δt and Δx. This will yield a family of controls that will be bounded as $\Delta t \to 0$, which constitutes a natural candidate to converge to the control of (3.1).

Inequality (3.28) is the discrete analogue of (3.5). In particular, note that, according to Taylor's formula $\phi_x(1,t) \sim (\phi(1,t) - \phi(1-\Delta x, t))/\Delta x$. Thus, at the discrete level and taking into account that, according to the boundary conditions, $\phi^n_{N+1} = 0$, we obtain $\phi_x(1,t) \sim -\phi^n_N/\Delta x$. Thus, the right-hand side of (3.28) represents a discrete version of the right-hand side term in the continuous observability inequality (3.5).

Inequality (3.28) may also be seen as a time-discretization of the semi-discrete observability inequality (3.11). Note that, in fact, the semi-discrete case corresponds to taking $\mu = 0$ in the fully discrete scheme.

According to Theorem 2.1, the spectral gap between two consecutive eigenvalues plays a very important role in the analysis of the uniform observability inequality (3.28).

It is important to distinguish two cases:

• In the particular case where $\Delta t = \Delta x := h$ ($\mu = 1$) we have

$$\lambda_k = \mathrm{sgn}(k)\frac{2}{h}\arcsin\left(\sin\frac{k\pi h}{2}\right) = \mathrm{sgn}(k)k\pi.$$

Thus,

$$\lambda_{k+1} - \lambda_k = \gamma = \pi.$$

But the condition (2.2) does not hold, because

$$\max_{k,l}|\lambda_k - \lambda_l| = \frac{2\pi - 2\pi\Delta t}{\Delta t}.$$

Note, however, that, in this particular case, due to the orthogonality properties of the family of complex discrete exponentials involved in the Fourier representation of solutions,

$$\sum_{n=0}^{M} e^{in\Delta t\pi(k-l)} = (M+1)\delta_{k,l},$$

where $\delta_{k,l}$ is Kronecker's delta, an inequality of type (2.3) holds immediately and the discrete Ingham inequality is not needed.

Indeed, denoting by $m_k = (-1)^k a_k \sin(k\pi\Delta x)/\Delta x$, the energy of the solutions (3.24) concentrated on the extreme $x = 1$ can be written as

$$\Delta t\sum_{n=0}^{M}\left|\frac{\phi_N}{\Delta x}\right|^2 = \Delta t\sum_{n=0}^{M}\left|\sum_{k=-N}^{N} m_k e^{in\Delta t\pi k}\right|^2 \tag{3.29}$$

and the total energy of the solutions is

$$E_0 = \frac{1}{2}\sum_{k=-N}^{N}|m_k|^2 \tag{3.30}$$

(see [20] for more details). Then, for $T = 2$ we have

$$\begin{aligned} h\sum_{n=0}^{M}\left|\frac{\phi_N}{h}\right|^2 &= h\sum_{n=0}^{M}\left|\sum_{k=-N}^{N} m_k e^{inh\pi k}\right|^2 \\ &= h\sum_{n=0}^{M}\sum_{k=-N}^{N}|m_k|^2 + h\sum_{n=0}^{M}\sum_{k=-N,k\neq l}^{N} m_k\bar{m}_l e^{inh\pi(k-l)} = 2\sum_{k=-N}^{N}|m_k|^2, \end{aligned}$$

and therefore

$$E_0 = \frac{1}{4}\left[h\sum_{n=0}^{M}\left|\frac{\phi_N^n}{h}\right|^2\right].$$

A similar identity holds for the continuous wave equation (3.3) in the minimal observability time $T = 2$. Namely

$$E = \frac{1}{4}\int_0^2 |\phi_x(1,t)|^2$$

for every solution ϕ of (3.3), where E is the energy of the solutions $\phi = \phi(x,t)$.

• In the case when $\mu < 1$ the gap between two consecutive eigenfrequencies decreases at high frequencies and it is of the order of Δx when $\Delta x \to 0$. Indeed, we have

$$
\begin{aligned}
|\lambda_{k+1} - \lambda_k| &= \left| \frac{2}{\Delta t} \left[\arcsin\left(\frac{\Delta t}{\Delta x} \sin \frac{(k+1)\pi\Delta x}{2} \right) - \arcsin\left(\frac{\Delta t}{\Delta x} \sin \frac{k\pi\Delta x}{2} \right) \right] \right| \\
&\le \left| \frac{\pi}{2} \frac{2}{\Delta t} \frac{\Delta t}{\Delta x} \left(\sin \frac{(k+1)\pi\Delta x}{2} - \sin \frac{k\pi\Delta x}{2} \right) \right| \\
&= \left| \frac{\pi}{2} \frac{2}{\Delta x} \left[\sin \frac{k\pi\Delta x}{2} \left(\cos \frac{\pi\Delta x}{2} - 1 \right) + \sin \frac{\pi\Delta x}{2} \cos \frac{k\pi\Delta x}{2} \right] \right| \\
&\le \left| \frac{\pi}{2} \frac{2}{\Delta x} \left| 1 - \cos \frac{\pi\Delta x}{2} \right| + \frac{\pi^2}{2} \cos \frac{k\pi\Delta x}{2} \right| \\
&= \left| \frac{\pi}{2} \frac{2}{\Delta x} 2 \sin^2 \frac{\pi\Delta x}{4} + \frac{\pi^2}{2} \cos \frac{k\pi\Delta x}{2} \right| \le \left| \frac{\pi^2}{2} \left[\frac{\pi\Delta x}{4} + \sin\left(\frac{((N+1)-k)\Delta x\pi}{2} \right) \right] \right|.
\end{aligned}
$$

In particular, the gap for the highest frequencies satisfies

$$
|\lambda_N - \lambda_{N-1}| \le \frac{\pi^2}{2} \left(\frac{\pi\Delta x}{4} + \frac{\pi\Delta x}{2} \right) = \frac{3\pi^3 \Delta x}{8} \to 0, \text{ when } \Delta x \to 0.
$$

So the uniform gap condition (2.1) is not satisfied and we cannot directly apply Theorem 2.1 to prove inequality (3.28). Therefore, as soon as $\mu < 1$, we are in the same situation as for the semi-discrete equation (3.9) in which $\mu = 0$: the lack of spectral gap may produce the degeneracy of the observability constant.

To remedy this lack of uniform estimates, we need to introduce a subclass of solutions of system (3.24) where the high frequency components have been filtered. To do that, given $\alpha \in (0,1)$, the so-called *filtering parameter*, we consider the class $\mathcal{C}_\alpha(\Delta x)$,

$$
\mathcal{C}_\alpha(\Delta x) = \left\{ \bar\phi^n \text{ sol. of (3.24)} \ : \bar\phi^n = \sum_{k=-\alpha N, k\neq 0}^{\alpha N} a_k e^{i\lambda_k n\Delta t} \bar\varphi_{|k|} \right\}, \tag{3.31}
$$

of solutions involving the eigenvalues $\{\lambda_k\}_{k\in[-\alpha N,\alpha N]}$, $k \neq 0$:

$$
\bar\phi^n = \sum_{k=-\alpha N, k\neq 0}^{\alpha N} a_k e^{i\lambda_k n\Delta t} \bar\varphi_{|k|}. \tag{3.32}
$$

Let us first check the gap condition. We have

$$
\begin{aligned}
\lambda_{k+1} - \lambda_k &= \frac{2}{\Delta t} \left[\arcsin\left(\frac{\Delta t}{\Delta x} \sin \frac{(k+1)\pi\Delta x}{2} \right) - \arcsin\left(\frac{\Delta t}{\Delta x} \sin \frac{k\pi\Delta x}{2} \right) \right] \\
&= \frac{\pi \cos \frac{\xi\Delta x}{2}}{\sqrt{1 - \left(\frac{\Delta t}{\Delta x} \sin \frac{\xi\Delta x}{2} \right)^2}} := \gamma_k,
\end{aligned} \tag{3.33}
$$

for every $k \in [-\alpha N, \alpha N]$ and for some $\xi \in [k\pi, (k+1)\pi]$. Therefore, in particular,

$$\lambda_{k+1} - \lambda_k \geq \frac{\pi \cos \frac{N\alpha\pi\Delta x}{2}}{\sqrt{1 - \left(\frac{\Delta t}{\Delta x} \sin \frac{\xi \Delta x}{2}\right)^2}} \geq \pi \cos \frac{N\alpha\pi\Delta x}{2} \geq \pi(1-\alpha).$$

Consequently, for any filtering parameter $\alpha \in (0,1)$, the gap condition (2.1) holds with

$$\gamma_\alpha := \min_{|k| \leq \alpha N} (\gamma_k) \geq \gamma(\alpha) = \pi \cos\left(\frac{N\alpha\pi\Delta x}{2}\right) \geq \pi(1-\alpha). \tag{3.34}$$

On the other hand, by the mean value theorem,

$$\begin{aligned} |\lambda_k - \lambda_l| &= \left| \frac{2}{\Delta t}\left(\arcsin\left(\frac{\Delta t}{\Delta x}\sin\left(\frac{k\pi\Delta x}{2}\right)\right) - \arcsin\left(\frac{\Delta t}{\Delta x}\sin\left(\frac{l\pi\Delta x}{2}\right)\right)\right)\right| \\ &= \left| \frac{2}{\Delta t} \frac{\frac{\Delta t}{\Delta x}\frac{\pi\Delta x}{2}\cos\left(\frac{\xi\pi\Delta x}{2}\right)(k-l)}{\sqrt{1 - \left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2 \frac{\xi\pi\Delta x}{2}}} \right| \leq \left| \frac{2N\alpha\pi\cos\left(\frac{\xi\pi\Delta x}{2}\right)}{\sqrt{1 - \left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2 \frac{\xi\pi\Delta x}{2}}} \right| \\ &\leq \left| \frac{2N\alpha\pi\cos\left(\frac{\xi\pi\Delta x}{2}\right)}{\sqrt{\left(\frac{\Delta t}{\Delta x}\right)^2 - \left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2 \frac{\xi\pi\Delta x}{2}}} \right| = \frac{2N\alpha\pi\Delta x}{\Delta t} = \frac{2\alpha\pi - 2\alpha\pi\Delta x}{\Delta t} \\ &\leq \frac{2\pi\alpha(1-\Delta t)}{\Delta t}. \end{aligned} \tag{3.35}$$

In view of (3.35), by choosing conveniently the filtering parameter α such that

$$\alpha \leq \alpha^*(\Delta t) := \frac{2\pi - (\Delta t)^p}{2\pi(1-\Delta t)}, \tag{3.36}$$

with $0 \leq p < 1/2$, hypothesis (2.2) of Theorem 2.1 is verified.

In practice it is convenient to fix the filtering parameter $0 < \alpha < 1$, independent of Δt. In this way (3.36) is automatically satisfied for Δt small enough, which is the relevant case in numerical approximation problems. On the other hand the gap condition (3.34) is also automatically and uniformly satisfied for the truncated sequence $\{\lambda_k\}_{|k| \leq N\alpha}$.

Note that the gap γ_α (respectively, the minimal observability/ control time $2\pi/\gamma_\alpha$) tends to π (respectively, to 2) when $\alpha \searrow 0^+$ while it converges to zero (respectively, to infinity) when $\alpha \nearrow 1^-$.

Note also that the minimal observability/control time can be taken to be any $T > 2\pi/\gamma_\alpha$ since the minimal time $T(\alpha) = 2\pi/\gamma_\alpha + \epsilon(\Delta t)$ tends to $2\pi/\gamma_\alpha$ as Δt tends to zero.

More precisely, the following theorem holds.

THEOREM 3.1. *For all Courant numbers $0 < \mu < 1$ and all values of the filtering parameter $0 < \alpha < 1$, the observability inequality below holds that*

$$E_0 \leq \frac{1}{2\cos^2 \frac{\alpha\pi}{2} C_1(T, \gamma_\alpha)} \left[\Delta t \sum_{n=0}^{M} \left| \frac{\phi_N}{\Delta x} \right|^2 \right] \tag{3.37}$$

for every solution of (3.24) *in the class $\mathcal{C}_\alpha(\Delta x)$, uniformly as $(\Delta t, \Delta x) \to (0,0)$ for any $T > T(\alpha) = 2\pi/\gamma_\alpha$, with $C_1(T, \gamma_\alpha)$ given by* (1.3). *Moreover,*

1. $T(\alpha) \nearrow \infty$ *as* $\alpha \nearrow 1^-$ *and* $T(\alpha) \searrow 2$ *as* $\alpha \searrow 0^+$.
2. $C_\alpha(T) := \dfrac{1}{2\cos^2 \frac{\alpha\pi}{2} C_1(T,\gamma_\alpha)} \searrow C(T) = \dfrac{1}{2C_1(T,\gamma)}$ *as* $\alpha \searrow 0^+$ *with* $C_1(T,\gamma)$ *given by* (1.3), *where* $C(T)$ *is the constant of the continuous observability inequality* (3.5).

Remark 3.2. This theorem allows the recovery of the uniform observability of the original system (3.3) as the limit when $(\Delta t, \Delta x) \to (0,0)$ of the observability of the solutions of discrete one (3.24) in the classes (3.31) by means of Fourier filtering; the statements in this theorem coincide with the predictions one may deduce from the analysis of the dispersion diagram of the numerical scheme [28], as we shall see in the next section.

Proof (Sketch of the proof). The energy of the solutions (3.26) of the discrete system (3.24), concentrated on $x = 1$ is given by (3.29) and the total energy (3.25) of the solutions is

$$E_0 = \frac{2}{(\Delta x)^2} \sum_k a_k^2 \sin^2 \frac{k\pi\Delta x}{2} = \frac{2}{(\Delta x)^2} \sum_k a_k^2 \frac{\sin^2(k\pi\Delta x)}{4\cos^2 \frac{k\pi\Delta x}{2}} = \frac{1}{2} \sum_k \mid m_k \mid^2 \frac{1}{\cos^2 \frac{k\pi\Delta x}{2}},$$

where $m_k = \sin(Nk\pi\Delta x)/\Delta x$.

For all $k \in [-\alpha N, \alpha N]$ we have $\cos(\alpha\pi/2) \le \cos(\alpha N\pi\Delta x/2) \le \cos(k\pi\Delta x/2) \le 1$ and, in this case,

$$\frac{1}{2} \sum_k \mid m_k \mid^2 \le E_0 \le \frac{1}{2\cos^2 \frac{N\alpha\pi\Delta x}{2}} \sum_k \mid m_k \mid^2 \le \frac{1}{2\cos^2 \frac{\alpha\pi}{2}} \sum_k \mid m_k \mid^2. \tag{3.38}$$

Applying Theorem 2.1 and the Fourier representation (3.32) of the solutions we obtain that, for all $T > 2\pi/\gamma_\alpha + \epsilon(\Delta t)$, there exist positive constants $C_j(\Delta t, T, \gamma_\alpha)$, $j = 1, 2$, such that

$$C_1(\Delta t, T, \gamma_\alpha) \sum_{k=-\alpha N}^{\alpha N} |m_k|^2 \le \Delta t \sum_{n=0}^{M} \left| \sum_{k=-\alpha N}^{\alpha N} m_k e^{in\Delta t\lambda_k} \right|^2 \le C_2(\Delta t, T, \gamma_\alpha) \sum_{k=-N}^{N} |m_k|^2.$$

Therefore, for every α as in (3.36), by (3.38), the following inequalities hold:

$$2\cos^2 \frac{\alpha\pi}{2} C_1(\Delta t, T, \gamma_\alpha) E_0 \le \Delta t \sum_{n=0}^{M} \left| \frac{\phi_N}{\Delta x} \right|^2 \le 2C_2(\Delta t, T, \gamma_\alpha) E_0, \tag{3.39}$$

with $C_j(\Delta t, T, \gamma_\alpha)$, $j = 1, 2$, defined by relations (2.4), for every truncated solution (3.32) of system (3.24) belonging to the class $\mathcal{C}_\alpha(\Delta x)$. □

The uniform observability inequality (3.39) implies uniform controllability results, as we shall prove in the next section, for the projection (over the subspace of unfiltered Fourier components) of solutions of the dual controlled system (3.23). In the limit as $\Delta t, \Delta x \to 0$ one recovers the sharp controllability results of the wave equation (3.1). For the details of the proof of convergence of controls we refer to [20] where the case $\Delta t = \Delta x$ was studied in detail. But, as mentioned above, for this particular one, because of the orthogonality of complex harmonic polynomials, the discrete Ingham inequality is not needed. We also refer to [16] where the convergence of controls for the semi-discretizations of the beam equation was analyzed in detail.

The usual centered finite-difference approximation of the wave equation we have considered here is only a simple example in which the discrete Ingham's theorem can be applied, together with some filtering mechanism, to get uniform observability inequalities. The discrete Ingham inequality can also be applied, for instance, to the implicit fully finite difference approximation of the wave equation, introduced in [18].

4. Uniform controllability of the filtered solutions. In this section, we apply the uniform observability results obtained above to analyze the controllability properties of the fully discrete system (3.23).

Let us define the Hilbert spaces of square summable sequences $\hbar^1$ and $\hbar^{-1}$ as follows:

$$\hbar^1 = \left\{ \{a_k\} \in \ell^2 \ : \quad \|a_k\|^2_{\hbar^1} = \sum_{k\in\mathbb{N}} |\ k\pi a_k\ |^2 < \infty \right\}, \tag{4.1}$$

$$\hbar^{-1} = \left\{ \{a_k\} \in \ell^2 \ : \quad \|a_k\|^2_{\hbar^{-1}} = \sum_{k\in\mathbb{N}} \left| \frac{a_k}{k\pi} \right|^2 < \infty \right\}, \tag{4.2}$$

where the discrete space ℓ^2 is given by (1.5).

For every $\alpha \in (0,1)$, we introduce the space $\mathcal{S}_\alpha$ generated by the eigenvectors $(\bar\varphi_k)$ involved in $\mathcal{C}_\alpha(\Delta x)$ of the filtered solutions of the homogeneous system (3.24) with filtering parameter α:

$$\mathcal{S}_\alpha = span\left\{ \bar\varphi_k : |k| \le \alpha N \right\}. \tag{4.3}$$

For every $s \in \mathbb{R}$, we denote by $\hbar^s_{\Delta x,\alpha}$ the space S_α endowed with the norm

$$\|v\|^2_{s,\Delta x} = \sum_{|k|\le N\alpha} \lambda_k^s \left|a_k\right|^2, \quad \text{for} \quad v \in S_\alpha : v = \sum_{|k|\le N\alpha} a_k \bar\varphi_k,$$

where λ_k are as in (3.27).

For every $\alpha \in (0,1)$ and $T > 0$, we consider the partial controllability problem for system (3.23) in the space $\ell^2 \times \hbar^{-1}$, which consists of finding a control $\bar v^n \in \mathbb{R}^M$ such that, for all initial data $(\bar u^0, \bar u^1) \in \ell^2 \times \hbar^{-1}$, the solution $\bar u^n$ of (3.23) satisfies

$$(\Pi_\alpha \bar u^M, \Pi_\alpha \bar u^{M+1}) = (0,0), \tag{4.4}$$

where Π_α is the orthogonal projection over S_α; i.e.,

$$\left(\Pi_\alpha \bar u^M, \Pi_\alpha \bar u^{M+1}\right) = \left(\sum_{|k|\le N\alpha} c_k \bar\varphi_k, \sum_{|k|\le N\alpha} d_k \bar\varphi_k \right),$$

where (c_k) and (d_k) are the Fourier coefficients of $(\bar u^M, \bar u^{M+1})$ in the basis of the eigenvectors $(\bar\varphi_k)_k$. Observe that we only require to control uniformly the projection Π_α of the solutions of the discrete system (3.23) over subspaces in which the high frequencies have been filtered.

As we shall see this result is a consequence of the partial observability results of the previous section in the class of filtered solutions $\mathcal{C}_\alpha(\Delta x)$.

Multiplying the first equation in (3.23) by an arbitrary solution $\bar\phi^n$ of (3.24) and adding in j and n, we get

$$\Delta t \sum_{n=1}^{M} v^n_{\Delta x} \frac{\phi^n_N}{\Delta x} + \frac{1}{\mu} \sum_{j=0}^{N} \left[u^1_j \phi^0_j - u^0_j \phi^1_j \right] = \frac{1}{\mu} \sum_{j=0}^{N} \left[u^{M+1}_j \phi^M_j - u^M_j \phi^{M+1}_j \right]. \tag{4.5}$$

The solution of system (3.23) may be characterized through a transposition argument based on the identity above. Indeed, given $M,\ N \in \mathbb{N}$, $\bar{v}_{\Delta t} \in \mathbb{R}^M$, and $(\bar{u}^0, \bar{u}^1) \in \mathbb{R}^N \times \mathbb{R}^N$, $\{\bar{u}^n\}$ solves (3.23) if for every $s \in [1, M]$ it holds that

$$L_s(\bar{\phi}^0, \bar{\phi}^1) = \frac{1}{\mu} \sum_{j=1}^{N} \left[u_j^s \phi_j^{s+1} - u_j^{s+1} \phi_j^s \right],$$

or equivalently

$$L_s(\bar{\phi}^0, \bar{\phi}^1) = \frac{1}{\mu} \left(\bar{u}^s, \bar{\phi}^{s+1} \right)_{\mathbb{R}^N} + \frac{1}{\mu} \left(\bar{u}^{s+1}, -\bar{\phi}^s \right)_{\mathbb{R}^N}, \tag{4.6}$$

for every solution $\{\bar{\phi}^n\}$ of the discrete problem (3.24), where the functional L_s: $\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is such that

$$L_s(\bar{\phi}^0, \bar{\phi}^1) = \frac{1}{\mu} \sum_{j=1}^{N} \left[u_j^0 \phi_j^1 - u_j^1 \phi_j^0 \right] - \Delta t \sum_{n=1}^{s} v_{\Delta x}^n \left[\frac{\phi_N^n}{\Delta x} \right].$$

The projection $\Pi_\alpha \bar{u}^n$ may be characterized by the same variational formulation (4.6), with the only difference being that the test functions in (4.6) are solutions of (3.24) in the class $\mathcal{C}_\alpha(\Delta x)$ (3.31).

Remark 4.1. Identity (4.4) is equivalent, by (4.5), to

$$\Delta t \sum_{n=1}^{M} v_{\Delta x}^n \left[\frac{\phi_N^n}{\Delta x} \right] = \frac{1}{\mu} \sum_{j=0}^{N} \left[u_j^0 \phi_j^1 - u_j^1 \phi_j^0 \right], \tag{4.7}$$

where $(\bar{\phi}^0, \bar{\phi}^1)$ are the initial data corresponding to the solution $\bar{\phi}^n \in \mathcal{C}_\alpha(\Delta x)$ of the discrete system (3.24).

Now let $\Delta x = 1/q$, $\Delta t = \mu/q$, $N = q - 1$, for some $q \in \mathbb{N}$ and $\mu < 1$. We have the following uniform (with respect to $(\Delta t, \Delta x) \to (0, 0)$) partial controllability property.

THEOREM 4.2. *Let $0 < \mu < 1$ and let us fix an arbitrary value of the filtering parameter $0 < \alpha < 1$. For every $T > T(\alpha) = 2\pi/\gamma_\alpha$, the system* (3.24) *is partially controllable on $\ell^2 \times \hbar^{-1}$ with controls $\bar{v}_{\Delta t}^n \in \mathbb{R}^M$ when $M = [Tq/\mu - 1]$. Moreover, the controls of minimal norm are uniformly bounded with respect to Δt. More precisely*

$$\left[\Delta t \sum_{n=0}^{M} |v_{\Delta x}^n|^2 \right]^{1/2} \leq C \|(\bar{u}^1, -\bar{u}^0)\|_{\hbar^{-1} \times \ell^2}, \tag{4.8}$$

where $C = C(T, \gamma_\alpha) > 0$ is a constant independent of $\Delta t \in (0, 1)$.

Proof. Let $(\bar{\phi}^n) \in \mathcal{C}_\alpha(\Delta x)$ be the solution of (3.24) with initial data $(\bar{\phi}^0, \bar{\phi}^1) \in S_\alpha \times S_\alpha$ and define the convex quadratic functional $J_{\Delta x} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, by

$$J_{\Delta x}(\bar{\phi}^0, \bar{\phi}^1) = \frac{\Delta t}{2} \sum_{n=0}^{M} \left| \frac{\phi_N^n}{\Delta x} \right|^2 - \frac{1}{\mu^2} \Delta x \sum_{j=0}^{N} \left(u_j^0 \frac{\phi_j^1 - \phi_j^0}{\Delta t} - \frac{u_j^1 - u_j^0}{\Delta t} \phi_j^0 \right). \tag{4.9}$$

For every $\bar{\phi}^n \in \mathcal{C}_\alpha(\Delta x)$ we have

$$\begin{aligned} \left| \sum_{j=0}^{N} \left(u_j^0 \phi_j^1 - u_j^1 \phi_j^0 \right) \right| &= \left| \left(\Pi_\alpha \bar{u}^1, \bar{\phi}^0 \right)_{\mathbb{R}^N} \right| + \left| \left(\Pi_\alpha \bar{u}^0, -\bar{\phi}^1 \right)_{\mathbb{R}^N} \right| \\ &\leq \|\Pi_\alpha \bar{u}^1\|_{\hbar^{-1}} \|\bar{\phi}^0\|_{\hbar^1} + \|\Pi_\alpha \bar{u}^0\|_{\ell^2} \|\bar{\phi}^1\|_{\ell^2} \\ &\leq \|\Pi_\alpha (\bar{u}^1, -\bar{u}^0)\|_{\hbar^{-1} \times \ell^2} \|(\bar{\phi}^0, \bar{\phi}^1)\|_{\hbar^1 \times \ell^2}. \end{aligned} \tag{4.10}$$

According to (4.10) and the direct observability inequality (the right-hand side term in (3.39)) we deduce that $J_{\Delta x}$ is continuous.

On the other hand, according to the observability inequality (3.37), $J_{\Delta x}$ is uniformly coercive in $\mathcal{C}_\alpha(\Delta x)$,

(4.11)
$$\left|J_{\Delta x}(\bar\phi^0,\bar\phi^1)\right| \geq \left\|(\bar\phi^0,\bar\phi^1)\right\|_{\hbar^1\times\ell^2}\left[C_1(T,\gamma_\alpha)\left\|(\bar\phi^0,\bar\phi^1)\right\|_{\hbar^1\times\ell^2} - \|\Pi_\alpha(\bar u^1,-\bar u^0)\|_{\hbar^{-1}\times\ell^2}\right].$$

Thus, there exists a unique minimizer $(\hat\phi^0,\hat\phi^1)$ of $J_{\Delta x}$,

$$J_{\Delta x}(\hat\phi^0,\hat\phi^1) = \min_{(\bar\phi^0,\bar\phi^1)\in S_\alpha\times S_\alpha} J_{\Delta x}(\bar\phi^0,\bar\phi^1).$$

Let $\hat\phi^n\in\mathcal{C}_\alpha(\Delta x)$ be the solution of the adjoint problem (3.24) with this minimizer as initial datum.

The pair $(\hat\phi^0,\hat\phi^1)$ satisfies the Euler–Lagrange equation

$$\Delta t\sum_{n=0}^{M}\frac{\hat\phi_N^n}{\Delta x}\frac{\phi_N^n}{\Delta x} = \frac{1}{\mu}\sum_{j=0}^{N}\left[u_j^0\phi_j^1 - u_j^1\phi_j^0\right], \tag{4.12}$$

for every initial data $(\bar\phi^0,\bar\phi^1)\in S_\alpha\times S_\alpha$ associated to the solution $\bar\phi^n\in\mathcal{C}_\alpha(\Delta x)$ of (3.24). Therefore, according to (4.7), the control we were looking for is $v_{\Delta x}^n = \hat\phi_N^n/\Delta x$.

To conclude the proof we check the uniform boundedness of the controls $v_{\Delta x}^n$. We have

$$J_{\Delta x}((\hat\phi^0,\hat\phi^1)) \leq J_{\Delta x}(0,0) = 0,$$

and, by (4.10), this implies

$$\frac{\Delta t}{2}\sum_{n=0}^{M}\left|\frac{\hat\phi_N^n}{\Delta x}\right|^2 \leq \|\Pi_\alpha(\bar u^1,-\bar u^0)\|_{\hbar^{-1}\times\ell^2}\|(\bar\phi^0,\bar\phi^1)\|_{\hbar^1\times\ell^2}. \tag{4.13}$$

The discrete energy E_0 of a solution $\bar\phi^n$ of (3.24) with initial data $(\bar\phi^0,\bar\phi^1)$ satisfies

$$E_0 = \frac{1}{2}\|(\bar\phi^0,\bar\phi^1)\|_{\hbar^1\times\ell^2}^2.$$

Now, using the Fourier development (3.32) of the solution $\bar\phi^n$ and applying the observability inequality (3.37) we get

$$\begin{aligned}\|(\bar\phi^0,\bar\phi^1)\|_{\hbar^1\times\ell^2}^2 = 2E_0 &\leq \frac{1}{\cos^2\frac{N\alpha\pi\Delta x}{2}C_1(T,\gamma_\alpha,\Delta t)}\Delta t\sum_{n=0}^{M}\left|\frac{\hat\phi_N^n}{\Delta x}\right|^2\\ &\leq \frac{1}{\cos^2\frac{\alpha\pi}{2}C_1(T,\gamma_\alpha)}\Delta t\sum_{n=0}^{M}\left|\frac{\hat\phi_N^n}{\Delta x}\right|^2.\end{aligned} \tag{4.14}$$

Therefore, in (4.13) we obtain

$$\left[\Delta t\sum_{n=0}^{M}\left|\frac{\hat\phi_N^n}{\Delta x}\right|^2\right]^{1/2} \leq \frac{1}{\sqrt{\cos^2\frac{\alpha\pi}{2}C_1(T,\gamma_\alpha)}}\|\Pi_\alpha(\bar u^1,-\bar u^0)\|_{\hbar^{-1}\times\ell^2}, \tag{4.15}$$

and then, the discrete controls $v^n_{\Delta x} = \hat{\phi}^n_N/\Delta x$ satisfy

$$\left[\Delta t \sum_{n=0}^{M} |v^n_{\Delta x}|^2\right]^{1/2} \leq C(T,\gamma_\alpha)\|\Pi_\alpha(\bar{u}^1, -\bar{u}^0)\|_{\hbar^{-1}\times\ell^2}, \tag{4.16}$$

as stated above. □

Remark 4.3. Note that, with the notations (3.32), the controls $(v^n_{\Delta x})$ are of the form

$$v^n_{\Delta x} = -\frac{\mu}{\Delta t}\sum_{k=-N\alpha}^{N\alpha} \cos(k\pi)\sin(k\pi\Delta t/\mu)\hat{a}_k e^{i\lambda_k n\Delta t}, \tag{4.17}$$

where $(\hat{a}_k)_k$ are the Fourier coefficients of the solution $\hat{\phi}^n \in \mathcal{C}_\alpha(\Delta x)$ of the adjoint problem (3.24), with initial data $(\hat{\phi}^0, \hat{\phi}^1)$ being the minimizer of the functional $J_{\Delta x}$.

Now we show the convergence of the controls $v^n_{\Delta x}$ of the discrete system (3.23) to the HUM control of the continuous one (3.1), as $\Delta t, \Delta x \to 0$.

Given an initial state $(u^0, u^1) \in L^2(0,1)\times H^{-1}(0,1)$ of the continuous system (3.1), we develop it in Fourier series

$$(u^0, u^1) = \sum_{k=1}^{\infty}(c_k, d_k)\varphi_k(x), \tag{4.18}$$

with

$$\sum_{k\in\mathbb{N}}\left[|c_k|^2 + \left|\frac{d_k}{k\pi}\right|^2\right] < \infty. \tag{4.19}$$

We now construct the initial states for the discrete system (3.23) by setting

$$(\bar{u}^0, \bar{u}^1) = \sum_{k=1}^{N}(c_k, c_k\cos(\lambda_k n\Delta t) + \frac{d_k}{\lambda_k}\sin(\lambda_k n\Delta t))\bar{\varphi}_k, \tag{4.20}$$

with λ_k given by (3.27). They may be rewritten as

$$(\bar{u}^0, \bar{u}^1) = \sum_{k=1}^{\infty}(c^N_k, c^N_k\cos(\lambda_k n\Delta t) + \frac{d^N_k}{\lambda_k}\sin(\lambda_k n\Delta t))\bar{\varphi}_k, \tag{4.21}$$

where

$$c^N_k = c_k\chi_N(k), \qquad d^N_k = d_k\chi_N(k),$$

χ_N being the characteristic function of the set $\{1,\dots,N\}$.

In view of Theorem 4.2, there exists a HUM control $(v^n_{\Delta x})$ for the discrete system (3.23), satisfying (4.4), with initial data (4.20).

Let us now prove that the sequence $(v^n_{\Delta x})$ converges (in a sense to be more precise below) to $v \in L^2(0,T)$, which is the HUM control for system (3.1) with initial data (4.18).

To better analyze the convergence of controls, we define the continuous extension of the discrete controls by setting

$$v_{\Delta x}(t) = -\frac{\mu}{\Delta t}\sum_{k\in\mathbb{Z}}\cos(k\pi)\sin(k\pi\Delta t/\mu)\hat{a}_k e^{i\lambda_k t},$$

where $\hat{a}_k$ are taken to be zero for $|k| > \alpha N$. This function, when restricted to the mesh, coincides with $(v^n_{\Delta x})$ (recall that $v^n_{\Delta x}$ is given by (4.17)).

The following convergence result holds.

THEOREM 4.4. *Let $\mu = \Delta t/\Delta x \leq 1$. Consider M, N as in Theorem* 4.2. *Fix $(u_0, u_1) \in L^2(0,1) \times H^{-1}(0,1)$ and consider the continuous and discrete controls v and $v_{\Delta x}$ as above, with the filtering parameter $\alpha \in (0,1)$ and $T > T_\alpha$. Then,*

$$v_{\Delta x}(\cdot) \to v(\cdot) \ \textit{strongly in} \ L^2(0,T) \ \text{as} \ \Delta t \to 0. \tag{4.22}$$

Proof (*Sketch of the proof*). In view of (4.8) it is easy to see that

$$\Delta t \sum_{n=0}^{M} |v^n_{\Delta x}|^2 \leq C, \tag{4.23}$$

and therefore

$$\int_0^T |v_{\Delta x}|^2 \, dt \leq C. \tag{4.24}$$

Then, up to the extraction of a subsequence that we still denote by $\{v_{\Delta x}\}_{\Delta t}$, we have

$$v_{\Delta x}(t) \rightharpoonup v(t) \ \text{ in } \ L^2(0,T) \ \text{ as } \ \Delta t \to 0. \tag{4.25}$$

By a Γ-convergence argument it can also be seen that the limit v is given by

$$v(t) = -\partial_x \hat{\phi}(1,t), \tag{4.26}$$

where $\hat{\phi}$ is the solution of the adjoint problem (3.3) with initial data $(\hat{\phi}^0, \hat{\phi}^1) \in H_0^1(0,1) \times L^2(0,1)$, the unique minimizer of the functional

$$J(\phi^0, \phi^1) = \frac{1}{2} \int_0^T \mid \partial_x \phi(1,t) \mid^2 dt - \int_0^1 u^0 \phi^1 - < u^1, \phi^0 >_{-1,1} \tag{4.27}$$

in the energy space $H_0^1(0,1) \times L^2(0,1)$. By taking limits in (4.7) and thanks to the construction of the initial data to be controlled for the discrete system we obtain

$$0 = \int_0^1 \left[u^1(x)\phi^0(x) - u^0(x)\phi^1(x) dx \right] + \int_0^T v(t) \partial_x \phi(1,t) dt \tag{4.28}$$

and this latter condition is equivalent to the fact that v, the limit in (4.25), is a control for system (3.1), driving the initial data (u^0, u^1) to rest; i.e., $v \in L^2(0,T)$ is the control of minimal L^2-norm.

The limit v being identified in a unique way, we deduce that the whole sequence $v_{\Delta x}$ converges.

Moreover, by the hypotheses of Theorem 4.4, the linear term of the discrete functional $J_{\Delta x}$ in (4.9) converges to the linear term of the functional defined in (4.27). Therefore, proving (4.22) is equivalent to proving that

$$J_{\Delta x}(\hat{\phi}^0_{\Delta x}, \hat{\phi}^1_{\Delta x}) \to J(\hat{\phi}^0, \hat{\phi}^1), \ \text{ as } \ \Delta x \to 0,$$

where $(\hat{\phi}^0_{\Delta x}, \hat{\phi}^1_{\Delta x}) \in S_\alpha \times S_\alpha$ minimizes (4.9) and $(\hat{\phi}^0, \hat{\phi}^1) \in H_0^1(0,1) \times L^2(0,1)$ minimizes (4.27). Indeed, taking into account the convergence of the linear terms in

this functional, and the structure of the functionals (4.9) and (4.27), we deduce the convergence of the norms of the controls that, together with the weak convergence, ensure strong convergence.

Thus, the controls $v_{\Delta x}$ and the controlled discrete solutions $u_{\Delta x}$ converge to the control and the controlled solution of the wave equation (3.1). It is important to note that the projections of the solutions of the controlled system end up covering the whole range of frequencies so that, in the limit, we recover the exact controllability property (3.2) of the continuous wave equation.

The details of the several steps of the proof are given in [19] and we omit them for brevity. □

5. Discrete Ingham inequalities and dispersion diagrams. In this section we discuss the observability results obtained in section 3 applying discrete Ingham inequalities in connection with the dispersion diagrams of the equations and numerical schemes under consideration. We also discuss the optimality of these results. First of all, we introduce and recall some classical concepts and notations.

Any time-dependent scalar, linear partial differential equation with constant coefficients admits plane wave solutions

$$\phi(x,t) = e^{i(\omega t - \xi x)}, \quad \xi \in \mathbb{R},\ \omega \in \mathbb{C}, \tag{5.1}$$

where ξ is the *wave number* and ω is the *frequency*. The relationship

$$\omega = \omega(\xi) \tag{5.2}$$

is known as the *dispersion relation* for the equation.

Any individual *"monochromatic wave"* (involving only one Fourier component) of (5.1) moves at the *phase velocity*

$$c(\xi,\omega) = \frac{\omega(\xi)}{\xi}. \tag{5.3}$$

When one superimposes two waves with nearby propagation velocities, there appear wave packets which can propagate with different velocities. The energy of wave packets propagates at the so-called *group velocity*

$$C(\xi,\omega) = \frac{d\omega(\xi)}{d\xi}. \tag{5.4}$$

In general, the dispersion relation for a partial differential equation is a polynomial relation between ξ and ω, while a discrete model amounts to a trigonometric approximation.

• *Continuous problem.* For the continuous wave equation (3.3) we have $\omega(\xi) = \xi$ and therefore $c(\xi) = C(\xi) = 1$.

• *Semi-discrete problem.* For the semi-discrete scheme (3.9) the dispersion relation is

$$\omega(\xi) = \frac{2}{\Delta x}\sin\frac{\xi\Delta x}{2}, \quad \xi \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]. \tag{5.5}$$

Note that, at the semi-discrete level, each dispersion relation is $2\pi/\Delta x$–periodic in ξ, and it is natural to take $\xi \in [-\pi/\Delta x, \pi/\Delta x]$ as a fundamental domain.

The phase velocity is in this case

$$c(\xi,\omega)=\frac{2}{\xi\Delta x}\sin\frac{\xi\Delta x}{2}. \tag{5.6}$$

The corresponding group velocity is

$$C(\xi,\omega)=\frac{d\omega(\xi)}{d\xi}=\cos\frac{\xi\Delta x}{2}. \tag{5.7}$$

• *Discrete problem.* The same analysis can be developed for fully discrete schemes. Considering numerical plane waves $\phi_j^n=e^{i(\omega n\Delta t-\xi j\Delta x)}$, for system (3.24), one obtains the dispersion relation

$$\omega(\xi)=\frac{2}{\Delta t}\arcsin\left(\frac{\Delta t}{\Delta x}\sin\frac{\xi\Delta x}{2}\right). \tag{5.8}$$

It is $2\pi/\Delta x$-periodic in ξ and $2\pi/\Delta t$-periodic in ω.

• When $\Delta t=\Delta x$ we obtain

$$\omega(\xi)=\xi. \tag{5.9}$$

This case is particularly interesting since (5.9) coincides with the dispersion relation for the continuous weave equation. In this case, $c(\xi,\omega)=C(\xi,\omega)=1$ and the discrete waves propagate at a constant velocity identically equal to one, like in the continuous case. But, as we shall see, this is a completely exceptional situation.

• When $\mu<1$, the phase velocity is given by

$$c(\xi,\omega)=\frac{2}{\xi\Delta t}\arcsin\left(\frac{\Delta t}{\Delta x}\sin\frac{\xi\Delta x}{2}\right) \tag{5.10}$$

and the group velocity is

$$C(\xi,\omega)=\frac{d\omega(\xi)}{d\xi}=\frac{\cos\frac{\xi\Delta x}{2}}{\sqrt{1-\left(\frac{\Delta t}{\Delta x}\sin\frac{\xi\Delta x}{2}\right)^2}}. \tag{5.11}$$

For $\Delta t=0$ the phase and group velocities in (5.10) and (5.11), which depend on ξ, coincide with those of the semi-discrete case (5.6) and (5.7), respectively, as expected.

Note that, as $\Delta x\to 0$, for all ξ we have

$$C(\xi,\omega)\le\frac{\cos\frac{\xi\Delta x}{2}}{\sqrt{1-\left(\frac{\Delta t}{\Delta x}\right)^2}}\to 0$$

when $\xi=\pi/\Delta x$.

In Figures 1–4 we describe the evolution of the group velocity diagrams starting with the semi-discrete case ($\mu=0$) up to $\mu=1$, for fixed $\Delta x=0.001$.

In general, any discrete dynamics generates spurious high-frequency oscillations that do no exist at the continuous level [23, 25]. Moreover, the interaction of waves with the grid produces a dispersion phenomenon and the velocity of propagation of these high frequency numerical waves may converge to zero when the mesh-size tends to zero. These spurious oscillations weakly converge to zero. Consequently, their

existence is compatible with the convergence of the numerical scheme for solving the initial-value problem. However, when we are dealing with the exact controllability or observability problems, a uniform time for the control of all numerical waves is needed. Since the velocity of propagation of some high frequency numerical waves may tend to zero as the mesh becomes finer and finer, uniform observability and therefore controllability properties of the discrete model may fail for all $T > 0$.

According to Theorem 2.1, the uniform gap between two consecutive eigenvalues is a sufficient (and actually also necessary) property for uniform (with respect to Δx and Δt) observability. On the other hand, the group velocity is the derivative of the eigenfrequencies λ_k and the spectral gap is, as we have seen, $\lambda_{k+1} - \lambda_k$. Both magnitudes are similar, and they become closer as $\Delta x \to 0$.

Thus, to efficiently observe at the point $x = 1$ a wave packet concentrated to the left of $x = 1$ that moves to the left (in the space variable) as t increases, and bounces back at $x = 0$ to eventually reach the observation point $x = 1$, the time needed is

$$T \geq 2/\min_{\xi}\{C(\xi, \omega)\}. \tag{5.12}$$

In the continuous case, (5.12) reduces to the well-known condition for observability $T \geq 2$ and it is uniform for all the frequencies. The minimal time $T = 2$ is the one one obtains in view of Ingham's theorem (1.2) because the gap is $\gamma = \pi$ in this case.

For the semi-discrete case, the observation time is

$$T \geq 2/\min_{\xi}(\cos(\xi \Delta x/2)). \tag{5.13}$$

But $\min_\xi(\cos(\xi\Delta x/2))$ is of the order of Δx, the same order as we have obtained in (3.18) for the spectral gap for the highest frequencies. Consequently, the observation time (5.13) diverges, $T \to \infty$, as $\Delta x \to 0$.

These facts confirm the necessity of filtering the high frequencies. Relation (5.13) shows that the time grows with the high frequencies, in the points where $\cos \xi\Delta x/2 \sim 0$ ($\xi \sim \pi/\Delta x$) and the same result is obtained applying the Ingham inequality.

For the fully discrete problem (3.24) the time needed for observation is

$$T \geq \max_{\xi} \frac{2\sqrt{1 - \left(\frac{\Delta t}{\Delta x} \sin \frac{\xi \Delta x}{2}\right)^2}}{\cos \frac{\xi \Delta x}{2}}. \tag{5.14}$$

Passing to the limit in (5.14) as $\Delta t \to 0$ for fixed Δx, one obtains the same time as in the semi-discrete case (5.13). The observation time grows with the high frequencies, except for the very particular case $\Delta t = \Delta x$, where the time obtained in the previous section, using the orthogonality of the time exponentials, is $T = 2$, which coincides with the observation time given by the group velocity (5.14).

Summarizing, when $0 < \mu < 1$, the sequence of eigenvalues has no uniform gap and the observability time (5.14) tends to infinity. Therefore, as in the semi-discrete case, a suitable filtering of the spurious numerical high frequencies is necessary. Theorem 2.1 provides a sharp result in this direction and its main result coincides with the predictions one may do in view of the structure of the dispersion diagram.

6. Proof of the discrete Ingham inequality. The proof of Theorem 2.1 uses in an essential way some classical properties of the discrete Fourier transform. We recall these properties in subsection 6.1 following [23].

FIG. 1. *Group velocity for the semi-discrete (–) and discrete (- -) cases with* $\mu = \Delta t/\Delta x = 0.1$ *(left),* $\mu = \Delta t/\Delta x = 0.3$ *(middle),* $\mu = \Delta t/\Delta x = 0.5$ *(right),* $\Delta x = 0.001$.

FIG. 2. *Group velocity for the semi-discrete (–) and discrete (- -) cases with* $\mu = \Delta t/\Delta x = 0.9$ *(left),* $\mu = \Delta t/\Delta x = 0.999$ *(middle),* $\mu = \Delta t/\Delta x = 1$ *(right),* $\Delta x = 0.001$.

6.1. The discrete Fourier transform. Let $h > 0$ be a real number and let $\dots, x_{-1}, x_0, x_1, \dots$ be defined by $x_j = jh$. Thus $\{x_j\} = h\mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. The ℓ_h^2-norm of a discrete function $\{v_j\}$ is defined as

$$\|v\|_h = \left[h \sum_{j=-\infty}^{\infty} |v_j|^2 \right]^{1/2}.$$

We denote by l_h^2 the Hilbert space $l_h^2 = \{v : \|v\|_h < \infty\}$, the space of discrete functions of finite $\|\cdot\|_h$ norm.

For any $v \in \ell_h^2$, the discrete Fourier transform of v is the function $\hat{v}$ defined by

$$\hat{v}(\xi) = h \sum_{j=-\infty}^{\infty} e^{-i\xi x_j} v_j, \qquad \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h} \right]. \tag{6.1}$$

This can be viewed as a discrete approximation of the continuous Fourier transform

$$\hat{u}(\xi) = \int_{-\infty}^{\infty} e^{-i\xi x} u(x) dx, \qquad \xi \in \mathbb{R},$$

if $u = u(x)$ is a sufficiently smooth function such that $u(x_j) = v_j$.

A priori, the sum in (6.1) defines a function $\hat{v}(\xi)$ for all $\xi \in \mathbb{R}$. The function $\hat{v}(\xi)$ is $2\pi/h$ periodic on $\mathbb{R}$ and therefore we analyze it only for $\xi \in [-\pi/h, \pi/h]$ to avoid aliasing.

Let us recall a standard definition. A function u defined on $\mathbb{R}$ is said to have *bounded variation* if there is a constant M such that for any finite m and any points

FIG. 3. *Dispersion relation for the continuous (-.-), semi-discrete (- -) and discrete (–) cases with* $\mu = \Delta t/\Delta x = 0.1$ *(left)* $\mu = \Delta t/\Delta x = 0.3$ *(middle),* $\mu = \Delta t/\Delta x = 0.5$ *(right),* $\Delta x = 0.001$.

FIG. 4. *Dispersion relation for the continuous (-.-), semi-discrete (- -) and discrete (–) cases with* $\mu = \Delta t/\Delta x = 0.9$ *(left),* $\mu = \Delta t/\Delta x = 0.999$ *(middle),* $\mu = \Delta t/\Delta x = 1$ *(right),* $\Delta x = 0.001$.

$x_0 < x_1 < \cdots < x_m$,

$$\sum_{j=1}^{m} \mid u(x_j) - u(x_{j-1}) \mid \leq M.$$

Now we give a fundamental result (see [23, p. 96]) which describes the effect of discretization in the Fourier transform.

THEOREM 6.1. *Suppose that* $u \in L^2(\mathbb{R})$ *is a sufficiently smooth function defined on* $\mathbb{R}$ *and let* $v \in \ell_h^2$ *be the discretization obtained by sampling* u *at the grid points* x_j, *i.e.,* $u(x_j) = v_j$.

Then, if u *has* $p-1$ *continuous derivatives in* $L^2(\mathbb{R})$ *for some* $p \geq 1$ *and a* p*th derivative in* L^2 *that has bounded variation, it follows that*

$$|\hat{v}(\xi) - \hat{u}(\xi)| = o(h^{p+1}), \quad \text{when } h \to 0, \tag{6.2}$$

uniformly on $\xi \in [-\pi/h, \pi/h]$.

Proof. Since u is continuous, apply *Poisson formula*

$$\hat{v}(\xi) = \sum_{j=-\infty}^{\infty} \hat{u}(\xi + 2\pi j/h), \qquad \xi \in [-\pi/h, \pi/h].$$

Thus, for every $u \in L^2(\mathbb{R})$ and $v \in \ell^2$, we obtain

$$\hat{v}(\xi) - \hat{u}(\xi) = \sum_{j=-\infty}^{\infty} \hat{u}(\xi + 2\pi j/h) - \hat{u}(\xi) = \sum_{j=1}^{\infty} [\hat{u}(\xi + 2\pi j/h) + \hat{u}(\xi - 2\pi j/h)],$$

with $\hat{u}$ and $\hat{v}$ the Fourier transforms of u and v, respectively.

If u verifies the hypothesis of Theorem 6.1, then

$$|\hat{u}(\xi)| \leq C_1 |\xi|^{-p-1}, \quad \text{when } \xi \to \infty$$

for some constant C_1. Therefore

$$|\hat{v}(\xi) - \hat{u}(\xi)| \leq C_1 \sum_{j=1}^{\infty} (j\pi/h)^{-p-1} = C_2 h^{p+1} \sum_{j=1}^{\infty} j^{-p-1}.$$

For every $p \geq 1$ this sum converges, which implies (6.2), as required. ∎

6.2. The discrete Fourier transform of Ingham's cut-off function. We study some general properties of a discrete Fourier transform that we shall use in the proof of the discrete Ingham inequality.

Given $M \in \mathbb{N}$ and $T > 0$ we consider the function $g : \mathbb{R} \to \mathbb{R}$

$$g(t) = \sin\left(\frac{t\pi}{T}\right) \chi_{(0,T)}, \tag{6.3}$$

where $\chi_{(0,T)}$ is the characteristic function of the interval $(0, T)$. Function (6.3) is precisely the same that Ingham [9] used in the proof of the continuous inequality (1.2). Its Fourier transform $G : \mathbb{R} \to \mathbb{R}$ is

$$G(\tau) = \int_{-\infty}^{\infty} g(t) e^{it\tau} dt = -2 \cos\frac{T\tau}{2} e^{\frac{iT\tau}{2}} \frac{\pi T}{(T^2\tau^2 - \pi^2)}. \tag{6.4}$$

We define the restriction of g to the grid

$$\ldots < t_{-1} < t_0 = 0 < t_1 < \cdots < t_{M+1} = T < \ldots,$$

with $t_n = n\Delta t$, i.e.,

$$h(n\Delta t) = g(t_n) = \sin(n\Delta t\pi/T)\chi_{M+1}(n),$$

χ_{M+1} being the characteristic function of the set $\{0, \ldots, M+1\}$.

For any $\tau \in \mathbb{R}$ we define the discrete Fourier transform of the discrete function h

$$H(\tau) := \Delta t \sum_{n=-\infty}^{\infty} h(n\Delta t) e^{in\Delta t\tau} \tag{6.5}$$

for all $\tau \in [-\pi/\Delta t, \pi/\Delta t]$.

LEMMA 6.2. *For all Δt, $T > 0$ and $k \in \mathbb{Z}$ we have*

$$H(\tau) = -\frac{\Delta t \cos\frac{T\tau}{2} e^{\frac{iT\tau}{2}} \sin\frac{\Delta t\pi}{T}}{2\sin(\frac{\Delta t}{2T}(T\tau + \pi))\sin(\frac{\Delta t}{2T}(T\tau - \pi))} \tag{6.6}$$

for any $\tau \neq (2k\pi)/\Delta t \pm \pi/T$ with $k \in \mathbb{Z}$ and

$$H\left(\frac{2k\pi}{\Delta t} \pm \frac{\pi}{T}\right) = \mp\frac{T}{2i}, \ k \in \mathbb{Z}. \tag{6.7}$$

The function H defined in (6.5) *is continuous and*

$$\lim_{\tau \to \frac{2k\pi}{\Delta t} \pm \frac{\pi}{T}} H(\tau) = \mp \frac{T}{2i}. \tag{6.8}$$

Proof (*Proof of Lemma* 6.2). We divide the proof into two steps: first, we prove that the explicit expression (6.5) of H is (6.6) and then we study the continuity of H.
• *Step* 1. *From the definition of the function H for all $\tau \neq 2k\pi/\Delta t \pm \pi/T$, $k \in \mathbb{Z}$, we have*

$$H(\tau) \quad = \Delta t \sum_{n=0}^{M} \sin \frac{n\pi\Delta t}{T} e^{in\Delta t\tau} = \Delta t \sum_{n=0}^{M} \frac{e^{\frac{in\pi\Delta t}{T}} - e^{-\frac{in\pi\Delta t}{T}}}{2i} e^{in\Delta t\tau}.$$

Hence

$$H(\tau) = \frac{\Delta t}{2i} \sum_{n=0}^{M} e^{\frac{in\Delta t}{T}(T\tau+\pi)} - \frac{\Delta t}{2i} \sum_{n=0}^{M} e^{\frac{in\Delta t}{T}(T\tau-\pi)}. \tag{6.9}$$

In order to obtain identity (6.5), it is useful to prove it only for any $|\tau| < 2k\pi/\Delta t - \pi/T$ with $k = 1$. Then, taking the periodicity properties of the complex exponentials into account, it is easy to obtain the same result for all $k \in \mathbb{Z}$, $\tau \neq 2k\pi/\Delta t \pm \pi/T$.

The first term on the right-hand side of (6.9) is

(6.10)
$$\begin{aligned}
&\sum_{n=0}^{M} e^{\frac{in\Delta t}{T}(T\tau+\pi)} = \frac{e^{\frac{i(M+1)\Delta t}{T}(T\tau+\pi)} - 1}{e^{\frac{i\Delta t}{T}(T\tau+\pi)} - 1} = \frac{e^{\frac{i(M+1)\Delta t}{T}(T\tau+\pi)} - 1}{\cos(\frac{\Delta t}{T}(T\tau+\pi)) + i\sin(\frac{\Delta t}{T}(T\tau+\pi)) - 1} \\
&= \frac{e^{i(T\tau+\pi)} - 1}{1 - 2\sin^2(\frac{\Delta t}{2T}(T\tau+\pi)) + 2i\sin(\frac{\Delta t}{2T}(T\tau+\pi))\cos(\frac{\Delta t}{2T}(T\tau+\pi)) - 1} \\
&= \frac{-e^{iT\tau} - 1}{2i\sin(\frac{\Delta t}{2T}(T\tau+\pi))(\cos(\frac{\Delta t}{2T}(T\tau+\pi)) + i\sin(\frac{\Delta t}{2T}(T\tau+1\pi)))} \\
&= \frac{-(e^{iT\tau}+1)e^{-\frac{i\Delta t}{2T}(T\tau+\pi)}}{2i\sin(\frac{\Delta t}{2T}(T\tau+\pi))}.
\end{aligned}$$

For the second one we have

(6.11)
$$\begin{aligned}
&\sum_{n=0}^{M} e^{\frac{in\Delta t}{T}(T\tau-\pi)} = \frac{e^{\frac{i(M+1)\Delta t}{T}(T\tau-\pi)} - 1}{e^{\frac{i\Delta t}{T}(T\tau-\pi)} - 1} = \frac{e^{\frac{i(M+1)\Delta t}{T}(T\tau-\pi)} - 1}{\cos\frac{\Delta t}{T}(T\tau-\pi) + i\sin\frac{\Delta t}{T}(T\tau-\pi) - 1} \\
&= \frac{e^{i(T\tau-\pi)} - 1}{-2\sin^2\frac{\Delta t}{2T}(T\tau-\pi) + 2i\sin\frac{\Delta t}{2T}(T\tau-\pi)\cos\frac{\pi\Delta t}{2T}(T\tau-\pi)} \\
&= \frac{-e^{iT\tau} - 1}{2i\sin\frac{\Delta t}{2T}(T\tau-\pi)e^{\frac{i\Delta t}{2T}(T\tau-\pi)}} = \frac{-(e^{iT\tau}+1)e^{-\frac{i\Delta t}{2T}(T\tau-\pi)}}{2i\sin\frac{\Delta t}{2T}(T\tau-\pi)}.
\end{aligned}$$

Substituting (6.10) and (6.11) into (6.9) we obtain

(6.12)

$$
\begin{aligned}
&H(\tau)\\
&= \frac{-\Delta t}{4}\left[\frac{(e^{iT\tau}+1)e^{-\frac{i\Delta t}{2T}(T\tau-\pi)}}{\sin\frac{\Delta t}{2T}(T\tau-\pi)} - \frac{(e^{iT\tau}+1)e^{-\frac{i\Delta t}{2T}(T\tau+\pi)}}{\sin(\frac{\Delta t}{2T}(T\tau+\pi))}\right]\\
&= \frac{-\Delta t(e^{iT\tau}+1)}{4}\left(\frac{\cos\frac{\Delta t}{2T}(T\tau-\pi) - i\sin\frac{\Delta t}{2T}(T\tau-\pi)}{\sin\frac{\Delta t}{2T}(T\tau-\pi)}\right.\\
&\left.-\frac{\cos\frac{\Delta t}{2T}(T\tau+\pi) - i\sin(\frac{\Delta t}{2T}(T\tau+\pi))}{\sin(\frac{\Delta t}{2T}(T\tau+\pi))}\right)\\
&= \frac{-\Delta t(e^{iT\tau}+1)}{4}\left(\frac{\cos\frac{\Delta t}{2T}(T\tau-\pi)}{\sin\frac{\Delta t}{2T}(T\tau-\pi)} - \frac{\cos\frac{\Delta t}{2T}(T\tau+\pi)}{\sin(\frac{\Delta t}{2T}(T\tau+\pi))}\right)\\
&= \frac{-\Delta t(e^{iT\tau}+1)}{4}\left(\frac{\cos\frac{\Delta t}{2T}(T\tau-\pi)\sin\frac{\Delta t}{2T}(T\tau+\pi) - \cos\frac{\Delta t}{2T}(T\tau+\pi)\sin\frac{\Delta t}{2T}(T\tau-\pi)}{\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin(\frac{\Delta t}{2T}(T\tau+\pi)}\right)\\
&= \frac{-\Delta t(e^{iT\tau}+1)}{4\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin\frac{\pi\Delta t}{2T}(T\tau+\pi)}\sin\frac{\pi\Delta t}{T}.
\end{aligned}
$$

Therefore, applying Euler's formula in (6.12) we obtain the following expression for H:

(6.13)

$$
\begin{aligned}
H(\tau) &= \frac{-\Delta t\left(\cos(T\tau)+i\sin(T\tau)+1\right)\sin\frac{\pi\Delta t}{T}}{4\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin(\frac{\Delta t}{2T}(T\tau+\pi))}\\
&= \frac{-\Delta t\left(2\cos^2\frac{T\tau}{2}-1+2i\sin\frac{T\tau}{2}\cos\frac{T\tau}{2}+1\right)\sin\frac{\pi\Delta t}{T}}{4\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin(\frac{\Delta t}{2T}(T\tau+\pi))}\\
&= \frac{-2\Delta t\cos\frac{T\tau}{2}\left(\cos\frac{T\tau}{2}+i\sin\frac{T\tau}{2}\right)\sin\frac{\pi\Delta t}{T}}{4\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin(\frac{\Delta t}{2T}(T\tau+\pi)}\\
&= \frac{-\Delta t\cos\frac{T\tau}{2}e^{\frac{iT\tau}{2}}\sin\frac{\pi\Delta t}{T}}{2\sin\frac{\Delta t}{2T}(T\tau-\pi)\sin(\frac{\pi\Delta t}{2T}(T\tau+\pi))}.
\end{aligned}
$$

Moreover, if $\tau = 2k\pi/\Delta t + \pi/T$, with $k \in \mathbb{Z}$, using the definition (6.5) of H, we deduce that

$$
\begin{aligned}
H(\tau) &= \Delta t\sum_{n=0}^{M}\sin\frac{n\pi\Delta t}{T}e^{in\Delta t\left(\frac{2k\pi}{\Delta t}+\frac{\pi}{T}\right)}\\
&= \Delta t\sum_{n=0}^{M}\frac{e^{\frac{in\Delta t\pi}{T}}-e^{-\frac{in\Delta t\pi}{T}}}{2i}e^{2k\pi in}e^{\frac{in\Delta t\pi}{T}} = \frac{\Delta t}{2i}\sum_{n=0}^{M}e^{\frac{2in\Delta t\pi}{T}} - \frac{\Delta t}{2i}\sum_{n=0}^{M}1\\
&= \frac{\Delta t}{2i}\frac{e^{\frac{2i(M+1)\Delta t\pi}{T}}-1}{e^{\frac{2i\Delta t\pi}{T}}-1} - \frac{\Delta t}{2i}(M+1) = \frac{\Delta t}{2i}\frac{e^{2\pi i}-1}{e^{\frac{2i\Delta t\pi}{T}}-1} - \frac{T}{2i} = -\frac{T}{2i}.
\end{aligned}
$$

For every $\tau = 2k\pi/\Delta t - \pi/T$, with $k \in \mathbb{Z}$,

$$\begin{aligned} H(\tau) &= \Delta t \sum_{n=0}^{M} \sin \frac{n\pi\Delta t}{T} e^{in\Delta t\left(\frac{2k\pi}{\Delta t} - \frac{\pi}{T}\right)} \\ &= \Delta t \sum_{n=0}^{M} \frac{e^{\frac{in\Delta t\pi}{T}} - e^{-\frac{in\Delta t\pi}{T}}}{2i} e^{2k\pi in} e^{\frac{-in\Delta t\pi}{T}} \\ &= \frac{\Delta t}{2i} \sum_{n=0}^{M} 1 - \frac{\Delta t}{2i} e^{\frac{-2i(M+1)\Delta t\pi}{T}} \sum_{n=0}^{M} e^{\frac{2in\Delta t\pi}{T}} \\ &= \frac{\Delta t}{2i}(M+1) - \frac{\Delta t}{2i} \frac{e^{\frac{2i(M+1)\Delta t\pi}{T}} - 1}{e^{\frac{2i\Delta t\pi}{T}} - 1} = \frac{T}{2i} - \frac{\Delta t}{2i} \frac{e^{2\pi i} - 1}{e^{\frac{2i\Delta t\pi}{T}} - 1} = \frac{T}{2i}. \end{aligned}$$

• *Step 2. It is easy to see that H is continuous on $\mathbb{R}\backslash\{\tau : \tau = 2k\pi/\Delta t \pm \pi/T\}$, $k \in \mathbb{Z}$. We now study the continuity of H at the singularities $\tau = 2k\pi/\Delta t \pm \pi/T$. For every $\tau \to 2k\pi/\Delta t \pm \pi/T$, we have $\tau = 2k\pi/\Delta t \pm \pi/T + \varepsilon\pi$, with $\varepsilon \to 0$.*

1. *The case $\tau = 2k\pi/\Delta t + \pi/T + \varepsilon\pi$ with $\varepsilon \to 0$.*
   Using the definition of H we have

$$\begin{aligned} H\left(\frac{2k\pi}{\Delta t} + \frac{\pi}{T} + \varepsilon\pi\right) &= \Delta t \sum_{n=0}^{M} \sin \frac{n\pi\Delta t}{T} e^{in\Delta t\pi\left(\frac{2k}{\Delta t} + \frac{1}{T} + \varepsilon\right)} \\ &= \Delta t \sum_{n=0}^{M} \frac{e^{\frac{in\Delta t\pi}{T}} - e^{-\frac{in\Delta t\pi}{T}}}{2i} e^{in\pi 2k} e^{in\Delta t\pi\left(\frac{1}{T} + \varepsilon\right)}. \end{aligned}$$

   Hence,

$$H\left(\frac{2k\pi}{\Delta t} + \frac{\pi}{T} + \varepsilon\pi\right) = \frac{\Delta t}{2i} \sum_{n=0}^{M} e^{\frac{in\Delta t\pi}{T}(2+T\varepsilon)} - \frac{\Delta t}{2i} \sum_{n=0}^{M} e^{in\Delta t\pi\varepsilon}. \tag{6.14}$$

   For every $|x| < 2T/\Delta t$, according to the classical formula for the sum of a geometric series, the following identity holds:

$$\sum_{n=0}^{M} e^{\frac{in\Delta t\pi}{T}x} = \frac{e^{\frac{i(M+1)\Delta t\pi}{T}x} - 1}{e^{\frac{i\Delta t\pi}{T}x} - 1} = \frac{e^{i\pi x} - 1}{e^{\frac{i\Delta t\pi}{T}x} - 1}. \tag{6.15}$$

   For the first sum entering on the right-hand term of (6.14) we have

$$\sum_{n=0}^{M} e^{\frac{in\Delta t\pi}{T}(2+T\varepsilon)} = \frac{e^{2\pi i} e^{iT\pi\varepsilon} - 1}{e^{\frac{i\Delta t\pi}{T}(2+T\varepsilon)} - 1} = \frac{e^{iT\pi\varepsilon} - 1}{e^{\frac{i\Delta t\pi}{T}(2+T\varepsilon)} - 1}. \tag{6.16}$$

   For the second sum on the right-hand term (6.14), using (6.15) and the fact that $\Delta t\pi/T(2 + T\varepsilon) < 2\pi$, we have

$$\sum_{n=0}^{M} e^{in\Delta t\pi\varepsilon} = \frac{e^{iT\pi\varepsilon} - 1}{e^{i\Delta t\pi\varepsilon} - 1}. \tag{6.17}$$

Finally, replacing (6.16) and (6.17) in (6.14) and taking the limit $\varepsilon \to 0$ we obtain

$$\lim_{\varepsilon\to 0} H\left(\frac{2k\pi}{\Delta t}+\frac{\pi}{T}+\varepsilon\pi\right)= \frac{\Delta t}{2i}\lim_{\varepsilon\to 0}\frac{e^{iT\pi\varepsilon}-1}{e^{\frac{i\Delta t\pi}{T}(2+T\varepsilon)}-1}-\frac{\Delta t}{2i}\lim_{\varepsilon\to 0}\frac{e^{iT\pi\varepsilon}-1}{e^{i\Delta t\pi\varepsilon}-1}$$

$$= 0-\frac{\Delta t}{2i}\lim_{\varepsilon\to 0}\frac{e^{iT\pi\varepsilon}-1}{e^{i\Delta t\pi\varepsilon}-1}=-\frac{\Delta t}{2i}\lim_{\varepsilon\to 0}\frac{i\pi Te^{iT\pi\varepsilon}}{i\Delta t\pi e^{i\Delta t\pi\varepsilon}}=-\frac{T}{2i}.$$

2. *The case $\tau = 2k\pi/\Delta t - \pi/T + \varepsilon\pi$ with $\varepsilon \to 0$.*
We have

$$H\left(\frac{2k\pi}{\Delta t}-\frac{\pi}{T}+\varepsilon\pi\right)=\Delta t\sum_{n=0}^{M}\sin\frac{n\pi\Delta t}{T}e^{in\Delta t\pi\left(\frac{2k}{\Delta t}-\frac{1}{T}+\varepsilon\right)}$$

$$=\Delta t\sum_{n=0}^{M}\frac{e^{\frac{in\Delta t\pi}{T}}-e^{-\frac{in\Delta t\pi}{T}}}{2i}e^{in\pi 2k}e^{in\Delta t\pi\varepsilon}e^{-\frac{in\Delta t\pi}{T}}.$$

Hence

$$H\left(\frac{2k\pi}{\Delta t}-\frac{\pi}{T}+\varepsilon\pi\right)=\frac{\Delta t}{2i}\sum_{n=0}^{M}e^{in\Delta t\pi\varepsilon}-\frac{\Delta t}{2i}\sum_{n=0}^{M}e^{\frac{in\Delta t\pi}{T}(T\varepsilon-2)}. \tag{6.18}$$

By (6.17), for the first sum entering on the right-hand term of (6.18) we have

$$\sum_{n=0}^{M}e^{in\Delta t\pi\varepsilon}=\frac{e^{iT\pi\varepsilon}-1}{e^{i\Delta t\pi\varepsilon}-1}. \tag{6.19}$$

Moreover, in (6.17), applying the identity (6.15) for the second sum on the right-hand term of (6.18), we obtain

$$\sum_{n=0}^{M}e^{-\frac{in\Delta t\pi}{T}(2-T\varepsilon)}=e^{-\frac{i(M+1)\Delta t\pi}{T}(2-T\varepsilon)}\sum_{n=0}^{M}e^{\frac{in\Delta t\pi}{T}(2-T\varepsilon)}$$

$$=\!= e^{-2\pi i}e^{iT\pi\varepsilon}\frac{e^{\frac{i(M+1)\Delta t\pi}{T}(2-T\varepsilon)}-1}{e^{\frac{i\Delta t\pi}{T}(2-T\varepsilon)}-1}=e^{iT\pi\varepsilon}\frac{e^{2\pi i}e^{-iT\pi\varepsilon}-1}{e^{\frac{i\Delta t\pi}{T}(2-T\varepsilon)}-1}$$

$$=\frac{1-e^{iT\pi\varepsilon}}{e^{\frac{i\Delta t\pi}{T}(2-T\varepsilon)}-1}\to 0,\quad \varepsilon\to 0. \tag{6.20}$$

Hence,

$$\lim_{\varepsilon\to 0}H\left(\frac{2k\pi}{\Delta t}-\frac{\pi}{T}+\varepsilon\pi\right)=\frac{T}{\Delta t}\frac{\Delta t}{2i}=\frac{T}{2i}.$$

This concludes the proof of Lemma 6.2. □

Remark 6.3. For $\tau = 0$ in (6.6) we have

$$H(0) \quad =-\frac{\Delta t\sin\frac{\Delta t\pi}{T}}{2\sin\frac{\pi\Delta t}{2T}\sin\left(-\frac{\pi\Delta t}{2T}\right)}=-\frac{\Delta t\sin\frac{\Delta t\pi}{T}}{-2\sin^2\frac{\pi\Delta t}{2T}}$$

$$=\frac{\Delta t2\sin\frac{\Delta t\pi}{2T}\cos\frac{\Delta t\pi}{2T}}{2\sin^2\frac{\pi\Delta t}{2T}}=\frac{\Delta t\cos\frac{\Delta t\pi}{2T}}{\sin\frac{\pi\Delta t}{2T}}=\Delta t\cot\frac{\Delta t\pi}{2T}. \tag{6.21}$$

Taking the limit $\Delta t \to 0$ in (6.6), for every τ fixed, we obtain

$$\lim_{\Delta t \to 0} H(\tau) = -2\cos\frac{T\tau}{2} e^{\frac{iT\tau}{2}} \frac{\pi T}{T^2\tau^2 - \pi^2} = G(\tau) \tag{6.22}$$

and this is the classical Fourier transform of g given by (6.4).

6.3. Proof of Theorem 2.1. *This section is devoted to the proof of the main result of this paper. The proof of the discrete inequality (2.3) follows the strategy used in [26, (pp. 162–163)] to prove the classical Ingham inequality (1.2).*

Proof (Proof of the first (so-called inverse) inequality in (2.3)). We prove the first inequality in (2.3), namely,

$$C_1(\Delta t, T, \gamma) \sum_{k=-N}^{N} |a_k|^2 \le \Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} \right|^2 .$$

Taking into account that $\sin(n\Delta t\pi/T) \le 1$, *we have*

$$\begin{aligned} \Delta t \sum_{n=0}^{M} \left| \sum_k a_k e^{in\Delta t\lambda_k} \right|^2 &\ge \Delta t \sum_{n=0}^{M} \sin\frac{n\Delta t\pi}{T} \left| \sum_k a_k e^{in\Delta t\lambda_k} \right|^2 \\ &= \Delta t \sum_{n=0}^{M} \sin\frac{n\Delta t\pi}{T} \sum_k \sum_l a_k \bar{a}_l e^{in\Delta t(\lambda_k - \lambda_l)}. \end{aligned}$$

The function H defined by (6.6) is continuous, hence

$$\begin{aligned} &\Delta t \sum_{n=0}^{M} \sin\frac{n\Delta t\pi}{T} \sum_k \sum_l a_k \bar{a}_l e^{in\Delta t(\lambda_k - \lambda_l)} \\ &= \sum_k \sum_l a_k \bar{a}_l H(\lambda_k - \lambda_l) = H(0) \sum_k |a_k|^2 + \sum_k \sum_{l, l\neq k} a_k \bar{a}_l H(\lambda_k - \lambda_l) \\ &\ge H(0) \sum_k |a_k|^2 - \frac{1}{2} \sum_k \sum_{l, l\neq k} \left(|a_k|^2 + |a_l|^2 \right) |H(\lambda_k - \lambda_l)| \\ &= H(0) \sum_k |a_k|^2 - \sum_k |a_k|^2 \sum_{l, l\neq k} |H(\lambda_k - \lambda_l)| . \end{aligned} \tag{6.23}$$

In the last term in (6.23) we have

$$\sum_{l, k\neq l} |H(\lambda_k - \lambda_l)| = \sum_{\substack{l, k\neq l \\ |\lambda_k - \lambda_l| \le \frac{\pi}{\Delta t}}} |H(\lambda_k - \lambda_l)| + \sum_{\substack{l, k\neq l \\ |\lambda_k - \lambda_l| > \frac{\pi}{\Delta t}}} |H(\lambda_k - \lambda_l)| . \tag{6.24}$$

Moreover, the function H is periodic with period $2\pi/\Delta t$. Consequently, for every $k, l \in \mathbb{Z}$ with $\pi/\Delta t < |\lambda_k - \lambda_l| < 2\pi/\Delta t$, there exist $m_{k,l} \in [-\pi/\Delta t, \pi/\Delta t]$ such that $|m_{k,l}| = 2\pi/\Delta t - |\lambda_k - \lambda_l|$ with the property $H(\lambda_k - \lambda_l) = H(m_{k,l})$. Therefore, using this periodicity property and applying (6.2) from Theorem 6.1 and (6.24) in (6.23),

we obtain

(6.25)
$$\Delta t\sum_{n=0}^{M}\left|\sum_k a_k e^{in\Delta t\lambda_k}\right|^2$$
$$\geq H(0)\sum_k |a_k|^2-\sum_k |a_k|^2\left(\sum_{\substack{l,k\neq l\\ |\lambda_k-\lambda_l|\leq \pi/\Delta t}}|H(\lambda_k-\lambda_l)|+\sum_{\substack{l,k\neq l\\ |m_{k,l}|\leq \pi/\Delta t}}|H(m_{k,l})|\right)$$
$$\geq H(0)\sum_k |a_k|^2-\sum_k |a_k|^2\left(\sum_{\substack{l,k\neq l\\ |\lambda_k-\lambda_l|\leq \pi/\Delta t}}|G(\lambda_k-\lambda_l)|+\sum_{\substack{l,k\neq l\\ |m_{k,l}|\leq \pi/\Delta t}}|G(m_{k,l})|\right)$$
$$+CN(\Delta t)^2\Big).$$

On the other hand, as pointed out in [26, p. 162], for every sequence $\{\lambda_k\}$ satisfying the gap condition (2.1), the function G satisfies

(6.26)
$$\sum_{l\neq k,l=-N}^{N}|G(\lambda_k-\lambda_l)|\leq 2\pi T\sum_{l=-\infty,l\neq k}^{\infty}\frac{1}{T^2\left(\lambda_k-\lambda_l\right)^2-\pi^2}$$
$$\leq 2\pi T\sum_{l=-\infty,l\neq k}^{\infty}\frac{1}{T^2\gamma^2\left(k-l\right)^2-\pi^2}=4\pi T\sum_{r\geq 1}\frac{1}{\frac{T^2\gamma^2}{4\pi^2}4\pi^2r^2-\pi^2}\leq\frac{16\pi}{T\gamma^2}\sum_{r\geq 1}\frac{1}{4r^2-1}$$
$$=\frac{8\pi}{T\gamma^2}\sum_{r\geq 1}\left(\frac{1}{2r-1}-\frac{1}{2r+1}\right)=\frac{8\pi}{T\gamma^2}.$$

Further, for the terms of the sequence $\{\lambda_k\}$ satisfying $\pi/\Delta t <\mid \lambda_k-\lambda_l \mid< (2\pi-(\Delta t)^p)/\Delta t$, (and then, $(\Delta t)^{p-1}\leq |m_{k,l}|\leq \pi/\Delta t$), $k\neq l$, we have

$$|G(m_{k,l})|\leq 2\pi T\frac{1}{T^2\left(m_{k,l}\right)^2-\pi^2}=2\pi T\frac{1}{T^2\left(\frac{2\pi}{\Delta t}-(\lambda_l-\lambda_l)\right)^2-\pi^2}$$
$$\leq 2\pi T\frac{1}{T^2\left(\frac{2\pi}{\Delta t}-\frac{2\pi-(\Delta t)^p}{\Delta t}\right)^2-\pi^2}=2\pi T\frac{\Delta t^2}{T^2(\Delta t)^{2p}-\pi^2\Delta t^2}$$

and it follows that

$$\sum_{l\neq k,l=-N}^{N}|G(m_{k,l})|\leq (N\Delta t)2\pi T\frac{(\Delta t)^{1-2p}}{T^2-\pi^2(\Delta t)^{2-2p}}. \tag{6.27}$$

Using the relations (6.26) and (6.27) in (6.25) we obtain

$$\Delta t\sum_{n=0}^{M}\sin\frac{n\Delta t\pi}{T}\sum_k\sum_l a_k\bar{a}_l e^{in\Delta t(\lambda_k-\lambda_l)}$$
$$\geq H(0)\sum_k |a_k|^2-\sum_k |a_k|^2\left[\frac{8\pi}{T\gamma^2}+NC\Delta t^2+CN\Delta t(\Delta t)^{1-2p}\right] \tag{6.28}$$

when $\Delta t \to 0$. *For the function* $H(0)$ *given by (6.21) we have* $\lim_{\Delta t \to 0} H(0) = 2T/\pi$, *which is equivalent to* $2T/\pi - \theta \le H(0) \le 2T/\pi + \theta$, *with* $\theta \to 0$ *when* $\Delta t \to 0$. *In order to ensure the positivity of all the coefficients* $|a_k|^2$ *in (6.28) it is necessary and sufficient to have*

$$C_1(\Delta t, T, \gamma) := H(0) - \frac{8\pi}{T\gamma^2} - (NC\Delta t^2 + CN\Delta t(\Delta t)^{1-2p}) > 0, \tag{6.29}$$

which is equivalent to

$$T^2 - \frac{T\pi}{2}(\theta + \varepsilon_1) - \frac{4\pi^2}{\gamma^2} > 0,$$

where $\varepsilon_1 = NC\Delta t^2 + CN\Delta t(\Delta t)^{1-2p}$, $C > 0$. *This condition holds for every*

$$T(\Delta t) > T_0(\Delta t) = \frac{\frac{\pi}{2}(\varepsilon_1 + \theta) + \sqrt{\frac{\pi^2}{4}(\theta + \varepsilon_1)^2 + \frac{16\pi^2}{\gamma^2}}}{2} := \frac{2\pi}{\gamma} + \epsilon(\Delta t)$$

with $\epsilon(\Delta t) = C(\Delta t + N\Delta t(\Delta t)^{1-2p})$. *Hence, the inequality (2.3) holds with the constant* $C_1(\Delta t, T, \gamma)$ *defined by the relation (6.29) where* $\delta_1(\Delta t) = -(\varepsilon_1 + \theta)$.

Proof of the second (so-called direct) inequality. We now prove the inequality

$$\Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} \right|^2 \le C_2(\Delta t, T, \gamma) \sum_{k=-N}^{N} |a_k|^2 .$$

We have

$$\Delta t \sum_{n=0}^{M} \left| \sum_{k} a_k e^{in\Delta t\lambda_k} \right|^2 = \Delta t \sum_{n=0}^{\left[\frac{M}{2}\right]} \left| \sum_{k} a_k e^{in\Delta t\lambda_k} \right|^2 + \Delta t \sum_{n=\left[\frac{M}{2}\right]+1}^{M} \left| \sum_{k} a_k e^{in\Delta t\lambda_k} \right|^2 . \tag{6.30}$$

Consider the first term on the right-hand side of (6.30),

$$\begin{aligned} \Delta t \sum_{n=0}^{\left[\frac{M}{2}\right]} \left| \sum_{k} a_k e^{in\Delta t\lambda_k} \right|^2 &= \Delta t \sum_{n=\left[\frac{M+1}{4}\right]+1}^{\left[\frac{M}{2}\right]+\left[\frac{M+1}{4}\right]+1} \left| \sum_{k} a_k e^{i\left(n-\left[\frac{M+1}{4}\right]-1\right)\Delta t\lambda_k} \right|^2 \\ &= \Delta t \sum_{n=\left[\frac{M+1}{4}\right]+1}^{\left[\frac{M}{2}\right]+\left[\frac{M+1}{4}\right]} \left| \sum_{k} a_k e^{i\left(n-\left[\frac{M+1}{4}\right]-1\right)\Delta t\lambda_k} \right|^2 + \Delta t \left| \sum_{k} a_k e^{i\left[\frac{M}{2}\right]\Delta t\lambda_k} \right|^2 . \end{aligned} \tag{6.31}$$

Using the properties of the entire part of a real number we have

$$\begin{aligned} &\left[\frac{M+1}{4}\right] \le \frac{M+1}{4} \le \left[\frac{M+1}{4}\right] + 1, \\ &\left[\frac{M}{2}\right] + \left[\frac{M+1}{4}\right] \le \left[\frac{3M+1}{4}\right], \\ &\left[\frac{3M+1}{4}\right] \le \frac{3M+1}{4} \le \left[\frac{3M+1}{4}\right] + 1. \end{aligned}$$

For every $n \in \mathbb{N}$ *with* $\left[\frac{M+1}{4}\right]+1 \leq n \leq \left[\frac{3M+1}{4}\right]$, *we have*

$$\frac{M+1}{4} \leq n \leq \frac{3M+1}{4} \tag{6.32}$$

and

$$\frac{\pi}{4} \leq \frac{n\pi\Delta t}{T} \leq \frac{3\pi}{4},$$

due to the fact that $(M+1)\Delta t = T$.

Therefore, for every $n \in \mathbb{N}$ *as in (6.32) we have* $\sin(n\pi\Delta t/T) \geq \sqrt{2}/2$ *and*

$$\begin{aligned}
&\Delta t \sum_{n=\left[\frac{M+1}{4}\right]+1}^{\left[\frac{M}{2}\right]+\left[\frac{M+1}{4}\right]} \left|\sum_k a_k e^{i\left(n-\left[\frac{M+1}{4}\right]-1\right)\Delta t\lambda_k}\right|^2 + \Delta t\left|\sum_k a_k e^{i\left[\frac{M}{2}\right]\Delta t\lambda_k}\right|^2 \\
&\leq 2\Delta t \sum_{n=\left[\frac{M+1}{4}\right]+1}^{\left[\frac{M}{2}\right]+\left[\frac{M+1}{4}\right]} \sin\frac{n\pi\Delta t}{T}\left|\sum_k a_k e^{i\left(n-\left[\frac{M+1}{4}\right]-1\right)\Delta t\lambda_k}\right|^2 + \Delta t\left|\sum_k a_k e^{i\left[\frac{M}{2}\right]\Delta t\lambda_k}\right|^2 \\
&\leq 2\Delta t \sum_{n=0}^{M} \sin\frac{n\pi\Delta t}{T}\left|\sum_k a_k e^{in\Delta t\lambda_k} e^{-i\left(\left[\frac{M+1}{4}\right]+1\right)\Delta t\lambda_k}\right|^2 + \Delta t\left|\sum_k a_k e^{i\left[\frac{M}{2}\right]\Delta t\lambda_k}\right|^2 \\
&= 2\Delta t \sum_{n=0}^{M} \sin\frac{n\pi\Delta t}{T} \sum_k\sum_l a_k\bar{a}_l e^{in\Delta t(\lambda_k-\lambda_l)} e^{-i\left(\left[\frac{M+1}{4}\right]+1\right)\Delta t(\lambda_k-\lambda_l)} \\
&+\Delta t \sum_k\sum_l a_k\bar{a}_l e^{2i\left[\frac{M}{2}\right]\Delta t(\lambda_k-\lambda_l)} \\
&= 2\sum_k\sum_l a_k\bar{a}_l H(\lambda_k-\lambda_l) e^{-i\left(\left[\frac{M+1}{4}\right]+1\right)\Delta t(\lambda_k-\lambda_l)} + \Delta t\sum_k\sum_l a_k\bar{a}_l e^{i\left[\frac{M}{2}\right]\Delta t(\lambda_k-\lambda_l)} \\
&= 2H(0)\sum_k |a_k|^2 + 2\sum_k\sum_{l,l\neq k} a_k\bar{a}_l H(\lambda_k-\lambda_l) e^{-i\left(\left[\frac{M+1}{4}\right]+1\right)\Delta t(\lambda_k-\lambda_k)} \\
&+\Delta t\sum_k |a_k|^2 + \Delta t\sum_k\sum_{l,l\neq k} a_k\bar{a}_l e^{i\left[\frac{M}{2}\right]\Delta t(\lambda_k-\lambda_k)} \\
&\leq 2H(0)\sum_k |a_k|^2 + \sum_k\sum_{l,l\neq k}\left(|a_k|^2+|a_l|^2\right)|H(\lambda_k-\lambda_l)| + \Delta t\sum_k |a_k|^2
\end{aligned}$$

$$\begin{aligned}
&+\frac{N\Delta t}{2}\sum_k\sum_{l,l\neq k}\left(|a_k|^2+|a_l|^2\right) \\
&\leq 2H(0)\sum_k |a_k|^2 + 2\sum_k |a_k|^2 \sum_{l,l\neq k}|H(\lambda_k-\lambda_l)| + \Delta t\sum_k |a_k|^2 + 2N\Delta t\sum_k |a_k|^2.
\end{aligned} \tag{6.33}$$

Using the same argument (6.2) as in the proof of the inverse inequality, for every $C > 0$, *we have*

$$\sum_{l,k\neq l} |H(\lambda_k - \lambda_l)| \leq \sum_{l\neq k, l=-N}^{N} |G(\lambda_k - \lambda_l)| + CN(\Delta t)^2, \tag{6.34}$$

when Δt *is small enough, with* G *the Fourier transform (6.4) satisfying (6.26) and (6.27).*

Therefore, for every k,

$$\begin{aligned}
&2H(0)\sum_k |a_k|^2 + \sum_k \sum_{l,l\neq k} \left(|a_k|^2 + |a_l|^2\right) |H(\lambda_k - \lambda_l)| + \Delta t \sum_k |a_k|^2 \\
&+\frac{N\Delta t}{2} \sum_k \sum_{l,l\neq k} \left(|a_k|^2 + |a_l|^2\right) \leq 2\Delta t \cot\frac{\Delta t\pi}{2T} \sum_k |a_k|^2 \\
&+2\sum_k |a_k|^2 \left(\frac{8\pi}{T\gamma^2} + NC\Delta t^2 + CN\Delta t(\Delta t)^{1-2p} + 2\Delta t\right) + 2N\Delta t \sum_k |a_k|^2 .
\end{aligned}$$

Hence

$$\Delta t \sum_{n=0}^{\left[\frac{M}{2}\right]} \left|\sum_k a_k e^{in\Delta t\lambda_k}\right|^2 \leq \sum_k |a_k|^2 \left(2\Delta t \cot\frac{\Delta t\pi}{2T} + \frac{16\pi}{T\gamma^2} + \varepsilon(\Delta t)\right), \tag{6.35}$$

with $\varepsilon(\Delta t) = 2NC\Delta t^2 + 2CN\Delta t(\Delta t)^{1-2p} + 2\Delta t + 2N\Delta t$.

For the second right-hand term of (6.30) we have

$$\begin{aligned}
&\Delta t \sum_{n=\left[\frac{M}{2}\right]+1}^{M} \left|\sum_k a_k e^{in\Delta t\lambda_k}\right|^2 = \Delta t \sum_{n=\left[\frac{M}{2}\right]-\left[\frac{M+1}{4}\right]}^{M-\left[\frac{M+1}{4}\right]-1} \left|\sum_k a_k e^{i(n+\left[\frac{M+1}{4}\right]+1)\Delta t\lambda_k}\right|^2 \\
&= \Delta t \sum_{n=\left[\frac{M}{2}\right]+1-\left[\frac{M+1}{4}\right]}^{M-\left[\frac{M+1}{4}\right]-1} \left|\sum_k a_k e^{i(n+\left[\frac{M+1}{4}\right]+1)\Delta t\lambda_k}\right|^2 + \Delta t \sum_k \left|a_k e^{i(\left[\frac{M}{2}\right]+1)\Delta t\lambda_k}\right|^2 .
\end{aligned}$$

Taking into account that

$$M - \left[\frac{M+1}{4}\right] - 1 \leq \frac{3M+1}{4} \text{ and } \left[\frac{M}{2}\right] - \left[\frac{M+1}{4}\right] + 1 \geq \frac{M+1}{4},$$

for every $n \in \mathbb{N}$, *with* $(M+1)/4 \leq n \leq (3M+1)/4$ *we have* $\sin(n\pi\Delta t/T) \geq \sqrt{2}/2$.

Thus,

$$\begin{aligned}
&\Delta t \sum_{n=\left[\frac{M}{2}\right]+1}^{M} \left| \sum_k a_k e^{in\Delta t\lambda_k} \right|^2 \\
&\leq 2\Delta t \sum_{n=\left[\frac{M}{2}\right]+1-\left[\frac{M+1}{4}\right]}^{M-\left[\frac{M+1}{4}\right]-1} \sin\frac{n\pi\Delta t}{T} \left| \sum_k a_k e^{i(n+\left[\frac{M+1}{4}\right]+1)\Delta t\lambda_k} \right|^2 \\
&+\Delta t \sum_k \left| a_k e^{i(\left[\frac{M}{2}\right]+1)\Delta t\lambda_k} \right|^2 \\
&\leq 2\Delta t \sum_{n=0}^{M} \sin\frac{n\pi\Delta t}{T} \left| \sum_k a_k e^{in\Delta t\lambda_k} e^{i(\left[\frac{M+1}{4}\right]+1)\Delta t\lambda_k} \right|^2 + \Delta t \sum_k \left| a_k e^{i(\left[\frac{M}{2}\right]+1)\Delta t\lambda_k} \right|^2 \\
&= 2\Delta t \sum_{n=0}^{M} \sin\frac{n\pi\Delta t}{T} \sum_k \sum_l a_k \bar{a}_l e^{in\Delta t(\lambda_k-\lambda_l)} e^{i\left(\left[\frac{M+1}{4}\right]+1\right)\Delta t(\lambda_k-\lambda_k)} \\
&+\Delta t \sum_k \left| a_k e^{i(\left[\frac{M}{2}\right]+1)\Delta t\lambda_k} \right|^2
\end{aligned}$$

and we obtain the estimate

(6.36)
$$\begin{aligned}
&\Delta t \sum_{n=\left[\frac{M}{2}\right]+1}^{M} \left| \sum_k a_k e^{in\Delta t\lambda_k} \right|^2 \\
&\leq 2H(0) \sum_k |a_k|^2 + 2 \sum_k \sum_{l,l\neq k} |a_k|^2 |H(\lambda_k-\lambda_l)| + 2N\Delta t \sum_k |a_k|^2 \\
&\leq \left(2\Delta t \cot\frac{\Delta t\pi}{2T} + \frac{16\pi}{T\gamma^2} + NC\Delta t^2 + CN\Delta t(\Delta t)^{1-2p} + 2N\Delta t \right) \sum_k |a_k|^2 .
\end{aligned}$$

For the function $H(0)$ defined by the relation (6.21) we have $\lim_{\Delta t\to 0} H(0) = 2T/\pi$.

From (6.30) and (6.35) we get

$$\Delta t \sum_{n=0}^{M} \left| \sum_k a_k e^{in\Delta t\lambda_k} \right|^2 \leq \left(\frac{8T}{\pi} + \frac{32\pi}{T\gamma^2} + \delta_2(\Delta t) \right) \sum_k |a_k|^2 , \tag{6.37}$$

with

$$\delta_2(\Delta t) = 4\varepsilon(\Delta t) + \theta,$$

($2T/\pi - \theta \leq H(0) \leq 2T/\pi + \theta$, with $\theta \to 0$ when $\Delta t \to 0$).

This concludes the proof of Theorem 2.1. □

Proof (Proof of Theorem 2.2). Following the same steps of the above proof we obtain the discrete version (2.5) of the L^1 Ingham's inequality (1.6) given by Theorem 2.2.

More precisely, we have

$$\Delta t \sum_{n=0}^{M} \sin \frac{n\Delta t\pi}{T} \left(\sum_k a_k e^{in\Delta t\lambda_k} \right) e^{-in\Delta t\lambda_k} = \sum_l a_l H(\lambda_k - \lambda_l), \tag{6.38}$$

where function H is defined by (6.6).

Taking $l = \nu$ in (6.38), where $|a_\nu|$ is the greatest $|a_n|$, we deduce

$$\left| \Delta t \sum_{n=0}^{M} \sin \frac{n\Delta t\pi}{T} \left(\sum_k a_k e^{in\Delta t\lambda_k} \right) e^{-in\Delta t\lambda_\nu} \right| \geq |a_\nu H(0)| - |a_\nu| \sum_{k,k\neq\nu} |H(\lambda_k - \lambda_\nu)| . \tag{6.39}$$

Function H is exactly the discrete Fourier transform (6.5) used in the proof of discrete Ingham's inequality (2.3). By (6.24) and the estimates for $|H(\lambda_k - \lambda_\nu)|$ used in the proof of Theorem 2.1, i.e.,

$$\sum_{k,k\neq\nu} |H(\lambda_k - \lambda_\nu)| \leq \left[\frac{8\pi}{T\gamma^2} + NC\Delta t^2 + CN\Delta t(\Delta t)^{1-2p} \right],$$

as in (6.28), we deduce that

$$\Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} a_k e^{in\Delta t\lambda_k} \right| \geq C_1(\Delta t, T, \gamma) \max |a_n| . \tag{6.40}$$

The direct inequality in (2.5) may be obtained using the same arguments and estimates and we omit the details. □

REFERENCES

[1] S. A. Avdonin and W. Moran, Ingham-type inequalities and Riesz bases of divided differences, *Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 803–820.*

[2] C. Baiocchi, V. Komornik, and P. Loreti, Ingham type theorems and applications to control theory, *Boll. Unione Mat. Ital., Sez. B Artic. Ric. Mat., B (8) 2 (1999), pp. 33–63.*

[3] C. Baiocchi, V. Komornik, and P. Loreti, Ingham-Beurling type theorems with weakened gap conditions, *Acta Math. Hungar., 97 (2002), pp. 55–95.*

[4] C. Baiocchi, V. Komornik, and P. Loreti, Généralisation d'un théorème de Beurling et application à la théorie du contrôle, *C. R. Acad. Sci. Paris, Sér. I, Math., 330 (2000), pp. 281–286.*

[5] J. M. Ball and M. Slemrod, Nonharmonic Fourier series and the stabilization of distributed semilinear control system, *Comm. Pure Appl. Math., 32 (1979), pp. 555–587.*

[6] L. Carleson, P. Malliavin, J. Neuberger, and J. Werner, eds., The Collected Works of Arne Beurling, *Vol. 2, Birkhäuser, Boston, 1989.*

[7] C. Castro and S. Micu, Boundary controllability of a linear semi-discrete 1-d wave equation derived from a mixed finite elements method, *preprint, 2005.*

[8] C. Castro and E. Zuazua, Une remarque sur les séries de Fourier non-harmoniques et son application à la contrôlabilité des cordes avec densité singulière, *C. R. Acad. Sci. Paris Ser., I Math., 323 (1996), pp. 365–370.*

[9] A. E. Ingham, Some trigonometrical inequalities with applications in the theory of series, *Math. Z., 41 (1936), pp. 367–379.*

[10] E. Isaacson and H. B. Keller, Analysis of Numerical Methods, *John Wiley and Sons, Inc., New York, 1966.*

[11] J. A. Infante and E. Zuazua, Boundary observability for the space semi-discretizations of the 1-D wave equation, M^2AN *Math. Model. Numer. Anal., 33 (1999), pp. 407–438.*

[12] S. Jaffard, M. Tucsnak, and E. Zuazua, On a theorem of Ingham, *J. Fourier Anal. Appl., 3 (1997), pp. 577–582.*

[13] S. Jaffard, M. Tucsnak, and E. Zuazua, Singular internal stabilization of the wave equation, *J. Differential Equations, 145 (1998), pp. 184–215.*

[14] R. Glowinski, C. H. Li, and J.-L. Lions, A numerical approach to the exact boundary controllability of the wave equation. *I. Dirichlet controls: Description of the numerical methods, Japan J. Appl. Math., (1990), pp. 1–76.*

[15] V. Komornik, Exact Controllability and Stabilization: The Multiplier Method, *Masson & John Wiley, Paris & Chichester, UK, 1994.*

[16] L. León and E. Zuazua, Boundary controllability of the finite-difference space semi-discretizations of the beam equation, *in ESAIM: Control Optim. Calc. Var., 8 (2002), pp. 827–862.*

[17] J. L. Lions, Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués. Tome 1 Contrôlabilité Exacte, *Masson, Paris, 1988.*

[18] A. Münch, Famille de schémas implicites uniformément contrôlables pour l'équation des ondes 1-D, *C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 733–738.*

[19] M. Negreanu, Métodos Numéricos para el Análisis de la Propagación, Observación y Control de Ondas, *Ph.D. thesis, Universidad Complutense de Madrid, Departamento de Matemática Aplicada, Madrid, Spain, 2004.*

[20] M. Negreanu and E. Zuazua, Uniform boundary controllability of a discrete 1-D wave equation, *Systems Control Lett., 48 (2003), pp. 261–279.*

[21] M. Negreanu and E. Zuazua, Discrete Ingham inequalities and applications, *C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 281–286.*

[22] R. E. A. C. Paley and N. Wiener, Fourier Transforms in the Complex Domain, *vol. XIX, AMS Coll. Publ., New York, 1934.*

[23] L. N. Trefethen, Finite difference and spectral methods for ordinary and partial differential equations, *unpublished text, 1996; available online at http://web.comlab.ox.ac.uk/oucl/work/nick.trefethen/pdetext.html.*

[24] L. N. Trefethen, Group velocity in the finite difference schemes, *SIAM Rev., 24 (1982), pp. 113–136.*

[25] R. Vichnevetsky and J. B. Bowles, Fourier Analysis of Numerical Approximations of Hyperbolic Equations, *SIAM Studies in Applied Mathematics 5, SIAM, Philadelphia, 1982.*

[26] R. M. Young, An Introduction to Nonharmonic Fourier Series, *Academic Press, San Diego, 1980.*

[27] E. Zuazua, Boundary observability for the finite-difference space semi-discretizations of the 2-d wave equation in the square, *J. Math. Pures Appl., 78 (1999), pp. 523–563.*

[28] E. Zuazua, Propagation, observation, control of waves approximated by finite difference methods, *SIAM Rev., 47 (2005), pp. 197–243.*

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 449–477

CONVERGENCE RATE OF A SCHWARZ MULTILEVEL METHOD FOR THE CONSTRAINED MINIMIZATION OF NONQUADRATIC FUNCTIONALS*

L. BADEA†

Abstract. In [L. Badea, *Convergence Rate of a Multiplicative Schwarz Method for Strongly Nonlinear Inequalities*, V. Barbu, I. Lasiecka, D. Tiba, and C. Varsan, eds., Kluwer Academic Publishers, Boston, 2003], the convergence of a subspace correction method applied to the constrained minimization of a functional in a general reflexive Banach space has been proved, provided that the convex set verifies a certain assumption. This assumption is weaker than that in which the convex set is decomposed according to the space decomposition as a sum of subsets. In the Sobolev spaces, the proposed method becomes a multiplicative Schwarz method for the solution of the variational inequalities coming from the minimization of nonquadratic functionals. We prove in this paper that this assumption holds for the one-, two- and multilevel multiplicative Schwarz methods in the finite element space, and we explicitly write the constants in the error estimations depending on the overlapping and mesh parameters. Our error estimates are similar with those obtained for the minimization of quadratic functionals in [L. Badea, X.-C. Tai, and J. Wang, *SIAM J. Numer. Anal.*, (2003), pp. 1052–1073], or with those obtained for the one-obstacle problem in [X.-C. Tai, *Numer. Math.*, 93 (2003), pp. 755–786].

Key words. domain decomposition methods, variational inequalities, nonquadratic minimization, multigrid and multilevel methods, finite element methods, nonlinear obstacle problems

AMS subject classifications. 65N55, 65N30, 65J15

DOI. 10.1137/S003614290342995X

1. Introduction.

Domain decomposition methods provide efficient numerical algorithms to solve very large-scale problems. The great interest in these methods comes from the fact that they are parallelizable on multiprocessor machines. Schwarz overlapping methods represent a typical example of such parallelizable methods, and they are traditionally being classified as multiplicative and additive. The main focus of this paper is the convergence of the multiplicative Schwarz method applied to the constrained minimization of nonquadratic convex functionals.

Naturally, most papers dealing with these methods are dedicated to linear problems. The multiplicative and additive Schwarz methods for elliptic linear problems have been studied by Lions [26, 27, 28]; Chan, Hou, and Lions [10]; P. Le Tallec [25]; A. Quarteroni and A. Valli [33]; Bramble, Pasciak, Wang, and Xu [8]; and Badea [1] for the multiplicative methods, and Dryja [12]; Dryja and Widlund [13], [14]; and Nepomnyaschikh [32] for the additive version.

For the application of the Schwarz method to the solution of the variational inequalities, we can cite the papers written by Hoffman and Zou [19]; Kuznetsov and Neittaanmäki [22]; Kuznetsov, Neittaanmäki, and Tarvainen [23], [24]; Lü, Liem, and Shih [29]; Zeng and Zhou [42]; Tai [35, 36, 37]; Tai and Tseng [39]; Badea and Wang [3]; Badea, Tai, and Wang [4]; and Badea [2], [6], [7].

*Received by the editors June 17, 2003; accepted for publication (in revised form) August 2, 2005; published electronically March 15, 2006. This work was supported by IMAR under contract ICA1-CT-2000-70022 with the European Commission.
http://www.siam.org/journals/sinum/44-2/42995.html
†Institute of Mathematics, Romanian Academy of Sciences, P. O. Box 1-764, RO-70700 Bucharest, Romania (lori.badea@imar.ro).

Also, the multilevel and multigrid methods can be viewed as domain decomposition methods and we can cite the results obtained by Kornhuber [21]; Mandel [31]; and Smith, Bjørstad, and Gropp [34].

However, very few papers deal with the application of these methods to nonlinear problems. We can cite in this direction the papers written by Tai and Espedal [38], Tai and Xu [40] for nonlinear equations, Hoffmann and Zhou [20], Lui [30], Zeng and Zhou in [43] for inequalities having nonlinear source terms, and Badea [5] for a general result concerning the convergence of the method for the constrained minimization of nonquadratic functionals. Evidently, the above lists of citations are not exhaustive and can be completed by many other papers.

Almost exclusively, the convergence of the domain decomposition methods for variational inequalities coming from the minimization of a functional is studied in the case when this functional is quadratic. Also, most papers consider the convex set decomposed according to the space decomposition as a sum of convex subsets. The main goal of this paper is to give error estimates for the one-, two- and multilevel Schwarz domain decomposition methods applied to the constrained minimization of the nonquadratic convex functionals over enough general convex sets.

The convergence of a domain decomposition algorithm solving variational inequalities coming from the minimization of quadratic functionals over convex sets is proved in [2]. In that paper, the convex set, defined by constraints on the function values at the points of the domain, is not supposed to be decomposed as a sum of convex subsets. In [40], a subspace correction method applied to the minimization without constraints of a differentiable and convex functional defined in a reflexive Banach space is introduced. Also, in [5], the convergence of an algorithm in a reflexive Banach space for the constrained minimization of convex functionals is proved. There, in order to prove the convergence, a weaker property than that given in [2] is imposed on the convex set. To the author's knowledge, there are no other papers dealing with the Schwarz method applied to the constrained minimization of nonquadratic functionals. Even if sometimes the conditions on the convex functional are general enough, the authors always consider the space H^1 and implicitly quadratic functionals. For instance, in [4], using the subspace correction techniques in [8] and [41], and more general conditions in [38] on the convex functional, the convergence rate for the one- and two-level algorithms of the method in [2] is given only for the minimization of quadratic functionals. Starting from the general convergence result given in cite [5], we generalize in this paper the results in [4] and [40] to the constrained minimization of nonquadratic functionals. Our error estimates are similar with those obtained for the minimization of quadratic functionals in [4] or [37].

The paper is organized as follows. In section 2, we state the multiplicative Schwarz method as a subspace correction method in a general reflexive Banach space for the constrained minimization of convex functionals. We also give the convergence theorem of this algorithm which has been proven in [5] provided that a certain assumption holds. In sections 3, 4, and 5 we prove that the introduced assumption holds and we estimate the error for the one-, two- and multilevel Schwarz methods, respectively, in the finite element spaces. In these cases, we are able to explicitly write the convergence rate depending on the mesh and domain decomposition parameters. The proof for the two- and multilevel methods is based on a lemma which can be viewed as a Friedrichs–Poincaré inequality for the finite element spaces. In subsection 5.1, we find the convergence rate of the multigrid method from the results obtained for the multilevel method.

Finally, for writing simplicity, we have considered the next sections problems in $W^{1,s}$, but all the obtained results hold reading $[W^{1,s}]^d$ in the place of $W^{1,s}$.

2. General convergence result. We enunciate in this section a general algorithm and give an error estimate theorem for it. This general theory, the proof of the theorem included, are given in detail in [5]. We consider that V is a reflexive Banach space and $V_1, \ldots, V_m$, are some closed subspaces of V. Also, let $K \subset V$ be a nonempty closed convex set, and we make the following

ASSUMPTION 2.1. *There exists a constant C_0 such that for any $w, v \in K$ and $w_i \in V_i$ with $w + \sum_{j=1}^{i} w_j \in K$, $i = 1, \ldots, m$, there exist $v_i \in V_i$, $i = 1, \ldots, m$, satisfying*

$$w + \sum_{j=1}^{i-1} w_j + v_i \in K \text{ for } i = 1, \ldots, m, \tag{2.1}$$

$$v - w = \sum_{i=1}^{m} v_i, \tag{2.2}$$

and

$$\sum_{i=1}^{m} ||v_i||^p \leq C_0^p \left(||v - w||^p + \sum_{i=1}^{m} ||w_i||^p \right). \tag{2.3}$$

This assumption looks complicated enough, but as we shall see in what follows, it is satisfied for a large kind of convex sets in Sobolev spaces. In our proofs, v is the exact solution, w is the solution of the iterative algorithm at a certain iteration, and w_i are its corrections on the subspaces $V_i, i = 1, \ldots, m$. In the case of the convex sets written as a sum of convex subsets, (2.1) and (2.2) are always satisfied. We point out that in the case of the problems without constraints or that of the one-obstacle problems, the above assumption can be taken with $w_i = 0$ (see [37], for instance), and for this reason (2.3) usually is known without the extra terms given by w_i.

We consider a Gâteaux differentiable functional $F : K \to R$, which is supposed to be coercive if K is not bounded, and we assume that for any real number $M > 0$ there exist two functions,

$$\alpha_M(\tau) = A_M \tau^p, \quad \beta_M(\tau) = B_M \tau^{q-1}, \tag{2.4}$$

such that

$$< F'(v) - F'(u), v - u > \geq \alpha_M(||v - u||), \text{ for any } u, v \in K, \ ||u||, ||v|| \leq M, \tag{2.5}$$

and

$$\beta_M(||v - u||) \geq ||F'(v) - F'(u)||_{V'}, \text{ for any } u, v \in K, \ ||u||, ||v|| \leq M, \tag{2.6}$$

where F' is the Gâteaux derivative of F, and $A_M > 0$, $B_M > 0$, $p > 1$, and $q > 1$ are some real constants. We have marked here that the constants A_M and B_M depend on M. It is evident that if (2.5) and (2.6) hold, then

$$\begin{aligned} &\alpha_M(||v - u||) \leq < F'(v) - F'(u), v - u > \leq \beta_M(||v - u||)||v - u||, \\ &\text{for any } u, v \in K, \ ||u||, ||v|| \leq M. \end{aligned} \tag{2.7}$$

It follows from (2.7) that we must take $p \geq q$. Following the way in [17, Lemmas 1.1 and 1.2] we can prove that

$$\begin{aligned} &< F'(u), v-u > +\lambda_M(||v-u||) \leq F(v) - F(u) \\ &\leq < F'(u), v-u > +\mu_M(||v-u||), \text{ for any } u, v \in K,\ ||u||, ||v|| \leq M, \end{aligned} \tag{2.8}$$

where

$$\lambda(\tau) = \tfrac{A_M}{p}\tau^p, \quad \mu(\tau) = \tfrac{B_M}{q}\tau^q. \tag{2.9}$$

It is well known (see [16]) that if V and F satisfy the above assumption, then the minimization problem

$$u \in K : F(u) \leq F(v), \text{ for any } v \in K \tag{2.10}$$

has a unique solution, and it is also the unique solution of the problem

$$u \in K : \ < F'(u), v-u > \geq 0, \text{ for any } v \in K. \tag{2.11}$$

From (2.8), for a given $M > 0$ such that the solution u of (2.11) satisfies $||u|| \leq M$, we have

$$\lambda_M(||v-u||) \leq F(v) - F(u), \text{ for any } v \in K,\ ||v|| \leq M. \tag{2.12}$$

The proposed algorithm corresponding to the subspaces $V_1, \ldots, V_m$ and the convex set K is written as follows.

ALGORITHM 2.1. *We start the algorithm with an arbitrary $u^0 \in K$. At iteration $n+1$, having $u^n \in K$, $n \geq 0$, we compute sequentially for $i = 1, \ldots, m$, $w_i^{n+1} \in V_i$ satisfying*

$$w_i^{n+1} = \arg \min_{\substack{u^{n+\frac{i-1}{m}} + v_i \in K \\ v_i \in V_i}} G(v_i), \ \textit{with } G(v_i) = F(u^{n+\frac{i-1}{m}} + v_i), \tag{2.13}$$

and then we update

$$u^{n+\frac{i}{m}} = u^{n+\frac{i-1}{m}} + w_i^{n+1}.$$

This algorithm does not assume a decomposition of the convex set K depending on the subspaces V_i, and it has been proposed in [2] in an equivalent form. The above form of this algorithm has been proposed in [4] for the constrained minimization of the quadratic functions. As for problem (2.10), since the subspaces V_i are reflexive Banach spaces, problem (2.13) has a unique solution and also satisfies the variational inequality

$$\begin{aligned} &w_i^{n+1} \in V_i,\ u^{n+\frac{i-1}{m}} + w_i^{n+1} \in K : \\ &< F'(u^{n+\frac{i-1}{m}} + w_i^{n+1}), v_i - w_i^{n+1} > \geq 0, \\ &\text{for any } v_i \in V_i,\ u^{n+\frac{i-1}{m}} + v_i \in K. \end{aligned} \tag{2.14}$$

The introduction of some parameters $\varepsilon_{ij} \geq 0$, $i, j = 1, \ldots, m$, is useful to obtain some sharper error estimations, especially in the case of minimization of the quadratic

functionals. Following this way, we assume that for a given $M > 0$, if $v \in K$, $||v|| \leq M$, and $v_i \in V_i$, satisfying $v + v_i \in K$, $||v + v_i|| \leq M$, $i = 1, \ldots, m$, then we have

$$< F'(v + v_i) - F'(v), w_j > \leq \varepsilon_{ij} B_M ||v_i||^{q-1} ||w_j|| \tag{2.15}$$

for any $w_i \in V_i$, $i = 1, \ldots, m$. Evidently, using (2.6), we may always take $\varepsilon_{ij} = 1$, $i, j = 1, \ldots, m$, in (2.15).

In [40], it is proved the convergence of the method for nonlinear equations. The following theorem extends this result to inequalities.

THEOREM 2.1. *We consider that V is a reflexive Banach, $V_1, \ldots, V_m$ are some closed subspaces of V, K is a nonempty closed convex subset of V, and F is a Gâteaux differentiable functional on K which is supposed to be coercive if K is not bounded. We assume that the functional F satisfies* (2.5) *and* (2.6)*, and we make Assumption* 2.1*. On these conditions, if u is the solution of problem* (2.10) *and u^n, $n \geq 0$ are its approximations obtained from Algorithm* 2.1*, then we have the following error estimations:*

(i) *if $p = q$ we have*

$$\begin{aligned} &F(u^n) - F(u) \leq \left(\frac{\hat{C}}{\hat{C}+1}\right)^n \left[F(u^0) - F(u)\right], \\ &||u^n - u||^p \leq \frac{\hat{C}+1}{\bar{C}} \left(\frac{\hat{C}}{\hat{C}+1}\right)^n \left[F(u^0) - F(u)\right]. \end{aligned} \tag{2.16}$$

(ii) *if $p > q$ we have*

$$\begin{aligned} &F(u^n) - F(u) \leq \frac{F(u^0) - F(u)}{\left[1 + n\,\tilde{C}(F(u^0) - F(u))^{\frac{p-q}{q-1}}\right]^{\frac{q-1}{p-q}}}, \\ &||u - u^n||^p \leq \frac{\hat{C}}{\bar{C}} \frac{\left(F(u^0) - F(u)\right)^{\frac{q-1}{p-1}}}{\left[1 + (n-1)\tilde{C}(F(u^0) - F(u))^{\frac{p-q}{q-1}}\right]^{\frac{(q-1)^2}{(p-1)(p-q)}}}. \end{aligned} \tag{2.17}$$

The constants $\hat{C}$, $\bar{C}$, and $\tilde{C}$ are written as

$$\begin{aligned} \hat{C} = \hat{C}(m, C_0, u^0) = B_M \left(\tfrac{p}{A_M}\right)^{\frac{q}{p}} |\varepsilon_{ij}| \Big[&(1 + 2C_0)\left(F(u^0) - F(u)\right)^{\frac{p-q}{p(p-1)}} \\ &+ \left(B_M \left(\tfrac{p}{A_M}\right)^{\frac{q}{p}} |\varepsilon_{ij}|\right)^{\frac{1}{p-1}} C_0^{\frac{p}{p-1}} / \eta^{\frac{1}{p-1}} \Big] / (1 - \eta), \end{aligned} \tag{2.18}$$

$$\bar{C} = \frac{(2 - \eta) A_M}{(1 - \eta) p}, \tag{2.19}$$

$$\tilde{C} = \frac{p - q}{(p - 1)\left(F(u^0) - F(u)\right)^{\frac{p-q}{q-1}} + (q - 1)\hat{C}^{\frac{p-1}{q-1}}}. \tag{2.20}$$

The value of η in the expressions of $\hat{C}$ and $\bar{C}$ can be arbitrary in $(0, 1)$. On the other hand, we see that the constants in the error estimations of $F(u^n) - -F(u)$ in (2.16) and (2.17) are some increasing functions of $\hat{C}$, and there is an $\eta_0 \in (0, 1)$ such that $\hat{C}(\eta_0) \leq \hat{C}(\eta)$ for any $\eta \in (0, 1)$. However, this value η_0 can be found by solving a nonlinear algebraic equation.

We point out that a convergence result can be found (see [5]) under weaker conditions on the functions α_M and β_M than those given in (2.4), and a weaker assumption than Assumption 2.1.

The above algorithm can be viewed as a multiplicative Schwarz method in a subspace correction variant if we use the Sobolev spaces. In this way, we consider for a domain Ω in $\mathbf{R}^d$, $d \geq 1$, with Lipschitz continuous boundary $\partial\Omega$, an overlapping decomposition

$$\Omega = \bigcup_{i=1}^{m} \Omega_i \tag{2.21}$$

in which the subdomains Ω_i have a Lipschitz continuous boundary, too. We associate with the domain Ω the space $V = W_0^{1,s}(\Omega)$, $1 < s < \infty$, and with the subdomains Ω_i the subspaces $V_i = W_0^{1,s}(\Omega_i)$, $i = 1, \ldots, m$. We assume that the convex set $K \subset V$ satisfies

PROPERTY 2.1. *If $v, w \in K$, and if $\theta \in C^1(\Omega)$ with $0 \leq \theta \leq 1$, then $\theta v + (1-\theta)w \in K$.*

For such a convex set, the following proposition has been proved in [5].

PROPOSITION 2.1. *If for the domain decomposition* (2.21) *there exist some continuously differentiable unity partitions $\{\theta_j^i\}_{j=i,\ldots,m}$ associated with $\cup_{j=i}^m \Omega_j$, $i = 1, \ldots, m$, (i.e., for any $i = 1, \ldots, m$,* supp $\theta_j^i \subset \Omega_j$, $\theta_j^i \in C^1(\Omega_j)$, *and $0 \leq \theta_j^i \leq 1$, for $j = i, \ldots, m$, and $\sum_{j=i}^m \theta_j^i = 1$ on $\cup_{j=i}^m \Omega_j$), then Assumption* 2.1 *holds for any convex set K having Property* 2.1.

Consequently, provided that functional F satisfies (2.5) and (2.6), Algorithm 2.1 converges and we can apply Theorem 2.1 to get the convergence rate. The above Sobolev spaces $W_0^{1,s}$ correspond to Dirichlet boundary conditions. Similar results can be obtained if we consider appropriate subspaces of $W^{1,s}$ for the mixed boundary conditions.

The constant C_0 in Assumption 2.1 depends on the domain decomposition parameters. Consequently, since the constants $\hat{C}$ and $\bar{C}$ in the error estimations in Theorem 2.1 depend on C_0, then these estimations will depend on domain decomposition parameters, too. The goal of the next sections is to prove, for the one-, two-level and multilevel multiplicative Schwarz methods, that Assumption 2.1 also holds for any closed convex K satisfying a similar property to that given in 2.1. In these cases we are able to explicitly write the dependence of C_0 on the domain decomposition and mesh parameters.

3. One-level multiplicative Schwarz method. First, let us consider that the domain $\Omega \subset \mathbf{R}^d$ has an overlapping domain decomposition $\{O_i\}_{1 \leq i \leq M}$ and a simplicial mesh partition $\mathcal{T}_h$ of mesh size h. We assume that $\mathcal{T}_h$ is regular (i.e., there exists a constant $C > 0$, independent of h, such that each τ in $\mathcal{T}_h$ contains a ball with the diameter of Ch, and, evidently, it is contained in a ball with the diameter of h; see [11], p. 124, for instance) and it supplies a mesh partition for each subdomain O_i, $i = 1, \ldots, M$, too. In addition, we suppose that there exists a positive constant δ, the overlapping parameter, such that for any $i = 1, \ldots, M$, we have

$$O_i \cap \partial\left(\bigcup_{j \neq i} O_j\right) \neq \emptyset \text{ and } \operatorname{dist}\left(\partial O_i \backslash \partial\Omega, O_i \cap \partial\left(\bigcup_{j \neq i} O_j\right)\right) \geq \delta. \tag{3.1}$$

Now, we assume that there exist m colors such that each subdomain O_i can be marked with one color, and the subdomains with the same color do not intersect with each

other. For suitable overlaps, one can always choose $m = 2$ if $d = 1$, $m \leq 4$ if $d = 2$, and $m \leq 8$ if $d = 3$. Let Ω_i be the union of the subdomains O_j having the color i. In this way, we have obtained an overlapping decomposition (2.21) with overlaps of size δ. Taking into account (3.1), we can assume that the unity partitions $\{\theta_j^i\}_{j=i,\ldots,m}$ associated with $\cup_{j=i}^m \Omega_j$ in Proposition 2.1 satisfy

$$|\partial_{x_k}\theta_j^i| \leq C/\delta, \text{ for any } i = 1,\ldots,m.j = i,\ldots,m, \text{ and } k = 1,\ldots,d, \tag{3.2}$$

As in (3.2), we denote in the following by C a generic constant which does not depend on either the mesh or the domain decomposition parameters.

In this section we prove for the finite element spaces a similar result to that given in Proposition 2.1 for general Sobolev spaces. The proof is also similar to that given in [4] for the minimization of the quadratic functionals. We consider the piecewise linear finite element space

$$V_h = \{v \in C^0(\bar{\Omega}) : v|_\tau \in P_1(\tau),\ \tau \in \mathcal{T}_h,\ v = 0 \text{ on } \partial\Omega\}, \tag{3.3}$$

and also, for $i = 1,\ldots,m$, we take

$$V_h^i = \{v \in V_h : v = 0 \text{ in } \Omega\backslash\Omega_i\} \tag{3.4}$$

as some subspaces of V_h corresponding to the domain decomposition $\Omega_1,\ldots,\Omega_m$. The spaces V_h and V_h^i, $i = 1,\ldots,m$, are considered subspaces of $W^{1,s}$, for some fixed $1 \leq s \leq \infty$. We denote by $||\cdot||_{0,s}$ the norm in L^s, and by $||\cdot||_{1,s}$ and $|\cdot|_{1,s}$ the norm and seminorm in $W^{1,s}$, respectively.

In the following, L_h will be the P_1-Lagrangian interpolation operator which uses the function values at the nodes of the mesh $\mathcal{T}_h$. The convex set K_h is defined as a subset of V_h satisfying the following property.

PROPERTY 3.1. *If $v, w \in K_h$, and if $\theta \in C^1(\Omega)$ with $0 \leq \theta \leq 1$, then $L_h(\theta v + (1-\theta)w) \in K_h$.*

In order to prove that Assumption 2.1 holds, we follow the same way as in [4] or [5]. Taking into account the additivity of the Lagrangian interpolation L_h, (2.1) and (2.2) in Assumption 2.1 can be recurrently proved. Indeed, first we write

$$v_1 = L_h\left(\theta_1^1(v-w) + (1-\theta_1^1)w_1\right), \tag{3.5}$$

and prove that

$$\begin{aligned}
&v_1 \in V_h^1 \text{ and } w + v_1 \in K_h,\\
&v - v_1 + w_1 \in K_h,\\
&v - w - v_1 \in W_0^{1,s}\left(\bigcup_{j=2}^m \Omega_j\right) \text{ and}\\
&v - w - v_1 = 0 \text{ in } \Omega - \overline{\cup_{j=2}^m \Omega_j}.
\end{aligned}$$

Next, for $i = 2,\ldots,m-1$, we write

$$v_i = L_h\left(\theta_i^i(v - w - \sum_{j=1}^{i-1} v_j) + (1-\theta_i^i)w_i\right), \tag{3.6}$$

and prove

$$v_i \in V_h^i \text{ and } w + \sum_{j=1}^{i-1} w_j + v_i \in K_h,$$
$$v - \sum_{j=1}^{i} v_j + \sum_{j=1}^{i} w_j \in K_h,$$
$$v - w - \sum_{j=1}^{i} v_j \in W_0^{1,s}\left(\bigcup_{j=i+1}^{m} \Omega_j\right) \text{ and}$$
$$v - w - \sum_{j=1}^{i} v_j = 0 \text{ in } \Omega - \overline{\cup_{j=i+1}^{m}\Omega_j},$$

assuming that these equations hold for $i-1$. Finally, we take

$$v_m = v - w - \sum_{j=1}^{m-1} v_j. \tag{3.7}$$

To prove inequality (2.3) in Assumption 2.1, we first note that, starting from v_1 given in (3.5) by the recurrent application of (3.6), and then taking v_m given in (3.7), we get that v_i, $i = 1, \ldots, m$, are of the form

$$v_i = L_h\left(\tau_0^i(v-w) + \sum_{j=1}^{i} \tau_j^i w_j\right),\ i = 1, \ldots, m. \tag{3.8}$$

By a simple calculus we get that

$\tau_0^1 = \theta_1^1,\ \tau_1^1 = 1 - \theta_1^1,$
$\tau_0^i = \theta_i^i(1-\theta_{i-1}^{i-1})\cdots(1-\theta_1^1),\ \tau_i^i = 1-\theta_i^i,\ \tau_j^i = -\theta_i^i(1-\theta_{i-1}^{i-1})\cdots(1-\theta_j^j),$
for $i = 2, \ldots, m-1,\ j = 1, \ldots, i-1,$
$\tau_0^m = (1-\theta_{m-1}^{m-1})\cdots(1-\theta_1^1),\ \tau_m^m = 0, \tau_{m-1}^m = -(1-\theta_{m-1}^{m-1}),$
$\tau_j^m = \theta_{m-1}^{m-1}(1-\theta_{m-2}^{m-2})\cdots(1-\theta_j^j),$ for $j = 1, \ldots, m-2.$

Consequently, from (3.2), we have

$$|\tau_j^i| \le 1 \text{ and } |\partial_{x_k}\tau_j^i| \le C(m-1)/\delta,\ i = 1, \ldots, m,\ j = 0, \ldots, i,\ k = 1, \ldots, d. \tag{3.9}$$

For a $v \in V_h$, we get (see, for instance [11, Theorem 3.1.6]) that

$$||\tau_j^i v - L_h(\tau_j^i v)||_{0,s} \le Ch|\tau_j^i v|_{1,s},\ ||\tau_j^i v - L_h(\tau_j^i v)||_{1,s} \le C|\tau_j^i v|_{1,s},$$

and therefore

$$||L_h(\tau_j^i v)||_{1,s} \le C||\tau_j^i v||_{1,s}, \text{ with } v \in V_h, \tag{3.10}$$

for any $i = 1, \ldots, m$, $j = 0, \ldots, i$. On the other hand, from (3.9) we get

$$||\tau_j^i v||_{0,s} \le ||v||_{0,s}, \quad |\tau_j^i v|_{1,s} \le C\left(|v|_{1,s} + \frac{m-1}{\delta}||v||_{0,s}\right), \text{ for any } v \in V_h, \tag{3.11}$$

and therefore, using (3.10), we get

$$||L_h(\tau_j^i v)||_{1,s} \leq C\left(||v||_{1,s} + \frac{m-1}{\delta}||v||_{0,s}\right), \text{ for any } v \in V_h. \tag{3.12}$$

Now, by a application of (3.12) to (3.8) we get

$$||v_i||_{1,s} \leq C\left(1 + \frac{m-1}{\delta}\right)\left(||v-w||_{1,s} + \sum_{j=1}^{i}||w_j||_{1,s}\right), \quad \text{for any } i = 1,\ldots,m. \tag{3.13}$$

Using the above equation we get (2.3) in Assumption 2.1, and have the following proposition.

PROPOSITION 3.1. *Let $\Omega_1,\ldots,\Omega_m$ be the overlapping decomposition of the domain Ω defined in this section. Then, Assumption* 2.1 *holds for the piecewise linear finite element spaces, $V = V_h$ and $V_i = V_h^i$, $i = 1,\ldots,m$, and for any convex set $K = K_h \subset V_h$ having Property* 3.1. *The constant in* (2.3) *of Assumption* 2.1 *can be written as*

$$C_0 = C(m+1)\left(1 + \frac{m-1}{\delta}\right), \tag{3.14}$$

where C is independent of the mesh parameter and the domain decomposition.

Remark 3.1. We notice that the number m of the subdomains Ω_i in the decomposition of Ω is in fact the number of colors of the overlapping domain decomposition $\{O_i\}_{1\leq i\leq M}$, and it depends only on the dimension d of the space $\mathbf{R}^d$. Consequently, error estimations (2.16) and (2.17) in Theorem 2.1 depend only on the size δ of the overlaps through the intermediary of the constant C_0 given in (3.14).

4. Two-level multiplicative Schwarz method. We consider a simplicial mesh partition $\mathcal{T}_h$ of the domain $\Omega \subset \mathbf{R}^d$ of a mesh size h, and a simplicial coarser mesh $\mathcal{T}_H$ with a mesh size H, $\mathcal{T}_h$ being a refinement of $\mathcal{T}_H$. The mesh size h is supposed to approach zero and we shall consider a family of mesh pairs (h, H). We assume that both the families, of fine and coarse meshes, are regular.

As in the previous section, we consider an overlapping decomposition $\Omega = \cup_{i=1}^{M} O_i$, the mesh partition $\mathcal{T}_h$ of Ω supplying a mesh partition for each O_i, $1 \leq i \leq M$. Also, we assume that the overlapping size is δ, i.e., (3.1) is satisfied. In addition, we suppose that there exists a constant C such that

$$\operatorname{diam}(O_i) \leq CH, \ i = 1,\ldots,M. \tag{4.1}$$

Now, we color the subdomains O_i, $i = 1,\ldots,M$, and obtain the subdomains Ω_i, $i = 1,\ldots,m$ as in the previous section. We point out that the domain Ω may be different from

$$\Omega_0 = \bigcup_{\tau\in\mathcal{T}_H} \tau, \tag{4.2}$$

but we assume that if a node of $\mathcal{T}_H$ lies on $\partial\Omega_0$, then it lies on $\partial\Omega$, too, and

$$\Omega\backslash\Omega_0 \subset \bigcup_{x^i \text{ node of } \mathcal{T}_H,\, x^i\in\partial\Omega} S_{x^i}, \tag{4.3}$$

where the sets S_{x^i} are defined as follows. We first denote by ω_i the union of all $\tau \in \mathcal{T}_H$ having x^i as a vertex. Then, S_{x^i} is the union of ω_i with all $\tau \in \mathcal{T}_h$, $\tau \not\subset \Omega_0$, which are contained in the smallest sphere centered at x^i and containing ω_i.

Now, we introduce the continuous, piecewise linear finite element space corresponding to the H-level,

$$V_H^0 = \left\{ v \in C^0(\bar{\Omega}_0) : v|_\tau \in P_1(\tau),\ \tau \in \mathcal{T}_H,\ v = 0 \text{ on } \partial\Omega_0 \right\}, \tag{4.4}$$

and extending the functions of V_H^0 with zero in $\Omega \backslash \Omega_0$, it becomes a subspace of V_h. The convex set $K_h \subset V_h$ is defined as a subset of V_h having Property 3.1.

The two-level Schwarz method is also obtained from Algorithm 2.1 in which we take $V = V_h$, $K = K_h$, and the subspaces $V_0 = V_H^0$, $V_1 = V_h^1$, $V_2 = V_h^2, \ldots, V_m = V_h^m$. As in the previous section, the spaces V_h, V_H^0, V_h^1, $V_h^2, \ldots, V_h^m$, are considered as subspaces of $W^{1,s}$ for $1 \le s \le \infty$. We note that this time the decomposition of the domain Ω contains m overlapping subdomains, but we use $m+1$ subspaces of V, V_0, $V_1, \ldots, V_m$, in Algorithm 2.1. Naturally, this algorithm will converge if Assumption 2.1, written for $m+1$ subspaces, will be satisfied for the previous choice of the convex set K and the subspaces $V_0, V_1, \ldots, V_m$, of V. As in the previous section, we prove that Assumption 2.1 holds and find the constant C_0 depending on the mesh and domain decomposition parameters. First, we have the following lemma in which inequality (4.5) can be viewed as a Friedrichs–Poincaré type for the finite element spaces.

LEMMA 4.1. *Let $\omega \subset \mathbf{R}^d$ be a domain of diameter H, and ω_i, $i = 0, 1, \ldots, N$, be an overlapping decomposition of it, $\omega = \cup_{i=0}^N \omega_i$. We consider a simplicial regular mesh partition $\mathcal{T}_h$ of ω and assume that it supplies a mesh partition for each ω_i, $i = 0, 1, \ldots, N$, too. Let $x^0 \in \bar{\omega}_0$ be a node of $\mathcal{T}_h$. We assume that the overlapping partition of ω satisfies:*

(i) *for any $x \in \bar{\omega}_0$, the line segment $[x^0, x]$ lies in $\bar{\omega}_0$,*

(ii) *for $N > 0$, if $\omega_i \cap \omega_j \neq \emptyset$, $0 \le i \ne j \le N$, then for any $x \in \bar{\omega}_i$, $y \in \bar{\omega}_j$ and $z \in \bar{\omega}_i \cap \bar{\omega}_j$, the line segments $[x, z]$ and $[y, z]$ lie in $\bar{\omega}_i$ and $\bar{\omega}_j$, respectively.*

On these conditions, if v is a continuous function which is linear on each $\tau \in \mathcal{T}_h$, and $v(x^0) = 0$, then

$$||v||_{0,s,\omega} \le C(N,s)C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega}, \tag{4.5}$$

where

$$C_{d,s}(H,h) = \begin{cases} 1 & \text{if } d = s = 1 \text{ or } 1 \le d < s \le \infty \\ \left(\ln \frac{H}{h} + 1\right)^{\frac{d-1}{d}} & \text{if } 1 < d = s < \infty \\ \left(\frac{H}{h}\right)^{\frac{d-s}{s}} & \text{if } 1 \le s < d < \infty, \end{cases} \tag{4.6}$$

$$C(d,s) = \begin{cases} C & \text{if } d = s = 1 \text{ or } 1 = s < d < \infty \\ C\left(d\frac{s-1}{s-d}\right)^{\frac{s-1}{s}} & \text{if } 1 \le d < s \le \infty \\ Cd^{\frac{d-1}{d}} & \text{if } 1 < d = s < \infty \\ C(d\frac{s-1}{d-s})^{\frac{s-1}{s}} & \text{if } 1 < s < d < \infty. \end{cases} \tag{4.7}$$

and

$$C(N,s) = \begin{cases} 1 & \text{if } N = 0 \\ \text{if } (N+1)\frac{C_\omega^{(N+1)/s}-1}{C_\omega^{1/s}-1} & \text{if } N \neq 0 \end{cases} \tag{4.8}$$

with

$$C_\omega = \max_{\omega_i \cap \omega_j \neq \emptyset} \frac{|\omega_i|}{|\omega_i \cap \omega_j|}. \tag{4.9}$$

In (4.9) *we have denoted by* $|\ \ |$ *the measure of a set, and have marked in* (4.5) *that the norm in* L^s *and the seminorm in* $W^{1,s}$, $1 \leq s \leq \infty$, *refer to the domain* ω. *The constant* C *in* (4.7) *is independent of* H, h, d, s, *and the decomposition of* ω.

Proof. Here we use the polar coordinates. The Jacobian determinant of the transformation from the rectangular coordinates to the polar coordinates can be written as

$$J(r, \varphi) = r^{d-1} E(\varphi),$$

where $E(\varphi)$ is an algebraic expression of cosines and sines of the component angles of φ.

We first consider that $N = 0$, i.e., the decomposition of ω in the statement of the lemma has only one element, $\omega_0 = \omega$. Consequently, for any $x \in \bar{\omega}$, the line segment $[x^0, x]$ lies in $\bar{\omega}$. We take the origin of the system of coordinates at the point x^0, and, using the polar coordinates, a point $x = (x_1, \ldots, x_d)$, will be written as $x = (r, \varphi)$, φ being the system of $d-1$ angles giving the direction of the vector x. We denote by r_φ the maximum size of the radius in the direction φ of the points in $\bar{\omega}$, and consequently, the points on $\partial\omega$ will be written as (r_φ, φ). We denote by o the union of the $\tau \in \mathcal{T}_h$ having a vertex at x^0; let r_0 be the distance from x^0 to $\partial o \backslash \partial\omega$. We consider the open ball with the center at x^0 of radius r_0, $B_{r_0}(x^0)$. For two points $x' = (r', \varphi) \in \omega \cap B_{r_0}(x^0)$ and $x = (r, \varphi) \in \omega \backslash \bar{B}_{r_0}(x^0)$, we have

$$\begin{aligned}
&|v(x)| = |v(r,\varphi)| \leq |v(r',\varphi)| + \left|\textstyle\int_{r'}^{r} \frac{\partial v}{\partial r}(\rho,\varphi) d\rho\right| \\
&= \left|\tfrac{\partial v}{\partial r}(r',\varphi)\right| r' + \left|\textstyle\int_{r'}^{r} \frac{\partial v}{\partial r}(\rho,\varphi) d\rho\right| \leq \left|\nu_1 \tfrac{\partial v}{\partial x_1}(r',\varphi) + \cdots + \nu_d \tfrac{\partial v}{\partial x_d}(r',\varphi)\right| r' \\
&+ \left|\textstyle\int_{r'}^{r_\varphi} \left(\nu_1 \frac{\partial v}{\partial x_1}(\rho,\varphi) + \cdots + \nu_d \frac{\partial v}{\partial x_d}(\rho,\varphi)\right) d\rho\right| \\
&\leq \left(\left|\tfrac{\partial v}{\partial x_1}(r',\varphi)\right| + \cdots + \left|\tfrac{\partial v}{\partial x_d}(r',\varphi)\right|\right) r' \\
&+ \textstyle\int_{r'}^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|\right) d\rho,
\end{aligned} \tag{4.10}$$

where $(\nu_1, \ldots, \nu_d)$ is the unity vector giving the direction of $x = (r, \varphi)$ in the rectangular system of coordinates $(x_1, \ldots, x_d)$. In the following, we find (4.5) for the various values of d and s starting from (4.10).

For $d = s = 1$ or $1 \leq d < s \leq \infty$, we take $r' = 0$ in (4.10). If $d = s = 1$ we get

$$|v(x)| = |v(r,\varphi)| \leq \int_0^{r_\varphi} \left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| d\rho.$$

Here, we may have $\varphi = 0$ and $\varphi = \pi$ if x^0 is a inner point in ω, and only $\varphi = 0$ or only $\varphi = \pi$ if $x^0 \in \partial\omega$. Integrating again from 0 to $r_\varphi \leq H$, we get (4.5) for $N = 0$ and $d = s = 1$. If $1 \leq d < s = \infty$, we have

$$|v(x)| \leq r_\varphi d \max_{1 \leq j \leq d} \sup_{0 \leq \rho \leq r_\varphi} \left|\frac{\partial v}{\partial x_j}(\rho,\varphi)\right| \leq CdH|v|_{1,\infty,\omega}.$$

If $1 \leq d < s < \infty$ we have

$$|v(x)|^s \leq d^{s-1} \left[\int_0^{r_\varphi} \rho^{\frac{1-d}{s-1}} d\rho\right]^{s-1} \int_0^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho.$$

Multiplying the above inequality by r^{d-1} and integrating from 0 to $r_\varphi \le H$ we get

$$\begin{aligned}&\textstyle\int_0^{r_\varphi} |v(r,\varphi)|^s r^{d-1} dr\\ &\le \left(d\frac{s-1}{s-d}\right)^{s-1} (CH)^s \int_0^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho.\end{aligned}$$

By multiplication of this equation with the Jacobian part depending on φ, $E(\varphi)$, and integrating over the $d-1$ dimensional domain of the angles φ, we get (4.5) for $N=0$ and $1 \le d < s < \infty$.

Now, from (4.10) for an arbitrary $0 < r' < r_0$, we get

$$\begin{aligned}|v(x)| = |v(r,\varphi)| &\le \left(\left|\frac{\partial v}{\partial x_1}(r',\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(r',\varphi)\right|\right) r_0\\ &+ \int_{r_0}^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|\right) d\rho.\end{aligned} \tag{4.11}$$

Also, since for a fixed φ, $\frac{\partial v}{\partial r}(r',\varphi)$ is constant for $r' \in (0, r_0)$, we have

$$\begin{aligned}&|v(x')|^s = |v(r',\varphi)|^s \le \tfrac{(r')^{s-d}}{d} \textstyle\int_0^{r_0} \left|\frac{\partial v}{\partial \rho}(\rho,\varphi)\right|^s \rho^{d-1} d\rho\\ &\le d^{s-2}(r')^{s-d} \textstyle\int_0^{r_0} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho.\end{aligned}$$

Multiplying the above inequality by $(r')^{d-1}$, and integrating from 0 to r_0, we get

$$\begin{aligned}&\textstyle\int_0^{r_0} |v(\rho,\varphi)|^s \rho^{d-1} d\rho\\ &\le \frac{d^{s-2}}{s} r_0^s \textstyle\int_0^{r_0} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho.\end{aligned} \tag{4.12}$$

Now, if $1 = s < d < \infty$ we get from (4.11),

$$\begin{aligned}&|v(x)| \le \tfrac{1}{d} r_0^{1-d} \textstyle\int_0^{r_0} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|\right) \rho^{d-1} d\rho\\ &+ r_0^{1-d} \textstyle\int_{r_0}^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|\right) \rho^{d-1} d\rho\\ &\le r_0^{1-d} \textstyle\int_0^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|\right) \rho^{d-1} d\rho.\end{aligned}$$

Using the regularity of the mesh $\mathcal{T}_h$, we have $\frac{r_\varphi}{r} \le C\frac{H}{h}$, and therefore,

$$\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)| \rho^{d-1} d\rho \le CH \left(\frac{H}{h}\right)^{d-1} \int_0^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right| + \cdots + \left|\left|\frac{\partial v}{\partial x_d}(\rho,\varphi\right|\right) \rho^{d-1} d\rho.$$

From this last inequality and (4.12) we get (4.5) for $N=0$ and $1 = s < d < \infty$ by a multiplication with $E(\varphi)$ and integrating over the domain of the angles φ.

Starting again from (4.11), for $1 < d = s < \infty$ or $1 < s < d < \infty$, we get

$$\begin{aligned}&|v(x)|^s \le (2d)^{s-1} \left(\left|\frac{\partial v}{\partial x_1}(r',\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(r',\varphi)\right|^s\right) r_0^s\\ &+ (2d)^{s-1} \left[\textstyle\int_{r_0}^{r_\varphi} \rho^{\frac{1-d}{s-1}} d\rho\right]^{s-1} \int_{r_0}^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho\\ &= 2^{s-1} d^s r_0^{s-d} \textstyle\int_0^{r_0} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho\\ &+ (2d)^{s-1} \left[\textstyle\int_{r_0}^{r_\varphi} \rho^{\frac{1-d}{s-1}} d\rho\right]^{s-1} \int_{r_0}^{r_\varphi} \left(\left|\frac{\partial v}{\partial x_1}(\rho,\varphi)\right|^s + \cdots + \left|\frac{\partial v}{\partial x_d}(\rho,\varphi)\right|^s\right) \rho^{d-1} d\rho.\end{aligned}$$

Consequently,

$$
\begin{aligned}
&\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)|^s \rho^{d-1} d\rho \\
&\leq (2d)^{s-1} r_\varphi^d r_0^{s-d} \textstyle\int_0^{r_0} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi) \right|^s + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^s \right) \rho^{d-1} d\rho \\
&+ 2^{s-1} d^{s-2} r_\varphi^d \left[\textstyle\int_{r_0}^{r_\varphi} \rho^{\frac{1-d}{s-1}} d\rho \right]^{s-1} \\
&\cdot \textstyle\int_{r_0}^{r_\varphi} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi) \right|^s + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^s \right) \rho^{d-1} d\rho.
\end{aligned} \tag{4.13}
$$

Now, from (4.13), if $1 < d = s < \infty$ we get

$$
\begin{aligned}
&\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)|^d \rho^{d-1} d\rho \\
&\leq 2^{d-1} r_\varphi^d \max \left\{ d^{d-1}, d^{d-2} \left(\ln \tfrac{r_\varphi}{r_0} \right)^{d-1} \right\} \\
&\textstyle\int_0^{r_\varphi} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi \right|^d + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^d \right) \rho^{d-1} d\rho.
\end{aligned}
$$

Using regularity of the mesh $\mathcal{T}_h$, we get

$$
\begin{aligned}
&\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)|^d \rho^{d-1} d\rho \\
&\leq d^{d-1} (CH)^d \left(\ln \tfrac{H}{h} + 1 \right)^{d-1} \textstyle\int_0^{r_\varphi} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi) \right|^d + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^s \right) \rho^{d-1} d\rho.
\end{aligned}
$$

This inequality together with (4.12) prove (4.5) for $N = 0$ and $1 < d = s < \infty$.

Finally, if $1 < s < d < \infty$, we get from (4.13),

$$
\begin{aligned}
&\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)|^s \rho^{d-1} d\rho \leq 2^{s-1} \max \left\{ d^{s-1} r_\varphi^d r_0^{s-d}, d^{s-2} r_\varphi^s (\frac{s-1}{d-s})^{s-1} \left[\left(\tfrac{r_\varphi}{r_0} \right)^{\frac{d-s}{s-1}} - 1 \right]^{s-1} \right\} \\
&\textstyle\int_0^{r_\varphi} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi) \right|^s + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^s \right) \rho^{d-1} d\rho,
\end{aligned}
$$

and, consequently,

$$
\begin{aligned}
&\textstyle\int_{r_0}^{r_\varphi} |v(\rho,\varphi)|^s \rho^{d-1} d\rho \\
&\leq (d\tfrac{s-1}{d-s})^{s-1} (CH)^s \left(\tfrac{H}{h} \right)^{d-s} \textstyle\int_0^{r_\varphi} \left(\left| \frac{\partial v}{\partial x_1}(\rho,\varphi) \right|^s + \cdots + \left| \frac{\partial v}{\partial x_d}(\rho,\varphi) \right|^s \right) \rho^{d-1} d\rho.
\end{aligned}
$$

Using again (4.12) and the last inequality, we get (4.5) for $N = 0$ and $1 < s < d < \infty$.

Assume now that $N > 0$, i.e., we have more than one subdomain ω_i, $i = 0, 1, \ldots, N$ in the overlapping decomposition of ω. Such a decomposition is considered when there exist points $x \in \bar{\omega}$ for which the line segment $[x^0, x]$ does not wholly lie in $\bar{\omega}$. Let ω_i and ω_j, $i \neq j$ be two fixed subdomains such that $\omega_i \cap \omega_j \neq \emptyset$. We consider a fixed point $z \in \omega_i \cap \omega_j$, and denoting by z^k and ϕ_k the nodes of $\mathcal{T}_h$ in $\bar{\omega}_i \cap \bar{\omega}_j$ and the corresponding functions in the nodal basis, respectively, for a given $1 \leq s < \infty$, we have

$$
\begin{aligned}
&||v||_{0,s,\omega_j} - \left| \sum_k v(z^k) \phi_k(z) \right| |\omega_j|^{1/s} \leq ||v - \sum_k v(z^k) \phi_k(z)||_{0,s,\omega_j} \\
&= \left\| \sum_k \left(v - v(z^k) \right) \phi_k(z) \right\|_{0,s,\omega_j} \leq \sum_k ||v - v(z^k)||_{0,s,\omega_j} \phi_k(z).
\end{aligned}
$$

Since $v - v(z^k)$ vanishes at z^k, we get from the first part of the proof and the last equation that

$$||v||_{0,s,\omega_j} - \left|\sum_k v(z^k)\phi_k(z)\right| |\omega_j|^{1/s} \le \sum_k C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_j}\phi_k(z)$$
$$= C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_j},$$

and integrating over $\omega_i \cap \omega_j$, we get

$$|\omega_i \cap \omega_j|||v||_{0,s,\omega_j} \le |\omega_j|^{1/s} \int_{\omega_i\cap\omega_j} |v| + |\omega_i \cap \omega_j|C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_j}$$
$$\le |\omega_j|^{1/s}|\omega_i \cap \omega_j|^{(s-1)/s}||v||_{0,s,\omega_i\cap\omega_j} + |\omega_i \cap \omega_j|C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_j}.$$

Consequently, we have

$$||v||_{0,s,\omega_j} \le \left(\frac{|\omega_j|}{|\omega_i \cap \omega_j|}\right)^{1/s} ||v||_{0,s,\omega_i} + C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_j}. \tag{4.14}$$

It is easy to see that (4.14) holds for $s = \infty$, too. Taking into account that

$$||v||_{0,s,\omega_0} \le C(d,s)HC_{d,s}(H,h)|v|_{1,s,\omega_0}, \tag{4.15}$$

from (4.14) and (4.15), we get (4.5) for $N > 0$. □

Remark 4.1. As stated at the beginning of this section, we are interested in the error estimation for a family of pairs (H,h). In general, since the mesh $\mathcal{T}_h$ is regular, the overlapping decomposition of ω in Lemma 4.1 can be taken such that the number N and the constant C_ω in (4.9) are bounded and independent of (H,h). In this point of view, the constants $C(d,s)$, $C(N,s)$ and C_ω, written in (4.7)–(4.9), can be considered as independent of H and h, and assimilated to the generic constant C. In the following we write (4.5) as

$$||v||_{0,s,\omega} \le CHC_{d,s}(H,h)|v|_{1,s,\omega}, \tag{4.16}$$

where $C = C(N,s)C(d,s)$ and $C_{d,s}(H,h)$ is given in (4.6).

The above lemma can be very useful in various error estimations. The following result, for instance, extends to $W^{1,s}$ that in [9, Lemma 2.3].

COROLLARY 4.1. *Let ω be a domain of diameter H and have a simplicial regular mesh partition $\mathcal{T}_h$. If v is a continuous function which is linear on each $\tau \in \mathcal{T}_h$, and $v = 0$ on $\partial\omega$, then for any $1 \le s \le \infty$ we have*

$$||v||_{0,\infty,\omega} \le CH^{\frac{s-d}{s}}C_{d,s}(H,h)|v|_{1,s,\omega}, \tag{4.17}$$

where $C_{d,s}(H,h)$ is given in (4.6)*, and C is independent of H and h.*

Proof. Let $x^0 \in \bar{\omega}$ be the point where $|v(x^0)| = ||v||_{0,\infty,\omega}$, and $x \in \omega$ a current point. We note that x^0 is a node of $\mathcal{T}_h$. For $1 \le s < \infty$, we have

$$|v(x^0)|^s \le 2^{s-1}|v(x^0) - v(x)|^s + 2^{s-1}|v(x)|^s,$$

and integrating it over ω, using (4.16), we get

$$|\omega|||v||^s_{0,\infty,\omega} \le 2^{s-1}||v(x^0) - v(x)||^s_{0,s,\omega} + 2^{s-1}||v(x)||^s_{0,s,\omega}$$
$$\le 2^{s-1}\left(CHC_{d,s}(H,h)\right)^s |v(x)|^s_{1,s,\omega} + 2^{s-1}||v(x)||^s_{0,s,\omega}.$$

Now, since $v = 0$ on $\partial\omega$, we can apply the classical Friedrichs–Poincaré inequality and obtain (4.17). If $s = \infty$, the proof is similar. ∎

Coming back to the two-level method, let us denote by x^i a node of $\mathcal{T}_H$, by ϕ_i the linear nodal basis function associated with x^i and $\mathcal{T}_H$, and by ω_i the support of ϕ_i. We point out that we consider all the nodal basis functions, including those corresponding to the nodes on $\partial\Omega_0$. Given a $v \in V_h$, let us write

$$I_i^- v = \min_{x\in\omega_i} v(x)^- \quad \text{and} \quad I_i^+ v = \min_{x\in\omega_i} v(x)^+, \tag{4.18}$$

where $v(x)^- = \max(0, -v(x))$ and $v(x)^+ = \max(0, v(x))$. Since v is piecewise linear, $I_i^- v$ or $I_i^+ v$ are attained at a node of $\mathcal{T}_h$ if they are not zero. For a $v \in V_h$, we define

$$I_H^- v := \sum_{x^i \text{ node of } \mathcal{T}_H} (I_i^- v)\phi_i(x) \quad \text{and} \quad I_H^+ v := \sum_{x^i \text{ node of } \mathcal{T}_H} (I_i^+ v)\phi_i(x), \tag{4.19}$$

and we write

$$I_H v = I_H^+ v - I_H^- v. \tag{4.20}$$

The following result extends to that given in [37], where similar operators to I_i^+ have been introduced.

LEMMA 4.2. *For any $v \in V_h$ we have*

$$||I_H v - v||_{0,s,\Omega_0} \le CHC_{d,s}(H,h)|v|_{1,s,\Omega_0} \tag{4.21}$$

and

$$||I_H v||_{0,s,\Omega_0} \le C||v||_{0,s,\Omega_0} \text{ and } |I_H v|_{1,s,\Omega_0} \le CC_{d,s}(H,h)|v|_{1,s,\Omega_0}, \tag{4.22}$$

where Ω_0 is the union of the simplexes in $\mathcal{T}_H$ written in (4.2)*, $C_{d,s}(H,h)$ is defined in* (4.6)*, and C is independent of H, h, and δ. Equations* (4.21) *and* (4.22) *also hold if Ω_0 is replaced by Ω. Moreover, if $\mathcal{K}$ is a convex and closed set in V_h having Property* 3.1*, with $0 \in \mathcal{K}$, then for any $v \in \mathcal{K}$ we have $I_H v \in \mathcal{K} \cap V_H^0$.*

Proof. Let us take an ω_i, the support of the linear basis function ϕ_i corresponding to the node x^i of $\mathcal{T}_H$, and a $v \in V_h$. If v vanishes at a point in ω_i, then $I_i^+ v = I_i^- v = 0$ and v^+ and v^- vanish at some nodes of $\mathcal{T}_h$ in ω_i. Applying Lemma 4.1, we get

$$\begin{aligned} &||v||_{0,s,\omega_i}^s = ||v^+ - v^-||_{0,s,\omega_i}^s = ||v^+||_{0,s,\omega_i}^s + ||v^-||_{0,s,\omega_i}^s \\ &\le [CHC_{d,s}(H,h)]^s \left[|v^+|_{1,s,\omega_i}^s + |v^-|_{1,s,\omega_i}^s\right] = [CHC_{d,s}(H,h)]^s |v|_{1,s,\omega_i}^s. \end{aligned} \tag{4.23}$$

Consequently,

$$||v - I_i^+ v + I_i^- v||_{0,s,\omega_i} \le CHC_{d,s}(H,h)|v|_{1,s,\omega_i}. \tag{4.24}$$

If $v \neq 0$ at any point of ω_i, then either $v^+ = I_i^+ v = 0$ or $v^- = I_i^- v = 0$. Consequently, there exits at least a node of $\mathcal{T}_h$ in ω_i at which $v - I_i^+ v + I_i^- v = v^+ - v^- - I_i^+ v + I_i^- v$ vanishes. From Lemma 4.1, since $I_i^+ v - I_i^- v$ is a constant, we again get (4.24). We notice that, since for any $x \in \omega_i$ the line segment $[x^i, x]$ lies in ω_i, we can take a decomposition of ω_i as in Lemma 4.1 having $N \le 1$. Assuming that $N = 1$, let ω_{i0} and $\omega_{i1} = \omega$ be this decomposition. Since ω_{i0} contains at least one $\tau \in \mathcal{T}_H$ and the mesh $\mathcal{T}_H$ is regular, then, according to (4.9), C_{ω_i} can be taken independent of H and

h. Consequently, $C(N,s)$ in (4.8) is independent of H and h. Now, using (4.24), we get

$$||I_H v - v||^s_{0,s,\omega_i} = \left\| \sum_{x^j \in \omega_i} \left[I_j^+ v - I_j^- v - v \right] \phi_j \right\|^s_{0,s,\omega_i}$$
$$\leq C \sum_{x^j \in \omega_i} ||I_j^+ v - I_j^- v - v||^s_{0,s,\omega_i \cap \omega_j} \leq [CHC_{d,s}(H,h)]^s \sum_{x^j \in \omega_i} |v|^s_{1,s,\omega_j}.$$

Above, x^j are nodes of $\mathcal{T}_H$, and we used the fact that, since the mesh is regular, the maximum number of ω_j which nonemptily intersects a given ω_i is bounded and independent of H. Now, we again use this property to obtain

$$||I_H v - v||^s_{0,s,\Omega_0} \leq \sum_{x^i \in \Omega_0} ||I_H v - v||^s_{0,s,\omega_i}$$
$$\leq [CHC_{d,s}(H,h)]^s \sum_{x^i \in \Omega_0} \sum_{x^j \in \omega_i} |v|^s_{1,s,\omega_j} \leq [CHC_{d,s}(H,h)]^s \sum_{x^i \in \Omega_0} |v|^s_{1,s,\omega_i}.$$

Also, from the regularity of the mesh, it follows that each ω_i contains a bounded number of simplexes of $\mathcal{T}_H$ which is independent of H. Consequently, we have

$$||I_H v - v||^s_{0,s,\Omega_0} \leq [CHC_{d,s}(H,h)]^s \sum_{\tau \in \mathcal{T}_H} |v|^s_{1,s,\tau},$$

and in this way, we get (4.21).

In order to prove (4.22), we notice first that, from the definition of $I_i^+ v$ and $I_i^- v$, we have for any $x \in \omega_i$,

$$\begin{aligned} &0 \leq I_i^+ v - I_i^- v \leq v(x) \text{ if } v(x) \geq 0, \text{ and} \\ &0 \geq I_i^+ v - I_i^- v \geq v(x) \text{ if } v(x) \leq 0, \end{aligned} \tag{4.25}$$

and therefore,

$$|I_i^+ v - I_i^- v| \leq |v(x)| \text{ for any } x \in \omega_i. \tag{4.26}$$

Using this inequality, we obtain

$$||I_H v||^s_{0,s,\omega_i} = \left\| \sum_{x^j \in \omega_i} (I_j^+ v - I_j^- v) \phi_j \right\|^s_{0,s,\omega_i}$$
$$\leq \int_{\omega_i} \left(\sum_{x^j \in \omega_i} |I_j^+ - I_j^-| \phi_j \right)^s = \int_{\omega_i} \left(\sum_{x^j \in \omega_i} |v(x)| \phi_j \right)^s = \int_{\omega_i} |v(x)|^s = ||v||^s_{0,s,\omega_i}.$$

Taking again into account the regularity of the mesh, we get

$$||I_H v||^s_{0,s,\Omega_0} \leq \sum_{x^i \in \Omega_0} ||I_H v||^s_{0,s,\omega_i} \leq \sum_{x^i \in \Omega_0} ||v||^s_{0,s,\omega_i} \leq C \sum_{\tau \in \mathcal{T}_H} ||v||^s_{0,s,\tau},$$

and therefore, the first equation in (4.22) holds. To prove the second equation in (4.22), first we write

$$|I_H v|^s_{1,s,\omega_i} = \left| \sum_{x^j \in \omega_i} (I_j^+ v - I_j^- v) \phi_j \right|^s_{1,s,\omega_i}$$
$$\leq C H^{d-s} \max_{x^k, x^l \in \omega_i, \omega_k \cap \omega_l \neq \emptyset} |(I_k^+ v - I_k^- v) - (I_l^+ v - I_l^- v)|^s.$$

Since $\omega_k \cap \omega_l \neq \emptyset$, taking into account the definition of $I_i^+ v$ and $I_i^- v$ in (4.18), we get that $I_k^+ v - I_k^- v$ and $I_l^+ v - I_l^- v$ cannot be both different from zero and have different signs. Therefore, if we write

$$|I_p^+ v - I_p^- v| - |I_q^+ v - I_q^- v| = \max_{x^k, x^l \in \omega_i, \omega_k \cap \omega_l \neq \emptyset} |(I_k^+ v - I_k^- v) - (I_l^+ v - I_l^- v)|,$$

using (4.26), we get

$$|I_H v|^s_{1,s,\omega_i} \leq C H^{d-s} (|I_p^+ v - I_p^- v| - |I_q^+ v - I_q^- v|)^s \leq C H^{d-s} (|v(x)| - |I_q^+ v - I_q^- v|)^s$$

for any $x \in \omega_p \cap \omega_q$. Since the mesh $\mathcal{T}_h$ is regular, we have $H^d \leq C|\omega_p \cap \omega_q|$, and integrating the above equation over $\omega_p \cap \omega_q$ we get

$$\begin{aligned} &|I_H v|^s_{1,s,\omega_i} \leq C H^{-s} \textstyle\int_{\omega_p \cap \omega_q} (|v(x)| - |I_q^+ v - I_q^- v|)^s \\ &= C H^{-s} \textstyle\int_{\omega_p \cap \omega_q} |v(x) - (I_q^+ v - I_q^- v)|^s \leq C H^{-s} \textstyle\int_{\omega_q} |v(x) - (I_q^+ v - I_q^- v)|^s. \end{aligned}$$

If there exists a point in ω_q at which v vanishes, then $I_q^+ v = I_q^- v = 0$, and, as in (4.23), we get

$$|I_H v|_{1,s,\omega_i} \leq C C_{d,s}(H,h) |v|_{1,s,\omega_q}.$$

Also, if $v > 0$ or $v < 0$ in ω_q, then there exists $x^q \in \omega_q$, node of $\mathcal{T}_h$, such that $v(x^q) = I_q^+ v - I_q^- v$, and we again get the above inequality applying Lemma 4.1. Finally, using again the fact that the mesh $\mathcal{T}_H$ is regular, we get the second equation in (4.22).

To prove that (4.21) and (4.22) hold on Ω, we see that $I_H v = 0$ on all the sets S_{x^i} introduced in (4.3). Therefore, (4.22) holds on all sets S_{x^i}. Also, since $v(x^i) = 0$, from Lemma 4.1, we get that (4.21) holds on S_{x^i}. Consequently, the above reasoning we made for Ω_0 can be done for Ω, too. From (4.25), (4.19), and (4.20), we get that for any $x \in \Omega$, we have

$$0 \leq I_H v(x) \leq v(x) \text{ if } v(x) \geq 0, \text{ and } 0 \geq I_H v(x) \geq v(x) \text{ if } v(x) \leq 0. \tag{4.27}$$

Therefore, we can find a $\theta(x) \in C^1(\Omega)$, $0 \leq \theta(x) \leq 1$, such that $\theta(x^i) = I_H v(x^i)/v(x^i)$ if $I_H v(x^i) \neq 0$, and $\theta(x^i) = 0$ if $I_H v(x^i) = 0$, at any node x^i of $\mathcal{T}_h$. Consequently, we can write $I_H v = L_h(\theta v + (1-\theta)0)$. Finally, if $0, v \in \mathcal{K}$, and $\mathcal{K}$ has Property 3.1, we get that $I_H v \in \mathcal{K}$. ∎

Now, we can prove the following proposition which shows that the constant C_0 in Assumption 2.1 is independent of the mesh and domain decomposition parameters if H/δ and H/h are constant. This result is similar to that given in [4] for the inequalities coming from minimization of the quadratic functionals. In the first part of the proof, the construction of v_i, $i = 1, \ldots, m$, is similar to that given for the one-level method. In the second part we define an appropriate v_0 using the previous lemma.

PROPOSITION 4.1. *Let $\Omega_1, \ldots, \Omega_m$ be the overlapping decomposition of the domain Ω defined in this section. Then Assumption* 2.1 *is verified for the piecewise linear finite element spaces $V = V_h$ and $V_0 = V_H^0$, $V_i = V_h^i$, and $i = 1, \ldots, m$, defined in* (3.3), (3.4), *and* (4.4), *respectively, and any convex set $K = K_h$ satisfying Property* 3.1. *The constant in* (2.3) *of Assumption* 2.1 *can be taken of the form*

$$C_0 = C(m+2)^{1-\frac{1}{p}} \left(1 + (m-1)\frac{H}{\delta}\right) C_{d,s}(H,h), \tag{4.28}$$

where C is independent of the mesh and domain decomposition parameters, and $C_{d,s}(H,h)$ is given in (4.6).

Proof. Let us consider $w \in K_h$, $w_0 \in V_H^0$, and $w_i \in V_h^i$ such that $w + \sum_{j=0}^{i} w_i \in K_h$, $i = 0, \ldots, m$, and let v be another element in K_h. In the following, we use unity partitions $(\theta_j^i)_{j=i,\ldots,m}$, of the domains $\cup_{j=i,m}^{m} \Omega_j$, $i = 1, \ldots, m$, having property (3.2).

Step 1. We assume that we have a $v_0 \in V_H^0$ satisfying

$$w + v_0,\ v + w_0 - v_0 \in K_h, \tag{4.29}$$

and we recursively construct $v_i \in V_h^i$, $i = 1, \ldots, m$, which satisfies (2.1) and (2.2) in Assumption 2.1. To this end, we define

$$v_1 = L_h\left(\theta_1^1(v - w - v_0) + (1 - \theta_1^1)w_1\right), \tag{4.30}$$

and, as in the previous section, we get

$$\begin{aligned} &v_1 \in V_h^1 \text{ and } w + w_0 + v_1 \in K_h,\\ &v - v_0 - v_1 + w_0 + w_1 \in K_h,\\ &v - w - v_0 - v_1 \in W_0^{1,s}\left(\bigcup_{j=2}^{m} \Omega_j\right) \text{ and}\\ &v - w - v_0 - v_1 = 0 \text{ in } \Omega - \overline{\cup_{j=2}^{m}\Omega_j}. \end{aligned}$$

Also, for $i = 2, \ldots, m-1$ we write

$$v_i = L_h\left(\theta_i^i\left(v - w - \sum_{j=0}^{i-1} v_j\right) + (1 - \theta_i^i)w_i\right), \tag{4.31}$$

and we prove

$$\begin{aligned} v_i \in V_h^i \text{ and } w + \sum_{j=0}^{i-1} w_j + v_i \in K_h,\\ v - \sum_{j=0}^{i} v_j + \sum_{j=0}^{i} w_j \in K_h,\\ v - w - \sum_{j=0}^{i} v_j \in W_0^{1,s}\left(\bigcup_{j=i+1}^{m} \Omega_j\right), \text{ and}\\ v - w - \sum_{j=0}^{i} v_j = 0 \text{ in } \Omega - \overline{\cup_{j=i+1}^{m}\Omega_j}, \end{aligned}$$

assuming that these equations hold for $i-1$. Finally, we take

$$v_m = v - w - \sum_{j=0}^{m-1} v_j \tag{4.32}$$

and we get that (2.1) and (2.2) in Assumption 2.1 hold.

Step 2. We define in this step a $v_0 \in V_H^0$ satisfying (4.29) and prove that condition (2.3) in Assumption 2.1 is satisfied with the constant C_0 given in (4.28). It is easy to see that (4.29) is equivalent with

$$v_0 - w_0 \in (K_h - (w + w_0)) \cap (v - K_h), \tag{4.33}$$

and also, since $v, w + w_0 \in K_h$, we get

$$v - w - w_0 \in (K_h - (w + w_0)) \cap (v - K_h)\,. \tag{4.34}$$

We write $\mathcal{K} = (K_h - (w + w_0)) \cap (v - K_h)$, and from the above equation and Lemma 4.2, we get that $I_H(v - w - w_0) \in \mathcal{K}$. From (4.21) and (4.22) we have

$$||v - w - w_0 - I_H(v - w - w_0)||_{0,s} \leq CHC_{d,s}(H,h)|v - w - w_0|_{1,s} \tag{4.35}$$

and

$$\begin{aligned} &||I_H(v - w - w_0)||_{0,s} \leq CC_{d,s}(H,h)||v - w - w_0||_{0,s} \\ &|I_H(v - w - w_0)|_{1,s} \leq CC_{d,s}(H,h)|v - w - w_0|_{1,s}, \end{aligned} \tag{4.36}$$

where $C_{d,s}(H,h)$ is defined in (4.6). Now we take

$$v_0 = w_0 + I_H(v - w - w_0), \tag{4.37}$$

and, from 4.34, the second part of Lemma 4.2, and (4.33), we get that it satisfies condition (4.29). To prove condition (2.3) in Assumption 2.1, we first notice that, starting from v_1 given in (4.30), by the recurrent application of (4.31), as in the proof of Proposition 3.1, we get v_i, $i = 1, \ldots, m$, of the form

$$v_i = L_h\left(\tau_0^i(v - w - v_0) + \sum_{j=1}^{i} \tau_j^i w_j\right),\ i = 1, \ldots, m, \tag{4.38}$$

where τ_j^i, $i = 1, \ldots, m$, $j = 0, \ldots, i$, satisfy (3.9). Using (3.10) and (3.11), we get

$$||L_h(\tau_j^i w_j)||_{1,s} \leq C||\tau_j^i w_j||_{1,s} \leq C\left(||w_j||_{1,s} + \frac{m-1}{\delta}||w_j||_{0,s}\right).$$

It follows from (4.1) that the diameters of the connected component of Ω_i are less than CH, and since $w_i \in V_h^i$, using the classical Friedrichs–Poincaré inequality, we get

$$||L_h(\tau_j^i w_j)||_{1,s} \leq C\left[1 + (m-1)\frac{H}{\delta}\right]|w_j|_{1,s},\ i = 1, \ldots, m,\ j = 1, \ldots, i. \tag{4.39}$$

On the other hand, taking into account (3.10), (3.11), (4.37), and (4.35), we get

$$\begin{aligned} &||L_h(\tau_0^i(v - w - v_0))||_{1,s} \leq C[|v - w - v_0|_{1,s} + \left(1 + \tfrac{m-1}{\delta}\right)||v - w - v_0||_{0,s}] \\ &= C[|v - w - v_0|_{1,s} + \left(1 + \tfrac{m-1}{\delta}\right)||v - w - w_0 - I_H(v - w - w_0)||_{0,s}] \\ &\leq C\left[|v - w - v_0|_{1,s} + (m-1)C_{d,s}(H,h)\tfrac{H}{\delta}|v - w - w_0|_{1,s}\right] \\ &\leq C(|v - w|_{1,s} + |v_0|_{1,s}) + C(m-1)C_{d,s}(H,h)\tfrac{H}{\delta}(|v - w|_{1,s} + |w_0|_{1,s}). \end{aligned}$$

Consequently, we have

$$\begin{aligned} &||L_h(\tau_0^i(v - w - v_0))||_{1,s} \leq C\left[1 + (m-1)\tfrac{H}{\delta}\right]C_{d,s}(H,h) \\ &\cdot(|v - w|_{1,s} + |w_0|_{1,s}) + C|v_0|_{1,s},\ i = 1, \ldots, m. \end{aligned} \tag{4.40}$$

Also, from (4.37) and (4.36), we get

$$|v_0|_{1,s} = |w_0 + I_H(v - w - w_0)|_{1,s} \le |w_0|_{1,s} + |I_H(v - w - w_0)|_{1,s}$$
$$\le |w_0|_{1,s} + CC_{d,s}(H,h)|v - w - w_0|_{1,s},$$

and therefore,

$$|v_0|_{1,s} \le CC_{d,s}(H,h)(|v - w|_{1,s} + |w_0|_{1,s}). \tag{4.41}$$

Now, from (4.40) and (4.41), we get

$$\begin{aligned} &||L_h(\tau_0^i(v - w - v_0))||_{1,s} \\ &\le C\left[1 + (m-1)\tfrac{H}{\delta}\right] C_{d,s}(H,h)(|v - w|_{1,s} + |w_0|_{1,s}),\ i = 1, \dots, m. \end{aligned} \tag{4.42}$$

Finally, from (4.38), (4.39), (4.41), and (4.42) we obtain that condition (2.3) in Assumption 2.1 holds with C_0 given in (4.28). ∎

Remark 4.2. As in Remark 3.1, we notice that, since the number m of the subdomains Ω_i is the number of colors of the overlapping domain decomposition $\{O_i\}_{1 \le i \le M}$, the error estimates in Theorem 2.1 depends only on C_0 given in (4.28). Therefore, if the overlapping size δ and the mesh sizes H and h are chosen such that H/h and H/δ are constant, then the convergence rate of the two-level multiplicative Schwarz method is independent of the mesh and domain decomposition parameters.

5. Multilevel multiplicative Schwarz method. We consider over the domain $\Omega \subset \mathbf{R}^d$ a family of regular meshes $\mathcal{T}_{h_j}$ of mesh sizes h_j, $j = 1, \dots, L$, such that $\mathcal{T}_{h_{j+1}}$ is a refinement of $\mathcal{T}_{h_j}$, $j = 1, \dots, L-1$. We write

$$\Omega_j = \bigcup_{\tau \in \mathcal{T}_{h_j}} \tau \tag{5.1}$$

and we assume that $\Omega = \Omega_L$. As in the previous section, we assume that, if a node of $\mathcal{T}_{h_j}$ lies on $\partial\Omega_j$, then it lies on $\partial\Omega_{j+1}$, too, that is, it lies on $\partial\Omega$. Also, for the nodes $x^j \in \partial\Omega$ of $\mathcal{T}_{h_j}$, $j = 1, \dots, L-1$, we consider the union of all $\tau \in \mathcal{T}_{h_j}$ having x^j as a vertex, ω_j, and define the set S_{x^j} as the union of ω_j with all $\tau \in \mathcal{T}_{h_{j+1}}$, $\tau \not\subset \Omega_j$, which are contained in the smallest sphere which is centered at x^j and contains ω_j. We assume that

$$\Omega_{j+1} \backslash \Omega_j \subset \bigcup_{x^j \text{ node of } \mathcal{T}_{h_j},\ x^j \in \partial\Omega} S_{x^j} \text{ for } j = 1, \dots, L-1. \tag{5.2}$$

Since the mesh $\mathcal{T}_{h_{j+1}}$ is a refinement of $\mathcal{T}_{h_j}$, we have $h_{j+1} \le h_j$, and assume that there exists a constant γ, independent of the number of meshes, L, such that

$$1 < \gamma \le \frac{h_j}{h_{j+1}}, \quad j = 1, \dots, L-1. \tag{5.3}$$

At each level $j = 1, \dots, L$, we consider an overlapping decomposition $\{O_j^i\}_{1 \le i \le M_j}$ of Ω_j, and assume that the mesh partition $\mathcal{T}_{h_j}$ of Ω_j supplies a mesh partition for each O_j^i, $1 \le i \le M_j$. Also, we assume that the overlapping size for the domain decomposition at the level $1 \le j \le L$ is δ_j, i.e.,

$$O_j^i \cap \partial\left(\bigcup_{l \ne i} O_j^l\right) \ne \emptyset \text{ and } \operatorname{dist}\left(\partial O_j^i \backslash \partial\Omega_j, O_j^i \cap \partial\left(\bigcup_{l \ne i} O_j^l\right)\right) \ge \delta_j \tag{5.4}$$

is satisfied. In addition, we suppose that there exists a constant C such that

$$\operatorname{diam}(O^i_{j+1}) \le Ch_j,\ j = 1, \dots, L-1,\ i = 1, \dots, M_j. \tag{5.5}$$

Now, at each level $j = 1, \dots, L$, we color the subdomains O^i_j, $i = 1, \dots, M_j$, and obtain the overlapping subdomains Ω^i_j, $i = 1, \dots, m_j$, as in the previous section. Finally, we assume that $m_1 = 1$, and write

$$m = \max_{j=1,\dots,L} m_j. \tag{5.6}$$

At each level $j = 1, \dots, L$, we introduce the linear finite element spaces,

$$V_{h_j} = \{v \in C^0(\bar{\Omega}_j) : v|_\tau \in P_1(\tau),\ \tau \in \mathcal{T}_{h_j},\ v = 0 \text{ on } \partial\Omega_j\}, \tag{5.7}$$

and, for $i = 1, \dots, m_j$, we write

$$V^i_{h_j} = \{v \in V_{h_j} : v = 0 \text{ in } \Omega_j \backslash \Omega^i_j\}. \tag{5.8}$$

The convex set will be a subset K_{h_L} of V_{h_L} having Property 3.1.

In order to prove that Assumption 2.1 holds for the convex set $K = K_{h_L}$ and the spaces $V = V_{h_L}$, $V^i_j = V^i_{h_j}$, $j = 1, \dots, L$, $i = 1, \dots, m_j$, and to find the constant C_0 in (2.3) as a function of the domain decomposition and mesh parameters, we need the following lemma. This result generalizes to more than two levels the second inequality (4.22) in Lemma 4.2. To this end, we introduce operators $I_{h_k} : V_{h_{k+1}} \to V_{h_k}$, $k = 1, \dots, L-1$, which are similar to the operator $I_H : V_h \to V_H$ defined in (4.20).

LEMMA 5.1. *For a given $1 \le j < L-1$, let $v_k,\ w_k \in V_{h_k}$, $k = j+1, \dots, L-1$, such that*

$$v_k = w_k + I_{h_k}(v_{k+1}). \tag{5.9}$$

Then,

$$\begin{aligned} |I_{h_j} v_{j+1}|_{1,s,\Omega_j} \le C(L-j)^{\frac{s-1}{s}} \Bigg\{ & \sum_{k=j+1}^{L-1} C_{d,s}(h_j, h_k)^s |w_k|^s_{1,s,\Omega_j} \\ & + C_{d,s}(h_j, h_L)^s |v_L|^s_{1,s,\Omega_j} \Bigg\}^{\frac{1}{s}}. \end{aligned} \tag{5.10}$$

Moreover, (5.10) *also holds if its seminorms over Ω_j are replaced with seminorms over Ω_k, for any $k = j+1, \dots, L$.*

Proof. Let ω_j be the support of the nodal basis function in V_{h_j} corresponding to the node x^j of $\mathcal{T}_{h_j}$. Then there exists two nodes of $\mathcal{T}_{h_j}$, $x^j_1, x^j_2 \in \omega_j$, such that

$$|I_{h_j} v_{j+1}|^s_{1,s,\omega_j} \le Ch_j^{d-s} |(I_{h_j} v_{j+1})(x^j_1) - (I_{h_j} v_{j+1})(x^j_2)|^s. \tag{5.11}$$

Starting from (5.11), we prove that, for each $k = j, \dots, L-1$, there exist two nodes of $\mathcal{T}_{h_{k+1}}$, $x^{k+1}_1 \in \omega^1_k$ and $x^{k+1}_2 \in \omega^2_k$, ω^1_k and ω^2_k being the supports of the nodal basis function in V_{h_k} corresponding to the nodes x^k_1 and x^k_2 of $\mathcal{T}_{h_k}$, respectively, such that

$$\begin{aligned} & |I_{h_j} v_{j+1}|^s_{1,s,\omega_j} \\ & \le Ch_j^{d-s} \left[\sum_{k=j+1}^{L-1} |w_k(x^k_1) - w_k(x^k_2)| + |v_L(x^L_1) - v_L(x^L_2)| \right]^s. \end{aligned} \tag{5.12}$$

First, we assume that, starting with x_1^j and x_2^j in (5.11), at each level $k = j, \dots, L-1$, the values of $(I_{h_k} v_{k+1})(x_1^k)$ and $(I_{h_k} v_{k+1})(x_2^k)$ are obtained as values of v_{k+1} at two nodes $x_1^{k+1} \in \omega_k^1$ and $x_2^{k+1} \in \omega_k^2$, respectively, that is, we have

$$(I_{h_k} v_{k+1})(x_1^k) = v_{k+1}(x_1^{k+1}) \text{ and } (I_{h_k} v_{k+1})(x_2^k) = v_{k+1}(x_2^{k+1}). \tag{5.13}$$

Therefore, starting from (5.11), using (5.9), for any $j+1 \le N \le L-1$, we get

$$\begin{aligned} &|I_{h_j} v_{j+1}|^s_{1,s,\omega_j} \le C h_j^{d-s} \\ &\cdot \left[\sum_{k=j+1}^{N} |w_k(x_1^k) - w_k(x_2^k)| + |(I_{h_N} v_{N+1})(x_1^N) - (I_{h_N} v_{N+1})(x_2^N)| \right]^s , \end{aligned} \tag{5.14}$$

and, consequently, we get (5.12). Now, we prove that there exist some nodes x_1^k and x_2^k of $\mathcal{T}_{h_k}$ such that (5.12) holds, even if (5.13) does not hold for all $k = j, \dots, L-1$. Let us assume that (5.13) holds for $k = j, \dots, N$, $j+1 \le N \le L-1$. Therefore we can get (5.14). From the definition of the operators I_i^+ and I_i^- in (4.18), it follows that if, for instance, $(I_{h_N} v_{N+1})(x_1^N) \ne v_{N+1}(x)$ for any node $x \in \omega_N^1$ of $\mathcal{T}_{h_{N+1}}$, then $(I_{h_N} v_{N+1})(x_1^N) = 0$ and v_{N+1} takes both positive and negative values at the nodes of $\mathcal{T}_{h_{N+1}}$ in ω_N^1. Consequently, if both $(I_{h_N} v_{N+1})(x_1^N) \ne v_{N+1}(x)$ for any node $x \in \omega_N^1$ of $\mathcal{T}_{h_{N+1}}$, and $(I_{h_N} v_{N+1})(x_2^N) \ne v_{N+1}(x)$ for any node $x \in \omega_N^2$ of $\mathcal{T}_{h_{N+1}}$, then $(I_{h_N} v_{N+1})(x_1^N) = (I_{h_N} v_{N+1})(x_2^N) = 0$, and we get that (5.12) holds for some arbitrary nodes of $\mathcal{T}_{h_k}$, $x_1^k \in \omega_{k-1}^1$, $x_2^k \in \omega_{k-1}^2$, $N+1 \le k \le L$. Also, if $(I_{h_N} v_{N+1})(x_1^N) \ne v_{N+1}(x)$ for any node $x \in \omega_N^1$ of $\mathcal{T}_{h_{N+1}}$, but there exists $x_2^{N+1} \in \omega_N^2$, node of $\mathcal{T}_{h_{N+1}}$, such that $(I_{h_N} v_{N+1})(x_2^N) = v_{N+1}(x_2^{N+1})$, then

$$|(I_{h_N} v_{N+1})(x_1^N) - (I_{h_N} v_{N+1})(x_2^N)| = |v_{N+1}(x_2^{N+1})| \le |v_{N+1}(x_1^{N+1}) - v_{N+1}(x_2^{N+1})|,$$

where $x_1^{N+1} \in \omega_N^1$ is an arbitrary node of $\mathcal{T}_{h_{N+1}}$ for which $v_{N+1}(x_1^{N+1})$ and $v_{N+1}(x_2^{N+1})$ have different signs. In this way we get that (5.14) holds for $N+1$, and can continue the same reasoning for $N+2 \le k \le L-1$.

If we write $\omega_{j-1}^1 = \omega_{j-1}^2 = \omega_j$, since, for $k = j, \dots, L$, the above nodes x_1^k and x_2^k of $\mathcal{T}_{h_k}$ belong to ω_{k-1}^1 and ω_{k-1}^2, respectively, and $\operatorname{diam}(\omega_{k-1}^1)$, $\operatorname{diam}(\omega_{k-1}^2) \le 2h_k$, then x_1^k and x_2^k, $k = j, \dots, L$, belong to the sphere centered at x^j and having the radius of $2h_j + h_{j+1} + h_{j+2} + \cdots + h_{L-1}$. Using (5.3), we get that they belong to the sphere centered at x^j with the radius of $\frac{2\gamma-1}{\gamma-1} h_j$. Consequently, if we write

$$\tilde{\omega}_j = \bigcup_{\tau \in \mathcal{T}_{h_j},\ \operatorname{dist}(\mathrm{x^j}, \tau) \le \frac{\gamma}{\gamma-1} \mathrm{h_j}} \tau, \tag{5.15}$$

then $x_1^k, x_2^k \in \tilde{\omega}_j$, $k = j, \dots, L$. For any $x \in \tilde{\omega}_j$, we get from (5.12),

$$\begin{aligned} &|I_{h_j} v_{j+1}|^s_{1,s,\omega_j} \le C h_j^{d-s} [2(L-j)]^{s-1} \\ &\left\{ \sum_{k=j+1}^{L-1} [|w_k(x_1^k) - w_k(x)|^s + |w_k(x_2^k) - w_k(x)|^s] \right. \\ &\left. + |v_L(x_1^L) - v_L(x)|^s + |v_L(x_2^L) - v_L(x)|^s \right\}, \end{aligned}$$

and integrating over $\tilde{\omega}_j$ we have,

$$(\tfrac{2\gamma-1}{\gamma-1}h_j)^d|I_{h_j}v_{j+1}|^s_{1,s,\omega_j} \le Ch_j^{d-s}[2(L-j)]^{s-1}$$
$$\left\{\sum_{k=j+1}^{L-1}[||w_k(x_1^k)-w_k||^s_{0,s,\tilde{\omega}_j}+||w_k(x_2^k)-w_k||^s_{0,s,\tilde{\omega}_j}]\right.$$
$$\left.+||v_L(x_1^L)-v_L||^s_{0,s,\tilde{\omega}_j}+||v_L(x_2^L)-v_L||^s_{0,s,\tilde{\omega}_j}\right\}.$$

From this inequality and (4.16), we get

$$|I_{h_j}v_{j+1}|^s_{1,s,\omega_j} \le C(L-j)^{s-1}\left(\tfrac{\gamma-1}{2\gamma-1}\right)^d$$
$$\left\{\sum_{k=j+1}^{L-1}C_{d,s}\left(2h_j\frac{2\gamma-1}{\gamma-1},h_k\right)^s|w_k|^s_{1,s,\tilde{\omega}_j}+C_{d,s}\left(2h_j\frac{2\gamma-1}{\gamma-1},h_L\right)^s|v_L|^s_{1,s,\tilde{\omega}_j}\right\},$$

and, taking into account the definition of $C_{d,s}$ in (4.6), we have

$$|I_{h_j}v_{j+1}|^s_{1,s,\omega_j} \le C(L-j)^{s-1}\left\{\sum_{k=j+1}^{L-1}C_{d,s}(h_j,h_k)^s|w_k|^s_{1,s,\tilde{\omega}_j}+C_{d,s}(h_j,h_L)^s|v_L|^s_{1,s,\tilde{\omega}_j}\right\}.$$

Finally, since the mesh $\mathcal{T}_{h_j}$ is regular and γ is independent of L and of the mesh parameters, then ω_j and $\tilde{\omega}_j$ contain a bounded number of simplexes in $\mathcal{T}_{h_j}$, which is also independent of L and of the mesh parameters. Consequently, we get (5.10). Since the nodes of $\mathcal{T}_{h_j}$ belonging to $\partial\Omega_j$ lie also on $\partial\Omega_{j+1}$, and $v_{j+1}=0$ on $\partial\Omega_{j+1}$, it follows that $I_{h_j}v_{j+1}=0$ on $\partial\Omega_j$. Consequently, they are extended with zero to Ω_k, $j+1\le k\le L$, and (5.10) holds for these domains, too. ∎

The following proposition shows that Assumption 2.1 holds for the multilevel method and writes the constant C_0 as a function of the domain decomposition and mesh parameters.

PROPOSITION 5.1. *Let, for each level $j=1,\dots,L$, $\Omega_j^1,\dots,\Omega_j^{m_j}$ be the overlapping decomposition of the domain Ω_j defined in this section with $\Omega_L=\Omega$ and $m_1=1$. Then Assumption* 2.1 *is verified for the piecewise linear finite element spaces, $V=V_{h_L}$ and $V_j^i=V_{h_j}^i$, $j=1,\dots,L$, $i=1,\dots,m_j$ defined in* (5.7) *and* (5.8)*, respectively, and any convex set $K=K_{h_L}\subset V_{h_L}$ with Property* 3.1. *The constant in* (2.3) *of Assumption* 2.1 *can be taken of the form*

$$C_0 = Cm^2(L+1)^{2-\frac{1}{p}-\frac{1}{s}}\sum_{j=1}^{L}\left[1+(m-1)\frac{h_{j-1}}{\delta_j}\right]C_{d,s}(h_{j-1},h_L) \tag{5.16}$$

in which we take $h_0=h_1$, C is independent of the mesh and domain decomposition parameters, and $C_{d,s}(H,h)$ is given in (4.6).

Proof. Let us consider $w\in K_{h_L}$, $w_j^i\in V_{h_j}^i$, $j=1,\dots,L$, $i=1,\dots,m_j$, such that $w+\sum_{j=1}^{k-1}\sum_{i=1}^{m_j}w_j^i+\sum_{i=1}^{l}w_k^i\in K_{h_L}$, $k=1,\dots,L$, $l=1,\dots,m_k$, and let v be another element in K_{h_L}. For $j=1,\dots,L$, we write

$$w_j^0=\sum_{i=1}^{m_j}w_j^i \quad\text{and}\quad w_j=\sum_{k=1}^{j}w_k^0=\sum_{k=1}^{j}\sum_{i=1}^{m_j}w_j^i.$$

Since $v, w + w_{L-2} \in K_{h_L}$, and also, $w + w_{L-2} + w^0_{L-1} \in K_{h_L}$ and $w + w_{L-2} + w^0_{L-1} + \sum_{i=1}^{l} w^i_L \in K_{h_L}$, $l = 1, \dots, m_L$, as in the proof of Proposition 4.1, we get that there exist $v^0_{L-1} \in V_{h_{L-1}}$ and $v^i_L \in V^i_{h_L}$, $i = 1, \dots, m_L$ such that

$$w + w_{L-2} + v^0_{L-1} \in K_{h_L}, \tag{5.17}$$

$$w + w_{L-2} + w^0_{L-1} + \sum_{i=1}^{l-1} w^i_L + v^l_L \in K_{h_L},\ l = 1, \dots, m_L, \tag{5.18}$$

$$v - w - w_{L-2} = v^0_{L-1} + \sum_{i=1}^{m_L} v^i_L, \tag{5.19}$$

and

$$\begin{aligned} &v^0_{L-1} = w^0_{L-1} + I_{h_{L-1}}(v - w - w_{L-2} - w^0_{L-1}), \\ &v^i_L = L_{h_L}(\tau^i_0 \left(v - w - w_{L-2} - v^0_{L-1}) + \sum_{l=1}^{i} \tau^i_l w^l_L \right),\ i = 1, \dots, m_L, \end{aligned} \tag{5.20}$$

where τ^i_j, $i = 1, \dots, m$, $j = 0, \dots, i$, satisfy (3.9). In this way, using (5.17), we get that $w + w_{L-3} + w^0_{L-2} + v^0_{L-1}, w + w_{L-3} \in K_{h_L}$, and also, $w + w_{L-3} + w^0_{L-2} \in K_{h_L}$ and $w + w_{L-3} + w^0_{L-2} + \sum_{i=1}^{l} w^i_{L-1} \in K_{h_L}$, $l = 1, \dots, m_{L-1}$. Consequently, there exist $v^0_{L-2} \in V_{h_{L-2}}$ and $v^i_{L-1} \in V^i_{h_{L-1}}$, $i = 1, \dots, m_{L-1}$ such that

$$w + w_{L-3} + v^0_{L-2} \in K_{h_L}, \tag{5.21}$$

$$w + w_{L-3} + w^0_{L-2} + \sum_{i=1}^{l-1} w^i_{L-1} + v^l_{L-1} \in K_{h_L},\ l = 1, \dots, m_{L-1}, \tag{5.22}$$

$$w^0_{L-2} + v^0_{L-1} = v^0_{L-2} + \sum_{i=1}^{m_{L-1}} v^i_{L-1}, \tag{5.23}$$

and

$$\begin{aligned} &v^0_{L-2} = w^0_{L-2} + I_{h_{L-2}}(v^0_{L-1}) \\ &v^i_{L-1} = L_{h_{L-1}} \left(\tau^i_0 (w^0_{L-2} + v^0_{L-1} - v^0_{L-2}) + \sum_{l=1}^{i} \tau^i_l w^l_{L-1} \right),\ i = 1, \dots, m_{L-1}. \end{aligned} \tag{5.24}$$

Starting with (5.21) we successively get for $j = 3, \dots, L-1$ that $w + w_{L-j-1} + w^0_{L-j} + v^0_{L-j+1}, w + w_{L-j-1} \in K_{h_L}$, and also, $w + w_{L-j-1} + w^0_{L-j} \in K_{h_L}$ and $w + w_{L-j-1} + w^0_{L-j} + \sum_{i=1}^{l} w^i_{L-j+1} \in K_{h_L}$, $l = 1, \dots, m_{L-j+1}$. Consequently, there exist $v^0_{L-j} \in V_{h_{L-j}}$ and $v^i_{L-j+1} \in V^i_{h_{L-j+1}}$, $i = 1, \dots, m_{L-j+1}$ such that

$$w + w_{L-j-1} + v^0_{L-j} \in K_{h_L}, \tag{5.25}$$

$$\begin{aligned} w + w_{L-j-1} + w^0_{L-j} + \sum_{i=1}^{l-1} w^i_{L-j+1} \\ + v^l_{L-j+1} \in K_{h_L},\ l = 1, \dots, m_{L-j+1}, \end{aligned} \tag{5.26}$$

$$w^0_{L-j} + v^0_{L-j+1} = v^0_{L-j} + \sum_{i=1}^{m_{L-j+1}} v^i_{L-j+1}, \tag{5.27}$$

and

$$
\begin{aligned}
&v^0_{L-j} = w^0_{L-j} + I_{h_{L-j}}(v^0_{L-j+1}) \\
&v^i_{L-j+1} = L_{h_{L-j+1}}\left(\tau^i_0(w^0_{L-j} + v^0_{L-j+1} - v^0_{L-j})\right. \\
&\left. + \sum_{l=1}^{i} \tau^i_l w^l_{L-j+1}\right),\ i = 1, \ldots, m_{L-j+1}.
\end{aligned} \tag{5.28}
$$

If we write $v^1_1 = v^0_1$, since $m_1 = 1$, then (5.18), (5.22), and (5.26) prove that (2.1) of Assumption 2.1 holds. Also, we get (2.2) of Assumption 2.1 from (5.19), (5.23), and (5.27). Now, if we write

$$
v^0_L = v - w - w_{L-1}, \tag{5.29}
$$

we get from (5.20), (5.24), and (5.28) that

$$
\begin{aligned}
&v^0_{j-1} = w^0_{j-1} + I_{h_{j-1}}(v^0_j), \\
&v^i_j = L_{h_j}\left(\tau^i_0(w^0_{j-1} + v^0_j - v^0_{j-1}) + \sum_{l=1}^{i} \tau^i_l w^l_j\right), \\
&\text{for } j = 2, \ldots, L,\ i = 1, \ldots, m_j.
\end{aligned} \tag{5.30}
$$

Similar to (4.39), we get that

$$
\begin{aligned}
&||L_{h_j}(\tau^i_l w^l_j)||_{1,s} \le C\left[1 + (m_j - 1)\tfrac{h_{j-1}}{\delta_j}\right]|w^l_j|_{1,s}, \\
&j = 2, \ldots, L,\ i = 1, \ldots, m_j,\ l = 1, \ldots, i.
\end{aligned} \tag{5.31}
$$

Replacing v^0_{j-1} given in the first equation of (5.30) into the second equation of (5.30), and using (4.21) and (4.22), we get

$$
\begin{aligned}
&||L_{h_j}(\tau^i_0(w^0_{j-1} + v^0_j - v^0_{j-1}))||_{1,s} = ||L_{h_j}(\tau^i_0(v^0_j - I_{h_{j-1}}v^0_j))||_{1,s} \\
&\le C\left[|v^0_j - I_{h_{j-1}}v^0_j|_{1,s} + \left(1 + \tfrac{m_{j-1}}{\delta_j}\right)||v^0_j - I_{h_{j-1}}v^0_j||_{0,s}\right] \\
&\le C\left\{[1 + C_{d,s}(h_{j-1}, h_j)]|v^0_j|_{1,s} + \left(1 + \tfrac{m_j - 1}{\delta_j}\right)h_{j-1}C_{d,s}(h_{j-1}, h_j)|v^0_j|_{1,s}\right\}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&||L_{h_j}(\tau^i_0(w^0_{j-1} + v^0_j - v^0_{j-1}))||_{1,s} \\
&\le C\left[1 + (m_j - 1)\tfrac{h_{j-1}}{\delta_j}\right]C_{d,s}(h_{j-1}, h_j)|v^0_j|_{1,s} \text{ for } j = 2, \ldots, L,\ i = 1, \ldots, m_j.
\end{aligned} \tag{5.32}
$$

From the second equation in (5.30), (5.31), and (5.32), for $j = 2, \ldots, L$ and $i = 1, \ldots, m_j$, we get

$$
\begin{aligned}
&||v^i_j||_{1,s} \le C[1 + (m_j - 1)\tfrac{h_{j-1}}{\delta_j}]C_{d,s}(h_{j-1}, h_j)|v^0_j|_{1,s} \\
&+ C\left[1 + (m_j - 1)\frac{h_{j-1}}{\delta_j}\right]\sum_{l=1}^{i}|w^l_j|_{1,s},
\end{aligned}
$$

and using (5.6), we have

$$
\begin{aligned}
&||v^i_j||_{1,s} \le C\left[1 + (m - 1)\tfrac{h_{j-1}}{\delta_j}\right]C_{d,s}(h_{j-1}, h_j)|v^0_j|_{1,s} \\
&+ C\left[1 + (m - 1)\frac{h_{j-1}}{\delta_j}\right]\sum_{l=1}^{m_j}|w^l_j|_{1,s} \text{ for } j = 2, \ldots, L.
\end{aligned} \tag{5.33}
$$

The first equation in (5.30) shows that the conditions of Lemma 5.1 are satisfied, and we get from (5.10) that for $j = 1, \dots, L-1$,

$$|v_j^0|_{1,s} \le C(L-j)^{\frac{s-1}{s}} \left[\sum_{k=j}^{L-1} C_{d,s}(h_j, h_k)^s |w_k^0|_{1,s}^s + C_{d,s}(h_j, h_L)^s |v_L^0|_{1,s}^s \right]^{\frac{1}{s}}.$$

Since $C_{d,s}(h_j, h_k) \le C_{d,s}(h_j, h_L)$, $j = 1, \dots L-1$, $j \le k \le L-1$, using (5.29), we get

$$|v_j^0|_{1,s} \le C(L-1)^{\frac{s-1}{s}} C_{d,s}(h_j, h_L) \left[\sum_{k=1}^{L-1} |w_k^0|_{1,s} + |v-w|_{1,s} \right] \quad \text{for } j = 1, \dots, L-1. \tag{5.34}$$

From (4.6), we have $C_{d,s}(h_{j-1}, h_j) C_{d,s}(h_j, h_L) \le C_{d,s}(h_{j-1}, h_L)$, and using it we get from (5.33) and (5.34),

$$||v_j^i||_{1,s} \le C(L-1)^{\frac{s-1}{s}} \left[1 + (m-1)^{\frac{h_{j-1}}{\delta_j}} \right] C_{d,s}(h_{j-1}, h_L) \left[\sum_{k=1}^{L} \sum_{l=1}^{m_k} |w_k^l|_{1,s} + |v-w|_{1,s} \right], \quad \text{for } j = 2, \dots, L.$$

Since $m_1 = 1$ and we have written $v_1^1 = v_1^0$ which vanishes on $\partial\Omega$, it follows from (5.34) that the above equation also holds for $j = 1$ with $h_0 = h_1$. From this equation we get

$$||v_j^i||_{1,s} \le C m^{\frac{p-1}{p}} (L+1)^{\frac{p-1}{p}} (L-1)^{\frac{s-1}{s}} \left[1 + (m-1)^{\frac{h_{j-1}}{\delta_j}} \right] C_{d,s}(h_{j-1}, h_L) \left[\sum_{k=1}^{L} \sum_{l=1}^{m_k} |w_k^l|_{1,s}^p + |v-w|_{1,s}^p \right]^{\frac{1}{p}}, \tag{5.35}$$

and (5.16) follows from it. □

5.1. Multigrid method. In the above multilevel method a mesh is the refinement of that on the previous level, but the domain decompositions are almost independent from one level to another. The multigrid method is obtained from the multilevel method by taking the subsets O_j^i of a particular form: we associate at each node x_j^i of $\mathcal{T}_{h_j}$, $j = 1, \dots, L$, $i = 1, \dots, M_j$, an O_j^i defined as the union of the simplexes in $\mathcal{T}_{h_j}$ having x_j^i as a vertex. Consequently, the subspaces $V_{h_j}^i$ will be direct sums of some one-dimensional spaces generated by the nodal basis functions associated with the nodes of $\mathcal{T}_{h_j}$. Evidently, all the previous assumptions on the domain decompositions are satisfied and we can take $\delta_j = h_j$. In the multigrid methods, the construction of a finer mesh from a coarse one, is made following the same procedure of division of the simplexes at each level. Therefore, we can replace (5.3) by

$$1 < \gamma \le \frac{h_j}{h_{j+1}} \le C\gamma, \quad j = 1, \dots, L-1, \tag{5.36}$$

where the constant C is independent of the number of meshes. Starting with the expression of the constant C_0 in (5.16), using (5.36), we have

$$\begin{aligned} & Cm^2(L+1)^{2-\frac{1}{p}-\frac{1}{s}} \sum_{j=1}^{L} \left[1+(m-1)\frac{h_{j-1}}{\delta_j}\right] C_{d,s}(h_{j-1},h_L) \\ & \leq Cm^2(L+1)^{2-\frac{1}{p}-\frac{1}{s}} L[1+(m-1)\gamma] C_{d,s}(h_1,h_L) \\ & \leq Cm^3 L^{3-\frac{1}{p}-\frac{1}{s}} \gamma C_{d,s}(h_1,h_L). \end{aligned}$$

If we write $h = h_1$ and denote by H the diameter of Ω, then the constant C_0 can be taken as

$$C_0 = CL^{3-\frac{1}{p}-\frac{1}{s}} \gamma C_{d,s}(H,h). \tag{5.37}$$

We point out that an iteration of Algorithm 2.1 using the one-dimensional spaces generated by the basis functions corresponding to the nodes of the L meshes represents half of a V-cycle multigrid iteration. Since a full V-cycle multigrid iteration uses these one-dimensional spaces more than once, in order to describe it we should repeat them in the list of the subspaces used by Algorithm 2.1. Consequently, for the multigrid method, only L in the expression of C_0 in (5.37) should be multiplicated by a constant. Therefore, C_0 given in (5.37) is valid for the multigrid method, too.

REFERENCES

[1] L. Badea, *A generalization of the Schwarz alternating method to an arbitrary number of subdomains*, Numer. Math., 55 (1989), pp. 61–81.

[2] L. Badea, *On the Schwarz alternating method with more than two subdomains for nonlinear monotone problems*, SIAM J. Numer. Anal., 28 (1991), pp. 179–204.

[3] L. Badea and J. Wang, *An additive Schwarz method for variational inequalities*, Math. Comp., 69 (2000), pp. 1341–1354.

[4] L. Badea, X.-C. Tai, and J. Wang, *Convergence rate analysis of a multiplicative Schwarz method for variational inequalities*, SIAM J. Numer. Anal., 41 (2003), pp. 1052–1073.

[5] L. Badea, *Convergence rate of a multiplicative Schwarz method for strongly nonlinear inequalities*, in Analysis and Optimization of Differential Systems, V. Barbu, I. Lasiecka, D. Tiba, and C. Varsan, eds., Kluwer Academic Publishers, Boston, 2003, also available from http://www.imar.ro/lbadea.

[6] L. Badea, *Domain decomposition Schwarz method for strongly nonlinear inequalities*, preprint, series of the Institute of Mathematics of the Romanian Academy, Bucharest, 2002.

[7] L. Badea, *On a multiplicative Schwarz domain decomposition method for variational inequalities*, in Current Topics in Continuum Mechanics, vol. II, Lazar Dragos, ed., Editura Academiei, Bucharest, 2003, pp. 11–40.

[8] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp., 57 (1991), pp. 1–21.

[9] J. H. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.

[10] T. Chan, T. Hou, and P. L. Lions, *Geometry related convergence results for domain decomposition algorithms*, SIAM J. Numer. Anal., 28 (1991), pp. 378–391.

[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[12] M. Dryja, *An additive Schwarz algorithm for two- and three-dimensional finite element elliptic problems*, in T. Chan et al., eds., Domain Decomposition Methods, SIAM, Philadelphia, 1989, pp. 168–172.

[13] M. Dryja and O. Widlund, *Some domain decomposition algorithms for elliptic problems*, in L. Hayes and D. Kincaid, eds., Iterative Methods for Large Systems, Academic Press, Boston, 1990, pp. 273–291.

[14] M. Dryja and O. Widlund, *Towards a unified theory of domain decomposition algorithms for elliptic problems*, in T. Chan et al., eds., Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1990, pp. 3–21.

[15] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[16] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[17] R. Glowinski, J. L. Lions, and R. Trémolières, *Analyse numérique des inéquations variationnelles*, Dunod, Paris, 1976.

[18] R. Glowinski and A. Marrocco, *Sur l'aproximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, Rev. Francaise Automat. Informat. Recherche Opérationnelle, Sér. Rouge Anal. Numér., 9, 1975, pp. 41–76.

[19] K. H. Hoffmann and J. Zou, *Parallel algorithms of Schwarz variant for variational inequalities*, Numer. Funct. Anal. Optim., 13 (1992), pp. 449–462.

[20] K. H. Hoffmann and J. Zou, *Parallel solution of variational inequality problems with nonlinear source terms*, IMA J. Numer. Anal. 16 (1996), pp. 31–45.

[21] R. Kornhuber, *Monotone multigrid methods for elliptic variational inequalities*, I, Numer. Math. 69 (1994), pp. 167–184.

[22] Y. Kuznetsov and P. Neittaanmäki, *Overlapping domain decomposition methods for the simplified Dirichlet–Signorini problem*, Computational and Applied Mathematics II, in W. Ames and P. van der Houwen, eds., North-Holland, Amsterdam, 1992, pp. 297–306.

[23] Y. Kuznetsov, P. Neittaanmäki, and P. Tarvainen, *Block relaxation methods for algebraic obstacle problems with M-matrices*, East-West J. Numer. Math., 2 (1994), pp. 75–89.

[24] Y. Kuznetsov, P. Neittaanmäki, and P. Tarvainen, *Overlapping domain decomposition methods for the obstacle problem*, in Domain Decomposition Methods in Science and Engineering, Y. Kuznetsov et al., eds., AMS, Providence, RI, 1994, pp. 271–277.

[25] P. Le Tallec, *Domain decomposition methods in computational mechanics*, in Comput. Mech. Adv., vol. 1, J. T. Oden, ed., North-Holland, Amsterdam, 1994, pp. 121–220.

[26] P. L. Lions, *On the Schwarz alternating method*. I, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski et al., eds., SIAM, Philadelphia, 1988, pp. 2–42.

[27] P. L. Lions, *On the Schwarz alternating method*. II., in Stochastic Interpretation and Order Properties, T. Chan et al., eds., Domain Decomposition Methods, Philadelphia, SIAM, 1989, pp. 47–70.

[28] P. L. Lions, *On the Schwarz alternating method*. A Variant for Nonoverlapping Domains III, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, Chan et al., eds., Philadelphia, SIAM, 1990, pp. 202–223.

[29] T. Lü, C. Liem, and T. Shih, *Parallel algorithms for variational inequalities based on domain decomposition*, System Sci. Math. Sci., 4 (1991), pp. 341–348.

[30] S-H Lui, *On monotone and Schwarz alternating methods for nonlinear elliptic Pdes*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1–15.

[31] J. Mandel, *A multilevel iterative method for symmetric, positive definite linear complementarity problems*, Appl. Math. Optim., 11 (1984), pp. 77–95.

[32] S. Nepomnyaschikh, *Application of domain decomposition to elliptic problems with discontinuous coefficients*, in eds., Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski et al., eds., SIAM, Philadelphia, 1991, pp. 242–251.

[33] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, New York, 1999.

[34] B. F. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[35] X.-C. Tai, *Parallel function and space decomposition methods. Part* I. *Function decomposition*, Beijing Math., 1 (1991), pp. 104–134.

[36] X.-C. Tai, *Parallel function and space decomposition methods. Part* II. *Space decomposition*, Beijing Math., 1 (1991), pp. 135–152.

[37] X.-C. Tai, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, Numer. Math., 93 (2003), pp. 755–786.

[38] X.-C. Tai and M. Espedal, *Rate of convergence of some space decomposition methods for linear and nonlinear problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1558–1570.

[39] X.-C. Tai and P. Tseng, *Convergence rate analysis of an asynchronous space decomposition method for convex minimization*, Math. Comput., 71 (2002), pp. 1105–1135.
[40] X.-C. Tai and J. Xu, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.
[41] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[42] J. Zeng and S. Zhou, *On monotone and geometric convergence of Schwarz methods for two-sided obstacle problems*, SIAM J. Numer. Anal., 35 (1998), pp. 600–616.
[43] J. Zeng and S. Zhou, *Schwarz algorithm for the solution of variational inequalities with nonlinear source terms*, Appl. Math. Comput., 97 (1998), pp. 23–35.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 478–493

A MULTIGRID PRECONDITIONER FOR THE MIXED FORMULATION OF LINEAR PLANE ELASTICITY*

JOSEPH E. PASCIAK† AND YANQIU WANG‡

Abstract. In this paper, we develop a multigrid preconditioner for the discrete system of linear equations that results from the mixed formulation of the linear plane elasticity problem using the Arnold–Winther elements. This, in turn, can be reduced to the problem of finding a multigrid preconditioner for the form $(\cdot,\cdot)+(\mathbf{div}\,\cdot,\mathbf{div}\,\cdot)$ in the symmetric matrix space resulting from Arnold–Winther elements. Since the form is not uniformly elliptic, a Helmholtz-type decomposition is essential. The Arnold–Winther finite element space gives rise to nonnested multilevel spaces adding difficulty to the analysis. We prove that for the variable V-cycle multigrid preconditioner, the condition number of the preconditioned system is independent of the number of levels. The results of numerical experiments are also presented.

Key words. multigrid, mixed finite element, linear elasticity

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/040617820

1. Introduction.

Mixed finite element methods [7, 16] have been widely used in solving partial differential equations. Compared to the primal-based methods, mixed finite element methods have some well-known advantages. For example, the dual variable (in this case the stress), which is often the variable of primary interest, is computed directly as a fundamental unknown. Mixed methods also have some obvious disadvantages, such as the necessity of constructing stable pairs of finite element spaces and the fact that the resulting discrete system is indefinite. The construction of stable pairs of finite element spaces and the development of efficient iterative solvers for the resulting discrete system remain two of the most important issues in the applications of mixed finite element methods.

For decades, extensive research has been carried out to explore the mixed formulation of the plane elasticity problem. Most of this research was focused on developing stable pairs of mixed finite element spaces, and several different solutions have been proposed [5, 6, 26]. As stated in those papers, the crux of the difficulty is that the stress tensor in the Hellinger–Reissner principle has to be symmetric. Indeed, this symmetry condition is so hard to satisfy that the authors of [5, 26] resort to composite elements. Only recently did Arnold and Winther construct a stable pair of mixed finite elements [6] which did not use composite elements. The Arnold–Winther finite element spaces consist of piecewise polynomials over a triangular mesh tied together by degrees of freedom resulting in $\boldsymbol{H}(\mathbf{div})$ conforming symmetric approximation subspaces.

We mention some alternative ways to circumvent the difficulty of constructing stable pairs of finite elements. One way is to reformulate the saddle-point problem by using Lagrangian functionals so that it does not require symmetric matrices [1, 4].

*Received by the editors October 28, 2004; accepted for publication (in revised form) July 18, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sinum/44-2/61782.html
†Department of Mathematics, Texas A&M University, College Station, TX 77843 (pasciak@math.tamu.edu).
‡Department of Mathematics, Purdue University, West Lafayette, IN 47907 (yqwang@math.purdue.edu).

Another way is to use the least-square formulation so that the classical discrete inf-sup condition is no longer needed [10, 17, 18]. Finally, other authors resort to the use of stabilizing techniques (see [22] and the references therein).

In this paper, we will focus on the lowest order Arnold–Winther finite element. The purpose is to develop and analyze a multigrid preconditioner for the resulting discrete system.

The discretization of the mixed formulation leads to a symmetric indefinite linear system. Generally speaking, there are three main approaches for solving large symmetric indefinite linear systems corresponding to mixed formulations. The first approach is to use Uzawa-type methods [9, 11, 20]. The second is the positive definite reformulation proposed by Bramble and Pasciak in [12] and [13]. The third is the preconditioned minimum residual method analyzed in [2, 27]. We adopt the idea of the preconditioned minimum residual method. An analysis similar to the one in [2] will show that the problem of constructing a preconditioner for the indefinite linear system derived from the mixed formulation of linear plane elasticity can be reduced to the problem of constructing a preconditioner for the $\boldsymbol{H}(\mathbf{div})$ problem on the Arnold–Winther finite element space on the symmetric matrix field.

In this paper, we construct and analyze a multigrid preconditioner for the $\boldsymbol{H}(\mathbf{div})$ problem. Multigrid methods provide efficient preconditioners for second order elliptic problems. A vast amount of research has been done in this area [15, 24, 29]. However, the classical techniques for the multigrid preconditioner do not work for the $\boldsymbol{H}(\mathbf{div})$ problem since the discrete operator which results from the $\boldsymbol{H}(\mathbf{div})$ problem is not uniformly elliptic. To deal with this difficulty, we follow the idea of using a Helmholtz-like decomposition [2, 3, 8, 21, 25] and decompose the Arnold–Winther finite element space into two orthogonal subspaces: the subspace of divergence-free functions and its orthogonal complement. Then, the analysis of our preconditioners can be done on these two subspaces separately. Our results show that for convex polygonal domains and the pure traction boundary problem, the condition number of the preconditioned system using the variable V-cycle multigrid preconditioner is independent of the number of levels.

The outline of the remainder of the paper is as follows. In section 2, we briefly introduce the mixed formulation of the elasticity problem, the Arnold–Winther mixed finite element for (2.3) and the technique for preconditioning a mixed system proposed in [2]. In section 3, the details of the multigrid preconditioner are explained, and the condition number of the preconditioned system is analyzed under certain assumptions on the smoother. In section 4, we construct a smoother and prove that it satisfies the assumptions stated in section 3. Finally, we give results of numerical experiments in section 5.

2. The mixed problem formulation, discretization, and preconditioning. In this section, we first state the mixed form of the linear elasticity problem. Next, we introduce the Arnold–Winther elements of lowest order. Finally, we briefly describe the idea of preconditioning the mixed system introduced in [2] which reduces the preconditioning problem to one on $\boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2)$ (defined below).

2.1. Mixed elasticity formulation. Let Ω be a convex polygon in $\mathbb{R}^2$. We use the usual notation $H^s(\Omega)$, where s is a real number, to denote the Sobolev space defined on Ω [19]. For $s = 0$, the space is also denoted by $L^2(\Omega)$. Define $H_0^s(\Omega)$ to be the closure of $C_0^\infty(\Omega)$ under the $H^s(\Omega)$ norm.

Let $\mathbb{R}^2$ be the space of two-dimensional vector functions and $\mathbb{S}_2$ be the space of symmetric 2×2 matrix functions defined on Ω. Throughout the paper, we adopt

the convention that bold Latin characters in lower case denote vectors and bold Greek characters denote 2×2 symmetric matrices. Let $\boldsymbol{\tau} = (\tau_{ij})_{1\le i,j\le 2} \in \mathbb{S}_2$, $\boldsymbol{v} = (v_i)_{1\le i\le 2} \in \mathbb{R}^2$, and q be a scalar function. Define $\operatorname{div} \boldsymbol{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}$ and

$$\mathbf{div}\,\boldsymbol{\tau} = \begin{pmatrix} \frac{\partial \tau_{11}}{\partial x} + \frac{\partial \tau_{12}}{\partial y} \\ \frac{\partial \tau_{21}}{\partial x} + \frac{\partial \tau_{22}}{\partial y} \end{pmatrix}, \qquad \mathbf{airy}\, q = \begin{pmatrix} \frac{\partial^2 q}{\partial y^2} & -\frac{\partial^2 q}{\partial x \partial y} \\ -\frac{\partial^2 q}{\partial x \partial y} & \frac{\partial^2 q}{\partial x^2} \end{pmatrix}. \tag{2.1}$$

Denote the inner product between vectors and the inner product between matrices by

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + u_2 v_2, \quad \text{and} \quad \boldsymbol{\sigma} : \boldsymbol{\tau} = \sum_{i,j=1}^{2} \sigma_{ij} \tau_{ij}.$$

We generalize the definition of the Sobolev space to the cases of vector functions and symmetric matrix functions. Define the spaces

$$\boldsymbol{H}^s(\Omega, \mathbb{R}^2) = (H^s(\Omega))^2, \qquad \boldsymbol{H}^s(\Omega, \mathbb{S}_2) = (H^s(\Omega))^3$$

with norms

$$\|\boldsymbol{v}\|_{\boldsymbol{H}^s(\Omega,\mathbb{R}^2)} = (\|v_1\|^2_{H^s(\Omega)} + \|v_2\|^2_{H^s(\Omega)})^{1/2},$$
$$\|\boldsymbol{\tau}\|_{\boldsymbol{H}^s(\Omega,\mathbb{S}_2)} = (\|\tau_{11}\|^2_{H^s(\Omega)} + 2\|\tau_{12}\|^2_{H^s(\Omega)} + \|\tau_{22}\|^2_{H^s(\Omega)})^{1/2}.$$

We define $\boldsymbol{L}^2(\Omega, \mathbb{R}^2)$ and $\boldsymbol{L}^2(\Omega, \mathbb{S}_2)$ in the same fashion. For simplicity, denote $\|\cdot\|_{s,\Omega}$ to be the H^s-norm over scalar, vector, or symmetric matrix fields, depending on the type of the function. We also use the notation $(\cdot, \cdot)$ for the L^2 inner product over scalar, vector, or matrix fields defined on Ω.

Define

$$\boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2) = \{\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega, \mathbb{S}_2) : \ \mathbf{div}\,\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega, \mathbb{R}^2) \text{ and } \boldsymbol{\tau}\boldsymbol{n}|_{\partial\Omega} = \mathbf{0}\},$$

where $\boldsymbol{n}$ is the outward normal vector on $\partial\Omega$. The norm on $\boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2)$ is defined to be

$$\|\boldsymbol{\tau}\|^2_{\boldsymbol{H}(\mathbf{div}\,,\Omega,\mathbb{S}_2)} = \|\boldsymbol{\tau}\|^2_{0,\Omega} + \|\mathbf{div}\,\boldsymbol{\tau}\|^2_{0,\Omega}.$$

$\boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2)$ is a Hilbert space with the inner product

$$\boldsymbol{\Lambda}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = (\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\mathbf{div}\,\boldsymbol{\sigma}, \mathbf{div}\,\boldsymbol{\tau}). \tag{2.2}$$

Next, we state the mixed formulation of the plane elasticity problem. We only consider the pure traction boundary problem [6, 16]: Find the stress $\boldsymbol{\sigma} \in \boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2)$ and the displacement $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega, \mathbb{R}^2)$ satisfying

$$\begin{cases} \int_\Omega \mathbb{A}\boldsymbol{\sigma} : \boldsymbol{\tau}\, d\boldsymbol{x} + \int_\Omega \mathbf{div}\,\boldsymbol{\tau} \cdot \boldsymbol{u}\, d\boldsymbol{x} = 0 & \text{for all } \boldsymbol{\tau} \in \boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2), \\ \int_\Omega \mathbf{div}\,\boldsymbol{\sigma} \cdot \boldsymbol{v}\, d\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \boldsymbol{v}\, d\boldsymbol{x} & \text{for all } \boldsymbol{v} \in \boldsymbol{L}^2(\Omega, \mathbb{R}^2). \end{cases} \tag{2.3}$$

Here the fourth order compliance tensor $\mathbb{A}$ is bounded, symmetric, and uniformly positive definite the body force per unit volume $\boldsymbol{g}$ is in $\boldsymbol{L}^2(\Omega, \mathbb{R}^2)$. For (2.3) to be well posed, we need a compatibility condition on $\boldsymbol{g}$. Let

$$RM := span\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -y \\ x \end{pmatrix} \right\}$$

be the space of infinitesimal rigid motions. By Korn's inequality, one can see that for any $\boldsymbol{g} \in \boldsymbol{L}^2(\Omega, \mathbb{R}^2)/RM$ (the orthogonal complement of RM in $\boldsymbol{L}^2(\Omega, \mathbb{R}^2)$), system (2.3) has a unique solution in $\boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2) \times \boldsymbol{L}^2(\Omega, \mathbb{R}^2)/RM$ [16].

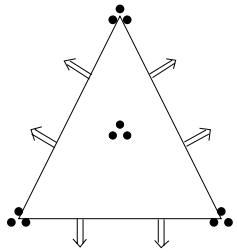


FIG. 2.1. *The Arnold–Winther finite element* $\boldsymbol{\Sigma}_T$.

2.2. Arnold–Winther elements. Let $\mathcal{T}$ be a quasi-uniform triangulation of Ω with characteristic mesh size h. On each triangle $T \in \mathcal{T}$ define

$$\begin{aligned}\boldsymbol{\Sigma}_T &= \{\text{symmetric matrices } \boldsymbol{\tau} \in (P_3(T))^3 \text{ such that } \mathbf{div}\,\boldsymbol{\tau} \in (P_1(T))^2\},\\ \boldsymbol{V}_T &= (P_1(T))^2,\end{aligned}$$

where $P_i(T)$ denotes the space consisting of polynomials of degree i or less. The degrees of freedom (dofs) for $\boldsymbol{\Sigma}_T$ are

- the nodal values of the three components of $\boldsymbol{\tau}(x)$ at each vertex of T (9 dofs);
- the moments of degree 0 and 1 of the two normal components of $\boldsymbol{\tau}$ on each edge of T (12 dofs);
- the moments of degree 0 of the three components of $\boldsymbol{\tau}$ on T (3 dofs).

The dofs of $\boldsymbol{V}_T$ are given as the zeroth and first order moments on T. Figure 2.1 illustrates the dofs for $\boldsymbol{\Sigma}_T$. The finite element spaces on the mesh $\mathcal{T}$ and domain Ω are defined as follows:

$$\begin{aligned}\boldsymbol{\Sigma}(\mathcal{T},\Omega) =&\{\boldsymbol{\tau} :\ \boldsymbol{\tau}|_T \in \boldsymbol{\Sigma}_T \text{ for each } T \in \mathcal{T},\ \boldsymbol{\tau} \text{ is continuous on the dofs}\\ &\qquad \text{on each vertex and each edge of } \mathcal{T} \text{ and } \boldsymbol{\tau}\boldsymbol{n}|_{\partial\Omega} = \mathbf{0}\},\\ \boldsymbol{V}(\mathcal{T},\Omega) =&\{\boldsymbol{v} \in \boldsymbol{L}_2(\Omega,\mathbb{R}^2) :\ \boldsymbol{v}|_T \in \boldsymbol{V}_T \text{ for each } T \in \mathcal{T}\}.\end{aligned}$$

The definition of $\boldsymbol{\Sigma}(\mathcal{T},\Omega)$ implies that $\boldsymbol{\Sigma}(\mathcal{T},\Omega) \subset \boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)$ (see [6, 16]). Note that the boundary condition $\boldsymbol{\tau}\boldsymbol{n}|_{\partial\Omega} = \mathbf{0}$ implies two linear relations among the three components of $\boldsymbol{\tau}$ on boundary nodes. Hence on the corner vertices where two boundary edges meet, we will have $\boldsymbol{\tau} = \mathbf{0}$. This fact was noticed by Arnold and Winther in [6]. Another immediate observation is that by Green's formula,

$$\mathbf{div}\,\boldsymbol{\tau} \in RM^{\perp_{\boldsymbol{V}(\mathcal{T},\Omega)}} \qquad \text{for all } \boldsymbol{\tau} \in \boldsymbol{\Sigma}(\mathcal{T},\Omega).$$

The discrete elasticity problem can be written as follows: find $\boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}(\mathcal{T},\Omega)$ and $\boldsymbol{u}_h \in \boldsymbol{V}(\mathcal{T},\Omega)$ such that

$$\begin{cases}(\mathbb{A}\boldsymbol{\sigma}_h,\boldsymbol{\tau}) + (\mathbf{div}\,\boldsymbol{\tau},\boldsymbol{u}_h) = 0 & \text{for all } \boldsymbol{\tau} \in \boldsymbol{\Sigma}(\mathcal{T},\Omega),\\ (\mathbf{div}\,\boldsymbol{\sigma}_h,\boldsymbol{v}) = (\boldsymbol{g},\boldsymbol{v}) & \text{for all } \boldsymbol{v} \in \boldsymbol{V}(\mathcal{T},\Omega).\end{cases} \tag{2.4}$$

Arnold and Winther have proved that the Arnold–Winther finite element spaces (without the essential boundary condition $\boldsymbol{\tau}\boldsymbol{n}|_{\partial\Omega} = \mathbf{0}$) satisfy the LBB condition [6]. In [28], it was proved that the Arnold–Winther finite element spaces $(\boldsymbol{\Sigma}(\mathcal{T},\Omega), \boldsymbol{V}(\mathcal{T},\Omega))$ (with the essential boundary condition $\boldsymbol{\tau}\boldsymbol{n}|_{\partial\Omega} = \mathbf{0}$) also satisfy the LBB condition.

Furthermore, the assumption on $\mathbb{A}$ implies that there exists a positive constant c such that

$$(\mathbb{A}\boldsymbol{\tau},\boldsymbol{\tau}) \geq c\|\boldsymbol{\tau}\|^2_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)} \quad \text{for all } \boldsymbol{\tau}\in \boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2) \text{ with } \mathbf{div}\,\boldsymbol{\tau}=\mathbf{0}. \tag{2.5}$$

Combining these results shows that problem (2.4) has a unique solution (for compatible $\boldsymbol{g}$) in $(\boldsymbol{\Sigma}(\mathcal{T},\Omega),\boldsymbol{V}(\mathcal{T},\Omega)/RM)$. Furthermore, if $(\boldsymbol{\sigma},\boldsymbol{u})$ is the solution of the weak problem (2.3) and $(\boldsymbol{\sigma}_h,\boldsymbol{u}_h)$ is the solution of the discrete problem (2.4), we have the following error estimates [6, 28]:

$$\begin{aligned} \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega} m &\leq ch^m\|\boldsymbol{\sigma}\|_{m,\Omega}, && 1\leq m\leq 3,\\ \|\mathbf{div}\,\boldsymbol{\sigma}-\mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega}, &\leq ch^m\|\mathbf{div}\,\boldsymbol{\sigma}\|_{m,\Omega}, && 0\leq m\leq 2,\\ \|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM}, &\leq ch^m\|\boldsymbol{u}\|_{m+1,\Omega}, && 1\leq m\leq 2, \end{aligned} \tag{2.6}$$

where c is a positive constant independent of h.

Next, we introduce the Argyris element, which plays an important role in later analysis. Let Q_T denote the Argyris element [19] defined on T. It is a quintic element and the dofs are

- the function value on each vertex (three dofs), the first derivatives at each vertex (six dofs), and the second derivatives at each vertex (nine dofs);
- the moments of degree 0 of the normal derivative on each edges of T (three dofs).

Define the space

$$\begin{aligned} \mathrm{Q}(\mathcal{T},\Omega) = \{q:\ & q|_T\in \mathrm{Q}_T \text{ for each } T\in\mathcal{T},\ q \text{ is continuous on the degrees of}\\ & \text{freedom on each vertex and each edge of } \mathcal{T} \text{ and } q|_{\partial\Omega}=0,\ \nabla q|_{\partial\Omega}=\mathbf{0}\}. \end{aligned}$$

Clearly $\mathrm{Q}(\mathcal{T},\Omega)\subset H^2_0(\Omega)$.

Similar to the De Rham sequence, it is elementary to see that the following exact sequence holds [6, 28]:

$$0 \xrightarrow{\subset} H^2_0(\Omega) \xrightarrow{\mathbf{airy}} \boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2) \xrightarrow{\mathbf{div}} \boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM \to 0.$$

Recall that operators in an exact sequence have the property that the range of the operator on the left equals the kernel of the operator on the right.

We can define an operator $\mathbf{div}^{-1}:\boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM\to \boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)/Ker(\mathbf{div})$ as follows. For $\boldsymbol{v}\in\boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM$, let $\boldsymbol{\sigma}\in\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)$ and $\boldsymbol{u}\in\boldsymbol{L}^2(\Omega,\mathbb{R}^2)$ satisfy

$$\begin{cases} (\boldsymbol{\sigma},\boldsymbol{\tau})+(\mathbf{div}\,\boldsymbol{\tau},\boldsymbol{u})=0 & \text{for all } \boldsymbol{\tau}\in\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2),\\ (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{w})=(\boldsymbol{v},\boldsymbol{w}) & \text{for all } \boldsymbol{w}\in\boldsymbol{L}^2(\Omega,\mathbb{R}^2). \end{cases} \tag{2.7}$$

Since $\mathbf{div}$ maps $\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)$ onto $\boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM$, system (2.7) admits a unique solution in $(\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2),\boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM)$ (see [16]). Then, set $\mathbf{div}^{-1}\boldsymbol{v}=\boldsymbol{\sigma}$. By definition, $\mathbf{div}^{-1}\boldsymbol{v}$ is orthogonal to any divergence free function in $\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)$ under both the $\boldsymbol{L}^2$ inner product and the $\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)$ inner product. Therefore, for all $\boldsymbol{\tau}\in\boldsymbol{H}_0(\mathbf{div},\Omega,\mathbb{S}_2)$, we have a unique orthogonal decomposition

$$\boldsymbol{\tau}=\mathbf{airy}\,q+\mathbf{div}^{-1}\boldsymbol{v},$$

where $q\in H^2_0(\Omega)$ and $\boldsymbol{v}=\mathbf{div}\,\boldsymbol{\tau}$. Furthermore, we have the regularity result (see [23]),

$$\mathbf{div}^{-1}\boldsymbol{v}\in\boldsymbol{H}^1(\Omega,\mathbb{S}_2) \quad\text{and}\quad \|\mathbf{div}^{-1}\boldsymbol{v}\|_{1,\Omega}\leq c\|\boldsymbol{v}\|_{0,\Omega}, \tag{2.8}$$

where c is a positive constant independent of $\boldsymbol{v}$.

Analogously, on the discrete level we have the following exact sequence:

$$0 \xrightarrow{\subset} \mathrm{Q}(\mathcal{T},\Omega) \xrightarrow{\mathbf{airy}} \boldsymbol{\Sigma}(\mathcal{T},\Omega) \xrightarrow{\mathbf{div}} \boldsymbol{V}(\mathcal{T},\Omega)/RM \to 0. \tag{2.9}$$

The exactness of this sequence for the Arnold–Winther finite element spaces follows from [6]. We define an operator $\mathbf{div}_{\mathcal{T}}^{-1} : \boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM \to \boldsymbol{\Sigma}(\mathcal{T},\Omega)/Ker(\mathbf{div})$ as follows. For $\boldsymbol{v} \in \boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM$, let $\boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}(\mathcal{T},\Omega)$ and $\boldsymbol{u}_h \in \boldsymbol{V}(\mathcal{T},\Omega)$ satisfy

$$\begin{cases} (\boldsymbol{\sigma}_h,\boldsymbol{\tau}) + (\mathbf{div}\,\boldsymbol{\tau},\boldsymbol{u}_h) = 0 & \text{for all } \boldsymbol{\tau} \in \boldsymbol{\Sigma}(\mathcal{T},\Omega), \\ (\mathbf{div}\,\boldsymbol{\sigma}_h,\boldsymbol{w}) = (\boldsymbol{v},\boldsymbol{w}) & \text{for all } \boldsymbol{w} \in \boldsymbol{V}(\mathcal{T},\Omega). \end{cases} \tag{2.10}$$

Since the Arnold–Winther finite element spaces satisfy the LBB condition, the solution to (2.10) exists and is unique in $(\boldsymbol{\Sigma}(\mathcal{T},\Omega), \boldsymbol{V}(\mathcal{T},\Omega)/RM)$. Define $\mathbf{div}_{\mathcal{T}}^{-1}\boldsymbol{v} = \boldsymbol{\sigma}_h$. Then, for all $\boldsymbol{\tau} \in \boldsymbol{\Sigma}(\mathcal{T},\Omega)$, there exists a unique discrete orthogonal decomposition

$$\boldsymbol{\tau} = \mathbf{airy}\, q + \mathbf{div}_{\mathcal{T}}^{-1}\boldsymbol{v},$$

where $q \in \mathrm{Q}(\mathcal{T},\Omega)$ and $\boldsymbol{v} = \mathbf{div}\,\boldsymbol{\tau}$.

By the approximation property (2.6) of the Arnold–Winther element and the regularity result (2.8), for all $\boldsymbol{v} \in \boldsymbol{L}^2(\Omega,\mathbb{R}^2)/RM$,

$$\|\mathbf{div}^{-1}\boldsymbol{v} - \mathbf{div}_{\mathcal{T}}^{-1}\boldsymbol{v}\|_{0,\Omega} \le ch\|\mathbf{div}^{-1}\boldsymbol{v}\|_{1,\Omega} \le ch\|\boldsymbol{v}\|_{0,\Omega}, \tag{2.11}$$

where c is a positive constant independent of $\boldsymbol{v}$.

2.3. A block diagonal preconditioner for the mixed system. For simplicity, let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\mathcal{T},\Omega)$ and $\boldsymbol{V} = \boldsymbol{V}(\mathcal{T},\Omega)/RM$. Let $\|\cdot\|_{\boldsymbol{\Sigma}}$ and $\|\cdot\|_{\boldsymbol{V}}$ be the norms on $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$, respectively, i.e., $\|\cdot\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}$ and $\|\cdot\|_{\boldsymbol{L}^2(\Omega,\mathbb{R}^2)}$. Let $\boldsymbol{\Sigma}^*$ and $\boldsymbol{V}^*$ be the dual spaces of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ with dual norms $\|\cdot\|_{\boldsymbol{\Sigma}^*}$ and $\|\cdot\|_{\boldsymbol{V}^*}$ and $<\cdot,\cdot>$ denote the duality pairing. Define the operators

$$\begin{cases} \mathcal{A} :\ \boldsymbol{\Sigma} \to \boldsymbol{\Sigma}^*, & <\mathcal{A}\boldsymbol{\sigma},\boldsymbol{\tau}> = (\mathbb{A}\boldsymbol{\sigma},\boldsymbol{\tau}) \quad \text{for all } \boldsymbol{\tau} \in \boldsymbol{\Sigma}, \\ \mathcal{B} :\ \boldsymbol{\Sigma} \to \boldsymbol{V}^*, & <\mathcal{B}\boldsymbol{\sigma},\boldsymbol{v}> = (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}. \end{cases}$$

Let $\mathcal{B}^t :\ \boldsymbol{V} \to \boldsymbol{\Sigma}^*$ be the adjoint of $\mathcal{B}$. Equation (2.4) can be rewritten as

$$\mathcal{M}\begin{pmatrix}\boldsymbol{\sigma}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\mathcal{A} & \mathcal{B}^t\\ \mathcal{B} & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\sigma}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}F\\ G\end{pmatrix}, \tag{2.12}$$

where $F \in \boldsymbol{\Sigma}^*$, $G \in \boldsymbol{V}^*$. The following lemma results from the LBB condition and (2.5). (See [16] for the proof.)

LEMMA 2.1. *The map $(F,G) \to (\boldsymbol{\sigma},\boldsymbol{u})$ defined by solving (2.12) with $F \in \boldsymbol{\Sigma}^*$ and $G \in \boldsymbol{V}^*$ is an isomorphism of $\boldsymbol{\Sigma}^* \times \boldsymbol{V}^*$ onto $\boldsymbol{\Sigma} \times \boldsymbol{V}$ and so*

$$c_0(\|F\|_{\boldsymbol{\Sigma}^*} + \|G\|_{\boldsymbol{V}^*}) \le \|\boldsymbol{\sigma}\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{u}\|_{\boldsymbol{V}} \le c_1(\|F\|_{\boldsymbol{\Sigma}^*} + \|G\|_{\boldsymbol{V}^*}),$$

where c_0 and c_1 are positive and independent of h.

Our purpose is to find a preconditioner for the operator $\mathcal{M}$. By Lemma 2.1, we only need to find an operator $\mathcal{S} :\ \boldsymbol{\Sigma}^* \times \boldsymbol{V}^* \to \boldsymbol{\Sigma} \times \boldsymbol{V}$ such that $\|\mathcal{S}\|_{\mathcal{L}(\boldsymbol{\Sigma}^*\times\boldsymbol{V}^*,\boldsymbol{\Sigma}\times\boldsymbol{V})}$ and $\|\mathcal{S}^{-1}\|_{\mathcal{L}(\boldsymbol{\Sigma}\times\boldsymbol{V},\boldsymbol{\Sigma}^*\times\boldsymbol{V}^*)}$ are bounded uniformly in h (see [2] for details). Indeed, we can consider an operator in the form $\mathcal{S} = \begin{pmatrix}\mathcal{S}_1 & 0\\ 0 & \mathcal{S}_2\end{pmatrix}$, where $\mathcal{S}_1 :\ \boldsymbol{\Sigma}^* \to \boldsymbol{\Sigma}$ and $\mathcal{S}_2 :\ \boldsymbol{V}^* \to \boldsymbol{V}$

and their inverses are bounded uniformly in h. Consider the following problem: find $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$ such that

$$\boldsymbol{\Lambda}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = F(\boldsymbol{\tau}) \qquad \text{for all } \ \boldsymbol{\tau} \in \boldsymbol{\Sigma}. \tag{2.13}$$

Clearly a good preconditioner for this problem will yield an ideal $\mathcal{S}_1$. Similarly, an ideal $\mathcal{S}_2$ will come from a good preconditioner for the following problem: find $\boldsymbol{u} \in \boldsymbol{V}$ such that

$$(\boldsymbol{u}, \boldsymbol{v}) = G(\boldsymbol{v}) \qquad \text{for all } \ \boldsymbol{v} \in \boldsymbol{V}. \tag{2.14}$$

The problem (2.14) is easy to solve efficiently. Indeed, we use the basis for $\boldsymbol{V}(\mathcal{T}, \Omega)$ in the implementation. (This, of course, provides a spanning set for $\boldsymbol{V}$.) First, we note that the functional G in original problem (2.12) is usually available as a functional $\widetilde{G}$ defined on $(\boldsymbol{V}(\mathcal{T}, \Omega))^*$ which vanishes on RM. This functional is naturally represented by its action on the basis functions for $\boldsymbol{V}(\mathcal{T}, \Omega)$ and provides the data for the first solve of (2.14). Subsequent solves of (2.14) involve this data plus the result of $\mathcal{B}$ applied to something in $\boldsymbol{\Sigma}$. Thus, at any step of the iteration, (2.14) will have to be solved with a known functional $\widetilde{G}$ on $(\boldsymbol{V}(\mathcal{T}, \Omega))^*$ which vanishes on RM. In this case, the solution of (2.14) coincides with the solution $\boldsymbol{u} \in \boldsymbol{V}(\mathcal{T}, \Omega)$ satisfying

$$(\boldsymbol{u}, \boldsymbol{v}) = \widetilde{G}(\boldsymbol{v}) \qquad \text{for all } \ \boldsymbol{v} \in \boldsymbol{V}(\mathcal{T}, \Omega). \tag{2.15}$$

The space $\boldsymbol{V}(\mathcal{T}, \Omega)$ consists of discontinuous linears on the triangles so the exact solution of (2.15) reduces to the inversion of a block diagonal matrix, with 3×3 diagonal blocks. Hence the problem of defining $\mathcal{S}$ reduces to the problem of constructing $\mathcal{S}_1$. In the remainder of this paper we will focus on constructing a multigrid preconditioner for problem (2.13).

3. The multigrid preconditioner. In this section, we construct and analyze a multigrid preconditioner for problem (2.13). To this end, let $\mathcal{T}_1$ be a unit-sized coarse triangulation of Ω. Subsequently finer triangulations are defined recursively. Given the kth level triangulation $\mathcal{T}_k$, define the $(k+1)$st level mesh $\mathcal{T}_{k+1}$ by breaking each triangle in $\mathcal{T}_k$ into four triangles by connecting the midpoints of the edges. Repeating this process gives a series of nested meshes $\mathcal{T}_1, \mathcal{T}_1, \ldots, \mathcal{T}_K$. Denote the characteristic mesh size of $\mathcal{T}_k$ as h_k. We clearly have $h_k = \frac{1}{2} h_{k-1} = O(2^{-k})$. For simplicity of notation, in the rest of this paper, we use $\lesssim$ to denote "less then or equal to" with a factor c independent of k or h_k.

Denote the finite element spaces on the kth level by

$$\mathrm{Q}_k = \mathrm{Q}(\mathcal{T}_k, \Omega), \qquad \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}(\mathcal{T}_k, \Omega), \qquad \boldsymbol{V}_k = \boldsymbol{V}(\mathcal{T}_k, \Omega)/RM.$$

Notice that we have $\mathrm{Q}_k \subset H_0^2(\Omega)$, $\boldsymbol{\Sigma}_k \subset \boldsymbol{H}_0(\mathbf{div}\,, \Omega, \mathbb{S}_2)$, and $\boldsymbol{V}_k \subset \boldsymbol{L}^2(\Omega, \mathbb{R}^2)$ for each k.

The bilinear form for the biharmonic problem will play an important role in the following analysis. It is defined on $H_0^2(\Omega)$ by

$$\begin{aligned} \mathrm{A}(q, p) &= \int_\Omega \left(\frac{\partial^2 q}{\partial x^2} \frac{\partial^2 p}{\partial x^2} + 2 \frac{\partial^2 q}{\partial x \partial y} \frac{\partial^2 p}{\partial x \partial y} + \frac{\partial^2 q}{\partial y^2} \frac{\partial^2 p}{\partial y^2} \right) d\boldsymbol{x} \\ &= (\mathbf{airy}\, q, \mathbf{airy}\, p). \end{aligned} \tag{3.1}$$

Define operators $\mathrm{A}_k : \mathrm{Q}_k \to \mathrm{Q}_k$ and $\boldsymbol{\Lambda}_k : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_k$ by

$$\begin{aligned}(\mathrm{A}_k q, p) &= \mathrm{A}(q,p) \qquad \text{for all } q, p \in \mathrm{Q}_k,\\ (\boldsymbol{\Lambda}_k \boldsymbol{\sigma}, \boldsymbol{\tau}) &= \boldsymbol{\Lambda}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \qquad \text{for all } \boldsymbol{\sigma}, \boldsymbol{\tau} \in \boldsymbol{\Sigma}_k,\end{aligned}$$

where the bilinear forms $\mathrm{A}(\cdot,\cdot)$ and $\boldsymbol{\Lambda}(\cdot,\cdot)$ were defined in (3.1) and (2.2), respectively.

The spaces $\{\mathrm{Q}_k\}$ and $\{\boldsymbol{\Sigma}_k\}$ are nonnested since, for example, a function $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_k$ is not necessarily continuous at the midpoints of the edges in the mesh $\mathcal{T}_k$ and a function $q \in \mathrm{Q}_k$ does not necessarily have continuous second order derivatives at the midpoints of the edges in the mesh $\mathcal{T}_k$. Hence we need to define interpolation operators $\mathcal{I}_k : \mathrm{Q}_{k-1} \to \mathrm{Q}_k$ and $\mathbf{I}_k : \boldsymbol{\Sigma}_{k-1} \to \boldsymbol{\Sigma}_k$. The easiest way to do this is by using the "local" nodal value interpolation on each triangle and then taking average on the discontinuous degrees of freedom at vertices.

Denote $\mathcal{N}_k$ to be the set of all nodes in the mesh $\mathcal{T}_k$. For any vertex $v \in \mathcal{N}_k$, let $S_{k-1}(v)$ be the set of all triangles in $\mathcal{T}_{k-1}$ which contain the vertex v and let $|S_{k-1}(v)|$ denote the number of triangles in $S_{k-1}(v)$. For $q \in \mathrm{Q}_{k-1}$ and $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_{k-1}$, define the dofs for $\mathcal{I}_k q$ and $\mathbf{I}_k \boldsymbol{\tau}$ to be identical to those for q and $\boldsymbol{\tau}$ for all dofs excluding the second order derivatives at the vertices for $\mathcal{I}_k q$ and the nodal values at the vertices for $\mathbf{I}_k \boldsymbol{\tau}$. On the excluded dofs we use

$$\begin{aligned}\mathbf{airy}\,(\mathcal{I}_k q)(v) &= \frac{1}{|S_{k-1}(v)|} \sum_{T_v \in S_{k-1}(v)} \mathbf{airy}\, q(v)|_{T_v} \quad \text{for } v \in \mathcal{N}_k,\\ \mathbf{I}_k \boldsymbol{\tau}(v) &= \frac{1}{|S_{k-1}(v)|} \sum_{T_v \in S_{k-1}(v)} \boldsymbol{\tau}(v)|_{T_v} \qquad \text{for } v \in \mathcal{N}_k.\end{aligned}$$

Combining the above gives the definition of $\mathcal{I}_k q$ and $\mathbf{I}_k \boldsymbol{\tau}$ on all dofs. We then have

$$\begin{aligned}\mathcal{I}_k q &= q + \tilde{q} \qquad \text{for all } q \in \mathrm{Q}_{k-1},\\ \mathbf{I}_k \boldsymbol{\tau} &= \boldsymbol{\tau} + \tilde{\boldsymbol{\tau}} \qquad \text{for all } \boldsymbol{\tau} \in \boldsymbol{\Sigma}_{k-1},\end{aligned}$$

where $\tilde{q} \in H_0^2(\Omega)$ and $\tilde{\boldsymbol{\tau}} \in \boldsymbol{H}_0(\mathbf{div}, \Omega, \mathbb{S}_2)$ satisfy

$$\begin{aligned}\mathbf{airy}\,\tilde{q}(v)|_T &= \left(\frac{1}{|S_{k-1}(v)|} \sum_{T_v \in S_{k-1}(v)} \mathbf{airy}\, q(v)|_{T_v}\right) - \mathbf{airy}\, q(v)|_T,\\ \tilde{\boldsymbol{\tau}}(v)|_T &= \left(\frac{1}{|S_{k-1}(v)|} \sum_{T_v \in S_{k-1}(v)} \boldsymbol{\tau}(v)|_{T_v}\right) - \boldsymbol{\tau}(v)|_T\end{aligned} \tag{3.2}$$

at each vertex v of any triangle $T \in \mathcal{T}_k$ and vanish at all the other dofs. Define $\mathcal{P}_{k-1} : \mathrm{Q}_k \to \mathrm{Q}_{k-1}$ to be the A-adjoint of $\mathcal{I}_k$ and $\mathbf{P}_{k-1} : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_{k-1}$ to be the $\boldsymbol{\Lambda}$-adjoint of $\mathbf{I}_k$.

LEMMA 3.1. *We have*

$$\boldsymbol{\Lambda}(\mathbf{I}_k \boldsymbol{\sigma}_{k-1}, \mathbf{I}_k \boldsymbol{\sigma}_{k-1}) \le \omega \boldsymbol{\Lambda}(\boldsymbol{\sigma}_{k-1}, \boldsymbol{\sigma}_{k-1}) \quad \textit{for all } \boldsymbol{\sigma}_{k-1} \in \boldsymbol{\Sigma}_{k-1},$$

where ω is independent of k. Consequently,

$$\boldsymbol{\Lambda}(\mathbf{P}_{k-1} \boldsymbol{\sigma}_k, \mathbf{P}_{k-1} \boldsymbol{\sigma}_k) \le \omega \boldsymbol{\Lambda}(\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k) \quad \textit{for all } \boldsymbol{\sigma}_k \in \boldsymbol{\Sigma}_k.$$

Proof. The proof follows from a standard scaling argument, the definition of $\mathbf{I}_k$, and the quasi-uniformity of the mesh. ∎

We have the following two lemmas concerning the interpolation operators $\mathcal{I}_k$ and $\mathbf{I}_k$ from [28].

LEMMA 3.2. *Let T be a triangle and v_i, $i = 1, 2, 3$, be its vertices. Let $\boldsymbol{\tau}_i$, $i = 1, 2, 3$, be given constant symmetric matrices. Define $q \in \mathrm{Q}_T$ and $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T$ such that*

$$\begin{aligned} \mathbf{airy}\, q(v_i) &= \boldsymbol{\tau}_i \qquad \text{for } i = 1, 2, 3, \\ \boldsymbol{\tau}(v_i) &= \boldsymbol{\tau}_i \qquad \text{for } i = 1, 2, 3, \end{aligned}$$

while vanishing on all the other dofs. Then $\mathbf{airy}\, q = \boldsymbol{\tau}$.

LEMMA 3.3. *The following commutative diagram of exact sequences holds:*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathrm{Q}_{k-1} & \xrightarrow{\mathbf{airy}} & \boldsymbol{\Sigma}_{k-1} & \xrightarrow{\mathbf{div}} & \boldsymbol{V}_{k-1}/RM & \to & 0 \\ & & \downarrow \mathcal{I}_k & & \downarrow \mathbf{I}_k & & \downarrow id & & \\ 0 & \longrightarrow & \mathrm{Q}_k & \xrightarrow{\mathbf{airy}} & \boldsymbol{\Sigma}_k & \xrightarrow{\mathbf{div}} & \boldsymbol{V}_k/RM & \to & 0. \end{array} \tag{3.3}$$

It is not our goal to study the general approximation properties of the interpolation operator $\mathbf{I}_k$. Instead, for the multigrid analysis, we require the specific results obtained in the following two lemmas.

LEMMA 3.4. *Let $\boldsymbol{\tau}_{k-1}$ be piecewise linear with respect to $\mathcal{T}_{k-1}$ on all components. Then,*

$$((\mathbf{I} - \mathbf{I}_k)\boldsymbol{\sigma}_{k-1}, \boldsymbol{\tau}_{k-1}) = 0 \quad \text{for all } \boldsymbol{\sigma}_{k-1} \in \boldsymbol{\Sigma}_{k-1}.$$

Proof. Let $T \in \mathcal{T}_{k-1}$ and v_i, $i = 1, 2, 3$, be the three midpoints of each edge of T. We note that $(\mathbf{I} - \mathbf{I}_k)\boldsymbol{\sigma}_{k-1}$ restricted to T is in $\boldsymbol{\Sigma}(\mathcal{T}_k, T)$ and has nonzero dofs only on the nodal values at v_i, $i = 1, 2, 3$. On each of the four finer triangles T_i, $i = 1, \ldots, 4$, making up T, we have

$$(\mathbf{I} - \mathbf{I}_k)\boldsymbol{\sigma}_{k-1} = \mathbf{airy}\, q_i$$

for q_i as defined in Lemma 3.2. By construction, these q_i share the same nodal values at v_j, $j = 1, 2, 3$, and thus the function q whose restriction is q_i on T_i is in $Q(\mathcal{T}_k, T) \subset C^1(T)$. Now, since $\boldsymbol{\sigma}_{k-1}$ has continuous normal components, we have

$$(\mathbf{I} - \mathbf{I}_k)\boldsymbol{\sigma}_{k-1}\, \boldsymbol{n}|_{\partial T} = \mathbf{airy}\, q\, \boldsymbol{n}|_{\partial T} = \mathbf{0},$$

i.e., $\frac{\partial^2 q}{\partial \boldsymbol{n} \partial \boldsymbol{s}} = \frac{\partial^2 q}{\partial \boldsymbol{s}^2} = 0$, where $\boldsymbol{n}$ is the outward normal vector and $\boldsymbol{s}$ is the normal tangential vector of ∂T. It follows that both q and ∇q vanish on ∂T and are continuous across ∂T_i. Thus, integration by parts gives that for any linear function f on T,

$$((\mathbf{I} - \mathbf{I}_k)\boldsymbol{\sigma}_{k-1}, f)_{L^2(T)} = (\mathbf{airy}\, q, f)_{L^2(T)} = 0.$$

This completes the proof of the lemma. ∎

LEMMA 3.5. *There exists a positive constant c such that for all $\boldsymbol{v}_{k-1} \in \boldsymbol{V}_{k-1}$,*

$$\|(\mathbf{I} - \mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}\|_{0,\Omega} \le c h_k \|\boldsymbol{v}_{k-1}\|_{0,\Omega}.$$

Here $\mathbf{div}_k^{-1} = \mathbf{div}_{\mathcal{T}_k}^{-1}$ as defined by (2.10).

Proof. Notice that $(\mathbf{I} - \mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}$ is divergence free by Lemma 3.2. Therefore

$$((\mathbf{I} - \mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}, \mathbf{div}^{-1}\boldsymbol{v}_{k-1}) = 0.$$

According to Lemma 3.4, for any $\boldsymbol{\tau}_{k-1} \in \boldsymbol{\Sigma}_{k-1}$ which is continuous and piecewise linear with respect to $\mathcal{T}_{k-1}$,

$$((\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}, \boldsymbol{\tau}_{k-1}) = 0.$$

Let $\boldsymbol{\tau}_{k-1}$ be the L^2 projection of $\mathbf{div}^{-1}\boldsymbol{v}_{k-1}$ into the space of continuous piecewise linear functions based on $\mathcal{T}_{k-1}$. Notice that $\mathbf{I}_k\boldsymbol{\tau}_{k-1} = \boldsymbol{\tau}_{k-1}$. By the regularity result (2.8) and the approximation result (2.11),

$$\begin{aligned}\|(\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}\|_{0,\Omega}^2 &= ((\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}, \mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1} - \mathbf{div}^{-1}\boldsymbol{v}_{k-1}) \\ &\quad - ((\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}, \mathbf{I}_k(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1} - \boldsymbol{\tau}_{k-1})) \\ &\lesssim \|(\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}\|_{0,\Omega}(h_k\|\boldsymbol{v}_{k-1}\|_{0,\Omega} + \|\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1} - \boldsymbol{\tau}_{k-1}\|_{0,\Omega}).\end{aligned}$$

Thus,

$$\begin{aligned}\|(\mathbf{I}-\mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1}\|_{0,\Omega} &\lesssim h_k\|\boldsymbol{v}_{k-1}\|_{0,\Omega} + \|\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_{k-1} - \mathbf{div}^{-1}\boldsymbol{v}_{k-1}\|_{0,\Omega} \\ &\quad + \|\mathbf{div}^{-1}\boldsymbol{v}_{k-1} - \boldsymbol{\tau}_{k-1}\|_{0,\Omega} \\ &\lesssim h_k\|\boldsymbol{v}_{k-1}\|_{0,\Omega}.\end{aligned}$$

This completes the proof of the lemma. ∎

Now we state the variable V-cycle multigrid preconditioner. Let $\mathbf{R}_k : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_k$ be a symmetric and positive definite linear operator which we call a smoother. A construction for $\mathbf{R}_k$ will be given in the next section. Let m_k, the number of smoothing steps on the kth level, satisfy

$$\beta_0 m_k \le m_{k-1} \le \beta_1 m_k, \qquad \text{where } 1 < \beta_0 \le \beta_1.$$

The choice of $\beta_0 = \beta_1 = 2$ is typical. Denote $\mathbf{I}_k^t : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_{k-1}$ to be the $\boldsymbol{L}^2$-adjoint of $\mathbf{I}_k$, i.e.,

$$(\mathbf{I}_k^t\boldsymbol{\sigma}_k, \boldsymbol{\tau}_{k-1}) = (\boldsymbol{\sigma}_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1}) \quad \text{for all } \boldsymbol{\tau}_{k-1} \in \boldsymbol{\Sigma}_{k-1}.$$

The variable V-cycle multigrid preconditioner $\mathbf{B}_k : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_k$ is defined inductively as follows.

ALGORITHM 1. *Set* $\mathbf{B}_1 = \boldsymbol{\Lambda}_1^{-1}$. *Assuming that* $\mathbf{B}_{k-1} : \boldsymbol{\Sigma}_{k-1} \to \boldsymbol{\Sigma}_{k-1}$ *has been defined, define* $\mathbf{B}_k : \boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}_k$ *as follows. For* $\boldsymbol{g} \in \boldsymbol{\Sigma}_k$, *set* $\boldsymbol{\tau}^0 = \mathbf{0}$ *and define*

(1) $\boldsymbol{\tau}^l = \boldsymbol{\tau}^{l-1} + \mathbf{R}_k(\boldsymbol{g} - \boldsymbol{\Lambda}_k\boldsymbol{\tau}^{k-1})$ *for* $l = 1, \ldots, m_k$;
(2) $\boldsymbol{\sigma}^{m_k} = \boldsymbol{\tau}^{m_k} + \mathbf{I}_k\mathbf{B}_{k-1}\mathbf{I}_k^t(\boldsymbol{g} - \boldsymbol{\Lambda}_k\boldsymbol{\tau}^{m_k})$;
(3) $\boldsymbol{\sigma}^l = \boldsymbol{\sigma}^{l-1} + \mathbf{R}_k(\boldsymbol{g} - \boldsymbol{\Lambda}_k\boldsymbol{\sigma}^{l-1})$ *for* $l = m_k + 1, \ldots, 2m_k$;
Set $\mathbf{B}_k\boldsymbol{g} = \boldsymbol{\sigma}^{2m_k}$.

Remark 1. It appears that one needs to solve linear systems involving the mass matrix for the computation of $\mathbf{I}_k^t$ and $\boldsymbol{\Lambda}_k$ in the above algorithm. This Gram matrix inversion is avoided in the implementation because of the judicious choice of $\mathbf{R}_k$. For these and other implementation issues, see [15, 28].

The following theorem and its proof is a straightforward variation of Theorem 7.4 in [15].

THEOREM 3.6. *Assume that*

(M.1) *the spectrum of* $\mathbf{I} - \mathbf{R}_k\boldsymbol{\Lambda}_k$ *lies inside the interval* $[0, 1)$;
(M.2) *there exist a constant* $0 < \alpha \le 1$ *and a constant* C_p *independent of* k *such that for all* $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_k$,

$$|\boldsymbol{\Lambda}((\mathbf{I} - \mathbf{I}_k\mathbf{P}_{k-1})\boldsymbol{\tau}, \boldsymbol{\tau})| \le C_p^{2\alpha}(\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\tau}, \boldsymbol{\Lambda}_k\boldsymbol{\tau})^{\alpha}\boldsymbol{\Lambda}(\boldsymbol{\tau}, \boldsymbol{\tau})^{1-\alpha}.$$

Then, the preconditioner $\mathbf{B}_k$ *is symmetric and positive definite. Furthermore,* $\mathbf{B}_k$ *satisfies*

$$\left(\frac{m_k^\alpha}{M+m_k^\alpha}\right)\mathbf{\Lambda}(\boldsymbol{\tau},\boldsymbol{\tau}) \le \mathbf{\Lambda}(\mathbf{B}_k\mathbf{\Lambda}_k\boldsymbol{\tau},\boldsymbol{\tau}) \le \left(\frac{M+m_k^\alpha}{m_k^\alpha}\right)\mathbf{\Lambda}(\boldsymbol{\tau},\boldsymbol{\tau}) \quad \textit{for all } \boldsymbol{\tau}\in\mathbf{\Sigma}_k,$$

where M *is a sufficiently large positive constant depending only on* C_p *and* α.

In the next section, we will construct an additive smoother and prove it satisfies assumptions (M.1) and (M.2).

4. An additive Schwarz smoother. Recall that $\mathcal{N}_k$ denotes the set of all nodes in the triangulation $\mathcal{T}_k$ (including the boundary nodes) and $S_k(v)$ denotes the set of triangles in $\mathcal{T}_k$ meeting at the vertex v for each $v\in\mathcal{N}_k$. The (interior of the) union of all triangles in $S_k(v)$ forms a subdomain which we denote $\Omega_{k,v}$. Clearly $\{\Omega_{k,v}\}_{v\in\mathcal{N}_k}$ is an overlapping decomposition of Ω such that each $x\in\Omega$ is in at most three subdomains in $\{\Omega_{k,v}\}_{v\in\mathcal{N}_k}$.

Let $\mathrm{Q}_{k,v}$ and $\mathbf{\Sigma}_{k,v}$ be the subspace of functions in Q_k and $\mathbf{\Sigma}_k$, respectively, which have support contained in $\overline{\Omega}_{k,v}$. It is easy to see that the span of $\{\mathrm{Q}_{k,v}\}$ (respectively, $\mathbf{\Sigma}_{k,v}$) is all of Q_k (respectively, $\mathbf{\Sigma}_k$). Let $\mathcal{P}_{k,v}:\mathrm{Q}_k\to\mathrm{Q}_{k,v}$ be the A-projection, $\mathbf{P}_{k,v}:\mathbf{\Sigma}_k\to\mathbf{\Sigma}_{k,v}$ be the $\mathbf{\Lambda}$-projection, and $\mathcal{I}_{k,v}^t:\mathrm{Q}_k\to\mathrm{Q}_{k,v}$, $\mathbf{I}_{k,v}^t:\mathbf{\Sigma}_k\to\mathbf{\Sigma}_{k,v}$ be the $\boldsymbol{L}^2$-projections. Define $\mathrm{A}_{k,v}:\mathrm{Q}_{k,v}\to\mathrm{Q}_{k,v}$ and $\mathbf{\Lambda}_{k,v}:\mathbf{\Sigma}_{k,v}\to\mathbf{\Sigma}_{k,v}$ by

$$\begin{aligned}(\mathrm{A}_{k,v}p,q) &= \mathrm{A}(p,q) \qquad &&\text{for all } p,q\in\mathrm{Q}_{k,v},\\ (\mathbf{\Lambda}_{k,v}\boldsymbol{\sigma},\boldsymbol{\tau}) &= \mathbf{\Lambda}(\boldsymbol{\sigma},\boldsymbol{\tau}) &&\text{for all } \boldsymbol{\sigma},\boldsymbol{\tau}\in\mathbf{\Sigma}_{k,v}.\end{aligned}$$

Clearly, we have $\mathrm{A}_{k,v}\mathcal{P}_{k,v}=\mathcal{I}_{k,v}^t\mathrm{A}_k$ and $\mathbf{\Lambda}_{k,v}\mathbf{P}_{k,v}=\mathbf{I}_{k,v}^t\mathbf{\Lambda}_k$. Define

$$\begin{aligned}\mathcal{R}_k &= \rho\sum_{v\in\mathcal{N}_k}\mathcal{P}_{k,v}\mathrm{A}_k^{-1}=\rho\sum_{v\in\mathcal{N}_k}\mathrm{A}_{k,v}^{-1}\mathcal{I}_{k,v}^t,\\ \mathbf{R}_k &= \rho\sum_{v\in\mathcal{N}_k}\mathbf{P}_{k,v}\mathbf{\Lambda}_k^{-1}=\rho\sum_{v\in\mathcal{N}_k}\mathbf{\Lambda}_{k,v}^{-1}\mathbf{I}_{k,v}^t,\end{aligned}\tag{4.1}$$

where $\rho>0$ is a scaling factor which will only depend on the finite overlapping constant, e.g., $\rho=1/3$. It is well known (see [30]) that since $\{\mathbf{\Sigma}_{k,v}\}$ spans $\mathbf{\Sigma}_k$, $\mathbf{R}_k$ is invertible and satisfies

$$(\mathbf{R}_k^{-1}\boldsymbol{\tau},\boldsymbol{\tau}) = \rho^{-1}\inf_{\substack{\boldsymbol{\tau}_v\in\mathbf{\Sigma}_{k,v}\\ \sum_v\boldsymbol{\tau}_v=\boldsymbol{\tau}}}\sum_{v\in\mathcal{N}_k}\mathbf{\Lambda}(\boldsymbol{\tau}_v,\boldsymbol{\tau}_v) \qquad \text{for all } \boldsymbol{\tau}\in\mathbf{\Sigma}_k.\tag{4.2}$$

Also, we note that $\mathcal{R}_k$ is defined purely for theoretical analysis and only $\mathbf{R}_k$ appears in the implementation. The implementation of $\mathbf{R}_k$ involves solving local problems on each $\Omega_{k,v}$.

Remark 2. The above smoother $\mathbf{R}_k$ is constructed by using an additive Schwarz scheme. A multiplicative version of the smoother can be constructed based on the same space decomposition.

In the remainder of this section, we prove that the smoother $\mathbf{R}_k$ satisfies assumptions (M.1) and (M.2). These results are gathered in the next two lemmas.

LEMMA 4.1. *For* $\rho\le 1/3$, *the smoother* $\mathbf{R}_k$ *satisfies assumption* (M.1).

Proof. The proof follows from the Cauchy–Schwarz inequality and the finite overlapping condition (see, e.g., [14]). ∎

LEMMA 4.2. *The smoother* $\mathbf{R}_k$ *satisfies assumption* (M.2).

Proof. As shown in section 2, there exists a decomposition $\boldsymbol{\sigma}_k = \mathbf{airy}\, q_k + \mathbf{div}_k^{-1}\boldsymbol{v}_k$ for $\boldsymbol{\sigma}_k \in \boldsymbol{\Sigma}_k$, where $q_k \in \mathrm{Q}_k$ and $\boldsymbol{v}_k = \mathbf{div}\,\boldsymbol{\sigma}_k \in \boldsymbol{V}_k/RM$. By Lemma 3.3,

$$(\mathbf{I} - \mathbf{I}_k\mathbf{P}_{k-1})\boldsymbol{\sigma}_k = \sum_{i=1}^{4} \boldsymbol{\sigma}_k^i,$$

where

$$\begin{aligned}
\boldsymbol{\sigma}_k^1 &= \mathbf{airy}\,(q_k - \mathcal{I}_k\mathcal{P}_{k-1}q_k),\\
\boldsymbol{\sigma}_k^2 &= \mathbf{I}_k(\mathbf{airy}\,\mathcal{P}_{k-1}q_k - \mathbf{P}_{k-1}\mathbf{airy}\,q_k),\\
\boldsymbol{\sigma}_k^3 &= \mathbf{div}_k^{-1}\boldsymbol{v}_k - \mathbf{I}_k\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k,\\
\boldsymbol{\sigma}_k^4 &= \mathbf{I}_k(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{P}_{k-1}\mathbf{div}_k^{-1}\boldsymbol{v}_k).
\end{aligned}$$

Notice that all $\boldsymbol{\sigma}_k^i$, $i = 1, 2, 3, 4$, are in $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\sigma}_k^1$ is divergence free. Thus

$$\begin{aligned}
|\boldsymbol{\Lambda}((\mathbf{I} - \mathbf{I}_k\mathbf{P}_{k-1})\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k)| &= |\boldsymbol{\Lambda}(\boldsymbol{\sigma}_k^1 + \boldsymbol{\sigma}_k^2 + \boldsymbol{\sigma}_k^3 + \boldsymbol{\sigma}_k^4, \boldsymbol{\sigma}_k)|\\
&\lesssim |\boldsymbol{\Lambda}(\boldsymbol{\sigma}_k^1, \mathbf{airy}\, q_k)| + \sum_{i=2}^{4} (\mathbf{R}_k^{-1}\boldsymbol{\sigma}_k^i, \boldsymbol{\sigma}_k^i)^{1/2} (\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k, \boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k)^{1/2}.
\end{aligned} \tag{4.3}$$

We will show that
(I) $|\boldsymbol{\Lambda}(\boldsymbol{\sigma}_k^1, \mathbf{airy}\, q_k)| \lesssim (\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k, \boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k)^{1/4}\boldsymbol{\Lambda}(\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k)^{3/4}$;
(II) $(\mathbf{R}_k^{-1}\boldsymbol{\sigma}_k^i, \boldsymbol{\sigma}_k^i) \lesssim \boldsymbol{\Lambda}(\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k)$ for $i = 2, 3, 4$.
Then, since assumption (M.1) implies $(\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k, \boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k) \le \boldsymbol{\Lambda}(\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k)$, assumption (M.2) with $\alpha = 1/4$ will follow from (4.3), (I), and (II).

To prove (I), first notice that for the biharmonic problem, we have (see [15])

$$\frac{1}{\tilde{\lambda}_k}\|\mathrm{A}_k q_k\|_{0,\Omega}^2 \lesssim (\mathcal{R}_k\mathrm{A}_k q_k, \mathrm{A}_k q_k) \quad \text{for all } q_k \in \mathrm{Q}_k,$$

where $\tilde{\lambda}_k = O(h_k^{-4})$ is the largest eigenvalue of the operator A_k.

Theorem 14.1 in [15] states that if Ω is a convex polygon, then

$$\mathrm{A}((\mathrm{I} - \mathcal{I}_k\mathcal{P}_{k-1})q_k, q_k) \lesssim (\mathrm{A}_k q_k, q_k)^{3/4} \left(\frac{\|\mathrm{A}_k q_k\|_{0,\Omega}^2}{\tilde{\lambda}_k}\right)^{1/4}.$$

Therefore,

$$\begin{aligned}
|\boldsymbol{\Lambda}(\boldsymbol{\sigma}_k^1, \mathbf{airy}\, q_k)| &= |\boldsymbol{\Lambda}(\mathbf{airy}\,(q_k - \mathcal{I}_k\mathcal{P}_{k-1}q_k), \mathbf{airy}\, q_k)|\\
&= |\mathrm{A}((\mathrm{I} - \mathcal{I}_k\mathcal{P}_{k-1})q_k, q_k)| \lesssim (\mathrm{A}_k q_k, q_k)^{3/4} \left(\frac{\|\mathrm{A}_k q_k\|_{0,\Omega}^2}{\tilde{\lambda}_k}\right)^{1/4}\\
&\lesssim \boldsymbol{\Lambda}(\boldsymbol{\sigma}_k, \boldsymbol{\sigma}_k)^{3/4} (\mathcal{R}_k\mathrm{A}_k q_k, \mathrm{A}_k q_k)^{1/4}.
\end{aligned}$$

Thus, to prove (I), we only need to show that

$$(\mathcal{R}_k\mathrm{A}_k q_k, \mathrm{A}_k q_k) \le (\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k, \boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k). \tag{4.4}$$

Notice that by the definition of $\mathcal{R}_k$ and $\mathbf{R}_k$,

$$\begin{aligned}
(\mathcal{R}_k\mathrm{A}_k q_k, \mathrm{A}_k q_k) &= \rho \sum_{v \in \mathcal{N}_k} \mathrm{A}(\mathcal{P}_{k,v} q_k, \mathcal{P}_{k,v} q_k),\\
(\mathbf{R}_k\boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k, \boldsymbol{\Lambda}_k\boldsymbol{\sigma}_k) &= \rho \sum_{v \in \mathcal{N}_k} \boldsymbol{\Lambda}(\mathbf{P}_{k,v}\boldsymbol{\sigma}_k, \mathbf{P}_{k,v}\boldsymbol{\sigma}_k).
\end{aligned}$$

Hence (4.4) will follow if for each $v \in \mathcal{N}_k$,

$$\mathrm{A}(\mathcal{P}_{k,v}q_k, \mathcal{P}_{k,v}q_k) = \mathbf{\Lambda}(\mathbf{airy}\,(\mathcal{P}_{k,v}q_k), \mathbf{airy}\,(\mathcal{P}_{k,v}q_k)) \le \mathbf{\Lambda}(\mathbf{P}_{k,v}\boldsymbol{\sigma}_k, \mathbf{P}_{k,v}\boldsymbol{\sigma}_k). \tag{4.5}$$

Notice that for any $p \in \mathrm{Q}_{k,v}$,

$$\begin{aligned}\mathbf{\Lambda}(\mathbf{P}_{k,v}\boldsymbol{\sigma}_k, \mathbf{airy}\,p) &= (\boldsymbol{\sigma}_k, \mathbf{airy}\,p) = (\mathbf{airy}\,q_k, \mathbf{airy}\,p)\\ &= (\mathbf{airy}\,(\mathcal{P}_{k,v}q_k), \mathbf{airy}\,p) = \mathbf{\Lambda}(\mathbf{airy}\,(\mathcal{P}_{k,v}q_k), \mathbf{airy}\,p).\end{aligned}$$

This implies that $\mathbf{airy}\,(\mathcal{P}_{k,v}q_k)$ is the $\mathbf{\Lambda}$-projection of $\mathbf{P}_{k,v}\boldsymbol{\sigma}_k$ into the subspace $\mathbf{airy}\,(\mathrm{Q}_{k,v})$ of $\mathbf{\Sigma}_{k,v}$. Therefore, (4.5) follows. This completes the proof of (I).

Next, we prove (II). For each $v \in \mathcal{N}_k$ let θ_v denote the piecewise continuous linear basis function associated with v. Clearly, $\sum_v \theta_v$ gives a partition of unity on Ω which satisfies

$$\begin{aligned}&(1)\ \theta_v|_T \in P_1(T) \text{ for any } T \in \mathcal{T}_k; \qquad (2)\ \mathrm{supp}(\theta_v) \subset \overline{\Omega}_{k,v};\\ &(3)\ |\theta_v|_{W^{j,\infty}(\Omega)} \lesssim h_k^{-j}, \quad j = 0, 1.\end{aligned}$$

Let Π_k denote the natural interpolation operator onto $\mathbf{\Sigma}_k$ associated with the dofs. Clearly Π_k is linear and preserves $\boldsymbol{\sigma}_k \in \mathbf{\Sigma}_k$. Notice that for each $\boldsymbol{\sigma}_k^i$, $\Pi_k(\theta_v\boldsymbol{\sigma}_k^i)$ is a well-defined function in $\mathbf{\Sigma}_{k,v}$ and $\boldsymbol{\sigma}_k^i = \sum_{v\in\mathcal{N}_k} \Pi_k(\theta_v\boldsymbol{\sigma}_k^i)$. Since the Arnold–Winther element is affine under the matrix Piola transformation [6], a simple scaling argument shows that

$$\|\Pi_k(\theta_v\boldsymbol{\tau})\|_{0,\Omega} \lesssim \|\theta_v\boldsymbol{\tau}\|_{0,\Omega}. \tag{4.6}$$

Also, it has been shown in [6] that $\mathbf{div}\,\Pi_k = \mathbf{P}_{\boldsymbol{V}_k}\mathbf{div}$, where $\mathbf{P}_{\boldsymbol{V}_k}$ is the $\boldsymbol{L}^2$ projection onto $\boldsymbol{V}_k$. Therefore

$$\|\mathbf{div}\,\Pi_k(\theta_v\boldsymbol{\tau})\|_{0,\Omega} = \|\mathbf{P}_{\boldsymbol{V}_k}\mathbf{div}\,(\theta_v\boldsymbol{\tau})\|_{0,\Omega} \le \|\mathbf{div}\,(\theta_v\boldsymbol{\tau})\|_{0,\Omega}.$$

By (4.2), (4.6), an inverse inequality, and the properties of θ_v, for $i = 2, 3, 4$,

$$\begin{aligned}(\mathbf{R}_k^{-1}\boldsymbol{\sigma}_k^i, \boldsymbol{\sigma}_k^i) &\le \rho^{-1}\sum_{v\in\mathcal{N}_k}(\|\Pi_k(\theta_v\boldsymbol{\sigma}_k^i)\|^2_{0,\Omega_{k,v}} + \|\mathbf{div}\,\Pi_k(\theta_v\boldsymbol{\sigma}_k^i)\|^2_{0,\Omega_{k,v}})\\ &\lesssim \sum_{v\in\mathcal{N}_k}(\|\theta_v\boldsymbol{\sigma}_k^i\|^2_{0,\Omega_{k,v}} + \|\mathbf{div}\,(\theta_v\boldsymbol{\sigma}_k^i)\|^2_{0,\Omega_{k,v}})\\ &\lesssim h_k^{-2}\|\boldsymbol{\sigma}_k^i\|^2_{0,\Omega} + \|\mathbf{div}\,\boldsymbol{\sigma}_k^i\|^2_{0,\Omega}.\end{aligned}$$

Hence the proof for (II) reduces to proving for $i = 2, 3, 4$ that

$$\begin{aligned}\|\boldsymbol{\sigma}_k^i\|_{0,\Omega} &\lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)},\\ \|\mathbf{div}\,\boldsymbol{\sigma}_k^i\|_{0,\Omega} &\lesssim \|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned} \tag{4.7}$$

For $\boldsymbol{\sigma}_k^2$ and any $\boldsymbol{\tau}_{k-1} = \mathbf{airy}\,p_{k-1} + \mathbf{div}_{k-1}^{-1}\boldsymbol{w}_{k-1} \in \mathbf{\Sigma}_{k-1}$,

$$\begin{aligned}&|\mathbf{\Lambda}(\mathbf{airy}\,\mathcal{P}_{k-1}q_k - \mathbf{P}_{k-1}\mathbf{airy}\,q_k, \boldsymbol{\tau}_{k-1})|\\ &\qquad = |(\mathbf{airy}\,\mathcal{P}_{k-1}q_k, \mathbf{airy}\,p_{k-1}) - (\mathbf{airy}\,q_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1})|.\end{aligned}$$

Now

$$(\mathbf{airy}\,\mathcal{P}_{k-1}q_k, \mathbf{airy}\,p_{k-1}) = (\mathbf{airy}\,q_k, \mathbf{I}_k\mathbf{airy}\,p_{k-1})$$

so

$$\begin{aligned}|\boldsymbol{\Lambda}(\mathbf{airy}\,\mathcal{P}_{k-1}q_k - \mathbf{P}_{k-1}\mathbf{airy}\,q_k, \boldsymbol{\tau}_{k-1})| &= |(\mathbf{airy}\,q_k, \mathbf{I}_k\mathbf{div}_{k-1}^{-1}\boldsymbol{w}_{k-1})| \\ &\le |(\mathbf{airy}\,q_k, (\mathbf{I}_k - \mathbf{I})\mathbf{div}_{k-1}^{-1}\boldsymbol{w}_{k-1})| + |(\mathbf{airy}\,q_k, \mathbf{div}_{k-1}^{-1}\boldsymbol{w}_{k-1} - \mathbf{div}^{-1}\boldsymbol{w}_{k-1})| \\ &\lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}\|\boldsymbol{\tau}_{k-1}\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned}$$

We used the Cauchy–Schwarz inequality, (2.11), and Lemma 3.5 for the last inequality above. Then, by setting $\boldsymbol{\tau}_{k-1} = \mathbf{airy}\,\mathcal{P}_{k-1}q_k - \mathbf{P}_{k-1}\mathbf{airy}\,q_k$ and using Lemma 3.1, we have

$$\begin{aligned}\|\boldsymbol{\sigma}_k^2\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)} &\lesssim \|\mathbf{airy}\,\mathcal{P}_{k-1}q_k - \mathbf{P}_{k-1}\mathbf{airy}\,q_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)} \\ &\lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned}$$

Therefore, $\boldsymbol{\sigma}_k^2$ satisfies (4.7).

Next, we consider $\boldsymbol{\sigma}_k^3$. Define $\mathbf{P}_{\boldsymbol{V}_{k-1}}$ to be the $\boldsymbol{L}^2$ projection onto $\boldsymbol{V}_{k-1}/RM$. Then

$$\|\mathbf{div}\,\boldsymbol{\sigma}_k^3\|_{0,\Omega} = \|\boldsymbol{v}_k - \mathbf{P}_{\boldsymbol{V}_{k-1}}\boldsymbol{v}_k\|_{0,\Omega} \le \|\boldsymbol{v}_k\|_{0,\Omega} \lesssim \|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}$$

and by (2.11), Lemma 3.5, and the fact that $h_{k-1} = 2h_k$,

$$\begin{aligned}\|\boldsymbol{\sigma}_k^3\|_{0,\Omega} &\lesssim \|\mathbf{div}_k^{-1}\boldsymbol{v}_k - \mathbf{div}^{-1}\boldsymbol{v}_k\|_{0,\Omega} + \|\mathbf{div}^{-1}\boldsymbol{v}_k - \mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k\|_{0,\Omega} \\ &\quad + \|(\mathbf{I} - \mathbf{I}_k)\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k\|_{0,\Omega} \\ &\lesssim h_k\|\boldsymbol{v}_k\|_{0,\Omega} \lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned}$$

Hence $\boldsymbol{\sigma}_k^3$ satisfies (4.7).

For $\boldsymbol{\sigma}_k^4$, let $\boldsymbol{\tau}_{k-1} \in \boldsymbol{\Sigma}_{k-1}$ be arbitrary. Then

(4.8)
$$\begin{aligned}&|\boldsymbol{\Lambda}(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{P}_{k-1}\mathbf{div}_k^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1})| = |\boldsymbol{\Lambda}(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1}) - \boldsymbol{\Lambda}(\mathbf{div}_k^{-1}\boldsymbol{v}_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1})| \\ &\quad = |(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1}) - (\mathbf{div}_k^{-1}\boldsymbol{v}_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1}) + (\mathbf{P}_{\boldsymbol{V}_{k-1}}\boldsymbol{v}_k - \boldsymbol{v}_k, \mathbf{div}\,\tau_{k-1})| \\ &\quad = |(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1}) - (\mathbf{div}_k^{-1}\boldsymbol{v}_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1})|.\end{aligned}$$

Since $(\mathbf{div}^{-1}\boldsymbol{v}_k, (\mathbf{I} - \mathbf{I}_k)\boldsymbol{\tau}_{k-1})$ is zero, by (4.8), (2.11), and Lemma 3.1, we have

$$\begin{aligned}|\boldsymbol{\Lambda}(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{P}_{k-1}\mathbf{div}_k^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1})| &= |(\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{div}^{-1}\boldsymbol{v}_k, \boldsymbol{\tau}_{k-1}) \\ &\quad + (\mathbf{div}^{-1}\boldsymbol{v}_k - \mathbf{div}_k^{-1}\boldsymbol{v}_k, \mathbf{I}_k\boldsymbol{\tau}_{k-1})| \\ &\lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}\|\boldsymbol{\tau}_{k-1}\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned}$$

Setting $\boldsymbol{\tau}_{k-1} = \mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{P}_{k-1}\mathbf{div}_k^{-1}\boldsymbol{v}_k$ and using Lemma 3.1 gives

$$\begin{aligned}\|\boldsymbol{\sigma}_k^4\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)} &\lesssim \|\mathbf{div}_{k-1}^{-1}\boldsymbol{v}_k - \mathbf{P}_{k-1}\mathbf{div}_k^{-1}\boldsymbol{v}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)} \\ &\lesssim h_k\|\boldsymbol{\sigma}_k\|_{\boldsymbol{H}(\mathbf{div},\Omega,\mathbb{S}_2)}.\end{aligned}$$

Therefore, $\boldsymbol{\sigma}_k^4$ satisfies (4.7).

Combining all the above shows that $\mathbf{R}_k$ satisfies assumption (M.2) with a constant C_p independent of k. ∎

TABLE 5.1
Condition number estimates for $\mathbf{\Lambda}_k$, $\mathbf{B}_k^R\mathbf{\Lambda}_k$, $\mathbf{B}_k\mathbf{\Lambda}_k$, *and* $\mathbf{B}_k^m\mathbf{\Lambda}_k$.

level	dofs	cond($\mathbf{\Lambda}_k$)	cond($\mathbf{B}_k^R\mathbf{\Lambda}_k$)	cond($\mathbf{B}_k\mathbf{\Lambda}_k$)	cond($\mathbf{B}_k^m\mathbf{\Lambda}_k$)
2	115	1.58e+04	6.37e+03	3.43	2.66
3	395	7.19e+04	3.90e+04	4.09	3.15
4	1459	2.97e+05	1.67e+05	4.23	3.41
5	5603	1.20e+06	6.82e+05	4.24	3.53

TABLE 5.2
Condition number estimates for $\mathbf{B}_k^V\mathbf{\Lambda}_k$.

level	2	3	4	5
cond($\mathbf{B}_k^V\mathbf{\Lambda}_k$)	3.43	4.03	4.20	4.22

5. Numerical results. We report some numerical results for the multigrid preconditioners for the $\boldsymbol{H}(\mathbf{div})$ problem (2.13). Let Ω be the unit square $(0,1)\times(0,1)$. We solve problem (2.13) by the preconditioned conjugate gradient method (PCG). The right-hand side is selected randomly.

Three different multigrid preconditioners are considered. For variable V-cycle preconditioners, we use $\beta_0 = \beta_1 = 2$ and one smoothing on the finest grid. First, we consider the variable V-cycle multigrid preconditioner with Richardson smoother (denoted by $\mathbf{B}_k^R$). Secondly, we experiment on the variable V-cycle multigrid preconditioner $\mathbf{B}_k$ with the additive Schwarz smoother built on the vertex-based subspaces, as defined in section 4. The scaling factor ρ in (4.1) is set to be $\frac{1}{3}$. Finally, we consider the variable V-cycle multigrid preconditioner $\mathbf{B}_k^m$ using the multiplicative Schwarz smoother as discussed in Remark 2. For all three preconditioners, we set the first level mesh by bisecting Ω using its negatively sloped diagonal.

Experiments show that $\mathbf{B}_k^R$ does not work well, as shown in Table 5.1. We report the condition number estimates for $\mathbf{B}_k\mathbf{\Lambda}_k$ in Table 5.1, together with the condition number estimates for $\mathbf{B}_k^m\mathbf{\Lambda}_k$. Both appear to be bounded independently of k. These results also indicate that $\mathbf{B}_k^m$ works better than $\mathbf{B}_k$, which is not surprising since multiplicative overlapping Schwarz methods have been observed to work better than additive overlapping Schwarz methods for many other applications.

Further experiments also suggest that the V-cycle multigrid preconditioner $\mathbf{B}_k^V$ with the additive Schwarz smoother as in $\mathbf{B}_k$ and one smoothing on each level is also optimal for this test problem (see Table 5.2). We are unable to explain this theoretically.

REFERENCES

[1] D. N. ARNOLD, F. BREZZI, AND J. J. DOUGLAS, *Peers: A new mixed finite element for plane elasticity*, Japan J. Appl. Math., 1 (1984), pp. 347–367.
[2] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Preconditioning in H(div) and applications*, Math. Comp., 66 (1997), pp. 957–984.
[3] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.
[4] D. N. ARNOLD AND R. S. FALK, *A new mixed formulation for elasticity*, Numer. Math., 53 (1988), pp. 13–30.
[5] D. N. ARNOLD, J. J. DOUGLAS, AND C. P. GUPTA, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.
[6] D. N. ARNOLD AND R. WINTHER, *Mixed finite element for elasticity*, Numer. Math., 92 (2002), pp. 401–419.

[7] D. N. Arnold, *Mixed finite element methods for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 82 (1990), pp. 281–300.
[8] T. M. Austin, T. A. Manteuffel, and S. McCormick, *A robust multilevel approach for minimizing H(div)-dominated functionals in an H1-conforming finite element space*, Numer. Linear Algebra Apps., 11 (2004), pp. 115–140.
[9] R. Bank, B. Welfert, and H. Yserentant, *A preconditioning technique for indefinite systems resulting from mixed approximation of elliptic problems*, Numer. Math., 56 (1990), pp. 645–666.
[10] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *Least-squares methods for linear elasticity based on a discrete negative norm*, Comput. Methods Appl. Mech. Engrg., 152 (2001), pp. 520–543.
[11] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997) pp. 1072–1092.
[12] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.
[13] J. H. Bramble and J. E. Pasciak, *A domain decomposition technique for Stokes problems*, Appl. Numer. Math., 6 (1990), pp. 251–261.
[14] J. H. Bramble and J. E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.
[15] J. Bramble and X. Zhang, *Handbook of numerical analysis*, in *Handbook of Numerical Analysis*, vol. VII, P. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 173–415.
[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[17] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.
[18] Z. Cai and G. Starke, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal., To appear.
[19] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[20] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.
[21] R. E. Ewing and J. Wang, *Analysis of the Schwarz algorithm for mixed finite element methods*, RAIRO Math. Model. Anal. Numer., 26 (1992), pp. 739–756.
[22] L. P. Franca, T. J. Hughes, A. F. Loula, and I. Miranda, *A new family of stable elements for nearly incompressible elasticity based on a mixed Petrov-Galerkin finite element formulation*, Numer. Math., 53 (1988), pp. 123–141.
[23] P. Grisvard, *Singularities in Boundary Value Problems*, Research Notes in Appl. Math. 22, Springer-Verlag, New York, 1992.
[24] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[25] R. Hiptmair, *Multigrid method for H(div) in three dimensions*, Electron. Trans. Numer. Anal., 6 (1997), pp. 133–152.
[26] C. Johnson and B. Mercier, *Some equilbrium finite element methods for two-dimensional elasticity problems*, Numer. Math., 30 (1978), pp. 103–116.
[27] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problem*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.
[28] Y. Wang, *Overlapping Schwarz preconditioner for the mixed formulation of plane elasticity*, Applied. Numer. Math., 54 (2005), pp. 292–309.
[29] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[30] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 494–514

LOCALIZED POINTWISE A POSTERIORI ERROR ESTIMATES FOR GRADIENTS OF PIECEWISE LINEAR FINITE ELEMENT APPROXIMATIONS TO SECOND-ORDER QUASILINEAR ELLIPTIC PROBLEMS*

ALAN DEMLOW†

Abstract. Two types of pointwise a posteriori error estimates are presented for gradients of finite element approximations of second-order quasilinear elliptic Dirichlet boundary value problems over convex polyhedral domains Ω in space dimension $n \geq 2$. We first give a residual estimator which is equivalent to $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ up to higher-order terms. The second type of residual estimator is designed to control $\nabla(u-u_h)$ locally over any subdomain of Ω. It is a novel a posteriori counterpart to the localized or weighted a priori estimates of [Sch98]. This estimator is shown to be equivalent (up to higher-order terms) to the error measured in a weighted global norm which depends on the subdomain of interest. All estimates are proved for general shape-regular meshes which may be highly graded and unstructured. The constants in the estimates depend on the unknown solution u in the nonlinear case, but in a fashion which places minimal restrictions on the regularity of u.

Key words. finite element methods, quasilinear elliptic problems, a posteriori error estimation, pointwise error analysis

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/040610064

1. Introduction and results.

1.1. Introduction. We consider finite element approximations to second-order quasilinear elliptic Dirichlet boundary value problems having the form

$$-\sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(x,u,\nabla u) + F_0(x,u,\nabla u) = 0 \text{ in } \Omega, \qquad u = b \text{ on } \partial\Omega. \tag{1.1}$$

Here Ω is a convex polyhedral domain in $\mathbb{R}^n$, $n \geq 2$, and we assume that $u \in C^{1,\alpha}(\overline{\Omega})$ for some $0 < \alpha \leq 1$. The coefficients $F_i(x,z,p)$ are assumed to be elliptic (although not necessarily uniformly so) and to satisfy minimal smoothness requirements given in detail later. Problems ranging from uniformly elliptic equations to highly nonlinear, nonuniformly elliptic equations take the form (1.1). Examples which may be treated with the techniques presented here include uniformly elliptic linear problems, where $F_i(x,z,p) = \sum_{j=1}^n a_{ij}(x)p_j$, $1 \leq i \leq n$, and $F_0(x,z,p) = \vec{b}(x)\cdot p + c(x)z - f(x)$; the prescribed mean curvature equation, where $F_i(x,z,p) = p_i/\sqrt{1+|p|^2}$, $1 \leq i \leq n$, and $F_0(x,z,p) = -H(x)$; and mildly nonlinear equations, where $F_i(x,z,p) = \sum_{j=1}^n a_{ij}(x,z)p_j$, $1 \leq i \leq n$, and $F_0(x,z,p) = -f(x)$.

In this paper we provide two types of computationally efficient residual-based pointwise a posteriori error estimators for the gradient error $\nabla(u-u_h)$ in the piecewise linear finite element approximation u_h to u. We first give estimators which are

*Received by the editors June 16, 2004; accepted for publication (in revised form) September 12, 2005; published electronically March 15, 2006. This material is based upon work partially supported under a National Science Foundation postdoctoral research fellowship.

http://www.siam.org/journals/sinum/44-2/61006.html

†Abteilung für Angewandte Mathematik, Hermann-Herder-Str. 10, 79104 Freiburg, Germany (demlow@mathematik.uni-freiburg.de).

equivalent up to constants and logarithmic factors to $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$. While most a posteriori error estimates in the literature similarly control global norms of the error, the quantity of interest in many practical calculations is dependent only on the solution in some subset D of Ω. The goal in these cases is to refine the mesh enough globally to ensure that the solution in $\Omega \setminus D$ does not "pollute" the solution in D while not overrefining in $\Omega \setminus D$. To this end, we prove an a posteriori error estimate for $\|\nabla(u-u_h)\|_{L_\infty(D)}$ which is a novel a posteriori counterpart to the weighted or localized a priori pointwise estimates proved in [Sch98]. The resulting estimators, which we call localized estimators, bound $\|\nabla(u-u_h)\|_{L_\infty(D)}$ and are essentially equivalent to a certain weighted global norm of $\nabla(u-u_h)$. Both types of estimates are valid on general shape-regular meshes and under reasonable regularity assumptions on coefficients and the solution u. To our knowledge, these estimates are the first to provide pointwise error control for gradients in either global or local norms on highly graded, unstructured meshes.

The W^1_∞ estimates we give here are in several senses an extension of the framework for a posteriori analysis of nonlinear problems in integral norms which was proposed in [Ver94]. As in that work, our estimates provide a theoretical basis for a posteriori error estimation and adaptive mesh refinement but also suffer from several drawbacks. The first is that we can prove reliability of our estimators only under the uncomputable condition that $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ is small enough. Second, a priori constants appear in our a posteriori upper bounds. Finally, the estimates presented here suffer from a "spectral gap" between the a posteriori upper and lower bounds when the maximum pointwise ratio of the largest and smallest eigenvalues of the coefficient matrix $[\frac{\partial}{\partial p_j} F_i(x,u,\nabla u)]$ is large. [FV03] proposes a method which essentially eliminates the first two of these problems in the context of a posteriori error estimation in the energy norm for equations of prescribed mean curvature. The third problem mentioned above, ill conditioning resulting from a spectral gap, seems to be an essential feature of residual-type estimators for linear as well as nonlinear problems; cf. [FV03] and [BV00].

In the present work we focus on presenting a basic theory for problems on polyhedral domains. Two important questions which we do not consider here are the case of smooth boundaries and the treatment of the constants arising in our estimates. The first of these questions is important for many nonlinear problems as even theoretical results are not always available on polygonal domains. However, the proper treatment of finite element approximations involving curved boundaries is somewhat technical even when considering a posteriori energy-norm bounds (cf. [DR98]), and we do not wish to clutter our presentation. Second, the constants in our a posteriori estimates depend on the unknown solution u in nonlinear problems and even in linear problems may depend on the coefficients in a fashion that will require a local weighting of the residuals. In [De05] we combine further theoretical results with computational experiments in order to investigate this problem.

1.2. Outline of results. Before outlining our results, we introduce some notation. First note that we shall restrict most of our presentation to a model problem having the form

$$\begin{aligned} -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(x,\nabla u) &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{1.2}$$

where Ω is convex and polygonal. Most of the examples mentioned in the previous

section are of the form (1.2). Extension to more general coefficients is fairly immediate under appropriate assumptions. In section 5 we sketch the necessary modifications and also provide a brief analysis of problems with nonhomogeneous Dirichlet boundary conditions.

Let $\mathcal{T}$ be a decomposition of Ω (now assumed to be polygonal) into shape-regular simplices. Let also $h_T = |T|^{1/n}$ for each element $T \in \mathcal{T}$, let $\underline{h} = \min_{T\in\mathcal{T}} h_T$, and let S_h be the continuous piecewise linear functions which are 0 on $\partial\Omega$. We emphasize that we place no restrictions on the mesh other than shape regularity, so that both highly graded and unstructured meshes are allowed throughout. A logarithmic factor which for technical reasons is different when $n = 2$ also appears in our estimates, and for convenience we define $\gamma(2) = 2$, $\gamma(n) = 1$ for $n > 2$, and $\ell_{\underline{h},n} = (\log(1/\underline{h}))^{\gamma(n)}$.

We next define a first-order maximum norm residual. Let S be a face shared by two elements T_1 and T_2, and let $\vec{n}$ be a unit normal on S (with arbitrary orientation). For $v_h \in S_h$, we then define

$$[v_h]_S(x) = \sum_{i=1}^{n} [F_i(x, \nabla v_h|_{T_1}) - F_i(x, \nabla v_h|_{T_2})]n_i$$

with $[v_h]_S(x) = 0$ when $S \subset \partial\Omega$. Dropping the subscript S as it will cause no confusion, we define the residual

$$\mathcal{E}_T = h_T \|f + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(\cdot, \nabla u_h)\|_{L_\infty(T)} + \|[u_h]\|_{L_\infty(\partial T)}. \tag{1.3}$$

Our first result is the global estimate

$$\frac{1}{\tilde{C}_1} \max_{T\in\mathcal{T}} (\mathcal{E}_T - R_1(T)) \le \|\nabla(u - u_h)\|_{L_\infty(\Omega)} \le C_1 \ell_{\underline{h},n} \max_{T\in\mathcal{T}} \mathcal{E}_T + R_2(\Omega), \tag{1.4}$$

where R_i, $i \ge 1$, denotes a higher-order term which will be defined more precisely later and C_1 and $\tilde{C}_1$ depend on the coefficients F_i, $\|\nabla u\|_{L_\infty(\Omega)}$, and the Dini continuity of ∇u in the nonlinear case. Thus our estimators are equivalent to the actual error up to constants, logarithmic factors, and higher-order terms, that is, they are efficient and reliable. We do not attempt to provide asymptotically exact estimators. Besides having obvious application to control of global norms of gradient errors, our global estimates may also be combined with the results of [Noc95] and [DDP00] to establish a posteriori estimates for $\|u - u_h\|_{L_\infty(\Omega)}$ for nonlinear problems on convex polygonal and polyhedral domains in two and three space dimensions.

In order to provide local error control, we present novel localized estimators which are inspired by the localized or weighted a priori pointwise estimates proved in [Sch98]. These estimates are valid for smooth linear Neumann problems on globally quasi-uniform meshes of size h. Defining the weight $\sigma_{x_0}(y) = \frac{h}{|y-x_0|+h}$, it was shown that

$$|\nabla(u - u_h)(x_0)| \le \|\sigma_{x_0} \nabla(u - u_h)\|_{L_\infty(\Omega)} \le C \min_{\chi\in S_h} \|\sigma_{x_0} \nabla(u - \chi)\|_{L_\infty(\Omega)}.$$

In our a posteriori results, we wish to control $\nabla(u - u_h)$ over any subset D of Ω, so we define the piecewise constant weight

$$\sigma_D(T) = \frac{h_T}{dist(T, D) + h_T}.$$

We prove the localized a posteriori estimate

$$\begin{aligned}(1.5)\quad \tfrac{1}{C_1}\max_{T\in\mathcal{T}}\sigma_D(T)(\mathcal{E}_T - R_1(T)) &\le \|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}\\ &\le C_2\ell_{\underline{h},n}\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T + \mathcal{E}_T^2) + R_3(\Omega).\end{aligned}$$

Here C_2 depends on the coefficients F_i and $\|u\|_{W^2_\infty(\Omega)}$ in the nonlinear case, and the term $\mathcal{E}_T^2$ may be dropped in the linear case. Note that the right-hand side of (1.5) also bounds $\|\nabla(u-u_h)\|_{L_\infty(D)}$ since $\sigma_D \equiv 1$ on D. Beyond the evaluation of standard residuals, the only requirement for the practical implementation of the estimator $\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T + \mathcal{E}_T^2)$ is the ability to efficiently compute the distance to the set D.

The constants C_1 and C_2 appearing in the estimates (1.4) and (1.5) depend on the unknown solution u in nonlinear situations, and these estimates are thus not strictly speaking a posteriori estimates. We do establish that C_1 depends only on weak regularity properties of u ($\|\nabla u\|_{L_\infty(\Omega)}$ and the Dini continuity of ∇u), and C_2 depends on moderate regularity properties of u ($\|u\|_{W^2_\infty(\Omega)}$). Although perhaps possible, tracing a more precise theoretical dependence of these constants on u would be difficult and, more important for practical purposes, unlikely to yield sharp results. In nonlinear problems especially, the estimators given here are therefore at most suitable for use as error indicators in adaptive mesh refinement.

We finally give a brief survey of relevant work related to pointwise a posteriori estimates and a posteriori estimates for nonlinear problems. In [Ver94], a general framework is given for a posteriori error estimation in canonical or energy norms for nonlinear problems. As mentioned in the introduction, our global estimates are maximum-norm analogues (in a rather more restricted situation) to the estimates presented in that work in that they yield reliable estimators only for u_h close enough to u, and the constants in the a posteriori estimates depend on the unknown solution u. The method of proof used here is partially inspired by that used in [Noc95] and [DDP00] to establish reliable and efficient a posteriori estimates for $\|u-u_h\|_{L_\infty(\Omega)}$ for linear problems on general shape regular grids on arbitrary polygonal domains in $\mathbb{R}^2$ and $\mathbb{R}^3$. A technique related to our localized estimates when D consists of a single point is the "dual weighted residual" method of [BR01], which involves solving a linear dual problem for each point for which one wishes to control the gradient error. Finally, in [HSWW01] and [SW04], localized a priori pointwise estimates are employed to provide sharply local and asymptotically exact pointwise control of the gradient via "gradient recovery" operators. These estimates have been shown to be valid only for smooth linear problems and on globally quasi-uniform meshes, however.

The outline of the paper is as follows. Section 2 contains further preliminaries and assumptions. In sections 3 and 4 we give precise results and proofs along with more detailed discussion of our global and localized a posteriori estimates, respectively. In section 5 we briefly discuss extensions to problems of the form (1.1).

2. Preliminaries. In this section we make a number of definitions and state some lemmas.

2.1. Finite element approximation and mesh. In addition to the notation and assumptions introduced in the previous section, we make the following definitions. By shape regular we mean that there exist positive constants r_1 and r_2 such that for each $T \in \mathcal{T}$, one may inscribe a sphere of radius $r_1 h_T$ in T and inscribe T in a sphere of radius $r_2 h_T$. Letting T_x be an arbitrary element containing the point x, we denote by $h(x)$ the quantity h_{T_x}. Additionally, we define the patches $P_T =$

$\cup_{\{T'\in\mathcal{T} \text{ such that } \bar{T}\cap\bar{T}'\neq\emptyset\}}\bar{T}'$, $P_T' = \cup_{\{T'\subset P_T\}}P_{T'}$, and $P_T'' = \cup_{\{T'\subset P_T'\}}P_{T'}$. Finally, we assume that there exists a finite element approximation $u_h \in S_h$ to u satisfying

$$\int_\Omega \sum_{i=1}^n F_i(x,\nabla u_h)\chi_{x_i}\,\mathrm{d}x = \int_\Omega f\chi\,\mathrm{d}x \;\forall\; \chi \in S_h. \tag{2.1}$$

The proof of our localized estimates requires a global growth condition on the mesh which is implied by shape regularity, a fact which we now formulate and prove.

PROPOSITION 2.1. *Assume the triangulation $\mathcal{T}$ is shape regular. Then there exists a constant $C_{\mathcal{T}}$ depending only on the shape regularity of $\mathcal{T}$ such that for the barycenter x_T of each element $T \in \mathcal{T}$, there holds for each point $y \in \Omega \setminus T$*

$$h(y) \le C_{\mathcal{T}}|x_T - y|. \tag{2.2}$$

Proof. First fix an element T with barycenter x_T. Shape regularity implies that there exists $0 < K_1$ such that

$$dist(x_T, \partial T) \ge K_1 h_T. \tag{2.3}$$

Next note that the elements contained in P_T are quasi-uniform, that is, there exist constants $K_3 \le 1 \le K_4$ such that for each $T' \subset P_T$,

$$K_3 h_T \le h_{T'} \le K_4 h_T.$$

We shall without loss of generality assume that $K_4 \ge K_1$. Finally, shape regularity implies that there exists $0 < K_2 \le 1$ such that for each point $y \in T$,

$$B_{K_2 h_T}(y) \subset P_T. \tag{2.4}$$

We now assert that (2.2) holds with $C_{\mathcal{T}} = \frac{K_4}{K_1K_2}$. Note from (2.3) that if $y \in P_T \setminus T$, then $|y - x_T| \ge dist(x_T, \partial T) \ge K_1 h_T$. Since $K_2 \le 1$, we thus have

$$h(y) \le K_4 h_T \le \frac{K_4}{K_1}|y - x_T| \le \frac{K_4}{K_1K_2}|y - x_T|.$$

Now assume that $y \notin P_T$, that is, $\bar{T} \cap \bar{T}_y = \emptyset$. In order to reach a contradiction, we assume that $h(y) > \frac{K_4}{K_1K_2}|y - x_T|$. Then since $\frac{K_4}{K_1} \ge 1$,

$$K_2 h(y) > \frac{K_4}{K_1}|x_T - y| \ge |x_T - y|,$$

that is, $x_T \in B_{K_2h(y)}(y)$. Thus by (2.4), $x_T \in B_{K_2h(y)}(y) \subset P_{T_y}$, that is, $\bar{T} \cup \bar{T}_y \neq \emptyset$. This is a contradiction, so our proposition is proved. ∎

We shall also employ the Scott–Zhang interpolation operator I_h defined in [SZ90] which preserves homogeneous boundary conditions and satisfies

$$\|v - I_h v\|_{L_1(T)} \le C h_T^{1+j} \|v\|_{W_1^{1+j}(P_T)},\; j = 1, 2, \tag{2.5}$$

and

$$\|v - I_h v\|_{W_1^1(T)} \le C h_T^{j} \|v\|_{W_1^{1+j}(P_T)},\; j = 0, 1. \tag{2.6}$$

Here and throughout, C is a constant which depends at most on Ω and the shape regularity of $\mathcal{T}$.

2.2. Auxiliary problems and assumptions on coefficients. We assume that the coefficients $F_i(x,p)$ are twice continuously differentiable in p and define $F_{ij}(x,p) = \frac{\partial}{\partial p_j}F_i(x,p)$ and $F_{ijk}(x,p) = \frac{\partial^2}{\partial p_j \partial p_k}F_i(x,p)$. We also require that $F_i(x,p)$, $1 \le i \le n$, have derivatives with respect to the x variable which are uniformly bounded with respect to both x and p. We note that the analysis of our global results requires only that $F_{ij}(x,p)$ be Dini-continuous with respect to the x variable, but $F_i(x,p)$ must possess bounded spatial derivatives over each element in order to guarantee that the residual (1.3) is computable. Finally, we assume the ellipticity condition

$$\sum_{i,j=1}^{n} F_{ij}(x,p)\xi_i\xi_j > 0 \quad \forall \quad x \in \overline{\Omega},\ \xi \in \mathbb{R}^n \setminus \{0\}, p \in \mathbb{R}^n. \tag{2.7}$$

Remark 2.2. The conditions placed on the coefficients F_i may be slightly relaxed at the expense of some complication in our presentation. First, for our global results it is necessary not that $F_{ijk}(x,p)$ exist but rather only that $F_{ij}(x,p)$ be Hölder continuous in p with Hölder exponent $0 < \alpha \le 1$. A perturbation term of the form $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ arising in our global results would be replaced in this situation by $\|\nabla(u-u_h)\|^{1+\alpha}_{L_\infty(\Omega)}$. Second, the ellipticity condition (2.7) must only hold for $p \in \overline{range(\nabla u)}$ and not for all p in $\mathbb{R}^n$. This latter observation would, for example, allow analysis of the β-Laplacian, where $F_i(x,p) = p_i|p|^{\beta-2}$ and $1 < \beta < \infty$, if it could be established a priori that $|\nabla u| \ge C > 0$. For W^2_∞ solutions of the β-Laplacian with $\beta > 2$, Lemma 4.2 of [BL93] establishes such an inequality if $|f| \ge C > 0$. However, we are not aware of any regularity results in the literature which would guarantee such smooth solutions of this problem.

Two auxiliary linear problems are used in our analysis of quasilinear problems. Following for example [FR78], we define

$$a^h_{ij} = \int_0^1 F_{ij}(x, \nabla u_h + t\nabla(u-u_h))\ \mathrm{d}t, \quad i,j = 1,\dots,n,$$

and

$$a_{ij} = F_{ij}(x, \nabla u), \quad i,j = 1,\dots,n.$$

Correspondingly, we define bilinear forms

$$A_h(v,w) = \int_\Omega \sum_{i,j=1}^{n} a^h_{ij} v_{x_j} w_{x_i}\,\mathrm{d}x \tag{2.8}$$

and

$$A(v,w) = \int_\Omega \sum_{i,j=1}^{n} a_{ij} v_{x_j} w_{x_i}\,\mathrm{d}x.$$

From the ellipticity and smoothness of the coefficients F_i and the boundedness of ∇u, we can conclude that $[a_{ij}]$ is uniformly elliptic in Ω, that is,

$$\lambda|\xi|^2 \le \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \le \Lambda|\xi|^2. \tag{2.9}$$

We emphasize that for nonlinear problems, λ and Λ in general depend on $\|\nabla u\|_{L_\infty(\Omega)}$.

The ellipticity of A_h, on the other hand, depends upon $\|\nabla u_h\|_{L_\infty(\Omega)}$, but A_h satisfies the error equation

$$A_h(u-u_h,\chi)=\int_\Omega \sum_{i=1}^n (F_i(x,\nabla u)-F_i(x,\nabla u_h))\chi_{x_i}\,\mathrm{d}x=0 \tag{2.10}$$

for $\chi\in S_h$, and in fact for general $v\in H_0^1(\Omega)$,

$$A_h(u-u_h,v)=\int_\Omega \sum_{i=1}^n (F_i(x,\nabla u)-F_i(x,\nabla u_h))v_{x_i}\,\mathrm{d}x. \tag{2.11}$$

We finally note that with S denoting the convex hull of $range(\nabla u)$ and $range(\nabla u_h)$,

$$\begin{aligned} |a_{ji}-a_{ji}^h| &= \Big|\int_0^1 F_{ji}(\nabla u)-F_{ji}(\nabla u_h+t\nabla(u-u_h))\,\mathrm{d}t\Big| \\ &\le \int_0^1 \sum_{k=1}^n \|F_{jik}\|_{L_\infty(S)}(1-t)|\nabla(u-u_h)|\,\mathrm{d}t \\ &\le C_F|\nabla(u-u_h)|. \end{aligned} \tag{2.12}$$

The essential estimate $\max_{1\le i,j,k\le n}\|F_{ijk}\|_{L_\infty(S)}\le C_F$ may be established here in one of two ways. It sometimes happens that F_{ijk} is bounded on $\Omega\times\mathbb{R}^n$, so that the bound is immediate and does not rely on ∇u and ∇u_h. If F_{ijk} is not globally bounded, then we must assume a priori that $\|\nabla u_h\|_{L_\infty(\Omega)}\le C$ or alternatively that $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}\le C$. C_F can then be taken to be the bound for $\max_{1\le i,j,k\le n}|F_{ijk}(x,p)|$ on the compact set $\{x\in\overline{\Omega},\ |p|\le\|\nabla u\|_{L_\infty(\Omega)}+C\}$. Thus we shall assume that either $F_{ijk}(x,p)$ is globally bounded in both x and p for all $1\le i,j,k\le n$ or that $\|\nabla u_h\|_{L_\infty(\Omega)}\le C$.

2.3. Green's function estimates. We denote by $G(x,y)$ the Green's function satisfying $A(G(x,\cdot),v)=v(x)$ for sufficiently smooth $v\in H_0^1(\Omega)$. The following estimate for the first and mixed second derivatives of G is essential to our proofs.

LEMMA 2.3. *Assume that the coefficients a_{ij} are Dini-continuous and satisfy the uniform ellipticity condition* (2.9)*, and let Ω be smooth or convex. Assume that $|\alpha|\le 1$ and $|\beta|\le 1$. Then for $n\ge 3$*

$$|D_x^\alpha D_y^\beta G(x,y)|\le C_G|x-y|^{2-n-|\alpha|-|\beta|} \tag{2.13}$$

and for $n=2$

$$|D_x^\alpha D_y^\beta G(x,y)|\le C_G|x-y|^{2-n-|\alpha|-|\beta|}\log\frac{1}{|x-y|}. \tag{2.14}$$

Here C_G depends on Ω, the Dini-continuity of the coefficients a_{ij}, and λ and Λ.

The estimate (2.13) for space dimension $n\ge 3$ may be found in [GW82] assuming that $\partial\Omega$ satisfies a uniform exterior sphere condition. This condition is met by both convex and smooth domains. The proof given in [GW82] does not carry directly over to $n=2$ due to the logarithmic nature of the singularity, but one may use the same method to obtain the suboptimal estimate (2.14) so long as the estimate

$$|G(x,y)|\le C(\lambda,\Lambda,\Omega)\log\frac{1}{|x-y|}$$

is known. This estimate is contained in [DM95] under the weak restrictions of L_∞ and uniformly elliptic coefficients and Lipschitz boundary $\partial\Omega$. The suboptimal estimate (2.14) will only add an additional logarithmic factor to our results in the case $n=2$.

3. Global estimate. In this section we state, discuss, and prove reliability and efficiency results for global estimators for $\nabla(u-u_h)$.

3.1. Reliability of global estimators. First we state the following upper bound for $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$.

THEOREM 3.1. *In addition to the assumptions of section 2, assume that $u \in C^{1,\nu}(\overline{\Omega})$ for some $0<\nu\le 1$. Then for any $0<\alpha\le\nu$ and any $\beta\ge 1$,*

$$\begin{aligned}\|\nabla(u-u_h)\|_{L_\infty(\Omega)} \le{}& C_1\beta^{\gamma(n)}\ell_{\underline{h},n}\bigl(\max_{T\in\mathcal{T}}\mathcal{E}_T + C_F\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}\bigr)\\ &+ C\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}.\end{aligned}\tag{3.1}$$

Here C_1 depends on C_G, the ellipticity coefficients λ and Λ, and the shape regularity of $\mathcal{T}$. In the nonlinear case, C_1 thus depends on $\|\nabla u\|_{L_\infty(\Omega)}$, the Dini-continuity of ∇u, and the coefficients F_i. In the linear case, C_1 does not depend on u and the term $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ in (3.1) does not appear.

Remark 3.2. The term $\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}$ may be omitted for $\underline{h}$ small enough if we make the nondegeneracy assumption $\underline{h}^{\epsilon}|u|_{C^{1,\alpha}(\overline{\Omega})}\le\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ for some $\epsilon>0$. We may then take $\beta=(\epsilon+1)/\alpha$ and $\underline{h}$ small enough to kick back the resulting term $C\underline{h}\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$. In [De04] we give a more precise nondegeneracy assumption which relies on lower bounds for polynomial approximations. In particular, assume that there exists a single point $\tilde{x}\in\Omega$ and $\eta>0$ such that $|D^2u(\tilde{x})|\ge\tilde{C}>0$ and $\|u\|_{W^3_\infty(B_\eta(\tilde{x}))}\le\tilde{C}'$. The term $\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}$ may then be removed at the expense of a weak preasymptotic a priori dependence in the logarithmic factor. We do not give the details here. This more precise nondegeneracy assumption leads to an estimate which for linear problems is reliable on coarse meshes, and in most practical situations we may thus omit this term.

In [Noc95] and [DDP00], the Hölder continuity of u (instead of ∇u) and an assumption similar to the condition $\underline{h}^{\epsilon}|u|_{C^{1,\alpha}(\overline{\Omega})}\le\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ above were used in the establishment of asymptotically reliable residual estimators for $\|u-u_h\|_{L_\infty(\Omega)}$ for linear problems on nonconvex polygonal domains. The technique we use in [De04] to remove the term $\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}$ is essentially a more rigorous and careful version of the argument contained in these works. A different and more sophisticated argument was used in [NSV03] to remove such nondegeneracy assumptions completely and thus prove L_∞ estimates which have no a priori dependence in the upper bounds and which are valid on coarse meshes. This technique does not appear to be applicable in the current context of W^1_∞ estimates, however.

Remark 3.3. The estimate (3.1) includes a logarithmic factor, whereas typical a priori estimates for $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ do not. It is possible to remove the logarithmic factor under the restriction that the mesh be quasi-uniform on balls of size $c\log(1/\underline{h})$ for any fixed $c>0$. Removing this computationally negligible factor would also lead to a stronger dependence of C_1 on u in the nonlinear case or on the coefficients a_{ij} in the linear case.

Next note that we may kick back the term $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ in (3.1) if

$$C_1C_F\beta^{\gamma(n)}\ell_{\underline{h},n}\|\nabla(u-u_h)\|_{L_\infty(\Omega)}\le C^*<1.\tag{3.2}$$

See [FR78] for an asymptotic a priori estimate for $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ on quasi-uniform meshes and under stricter regularity assumptions than we have made here. Using (3.2) and Remark 3.2 while noting that $\|u_h\|_{L_\infty(\Omega)}\le C$ if (3.2) holds, we may formulate an asymptotic reliability result which yields a computable estimator.

COROLLARY 3.4. *Assume that $\mathcal{T}$ is shape regular, $u \in C^{1,\nu}(\overline{\Omega})$, $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ is small enough, $\underline{h} \leq C$, and that $\|\nabla(u-u_h)\|_{L_\infty(\Omega)} \geq C\underline{h}^{\epsilon}|u|_{C^{1,\nu}(\overline{\Omega})}$ for some $\epsilon > 0$. Then*

$$\|\nabla(u-u_h)\|_{L_\infty(\Omega)} \leq C_1 \left(\frac{\epsilon+1}{\nu}\right)^{\gamma(n)} \ell_{\underline{h},n} \max_{T\in\mathcal{T}} \mathcal{E}_T. \tag{3.3}$$

Here C_1 is as in Theorem 3.1.

As stated in the introduction, the condition that $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ is small enough is an a priori and uncomputable condition. We refer again to [Ver94] for a posteriori estimates for nonlinear problems in integral norms which are reliable only for u_h close enough to u and which have reliability constants depending on the unknown solution u. For energy norms, an alternative approach to that of [Ver94] is to bound the error in a weighted problem-dependent norm depending on u_h. In some cases one may thus avoid the problem of unknown a priori constants replace the requirement that ∇u and ∇u_h be close enough with a computable condition whose fulfillment ensures reliability of the a posteriori upper bound; cf. [FV03] for such an example.

We finally note that our results may easily be combined with those of [Noc95] and [DDP00] to establish a bound for $\|u-u_h\|_{L_\infty(\Omega)}$ for quasilinear problems on convex polyhedral domains in two and three space dimensions.

COROLLARY 3.5. *Assume that the conditions of Theorem* 3.1 *are satisfied and that in addition the coeffcients F_i are nonlinear and $u \in W^2_\infty(\Omega)$. Then for any $\alpha > 0$ and $\beta > 0$,*

$$\begin{aligned}\|u-u_h\|_{L_\infty(\Omega)} \leq{}& \tilde{C}\beta^{\gamma(n)}\ell_{\underline{h}}[\max_{T\in\mathcal{T}} h_T\mathcal{E}_T + \max_{T\in\mathcal{T}} \mathcal{E}_T^2 \\ &+ \|\nabla(u-u_h)\|^4_{L_\infty(\Omega)}] + C[\underline{h}^{\alpha\beta}|u|_{C^\alpha(\overline{\Omega})} + \underline{h}^{2\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}],\end{aligned} \tag{3.4}$$

where $\tilde{C}$ depends on $\|u\|_{W^2_\infty(\Omega)}$ and the coefficients F_i, and $\ell_{\underline{h}}$ is a generic logarithmic factor.

Dropping the higher-order terms in the second line yields an asymptotically reliable estimator for $\|u-u_h\|_{L_\infty(\Omega)}$.

3.2. Efficiency of global estimators. Before stating our efficiency result, we define P_h to be the L_2 projection onto the functions which are piecewise constant on $\mathcal{T}$ and let $\tilde{P}_h$ be the L_2 projection onto the set of functions which are piecewise constant on the edges of elements in $\mathcal{T}$.

THEOREM 3.6. *Assume that either F_{ij} is globally bounded for $1 \leq i,j \leq n$, or that $\|\nabla u_h\|_{L_\infty(\Omega)} \leq C$. Then for any element $T \in \mathcal{T}$,*

$$\begin{aligned}\mathcal{E}_T \leq{}& \tilde{C}_1\|\nabla(u-u_h)\|_{L_\infty(P_T)} + Ch_T\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(P_T)} \\ &+ C\|[u_h] - \tilde{P}_h[u_h]\|_{L_\infty(\partial T)}.\end{aligned} \tag{3.5}$$

Here $\tilde{f}_h(x) = f(x) + \sum_{i=1}^n \frac{\partial}{\partial x_i} F_i(x,\nabla u_h)$ and $\tilde{C}_1 = C\|a^h_{ij}\|_{L_\infty(\Omega)}$ is bounded independent of u_h under the assumptions of this theorem.

Remark 3.7. If the coefficients $F_i(x,p)$ do not depend on x (as is the case for example for the prescribed mean curvature problem), then the higher-order term $h_T\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(P_T)} + \|[u_h] - \tilde{P}_h[u_h]\|_{L_\infty(\partial T)}$ reduces to $\|f - P_h f\|_{L_\infty(P_T)}$.

3.3. Proof of reliability. In our proofs we shall use a discrete δ-function. Modifying the technique used in [Noc95], we fix a point x and define a function δ_x as follows. Let $x \in T \in \mathcal{T}$. We then fix a simplex $\tilde{T}$ such that $x \in \tilde{T} \subset T$ and $\tilde{T}$ is shape regular

with diameter $\rho = \underline{h}^\beta$ for $\beta \geq 1$, where β is as given in the statement of Theorem 3.1. We then let $\delta_x \in C_0^\infty(\tilde{T})$ be a nonnegative function such that $\int_{\tilde{T}} \delta_x \, \mathrm{d}y = 1$,

$$\|\delta_x\|_{W_p^k(\tilde{T})} \leq C\rho^{n(1-1/p)-k}, \tag{3.6}$$

and $dist(\mathrm{supp}(\delta_x), \partial\tilde{T}) \geq c\rho$ for some $c > 0$. Such a function δ_x is easy to define by scaling and translation to $\tilde{T}$ from a reference element.

Denote by ∂ a first-order directional differential operator, that is, $\partial = \nabla \cdot \vec{v}$ for some $\vec{v}$ with $|\vec{v}| = 1$. Let $x_0 \in T$ and ∂ be such that $\|\nabla(u - u_h)\|_{L_\infty(\Omega)} \leq C|\partial(u - u_h)(x_0)|$, and let $\overline{\partial u} = \frac{1}{|\tilde{T}|}\int_{\tilde{T}} \partial u \, \mathrm{d}x$. Noting that ∂u_h is constant on $\tilde{T}$ and that $\overline{\partial u} = \partial u(x_1)$ for some $x_1 \in \tilde{T}$, we compute

$$\begin{aligned}
\|\nabla(u - u_h)\|_{L_\infty(\Omega)} &\leq C|\partial(u - u_h)(x_0)| \\
&\leq C(|\partial u(x_0) - \partial u(x_1)| + |\overline{\partial u} - \partial u_h|) \\
&\leq C(\rho^\alpha |u|_{C^{1,\alpha}(\tilde{T})} + |(\overline{\partial u} - \partial u_h, \delta_{x_0})|) \\
&\leq C(\rho^\alpha |u|_{C^{1,\alpha}(\tilde{T})} + |(\overline{\partial u} - \partial u, \delta_{x_0})| + |(\partial(u - u_h), \delta_{x_0})|) \\
&\leq C(\rho^\alpha |u|_{C^{1,\alpha}(\tilde{T})} + \|\partial u(x_1) - \partial u\|_{L_\infty(\tilde{T})} + |(u - u_h, \partial\delta_{x_0})|) \\
&\leq C(\rho^\alpha |u|_{C^{1,\alpha}(\Omega)} + |(u - u_h, \partial\delta_{x_0})|) \\
&\leq C(\underline{h}^{\alpha\beta} |u|_{C^{1,\alpha}(\Omega)} + |(u - u_h, \partial\delta_{x_0})|).
\end{aligned} \tag{3.7}$$

We next introduce a discrete Green's function. With ∂ and x_0 as above, we define $g^{x_0} \in H_0^1(\Omega)$ as the unique function satisfying

$$A(v, g^{x_0}) = (\partial\delta_{x_0}, v) \tag{3.8}$$

for all $v \in H_0^1(\Omega)$. Since the bound given below does not depend upon x_0, we shall suppress the dependence of g and δ on x_0 for the rest of this section. The heart of our proof consists of proving the following bound for $\|g\|_{W_1^1(\Omega)}$.

LEMMA 3.8. *If the conditions of Theorem* 3.1 *are satisfied, then*

$$\|g\|_{W_1^1(\Omega)} \leq C_g \beta^{\gamma(n)} \ell_{\underline{h},n}, \tag{3.9}$$

where $C_g = C(C_G, \lambda, \Lambda)$.

In order to complete the proof of Theorem 3.1 given Lemma 3.8, we use (3.8), (2.10), and (2.11) to find that

$$\begin{aligned}
(u - u_h, \partial\delta) &= A(u - u_h, g) = (A - A_h)(u - u_h, g) + A_h(u - u_h, g) \\
&= (A - A_h)(u - u_h, g) + A_h(u - u_h, g - I_h g) \\
&\leq C_F \|\nabla(u - u_h)\|_{L_\infty(\Omega)}^2 \|g\|_{W_1^1(\Omega)} \\
&\quad + \left| \int_\Omega \sum_{i=1}^n (F_i(x, \nabla u) - F_i(x, \nabla u_h))(g - I_h g)_{x_i} \, \mathrm{d}x \right|.
\end{aligned} \tag{3.10}$$

Note that $A_h = A$ if (1.2) is a linear problem, so the term $\|\nabla(u - u_h)\|_{L_\infty(\Omega)}^2$ is dropped in this case as claimed. We use the easily-proven scaled trace inequality

$$\|v\|_{L_1(\partial T)} \leq C(h_T^{-1}\|v\|_{L_1(T)} + \|\nabla v\|_{L_1(T)})$$

and integrate the last term in (3.10) by parts elementwise to find

$$
\begin{aligned}
&\left|\int_\Omega \textstyle\sum_{i=1}^n (F_i(x,\nabla u)-F_i(x,\nabla u_h))(g-I_hg)_{x_i}\,\mathrm{d}x\right| \\
&\quad=\left|\sum_{T\in\mathcal{T}}\int_T\sum_{i=1}^n (F_i(x,\nabla u)-F_i(x,\nabla u_h))(g-I_hg)_{x_i}\,\mathrm{d}x\right| \\
&\quad=\left|\sum_{T\in\mathcal{T}}\int_T(-\sum_{i=1}^n\frac{\partial}{\partial x_i}(F_i(x,\nabla u)-F_i(x,\nabla u_h))(g-I_hg)\,\mathrm{d}x\right. \\
&\qquad\left.+\int_{\partial T}\sum_{i=1}^n (F_i(x,\nabla u)-F_i(x,\nabla u_h))n_i(g-I_hg)\,\mathrm{d}s\right| \qquad (3.11)\\
&\quad\le\sum_{T\in\mathcal{T}}\|f+\sum_{i=1}^n\frac{\partial}{\partial x_i}F_i(\cdot,\nabla u_h)\|_{L_\infty(T)}\|g-I_hg\|_{L_1(T)} \\
&\qquad+\|[u_h]\|_{L_\infty(\partial T)}\|g-I_hg\|_{L_1(\partial T)} \\
&\quad\le\sum_{T\in\mathcal{T}}\left\|f+\sum_{i=1}^n\frac{\partial}{\partial x_i}F_i(\cdot,\nabla u_h)\right\|_{L_\infty(T)}\|g-I_hg\|_{L_1(T)} \\
&\qquad+\|[u_h]\|_{L_\infty(\partial T)}(h_T^{-1}\|g-I_hg\|_{L_1(T)}+\|\nabla(g-I_hg)\|_{L_1(T)}).
\end{aligned}
$$

Finally, we apply the approximation results (2.5) and (2.6) and thus obtain

$$
\begin{aligned}
&\sum_{T\in\mathcal{T}}\left\|f+\textstyle\sum_{i=1}^n\frac{\partial}{\partial x_i}F_i(\nabla u_h)\|_{L_\infty(T)}\|g-I_hg\right\|_{L_1(T)} \\
&\qquad+\|[u_h]\|_{L_\infty(\partial T)}(h_T^{-1}\|g-I_hg\|_{L_1(T)}+\|\nabla(g-I_hg)\|_{L_1(T)}) \qquad (3.12)\\
&\quad\le\sum_{T\in\mathcal{T}}\mathcal{E}_T\|g\|_{W_1^1(P_T)} \\
&\quad\le C\|g\|_{W_1^1(\Omega)}\max_{T\in\mathcal{T}}\mathcal{E}_T.
\end{aligned}
$$

Combining (3.12), (3.11), (3.10), and (3.7) and finally applying (3.9) completes the proof of Theorem 3.1 assuming Lemma 3.8.

To begin the proof of Lemma 3.8, we first note the elementary inequality

$$
\|g\|_{H^1(\Omega)}\le\frac{C(\Omega)}{\lambda}\|\delta\|_{L_2(\Omega)}. \qquad (3.13)
$$

In order to prove (3.9) for $n\ge 3$, we then note that if $|x-x_0|>2\rho$, we may apply (2.13) and (3.6) to find that

$$
\begin{aligned}
|\nabla_x g(x)|&=\left|\int_{\mathrm{supp}(\delta)}\nabla_x G(x,y)\partial\delta(y)\,\mathrm{d}y\right| \\
&=\left|\int_{\mathrm{supp}(\delta)}\partial_y\nabla_x G(x,y)\delta\,\mathrm{d}y\right| \qquad (3.14)\\
&\le C_G|x_0-y|^{-n}\|\delta\|_{L_1(\Omega)}\le C_G|x_0-x|^{-n},
\end{aligned}
$$

and similarly,

$$|g(x)| \le C_G |x_0 - x|^{1-n}. \tag{3.15}$$

We then use (3.13), (3.6), (3.14), and (3.15) to compute

$$\begin{aligned}
\|g\|_{W_1^1(\Omega)} &\le C\rho^{n/2}\|g\|_{H^1(B_{3\rho(x_0)})} + \|g\|_{W_1^1(\Omega\setminus B_{3\rho}(x_0))} \\
&\le \rho^{n/2}\frac{C(\Omega)}{\lambda}\|\delta\|_{L_2(\Omega)} + C_G\int_{C\rho}^{\operatorname{diam}(\Omega)} r^{-n}r^{n-1}dr \\
&\le \frac{C(\Omega)}{\lambda}\rho^{n/2}\rho^{-n/2} + C_G\log(1/\rho) \\
&\le C\left(\frac{1}{\lambda} + C_G\right)\beta\log(1/\underline{h}).
\end{aligned}$$

If $n = 2$, we must apply (2.14) instead of (2.13). Since we always apply (2.14) with $|x - y| \ge C\rho$, this results in an extra factor of $\log(1/\rho)$ in our estimates. ∎

Proof of Corollary 3.5. Letting $\delta_{x_0} = \delta$ be as above, it is easy to compute that for some $x_0 \in \Omega$, $\|u - u_h\|_{L_\infty(\Omega)} \le C\underline{h}^{\alpha\beta}|u|_{C^\alpha(\overline{\Omega})} + |(u - u_h, \delta)|$. Let then $\overline{g} \in H_0^1(\Omega)$ satisfy $A(v, \overline{g}) = (v, \delta)$ for all $v \in H_0^1(\Omega)$. Then

$$\begin{aligned}
|(u - u_h, \delta)| &= |A(u - u_h, \overline{g})| \le |(A - A_h)(u - u_h, \overline{g})| + |A_h(u - u_h, \overline{g} - I_h\overline{g})| \\
&\le C(\|\nabla(u - u_h)\|_{L_\infty(\Omega)}^2 + \max_{T\in\mathcal{T}} h_T\mathcal{E}_T)\|\overline{g}\|_{W_1^2(\Omega)}.
\end{aligned} \tag{3.16}$$

Inserting the bounds established in Theorem 3.1 of [Noc95] ($n = 2$) and Corollary 2.3 of [DDP00] ($n = 3$) for $\|\overline{g}\|_{W_1^2(\Omega)}$ and also (3.1) into (3.16) yields (3.4). ∎

3.4. Proof of efficiency. We follow here the local argument given in [Ver89] and adapted to the maximum norm case in [Noc95] and [DDP00]. Recalling the definition of $\tilde{f}_h$ from Theorem 3.6, we note first that for any $v \in H_0^1(\Omega)$,

$$\sum_{T\in\mathcal{T}}\int_T \tilde{f}_h v\,\mathrm{d}x + \frac{1}{2}\int_{\partial T}[u_h]v\,\mathrm{d}s = A_h(u - u_h, v). \tag{3.17}$$

Now fix an element T and choose $v = b_T$, where b_T is the polynomial bubble function of degree $n + 1$ which is obtained by multiplying the barycentric coordinates and scaling so that b_T is 1 at the barycenter of T. By transforming from a reference element, we see that $\int_T b_T\,\mathrm{d}x = Ch_T^n$, $\|b_T\|_{L_1(T)} \le Ch_T^n$, and $\|\nabla b_T\|_{L_1(T)} \le Ch_T^{n-1}$. Since $b_T = 0$ on ∂T, we may thus compute from (3.17) that

$$\begin{aligned}
\tilde{C}_1 h_T^{n-1}\,&\|\nabla(u - u_h)\|_{L_\infty(T)} \ge |A_h(u - u_h, b_T)| \\
&= \left|\int_T \tilde{f}_h b_T\,\mathrm{d}x + \frac{1}{2}\int_{\partial T}[u_h]b_T\,\mathrm{d}s\right| = \left|\int_T \tilde{f}_h b_T\,\mathrm{d}x\right| \\
&= \left|\int_T (\tilde{f}_h - P_h\tilde{f}_h)b_T\,\mathrm{d}x + P_h\tilde{f}_h|_T\int_T b_T\,\mathrm{d}x\right| \\
&\ge Ch_T^n|P_h\tilde{f}_h|_T| - C\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(T)} \\
&\ge Ch_T^n(|\|\tilde{f}_h\|_{L_\infty(T)} - \|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(T)}| - C\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(T)}) \\
&\ge Ch_T^n(\|\tilde{f}_h\|_{L_\infty(T)} - C\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(T)})
\end{aligned}$$

so that

$$h_T\|\tilde{f}_h\|_{L_\infty(T)} \le \tilde{C}_1\|\nabla(u - u_h)\|_{L_\infty(T)} + Ch_T\|\tilde{f}_h - P_h\tilde{f}_h\|_{L_\infty(T)}. \tag{3.18}$$

Next let $S = \overline{T} \cap \overline{T'}$ be a face of T not contained in $\partial\Omega$. We then define q_S to be the continuous piecewise polynomial of degree n which is 0 on $\partial(T \cup T')$ and 1 at the barycenter of S. Note that $\|q_S\|_{L_1(T\cup T')} \leq Ch_T^n$, $\|\nabla q_S\|_{L_1(T\cup T')} \leq Ch_T^{n-1}$, $\|q_S\|_{L_1(S)} \leq Ch_T^{n-1}$, and $\int_S q_S \mathrm{d}s = Ch_T^{n-1}$. Again computing using (3.17), we find that

$$\begin{aligned}\tilde{C}_1 h_T^{n-1}\|\nabla(u-u_h)\|_{L_\infty(T\cup T')} &\geq A_h(u-u_h, q_S)\\ &= \int_T \tilde{f}_h q_S \,\mathrm{d}x + \int_S ([u_h] - \tilde{P}_h[u_h])q_S \mathrm{d}s + \int_S \tilde{P}_h[u_h] q_S \mathrm{d}s\\ &\geq Ch_T^{n-1}\|[u_h]\|_{L_\infty(S)}\\ &\quad - C(h_T^n\|\tilde{f}_h\|_{L_\infty(T\cup T')} + h_T^{n-1}\|[u_h] - \tilde{P}_h[u_h]\|_{L_\infty(S)})\end{aligned}$$

and

$$\begin{aligned}\|[u_h]\|_{L_\infty(S)} &\leq \tilde{C}_1\|\nabla(u-u_h)\|_{L_\infty(P_T)}\\ &\quad + Ch_T\|\tilde{f}_h\|_{L_\infty(P_T)} + C\|[u_h] - \tilde{P}_h[u_h]\|_{L_\infty(S)}.\end{aligned} \tag{3.19}$$

Recalling that $\mathcal{E}_T = h_T\|\tilde{f}_h\|_{L_\infty(T)} + \|[u_h]\|_{L_\infty(\partial T)}$ and combining (3.18) with (3.19) completes the proof of (3.5). ∎

4. Localized estimates.

4.1. Reliability of localized estimators. We first give an a posteriori bound for $\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}$.

THEOREM 4.1. *Let $D \subset \Omega$. In addition to the assumptions of section 2, assume that $u \in C^{1,\nu}(\overline{\Omega})$ for some $0 < \nu \leq 1$ if the coefficients F_i are linear and $u \in W^2_\infty(\Omega)$ if the coefficients F_i are nonlinear. Then for any $0 < \alpha \leq \nu$ and any $\beta \geq 1$,*

$$\begin{aligned}\|\nabla(u-u_h)\|_{L_\infty(D)} &\leq \|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}\\ &\leq \beta^{\gamma(n)}\ell_{\underline{h},n}[C_2 \max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T + C_1C_F\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}]\\ &\quad + C\underline{h}^{\alpha\beta}\max_{T\in\mathcal{T}}\sigma_D(T)|u|_{C^{1,\alpha}(\overline{T})}.\end{aligned} \tag{4.1}$$

Here C_1 is as in Theorem 3.1 and C_2 depends on C_G, $\|a_{ij}\|_{W^1_\infty(\Omega)}$, λ, Λ, and the shape regularity of $\mathcal{T}$. In the nonlinear case, C_2 thus depends on $\|u\|_{W^2_\infty(\Omega)}$ and the coefficients F_i. In the linear case, C_2 does not depend on u and the term $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ in (4.1) does not appear.

Note that the term $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ in (4.1) is not generally of higher order, in contrast to the situation which arises when the global estimate (3.1) is applied. One may insert (3.3) into (4.1) in the nonlinear case to yield the following asymptotic reliability result.

COROLLARY 4.2. *Assume that $u \in W^2_\infty(\Omega)$, $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ and $\underline{h}$ are small enough and $\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)} \geq C\underline{h}^\epsilon \max_{T\in\mathcal{T}}\sigma_D(T)|u|_{C^{1,\nu}(\overline{T})}$ for some positive ν and ϵ. Then*

$$\begin{aligned}&\|\nabla(u-u_h)\|_{L_\infty(D)}\\ &\leq C_2\left(\frac{1+\epsilon}{\nu}\right)^{\gamma(n)}\ell_{\underline{h},n}\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T + C_1^3C_F\left(\frac{1+\epsilon}{\nu}\right)^{2\gamma(n)}\ell^3_{\underline{h},n}\max_{T\in\mathcal{T}}\mathcal{E}_T^2\\ &\leq C(C_1,C_2,\epsilon)\ell_{\underline{h}}\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T + \mathcal{E}_T^2).\end{aligned} \tag{4.2}$$

Here C_1 and C_2 are as in Theorems 3.1 and 4.1, $\ell_{\underline{h}}$ is a generic logarithmic factor, and the term $\mathcal{E}_T^2$ may be dropped in the linear case.

As in the case of our global estimator, the constants C_1 and C_2 potentially make it difficult to apply (4.2) efficiently and accurately as an error estimator in the nonlinear case. Even if we only wish to apply (4.2) as an error indicator, it likely will be necessary in most situations to gain some knowledge of the relative sizes of C_1 and C_2 as the terms in (4.2) could be weighted improperly otherwise. As stated in the introduction, a purely theoretical determination of these constants appears difficult, and their investigation is the subject of ongoing work.

One application of (4.2) is the computation of a gradient at a single point $x_0 \in \Omega$ to within a given tolerance without requiring that ∇u_h approximate ∇u to the same tolerance globally (as would be the case if a global estimator were used). Here the localized estimate (4.2) is an alternative to the "dual weighted residual" approach of [BR01], which in this case essentially involves computing a finite element approximation to the discrete Green's function g^{x_0} and inserting this approximation (using appropriate methods such as difference quotients to approximate second derivatives) into the appropriate residual equation, which for linear problems is

$$|\partial(u-u_h)(x_0)| \le C \sum_{T\in\mathcal{T}} h_T \mathcal{E}_T |g^{x_0}|_{W_1^2(T)}.$$

Note that our localized analysis essentially involves bounding $|g^{x_0}|_{W_1^2(T)}$ a priori instead of a posteriori as in the dual residual method. Since more of the work is done ahead of time, so to speak, localized estimators may be applied more easily and over larger subdomains than dual estimators, but potentially at the expense of some sharpness and unknown constants as compared with the dual weighted residual method. The advantages of localized estimators are their lower computational cost (the local nature of the discrete Green's function is employed a priori instead of being computed a posteriori) and the fact that they can easily be applied over larger subdomains.

4.2. Efficiency of localized estimators. We shall show that our localized estimator is efficient (up to higher-order terms) in the linear case and in a certain sense also in the nonlinear case.

THEOREM 4.3. *Under the same conditions as are assumed in Theorem* 3.6,

$$\begin{aligned}\sigma_D(T)\mathcal{E}_T \le{}& \tilde{C}_1 \|\sigma_D \nabla(u-u_h)\|_{L_\infty(P_T)} + C\|h\sigma_D(\tilde{f}_h - P_h\tilde{f}_h)\|_{L_\infty(P_T)} \\ &+ C\|\sigma_D([u_h]-\tilde{P}[u_h])\|_{L_\infty(\partial T)}.\end{aligned} \tag{4.3}$$

Here $\tilde{C}_1 = \|a_{ij}^h\|_{L_\infty(\Omega)}$, *and* C *only depends on* $\mathcal{T}$.

Proof. To prove (4.3), we simply distribute the weight $\sigma_D(T)$ through (3.5) while noting that h and σ_D are always equivalent on adjacent elements (and in particular on P_T). ∎

Remark 4.4. In the linear case, (4.3) establishes immediately that up to higher-order terms,

$$\frac{1}{\tilde{C}_1} \max_{T\in\mathcal{T}} \sigma_D(T)\mathcal{E}_T \le \|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)} \le C_2 \beta^{\gamma(n)} \ell_{\underline{h},n} \max_{T\in\mathcal{T}} \sigma_D(T)\mathcal{E}_T.$$

In the nonlinear case, the perturbation term $\max_{T\in\mathcal{T}} \mathcal{E}_T^2$ "morally" should behave as $\|h_T\nabla(u-u_h)\|_{L_\infty(\Omega)}$, which is bounded by $\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}$. However, one would have to resort to a priori estimates to prove such a statement. Instead, we combine the global reliability and efficiency estimates (3.1) and (3.5) with the localized estimates (4.2) and (4.3) while consolidating constants and ignoring higher-order terms to yield

the estimate

$$\begin{aligned}\frac{1}{\overline{C}}\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T+\mathcal{E}_T^2)&\le\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}+\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}\\&\le\overline{C}\ell_{\underline{h}}\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T+\mathcal{E}_T^2).\end{aligned}$$

Thus the estimator $\max_{T\in\mathcal{T}}(\sigma_D(T)\mathcal{E}_T+\mathcal{E}_T^2)$ reliably and efficiently estimates the quantity $\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}+\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$ instead of just the weighted norm $\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}$ as originally intended.

4.3. Proof of Theorem 4.1. First we assume that D is a single point $x_0\in\Omega$. We begin by picking a point $x_1\in\Omega$ and a first-order directional derivative ∂ such that $\|\sigma_{x_0}\nabla(u-u_h)\|_{L_\infty(\Omega)}\le C\sigma_{x_0}(x_1)|\partial(u-u_h)(x_1)|$. Here we have abused notation slightly by letting $\sigma_{x_0}(x_1)=\sigma_{x_0}(T_{x_1})$, where T_{x_1} is any element with $x_1\in\overline{T}_{x_1}$. Proceeding as in (3.7) while noting that $\|\sigma_{x_0}\|_{L_\infty(\Omega)}=1$, we obtain

$$\begin{aligned}\|\sigma_{x_0}\nabla(u-u_h)\|_{L_\infty(\Omega)}&\le C\sigma_{x_0}(x_1)|\partial(u-u_h)(x_1)|\\&\le C\sigma_{x_0}(x_1)(|(u-u_h,\partial\delta_{x_1})|+\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{T_{x_1}})}).\end{aligned}\tag{4.4}$$

We now compute as in (3.10) and (3.11) that

$$\begin{aligned}|(u-u_h,\partial\delta_{x_1})|&\le|(A-A_h)(u-u_h,g^{x_1})|+|A_h(u-u_h,g^{x_1}-I_hg^{x_1})|\\&\le C_F\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}\|g^{x_1}\|_{W_1^1(\Omega)}\\&\quad+\sum_{T\in\mathcal{T}}\Bigg[\left\|f+\sum_{i=1}^n\frac{\partial}{\partial x_i}F_i(x,\nabla u_h)\right\|_{L_\infty(T)}\|g^{x_1}-I_hg^{x_1}\|_{L_1(T)}\\&\quad+\|[u_h]\|_{L_\infty(\partial T)}(h_T^{-1}\|g^{x_1}-I_hg^{x_1}\|_{L_1(T)}+\|\nabla(g^{x_1}-I_hg^{x_1})\|_{L_1(T)})\Bigg].\end{aligned}\tag{4.5}$$

Next we note that by shape regularity, the elements in P''_T are quasi-uniform. Also, the weight σ_{x_1} is equivalent to 1 on $P''_{T_{x_1}}$ and is always equivalent on adjacent elements. Using these facts, we then apply (2.2) along with (2.5) and (2.6) to obtain

$$\begin{aligned}&\sum_{T\in\mathcal{T}}\left\|f+\sum_{i=1}^n\frac{\partial}{\partial x_i}F_i(\cdot,\nabla u_h)\right\|_{L_\infty(T)}\|g^{x_1}-I_hg^{x_1}\|_{L_1(T)}\\&\quad+\|[u_h]\|_{L_\infty(\partial T)}(h_T^{-1}\|g^{x_1}-I_hg^{x_1}\|_{L_1(T)}+\|\nabla(g^{x_1}-I_hg^{x_1})\|_{L_1(T)})\\&\le\sum_{T\subset P''_{T_{x_1}}}\mathcal{E}_T\|g^{x_1}\|_{W_1^1(P_T)}\\&\quad+\sum_{T\cap P''_{T_{x_1}}=\emptyset}\frac{h_T}{\operatorname{dist}(x_1,T)+h_T}\mathcal{E}_T(\operatorname{dist}(x_1,T)+h_T)|g^{x_1}|_{W_1^2(P_T)}\\&\le C(\|g^{x_1}\|_{W_1^1(\Omega)}+\sum_{T\cap P'_{T_{x_1}}=\emptyset}(\operatorname{dist}(x_1,T)+h_T)|g^{x_1}|_{W_1^2(T)})\\&\quad\cdot\max_{T\in\mathcal{T}}\sigma_{x_1}(T)\mathcal{E}_T\\&\le C(\|g^{x_1}\|_{W_1^1(\Omega)}+\int_{\Omega\setminus P'_{T_{x_1}}}|x-x_1||D^2g^{x_1}|\,\mathrm{d}x)\max_{T\in\mathcal{T}}\sigma_{x_1}(T)\mathcal{E}_T.\end{aligned}\tag{4.6}$$

We next state the fundamental lemma in the proof of our localized estimate.

LEMMA 4.5. *If the conditions of Theorem* 4.1 *are satisfied, then for any* $x_1 \in \Omega$,

$$\int_\Omega |x - x_1||D^2 g^{x_1}(y)|\,\mathrm{d}x \le C_g' \beta^{\gamma(n)} \ell_{\underline{h},n}, \tag{4.7}$$

where C_g' *depends on* λ, Λ, $\|a_{ij}\|_{W^1_\infty(\Omega)}$, *and the constant* C_g *from Lemma* 3.8.

Assuming (4.7), we apply (3.9) and combine (4.4), (4.5), and (4.6) to find that

$$\begin{aligned}\|\sigma_{x_0}\nabla(u-u_h)\|_{L_\infty(\Omega)} \le \beta^{\gamma(n)}\ell_{\underline{h},n}[C_g' \max_{T\in\mathcal{T}} \sigma_{x_0}(T_{x_1})\sigma_{x_1}(T)\mathcal{E}_T \\ + C_F C_g \|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}] + C\max_{T\in\mathcal{T}}\sigma_D(T)|u|_{C^{1,\alpha}(\overline{T})}.\end{aligned} \tag{4.8}$$

In order to complete our proof, we thus must show that for $T \in \mathcal{T}$,

$$\sigma_{x_0}(x_1)\sigma_{x_1}(T) \le \sigma_{x_0}(T). \tag{4.9}$$

We shall compute with the weight $\frac{h(x)}{|x_0-x|+h(x)}$, which is equivalent to but more convenient than $\sigma_{x_0}(x)$. Thus for any $T \in \mathcal{T}$ and $x_2 \in T$,

$$\begin{aligned}\sigma_{x_0}(x_1)\sigma_{x_1}(T) &\le C\frac{h(x_1)h(x_2)}{(|x_0-x_1|+h(x_1))(|x_1-x_2|+h(x_2))} \\ &= C\frac{h(x_2)}{|x_0-x_2|+h(x_2)}\frac{h(x_1)(|x_0-x_2|+h(x_2))}{(|x_0-x_1|+h(x_1))(|x_1-x_2|+h(x_2))} \\ &\le C\sigma_{x_0}(x_2)\frac{h(x_1)(|x_0-x_2|+h(x_2))}{(|x_0-x_1|+h(x_1))(|x_1-x_2|+h(x_2))}.\end{aligned} \tag{4.10}$$

Using the triangle inequality and noting from (2.2) that $h(x_1) \le C(|x_1-x_2|+h(x_2))$, we next compute that

$$\begin{aligned}&h(x_1)(|x_0-x_2|+h(x_2)) \\ &\quad\le h(x_1)(|x_0-x_1|+|x_1-x_2|+h(x_2)) \\ &\quad= h(x_1)|x_0-x_1| + h(x_1)(|x_1-x_2|+h(x_2)) \\ &\quad\le C(|x_1-x_2|+h(x_2))|x_0-x_1| + h(x_1)(|x_1-x_2|+h(x_2)).\end{aligned} \tag{4.11}$$

Noting that the expression above is bounded by C times the denominator of (4.10), we combine (4.10) and (4.11) to obtain (4.9). Inserting (4.9) into (4.8) completes the proof of (4.1) for $D = x_0$ assuming Lemma 4.5. Taking the maximum of (4.1) over $x_0 \in D$ while recalling (4.9) completes the proof of (4.1) for arbitrary $D \subset \Omega$.

In order to prove Lemma 4.5, we shall need the linear H^2_2 regularity result

$$\|g\|_{H^2_2(\Omega)} \le C_{reg}\|\partial\delta\|_{L_2(\Omega}, \tag{4.12}$$

where $C_{reg} = C(\lambda, \Lambda, \|a_{ij}\|_{W^1_\infty(\Omega)})$. This result is standard for smooth domains and may be found in [Gr85] for convex (including convex polyhedral) domains. Here and in what follows we suppress the dependence of g and δ on x_1.

We now decompose Ω into dyadic annuli. Let $\Omega_0 = B_{3\rho}(x_1)$, so that according to our definitions $\mathrm{dist}(supp(\delta), \partial\Omega_0) > C\rho$. We then define $d_j = 2^j 3\rho$, $j = 0, \dots, N$, $\tilde{\Omega}_j = \{x \in \mathbb{R}^n \text{ such that } d_{j-1} \le |x - x_1| \le d_j\}$, $\Omega_j = \tilde{\Omega}_j \cap \Omega$, and $\Omega_j' = \Omega_{j-1} \cup \Omega_j \cup \Omega_{j+1}$. Note that $\Omega = \bigcup_{j=0}^N \Omega_j$ with $N \le C\log(1/\rho)$. Finally, we let $\omega_j \in C_0^\infty(\tilde{\Omega}_{j-1} \cup \tilde{\Omega}_j \cup \tilde{\Omega}_{j+1})$ be a cutoff function which is 1 on $\tilde{\Omega}_j$ and which satisfies

$\|\omega_j\|_{W^k_\infty(\Omega)} \leq C d_j^{-k}$, $k = 0, 1, 2$. Then

$$\begin{aligned}\int_\Omega |x - x_1||D^2 g^{x_1}|\,\mathrm{d}x &\leq C d_1^{n/2+1}\|D^2 g\|_{L_2(\Omega_1')} + \sum_{i=2}^N d_j \|D^2 g\|_{L_1(\Omega_j)} \\ &\leq C_{reg}\rho^{n/2+1}\|\partial\delta\|_{L_2(\Omega)} + \sum_{i=2}^N d_j \|D^2(\omega_j g)\|_{L_1(\Omega)} \\ &\leq C_{reg} + \sum_{i=2}^N d_j \|D^2(\omega_j g)\|_{L_1(\Omega)}.\end{aligned} \tag{4.13}$$

Abusing notation slightly by letting A denote the matrix of coefficients $[a_{ij}]$, we compute that for $v \in H_0^1(\Omega)$ and $j > 1$,

$$\begin{aligned}A(v, \omega_j g) &= (v, -\operatorname{div}(A^*\nabla(\omega_j g))) \\ &= (v, -\operatorname{div}(A^*(g\nabla\omega_j + \omega_j \nabla g))) \\ &= (v, -g\operatorname{div}(A^*\nabla\omega_j) - A^*\nabla\omega_j \cdot \nabla g - A^*\nabla g \cdot \nabla\omega_j - \omega_j \operatorname{div}(A^*\nabla g)) \\ &= (v, -g\operatorname{div}(A^*\nabla\omega_j) - A^*\nabla\omega_j \cdot \nabla g - A^*\nabla g \cdot \nabla\omega_j - \omega_j\delta) \\ &= (v, -g\operatorname{div}(A^*\nabla\omega_j) - A^*\nabla\omega_j \cdot \nabla g - A^*\nabla g \cdot \nabla\omega_j)\end{aligned}$$

since ω_j and δ have disjoint support for $j > 1$. Then applying the regularity result (4.12) to $\omega_j g$, we find that

$$\|D^2(\omega_j g)\|_{L_2(\Omega)} \leq C_{reg}\|a_{ij}\|_{W^1_\infty(\Omega)} \left(\frac{1}{d_j^2}\|g\|_{L_2(\Omega_j')} + \frac{1}{d_j}\|\nabla g\|_{L_2(\Omega_j')}\right). \tag{4.14}$$

We then insert (4.14) into (4.13) while recalling (3.14) and (3.15) to find that for $n \geq 3$,

(4.15)

$$\begin{aligned}\int_\Omega |x - x_1|\,|D^2 g^{x_1}|\,\mathrm{d}x &\leq C(C_{reg} + \sum_{i=1}^N d_j^{n/2+1}\left(\frac{1}{d_j^2}\|g\|_{L_2(\Omega_j)} + \frac{1}{d_j}\|\nabla g\|_{L_2(\Omega_j)})\right) \\ &\leq C_{reg}\|a_{ij}\|_{W^1_\infty(\Omega)}\left(1 + \sum_{i=1}^N (d_j^{n-1}\|g\|_{L_\infty(\Omega_j)} + d_j^n\|\nabla g\|_{L_\infty(\Omega_j)})\right) \\ &\leq C_{reg}\|a_{ij}\|_{W^1_\infty(\Omega)} C_G\left(1 + \sum_{i=1}^N (d_j^{n-1}d_j^{1-n} + d_j^n d_j^{-n})\right) \\ &\leq C_{reg}\|a_{ij}\|_{W^1_\infty(\Omega)} C_G(1 + \log(1/\rho)) \\ &\leq C_{reg}\|a_{ij}\|_{W^1_\infty(\Omega)} C_G \beta \log(1/\underline{h}).\end{aligned}$$

When $n = 2$, an extra factor of $\log(1/\rho)$ enters the estimate (4.15) as before. Thus the proof of Lemma 4.5 is completed. □

5. Extension of results to the general quasilinear equation (1.1). In this section we outline the steps necessary to extend our results to operators of the form (1.1). We first consider the treatment of nonhomogeneous Dirichlet conditions in a model problem of the form (1.2), then we consider general operators of the form (1.1) with homogeneous boundary conditions.

5.1. Nonhomogeneous boundary conditions. We consider here the model problem (1.2), but now with the more general Dirichlet boundary condition $u = b$ on $\partial\Omega$ for some $b \in W^1_\infty(\Omega)$. We also assume that b_h is a piecewise linear finite element approximation to b and that u_h with $u_h - b_h \in S_h$ solves (2.1). The following is a corollary to Theorem 3.1 and Theorem 4.1.

COROLLARY 5.1. *Under the conditions of Theorem* 3.1*,*

$$\begin{aligned}\|\nabla(u-u_h)\|_{L_\infty(\Omega)} \le C_1\beta^{\gamma(n)}\ell_{\underline{h},n}[\max_{T\in\mathcal{T}}\mathcal{E}_T + C_F\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)} \\ + \|\nabla(b-b_h)\|_{L_\infty(\Omega)}] + C\underline{h}^{\alpha\beta}|u|_{C^{1,\alpha}(\overline{\Omega})}.\end{aligned} \tag{5.1}$$

Under the conditions of Theorem 4.1*,*

$$\begin{aligned}&\|\nabla(u-u_h)\|_{L_\infty(D)} \\ &\quad\le \beta^{\gamma(n)}\ell_{\underline{h},n}[C_2\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T + C_1C_F\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}] \\ &\qquad + C_1\|(b-b_h)[dist(\cdot,D)+\underline{h}^\beta]^{-n}\|_{L_1(\partial\Omega)} \\ &\qquad + C\underline{h}^{\alpha\beta}\max_{T\in\mathcal{T}}\sigma_D(T)|u|_{C^{1,\alpha}(\overline{T})}.\end{aligned} \tag{5.2}$$

Sketch of Proof. We procede as in (3.7) through (3.10), then let $\vec{n}$ be the outward normal on $\partial\Omega$ and compute that for δ and g defined with respect to the point x_0,

$$(u-u_h,\partial\delta) = A(u-u_h,g) - \int_{\partial\Omega}(b-b_h)(A\nabla g\cdot\vec{n})\,\mathrm{d}\sigma. \tag{5.3}$$

To prove (5.1), we bound $A(u-u_h,g)$ precisely as before and compute

$$\begin{aligned}\left|\int_{\partial\Omega}(b-b_h)(A\nabla g\cdot\vec{n})\,\mathrm{d}\sigma\right| &= |(b-b_h,\partial\delta) - A(b-b_h,g)| \\ &\le |(\partial(b-b_h),\delta)| + \Lambda\|\nabla(b-b_h)\|_{L_\infty(\Omega)}\|\nabla g\|_{L_1(\Omega)}.\end{aligned} \tag{5.4}$$

and then apply (3.6) with $p=1$ and $k=1$ along with Lemma 3.8.

In order to prove (5.2), we let $x_0 \in D$ be such that $\|\nabla(u-u_h)\|_{L_\infty(D)} = |\nabla(u-u_h)(x_0)|$. Recall that δ_{x_0} may always be defined so that $dist(\mathrm{supp}(\delta),\partial\Omega) \ge c\rho$. A calculation similar to (3.14) then yields $|(A\nabla g\cdot\vec{n})(y)| \le C_1[|\rho+|x_0-y|]^{-n}$ for $y\in\partial\Omega$. Inserting this inequality into (5.3), recalling that $\rho = \underline{h}^\beta$, and bounding $A(u-u_h,g)$ as in (4.5) and following completes the proof. ∎

In [DR98] an a posteriori energy-norm bound is given which treats Dirichlet data in a fashion similar to (5.1). The term $\|(b-b_h)[dist(\cdot,D)+\underline{h}^\beta]^{-n}\|_{L_1(\partial\Omega)} \le Ch^{-1}\|b-b_h\|_{L_\infty(\partial\Omega)}$ in (5.2) is very similar to one appearing in the a priori estimates given in Theorem A.1 of [BTW03]. One may easily compute that

$$\|(b-b_h)[dist(\cdot,D)+\underline{h}^\beta]^{-n}\|_{L_1(\partial\Omega)} \le C\min(dist(D,\partial\Omega)^{-1},\underline{h}^{-\beta})\|b-b_h\|_{L_\infty(\partial\Omega)}. \tag{5.5}$$

If $D \subset\subset \Omega$ this term is thus of higher order, reflecting the localization of the error to D. If D abuts $\partial\Omega$, however, the term $\underline{h}^\beta$ leads to suboptimality if $\beta > 1$. (Note that this problem is not encountered in the a priori estimates of [BTW03] on quasi-uniform meshes, where $\underline{h}^\beta$ may be replaced by the mesh size h.) One may in this case instead estimate the error in approximating the Dirichlet data by $\|\nabla(b-b_h)\|_{L_\infty(\Omega)}$ as in (5.1), but this estimate does not reflect the more local nature of the error. Thus (5.2) could likely be improved, although it appears difficult to do so using the present techniques.

5.2. Theoretical comments on more general operators. We assume that u solves (1.1) with $u = 0$ on $\partial\Omega$. As before, we assume Ω to be convex and polygonal. We also assume that $u_h \in S_h$ satisfies

$$\int_\Omega \sum_{i=1}^n F_i(x, u_h, \nabla u_h) v_{h,x_i} + F_0(x, u_h, \nabla u_h) v_h \, \mathrm{d}x = 0, \ v_h \in S_h.$$

The essential linear auxiliary operators A and A_h defined in section 2.2 may be easily modified to aid in the analysis of problems of the form (1.1). Letting $F_{j0} = \frac{\partial}{\partial z} F_j(x, z, p)$ and $v_{x_0} = v$, we have for $i = 0, \ldots, n$ that

$$\begin{aligned} &F_i(x, u, \nabla u) - F_i(x, u_h, \nabla u_h) \\ &\quad = \sum_{j=0}^n \int_0^1 F_{ij}(x, u_h + t(u - u_h), \nabla u_h + t\nabla(u - u_h))(u - u_h)_{x_j} \, \mathrm{d}t. \end{aligned}$$

For $0 \le i, j \le n$, we then make the definitions

$$\begin{aligned} a_{ij}^h(x) &= \int_0^1 F_{ij}(x, u_h + t(u - u_h), \nabla(u_h + t\nabla(u - u_h)) \, \mathrm{d}t, \\ a_{ij}(x) &= F_{ij}(x, u, \nabla u), \\ A_h(v, w) &= \int_\Omega \sum_{i,j=0}^n a_{ij}^h(x) v_{x_j} w_{x_i} \, \mathrm{d}x, \\ A(v, w) &= \int_\Omega \sum_{i,j=0}^n a_{ij}(x) v_{x_j} w_{x_i} \, \mathrm{d}x. \end{aligned}$$

Note also that

$$|a_{ij}(x) - a_{ij}^h(x)| \le \sum_{k=0}^n \|F_{ijk}\|_{L_\infty} |(u - u_h)_{x_k}(x)|. \tag{5.6}$$

A and A_h as defined here differ from their previous incarnations mainly in that some lower-order terms are now included. (Note that summation indices now run from 0 to n instead of from 1 to n.) Finally, the residual $\mathcal{E}_T$ must be modified to reflect the presence of lower-order terms. Thus we now define

$$\mathcal{E}_T = h_T \left\| \sum_{i=1}^n \frac{\partial}{\partial x_i} F_i(\cdot, u_h, \nabla u_h) - F_0(\cdot, u_h, \nabla u_h) \right\|_{L_\infty(T)} + \|[u_h]\|_{L_\infty(\partial T)}.$$

The analytical assumptions of section 2 must be modified only slightly. We must still assume that the operator A is uniformly elliptic in Ω, a fact which may be established exactly as in seciton 2.2. Second, we must assume that A admits unique and sufficiently regular solutions for homogeneous Dirichlet problems. Note that establishing existence and uniqueness of solutions of such problems is potentially complicated by the presence of lower-order terms. In nonlinear problems, we must as before assume some regularity of u as well, although lower regularity may be required of u in the case of mildly nonlinear problems, as we show below. Next, the Green's function estimates of Lemma 2.3 must hold. [GW82] states such results only for divergence form operators with no lower-order terms, although the same techniques should apply when lower-order terms are present. Finally, the constant C_F arising in (2.12) must be bounded as before.

5.3. Example: A mildly nonlinear problem. Consider the mildly nonlinear problem

$$-\sum_{i,j=0}^{n}\frac{\partial}{\partial x_i}(\tilde{a}_{ij}(x,u)\frac{\partial u}{\partial x_j}) = f(x) \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega, \tag{5.7}$$

where we assume that the coefficients $F_i(x,u,\nabla u) = \sum_{j=1}^{n}\tilde{a}_{ij}(x,u)u_{x_j}$ satisfy the requirements outlined in the previous subsection. Note that the term $(A-A_h)(u-u_h,g)$ appearing in (3.10) and (4.5) of the proofs of Theorem 3.1 and Theorem 4.1 led there to a nonlinear perturbation error term of the form $\|\nabla(u-u_h)\|^2_{L_\infty(\Omega)}$. Since $F_i(x,u,\nabla u)$ now depends only linearly on ∇u, however, (5.6) yields

$$|a_{ij}-a^h_{ij}| \le C_F|u-u_h|$$

so that

$$|(A-A_h)(u-u_h,g)| \le C\|u-u_h\|_{L_\infty(\Omega)}\|\nabla(u-u_h)\|_{L_\infty(\Omega)}\|g\|_{W^1_1(\Omega)}.$$

Thus the nonlinear perturbation term is now of higher order than it generally is for approximations of u solving (1.1).

Using this observation, we may obtain estimates similar to, but often simpler than, those in Theorem 3.1, Corollary 3.5, and Theorem 4.1. We assume here that a nondegeneracy condition as outlined in Remark 3.2 is satisfied and that u possess sufficient regularity. First, analogous to Theorem 3.1 and Corollary 3.4, we find that if $\|u-u_h\|_{L_\infty(\Omega)}$ is small enough, then

$$\begin{aligned}&\|\nabla(u-u_h)\|_{L_\infty(\Omega)}\\ &\quad\le C_3\ell_{\underline{h},n}(\max_{T\in\mathcal{T}}\mathcal{E}_T + C_F\|u-u_h\|_{L_\infty(\Omega)}\|\nabla(u-u_h)\|_{L_\infty(\Omega)})\\ &\quad\le 2C_3\ell_{\underline{h},n}\max_{T\in\mathcal{T}}\mathcal{E}_T.\end{aligned}$$

Analogous to Corollary 3.5, we obtain for $\|\nabla(u-u_h)\|_{L_\infty(\Omega)}$ small enough that

$$\begin{aligned}\|u-u_h\|_{L_\infty(\Omega)} &\le C_4\ell_{\underline{h}}(\max_{T\in\mathcal{T}}h_T\mathcal{E}_T + C_F\|u-u_h\|_{L_\infty(\Omega)}\|\nabla(u-u_h)\|_{L_\infty(\Omega)})\\ &\le 2C_4\ell_{\underline{h}}\max_{T\in\mathcal{T}}h_T\mathcal{E}_T.\end{aligned} \tag{5.8}$$

Finally, we employ (5.8) and note that $h_T \le C\sigma_D(T)$ for $D\subset\Omega$ to find that for $\|\nabla(u-u_h)\|$ small enough,

$$\begin{aligned}&\|\sigma_D\nabla(u-u_h)\|_{L_\infty(\Omega)}\\ &\quad\le C_5\ell_{\underline{h},n}(\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T + C_F\|\nabla(u-u_h)\|_{L_\infty(\Omega)}\|u-u_h\|_{L_\infty(\Omega)})\\ &\quad\le C_5\ell_{\underline{h},n}(\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T + C_F\|\nabla(u-u_h)\|_{L_\infty(\Omega)}C_4\ell_{\underline{h}}\max_{T\in\mathcal{T}}h_T\mathcal{E}_T)\\ &\quad\le C_5\ell_{\underline{h},n}\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T(1 + C_F\|\nabla(u-u_h)\|_{L_\infty(\Omega)}C_4\ell_{\underline{h}})\\ &\quad\le 2C_5\ell_{\underline{h},n}\max_{T\in\mathcal{T}}\sigma_D(T)\mathcal{E}_T.\end{aligned} \tag{5.9}$$

Since the coefficients of the dual linear operator A now have essentially the same regularity as u instead of as ∇u, the constants C_3, C_4, and C_5 above depend more weakly on the regularity of u than do the corresponding constants in Theorem 3.1, Corollary 3.5, and Theorem 4.1. Indeed, C_3 depends on $\|u\|_{L_\infty(\Omega)}$ and the Dini-continuity of u as opposed to C_1 from Theorem 3.1, which depends on $\|\nabla u\|_{L_\infty(\Omega)}$

and the Dini-continuity of ∇u. C_4 and C_5 depend only on $\|u\|_{W^1_\infty(\Omega)}$ as opposed to $\tilde{C}$ and C_2 from Corollary 3.5 and Theorem 4.1, which both depend on $\|u\|_{W^2_\infty(\Omega)}$. We also note that (5.8) and (5.9) only require that $u \in C^{1,\nu}(\overline{\Omega})$ for some $\nu > 0$ in order to hold, whereas the corresponding estimates in Corollary 3.5 and Theorem 4.1 require $u \in W^2_\infty(\Omega)$.

Acknowledgments. The author would like to thank Gerhard Dziuk and two anonymous referees for their many helpful comments.

REFERENCES

[BTW03] N. Y. Bakaev, V. Thomée, and L. B. Wahlbin, *Maximum-norm estimates for resolvents of elliptic finite element operators*, Math. Comp., 72 (2003), pp. 1597–1610.

[BL93] J. W. Barrett and W. B. Liu, *Finite element approximation of the p-Laplacian*, Math. Comp., 61 (1993), pp. 523–537.

[BR01] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.

[BV00] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.

[DDP00] E. Dari, R. G. Durán, and C. Padra, *Maximum norm error estimators for three-dimensional elliptic problems*, SIAM J. Numer. Anal., 37 (2000), pp. 683–700.

[De04] A. Demlow, *Local a posteriori estimates for pointwise gradient errors in finite element methods for elliptic problems*, Math. Comp., to appear.

[De05] A. Demlow, *Weighted residual estimators for a posteriori estimation of pointwise gradient errors in quasilinear elliptic problems*, Preprint Nr [05-12], Mathematisches Institut der Albert-Ludwigs-Universität Freiburg, 2005.

[DM95] G. Dolzmann and S. Müller, *Estimates for Green's matrices of elliptic systems by L^p theory*, Manuscripta Math., 88 (1995), pp. 261–273.

[DR98] W. Dörfler and M. Rumpf, *An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation*, Math. Comp., 67 (1998), pp. 1361–1382.

[FV03] F. Fierro and A. Veeser, *On the a posteriori error analysis for equations of prescribed mean curvature*, Math. Comp., 72 (2003), pp. 1611–1634.

[FR78] J. Frehse and R. Rannacher, *Asymptotic L^∞-error estimates for linear finite element approximations of quasilinear boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 418–431.

[Gr85] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Publishing, Marshfield, MA, 1985.

[GW82] M. Grüter and K.-O. Widman, *The Green function for uniformly elliptic equations*, Manuscripta Math., 37 (1982), pp. 303–342.

[HSWW01] W. Hoffmann, A. H. Schatz, L. B. Wahlbin, and G. Wittum, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes.* I. *A smooth problem and globally quasi-uniform meshes.*, Math. Comp., 70 (2001), pp. 897–909.

[Noc95] R. H. Nochetto, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp., 64 (1995), pp. 1–22.

[NSV03] R. H. Nochetto, K. G. Siebert, and A. Veeser, *Pointwise a posteriori error control for elliptic obstacle problems*, Numer. Math., 95 (2003), pp. 163–195.

[Sch98] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids.* I. *Global estimates.*, Math. Comp., 67 (1998), pp. 877–899.

[SW04] A. H. Schatz and L. B. Wahlbin, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes.* II. *The piecewise linear case*, Math. Comp., 73 (2004), pp. 517–523.

[SZ90] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[Ver89] R. Verfürth, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.

[Ver94] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 515–539

MAXIMUM PRINCIPLE AND CONVERGENCE ANALYSIS FOR THE MESHFREE POINT COLLOCATION METHOD*

DO WAN KIM† AND WING KAM LIU‡

Abstract. The discrete Laplacian operator is considered in the sense of the meshfree point collocation method which will be called the strong meshfree Laplacian operator. To define the strong meshfree Laplacian operator, we use the fast version of the generalized moving least square approximation, which can calculate the approximated derivatives of shape functions. Some types of the locally layered node distribution are defined in this paper, and two specific domains are constructed onto which we can distribute locally layered nodes. At such nodes, the discrete maximum principle can be shown to hold through the representation formula for the strong meshfree Laplacian operator. The discrete maximum principle, together with the reproducing property of the meshfree approximations, results in an a priori estimate for the strong meshfree Laplacian operator in the nodal solution space. Furthermore, the a priori estimate we have obtained guarantees the existence and the uniqueness of the numerical solution and plays a central role in achieving converged results for the Poisson problem with Dirichlet boundary conditions in the nodal solution space. The order of convergence of the nodal solutions can be raised up to $O(h^2)$ at the proposed type of nodes in specific domains. For generally shaped domains immersed in the previously mentioned domains, we can obtain the first order convergence result of $O(h)$.

Key words. strong meshfree Laplacian operator, discrete maximum principle, a priori estimate, meshfree point collocation method, generalized moving least square approximation, convergence analysis

AMS subject classifications. 65D25, 65M15, 65M12, 65M70

DOI. 10.1137/04060809X

1. Introduction.

In the field of numerical computations, meshfree methods have been developed for more than a decade. In order to solve many physical problems represented by partial differential equations, researchers and scientists have proposed meshfree approximations, examples of which include the element free Galerkin method [3], the moving least square reproducing kernel method [10, 17], the partition of unity finite element method [2], the reproducing kernel hierarchical partition of unity [11, 12, 13, 15], the reproducing kernel element method [14, 16, 18, 20], etc.

The above pioneering work has presented a common framework for meshfree methodologies and shown the potential of meshfree methods. In many cases, the work in meshfree fields has been based on the weak formulation of the model equation, but only a few papers supply the mathematical convergence for numerical solutions in the one-dimensional (1-D) case [2, 12].

In this paper, we focus on uniform convergence analysis for the numerical solution of the strong formulation using a meshfree approximation. Here we use the generalized moving least square approximation for efficient calculation of higher order

*Received by the editors May 12, 2004; accepted for publication (in revised form) August 8, 2005; published electronically March 15, 2006. This work is supported by the Korea Research Foundation grant KRF-2003-015-C00068 and it is made possible by support from NSF under grant DMI-0115079 to Northwestern University.

http://www.siam.org/journals/sinum/44-2/60809.html

†Department of Applied Mathematics, College of Science and Technology, Hanyang University, Ansan, Kyeonggi 426-791, Republic of Korea. Visiting Scholar, Department of Mechanical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3111 (d-kiml@northwestern.edu).

‡Corresponding author. Department of Mechanical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3111 (w-liu@northwestern.edu).

shape function derivatives, which stem from the reproducing kernel hierarchical partition of unity method by Li and Liu [11, 12], and from the fast moving least square reproducing kernel approximation method by D. W. Kim and Y. Kim [8, 9]. Such approximations are desired to convert the higher order differential operator into a discrete one by attacking the strong formulation and utilizing the point collocation method. The meshfree point collocation method (MPCM) follows the philosophy of the meshfree method in which no structured meshes are used.

In mathematical analyses of the Galerkin formulation using meshfree approximations, difficulties arise mainly in the treatment of Dirichlet boundary conditions. The construction of a test function space which belongs to the Sobolev space $H_0^1(\Omega)$ is a challenging issue in meshfree Galerkin formulations, particularly for higher dimensions. However, once one can overcome this difficulty, then the remaining part of the convergence analysis follows similarly from the mathematical theories of the finite element method. For the finite element method, uniform convergence has been shown by Ciarlet and Raviart [4] for second order elliptic models under some specially shaped meshes. To achieve first order uniform convergence, they used the discrete maximum principle and obtained an a priori estimate for the discrete solution of the Poisson-type problem. For local pointwise error estimates in finite element methods, one can see the important results for second order elliptic problems in [5] written by L. B. Wahlbin.

The meshfree point collocation methods, in contrast to the Galerkin formulation, have few mathematical results, as the theory of function spaces is not directly available. Thus, the objective of this paper is to build the underlying theories for the MPCM, particularly for the discrete Laplacian operator, and based on those theories to prove uniform convergence of the nodal solutions of the Poisson problem with Dirichlet boundary conditions. For the convergence estimate in the MPCM, the first step is to define the rigorous point collocation scheme—an important portion of the mathematical analysis. Next, we will show that the discrete Laplacian operator satisfies the discrete maximum principle for some classes of nodes, and then obtain an a priori estimate for the strong meshfree Laplacian operator on the nodal solution space, provided the discrete maximum principle holds.

As for the discrete maximum principle itself, many researchers are interested in cases in which it occurs and their applications [1, 4, 7, 19, 21, 22]. The discrete maximum principle for the discretized Laplacian operator in the finite difference method on evenly spaced grid points is well-known and is closely related to the mean-value property for the Laplace solutions. This means that the average value on the surrounding four points in a five-point stencil for the Laplacian operator is equal to the center value. Inspired by the difference scheme for the Laplacian operator in the finite difference method, we can obtain the representation formula at each node for the strong meshfree Laplacian operator which is followed by the discrete maximum principle.

As a result of the discrete maximum principle, an a priori estimate for the strong meshfree Laplacian is derived in the nodal solution space. The a priori estimate guarantees the existence and the uniqueness of the numerical solution governed by the point collocation scheme. We finally achieve convergence for the numerical solutions of the Poisson problem with Dirichlet boundary conditions. The convergence order can be up to second order on some specific domains, while we have first order convergence for general domains immersed in the specific domains.

We know that finite difference methods and finite element methods have discrete maximum principle for elliptic partial differential equations. However, for meshfree

strong form collocation methods, the authors are not aware of any previous theoretical results. This is an important aspect, and this is the first paper to the authors knowledge that deals with the theoretical foundation of the meshfree collocation method.

2. Generalized moving least square reproducing kernel approximation. To make this paper self-contained, we will describe how to obtain the meshfree approximation of the Laplacian operator. For a moment, we will make general statements on the moving least square reproducing kernel approximation as we can see similar content in the literature [8, 11, 12].

Let Ω be a bounded domain in $\mathbb{R}^n$ and also $\Lambda \equiv \{\mathbf{x}_I \in \overline{\Omega} \,|\, I = 1, \ldots, N\}$ where Λ is a set of distributed nodes in $\overline{\Omega}$. Throughout the paper, the multi-index notation and related definitions are employed as follows:

$$\alpha = (\alpha_1, \ldots, \alpha_n), \quad |\alpha| \equiv \sum_{i=1}^{n} \alpha_i, \quad \alpha! \equiv \alpha_1! \alpha_2! \cdots \alpha_n!, \tag{2.1}$$

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \mathbf{x}^\alpha \equiv x_1^{\alpha_1} \ldots x_n^{\alpha_n}, \quad D^\alpha_{\mathbf{x}} \equiv \partial^{\alpha_1}_{x_1} \ldots \partial^{\alpha_n}_{x_n}, \tag{2.2}$$

where α_k's are nonnegative integers and α is called the multi-index. We consider a vector of complete basis functions of order m in $\mathbb{R}^n$ such that

$$\mathbf{B}_m(\mathbf{x}) = (b_{\beta_1}(\mathbf{x}), b_{\beta_2}(\mathbf{x}), \ldots, b_{\beta_L}(\mathbf{x}))^T, \quad |\beta_k| \leq m, \tag{2.3}$$

where β_k's are all multi-indices in lexicographical order. Here we note that the number of β_k's is $L \equiv \frac{(m+n)!}{m!n!}$ and the complete basis of order m means that the $L \times L$ matrix $J_{\mathbf{B}_m}(\mathbf{0})$ is invertible if we define the Jacobian of $\mathbf{B}_m(\mathbf{x})$ at $\mathbf{0}$ as

$$J_{\mathbf{B}_m}(\mathbf{0}) \equiv \lim_{\mathbf{x} \to \mathbf{0}} \left(D^\alpha_{\mathbf{x}} \, b_\beta(\mathbf{x}) \right), \quad |\alpha|, |\beta| \leq L. \tag{2.4}$$

Let $B_r(\mathbf{z}) \equiv \{\mathbf{y} \,|\, \|\mathbf{y} - \mathbf{z}\| < r\}$ be the r-ball in $\mathbb{R}^n$ with center $\mathbf{z}$. We introduce the continuous nonnegative window function with its support on $\overline{B_1(\mathbf{0})}$ of the following type

$$W(\mathbf{x}) = (1 - \|\mathbf{x}\|^{\frac{1}{2}})^2 \quad \text{for} \quad \|\mathbf{x}\| < 1, \, \mathbf{x} \in \mathbb{R}^n \tag{2.5}$$

and the continuous positive dilation function

$$\rho(\mathbf{x}) > 0 \quad \text{on} \quad \overline{\Omega}. \tag{2.6}$$

For brevity, we will use $\rho_{\mathbf{x}}$ instead of $\rho(\mathbf{x})$.

Remark 1. The decreasing rate of the window function values apart from the origin is essential in proving the discrete maximum principle for the strong meshfree Laplacian operator introduced in a later section. The window function of the form (2.5) meets this decreasing rate. The support of the window function in this paper has the n-dimensional unit ball shape.

Remark 2. The dilation parameter used in most meshfree methods can be replaced with the dilation function defined on the whole domain $\overline{\Omega}$. The required regularity for the dilation function is only the continuity to be well defined when the center of window function moves to the evaluation point. The dilation function controls the support and its size of shape functions, and thus is directly available to the geometrically multiple scale problems [9].

The subsequent procedure to make the shape functions and the approximated derivative operators is addressed in detail in Appendix I. This includes the generalized

reproducing properties of meshfree shape functions and the proposal of a sufficient condition to regenerate the dilated basis functions. These are novel differences from the standard moving least square reproducing kernel approximation.

From this point forward, we restrict our attention to polynomial basis functions; that is, if there is no comment, then the basis functions will be maintained as complete polynomials up to order m

$$\mathbf{B}_m(\mathbf{x}) = (\mathbf{x}^{\beta_1}, \mathbf{x}^{\beta_2}, \ldots, \mathbf{x}^{\beta_L}), \quad |\beta_k| \le m \tag{2.7}$$

throughout the mathematical analysis.

For the subsequent analysis, we require the definition of the proper node distributions.

DEFINITION 1 (proper triple). *Let $(\Omega, \Lambda, \rho_{\mathbf{x}})$ be the triple of a domain, a set of distributed nodes on $\overline{\Omega}$, and a dilation function. The triple $(\Omega, \Lambda, \rho_{\mathbf{x}})$ is said to be proper if the moment matrix $M^{\rho_{\mathbf{x}}}(\mathbf{x}_I)$ is invertible for every interior node $\mathbf{x}_I \in \Lambda \cap \Omega$ under the dilation function $\rho_{\mathbf{x}}$.*

This definition is preventive of the degenerate distribution of nodes to approximate functions in the meshfree method.

3. Problem statement and the definition of the discrete problem. We will now consider the discretization of the Poisson problem as the popular model in the second order elliptic problem with Dirichlet boundary conditions and prepare the terminology for its convergence analysis. The Poisson equation uses the Laplacian operator, the principal operator in most physical models. Furthermore, the Laplacian is an interesting operator in itself, since it has the salient feature referred to as the maximum principle. Many mathematical theories have been developed based on this property. Among them, the regularity and the uniqueness of solutions of the Poisson equation is highly involved with the maximum principle. For the discrete case analogous to the continuous one, the discrete maximum principle has been reported not only for the Galerkin formulation [4] in the finite element method but also for the solution of some algebraic systems [7].

We consider the Poisson problem with Dirichlet data on the boundary of a domain Ω and propose the corresponding discrete problem using the point collocation approach based on the generalized meshfree approximation operators described in the previous section and Appendix I in detail. The model problem considered in this paper is governed by the following equations:

$$(\mathbf{CP}) \begin{cases} \Delta u = f, & \text{in } \Omega \\ u = g, & \text{on } \Gamma, \end{cases} \tag{3.1}$$

where $\Gamma \equiv \partial\Omega$ represents the boundary of the open bounded domain Ω. According to Theorem 6.13 in [6] for the general existence and regularity of a unique solution of (**CP**), if Ω is a bounded domain satisfying an exterior sphere condition at every boundary point and we have $f \in C^{s-2,\,\alpha}(\Omega)$ for $s = 3, 4$ and $g \in C(\partial\Omega)$, then the Dirichlet problem (**CP**) has a unique solution $u \in C^0(\overline{\Omega}) \cap C^{s,\,\alpha}(\Omega)$, where $C^0(\overline{\Omega})$ is the vector space to consist of all bounded and uniformly continuous functions on Ω and $C^{s,\alpha}(\Omega)$ represents the Hölder space of exponent $0 < \alpha \le 1$ equipped with the norm

$$\|v\|_{C^{s,\alpha}(\Omega)} \equiv \max_{0 \le |\beta| \le s} \sup_{\mathbf{x} \in \Omega} |D^\beta v(\mathbf{x})| + \max_{0 \le |\beta| \le s} \sup_{\mathbf{x},\mathbf{y} \in \Omega, \mathbf{x} \ne \mathbf{y}} \frac{|D^\beta v(\mathbf{x}) - D^\beta v(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\alpha}. \tag{3.2}$$

To discretize the problem (**CP**) in terms of meshfree point collocation method, we focus on the second order meshfree approximation ($m = 2$); that is, the complete polynomial basis functions up to second order are adopted to obtain all the shape functions (see Appendix I). It is taken to satisfy the minimum order of consistency for the discretization of second order partial differential equations since we approximate the Laplace operator in a pointwise manner. Higher order approximation could be better than the second order one in general but, since the focus in this paper is on analyzing the structure of the meshfree Laplace operator, the second order meshfree approximation must be the starting point. We also consider the 2-D space ($n = 2$) and hence the relevant multi-index $\beta_k\,(k = 1, \ldots, 6)$ appearing in the basis polynomials (2.7) which are fixed in lexicographical order as follows:

$$(0,0), \quad (1,0), \quad (0,1), \quad (2,0), \quad (1,1), \quad (0,2). \tag{3.3}$$

The analysis in this paper is expected to hold for the higher dimensions as well as for higher orders. It could depend on the construction of the local nodes and the adequate dilation function.

In order to define the discrete counterpart of the continuous problem (3.1), we assume Λ is a set of well distributed nodes on the domain Ω and its boundary, so that $(\Omega, \Lambda, \rho_{\mathbf{x}})$ becomes the proper triple. Let $C(\overline{\Omega})$ be the space of continuous functions up to the boundary of Ω, and V be the finite dimensional space of functions defined on Λ. We will call the function space V the nodal solution space if V is equipped with the following seminorm:

$$\|v\|_{\infty,A} \equiv \max_{\mathbf{x}_K \in A} |v_K| \quad \text{when} \quad v \in V, \tag{3.4}$$

where A is a nonempty subset of Λ. In the case when $A = \Lambda$, the seminorm becomes the norm on V.

If the restriction map $i : C(\overline{\Omega}) \to V$ to Λ is defined such that, for any $u \in C(\overline{\Omega})$,

$$i(u)(\mathbf{x}_I) \equiv u(\mathbf{x}_I) \quad \text{for any } \mathbf{x}_I \in \Lambda, \tag{3.5}$$

then the point collocation Laplacian operator Δ^ρ can be defined on V into itself such that if $v \in V$, then

$$(\Delta^\rho v)(\mathbf{x}_I) \equiv \sum_{\mathbf{x}_J \in \Lambda} v(\mathbf{x}_J)\, \psi_J^{\Delta}(\mathbf{x}_I) \quad \text{for any } \mathbf{x}_I \in \Lambda, \tag{3.6}$$

where the function $\psi_J^{\Delta}(\mathbf{x}_I)$ will be called the Laplacian shape function at $\mathbf{x}_I$ and is defined by

$$\psi_J^{\Delta}(\mathbf{x}_I) \equiv \psi_J^{[(2,0)]}(\mathbf{x}_I) + \psi_J^{[(0,2)]}(\mathbf{x}_I) \tag{3.7}$$

which is the sum of the $(2,0)$th and $(0,2)$th approximate derivatives whose definition comes from (7.3) in Appendix I. In fact, the operator Δ^ρ stems from the meshfree approximated Laplacian operators $D^{(2,0)}_{m,\rho_{\mathbf{x}}} + D^{(0,2)}_{m,\rho_{\mathbf{x}}} \sim \Delta$. Hereafter, we will often use the symbol u_J instead of $u(\mathbf{x}_J)$ if $u \in V$ and $\mathbf{x}_J \in \Lambda$.

Using these operators i and Δ^ρ, we define the meshfree point collocation discretization of Poisson problem (**CP**) as the following:

$$u_h \in V \,:\, \begin{cases} \Delta^\rho u_h = i(f), & \text{on } \Lambda^o \\ u_h = g, & \text{on } \Lambda^b \end{cases}, \tag{3.8}$$

where $\Lambda = \Lambda^o \cup \Lambda^b$ and Λ^o and Λ^b are sets of interior nodes and Dirichlet boundary nodes, respectively. Consequently, our discrete problem for (**CP**) results in finding the nodal solution $u_h \in V$ such that

$$(\mathbf{DP}) \begin{cases} u_h \in V_g \equiv \{v_J \in \mathbb{R} \,|\, v_K = g(\mathbf{x}_K) \text{ for all } \mathbf{x}_K \in \Lambda^b\} \subset V \\ \Delta^\rho u_h = i(f), \quad \text{on } \Lambda^o. \end{cases} \tag{3.9}$$

This final formulation will be called *the discrete Poisson problem* (**DP**) and the operator Δ^ρ will be called *the strong meshfree Laplacian operator*.

In order to attain the error estimate, we begin discussion of the discrete maximum principle for the strong meshfree Laplacian operator Δ^ρ.

4. Discrete maximum principle for the strong meshfree Laplacian operator $\boldsymbol{\Delta^\rho}$. Let $(\Omega, \Lambda, \rho_{\mathbf{x}})$ be the proper triple. For convenience sake, the r-neighbor nodes of $\mathbf{x}$ are assumed to be the following set:

$$\Lambda_r(\mathbf{x}) \equiv \{\mathbf{x}_K \in \Lambda \,|\, \mathbf{x}_K \in B_r(\mathbf{x})\}, \quad r > 0 \tag{4.1}$$

and the symbol A^* for a subset $A \subset \Lambda$ implies the set defined by

$$A^* \equiv \bigcup_{\mathbf{x}_J \in A} \Lambda_{\rho_{\mathbf{x}_J}}(\mathbf{x}_J). \tag{4.2}$$

If there is no confusion, we briefly write $\Lambda(\mathbf{x}_K)$ instead of $\Lambda_{\rho_{\mathbf{x}_K}}(\mathbf{x}_K)$ for any node $\mathbf{x}_K \in \Lambda$.

We now state the definition of the discrete maximum principle.

DEFINITION 2 (discrete maximum principle for the operator Δ^ρ). *Assume the proper triple $(\Omega, \Lambda, \rho_{\mathbf{x}})$ is given. We will say the strong meshfree Laplacian Δ^ρ satisfies the discrete maximum principle at a node $\mathbf{x}_I \in \Lambda$ if the condition $(\Delta^\rho v)(\mathbf{x}_I) \geq 0$ for $v \in V$ implies that either $v_I < \max_{\mathbf{x}_K \in \Lambda(\mathbf{x}_I) \setminus \{\mathbf{x}_I\}} v_K$ or $v_K = v_I$ for all $\mathbf{x}_K \in \Lambda(\mathbf{x}_I)$. We also will say the operator Δ^ρ satisfies the discrete maximum principle on a subset $A \subset \Lambda$ if it satisfies the discrete maximum principle at all nodes in A.*

In fact, the discrete maximum principle for the discrete Laplace operator is known to depend on the geometry of the mesh in the finite element method and the orthogonal grid in the finite difference method, respectively. For example, if all the angles of the triangles of the triangulation on a domain are less than or equal to $\frac{\pi}{2}$, then the discrete maximum principle is known to hold in the finite element method [4]. Hence it can also be expected that the relative attitude between nodes strongly affects this kind of phenomenon in the meshfree area. Therefore, we are interested in finding such node distributions from the meshfree point of view. On the other hand, to perform the convergence analysis on such nodes, we have to inspect closely the moment matrix and its inverse, since it is located in the core of Laplacian shape functions in (3.7).

The moment matrix for the given set Λ of nodes has the following form in the generalized moving least square reproducing kernel approximation [17] (see also (7.5) in Appendix I)

$$M^{\rho_{\mathbf{x}}}(\mathbf{x}) = \sum_{\mathbf{x}_I \in \Lambda} \mathbf{B}_m \left(\frac{\mathbf{x}_I - \mathbf{x}}{\rho_{\mathbf{x}}} \right) \mathbf{B}_m^T \left(\frac{\mathbf{x}_I - \mathbf{x}}{\rho_{\mathbf{x}}} \right) W \left(\frac{\mathbf{x}_I - \mathbf{x}}{\rho_{\mathbf{x}}} \right), \tag{4.3}$$

where $\mathbf{B}_m \left(\frac{\mathbf{y} - \mathbf{x}}{\rho_{\mathbf{x}}} \right)$ is the normalized basis polynomial up to order m at the center

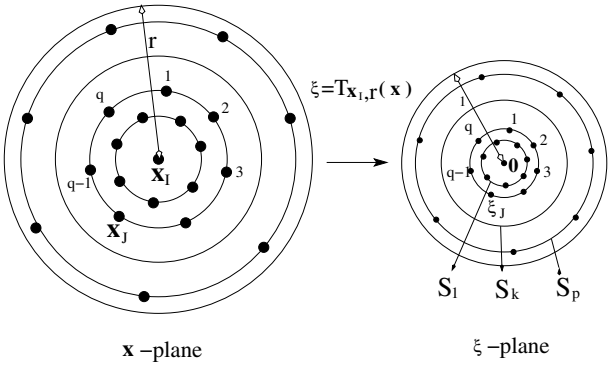


FIG. 4.1. *The locally (p,q)-layered nodes ($\mathbf{x}$-plane) and the normalizing ones (ξ-plane) by $T_{\mathbf{x}_I,r}(\mathbf{x})$. S_K's are the layers of the normalized nodes.*

point $\mathbf{x} \in B_{\rho_{\mathbf{x}}}(\mathbf{x})$ such that

$$\mathbf{B}_m\left(\frac{\mathbf{y}-\mathbf{x}}{\rho_{\mathbf{x}}}\right) = \left(\left(\frac{\mathbf{y}-\mathbf{x}}{\rho_{\mathbf{x}}}\right)^{\beta_1}, \left(\frac{\mathbf{y}-\mathbf{x}}{\rho_{\mathbf{x}}}\right)^{\beta_2}, \ldots, \left(\frac{\mathbf{y}-\mathbf{x}}{\rho_{\mathbf{x}}}\right)^{\beta_L}\right), \quad |\beta_k| \le m. \tag{4.4}$$

To calculate the moment matrix and its inverse concretely, we need to focus on some class of node distributions. In order to define some classes of nodes, we must first introduce the normalizing transformation $T_{\mathbf{x},r}(\mathbf{y}) : B_r(\mathbf{x}) \to B_1(\mathbf{0})$ such that, as shown in Figure 4.1,

$$\xi = T_{\mathbf{x},r}(\mathbf{y}) \equiv \frac{\mathbf{y}-\mathbf{x}}{r}. \tag{4.5}$$

DEFINITION 3 (layered node distribution). *Let $A_r(\mathbf{x}_I) \equiv \{\mathbf{x}_K \,|\, \mathbf{x}_K \in B_r(\mathbf{x}_I)\}$ be the finite subset of nodes within the distance r around $\mathbf{x}_I$. The set of nodes $A_r(\mathbf{x}_I)$ is said to be the locally (p,q)-layered at $\mathbf{x}_I$ if all the normalized nodes in $T_{\mathbf{x}_I,r}(A_r(\mathbf{x}_I))$ remain on the p-layer sets $S_1, \ldots, S_p$ in the increasing radial direction from the origin and the q nodes are distributed evenly on each layer. All the layer sets S_k's have the spherical shape only. Furthermore, we say that the node set Λ is possibly layered if, for any interior node $\mathbf{x}_I \in \Lambda$, $\Lambda(\mathbf{x}_I)$ is the locally (p,q)-layered at $\mathbf{x}_I$ for some $p,q > 0$.*

As a matter of fact, the possibly layered distribution of nodes is not a simple matter since the property of the locally (p,q)-layered at every neighboring node has to be achieved. Thus, we will propose two kinds of available distribution of nodes and show that they are the possibly layered. On such types of the possibly layered nodes, the discrete maximum principle for the strong meshfree Laplacian will be proven.

We begin with the calculation of the moment matrix that will play an essential role in proving the discrete maximum principle on some possibly layered nodes. If the subset of nodes $\Lambda(\mathbf{x}_I) \subset \Lambda$ is the locally (p,q)-layered at $\mathbf{x}_I$, then the moment matrix at $\mathbf{x}_I$ can be calculated from the following manner:

$$M^{\rho_{\mathbf{x}}}(\mathbf{x}_I) = W(\mathbf{0})\,\mathbf{B}_m(\mathbf{0})\mathbf{B}_m(\mathbf{0})^T + \sum_{K=1}^{p} \delta_K\,\mathbf{D}_K \left(\sum_{\xi_J \in S_K} \mathbf{B}_m(\zeta_J)\mathbf{B}_m(\zeta_J)^T\right) \mathbf{D}_K, \tag{4.6}$$

where S_K is the Kth layer set in the definition and for any nonzero $\xi_J \in S_K$ we use the symbols

$$\xi_J \equiv T_{\mathbf{x}_I}(\mathbf{x}_J), \quad \tau_K \equiv |\xi_1| = \cdots = |\xi_q| < 1, \quad \zeta_J \equiv \frac{\xi_J}{\tau_K}, \tag{4.7}$$

$$\delta_K \equiv W(\xi_1) = \cdots = W(\xi_q), \tag{4.8}$$

and $\mathbf{D}_K$ is the diagonal matrix such that

$$\mathbf{D}_K \equiv \mathrm{Diag}(\tau_K^{|\alpha_1|}, \tau_K^{|\alpha_2|}, \ldots, \tau_K^{|\alpha_L|}). \tag{4.9}$$

Since we have assumed $n = 2$, we have, without loss of generality, the ζ_J's distributed evenly on the layer S_K and represented by

$$\zeta_j = \left(\cos\left(\theta_K + j\,\frac{2\pi}{q}\right), \sin\left(\theta_K + j\,\frac{2\pi}{q}\right)\right), \quad j = 0, 1, \ldots, q-1, \tag{4.10}$$

where θ_K is the angle of the starting node ζ_1 on S_K. If the distribution of nodes around $\mathbf{x}_I$ is assumed to be the locally (p, q)-layered at $\mathbf{x}_I$, then, from the trigonometric identities in Appendix II, the term $\sum_{\xi_J \in S_K} \mathbf{B}(\zeta_J)\mathbf{B}(\zeta_J)^T$ in (4.6) has the following forms for the cases when $q = 4$ and $q \geq 5$:

- $\sum_{\xi_J \in S_K} \mathbf{B}(\zeta_J)\mathbf{B}(\zeta_J)^T$ when $q = 4$,

$$4\begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{3}{8} + \frac{1}{8}\cos 4\theta_K & \frac{1}{8}\sin 4\theta_K & \frac{1}{8} - \frac{1}{8}\cos 4\theta_K \\ 0 & 0 & 0 & \frac{1}{8}\sin 4\theta_K & \frac{1}{8} - \frac{1}{8}\cos 4\theta_K & -\frac{1}{8}\sin 4\theta_K \\ \frac{1}{2} & 0 & 0 & \frac{1}{8} - \frac{1}{8}\cos 4\theta_K & -\frac{1}{8}\sin 4\theta_K & \frac{3}{8} + \frac{1}{8}\cos 4\theta_K \end{bmatrix}; \tag{4.11}$$

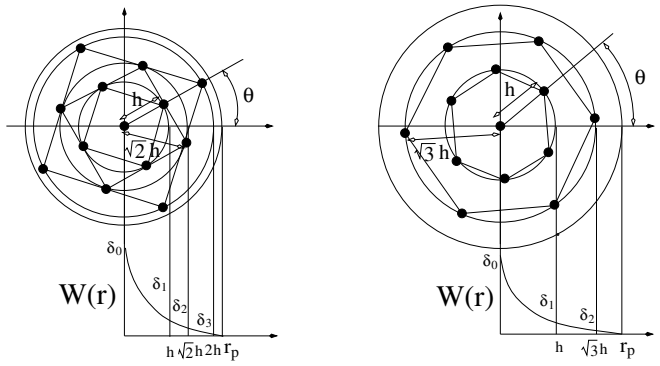
- $\sum_{\xi_J \in S_K} \mathbf{B}(\zeta_J)\mathbf{B}(\zeta_J)^T$ when $q \geq 5$

$$q\begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{3}{8} & 0 & \frac{1}{8} \\ 0 & 0 & 0 & 0 & \frac{1}{8} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{8} & 0 & \frac{3}{8} \end{bmatrix}. \tag{4.12}$$

Here we note that, if the number of nodes on the layer S_K is greater than or equal to 5, then the moment matrix does not depend on θ_K. This means that the (p, q)-layered node distributions for $q \geq 5$ makes the rotation invariant moment matrix.

The strong meshfree Laplacian operator Δ^ρ at $\mathbf{x}_I$ on the (p, q)-layered node set $\Lambda(\mathbf{x}_I)$ can be calculated from the equivalent form:

$$\begin{aligned} \sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} u_h(\mathbf{x}_J)\, \psi_J^{\Delta}(\mathbf{x}_I) &= \mathbf{d}_\Delta\, M^{\rho\star}(\mathbf{x}_I)^{-1}\, \mathbf{B}_m(\mathbf{0})\, W(\mathbf{0})\, u_h(\mathbf{0}) \\ &+ \mathbf{d}_\Delta\, M^{\rho\star}(\mathbf{x}_I)^{-1} \sum_{K=1}^{p} \delta_K\, \mathbf{D}_K \left[\mathbf{B}_m\left(\zeta_1^K\right)\ \mathbf{B}_m\left(\zeta_2^K\right)\ \cdots\ \mathbf{B}_m\left(\zeta_q^K\right)\right] \mathbf{u}_h^K, \end{aligned} \tag{4.13}$$



(I) Type I : Locally (p, 4)–layered nodes p = 2, 3 (II) Type II : Locally (p, 6)–layered nodes p = 1, 2

FIG. 4.2. *Two types of locally* (p, q)*-layered node distribution:* (I) *the set of Type* I *and,* (II) *the set of Type* II.

where $\mathbf{u}_h^K \equiv [u_h(\xi_1^K), \ldots, u_h(\xi_q^K)]^T$ is a column vector, the superscript of ξ_J^K means that ξ_J is on S_K, and when $m = 2$ the symbol $\mathbf{d}_\Delta$ designates the following row vector:

$$\mathbf{d}_\Delta \equiv \left[0,\ 0,\ 0,\ \frac{(2,0)!}{\rho_{\mathbf{x}_I}^{|(2,0)|}},\ 0,\ \frac{(0,2)!}{\rho_{\mathbf{x}_I}^{|(0,2)|}}\right]. \tag{4.14}$$

We now consider the two kinds of locally (p, q)-layered node distributions. The first one is composed of orthogonally positioned nodes and the other comes from a hexagonal structure.

In constructing specific types of nodes, we use the symbol δ_K as defined in (4.8) including $\delta_0 \equiv W(0) = 1$, or equivalently we have $\delta_K \equiv W(\tau_K)$ since our window function W in (2.5) depends only on the radial values. The terminology of the multi-index β_k defined in (3.3) is also utilized.

Let $h > 0$ and θ be the given angle.

4.1. Type I: The locally $(p, 4)$-layered nodes $(p = 2, 3)$. Let $A_{r_p}(\mathbf{0})$ be the set consisting of the following nodes as shown in Figure 4.2(I):

$$\{(0,0)\} \cup \bigcup_{K=1}^{p} \left\{ \left(t_K\, h\, \cos\left(\theta_K + i\,\frac{2\pi}{4}\right),\ t_K\, h\, \sin\left(\theta_K + i\,\frac{2\pi}{4}\right) \right) \,\middle|\, i = 0, 1, 2, 3 \right\}, \tag{4.15}$$

where $t_K = \sqrt{2}^{K-1}$, $\theta_K = \theta + (K-1)\frac{\pi}{4}$, and

$$r_p = h\,\frac{\sqrt{2}+2}{2},\ h\,\frac{2+\sqrt{5}}{2},\ \text{respectively when } p = 2, 3. \tag{4.16}$$

If A is a subset of nodes with $\mathbf{x}_I$ as its center node and it has the same property as $A_{r_p}(\mathbf{0})$ for $p = 2, 3$ under the normalizing transform (4.5), then it is said to be *the set of Type I at the node* $\mathbf{x}_I$. In this case, the values of τ_K's in (4.7) are calculated as the following:

$$\tau_K = \frac{h\, t_K}{r_p}, \quad 1 \le K \le p = 2, 3. \tag{4.17}$$

When $p = 3$, the determinant of the moment matrix in (4.6) at the center node can be calculated such as

$$|M^{r_p}(\mathbf{0})| = 2^6 \tau_1^{2\sum_{k=1}^{6}|\beta_k|}\, \delta_2\,(\delta_1 + 2\,\delta_2 + 4\,\delta_3)^2\,(\delta_1 + 16\,\delta_3)\,A_R \neq 0, \tag{4.18}$$

where the symbol A_R stands for the following positive value:

$$A_R = \delta_1 + 4\,\delta_2 + 16\,\delta_3 + 4\,\delta_1\delta_2 + 16\,\delta_2\delta_3 + 36\,\delta_1\delta_3. \tag{4.19}$$

When $p = 2$, we may simply set $\delta_3 = 0$.

On this kind of local node distribution, the strong meshfree Laplacian operator at the origin is calculated from the equation (4.13) as follows:

$$\psi^{\Delta}_{\mathbf{0}}(\mathbf{0}) = -\frac{4}{h^2}\,\frac{\delta_0\,(\delta_1 + 2\,\delta_2 + 4\,\delta_3)}{A_R}, \tag{4.20}$$

$$\psi^{\Delta}_{\xi_J \in S_1}(\mathbf{0}) = \frac{1}{h^2}\,\frac{\delta_1\,(1 - 4\,\delta_2 - 12\,\delta_3)}{A_R} \equiv \frac{1}{h^2}\,A_1, \tag{4.21}$$

$$\psi^{\Delta}_{\xi_J \in S_2}(\mathbf{0}) = \frac{1}{h^2}\,\frac{2\,\delta_2\,(1 + 2\,\delta_1 - 4\,\delta_3)}{A_R} \equiv \frac{1}{h^2}\,A_2, \tag{4.22}$$

$$\psi^{\Delta}_{\xi_J \in S_3}(\mathbf{0}) = \frac{1}{h^2}\,\frac{4\,\delta_3\,(1 + 3\,\delta_1 + 2\,\delta_2)}{A_R} \equiv \frac{1}{h^2}\,A_3, \tag{4.23}$$

where S_K $(K = 1, 2, 3)$ is the Kth layer and there are 4-nodes on each S_K. Therefore, from (4.20), (4.21), (4.22), and (4.23) and the fact that $A_1 + A_2 + A_3 = \frac{\delta_1 + 2\,\delta_2 + 4\,\delta_3}{A_R}$ since $\delta_0 \equiv W(\mathbf{0}) = 1$, we have, for any $v \in V$,

$$h^2 \sum_{\mathbf{x}_J \in A_r(\mathbf{0})} v(\mathbf{x}_J)\,\psi^{\Delta}_J(\mathbf{0}) = \sum_{k=1}^{3} A_k \left(-4\,v(\mathbf{0}) + \sum_{\xi_J \in S_k} v(\xi_J) \right). \tag{4.24}$$

If we set $\delta_3 = 0$ in the above formula, then we obtain the case when $p = 2$. With these types of nodes, the moment matrix depends on the rotation of nodes (θ) but the discrete Laplacian shape function $\psi^{\Delta}_J(\mathbf{0})$ is invariant under the rotation.

Summarizing the above discussion, we have the following lemma on the set of Type I at $\mathbf{x}_I$.

LEMMA 1. *Let $(\Omega, \Lambda, \rho_{\mathbf{x}})$ be a proper triple and $\Lambda(\mathbf{x}_I) \subset \Lambda$ be the set of Type I at the node $\mathbf{x}_I$. Then we have the following properties:*

1. *The following representation formula for the strong meshfree Laplacian operator holds:*

$$\sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} v(\mathbf{x}_J)\,\psi^{\Delta}_J(\mathbf{x}_I) = \frac{1}{h^2} \sum_{K=1}^{p} A_K \left(-4\,v(\mathbf{x}_I) + \sum_{\xi_J \in S_K} v(\xi_J) \right) \tag{4.25}$$

for some coefficients A_K which depend on the window function.

2. *If the coefficients A_K are positive, then the following type of inverse inequality for the Laplacian shape functions holds:*

$$\sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} \left|\, \psi^{\Delta}_J(\mathbf{x}_I) \,\right| \leq \frac{8}{h^2}. \tag{4.26}$$

Proof. The first property directly comes from (4.24). For the second property, if $A_K > 0$ for $K = 1, \ldots, p$ where $p = 2, 3$, then we have the following bound from (4.20), (4.21), (4.22), and (4.23):

$$\sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} \left|\psi_J^{\Delta}(\mathbf{x}_I)\right| = -\psi_I^{\Delta}(\mathbf{x}_I) + \sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)\setminus\{\mathbf{x}_I\}} \psi_J^{\Delta}(\mathbf{x}_I) = \frac{8}{h^2} \sum_{K=1}^{p} A_K \le \frac{8}{h^2} \tag{4.27}$$

since $\psi_I^{\Delta}(\mathbf{x}_I)$ is the only nonpositive term among $\psi_J^{\Delta}(\mathbf{x}_I)$ for all J, $\mathbf{x}_J \in \Lambda(\mathbf{x}_I)$. ∎

4.2. Type II: The locally $(p, 6)$-layered nodes ($p = 1, 2$). Let $A_{r_p}(\mathbf{0})$ be the set consisting of the following nodes as shown in Figure 4.2(II):

$$\{(0,0)\} \cup \bigcup_{K=1}^{p} \left\{ \left(t_K\, h\, \cos\left(\theta_K + i\,\frac{2\pi}{6}\right),\ t_K\, h\, \sin\left(\theta_K + i\,\frac{2\pi}{6}\right) \right) \,\middle|\, i = 0, 1, 2, 3, 4, 5 \right\}, \tag{4.28}$$

where $t_K = \sqrt{3}^{K-1}$, $\theta_K = \theta + (K-1)\frac{\pi}{6}$, and

$$r_p = h\frac{\sqrt{3}^{p-1} + \sqrt{3}^{p}}{2}, \quad p = 1, 2. \tag{4.29}$$

If A is a subset of nodes with $\mathbf{x}_I$ as its center node and it has the same property as $A_{r_p}(\mathbf{0})$ for $p = 1, 2$ under the normalizing transform (4.5), then it is said to be *the set of Type II at the node* $\mathbf{x}_I$. Here, we can see that

$$\tau_K = \frac{h\, t_K}{r_p}, \quad 1 \le K \le p = 1, 2. \tag{4.30}$$

When $p = 2$, the determinant of the moment matrix in (4.6) at the center node in this case can be calculated as follows:

$$|M^{r_p}(\mathbf{0})| = 3^5 \tau_1^{\,2} \sum_{k=1}^{6} |\beta_k| \,(\delta_1 + 3\,\delta_2)^2 (\delta_1 + 9\,\delta_2)^2\, A_H \neq 0, \tag{4.31}$$

where the symbol A_H means the following positive value:

$$A_H = \delta_1 + 9\,\delta_2 + 24\,\delta_1\delta_2. \tag{4.32}$$

In the case when $p = 1$, we can set $\delta_2 = 0$.

On this kind of local node distribution, the strong meshfree Laplacian operator at the origin is derived from (4.13) as follows:

$$\psi_{\mathbf{0}}^{\Delta}(\mathbf{0}) = -\frac{4}{h^2}\,\frac{\delta_0\,(\delta_1 + 3\,\delta_2)}{A_H}, \tag{4.33}$$

$$\psi_{\xi_J \in S_1}^{\Delta}(\mathbf{0}) = \frac{1}{h^2}\,\frac{\frac{2}{3}\delta_1\,(1 - 12\,\delta_2)}{A_H} \equiv \frac{1}{h^2} A_1, \tag{4.34}$$

$$\psi_{\xi_J \in S_2}^{\Delta}(\mathbf{0}) = \frac{1}{h^2}\,\frac{2\delta_2\,(1 + 4\,\delta_1)}{A_H} \equiv \frac{1}{h^2} A_2, \tag{4.35}$$

where $S_K\,(K = 1, 2)$ is the Kth layer and there are 6-nodes in each S_K. Thus, from (4.33), (4.34), (4.35), and the identity $A_1 + A_2 = \frac{\frac{2}{3}(\delta_1 + 3\,\delta_2)}{A_H}$, we also obtain, for any $v \in V$,

$$h^2 \sum_{\mathbf{x}_J \in A_r(\mathbf{0})} v(\mathbf{x}_J)\, \psi_J^{\Delta}(\mathbf{0}) = \sum_{k=1}^{2} A_k \left(-6\, v(\mathbf{0}) + \sum_{\xi_J \in S_k} v(\xi_J) \right). \tag{4.36}$$

In the case when $p = 1$, we can only set $\delta_2 = 0$.

Summarizing the above discussion, on the set of Type II at $\mathbf{x}_I$, we have the following lemma similar to Lemma 1.

LEMMA 2. *Let $(\Omega, \Lambda, \rho_{\mathbf{x}})$ be a proper triple and $\Lambda(\mathbf{x}_I) \subset \Lambda$ be the set of Type II at the node $\mathbf{x}_I$. Then we have the following properties:*

1. *The following representation formula for the strong meshfree Laplacian operator holds:*

$$\sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} v(\mathbf{x}_J)\, \psi_J^{\Delta}(\mathbf{x}_I) = \frac{1}{h^2} \sum_{K=1}^{p} A_K \left(-6\, v(\mathbf{x}_I) + \sum_{\xi_J \in S_K} v(\xi_J) \right) \tag{4.37}$$

for some coefficients A_K which depend on the window function.

2. *If the coefficients A_K are positive, then the following inverse inequality for the Laplacian shape functions holds:*

$$\sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} \left| \psi_J^{\Delta}(\mathbf{x}_I) \right| \le \frac{8}{h^2}. \tag{4.38}$$

Proof. The proof is similar to that in Lemma 1, so we omit the proof. ∎

Remark 3. The set of Type I and Type II belong to the locally $(p, 4)$-layered and locally $(p, 6)$-layered class, respectively. The significant feature of the set of Type I and Type II nodes is the staggered distribution of nodes across layers. Particularly, in the case of Type I, it is essential for the invertibility of the moment matrix at the center node since the matrix (4.11) derived from the nodes in each layer is singular with kernel dimension 1. However, in the case of Type II, the nodes do not have to be staggered through layers since one can see the nonsingular matrix (4.12) is independent of the attitude of nodes in each layer. Hence, Type II is more natural than Type I in the meshfree approximation.

4.3. Two possibly layered node distributions on specific domains. We propose two kinds of evenly spaced nodes on some domains. The size, the rotation, and the translation of the domain we are to construct are not critical in the subsequent analysis (i.e., the subsequent analysis is independent of the similarity transfromation).

As shown in Figure 4.3(a), we first consider the open square domain Ω_R with 4 vertices at

$$(1, 1),\ (-1, 1),\ (-1, -1),\ (1, -1). \tag{4.39}$$

In this case, the nodes can be distributed on $\overline{\Omega_R}$ to be Type I (i.e., staggered locally $(p, 4)$-layered $(p = 2, 3)$) at each interior node. The set of such nodes on Ω_R is written by the symbol Λ_R.

As depicted in Figure 4.3(b), the hexagonal domain Ω_H with the six vertices located at

$$(1, 0),\ \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right),\ \left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right),\ (-1, 0),\ \left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right),\ \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) \tag{4.40}$$

is taken as the second open domain. On the domain $\overline{\Omega_H}$, the nodes can be entirely distributed to be Type II (i.e., locally $(p, 6)$-layered $(p = 1, 2)$) at each interior node. Such a set of nodes is denoted by Λ_H.

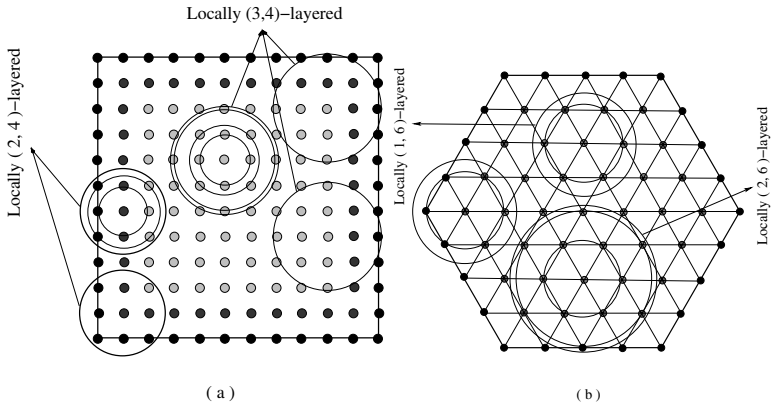


FIG. 4.3. *Possibly layered node distribution:* (*a*) Ω_R, *and* (*b*) Ω_H.

In both cases, the minimum node distance is set by $h \equiv \frac{2}{n}$ where n is regarded as the number of divisions. Now, we determine the dilation function $\rho_{\mathbf{x}}$ for each case. We only need the values of $\rho_{\mathbf{x}}$ at nodes. First of all, we notice that the value $\rho_{\mathbf{x}_I}$ at each node $\mathbf{x}_I$ on Λ_R (or Λ_H) depends on how we take the p number in the locally (p,q)-layered set $\Lambda(\mathbf{x}_I)$ at $\mathbf{x}_I$. As illustrated in Figure 4.3(a) and (b), we choose them in the following way:

$$\rho^R_{\mathbf{x}_I} = \begin{cases} h\,\frac{\sqrt{2}+2}{2}, & \mathbf{x}_I \notin \partial\Omega_R,\ \mathrm{dist}(\mathbf{x}_I, \partial\Omega_R) < \frac{3}{2}h \\ h\,((3-p)\frac{\sqrt{2}+2}{2} + (p-2)\frac{2+\sqrt{5}}{2}), & \mathbf{x}_I \notin \partial\Omega_R,\ \mathrm{dist}(\mathbf{x}_I, \partial\Omega_R) \geq \frac{3}{2}h \end{cases}, \tag{4.41}$$

$$\rho^H_{\mathbf{x}_I} = \begin{cases} h\,\frac{1+\sqrt{3}}{2}, & \mathbf{x}_I \notin \partial\Omega_H,\ \mathrm{dist}(\mathbf{x}_I, \partial\Omega_H) < \frac{3}{2}h \\ h\,((2-p)\frac{1+\sqrt{3}}{2} + (p-1)\frac{\sqrt{3}+3}{2}), & \mathbf{x}_I \notin \partial\Omega_H,\ \mathrm{dist}(\mathbf{x}_I, \partial\Omega_H) \geq \frac{3}{2}h \end{cases} \tag{4.42}$$

in which $\mathrm{dist}(\mathbf{x}_I, B) \equiv \min_{\mathbf{y}\in B} \|\mathbf{x}_I - \mathbf{y}\|$ represents the distance between $\mathbf{x}_I$ and the closed set B as usual and $p = 2, 3$ and $p = 1, 2$, respectively, in (4.41) and (4.42). The dilation function values on the interior nodes, taken by (4.41) and (4.42), make the node sets Λ_R and Λ_H be the possibly layered. Furthermore, every $\Lambda_{\rho_{\mathbf{x}_I}}(\mathbf{x}_I)$ for interior node $\mathbf{x}_I$ becomes the set of Type I or Type II at $\mathbf{x}_I$.

On the boundary nodes for both cases, the dilation function values can be assigned arbitrarily but they must be large enough to ensure the inverse of the moment matrices at the nodes themselves. Actually, the dilation function values on the boundary nodes do not affect the subsequent theorems for the convergence proof.

From the construction of two triples, namely, $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$ and $(\Omega_H, \Lambda_H, \rho^H_{\mathbf{x}})$, we can see that both Λ_R and Λ_H are the possibly layered and hence the two triples become the proper triples attributed to (4.24) and (4.36). Although we could not say how many possibly layered node distributions exist, we find at least two types of the possibly layered set of nodes.

4.4. Discrete maximum principle on the set of Type I or Type II. For these locally (p,q)-layered nodes of Type I and Type II, we should pay attention to the discretized form (4.24) and (4.36). If the window function is suitably chosen so that all the coefficients A_k may be strictly positive, then we can prove the discrete maximum principle at the center node.

LEMMA 3. *Let* $(\Omega, \Lambda, \rho_{\mathbf{x}})$ *be the triple. If the local node set* $\Lambda(\mathbf{x}_I) \subset \Lambda$ *is of either Type I or Type II at* $\mathbf{x}_I$*, then the strong meshfree Laplacian operator* Δ^{ρ} *satisfies the*

discrete maximum principle at the center node $\mathbf{x}_I$.

Proof. Let $\mathbf{x}_I \in \Lambda$ be the center node of $\Lambda(\mathbf{x}_I)$ that is either Type I or Type II at $\mathbf{x}_I$. Then the set $\Lambda(\mathbf{x}_I)$ is obviously the locally (p,q)-layered where $p = 2, 3$ for $q = 4$ or $p = 1, 2$ for $q = 6$. The discrete Laplacian shape functions have been calculated in (4.24) and (4.36) for both cases. First, we will show that all the coefficients A_K of the representation formula in Lemmas 1 and 2 are strictly positive. In the case of Type I which is the $(p,4)$-layered, we claim that, when $p = 3$,

$$1 - 4\,\delta_2 - 12\,\delta_3 > 0, \quad 1 + 2\,\delta_1 - 4\,\delta_3 > 0, \quad 1 + 3\,\delta_1 + 2\,\delta_2 > 0 \tag{4.43}$$

and, when $p = 2$,

$$1 - 4\,\delta_2 > 0, \quad 1 + 2\,\delta_1 > 0. \tag{4.44}$$

If $\delta_2 < \frac{1}{16}$ and $\delta_3 < \frac{1}{36}$, then all the left terms in (4.43) stay positive. Indeed, when $p = 3$, from (4.17) we have $\delta_2 = W(\frac{2\sqrt{2}}{2+\sqrt{5}}) < \frac{1}{16}$ and $\delta_3 = W(\frac{4}{2+\sqrt{5}}) < \frac{1}{36}$. In the other case when $p = 2$, it is true from (4.17) that $\delta_2 = W(\frac{2\sqrt{2}}{2+\sqrt{2}}) < \frac{1}{16}$. Therefore, we are done with the proof for the case of the locally $(p,4)$-layered nodes $(p = 2, 3)$.

On the other hand, when $\Lambda(\mathbf{x}_I)$ is of Type II which is the locally $(p,6)$-layered, all the coefficients A_1 and A_2 of the representation formula in Lemma 2 are positive since we know from (4.30) that $\delta_2 = W(\frac{2\sqrt{3}}{3+\sqrt{3}}) < \frac{1}{16}$ when $p = 2$. For the case when $\Lambda(\mathbf{x}_I)$ is the locally $(1,6)$-layered at $\mathbf{x}_I$, we trivially have $A_1 > 0$.

Let $\Lambda(\mathbf{x}_I)$ be the set of Type I or Type II at $\mathbf{x}_I \in \Lambda$ as mentioned in this Lemma. To prove the discrete maximum principle, suppose $(\Delta^\rho v)(\mathbf{x}_I) \geq 0$ for some $v \in V$. Due to the window function of type (2.5), the coefficients A_i in Lemmas 1 or 2 are proved to be positive in the above. From the positivity of the coefficients of the representation formula in both Lemmas 1 and 2, it never happens under this assumption that the center nodal value v_I of v at $\mathbf{x}_I$ is strictly greater than all the other nodal values v_K at the node $\mathbf{x}_K \in \Lambda_{\rho_{\mathbf{x}_I}}(\mathbf{x}_I)$ and therefore we have

$$v_I \leq \max_{\mathbf{x}_K \in \Lambda(\mathbf{x}_I) \setminus \{\mathbf{x}_I\}} v_K. \tag{4.45}$$

If the equality holds, then the event $v_I > v_K$ for some $\mathbf{x}_K, K \neq I$ makes $(\Delta^\rho v)(\mathbf{x}_I) < 0$. Hence, all v_K's must be equal to v_I. Therefore, the operator Δ^ρ satisfies the discrete maximum principle at the node $\mathbf{x}_I$ and this completes the proof. ∎

4.5. A priori estimate for the strong meshfree Laplacian operator. For the set of nodes on which the discrete maximum principle holds, we can obtain the general results in the meshfree regime.

LEMMA 4. *Let* $(\Omega, \Lambda, \rho_{\mathbf{x}})$ *be a proper triple. Assume the operator* Δ^ρ *satisfies the discrete maximum principle on a finite subset* $A \subset \Lambda$. *Then we have the following inequality:*

$$\max_{\mathbf{x}_J \in A} v_J \leq \max_{\mathbf{x}_K \in A^* \setminus A} v_K \tag{4.46}$$

whenever $v \in V$ *and* $\Delta^\rho v \geq 0$ *on* A.

Proof. Let us assume $v \in V$ and $\Delta^\rho v \geq 0$ on A. Suppose the maximum of nodal values over A occurs at the node $\mathbf{x}_{K^*} \in A$; that is,

$$v_{K^*} = \max_{\mathbf{x}_J \in A} v_J. \tag{4.47}$$

Then only two cases are possible. The first case is when $\Lambda(\mathbf{x}_{K^*})\backslash A \neq \emptyset$. In this case, from the maximum principle we have nothing to prove. In the other case, we have $\Lambda(\mathbf{x}_{K^*}) \subset A$. For this case, all v_K's in $\Lambda(\mathbf{x}_{K^*})$ are the same as v_{K^*}. If we set $A_0 \equiv \Lambda(\mathbf{x}_{K^*})$, then we can construct the set $A_1 \equiv A_0^*$ strictly larger that A_0 (i.e., A_1 contains at least one node not in A_0). If $A_1\backslash A \neq \emptyset$, then this lemma is proved. If not, all coefficients v_K in A_1 must have the same value v_{K^*}. Continuing this process, we can construct $A_2 = A_1^*, A_3 = A_2^*, \ldots$. However, this process has to stop in a number of finite steps since the number of nodes in A is finite. Therefore, we have proven this lemma. ∎

THEOREM 1 (a priori estimate for the strong meshfree Laplacian operator). *Let $(\Omega, \Lambda, \rho_{\mathbf{x}})$ be a proper triple. Assume the strong meshfree Laplacian operator Δ^ρ satisfies the discrete maximum principle on a finite subset $A \subset \Lambda$. Then, we have the following a priori estimate*

$$\|v\|_{\infty,A} \leq C(A)\,\|\Delta^\rho v\|_{\infty,A} + \|v\|_{\infty,A^*\backslash A} \quad \textit{whenever} \quad v \in V, \tag{4.48}$$

where $C(A) = \min_{\mathbf{x}_*} \max_{\mathbf{x}_L \in A^*\backslash A} \frac{1}{4}\|\mathbf{x}_L - \mathbf{x}_*\|^2$.

Proof. Let $v \in V$ be assumed to be the nodal function on Λ. Then from the definition of the strong meshfree Laplacian Δ^ρ, we have

$$(\Delta^\rho v)(\mathbf{x}_K) = \sum_{\mathbf{x}_J \in \Lambda} v_J\, \psi_J^{\Delta}(\mathbf{x}_K), \quad \mathbf{x}_K \in A. \tag{4.49}$$

Obviously we see that

$$-\|\Delta^\rho v\|_{\infty,A} \leq (\Delta^\rho v)(\mathbf{x}_K) \leq \|\Delta^\rho v\|_{\infty,A} \quad \text{for any} \quad \mathbf{x}_K \in A. \tag{4.50}$$

Owing to the reproducing property for polynomials up to second order, we have, for any $\mathbf{x}_K \in A$,

$$\Delta^\rho \left(\|\Delta^\rho v\|_{\infty,A}\, i\left(\frac{1}{4}\|\mathbf{x} - \mathbf{x}_*\|^2\right)\right) = \sum_{\mathbf{x}_J \in \Lambda} \left(\|\Delta^\rho v\|_{\infty,A}\, \frac{1}{4}\|\mathbf{x}_J - \mathbf{x}_*\|^2 \right) \psi_J^{\Delta}(\mathbf{x}) \tag{4.51}$$

$$= \|\Delta^\rho v\|_{\infty,A}, \tag{4.52}$$

where $\mathbf{x}_*$ is an arbitrary point. The first equality (4.51) comes from the definition of the operator Δ^ρ on V. The last identity (4.52) enables us to derive the following inequalities due to (4.50). For any $\mathbf{x}_K \in A$,

$$\sum_{\mathbf{x}_J \in \Lambda} \left[v_J + \left(\|\Delta^\rho v\|_{\infty,A}\, \frac{1}{4}\|\mathbf{x}_J - \mathbf{x}_*\|^2 \right)\right] \psi_J^{\Delta}(\mathbf{x}_K) \geq 0, \tag{4.53}$$

$$\sum_{\mathbf{x}_J \in \Lambda} \left[-v_J + \left(\|\Delta^\rho v\|_{\infty,A}\, \frac{1}{4}\|\mathbf{x}_J - \mathbf{x}_*\|^2 \right)\right] \psi_J^{\Delta}(\mathbf{x}_K) \geq 0. \tag{4.54}$$

From the discrete maximum principle (4.46) in Lemma 4 and both inequalities (4.53) and (4.54), we can conclude that

$$v_K \leq \max_{\mathbf{x}_L \in A^*\backslash A} \left(v_L + \left(\|\Delta^\rho v\|_{\infty,A}\, \frac{1}{4}\|\mathbf{x}_L - \mathbf{x}_*\|^2 \right)\right), \tag{4.55}$$

$$-v_K \leq \max_{\mathbf{x}_L \in A^*\backslash A} \left(-v_L + \left(\|\Delta^\rho v\|_{\infty,A}\, \frac{1}{4}\|\mathbf{x}_L - \mathbf{x}_*\|^2 \right)\right) \tag{4.56}$$

for all $\mathbf{x}_K \in A$. Therefore, the following estimate holds:

$$|v_K| \le \max_{\mathbf{x}_L \in A^* \setminus A} |v_L| + \max_{\mathbf{x}_L \in A^* \setminus A} \left(\|\Delta^\rho v\|_{\infty, A} \frac{1}{4} \|\mathbf{x}_L - \mathbf{x}_*\|^2 \right). \tag{4.57}$$

This completes the proof. ∎

5. Error estimate for Poisson problem on specific domains Ω_R and Ω_H. From here, we will achieve the convergence of the numerical solutions using the meshfree point collocation approach (**DP**) in (3.1) for the Poisson equation with Dirichlet data on two specific domains—Ω_R and Ω_H. Through the convergence proof, we can also understand the basic phenomena on the strong meshfree Laplacian operator and can view the structure of the meshfree approximations.

For the numerical solution of the problem (**DP**), the meshfree point collocation scheme is proposed in the manner of (3.9). The existence and uniqueness of the numerical solutions of (**DP**) follows immediately from the a priori estimate in Theorem 1.

THEOREM 2 (existence and uniqueness). *Assume that* $(\Omega, \Lambda, \rho_{\mathbf{x}})$ *is either* $(\Omega_R, \Lambda_R, \rho_{\mathbf{x}}^R)$ *or* $(\Omega_H, \Lambda_H, \rho_{\mathbf{x}}^H)$. *Then there exists the unique solution of the problem* (**DP**) *on* V.

Proof. Let us introduce the linear mapping $\widehat{\Delta^\rho} : V \to V$ defined by

$$(\widehat{\Delta^\rho}\, v)(\mathbf{x}_K) \equiv \begin{cases} (\Delta^\rho v)(\mathbf{x}_K), & \mathbf{x}_K \in \Lambda^o \\ v_K, & \mathbf{x}_K \in \Lambda^b \end{cases} \quad \text{for all } v \in V, \tag{5.1}$$

where $\Lambda^o \equiv \Lambda \cap \Omega$ and $\Lambda^b \equiv \Lambda \setminus \Lambda^o$. Then our discrete problem (**DP**) is equivalent to the following:

$$\text{find } v \in V \text{ such that } \quad \widehat{\Delta^\rho}\, v = \begin{cases} i(f) & \text{on } \Lambda^o \\ g & \text{on } \Lambda^b \end{cases}. \tag{5.2}$$

Since Λ is the possibly layered, the discrete maximum principle holds on Λ^o. Applying the a priori estimate in Theorem 1 to the problem (5.2), we have

$$\|v\|_{\infty, \Lambda \cap \Omega} \le C \, \|i(f)\|_{\infty, \Lambda \cap \Omega} + \|g\|_{\infty, \Lambda \cap \partial\Omega}. \tag{5.3}$$

We claim that the linear mapping $\widehat{\Delta^\rho}$ is one-to-one and onto. It suffices to show that the mapping is one-to-one since solution space V has finite dimension.

Suppose $\widehat{\Delta^\rho}\, v = \mathbf{0}$. This means that f and g become zero on the right-hand side of (5.3). Consequently, we have $\|v\|_{\infty, \Lambda \cap \Omega} = 0$ and hence $v = 0$ on $\Lambda \cap \Omega$. This implies $v = 0$ on Λ since $g = 0$ on $\Lambda \cap \partial\Omega$. Therefore, the mapping $\widehat{\Delta^\rho}$ is injective. From the fact that $\operatorname{Im} \widehat{\Delta^\rho} = (\operatorname{Ker} \widehat{\Delta^\rho})^\perp = V$, we also are done with the surjective proof. ∎

Furthermore, the following error estimate of the unique nodal solution of the problem (**DP**) holds on two specific domains Ω_R and Ω_H under the regularity assumption of the continuous problem (**CP**).

THEOREM 3. *Let* $(\Omega, \Lambda, \rho_{\mathbf{x}})$ *be either the triple* $(\Omega_R, \Lambda_R, \rho_{\mathbf{x}}^R)$ *or* $(\Omega_H, \Lambda_H, \rho_{\mathbf{x}}^H)$. *Assume* $u \in C^0(\overline{\Omega}) \cap C^{s,\alpha}(\Omega)\, (s = 3, 4)$ *is the classical solution of Poisson problem* (**CP**) *with Dirichlet data and* $u_h \in V$ *is the nodal solution of the discrete Poisson problem* (**DP**) *on the node set* Λ *corresponding to* Ω. *If* $\Lambda \cap \Omega$ *is the interior nodes of* Λ, *then we have the following error estimate:*

$$\|i(u) - u_h\|_{\infty, \Lambda \cap \Omega} \le K \, h^{s-2} \, \|u\|_{C^{s,\alpha}(\Omega)} \tag{5.4}$$

for some constant $K > 0$ independent of h.

Proof. We note that the set of nodes $\Lambda \equiv \Lambda_R$ (or Λ_H) of the proper triple $(\Omega_R, \Lambda_R, \rho_{\mathbf{x}}^R)$ (or $(\Omega_H, \Lambda_H, \rho_{\mathbf{x}}^H)$) is obviously the possibly layered from the construction. Thus we can calculate the operator Δ^ρ at every interior node $\mathbf{x}_I \in \Lambda \cap \Omega$. Let $\Lambda^o \equiv \Lambda \cap \Omega$ be the interior nodes of Λ. If $u_h \in V$ is the nodal solution of (**DP**) and $u \in C^0(\overline{\Omega}) \cap C^{s,\alpha}(\Omega)\,(s = 3, 4)$ is the solution of (**CP**), then we can derive the error equation on Λ^o such that

$$(\Delta^\rho u_h)(\mathbf{x}_I) - \Delta u(\mathbf{x}_I) = 0 \quad \text{for all} \quad \mathbf{x}_I \in \Lambda^o. \tag{5.5}$$

From the error equation (5.5), we can obtain

$$\sum_{\mathbf{x}_J \in \Lambda} (u_J^h - u(\mathbf{x}_J))\, \psi_J^\Delta(\mathbf{x}_I) = \Delta u(\mathbf{x}_I) - \sum_{\mathbf{x}_J \in \Lambda} u(\mathbf{x}_J)\, \psi_J^\Delta(\mathbf{x}_I) \tag{5.6}$$

for all $\mathbf{x}_I \in \Lambda^o$. Since the domain Ω is convex, we can obtain the following Taylor expansions for $u(\mathbf{x})$ at $\mathbf{x}_I \in \Lambda^o$. For every $\mathbf{x}_J \in \Lambda(\mathbf{x}_I)$,

$$\begin{aligned} u(\mathbf{x}_J) = &\sum_{|\beta| \le s-1} \frac{1}{\beta!} (\mathbf{x}_J - \mathbf{x}_I)^\beta D^\beta u(\mathbf{x}_I) \\ &+ \sum_{|\beta| = s} \frac{1}{\beta!} \int_0^1 (1-\tau)^{s-1} D^\beta u(\mathbf{x}_I + \tau(\mathbf{x}_J - \mathbf{x}_I))\, d\tau (\mathbf{x}_J - \mathbf{x}_I)^\beta. \end{aligned} \tag{5.7}$$

Since $\Lambda(\mathbf{x}_I)$ is the locally (p, q)-layered $(q = 4, 6)$ at $\mathbf{x}_I \in \Lambda^o$, we can observe the symmetric node structure such that $-(\mathbf{x}_J - \mathbf{x}_I)$ and $\mathbf{x}_J - \mathbf{x}_I$ are on the same layer for all $\mathbf{x}_J \in \Lambda(\mathbf{x}_I)$. This implies that, when $|\beta| = 3$,

$$\sum_{\mathbf{x}_J \in \Lambda} (\mathbf{x}_J - \mathbf{x}_I)^\beta\, \psi_J^\Delta(\mathbf{x}_I) = 0. \tag{5.8}$$

Thus, inserting the expansions (5.7) into the right-hand side of (5.6), the following is obtained from the second order reproducing property and the symmetric factor (5.8):

$$\sum_{\mathbf{x}_J \in \Lambda} (u_J^h - u(\mathbf{x}_J))\, \psi_J^\Delta(\mathbf{x}_I) = \sum_{\mathbf{x}_J \in \Lambda} c_{IJ} \psi_J^\Delta(\mathbf{x}_I) \quad \text{for all} \quad \mathbf{x}_I \in \Lambda^o, \tag{5.9}$$

where the coefficients c_{IJ} are defined as

$$c_{IJ} = -\sum_{|\beta| = s} \frac{1}{\beta!} \int_0^1 (1-\tau)^{s-1} D^\beta u(\mathbf{x}_I + \tau(\mathbf{x}_J - \mathbf{x}_I))\, d\tau (\mathbf{x}_J - \mathbf{x}_I)^\beta. \tag{5.10}$$

Since the left-hand side of (5.9) is the image of the strong meshfree Laplacian operator Δ^ρ of $u^h - i(u) \in V$, the a priori estimate (4.48) due to the maximum principle on Λ^o leads to the following estimate

$$\max_{\mathbf{x}_J \in \Lambda^o} |u_J^h - u(\mathbf{x}_J)| \le C(\Lambda^o) \max_{\mathbf{x}_I \in \Lambda^o} \sum_{\mathbf{x}_J \in \Lambda(\mathbf{x}_I)} |c_{IJ}| \left|\psi_J^\Delta(\mathbf{x}_I)\right| + \max_{\mathbf{x}_J \in \Lambda^{o*} \backslash \Lambda^o} |u_J^h - u(\mathbf{x}_J)|. \tag{5.11}$$

In the case of the node distributions assumed, we know that $\Lambda^b = \Lambda^{o*} \backslash \Lambda^o$ and $u_J^h - u(\mathbf{x}_J) = 0$ on Λ^b because of the Dirichlet boundary conditions and hence the second

term on the right-hand side of the inequality (5.11) vanishes. For the estimate of the first term on the right-hand side of the inequality (5.11), we need the estimate of $|c_{IJ}|$ for all $\mathbf{x}_J \in \Lambda(\mathbf{x}_I)$ as follows. For each $\mathbf{x}_I \in \Lambda^o$,

$$|c_{IJ}| \le \left(\max_{|\beta|=s} \sup_{\mathbf{x}\in\Omega} |D^\beta u(x)|\right) \left(\max_{\mathbf{x}_J\in\Lambda(\mathbf{x}_I)} |\mathbf{x}_J - \mathbf{x}_I|\right)^s \frac{1}{s} \sum_{|\beta|=s} \frac{1}{\beta!} \tag{5.12}$$

$$\le K_s \, \|u\|_{C^{s,\,\alpha}} \, \|i(\rho)\|^s_{\infty,\Lambda(\mathbf{x}_I)}, \tag{5.13}$$

where $K_s = \frac{1}{s}\sum_{|\beta|=s} \frac{1}{\beta!}$ and ρ is the dilation function. Therefore, we have the following error bound

$$\max_{\mathbf{x}_J\in\Lambda^o} |u^h_J - u(\mathbf{x}_J)| \le K_s \, C(\Lambda^o) \, \|i(\rho)\|^s_{\infty,\Lambda^o} \, \|u\|_{C^{s,\,\alpha}} \max_{\mathbf{x}_I\in\Lambda^o} \sum_{\mathbf{x}_J\in\Lambda(\mathbf{x}_I)} \left|\psi^\Delta_J(\mathbf{x}_I)\right|. \tag{5.14}$$

On the other hand, the constant $C(\Lambda^o)$ is bounded by the diameter of the domain Ω, and the dilation function ρ in the assumed triple $(\Omega, \Lambda, \rho_{\mathbf{x}})$ satisfies

$$h < \|i(\rho)\|_{\infty,\Lambda^o} < C\, h \tag{5.15}$$

for some constant C independent of h. Furthermore, from Lemmas 1 and 2,

$$\sum_{\mathbf{x}_J\in\Lambda(\mathbf{x}_I)} \left|\psi^\Delta_J(\mathbf{x}_I)\right| \le \frac{8}{h^2}. \tag{5.16}$$

Consequently, we obtain the error estimate derived from (5.14):

$$\|i(u) - u_h\|_{\infty,\Lambda\cap\Omega} \le K\, h^{s-2} \, \|u\|_{C^{s,\alpha}(\Omega)} \tag{5.17}$$

for some $K > 0$ independent of h. □

Remark 4. As seen in the proof of Theorem 3, the convergence order of the numerical solution to the exact one can be proven only to be 2, although the regularity index s of the solution becomes greater than 4. The higher order of basis polynomials in the fast version of the generalized moving least square meshfree approximation is directly related to a lift in the convergence order (see [8]). Its proof seems to need the boundary error estimate for the numerical solution without the discrete maximum principle, while the interior error estimate is the same as ours.

The error ratio of about 4 in Table 5.1 implies the second order convergence even for the less-regularity case. The numerical result not only attests the validation of the error estimate but also shows the numerical scheme proposed could be more accurate than we anticipated. A numerical example is proposed to verify the theoretical convergence result. The solution $u(x, y)$ is assumed to be defined on both domains $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$ and $(\Omega_H, \Lambda_H, \rho^H_{\mathbf{x}})$ as follows:

$$u(x, y) = e^{x+y-1} \left|x - \frac{1}{2}\right| \left(x - \frac{1}{2}\right)^2. \tag{5.18}$$

Applying the Laplacian operator to this solution, the corresponding force is given as follows:

$$f(x, y) = 2 \left|x - \frac{1}{2}\right| e^{x+y-1} \left(x^2 + 2x + \frac{7}{4}\right). \tag{5.19}$$

TABLE 5.1

Numerically experimental result on the relative error $(\|u^h - i(u)\|_{\Lambda,\infty}/\|i(u)\|_{\Lambda,\infty})$ *and the convergence rate of the numerical solutions for* $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$ *and* $(\Omega_H, \Lambda_H, \rho^H_{\mathbf{x}})$*, where* Λ *is either* Λ_R *or* Λ_H*, and* ρ^R *and* ρ_H *are taken as the value* $1.8 * h$ *at each interior node so that* Λ_R *and* Λ_H *can be locally* $(2,4)$*-layered and locally* $(1,6)$*-layered, respectively.*

h	$(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$	Error ratio	h	$(\Omega_H, \Lambda_H, \rho^H_{\mathbf{x}})$	Error ratio
$\frac{2}{20}$	4.2660×10^{-3}	—	$\frac{2}{10}$	1.1900×10^{-2}	—
$\frac{2}{40}$	1.0726×10^{-3}	3.98	$\frac{2}{20}$	2.9148×10^{-3}	4.08
$\frac{2}{80}$	2.6857×10^{-4}	4.00	$\frac{2}{40}$	7.1881×10^{-4}	4.06
$\frac{2}{160}$	6.7162×10^{-5}	4.00	$\frac{2}{80}$	1.7833×10^{-4}	4.03

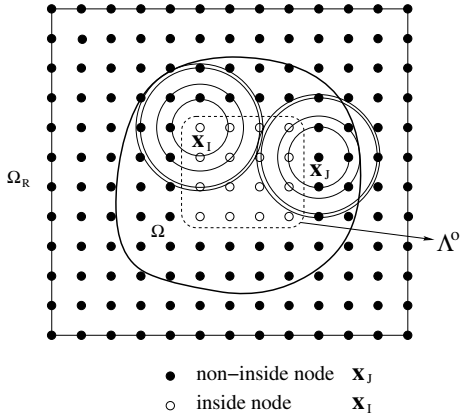


FIG. 5.1. *The immersed domain* Ω *in* Ω_R.

In this case, Dirichlet boundary condition on either $\partial\Omega_R$ or $\partial\Omega_H$ is presumed from the exact solution $u(x, y)$. The function u, in fact, belongs to $C^{2,1}$-class of functions on considered domains whose regularity is weaker than that stated in Theorem 3; nevertheless, the numerical example produces the second order convergence result as seen in Table 5.1.

6. Error estimate in the general domain immersed in Ω_R or Ω_H. We will try to analyze the convergence of our discrete problem (**DP**) with the boundary condition zero on a domain Ω which is immersed in the larger domain, for example, $\Omega \subset \hat{\Omega}_R$ (or $\Omega \subset \hat{\Omega}_H$) where $\hat{\Omega}_R$ (or $\hat{\Omega}_H$) is the image domain transformed from Ω_R (or Ω_H) by the similarity map. For brevity, we will rename it by Ω_R (or Ω_H).

Let $\Lambda \equiv \Lambda_R$ be the set of nodes in the triple $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$. Assume that the nodal solution space V in this case is defined on Λ. As shown in Figure 5.1, we separate the set of nodes into two parts—the set Λ^o of inside nodes and the set Λ^b of non-inside nodes which are defined as the following:

$$\Lambda^o \equiv \{\mathbf{x}_J \in \Lambda \,|\, \Lambda(\mathbf{x}_J) \subset \Omega\}, \quad \Lambda^b \equiv \Lambda \setminus \Lambda^o. \tag{6.1}$$

The inside node $\mathbf{x}_I \in \Lambda$ implies that the $\rho_{\mathbf{x}_I}$-neighbor nodes are contained in the open set Ω while the nonside node is anything else. The following is the immersed meshfree

Poisson problem:

$$(\mathbf{IMP})\begin{cases} u_h \in V_0 \equiv \{v_J \in \mathbb{R} \,|\, v_K = 0 \text{ for all } \mathbf{x}_K \in \Lambda^b\} \subset V \\ \Delta^\rho u_h = i(f) \quad \text{on } \Lambda^o \end{cases}. \tag{6.2}$$

Let the solution u of the Poisson problem (**CP**) with zero Dirichlet data belong to $C^0(\overline{\Omega}) \cap C^{3,\alpha}(\Omega)$ and $u_h \in V_0$ be the nodal solution of the immersed meshfree Poisson problem (**IMP**). If we extend u to $\Omega_R \setminus \Omega$ by zero, then we conjecture that

$$\|i(u) - u_h\|_{\infty,\Lambda\cap\Omega} \le K\, h\, \|u\|_{C^{3,\alpha}(\Omega)}, \tag{6.3}$$

where the constant K is independent of h.

THEOREM 4 (existence and uniqueness). *Let* $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$ *and* $(\Omega_H, \Lambda_H, \rho^H_{\mathbf{x}})$ *be the triples. Assume* Ω *is immersed in either* Ω_R *or* Ω_H*. Then there exists the unique solution of the problem* (**IMP**) *on* V*.*

The proof of the theorem is similar to that of Theorem 2 and thus it is omitted.

In order to prove the convergence result (6.3) for the immersed meshfree Poisson problem (**IMP**), let $(\Omega_R, \Lambda_R, \rho^R_{\mathbf{x}})$ be the triple defined in section 5. First, we have the following error:

$$\|i(u) - u_h\|_{\infty,\Lambda^{o*}\setminus\Lambda^o} = \|i(u)\|_{\infty,\Lambda^{o*}\setminus\Lambda^o} \tag{6.4}$$

since $\Lambda^{o*} \setminus \Lambda^o \subset \Omega$ and $u_h(\mathbf{x}_J) = 0$ on it. Through the similar procedure to the proof of Theorem 3 and from the fact that $\Lambda(\mathbf{x}_I) \subset \Omega$ for any $\mathbf{x}_I \in \Lambda^o$, we can obtain the following error equation:

$$\sum_{\mathbf{x}_J\in\Lambda} (u^h_J - u(\mathbf{x}_J))\, \psi^\Delta_J(\mathbf{x}_I) = \sum_{\mathbf{x}_J\in\Lambda} c_{IJ}\psi^\Delta_J(\mathbf{x}_I) \quad \text{for all} \quad \mathbf{x}_I \in \Lambda^o, \tag{6.5}$$

where the coefficients c_{IJ} are calculated as follows:

$$c_{IJ} = -\sum_{|\beta|=3} \frac{1}{\beta!} \int_0^1 (1-\tau)^2 D^\beta u(\mathbf{x}_I + \tau(\mathbf{x}_J - \mathbf{x}_I))\, d\tau (\mathbf{x}_J - \mathbf{x}_I)^\beta. \tag{6.6}$$

From a priori estimate in Theorem 1 due to the discrete maximum principle on Λ^o and from the identity (6.4), we obtain the following estimates:

$$\begin{aligned} \|i(u) - u_h\|_{\infty,\Lambda^o} &\le C(\Lambda^o) \max_{\mathbf{x}_I\in\Lambda^o} \left| \sum_{\mathbf{x}_J\in\Lambda} c_{IJ}\psi^\Delta_J(\mathbf{x}_I) \right| + \|i(u) - u_h\|_{\infty,\Lambda^{o*}\setminus\Lambda^o} \\ &\le K_1\, h\, \|u\|_{C^{3,\alpha}(\Omega)} + \|i(u)\|_{\infty,\Lambda^{o*}\setminus\Lambda^o} \end{aligned} \tag{6.7}$$

for some constant $K_1 > 0$ independent of h. On the other hand, let us pay attention to the fact that

$$\|i(u) - u_h\|_{\infty,\Lambda^b\cap\Omega} = \|i(u)\|_{\infty,\Lambda^b\cap\Omega}. \tag{6.8}$$

Then, from this fact and (6.7), the nodal error on the nodes in Ω is bounded by

$$\|i(u) - u_h\|_{\infty,\Lambda\cap\Omega} \le K_1\, h\, \|u\|_{C^{3,\alpha}(\Omega)} + \|i(u)\|_{\infty,\Lambda^b\cap\Omega} \tag{6.9}$$

since $\Lambda^{o*} \setminus \Lambda^o \subset \Lambda^b \cap \Omega$. Here, the second term on the right-hand side of (6.9) is bounded by

$$\|i(u)\|_{\infty,\Lambda^b\cap\Omega} \le \left(\max_{\mathbf{x}_K\in\Lambda^b\cap\Omega} \operatorname{dist}(\mathbf{x}_K,\partial\Omega)\right) \|u\|_{C^{1,\alpha}(\Omega)} \le K_2\, h\, \|u\|_{C^{3,\alpha}(\Omega)} \tag{6.10}$$

for some constant $K_2 > 0$ independent of h. Therefore, we have obtained the following theorem.

THEOREM 5. *Let $\Omega \subset \mathbb{R}^2$ be an open bounded domain which is immersed in either Ω_R or Ω_H. Assume that $u \in C^0(\overline{\Omega}) \cap C^{3,\alpha}(\Omega)$ is the solution of* (**CP**) *and $u_h \in V_0$ is the nodal solution of* (**IMP**). *Then we have the following error estimate:*

$$\|i(u) - u_h\|_{\infty,\Lambda\cap\Omega} \le K\, h\, \|u\|_{C^{3,\alpha}(\Omega)}, \tag{6.11}$$

where the node set Λ is either Λ_R or Λ_H and K is constant independent of h.

7. Conclusion. The generalized moving least square approximation is introduced, and based on this, we can define the strong meshfree Laplacian operator in the sense of a point collocation strategy. From the mathematical point of view, the discrete maximum principle for the strong meshfree Laplacian operator is presented for several types of layered node distributions. Using this principle, we perform convergence analysis for the nodal solutions of the Poisson problem with Dirichlet data on the boundary. As a result, second order convergence is achieved on the specific nodes in two typical domains, while the generally shaped domain immersed in these domains produces first order convergence of the nodal solution. An a priori estimate for the strong Laplacian operator in the meshfree regime is newly obtained via the discrete maximum principle and it is located in the core of the convergence proof together with the point collocation scheme proposed in this paper.

Appendix I: Generalized moving least square reproducing operators. For a given window function $W(\mathbf{x})$ and a dilation function $\rho_{\mathbf{x}}$, we find the vector $\mathbf{a}$ to minimize the following weighted square functional at $\bar{\mathbf{x}} \in \overline{\Omega}$:

$$J(\mathbf{a};\bar{\mathbf{x}},u) \equiv \sum_{\mathbf{x}_I\in\Lambda} |u(\mathbf{x}_I) - \mathbf{U}_m^{\rho_{\bar{\mathbf{x}}}}(\mathbf{x}_I;\bar{\mathbf{x}},\mathbf{a})|^2\, W\left(\frac{\mathbf{x}_I - \bar{\mathbf{x}}}{\rho_{\bar{\mathbf{x}}}}\right), \tag{7.1}$$

where $u(\mathbf{x})$ is a continuous function defined in $\overline{\Omega}$ and $\mathbf{U}_m^{\rho_{\bar{\mathbf{x}}}}(\mathbf{x};\bar{\mathbf{x}},\mathbf{a}) \equiv \mathbf{B}_m(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\rho_{\bar{\mathbf{x}}}}) \cdot \mathbf{a}$. Then the minimizer $\mathbf{a}$ should be a function of $\bar{\mathbf{x}}$ and u, and we can make the following approximation operators for u by limiting process

$$D_{m,\rho_{\mathbf{x}}}^{\beta_k} u(\mathbf{x}) \equiv \lim_{\bar{\mathbf{x}}\to\mathbf{x}} D_{\mathbf{x}}^{\beta_k} \mathbf{U}_m^{\rho_{\bar{\mathbf{x}}}}(\mathbf{x};\bar{\mathbf{x}},\mathbf{a}(\bar{\mathbf{x}},u)), \quad |\beta_k| \le m. \tag{7.2}$$

As a matter of fact, the operators $D_{m,\rho_{\mathbf{x}}}^{\beta_k}$ are linear in $u(\mathbf{x})$. We call the operator $D_{m,\rho_{\mathbf{x}}}^{\beta}$ ($|\beta| \le m$) the βth meshfree approximated derivative operator equipped with $\rho_{\mathbf{x}}$.

Suppose $\{u_I(\mathbf{x}) \,|\, u_I(\mathbf{x}_J) = \delta_{IJ},\ \mathbf{x}_I, \mathbf{x}_J \in \Lambda\}$ is a set of continuous functions. We define the following functions:

$$\psi_I^{\rho_{\mathbf{x}},[\beta_k]}(\mathbf{x}) \equiv D_{m,\rho_{\mathbf{x}}}^{\beta_k} u_I(\mathbf{x}). \tag{7.3}$$

Then the functions $\psi_I^{\rho_{\mathbf{x}},[\beta_k]}(\mathbf{x})$ can be characterized as follows:

$$\begin{pmatrix} {\rho_{\mathbf{x}}}^{|\beta_1|}\,\psi_I^{\rho_{\mathbf{x}},[\beta_1]}(\mathbf{x}) \\ {\rho_{\mathbf{x}}}^{|\beta_2|}\,\psi_I^{\rho_{\mathbf{x}},[\beta_2]}(\mathbf{x}) \\ \vdots \\ {\rho_{\mathbf{x}}}^{|\beta_L|}\,\psi_I^{\rho_{\mathbf{x}},[\beta_L]}(\mathbf{x}) \end{pmatrix} = J_{\mathbf{B}_m}(\mathbf{0})\, M^{\rho_{\mathbf{x}}}(\mathbf{x})^{-1}\, \mathbf{B}_m\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right) W\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right), \tag{7.4}$$

where $M^{\rho_{\mathbf{x}}}(\mathbf{x})$ is called the moment matrix and is defined such that

$$M^{\rho_{\mathbf{x}}}(\mathbf{x}) \equiv \sum_{\mathbf{x}_I\in\Lambda} \mathbf{B}_m\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right)\mathbf{B}_m^T\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right) W\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right). \tag{7.5}$$

We call the function $\psi_I^{\rho_{\mathbf{x}},[\beta]}(\mathbf{x})$ *the βth shape function* associated with the window function W and the dilation function $\rho_{\mathbf{x}}$, or briefly call it *the βth shape function* if no confusion arises. As a consequence of (7.3), the operator $D^{\beta_k}_{m,\rho_{\mathbf{x}}}$ defined by (7.2) can be rewritten as follows:

$$D^{\beta}_{m,\rho_{\mathbf{x}}} u(\mathbf{x}) = \sum_{\mathbf{x}_J\in\Lambda} u(\mathbf{x}_J)\,\psi_J^{\rho_{\mathbf{x}},[\beta]}(\mathbf{x}), \quad |\beta|\le m. \tag{7.6}$$

We also call this operator the βth meshfree approximated derivative operator and the following properties of this operator can be justified.

THEOREM 6 (generalized mth order consistency). *We have the following identities:*

$$\sum_{\mathbf{x}_I\in\Lambda} b_\alpha\left(\frac{\mathbf{x}_I-\mathbf{x}}{\rho_{\mathbf{x}}}\right)\psi_I^{\rho_{\mathbf{x}},[\beta]}(\mathbf{x}) = \frac{1}{{\rho_{\mathbf{x}}}^{|\beta|}}\,\frac{\partial^\beta}{\partial\mathbf{x}^\beta} b_\alpha(\mathbf{0}). \tag{7.7}$$

Proof. To the matrix equation (7.4) for *the βth shape functions*, multiplying $\mathbf{B}_m\left(\mathbf{x}_I-\mathbf{x}/\rho_{\mathbf{x}}\right)$ to the right on both sides and summing it over the whole nodes $\mathbf{x}_I$, we obtain the matrix equation. If we rewrite it in element-wise manner, then we have the resultant (7.7). □

The above theorem does not promise the βth meshfree approximated derivative operator to reproduce automatically all of the derivatives for the basis functions. However, for some useful class of functions including the polynomial class up to order m, we can have the generalized reproducing property which will play an important role in the convergence of approximations. For the class of the given basis functions to have such a generalized reproducing property, it is sufficient to satisfy the following condition.

COROLLARY 1 (sufficient condition for the generalized reproducing property). *Under the constant dilation function such that $\rho_{\mathbf{x}}\equiv\rho$, we assume that the basis functions satisfy the following relationships:*

$$b_\beta\left(\frac{\mathbf{y}}{\rho}\right) = \sum_{|\gamma|\le m} c_{\gamma\,\beta}\left(\frac{\mathbf{x}}{\rho}\right) b_\gamma\left(\frac{\mathbf{y}-\mathbf{x}}{\rho}\right), \quad |\beta|\le m \tag{7.8}$$

and the coefficient matrix is calculated from the equation

$$\left[c_{\alpha\,\beta}\left(\frac{\mathbf{x}}{\rho}\right)\right] \equiv C\left(\frac{\mathbf{x}}{\rho}\right) = J_{\mathbf{B}_m}(0)^{-1}\, J_{\mathbf{B}_m}\left(\frac{\mathbf{x}}{\rho}\right), \tag{7.9}$$

where $J_{\mathbf{B}_m}\left(\frac{\mathbf{x}}{\rho}\right)$ *is the Jacobian matrix defined as*

$$J_{\mathbf{B}_m}\left(\frac{\mathbf{x}}{\rho}\right) \equiv \left[\left(D^{\alpha} b_{\beta}\right)\left(\frac{\mathbf{x}}{\rho}\right)\right]. \tag{7.10}$$

Then the basis functions scaled by ρ *are exactly reproduced by the meshfree approximated derivative operators, i.e.,*

$$D^{\beta}_{m,\rho_{\mathbf{x}}} b_{\beta}\left(\frac{\mathbf{x}}{\rho}\right) = D^{\beta}_{\mathbf{x}} b_{\beta}\left(\frac{\mathbf{x}}{\rho}\right), \quad |\beta| \leq m. \tag{7.11}$$

Proof. Assume that $b_{\alpha}(\mathbf{x})$'s for $|\alpha| \leq m$ are the basis functions satisfying both conditions of (7.8) and (7.9). If we directly enforce (7.7) on these basis functions, then we obtain the result of (7.11). ∎

Corollary 1 provides us the opportunity of taking the general basis functions which can be reproduced in a dilated form. It is worth noting that the reproducing property does not happen in general if we take an arbitrary set of basis functions. That is why we propose the sufficient condition to ensure the reproducing condition for the dilated basis functions. According to the sufficient condition of (7.8) and (7.9) for the reproducing of basis functions, the class of polynomial basis up to order m can be shown to satisfy the exact reproducing property. That is, all of the derivatives of the basis itself are reproducible even in the case when involving the dilation function.

Corollary 2. *If we take the polynomials up to order* m *as basis functions, then the* β*th meshfree approximated derivative operator* $D^{\beta}_{m,\rho_{\mathbf{x}}}$ *is exactly the same as the differential operator* D^{β} *on the polynomial space up to order* m*. That is,*

$$D^{\beta}_{m,\rho_{\mathbf{x}}} u(\mathbf{x}) = D^{\beta}_{\mathbf{x}} u(\mathbf{x}) \tag{7.12}$$

whenever $u(\mathbf{x})$ *is a polynomial of order up to* m*.*

Proof. We can replace all $\rho_{\mathbf{x}}$ in Theorem 6 and all ρ in Corollary 1 with the number 1 for the case of polynomial basis up to order m. This fact suffices to prove this lemma. ∎

This corollary can be understood by recognizing that the βth meshfree approximated derivative operator $D^{\beta}_{m,\rho_{\mathbf{x}}}$ behaves in the same way as the exact derivative operator $D^{\beta}_{\mathbf{x}}$ at least on the polynomial function space up to order m.

Appendix II: Trigonometric identities. Let θ be an angle fixed. Then we have the following trigonometric identities for any natural number $n \geq 4$:

(7.13)
$$\sum_{k=0}^{n-1} \cos\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \sin\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \cos\left(\theta + k\,\frac{2\pi}{n}\right) \sin\left(\theta + k\,\frac{2\pi}{n}\right) = 0,$$

(7.14)
$$\sum_{k=0}^{n-1} \cos^2\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \sin^2\left(\theta + k\,\frac{2\pi}{n}\right) = \frac{n}{2},$$

(7.15)
$$\sum_{k=0}^{n-1} \cos^3\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \sin^3\left(\theta + k\,\frac{2\pi}{n}\right) = 0,$$

(7.16)
$$\sum_{k=0}^{n-1} \cos^2\left(\theta + k\,\frac{2\pi}{n}\right) \sin\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \cos\left(\theta + k\,\frac{2\pi}{n}\right) \sin^2\left(\theta + k\,\frac{2\pi}{n}\right) = 0,$$

(7.17)
$$\sum_{k=0}^{n-1} \cos^4\left(\theta + k\,\frac{2\pi}{n}\right) = \sum_{k=0}^{n-1} \sin^4\left(\theta + k\,\frac{2\pi}{n}\right) = \begin{cases} \frac{3}{8}n + \frac{1}{8}n\,\cos 4\theta, & n = 4 \\ \frac{3}{8}n, & n \neq 4 \end{cases},$$

(7.18)
$$\sum_{k=0}^{n-1} \cos^2\left(\theta + k\,\frac{2\pi}{n}\right) \sin^2\left(\theta + k\,\frac{2\pi}{n}\right) = \begin{cases} \frac{1}{8}n - \frac{1}{8}n\,\cos 4\theta, & n = 4 \\ \frac{1}{8}n, & n \neq 4 \end{cases},$$

(7.19)
$$\sum_{k=0}^{n-1} \cos^3\left(\theta + k\,\frac{2\pi}{n}\right) \sin\left(\theta + k\,\frac{2\pi}{n}\right) = \begin{cases} \frac{1}{8}n\,\sin 4\theta, & n = 4 \\ 0, & n \neq 4 \end{cases},$$

(7.20)
$$\sum_{k=0}^{n-1} \cos\left(\theta + k\,\frac{2\pi}{n}\right) \sin^3\left(\theta + k\,\frac{2\pi}{n}\right) = \begin{cases} -\frac{1}{8}n\,\sin 4\theta, & n = 4 \\ 0, & n \neq 4 \end{cases}.$$

Acknowledgments. We would like to give thanks particularly to Professor Shaofan Li for invaluable suggestions and David Farrel for correction of proofs, and we are also thankful to the referees for their valuable comments.

REFERENCES

[1] R. P. Agarwal, *Discrete polynomial interpolation, Green's functions, maximum principles, error bounds and boundary value problems*, Comput. Math. Appl., 25 (1993), pp. 3–39.
[2] I. Babuška and J. M. Melenk, *The partition of unity finite element method: Basic theory and applications*, Comput. Methods Appl. Mech. Engrg., 139 (1996), pp. 289–314.
[3] T. Belytschko, Y. Y. Lu, and L. Gu, *Element free Galerkin methods*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 229–256.
[4] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.

[5] L. B. Wahlbin, *Local behavior in finite element methods*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North Holland, 1991, pp. 353–522.
[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, New York, 1983.
[7] K. Ishihara, *Strong and weak discrete maximum principles for matrices associated with elliptic problems*, Linear Algebra Appl., 88-89 (1987), pp. 431–448.
[8] D. W. Kim and Y. Kim, *Point collocation methods using the fast moving least-square reproducing kernel approximation*, Internat. J. Numer. Methods Engrg., 56 (2003), pp. 1445–1464.
[9] D. W. Kim, Y. Kim, Y. C. Kim, H. S. Kim, S. Ahn, Y. Y. Park, and D.-W. Kim, *A miniaturized electron beam column simulation by the fast moving least square reproducing kernel point collocation method*, Jpn. J. Appl. Phys., 42 (2003), pp. 3842–3848.
[10] S. Li and W. K. Liu, *Moving least-square reproducing kernel method Part II: Fourier analysis*, Comput. Methods Appl. Mech. Engrg., 139 (1996), pp. 159–193.
[11] S. Li and W. K. Liu, *Reproducing kernel hierarchical partition of unity. Part I–formulation and theory*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 251–288.
[12] S. Li and W. K. Liu, *Reproducing kernel hierarchical partition of unity. Part II–applications*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 289–317.
[13] S. Li and W. K. Liu, *Meshfree particle methods and their applications*, Applied Mech. Rev., 54 (2002), pp. 1–34.
[14] S. Li, H. Lu, W. Han, and W. K. Liu, *Reproducing kernel element method. Part II. Globally conforming I^m/C^n hierarchies*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 953–987.
[15] W. K. Liu, Y. Chen, R. A. Uras, and C. T. Chang, *Generalized multiple scale reproducing kernel particle methods*, Comput. Methods Appl. Mech. Engrg., 139 (1996), pp. 91–157.
[16] W. K. Liu, W. Han, H. Lu, and S. Li, *Reproducing kernel element method. Part I. Theoretical formulation*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 933–951.
[17] W. K. Liu, S. Li, and T. Belytschko, *Moving least-square reproducing kernel method. Part I. Methodology and convergence*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 113–154.
[18] H. Lu, S. Li, D. C. Simkins, W. K. Liu, and J. Cao, *Reproducing kernel element method. Part III. Generalized enrichment and applications*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 989–1011.
[19] A. Mizukami, *A Petrov-Galerkin finite element method for convection-dominated flow: An accurate upwinding technique for satisfying the maximum principle*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 181–193.
[20] D. C. Simkins, S. Li, H. Lu, and W. K. Liu, *Reproducing kernel element method. Part IV. Globally conforming $C^n (n > 1)$ triangular hierarchy*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1013–1034.
[21] E. G. Yanik, *A discrete maximum principle for collocation methods*, Comput. Methods Appl. Mech. Engrg., 14 (1987), pp. 459–464.
[22] E. G. Yanik, *Sufficient conditions for a discrete maximum principle for higher order collocation methods*, Comput. Math. Appl., 17 (1989), pp. 1431–1434.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 540–560

A LEAST-SQUARES FINITE ELEMENT METHOD FOR THE LINEAR BOLTZMANN EQUATION WITH ANISOTROPIC SCATTERING*

TRAVIS M. AUSTIN† AND THOMAS A. MANTEUFFEL‡

Abstract. Least-squares methods have been applied to a wide range of differential equations and have been established to be competitive with other existing discretization strategies [P. B. Bochev and M. D. Gunzburger, SIAM Rev., 40 (1998), pp. 789–837]. In this article, we consider a least-squares method for the linear Boltzmann equation with anisotropic scattering. A similar method has already been developed, and extensively examined, for the linear Boltzmann equation with isotropic scattering. The success of the least-squares method for isotropic scattering depends on scaling the linear Boltzmann equation so that minimization of the least-squares functional in a discrete space always yields accurate discrete solutions. A similar scaling of the linear Boltzmann equation is employed for anisotropic scattering. In the previous work for isotropic scattering, coercivity and continuity results were established for the scaled least-squares functional relative to a physically reasonable norm. In this paper, we extend the previous coercivity and continuity results so that they hold in this more general case of anisotropic scattering. Additionally, we extend the bounds for the discretization error for the thin regime and for the thick regime. For the thick regime, we establish optimal error estimates for the case of highly anisotropic scattering.

Key words. least-squares, neutron transport, anisotropic scattering, finite elements

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/040610519

1. Introduction.

In this paper, we examine a least-squares method that is used to obtain discrete solutions to the single-group, steady-state linear Boltzmann equation with anisotropic scattering. The least-squares method for isotropic scattering was carefully analyzed for slab geometry in [12] and for xyz-geometry in [13]. Here, in the context of xyz-geometry, we extend the generality of the least-squares method by allowing for anisotropic scattering, whereby a particle has a preferential direction of scatter after collision.

For isotropic scattering, where particles have no preferential direction of scatter, it was proved in [12, 13] that the least-squares method yields discrete solutions that exhibit the correct asymptotic behavior in the diffusion limit. In this limit, the leading-order asymptotic solution of the Boltzmann equation converges to the solution of a diffusion equation. In [13], ellipticity of the least-squares functional was proved and the existence of optimal error estimates for a P_N angular discretization and a finite element spatial discretization was established. In [14], the authors enhanced the least-squares approach by adding a boundary functional to the least-squares functional, thus, weakly imposing the boundary conditions.

In [5], anisotropic scattering in the scaled least-squares approach was first considered in the context of multigroup transport. A scaling operator for the least-squares

*Received by the editors June 24, 2004; accepted for publication (in revised form) November 17, 2005; published electronically March 15, 2006.

http://www.siam.org/journals/sinum/44-2/61051.html

†Bioengineering Institute, University of Auckland, Private Bag 92019, Auckland, New Zealand (t.austin@auckland.ac.nz). This work is also referred to as LAUR-04-2523. Parts of this work were sponsored by the National Science Foundation under grant number DMS-8704169 and the Department of Energy, Applied Math Program grant DE-FG03-94ER25217.

‡Dept. of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@colorado.edu).

approach with anisotropic scattering in the multigroup context was introduced. Additionally, convergence results for a multilevel solution algorithm for the multigroup version of anisotropic scattering were presented. However, ellipticity of the scaled least-squares functional or estimates of the least-squares error were not considered.

For the single group equation, we set a foundation for the scaled least-squares approach introduced in [5] by proving ellipticity of the least-squares functional. We employ a scaling operator that is a single-group form of the scaling operator used in [5]. The proof of ellipticity for the thick regime with large absorption is identical to the equivalent case for isotropic scattering. The remaining cases depend on new techniques that were not used in [14]. Moreover, the proofs for the thin regime and the thick regime with small absorption have as special cases the proofs for the isotropic scattering [14]. It is the opinion of the authors that the proofs presented here are simpler and clearer than the proofs of [14], albeit resulting in small coercivity bounds.

The ellipticity results imply that the least-squares variational problem is well posed in an appropriate norm with ellipticity constants that are independent of the problem parameters. This ellipticity allows us to use Céa's lemma in establishing error bounds. Thus, once we introduce the discretization scheme (drawn from [5, 13, 14]), we use Céa's lemma to illustrate optimal bounds on the discretization error in the context of anisotropic scattering for the thin and thick regimes. For the thin regime, the proof from [14] can be invoked. For the thick regime with mildly anisotropic scattering, we merely indicate that the results are of the same form as [14]. There will, however, be new results for the thick regime with highly anisotropic scattering. These results will depend on an asymptotic expansion from Larsen and Pomraning in [10].

Most of the research on numerical methods for the linear Boltzmann equation with anisotropic scattering has focused on devising a plan to speed up source iteration, which is the standard iterative solution method used to solve isotropic transport problems [11]. Research has not focused on tailoring the discretization schemes used for isotropic scattering problems to anisotropic scattering problems because, in general, the same discretization techniques may be used [2, 15, 16]. Here, we focus on the formulation and discretization using a least-squares approach. We will not address the issue of what is the appropriate method for solving the resulting system of equations. For now we refer the reader to [5]. Since the approach for anisotropic scattering first described in [5] has not been studied theoretically, we concentrate on placing the method on firm ground. To this end, we proceed in the following way.

In section 2, we present the necessary preliminaries. Previous results for isotropic scattering are described in section 3. The scattering operator is presented in section 4 along with ellipticity results. In section 5, we describe the spatial and angular discretization scheme and present error estimates. In the final section, we discuss future work and further extensions of the least-squares method.

2. Preliminaries. As discussed in [11], the single-group, steady-state linear Boltzmann equation with anisotropic scattering is given by

$$\begin{aligned} [\boldsymbol{\Omega}\cdot\nabla+\sigma_t\mathcal{I}-\sigma_s\mathcal{K}]\,\psi(\mathbf{x},\boldsymbol{\Omega}) &= q \quad \text{for } (\mathbf{x},\boldsymbol{\Omega})\in R\times S^2\,, \\ \psi(\mathbf{x},\boldsymbol{\Omega}) &= g \quad \text{for } \mathbf{x}\in\partial R \text{ with } \mathbf{n}\cdot\boldsymbol{\Omega}<0, \end{aligned} \tag{2.1}$$

where σ_t is the *total cross section*, σ_s is the *scattering cross section*, and ψ is the *angular flux* to be determined for all points $\mathbf{x}\in R\subset\Re^3$ and all possible travel directions $\boldsymbol{\Omega}=(\sin\theta\cos\varphi\,,\sin\theta\sin\varphi,\cos\theta)\in S^2$. Spatial domain R is assumed to be an open connected set with $diam(R)=1$ and to have a piecewise $C^{1,1}$ boundary denoted by $\Gamma:=\partial R$.

To define anisotropic scattering operator $\mathcal{K}$, we must recall the normalized spherical harmonics from [3], given by

$$Y_{\ell m}(\mathbf{\Omega}) := Y_{\ell m}(\theta,\varphi) = (-1)^m \sqrt{\frac{(2\ell+1)(\ell-m)!}{(\ell+m)!}}\, P_\ell^m(\cos\,\theta)\, e^{im\varphi}, \tag{2.2}$$

where $P_\ell^m(\cdot)$ corresponds to the (ℓm)th associated Legendre moment. Normalization

$$\mathrm{d}\mathbf{\Omega} = \frac{\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\varphi}{4\pi}$$

allows us to expand the scattering operator $\mathcal{K}$ as

$$(\mathcal{K}v)(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \sigma_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}')\; v(\mathbf{x},\mathbf{\Omega}')\; \mathrm{d}\mathbf{\Omega}', \tag{2.3}$$

where $\sigma_\ell \in [0,1]$ for all $l > 0$ (with $\sigma_0 \equiv 1$) and $Y_{\ell m}^*$ is the complex conjugate of $Y_{\ell m}$. This infinite sum is truncated, in practice, such that for some $N_S \geq 0$,

$$(\mathcal{K}v)(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{N_S} \sigma_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}')\; v(\mathbf{x},\mathbf{\Omega}')\; \mathrm{d}\mathbf{\Omega}'. \tag{2.4}$$

Note that N_S depends on the degree of anisotropy in the scattering, and that when $N_S = 0$ in (2.4), we have

$$(\mathcal{K}v)(\mathbf{x},\mathbf{\Omega}) = \int_{S^2} v(\mathbf{x},\mathbf{\Omega}')\; \mathrm{d}\mathbf{\Omega}', \tag{2.5}$$

resulting in the isotropic transport operator. For the remainder, we refer to the operator in (2.5) as $\mathcal{P}$. Note that for subset $\Xi \subset \mathbb{N} \equiv \{0,1,2,3,\dots\}$ we can define the more general operator

$$(\mathcal{P}_\Xi\, v)(\mathbf{x},\mathbf{\Omega}) := \sum_{\ell\in\Xi} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}')\; v(\mathbf{x},\mathbf{\Omega}')\; \mathrm{d}\mathbf{\Omega}'. \tag{2.6}$$

Next, to distinguish between the isotropic and anisotropic transport operators, we introduce the notation $\mathcal{L}_I$ and $\mathcal{L}_A$ such that, for $v : R \times S^2 \to \Re$,

$$\mathcal{L}_I\, v := \mathbf{\Omega}\cdot\nabla\, v + \sigma_t\,(\mathcal{I}-\mathcal{P})\, v + \sigma_a\,\mathcal{P}\, v \tag{2.7}$$

and

$$\mathcal{L}_A\, v := \mathbf{\Omega}\cdot\nabla\, v + \sigma_t\,(\mathcal{I}-\mathcal{K})\, v + \sigma_a\,\mathcal{K}\, v, \tag{2.8}$$

where $\sigma_a := \sigma_t - \sigma_s$ represents the *absorption cross section*. Also, at times, the scattering term of the anisotropic transport operator will be represented by

$$\mathcal{S} = \sigma_t(\mathcal{I}-\mathcal{K}) + \sigma_a\,\mathcal{K}, \tag{2.9}$$

or by

$$\mathcal{S}v(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \mu_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}')\; v(\mathbf{x},\mathbf{\Omega}')\; \mathrm{d}\mathbf{\Omega}', \tag{2.10}$$

where

$$\mu_\ell = \sigma_t(1-\sigma_\ell) + \sigma_a\sigma_\ell. \tag{2.11}$$

Next, let the standard L^2 inner product and norm be denoted by

$$\langle u, v\rangle := \int_{S^2}\int_R u\, v^*\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{\Omega}' \qquad \text{and} \qquad \|u\| := \sqrt{\langle u, u\rangle},$$

where v^* again is the complex conjugate of v. Denote by $\mathrm{L}^2(S^2 \times R)$ the set of functions that are L^2-integrable on $S^2 \times R$. Any function in $\mathrm{L}^2(S^2 \times R)$ has a unique expression in terms of the spherical harmonics since the spherical harmonics are an orthonormal basis for $\mathrm{L}^2(S^2)$. Specifically, every $v \in \mathrm{L}^2(S^2 \times R)$ has the expansion

$$v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}), \tag{2.12}$$

with moments $\phi_{\ell m}(\mathbf{x})$ given by

$$\phi_{\ell m}(\mathbf{x}) = \int_{S^2} Y^*_{\ell m}(\mathbf{\Omega}')v(\mathbf{x}, \mathbf{\Omega}')\, \mathrm{d}\mathbf{\Omega}'. \tag{2.13}$$

3. Previous results for isotropic scattering. In [13], a scaling operator of the form $\mathcal{R} := a(\mathcal{I} - \mathcal{P}) + b\mathcal{P}$ was defined, with a and b depending on σ_t and σ_a. The isotropic form of (2.1) was then restated as the minimization of the least-squares functional

$$G_0(\psi; q) := \left\| \mathcal{R}^{-1/2}(\mathcal{L}_I\, \psi - q) \right\|^2. \tag{3.1}$$

The main result presented in [13] showed coercivity and continuity of the bilinear form $\left\langle \mathcal{R}^{-1}\mathcal{L}_I\, u, \mathcal{L}_I\, v \right\rangle$ with respect to

$$\|v\|_V^2 := \left\langle \mathcal{R}^{-1}\mathbf{\Omega}\cdot\nabla v, \mathbf{\Omega}\cdot\nabla v\right\rangle + \left\langle \mathcal{R}v, v\right\rangle. \tag{3.2}$$

To be more precise, defining V as the space of functions bounded in the V-norm and V_0 as the subspace of V with homogeneous inflow boundary conditions, the authors established V-ellipticity, i.e., constants C_e and C_c, independent of σ_t and σ_a, such that

$$C_e\, \|v\|_V^2 \le \left\langle \mathcal{R}^{-1}\mathcal{L}_I v, \mathcal{L}_I v\right\rangle \le C_c\, \|v\|_V^2 \tag{3.3}$$

for any $v \in V_0$. In [14], they extended this work by adding a boundary functional to the V-norm and the least-squares functional, and again proved ellipticity.

To describe the work in [14], it is necessary to describe the boundary functional. For each $\mathbf{x} \in \Gamma$, define $\mathbf{n}(\mathbf{x})$ to be the outward unit normal, define

$$\Gamma_I(\mathbf{\Omega}) := \{\mathbf{x} \in \Gamma \ : \ \mathbf{n}\cdot\mathbf{\Omega} < 0\}, \tag{3.4}$$

and define $\Gamma_O(\mathbf{\Omega}) := \Gamma/\ \Gamma_I(\mathbf{\Omega})$ to be the set of inflow and outflow particle travel directions. By defining $D := R \times S^2$, we then denote the inflow and outflow boundary of D by

$$\partial D_I := \{(\mathbf{x}, \mathbf{\Omega}) \in D \ : \ \mathbf{x} \in \Gamma_I(\mathbf{\Omega})\} \tag{3.5}$$

and

$$\partial D_O := \{(\mathbf{x}, \mathbf{\Omega}) \in D \ : \ \mathbf{x} \in \Gamma_O(\mathbf{\Omega})\}. \tag{3.6}$$

Corresponding to the inflow and outflow boundary of D are

$$b_I(u,v) := \int_{\partial R} \int_{\mathbf{n}\cdot\mathbf{\Omega}<0} uv|\mathbf{n}\cdot\mathbf{\Omega}| \,\mathrm{d}\mathbf{\Omega}\,\mathrm{d}\sigma \tag{3.7}$$

and

$$b_O(u,v) := \int_{\partial R} \int_{\mathbf{n}\cdot\mathbf{\Omega}>0} uv|\mathbf{n}\cdot\mathbf{\Omega}| \,\mathrm{d}\mathbf{\Omega}\,\mathrm{d}\sigma. \tag{3.8}$$

Associated with $b_I(\cdot,\cdot)$ is the inflow norm

$$\|v\|^2_{B_I} := b_I(v,v) \tag{3.9}$$

and the corresponding Sobolev space

$$B_I := \overline{\left\{v \in C^\infty(\overline{\partial R_I}) : \|v\|^2_{B_I} < \infty\right\}}. \tag{3.10}$$

For $q \in \mathrm{L}^2$ and $g \in B_I$, the least-squares functional studied in [14] is given by

$$G_I(\psi; q, g) := G_0(\psi; q) \ + \ 2b_I(\psi - g, \psi - g). \tag{3.11}$$

The authors obtained ellipticity results for (3.11) with respect to

$$\|v\|^2_{V_1} := \|v\|^2_V + b_I(v,v) \tag{3.12}$$

and the space V_1 consisting of functions bounded in the V_1-norm. Since G_I offers a more robust approximation of boundary conditions than G_0, we work exclusively in this paper with a least-squares functional that is similar in form to (3.12).

4. New results for anisotropic scattering. The scaling operator for anisotropic scattering is given by

$$\mathcal{R} := \begin{cases} \mathcal{I} & \text{in Region I}, \\ \sigma_t(\mathcal{I} - \mathcal{K}) + \sigma_a\mathcal{K} & \text{in Region II}, \\ \sigma_t(\mathcal{I} - \mathcal{K}) + \frac{1}{\sigma_t}\mathcal{K} & \text{in Region III}, \end{cases} \tag{4.1}$$

where Regions I, II, and III are defined in Figure 4.1. Note that $\mathcal{R}$ is a continuous function in σ_t and σ_a for fixed $\mathcal{K}$ and can be alternatively expressed as

$$\mathcal{R}v := \sum_{\ell=0}^{\infty} \nu_\ell \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x})\, Y_{\ell m}(\mathbf{\Omega}), \tag{4.2}$$

using $\phi_{\ell m}$ in (2.13) and

$$\nu_\ell := \begin{cases} 1 & \text{for Region I}, \\ \sigma_t\,(1 - \sigma_\ell) + \sigma_a\,\sigma_\ell & \text{for Region II}, \\ \sigma_t\,(1 - \sigma_\ell) + \frac{1}{\sigma_t}\,\sigma_\ell & \text{for Region III}. \end{cases} \tag{4.3}$$

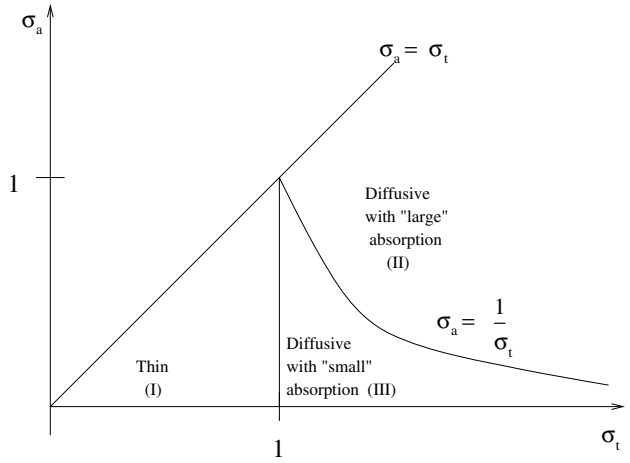


FIG. 4.1. *Division of parameters into Regions* I, II, *and* III.

In the remainder of this section, we develop V_1-ellipticity proofs for the case of anisotropic scattering using the scaling operator (4.1). Firstly, though, we note that the V_1-norm, defined by (3.12), does not change for anisotropic scattering once $\mathcal{R}$ is defined in terms of (4.1). Secondly, for $q \in \mathrm{L}^2$ and $g \in B_I$, we note that the solution of (2.1) can be expressed as

$$\psi = \arg\min_{v \in V_1} G_A(v;\, q, g), \tag{4.4}$$

where $G_A(v; q, g)$ is the anisotropic equivalent of $G_I(v; q, g)$. The corresponding variational form is: find $\psi \in V_1$ such that

$$a(\psi, v) := \left\langle \mathcal{R}^{-1}\mathcal{L}_A\,\psi, \mathcal{L}_A\,v \right\rangle + 2b_I(\psi, v) = \left\langle \mathcal{R}^{-1}q, \mathcal{L}_A\,v \right\rangle + 2b_I(g, v) \tag{4.5}$$

for every $v \in V_1$. Once we establish V_1-ellipticity results for $a(\psi, v)$, we will have established that (4.5) is well posed. This well posedness will imply that, for each pair $q \in \mathrm{L}^2$ and $g \in B_I$, there exists a unique $\psi \in V_1$ satisfying (2.1). Moreover, a standard stability result (cf. [4]) implies that we get the a priori estimate:

$$\|\psi\|_{V_1} \le C_e^{-1}\left(\left\|\mathcal{R}^{-1/2}q\right\| + b_I(g, g)^{1/2}\right), \tag{4.6}$$

where C_e is the coercivity bound.

4.1. Auxiliary lemmas. In this section, we present two lemmas. Most of the first lemma is a restatement of Lemma 3.1 from [14]. The second lemma is used in the thin and thick regime ellipticity proofs.

First, we define an operator that arises in Lemma 4.1, and the ellipticity proof for the thick regime with small absorption. To define the operator, we define $s = (\sigma_t - \frac{1}{\sigma_t})/(\sigma_t - \sigma_a)$ and split the moments into the two disjoint sets,

$$\Upsilon = \{\, \ell \in \mathbb{N} \;:\; \sigma_l > s\} = \{\ell \in \mathbb{N} \;:\; \mu_\ell < 1/\sigma_t\}, \tag{4.7}$$

$$\widehat{\Upsilon} = \{\, \ell \in \mathbb{N} \;:\; \sigma_l \le s\} = \{\ell \in \mathbb{N} \;:\; \mu_\ell \ge 1/\sigma_t\}. \tag{4.8}$$

We can then define the projection operator

$$\mathcal{P}_\Upsilon v(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell \in \Upsilon} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

according to (2.6). A similar operator can be defined for $\widehat{\Upsilon}$. Furthermore, we introduce the operator $\mathcal{D}$ given by

$$\mathcal{D}\, v(\mathbf{x},\mathbf{\Omega}) := \mathcal{P}_\Upsilon v(\mathbf{x},\mathbf{\Omega}) + \sum_{\ell\in\widehat{\Upsilon}} \zeta_l \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}), \tag{4.9}$$

where $\zeta_\ell := (1-\sigma_\ell) + \frac{\sigma_a \sigma_\ell}{\sigma_t}$. Notice that $\mathcal{D}^s$ is a meaningful operator for any $s \in \Re$ and that $\|\mathcal{D}\| \leq 1$ since $\zeta_\ell \leq 1$. In the following, we use the notation, $\sum_\ell = \sum_\ell^\infty$.

LEMMA 4.1. *For $v \in V_1$, we have*

(i) $2\,\langle \mathbf{\Omega}\cdot\nabla v, v\rangle = b_O(v,v) - b_I(v,v) \geq -b_I(v,v)$;

(ii) *the Poincaré-Friedrichs inequality*

$$\|v\|^2 \leq 2\,\mathrm{diam}(R)^2\ \|\mathbf{\Omega}\cdot\nabla v\|^2 + 2\,\mathrm{diam}(R)\, b_I(v,v)\,; \tag{4.10}$$

(iii) *for* $\mathrm{diam}(R) = 1$ *and* $\sigma_t \geq 1$, *we have*

$$\|\mathcal{P}_\Upsilon\, v\|^2 \leq 2\left\|\mathcal{Q}^{-1/2}\,\mathbf{\Omega}\cdot\nabla v + \sigma_t\,\mathcal{P}_\Upsilon\,\mathbf{\Omega}\cdot\nabla v\right\|^2 + 2b_I(v,v)\,, \tag{4.11}$$

where $\mathcal{Q} := \mathcal{D}\,(\mathcal{I}-\mathcal{P}_\Upsilon)$ *and* $\mathcal{Q}^s := \mathcal{D}^s\,(\mathcal{I}-\mathcal{P}_\Upsilon)$ *for* $s \in \Re$.

Proof. The proofs of (i) and (ii) are found in Lemma 3.1 of [14], while (iii) is proved by assuming (ii), and noting that $\|\mathcal{P}_\Upsilon\, v\|^2 \leq \|v\|^2$ and

$$\|\mathbf{\Omega}\cdot\nabla\, v\|^2 \leq \left\|\mathcal{D}^{-1/2}\,(\mathcal{I}-\mathcal{P}_\Upsilon)\,\mathbf{\Omega}\cdot\nabla v + \sigma_t\,\mathcal{P}_\Upsilon\,\mathbf{\Omega}\cdot\nabla v\right\|^2. \qquad \square$$

LEMMA 4.2. *Given $\lambda_\ell > 0$ and $\omega_\ell > 0$, the minimum of*

$$I(d) := \sum_\ell (\lambda_\ell - d)^2 \omega_\ell \tag{4.12}$$

for $d \in [0,1]$ is achieved at

$$d_m = \frac{\sum_\ell \lambda_\ell \omega_\ell}{\sum_\ell \omega_\ell},$$

and furthermore,

$$I(d_m) = \sum_\ell \lambda_\ell^2 \omega_\ell - \frac{(\sum_\ell \lambda_\ell \omega_\ell)^2}{\sum_\ell \omega_\ell}.$$

Proof. The result is established by differentiating $I(d)$ with respect to d. $\square$

The thin regime and thick regime with small absorption ellipticity proofs that follow make use of the projection operator

$$\mathcal{P}_\ell v := \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}), \tag{4.13}$$

implying that (2.10) can be expressed as

$$\mathcal{S}v(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \mu_\ell\, \mathcal{P}_\ell\, v(\mathbf{x},\mathbf{\Omega}). \tag{4.14}$$

Note the observation, $\sum_\ell \mathcal{P}_\ell = \mathcal{I}$, that we need in the following.

4.2. Thin regime ($0 \le \sigma_a \le \sigma_t \le 1$). For the thin regime, we have $\mathcal{R} = \mathcal{I}$ and

$$\|u\|_{V_1}^2 = \|\mathbf{\Omega} \cdot \nabla v\|^2 + \|v\|^2 + b_I(v,v).$$

THEOREM 4.3 (continuity and V_1-ellipticity for thin regime). *Assume that* $0 \le \sigma_a \le \sigma_t \le 1$. *Then, for all* $u, v \in V_1$, *we have*

$$\begin{aligned} |a(u,v)| = |\langle \mathcal{L}_A\, u, \mathcal{L}_A\, v\rangle + 2b_I(u,v)| &\le C_c\, \|u\|_{V_1}\, \|v\|_{V_1}, \\ a(v,v) = \langle \mathcal{L}_A\, v, \mathcal{L}_A\, v\rangle + 2b_I(v,v) &\ge C_e\, \|v\|_{V_1}^2, \end{aligned} \tag{4.15}$$

with $C_c \le 2$ *and* $C_e \ge 0.06574145$.

Proof. Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} |a(u,v)| &\le \|\mathcal{L}_A\, u\|\, \|\mathcal{L}_A\, v\| + 2b_I(u,u)^{\frac{1}{2}} b_I(v,v)^{\frac{1}{2}} \\ &\le \left(\|\mathcal{L}_A\, u\|^2 + 2b_I(u,u)\right)^{\frac{1}{2}} \left(\|\mathcal{L}_A\, v\|^2 + 2b_I(v,v)\right)^{\frac{1}{2}}. \end{aligned}$$

Given that $\sigma_t(1-\sigma_\ell) \le 1$ and $\sigma_a \sigma_\ell \le 1$, we have

$$\|\mathcal{L}_A\, u\|^2 \le 2\left(\|\mathbf{\Omega} \cdot \nabla u\|^2 + \|(\mathcal{I}-\mathcal{P})u\|^2 + \|\mathcal{P}u\|^2\right) = 2\left(\|\mathbf{\Omega} \cdot \nabla u\|^2 + \|u\|^2\right)$$

so that $\|\mathcal{L}_A\, u\|^2 + 2b_I(u,u) \le 2\, \|u\|_{V_1}^2$. This proves continuity of $a(\cdot,\cdot)$.

To prove ellipticity, we note that $0 \le \mu_\ell \le 1$ and refer the reader to (4.14). Using this definition of $\mathcal{S}$, we have

$$\begin{aligned} a(v,v) &= \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v\rangle + 2\langle \Omega \cdot \nabla v, \mathcal{S}v\rangle + \langle \mathcal{S}v, \mathcal{S}v\rangle + 2\, b_I(v,v) \\ &= \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v\rangle + 2\sum_\ell \mu_\ell \langle \Omega \cdot \nabla v, \mathcal{P}_\ell v\rangle + \sum_\ell \mu_\ell^2 \langle \mathcal{P}_\ell v, v\rangle + 2\, b_I(v,v). \end{aligned} \tag{4.16}$$

For any $d \in [0,1]$, adding the identity

$$2d\left[\langle \Omega \cdot \nabla v, v\rangle - \sum_\ell \langle \Omega \cdot \nabla v, \mathcal{P}_\ell v\rangle\right] = 0$$

to (4.16) and using Lemma 4.1(i) yields

$$\begin{aligned} a(v,v) \ge \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v\rangle &+ 2\sum_\ell (\mu_\ell - d)\langle \Omega \cdot \nabla v, \mathcal{P}_\ell v\rangle \\ &+ \sum_\ell \mu_\ell^2 \langle \mathcal{P}_\ell v, v\rangle + (2-d)\, b_I(v,v). \end{aligned} \tag{4.17}$$

For convenience, we put $A_\ell = \|\mathcal{P}_\ell \mathbf{\Omega} \cdot \nabla v\|^2$, $A = \|\mathbf{\Omega} \cdot \nabla v\|^2$, $B_\ell = \|\mathcal{P}_\ell v\|^2$, and $B = \|v\|^2$. We also define $\gamma_\ell = A_\ell/A$ and $\delta_\ell = B_\ell/B$ and note that

$$\sum_\ell \delta_\ell = \sum_\ell \gamma_\ell = 1.$$

Thus we seek a proof of

$$a(v,v) \ge C\,(A + B + b_I(v,v))$$

for all $v \in V_1$. Applying the arithmetic-geometric inequality to the cross product term in (4.17) with $\eta > 0$ yields

$$\begin{aligned} a(v,v) &\geq \left(A - \eta \sum_\ell A_\ell \right) + \left(\sum_\ell \mu_\ell^2 B_\ell - \frac{1}{\eta} \sum_\ell (\mu_\ell - d)^2 B_\ell \right) + (2-d)\, b_I(v,v) \\ &= (1-\eta)\, A + \left(\sum_\ell \mu_\ell^2 \delta_\ell - \frac{1}{\eta} \sum_\ell (\mu_\ell - d)^2 \delta_\ell \right) B + (2-d)\, b_I(v,v) \end{aligned}$$

since $B_\ell = \delta_\ell B$. Making use of Lemma 4.2 we choose $d = \sum_\ell \mu_\ell \delta_\ell \leq \sum_\ell \delta_\ell = 1$, implying that

$$\begin{aligned} a(v,v) &\geq [1-\eta]\ A + \left[\sum_\ell \mu_\ell^2 \delta_\ell - \frac{1}{\eta} \left(\sum_\ell \mu_\ell^2 \delta_\ell - \left(\sum_\ell \mu_\ell \delta_\ell \right)^2 \right) \right] B + b_I(v,v) \\ &= [1-\eta]\, A + \left[\left(1 - \frac{1}{\eta} \right) \sum_\ell \mu_\ell^2 \delta_\ell + \frac{1}{\eta} \left(\sum_\ell \mu_\ell \delta_\ell \right)^2 \right] B + b_I(v,v). \end{aligned}$$

For convenience, define $\delta = \sum_\ell \mu_\ell^2 \delta_\ell \leq \sum_\ell \mu_\ell \delta_\ell \leq 1$ such that

$$\begin{aligned} a(v,v) &\geq [1-\eta]\, A + \left[\left(1 - \frac{1}{\eta} \right) \delta + \frac{1}{\eta} \delta^2 \right] B + b_I(v,v) \\ &= [1-\eta]\, A + \left[\delta - \frac{1}{\eta} \delta (1-\delta) \right] B + b_I(v,v). \end{aligned}$$

Using Lemma 4.1(ii) (assuming $\operatorname{diam}(R) = 1$) we get

$$a(v,v) \geq [1-\eta-\beta]\, A + \left[\delta - \frac{1}{\eta} \delta (1-\delta) + \frac{\beta}{2} \right] B + [1-\beta]\, b_I(v,v)$$

for any $\beta \geq 0$. Define

$$\begin{aligned} C_1 &= 1 - \eta - \beta, \\ C_2 &= \delta - \frac{1}{\eta} \delta (1-\delta) + \frac{\beta}{2}, \\ C_3 &= 1 - \beta. \end{aligned}$$

Next, set $\eta = \sqrt{\delta(1-\delta)}$ and choose β to make $C_1 = C_2$. This requires

$$1 - \sqrt{\delta(1-\delta)} - \beta = \delta - \sqrt{\delta(1-\delta)} + \frac{\beta}{2},$$

which implies

$$\beta = \frac{2}{3}(1-\delta) \geq 0.$$

Plugging back into C_1 yields

$$\begin{aligned} C_1 = C_2 &= 1 - \sqrt{\delta(1-\delta)} - \frac{2}{3}(1-\delta), \\ C_3 &= 1 - \frac{2}{3}(1-\delta) \geq \frac{1}{3}. \end{aligned}$$

Numerically we find that the minimum value of C_1 occurs at $\delta = \frac{1-\sqrt{16/52}}{2} \approx 0.2226499$ and yields $C_1 = C_2 \approx 0.06574145$. Comparing to the bound on C_3, we see that $C_e \geq 0.06574145$. ∎

4.3. Thick regime with "large" absorption ($1 \leq \boldsymbol{\sigma_t} < \infty$; $\frac{1}{\boldsymbol{\sigma_t}} \leq \boldsymbol{\sigma_a} \leq \boldsymbol{\sigma_t}$). For the thick regime with "large" absorption, the scaling is given by

$$\mathcal{R} = \sigma_t \left(\mathcal{I} - \mathcal{K}\right) + \sigma_a \mathcal{K}, \tag{4.18}$$

which implies that

$$\|v\|_V^2 := \left\langle \mathcal{R}^{-1} \boldsymbol{\Omega} \cdot \nabla v, \boldsymbol{\Omega} \cdot \nabla v \right\rangle + \langle \mathcal{R}\, v, v \rangle$$

and

$$\|v\|_{V_1}^2 = \|v\|_V^2 + b_I(v, v).$$

THEOREM 4.4 (continuity and V_1-ellipticity for the thick regime with "large" absorption). *Assume that* $1 \leq \sigma_t < \infty$ *and* $\frac{1}{\sigma_t} \leq \sigma_a \leq \sigma_t$. *Then, for all* $u, v \in V_1$, *we have*

$$\begin{aligned} |a(u,v)| &\leq 2 \, \|u\|_{V_1} \, \|v\|_{V_1}, \\ a(v,v) &\geq \|v\|_{V_1}^2 . \end{aligned} \tag{4.19}$$

Proof. See the proof establishing coercivity and continuity in thick regime with "large" absorption from [14]. Proof for coercivity is the same because, as in [14], scaling operator $\mathcal{R}$, in (4.18) is equal to scattering operator $\mathcal{S}$. ∎

4.4. Thick regime with "small" absorption ($1 \leq \boldsymbol{\sigma_t} \leq \infty$; $\boldsymbol{\sigma_a} \leq \frac{1}{\boldsymbol{\sigma_t}}$). For the thick regime with "small" absorption, we must define the scaling operator so that it does not become singular as $\sigma_a \to 0$. Hence, the scaling operator defined in (4.18) is recast as

$$\mathcal{R} = \sigma_t(\mathcal{I} - \mathcal{K}) + \frac{1}{\sigma_t}\mathcal{K}, \tag{4.20}$$

while the V-norm and the V_1-norm have the same dependence on $\mathcal{R}$.

One of the key ingredients of the following proof is the intermediate scaling operator defined below. Recall Υ and $\widehat{\Upsilon}$ defined by (4.7) and (4.8), and $\mathcal{Q}$ defined by (4.11). The intermediate scaling operator is

$$\mathcal{T} := \sigma_t \mathcal{Q} + \frac{1}{\sigma_t}\mathcal{P}_\Upsilon = \mathcal{S}(\mathcal{I} - \mathcal{P}_\Upsilon) + \frac{1}{\sigma_t}\mathcal{P}_\Upsilon.$$

Note that $\mathcal{Q} \leq \mathcal{I}$, and additionally, that $\sigma_a \mathcal{P}_\Upsilon \leq \mathcal{S}\mathcal{P}_\Upsilon \leq \frac{1}{\sigma_t}\mathcal{P}_\Upsilon$, which is equivalent to $\sigma_a \leq \mu_\ell \leq \frac{1}{\sigma_t}$ for $\ell \in \Upsilon$. The first inequality is true by definition of μ_ℓ, and the second is true by definition of Υ. We also introduce τ_ℓ according to

$$\tau_\ell := \begin{cases} \frac{1}{\sigma_t} & \text{for } \ell \in \Upsilon \\ \mu_\ell & \text{for } \ell \in \widehat{\Upsilon} \end{cases} \tag{4.21}$$

such that, from (4.13),

$$\mathcal{T} v(\mathbf{x}, \boldsymbol{\Omega}) = \sum_{\ell=0}^{\infty} \tau_\ell \mathcal{P}_\ell v(\mathbf{x}, \boldsymbol{\Omega}).$$

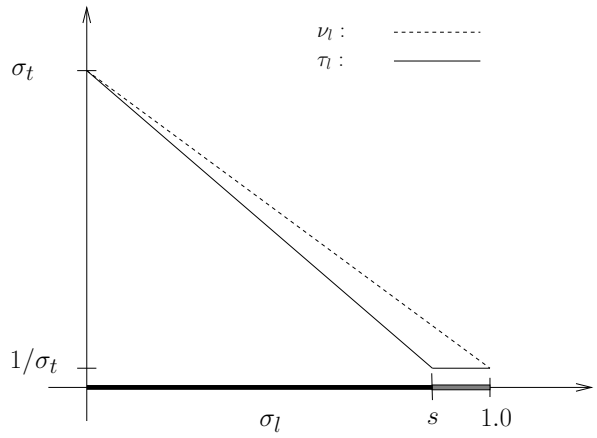


FIG. 4.2. *Graph of τ_ℓ and ν_ℓ, where the black region on $\sigma_\ell \in [0,1]$ denotes $\widehat{\Upsilon}$ and the grey region denotes Υ. From the graph we can see that the largest ratio occurs at the interface between $\widehat{\Upsilon}$ and Υ, which is where $\sigma_\ell = s$.*

For future reference, we note that

$$\mathcal{T}^{-1}\mathcal{S} = \mathcal{S}\mathcal{T}^{-1} = (\mathcal{I} - \mathcal{P}_\Upsilon) + \sigma_t \mathcal{S}\mathcal{P}_\Upsilon, \tag{4.22}$$

$$\mathcal{T}^{-1}\mathcal{S}^2 = \mathcal{S}^2\mathcal{T}^{-1} = \sigma_t \mathcal{Q} + \sigma_t \mathcal{S}^2\mathcal{P}_\Upsilon, \tag{4.23}$$

and $\mathcal{S}$ can be expressed as

$$\mathcal{S} = \mathcal{S}(\mathcal{I} - \mathcal{P}_\Upsilon) + \mathcal{S}\mathcal{P}_\Upsilon = \sigma_t \mathcal{Q} + \mathcal{S}\mathcal{P}_\Upsilon.$$

Note the inequality

$$\langle \mathcal{S}v, v\rangle \le \langle \mathcal{T}v, v\rangle \le \langle \mathcal{R}v, v\rangle \tag{4.24}$$

and the related inequality

$$\left\langle \mathcal{T}^{-1}v, v\right\rangle \ge \left\langle \mathcal{R}^{-1}v, v\right\rangle. \tag{4.25}$$

We also need the following lemma, which further relates $\mathcal{T}$ to $\mathcal{R}$.

LEMMA 4.5. *For $\sigma_t \ge 1$ and $\sigma_a \le 1/\sigma_t$, we have*

$$2\langle \mathcal{T}v, v\rangle \ge \langle \mathcal{R}v, v\rangle \qquad \text{and} \qquad \left\langle \mathcal{T}^{-1}v, v\right\rangle \le 2\left\langle \mathcal{R}^{-1}v, v\right\rangle.$$

Proof. To prove both results, we establish a bound relating τ_ℓ to ν_ℓ, where ν_ℓ is defined by (4.3). By observing the graph of τ_ℓ and ν_ℓ in Figure 4.2, we see that the ratio ν_ℓ/τ_ℓ is maximized at $s = (\sigma_t - \frac{1}{\sigma_t})/(\sigma_t - \sigma_a)$. Evaluating both ν_ℓ and τ_ℓ at s yields

$$\frac{\nu_\ell}{\tau_\ell} = \frac{\sigma_t(1-s) + s/\sigma_t}{1/\sigma_t} = \sigma_t^2(1-s) + s = \frac{2\sigma_t - \sigma_a\sigma_t^2 - 1/\sigma_t}{\sigma_t - \sigma_a}.$$

As $\sigma_a \to 0$, the value of s decreases to its minimum (dependent on σ_a) of $1 - 1/\sigma_t^2$. Since $\nu_\ell/\tau_\ell = \sigma_t^2(1-s) + s$, the maximum of the ratio is where s is at its minimum, i.e., where $\sigma_a = 0$. This implies

$$\frac{\nu_\ell}{\tau_\ell} \le 2 - 1/\sigma_t^2 \le 2.$$

Both results follow from this inequality. ∎

THEOREM 4.6 (continuity and V_1-ellipticity for thick regime with "small" absorption). *Assume that* $1 \leq \sigma_t < \infty$, $0 \leq \sigma_a \leq \frac{1}{\sigma_t}$. *Then, for all* $u, v \in V_1$, *we have*

$$|a(u,v)| = |\left\langle \mathcal{R}^{-1}\mathcal{L}_A\, u, \mathcal{L}_A\, v\right\rangle + 2b_I(u,v)| \leq C_c\ \|u\|_{V_1}\ \|v\|_{V_1}\,, \tag{4.26}$$

$$a(v,v) = \left\langle \mathcal{R}^{-1}\mathcal{L}_A\, v, \mathcal{L}_A\, v\right\rangle + 2b_I(v,v) \geq C_e\ \|v\|^2_{V_1}\,, \tag{4.27}$$

with $C_c \leq 2$ *and* $C_e \geq 0.01667$, *independent of* σ_t *and* σ_a.

Proof. The proof for continuity follows from the same reasoning as used in Theorem 4.4 since

$$|a(u,v)| \leq \left(\left\|\mathcal{R}^{-\frac{1}{2}}\mathcal{L}_A\, u\right\| + 2b_I(u,u)\right)^{\frac{1}{2}}\ \left(\left\|\mathcal{R}^{-\frac{1}{2}}\mathcal{L}_A\, v\right\| + 2b_I(v,v)\right)^{\frac{1}{2}}.$$

The observation that $\mu_\ell \leq \nu_\ell$ implies

$$\left\|\mathcal{R}^{-\frac{1}{2}}\mathcal{S}u\right\| \leq \left\|\mathcal{R}^{\frac{1}{2}}u\right\|.$$

One can easily show then that $\left\|\mathcal{R}^{-\frac{1}{2}}\mathcal{L}_A\, u\right\|^2 \leq 2\,\|u\|^2_V$.

To establish coercivity, we proceed as follows. We first define

$$\widetilde{a}(v,v) := \left\langle \mathcal{T}^{-1}\mathcal{L}_A v, \mathcal{L}_A v\right\rangle + 2\, b_I(v,v), \tag{4.28}$$

and through Lemma 4.5, we get

$$a(v,v) \geq \frac{1}{2}\,\widetilde{a}(v,v). \tag{4.29}$$

Assume for now that we have

$$\widetilde{a}(v,v) \geq \widetilde{C}_e\,\left(\left\|\mathcal{T}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|^2 + \left\|\mathcal{T}^{1/2}v\right\|^2 + b_I(v,v)\right). \tag{4.30}$$

Using inequality (4.25) and Lemma 4.5, we can bound the first two terms on the right-hand side of (4.30) from below to get

$$\widetilde{a}(v,v) \geq \frac{\widetilde{C}_e}{2}\,\left(\left\|\mathcal{R}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|^2 + \left\|\mathcal{R}^{1/2}v\right\|^2 + b_I(v,v)\right). \tag{4.31}$$

With (4.29) we get as an ellipticity constant in (4.27) the value of $\widetilde{C}_e/4$. Thus, we are only left to prove (4.30), and determine $\widetilde{C}_e$.

Noting (4.22) and (4.23) we write

$$\begin{aligned}
\widetilde{a}(v,v) &= \left\langle \mathcal{T}^{-1}\boldsymbol{\Omega}\cdot\nabla v, \boldsymbol{\Omega}\cdot\nabla v\right\rangle + \left\langle \mathcal{S}^2\mathcal{T}^{-1}v, v\right\rangle + 2\left\langle \mathcal{S}\mathcal{T}^{-1}\boldsymbol{\Omega}\cdot\nabla v, v\right\rangle + 2b_I(v,v)\\
&= \frac{1}{\sigma_t}\left\langle \mathcal{Q}^{-1}\boldsymbol{\Omega}\cdot\nabla v, \boldsymbol{\Omega}\cdot\nabla v\right\rangle + \sigma_t\left\langle \mathcal{P}_\Upsilon\boldsymbol{\Omega}\cdot\nabla v, \boldsymbol{\Omega}\cdot\nabla v\right\rangle + \sigma_t\left\langle \mathcal{Q}v, v\right\rangle\\
&\quad + \sigma_t\sum_{\ell\in\Upsilon}\mu_\ell^2\left\langle \mathcal{P}_\ell v, v\right\rangle + 2\left\langle \boldsymbol{\Omega}\cdot\nabla v, (\mathcal{I}-\mathcal{P}_\Upsilon)v\right\rangle + 2\sigma_t\sum_{\ell\in\Upsilon}\mu_\ell\left\langle \boldsymbol{\Omega}\cdot\nabla v, \mathcal{P}_\ell v\right\rangle\\
&\quad + 2b_I(v,v).
\end{aligned}$$

For any $d \in [0,1]$, adding the identity

$$2d\,\left[\langle\boldsymbol{\Omega}\cdot\nabla v, v\rangle - \langle\boldsymbol{\Omega}\cdot\nabla v, (\mathcal{I}-\mathcal{P}_\Upsilon)v\rangle - \langle\boldsymbol{\Omega}\cdot\nabla v, \mathcal{P}_\Upsilon v\rangle\right] = 0$$

to the last line above, and using the inequality from Lemma 4.1(ii), yields

$$\begin{aligned}\widetilde{a}(v,v) \geq \frac{1}{\sigma_t}\left\langle \mathcal{Q}^{-1}\boldsymbol{\Omega}\cdot\nabla v, \boldsymbol{\Omega}\cdot\nabla v\right\rangle + \sigma_t\left\langle \mathcal{P}_\Upsilon\boldsymbol{\Omega}\cdot\nabla v, \boldsymbol{\Omega}\cdot\nabla v\right\rangle + \sigma_t\left\langle \mathcal{Q}v, v\right\rangle \\ + \sigma_t\sum_{\ell\in\Upsilon}\mu_\ell^2\left\langle \mathcal{P}_\ell v, v\right\rangle - 2|1-d|\ |\langle \boldsymbol{\Omega}\cdot\nabla v, (\mathcal{I}-\mathcal{P}_\Upsilon)v\rangle| \\ -2\sum_{\ell\in\Upsilon}|\sigma_t\mu_\ell - d|\ |\langle \boldsymbol{\Omega}\cdot\nabla v, \mathcal{P}_\ell v\rangle| + (2-d)b_I(v,v).\end{aligned}$$

It is convenient to note that we may write

$$|\left\langle \boldsymbol{\Omega}\cdot\nabla v, (\mathcal{I}-\mathcal{P}_\Upsilon)v\right\rangle| \leq \left\|\frac{1}{\sqrt{\sigma_t}}Q^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|\ \left\|\sqrt{\sigma_t}Q^{1/2}v\right\|,$$

and for $\ell\in\Upsilon$,

$$|\left\langle \boldsymbol{\Omega}\cdot\nabla v, \mathcal{P}_\ell\, v\right\rangle| \leq \|\sqrt{\sigma_t}\,\mathcal{P}_\ell\,\boldsymbol{\Omega}\cdot\nabla v\|\ \left\|\frac{1}{\sqrt{\sigma_t}}\,\mathcal{P}_\ell\, v\right\|.$$

Let's now define

$$\delta_\ell := \frac{\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\ell v\right\|^2}{\left\|\sqrt{\sigma_t}\mathcal{Q}^{1/2}v + \frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\Upsilon v\right\|^2} \quad\text{and}\quad \gamma_\ell := \frac{\left\|\sqrt{\sigma_t}\mathcal{P}_\ell\boldsymbol{\Omega}\cdot\nabla v\right\|^2}{\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{Q}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v + \sqrt{\sigma_t}\mathcal{P}_\Upsilon\boldsymbol{\Omega}\cdot\nabla v\right\|^2}$$

for $\ell\in\Upsilon$. Additionally, let $\delta_0 = 1-\sum_{\ell\in\Upsilon}\delta_\ell$ and $\gamma_0 = 1-\sum_{\ell\in\Upsilon}\gamma_\ell$. For convenience, set

$$A = \left\|\mathcal{T}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|^2 = \left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{Q}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v + \sqrt{\sigma_t}\mathcal{P}_\Upsilon\boldsymbol{\Omega}\cdot\nabla v\right\|^2$$

and

$$B = \left\|\mathcal{T}^{1/2}v\right\|^2 = \left\|\sqrt{\sigma_t}\mathcal{Q}^{1/2}v + \frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\Upsilon v\right\|^2.$$

We can now use the arithmetic-geometric inequality to write, for any $\eta>0$,

$$\begin{aligned}\widetilde{a}(v,v) \geq \left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{Q}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|^2 + \sum_{\ell\in\Upsilon}\|\sqrt{\sigma_t}\mathcal{P}_\ell\boldsymbol{\Omega}\cdot\nabla v\|^2 + \left\|\sqrt{\sigma_t}\mathcal{Q}^{1/2}v\right\|^2 \\ + \sum_{\ell\in\Upsilon}(\sigma_t\mu_\ell)^2\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\ell v\right\|^2 - \eta\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{Q}^{-1/2}\boldsymbol{\Omega}\cdot\nabla v\right\|^2 \\ - \frac{(1-d)^2}{\eta}\left\|\sqrt{\sigma_t}\mathcal{Q}^{1/2}v\right\|^2 - \sum_{\ell\in\Upsilon}\eta\,\|\sqrt{\sigma_t}\mathcal{P}_\ell\boldsymbol{\Omega}\cdot\nabla v\|^2 \\ - \sum_{\ell\in\Upsilon}\frac{(\sigma_t\mu_\ell-d)^2}{\eta}\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\ell v\right\|^2 + (2-d)b_I(v,v)\end{aligned}$$

$$= \left[(1-\eta)\gamma_0 + \sum_{\ell\in\Upsilon}(1-\eta)\gamma_\ell\right]A + \left[\left(1-\frac{(1-d)^2}{\eta}\right)\delta_0 \tag{4.32}$$

$$+ \sum_{\ell\in\Upsilon}\left((\sigma_t\mu_\ell)^2 - \frac{(\sigma_t\mu_\ell-d)^2}{\eta}\right)\delta_\ell\right]B + (2-d)b_I(v,v). \tag{4.33}$$

We choose d to make the coefficient of B as large as possible. First we simplify this expression by temporarily setting $\mu_0 = 1/\sigma_t$ so that the coefficient on B can be expressed as

$$\sum_{\ell\in 0\cup\Upsilon} (\sigma_t\mu_\ell)^2\delta_\ell - \frac{1}{\eta}\sum_{\ell\in 0\cup\Upsilon} (\sigma_t\mu_\ell - d)^2\delta_\ell. \tag{4.34}$$

Then we see that d should be chosen to minimize the second sum, which by Lemma 4.2, yields

$$d = \sum_{\ell\in 0\cup\Upsilon} (\sigma_t\mu_\ell)\,\delta_\ell, \tag{4.35}$$

making use of the fact that $\sum_{\ell\in 0\cup\Upsilon}\delta_\ell = 1$. Note that $d\in[0,1]$ because $(\sigma_t\mu_\ell)\in[0,1]$.

Substituting d defined by (4.35) into (4.34) yields

$$\delta_0 + \sum_{\ell\in\Upsilon} (\sigma_t\mu_\ell)^2\delta_\ell - \frac{1}{\eta}\left(\left(\delta_0 + \sum_{\ell\in\Upsilon}(\sigma_t\mu_\ell)^2\,\delta_\ell\right) - \left(\delta_0 + \sum_{\ell\in\Upsilon}(\sigma_t\mu_\ell)\,\delta_\ell\right)^2\right). \tag{4.36}$$

For convenience, we rewrite (4.36) as

$$\Delta_0 - \frac{1}{\eta}\left(\Delta_0 - \Delta_1^2\right) = \Delta_0\left(1-\frac{1}{\eta}\right) + \frac{\Delta_1^2}{\eta},$$

where

$$\Delta_0 = \delta_0 + \sum_{\ell\in\Upsilon}(\sigma_t\mu_\ell)^2\delta_\ell \qquad \text{and} \qquad \Delta_1 = \delta_0 + \sum_{\ell\in\Upsilon}(\sigma_t\mu_\ell)\delta_\ell.$$

Note that Lemma 4.2 implies that $\Delta_0 \geq \Delta_1^2$. Also, $(\sigma_t\mu_\ell)^2 \leq (\sigma_t\mu_\ell) \leq 1$ for $\ell\in\Upsilon$ implies that $\Delta_0 \leq \Delta_1 \leq 1$. Hence, we get

$$\Delta_0 - \frac{1}{\eta}\left(\Delta_0 - \Delta_1^2\right) = \Delta_0\left(1-\frac{1}{\eta}\right) + \frac{\Delta_1^2}{\eta} \geq \Delta_0\left(1-\frac{1}{\eta}\right) + \frac{\Delta_0^2}{\eta}. \tag{4.37}$$

Lastly, notice that $\Delta_0 \geq \delta_0$.

Next, using (4.37) and the fact that $\gamma_0 = 1 - \sum_{\ell\in\Upsilon}\gamma_\ell$, we obtain from (4.32)–(4.33)

$$\begin{aligned}\widetilde{a}(v,v) &\geq [1-\eta]\,A + \left[\Delta_0\left(1-\frac{1}{\eta}\right) + \frac{\Delta_0^2}{\eta}\right]B + b_I(v,v)\\ &\geq [1-\eta]\,A + \left[\Delta_0 - \frac{\Delta_0(1-\Delta_0)}{\eta}\right]B + b_I(v,v).\end{aligned}$$

If we set $\delta = (1-\Delta_0) \leq (1-\delta_0)$, then using Lemma 4.1(iii), we can say

$$\begin{aligned}\frac{\beta\delta}{2}B \leq \frac{\beta(1-\delta_0)}{2}B &= \frac{\beta}{2}\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{P}_\Upsilon v\right\|^2\\ &\leq \beta\left\|\frac{1}{\sqrt{\sigma_t}}\mathcal{Q}^{-1/2}\mathbf{\Omega}\cdot\nabla v + \sqrt{\sigma_t}\mathcal{P}_\Upsilon\mathbf{\Omega}\cdot\nabla v\right\|^2 + \frac{\beta}{\sigma_t}b_I(v,v) = \beta A + \frac{\beta}{\sigma_t}b_I(v,v).\end{aligned}$$

This implies

$$\widetilde{a}(v,v) \geq [1-\eta-\beta]A + \left[(1-\delta) - \frac{\delta(1-\delta)}{\eta} + \frac{\beta\delta}{2}\right]B + \left[1-\frac{\beta}{\sigma_t}\right]b_I(v,v), \tag{4.38}$$

where $\beta\geq 0$ and $\eta\geq 0$ are to be determined.

Our choice of β is given by

$$\beta = \frac{2}{2+\delta}\left(\delta + \frac{\delta(1-\delta)}{\eta} - \eta\right), \tag{4.39}$$

which is found by setting $C_1 = C_2$, that is, setting

$$1 - \eta - \beta = (1-\delta) - \frac{\delta(1-\delta)}{\eta} + \frac{\beta\delta}{2}.$$

Plugging (4.39) into the coefficient on A generates the lower bound

$$C = 1 - \eta - \beta = \left(\frac{(2-\delta)}{(2+\delta)} - \frac{\delta}{2+\delta}\left(\eta + \frac{2(1-\delta)}{\eta}\right)\right). \tag{4.40}$$

This is maximized for $\eta = \sqrt{2(1-\delta)}$, which yields

$$C = \frac{(2-\delta) - 2\delta\sqrt{2(1-\delta)}}{2+\delta}. \tag{4.41}$$

Note that this is only valid when the corresponding $\beta \geq 0$, that is, (4.41) is valid only for $\delta \geq \delta_c$, where δ_c is the root of

$$\beta = \frac{2}{2+\delta}\left(\delta + \frac{\delta(1-\delta)}{\sqrt{2(1-\delta)}} - \sqrt{2(1-\delta)}\right) = \frac{2\delta - (2-\delta)\sqrt{2(1-\delta)}}{2+\delta} = 0.$$

This root is the only real root of the polynomial

$$\delta^3 - 3\delta^2 + 8\delta - 4 = 0.$$

Numerically we find that $\delta_c \approx 0.6117$, implying $C_c \approx 0.1188$. We also find numerically the minimum of C in (4.41) on $[\delta_c, 1]$ to be

$$C_m = \min_{\delta\in[\delta_c,1]} C \approx 0.06667, \tag{4.42}$$

which occurs at $\delta_m \approx 0.7836$.

Now, for $\delta \leq \delta_c$, we set $\beta = 0$ and choose η such that

$$1 - \eta = (1-\delta) - \frac{\delta(1-\delta)}{\eta},$$

resulting in

$$\eta = \frac{\delta + \sqrt{4\delta - 3\delta^2}}{2},$$

and, subsequently,

$$C = \frac{2 - \delta - \sqrt{4\delta - 3\delta^2}}{2}.$$

Clearly, C is a decreasing function of $\delta \in [0, \delta_c]$ implying it takes on its smallest value at δ_c. As mentioned previously $C \approx 0.1188$ for this value of δ.

We finally complete the proof with a bound on the coefficient of $b_I(v,v)$ in (4.38). The minimization of this term is only considered on $[\delta_c, 1]$ since $\beta = 0$ on $[0, \delta_c]$. Plotting β on $[\delta_c, 1]$ shows that β is an increasing function of δ on this interval implying $\beta \leq 2/3$. Thus, we have that $[1 - \beta/\sigma_t] \geq 1/3$. We finally conclude that

$$\widetilde{a}(v,v) \geq \widetilde{C}_e(A + B + b_I(v,v)),$$

where $\widetilde{C}_e = C_m \geq 0.06666$ coming from (4.42). Since $A \geq \left\|\mathcal{R}^{-1/2}\mathbf{\Omega}\cdot\nabla v\right\|^2$ and $2\,B \geq \left\|\mathcal{R}^{1/2}v\right\|^2$, we get

$$\widetilde{a}(v,v) \geq 0.03333\,\|v\|^2_{V_1},$$

implying that $C_e = 0.03333/2 = 0.01667$ from (4.31). ∎

5. Discretization and error bounds. Any finite dimensional subspace of V_1 may be used to construct an approximation to the solution of ψ. One approach, which is a subject of future research, is using a tessellation of the sphere to represent angular dependence and nonconforming finite elements to describe spatial variability. However, in this paper, we develop error bounds associated with a P_N approximation in angle and standard H^1 conforming finite elements in space. The angular approximation is represented by a truncated expansion of (2.12), which must be of greater order than the finite sum that represents the scattering kernel (i.e., $N \geq N_S$). A finite element approximation of the moments is defined on a triangulation $\mathcal{T}_h$ of R into hexahedrals or tetrahedrons.

Let $\mathbb{P}_k(\mathcal{T}_h)$ denote the space of piecewise polynomials of degree $\leq k$ on $\mathcal{T}_h$, let Π_h be the corresponding interpolation operator on $\mathbb{P}_k(\mathcal{T}_h)$, and let the truncation operator Π_N be defined by

$$\Pi_N\, v(\mathbf{x},\mathbf{\Omega}) := \sum_{\ell=0}^{N}\sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}). \tag{5.1}$$

Then the discrete space V^h is defined by

$$V^h := \left\{ v_h \in V \;:\; v_h = \sum_{\ell=0}^{N}\sum_{m=-\ell}^{\ell} \phi^h_{\ell m}(\mathbf{x}), Y_{\ell m}(\mathbf{\Omega});\quad \phi^h_{\ell m}(\mathbf{x}) \in \mathbb{P}_k(\mathcal{T}_h) \right\}. \tag{5.2}$$

The definition of V^h yields the discrete problem: find $\psi^h \in V^h$ such that

$$a(\psi^h, v^h) = \left\langle \mathcal{R}^{-1}\, q, \mathcal{L}_A\, v^h \right\rangle + 2 b_I(g, v^h) \tag{5.3}$$

for all $v^h \in V^h$.

Bounds for the discretization error are obtained by following the procedure outlined in [14]. Thus, let the components of $\mathbf{\Omega} \in S^2$ and $\mathbf{x} \in R$ be denoted by $\mathbf{\Omega} = (\Omega_1, \Omega_2, \Omega_3)$, $\mathbf{x} = (x_1, x_2, x_3)$, respectively, and let β, γ be a multi-index such that $D^\beta_{\mathbf{x}} := \frac{\partial^{|\beta|}}{\partial^{\beta_1}_{x_1}\partial^{\beta_2}_{x_2}\partial^{\beta_3}_{x_3}}$, and $D^\gamma_{\mathbf{\Omega}} := \frac{\partial^{|\gamma|}}{\partial^{\gamma_1}_{\Omega_1}\partial^{\gamma_2}_{\Omega_2}\partial^{\gamma_3}_{\Omega_3}}$. Recall that the standard norms [1] of $H^k(R)\times H^l(S^2)$ and $H^k(\partial R)\times H^l(S^2)$ are given by

$$\|v\|^2_{k,l} := \sum_{|\beta|\leq k}\sum_{|\gamma|\leq l} \left\| D^\gamma_{\mathbf{\Omega}} D^\beta_{\mathbf{x}} v \right\|,$$

$$\|v\|^2_{k,l,\partial R} := \sum_{|\beta|\leq k}\sum_{|\gamma|\leq l} \int_{\partial R}\int_{S^2} |D^\beta_{\mathbf{x}} v|^2 \,\mathrm{d}\mathbf{\Omega}\,\mathrm{d}\sigma.$$

Note also the following bounds for the interpolation error (see [4]):

$$\begin{aligned} \|v - \Pi_h v\|_{p,0} &\leq Ch^{k+1-p} \|v\|_{k+1,0} \ \forall v \in H^{k+1}(R) \times H^l(S^2), \\ \|v - \Pi_h v\|_{p,0,\partial R} &\leq Ch^{k+1-p} \|v\|_{k+1,0,\partial R} \ \forall v \in H^{k+1}(\partial R) \times H^l(S^2) \end{aligned} \tag{5.4}$$

for $p \in 0, 1$. We also define

$$\mathcal{E}_h(v) := v - \Pi_h v \tag{5.5}$$

and

$$\mathcal{E}_N(v) := v - \Pi_N v \tag{5.6}$$

for all $v \in \mathrm{L}^2(S^2 \times R)$.

To bound the error of the truncated expansion (5.1), we recall that the spherical harmonics are eigenfunctions of the Laplacian operator on the unit sphere, which implies

$$\begin{aligned} \Delta_{\mathbf{\Omega}} Y_{\ell m}(\Omega) &= \left[\tfrac{1}{\sin\theta}\tfrac{\partial}{\partial\theta}\left(\sin\theta\tfrac{\partial}{\partial\theta}\right) + \tfrac{1}{\sin^2\theta}\tfrac{\partial^2}{\partial\varphi^2}\right] Y_{\ell m}(\mathbf{\Omega}) \\ &= -l(l+1) Y_{\ell m}(\mathbf{\Omega}) \end{aligned} \tag{5.7}$$

for $l \geq 0$ and $m = -l, \ldots, 0, \ldots, l$. Next, for the reader's convenience, we include Lemma 4.1 from [14], as it is used throughout the remaining proofs.

LEMMA 5.1. *For $N \geq 1$, $|\beta| \leq k+1$, and $v \in V \cap (H^{k+1}(R) \times H^2(S^2))$ with $v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega})$, we have the following:*

(i) $\|\mathbf{\Omega} \cdot \nabla v\| \leq \sqrt{3} \sum_{i=1}^{3} \left\| \frac{\partial v}{\partial x_i} \right\|$.

(ii) $\left\| D_{\mathbf{x}}^{\beta} \phi_{\ell m}(\mathbf{x}) \right\|^2 \leq \frac{1}{[l(l+1)]^2} \int_{S^2} |\Delta_{\mathbf{\Omega}} D_{\mathbf{x}}^{\beta} v(\mathbf{x}, \mathbf{\Omega})|^2 \, \mathrm{d}\mathbf{\Omega}$.

(iii) $\left\| D_{\mathbf{x}}^{\beta} \mathcal{E}_N(v) \right\| \leq \frac{2}{N+1} \left\| \Delta_{\mathbf{\Omega}} D_{\mathbf{x}}^{\beta} v \right\|$.

(iv) $b_I(v, v) \leq \|v\|_{0,0,\partial R}^2$.

(v) $\|\mathcal{E}_N(v)\|_{0,0,\partial R} \leq \frac{2}{N+1} \|\Delta_{\mathbf{\Omega}} v\|_{0,0,\partial R}$.

(vi) *If, in addition, v satisfies the asymptotic expansion* (5.9)*, then*

$$\|\Pi_N(\mathcal{E}_h(v))\|_{0,0,\partial R} \leq C \frac{1}{\sigma_t} h^k \|\Delta_{\mathbf{\Omega}} \phi_R\|_{k+1,\partial R}.$$

Proof. See the proof in [14]. ∎

In the following, we present theorems for the thin regime (without proof) and for the thick regime with highly anisotropic scattering.

THEOREM 5.2 (thin regime). *Suppose that $N > N_S \geq 1$, $0 \leq \sigma_a \leq \sigma_t \leq 1$ and that $\|\cdot\|_{V_1}$ is defined as in* (3.12)*. Let $\psi \in V_1 \cap (H^{k+1}(R) \times H^2(S^2))$ be the solution of* (4.5)*, and let ψ^h be the solution of* (5.3) *with V^h defined by* (5.2)*. Then we have*

$$\left\|\psi - \psi^h\right\|_{V_1} \leq \frac{C_1}{N+1} \left(\|\Delta_{\mathbf{\Omega}} \psi\|_{1,0} + \|\Delta_{\mathbf{\Omega}} \psi\|_{0,0,\partial R} \right) + C_2 h^k \left(\|\psi\|_{k+1,0} + \|\psi\|_{0,\partial R} \right)$$

with C_1 and C_2 independent of σ_t and σ_a.

Proof. See the proof of the isotropic case in [14]. ∎

The error bounds for the thick regime require considering the asymptotic limit defined by $\sigma_t \to \infty$. These bounds depend on the asymptotic form of ψ, which itself depends on assumptions regarding material parameters. For the case of isotropic

scattering, this limit has been extensively examined, where it is assumed that $\sigma_a = \zeta/\sigma_t$, where ζ is bounded independently of σ_t as $\sigma_t \to \infty$ (see [7, 6, 9, 8, 17]). In this limit, referred to as the diffusion limit, the solution to (2.1) with isotropic scattering can then be expressed as

$$\psi(\mathbf{x}, \mathbf{\Omega}) = \phi_D(\mathbf{x}) + \frac{1}{\sigma_t}\phi_R(\mathbf{x}, \mathbf{\Omega}),$$

with ϕ_R bounded independently of σ_t and the leading-order term ϕ_D satisfying a diffusion equation.

In [10], for anisotropic scattering Larsen and Pomraning presented two different asymptotic limits for $\sigma_t \to \infty$. These two different limits rely on different assumptions on the degree of anisotropy in the scattering. One is for the case of mildly anisotropic scattering, and the other is for highly anisotropic scattering. In the following, we only examine the case of highly anisotropic scattering because the case of mildly anisotropic scattering yields results identical to Theorem 4.3 of [14].

To define the asymptotic limit, we let ζ_l and ω_l be $O(1)$ constants. Then we define μ_ℓ in terms of these constants as

$$\mu_\ell = \begin{cases} \zeta_l, & 0 \le l \le N_S, \\ \frac{\sigma_t}{\omega_l + 1}, & l > N_S. \end{cases} \tag{5.8}$$

Note that with these assumptions, we get that $\sigma_a = O(1)$ and

$$\sigma_l = \begin{cases} \frac{\sigma_t - \zeta_l}{\sigma_t - \sigma_a}, & 0 \le l \le N_S, \\ \frac{\sigma_t \omega_l}{\sigma_s(\omega_l + 1)}, & l > N_S. \end{cases}$$

Under these assumptions, Larsen and Pomraning in [10] illustrated that ψ can be expressed as

$$\psi(\mathbf{x}, \mathbf{\Omega}) = \hat{\phi}_D(\mathbf{x}, \mathbf{\Omega}) + \frac{1}{\sigma_t}\hat{\phi}_R(\mathbf{x}, \mathbf{\Omega}), \tag{5.9}$$

where

$$\hat{\phi}_D(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell=0}^{N_S} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega})$$

satisfies the first-order P_{N_S} equations and $\hat{\phi}_R(\mathbf{x}, \mathbf{\Omega})$ can be bounded independently of σ_t.

Remark. The P_{N_S} equations are a set of $(N_S + 1)^2$ differential equations for $(N_S + 1)^2$ unknowns, which are obtained by substituting $\psi_{N_S} := \Pi_{N_S}\psi$ for ψ in (2.1) and setting the resulting equation orthogonal to all spherical harmonics up to order N_S. Furthermore, as was illustrated in [14], the least-squares formulation described here is nearly identical to a least-squares minimization of the P_{N_S+1} equations.

The two components, $\hat{\phi}_D$ and $\hat{\phi}_R$, of (5.9) are not orthogonal in $\mathrm{L}^2(S^2)$. But we can rewrite (5.9) such that this condition holds. Note that this condition is employed in the proof of Theorem 5.3. This new expression is

$$\psi(\mathbf{x}, \mathbf{\Omega}) = \phi_D(\mathbf{x}, \mathbf{\Omega}) + \frac{1}{\sigma_t}\phi_R(\mathbf{x}, \mathbf{\Omega}), \tag{5.10}$$

where $\mathcal{P}_\Sigma \phi_R = 0$ for $\Sigma = \{l \in \mathbb{N} \ : \ l \leq N_S\}$. Before proving Theorem 5.3, we introduce the notation $\|a\| \lesssim \|b\|$ meaning $\|a\| \leq C \|b\|$, where C denotes an arbitrary, parameter-independent, positive constant.

THEOREM 5.3. (diffusive regime with highly anisotropic scattering) *Suppose that $N > N_S \geq 1$, $1 \leq \sigma_t < \infty$ and that $\|\cdot\|_{V_1}$ is defined as in* (3.12). *Let $\psi \in V_1 \cap (H^{k+1}(R) \times H^2(S^2))$ be the solution of* (4.5), *and let ψ^h be the solution of* (5.3) *with V^h as defined in* (5.2) *using the scaling operator* (4.18). *Assuming that $\zeta_\ell \in (\zeta_m, \zeta_M)$ and $\omega_\ell \in (\omega_m, \omega_M)$ $\forall l$, where the minimum and maximum terms are $O(1)$ and independent of σ_t, and assuming that ψ satisfies the expansion* (5.10), *then*

$$\left\|\psi - \psi^h\right\|_{V_1} \quad \leq \quad \frac{C_1 D_1(\sigma_t, \phi_R)}{\sigma_t^{1/2}(N+1)} + C_2 D_2(\sigma_t, \phi_D, \phi_R)\, h^k$$

with C_1 and C_2 independent of σ_t and σ_a, and

$$D_1(\sigma_t, \phi_R) := \frac{1}{\sigma_t} \sum_{i=1}^{3} \left\| \Delta_{\mathbf{\Omega}} \frac{\partial \phi_R}{\partial x_i} \right\| + \|\Delta_{\mathbf{\Omega}} \phi_R\| + \frac{1}{\sigma_t^{1/2}} \|\Delta_{\mathbf{\Omega}}\, \phi_R\|_{0,0,\partial R}\,,$$

and

$$D_2(\sigma_t, \phi_D, \phi_R) := \|\phi_D\|_{k+1,0} + \|\Delta_{\mathbf{\Omega}} \phi_D\|_{k+1,0,\partial R} + \frac{1}{\sigma_t^{3/2}} \|\phi_R\|_{k+1,0} + \frac{1}{\sigma_t} \|\Delta_{\mathbf{\Omega}} \phi_R\|_{k+1,0,\partial R}\,.$$

Proof. Combining Céa's lemma with Theorem 4.4 yields

$$\|\psi - \psi_h\|_{V_1} \lesssim \left(\|\mathcal{E}_N(\psi)\|_{V_1} + \|\Pi_N(\psi - \Pi_h \psi)\|_{V_1} \right).$$

Using

$$\|v\|_{V_1} \leq \left(\|v\|_V^2 + \|v\|_{0,0,\partial R}^2 \right)^{1/2} \leq \|v\|_V + \|v\|_{0,0,\partial R}\,,$$

which is obtained from

$$b_I(v, v) \leq \int_{\partial R} \int_{S^2} |v|^2 = \|v\|_{0,0,\partial R}^2\,,$$

we have

(5.11)

$$\|\psi - \psi_h\|_{V_1} \lesssim \left(\|\mathcal{E}_N(\psi)\|_V + \|\mathcal{E}_N(\psi)\|_{0,0,\partial R} + \|\Pi_N(\mathcal{E}_h(\psi))\|_V + \|\Pi_N(\mathcal{E}_h(\psi))\|_{0,0,\partial R} \right).$$

Next, we note that $\sigma_t \mathcal{E}_N(\psi) = \mathcal{E}_N(\phi_R)$ because of the fact that ψ satisfies (5.10). $\Pi_{N_S}\left(\mathbf{\Omega} \cdot \nabla(\mathcal{E}_N(\phi_R))\right) = 0$ and $\Pi_{N_S}(\mathcal{E}_N(\phi_R)) = 0$ because of our our assumptions on N. Now, we bound the first term of (5.11) as

$$\begin{aligned}
\|\mathcal{E}_N(\psi)\|_V &= \frac{1}{\sigma_t} \|\mathcal{E}_N(\phi_R)\|_V \\
&\lesssim \frac{1}{\sigma_t^{3/2}} \|\mathcal{E}_{N_S}((\mathbf{\Omega} \cdot \nabla)\, \mathcal{E}_N(\phi_R))\| + \frac{1}{\sigma_t^{1/2}} \|\mathcal{E}_{N_S}(\mathcal{E}_N(\phi_R))\| \\
&\lesssim \frac{1}{\sigma_t^{1/2}} \left(\frac{1}{\sigma_t} \sum_{i=1}^{3} \left\| \frac{\partial}{\partial x_i} \mathcal{E}_N(\phi_R) \right\| + \|\mathcal{E}_N(\phi_R)\| \right) \\
&\lesssim \frac{1}{\sigma_t^{1/2}(N+1)} \left(\frac{1}{\sigma_t} \sum_{i=1}^{3} \left\| \Delta_{\mathbf{\Omega}} \frac{\partial \phi_R}{\partial x_i} \right\| + \|\Delta_{\mathbf{\Omega}} \phi_R\| \right),
\end{aligned}$$

where we used (i) and (iii) of Lemma 5.1 and the fact that $(\mathcal{I} - \Pi_{N_S})$ is an $\mathrm{L}^2(S^2)$ orthogonal projection. We bound the second term of (5.11) according to

$$\|\mathcal{E}_N(\psi)\|_{0,0,\partial R} = \frac{1}{\sigma_t} \|\mathcal{E}_N(\phi_R)\|_{0,0,\partial R} \leq \frac{2}{\sigma_t(N+1)} \|\Delta_{\mathbf{\Omega}}\phi_R\|_{0,0,\partial R},$$

given (v) of Lemma 5.1.

For the third term of (5.11), we first need to introduce $\mathcal{O}_{N_S} := \Pi_{N_S}(\mathbf{\Omega}\cdot\nabla)$. Since $\sigma_a > 1/\sigma_t$, we have

$$||\Pi_N(\mathcal{E}_h(\psi))||_V = ||\mathcal{E}_h(\phi_D)||_V + \frac{1}{\sigma_t} ||\Pi_N(\mathcal{E}_h(\phi_R))||_V .$$

We then say

$$\begin{aligned}
||\mathcal{E}_h(\phi_D)||_V &\lesssim \frac{1}{\sigma_t^{1/2}} ||\mathcal{E}_{N_S}((\mathbf{\Omega}\cdot\nabla)\,\mathcal{E}_h(\phi_D))|| + ||\mathcal{O}_{N_S}\mathcal{E}_h(\phi_D)|| + ||\mathcal{E}_h(\phi_D)|| \\
&\lesssim ||\mathbf{\Omega}\cdot\nabla(\mathcal{E}_h(\phi_D))|| + ||\mathcal{E}_h(\phi_D)|| \\
&\lesssim \sum_{i=1}^{3} \left|\left| \frac{\partial}{\partial x_i}\mathcal{E}_h(\phi_D) \right|\right| + ||\mathcal{E}_h(\phi_D)|| \\
&\lesssim h^k \left(\|\phi_D\|_{k+1,0} + h\,\|\phi_D\|_{k+1,0} \right) \\
&\lesssim h^k \|\phi_D\|_{k+1,0}
\end{aligned}$$

and

$$\begin{aligned}
||\Pi_N(\mathcal{E}_h(\phi_R))||_V &\lesssim \frac{1}{\sigma_t} ||\mathcal{O}_{N_S}\mathcal{P}_{N_S+1}(\mathcal{E}_h(\phi_R))|| + \frac{1}{\sigma_t^{3/2}} ||\mathcal{E}_{N_S}((\mathbf{\Omega}\cdot\nabla)\,\Pi_N\mathcal{E}_h(\phi_R))|| \\
&\quad + \frac{1}{\sigma_t^{1/2}} ||\Pi_N\mathcal{E}_h(\phi_R)|| \\
&\lesssim \frac{1}{\sigma_t} ||\mathbf{\Omega}\cdot\nabla(\mathcal{E}_h(\phi_R))|| + \frac{1}{\sigma_t^{1/2}} ||\Pi_N\mathcal{E}_h(\phi_R)|| \\
&\lesssim \sum_{i=1}^{3} \frac{1}{\sigma_t} \left|\left| \frac{\partial}{\partial x_i}\mathcal{E}_h(\phi_R) \right|\right| + \frac{1}{\sigma_t^{1/2}} ||\mathcal{E}_h(\phi_R)|| \\
&\lesssim \frac{h^k}{\sigma_t^{1/2}} \left(\frac{1}{\sigma_t^{1/2}} \|\phi_R\|_{k+1,0} + h\,\|\phi_R\|_{k+1,0} \right) \\
&\lesssim \frac{h^k}{\sigma_t^{1/2}} \|\phi_R\|_{k+1,0} .
\end{aligned}$$

Subsequently,

$$||\Pi_N(\mathcal{E}_h(\psi))||_V \lesssim h^k \left(\|\phi_D\|_{k+1,0} + \frac{1}{\sigma_t^{3/2}} \|\phi_R\|_{k+1,0} \right).$$

Last, we can bound the fourth term of (5.11) according to

$$\|\Pi_N(\mathcal{E}_h(\psi))\|_{0,0,\partial R} \lesssim h^k \|\Delta_{\mathbf{\Omega}}\psi\|_{k+1,0,\partial R} \lesssim \|\Delta_{\mathbf{\Omega}}\phi_D\|_{k+1,0,\partial R} + \frac{1}{\sigma_t} \|\Delta_{\mathbf{\Omega}}\phi_R\|_{k+1,0,\partial R},$$

where we have used (iv) of Lemma 5.1. ∎

6. Final remarks. In this paper, we have extended the least-squares method for the linear Boltzmann equation to the case of anisotropic scattering by establishing uniqueness and existence of the minimization problem (4.5). Furthermore, the ellipticity is with respect to a physically meaningful norm, and the ellipticity constants are independent of the problem parameters. Using the ellipticity constants, we have also established error bounds in all three parameter regimes.

Future work consists of examining the least-squares approach with respect to more complex discretization approaches. Besides spherical harmonics approximations, one can use tessellations of the sphere as a finite element representation of the angular dependency. One of the main advantages to this approach is that there is a reduced coupling among moments as compared to the spherical harmonics approach used here. For the spatial domain, we plan to investigate nonconforming finite elements so as to better approximate problems having discontinuous solutions. Lastly, since this work provides a partial foundation for [5], we hope to provide a complete foundation by extending the results introduced here to the case of multiple energy groups.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] B. Akherraz, C. Fedon-Magnaud, J. Lautard, and R. Sanchez, *Anisotropic scattering treatment for the neutron-transport equation with primal finite-elements*, Nuclear Sci. Engrg., 120 (1995), pp. 187–198.
[3] G. Arfken, *Mathematical Methods for Physicists*, 3rd ed., Academic Press, San Diego, 1985.
[4] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 1997.
[5] B. Chang and B. Lee, *A multigrid algorithm for solving the multigroup, anisotropic scattering Boltzmann equation using first-order system least-squares methodology*, Electron. Trans. Numer. Anal., 15 (2001), pp. 132–151.
[6] E. Larsen, *Diffusion theory as an asymptotic limit of transport theory for nearly critical systems with small mean free path*, Ann. of Nuclear Energy, 7 (1980), pp. 249–255.
[7] E. Larsen, *On numerical solutions of transport problems in the diffusion limit*, Nuclear Sci. Engrg., 83 (1983), pp. 90–99.
[8] E. Larsen and J. Morel, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II*, J. of Comput. Phys., 83 (1989), p. 212.
[9] E. Larsen, J. Morel, and W. Miller, *Asymptotic solutions of numerical transport problems in optically thick diffusive regimes*, J. of Comput. Phys., 69 (1987), pp. 283–324.
[10] E. Larsen and G. Pomraning, *The P_N theory as an asymptotic limit of transport theory in planar geometry - I: Analysis*, Nuclear Sci. Engrg., 109 (1991), pp. 49–75.
[11] E. Lewis and W. Miller, *Computational Methods of Neutron Transport*, American Nuclear Society, La Grange Park, IL, 1993.
[12] T. Manteuffel and K. Ressel, *Multigrid methods for transport equations in diffusive regimes*, in Proceedings on the Copper Mountain Conference on Multigrid Methods, 1995.
[13] T. Manteuffel and K. Ressel, *Least-squares finite element solution of the neutron transport equation in diffusive regimes*, SIAM J. Numer. Anal., 35 (1998), pp. 806–835.
[14] T. Manteuffel, K. Ressel, and G. Starke, *A boundary functional for the least-squares finite element solution of neutron transport problems*, SIAM J. Numer. Anal., 37 (2000), pp. 556–586.
[15] J. E. Morel, *Fokker-Planck calculations using standard discrete ordinates transport codes*, Nuclear Sci. Engrg., 79 (1981), pp. 340–356.
[16] G. Palmiotti, C. B. Carrico, and E. E. Lewis, *Variational nodal transport methods with anisotropic scattering*, Nuclear Sci. Engrg., 115 (1993), pp. 233–243.
[17] G. Pomraning, *Diffusive limits for linear transport equations*, Nuclear Sci. Engrg., 112 (1992), pp. 239–255.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 561–585

FOURIER SPECTRAL APPROXIMATION TO LONG-TIME BEHAVIOR OF DISSIPATIVE GENERALIZED KdV-BURGERS EQUATIONS*

SHUJUAN LÜ† AND QISHAO LU†

Abstract. In this paper, we consider a generalized KdV-Burgers equation with periodic initial value condition. Firstly, a fully discrete Galerkin–Fourier spectral approximation scheme, which is a linear one, is constructed. Next, the dynamical properties of the discrete system are analyzed for the autonomous case. The existence and the convergence of the global attractors of the discrete system are obtained by a priori estimates and the error estimates of the discrete solution. The stability of the discrete scheme is also proved. Finally, the long-time stability and the convergence of the discrete scheme are proved for the nonautonomous case. All results in this paper are obtained without any restriction on the time step size.

Key words. long-time stability, long-time convergence, global attractors, spectral methods, KdV-Burgers equation

AMS subject classifications. 65M60, 65N35, 65N30

DOI. 10.1137/S0036142903426671

1. Introduction and description of the method.

In the wake of the development in the study of infinite-dimensional dynamical systems, the long-time behavior of dissipative nonlinear partial differential equations has attracted more and more attention of scientists—for example, the existence and the estimate of dimension of global attractor, the inertial manifolds and the approximate inertial manifolds, the structure of global attractor, and so on (see [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]). However, these studies depended on the results of numerical experimentation to a great extent. For this reason, it is worth studying whether the numerical results are reliable and the calculation schemes are suitable. This work began in the late 1980s (see [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]). Most of the existing classical Galerkin methods are nonlinear schemes, and the calculation is complex. The nonlinear Galerkin methods sometimes can be linear schemes, but the dimension of which is many times that of the classical Galerkin ones; consequently, the computation amount is very heavy. In addition, the stability and convergence of these discrete schemes were established under certain restrictions on the time step size.

In this paper, we construct a fully discrete classical Galerkin spectral scheme, which is a linear scheme. On the one hand, in comparison with the existing nonlinear Galerkin methods with linear schemes or classical Galerkin methods with nonlinear schemes, the computation amount of this scheme can be greatly reduced. On the other hand, without any restriction on the time step size, the results about the uniform stability and convergence of this discrete scheme are obtained.

To be more specific in this paper, we restrict ourselves to a class of damped generalized KdV-Burgers equations. More generally, similar schemes and analysis are applicable to other dissipative dynamical systems.

*Received by the editors April 22, 2003; accepted for publication September 2, 2005; published electronically March 17, 2006. This work was supported by the National Natural Science Foundation of China (10432010 and 10571010).

http://www.siam.org/journals/sinum/44-2/42667.html

†Department of Mathematics and LIMB, Beijing University of Aeronautics and Astronautics, Beijing 100083, P. R. China (lsj@buaa.edu.cn, qishaolu@hotmail.com).

The damped KdV-Burgers equation is a kind of important nonlinear evolution equations. They are proposed in many physical problems; for example, see [21, 22]. For this reason, the research of them is of theoretical and practice significance. In [23, 24], the existence and the uniqueness for the global smooth solution of this kind of equation have been proved. In [25] the existence of chaotic phenomena for a KdV-Burgers equation has been found. In this paper, we consider the periodic initial value problem of the damped generalized KdV-Burgers equation

$$u_t - \alpha u_{xx} + \beta u_{xxx} + f(u)_x = g(u) + h, \quad x \in R, t > 0, \tag{1.1}$$

$$u(x + 2\pi, t) = u(x, t), \qquad x \in R, t \geq 0, \tag{1.2}$$

$$u(x, 0) = u_0(x), \qquad x \in R, \tag{1.3}$$

where $\alpha,\ \beta$ are real constants and $\alpha > 0,\ \beta \neq 0$; $f(u),\ g(u),\ h$ are real value functions.

In [6, 20] the existence of global attractors and the estimates of upper bounds of their Hausdorff and fractal dimensions were proved for problem (1.1)–(1.3). In [19, 20] the Fourier discrete schemes, which are nonlinear schemes, were constructed, and the existence and the convergence of approximate attractors were proved. In this paper, we first construct a fully discrete Fourier spectral approximation scheme, which is a linear one. Then the existence and the convergence of approximate attractors, as well as the stability of discrete scheme, are proved in the autonomous case. Furthermore, the long-time stability and the convergence of discrete scheme are obtained in the nonautonomous case.

Throughout this paper we use the following notation: $\Omega = [0, 2\pi]$; $(\cdot, \cdot)$ denotes the inner product of $L^2(\Omega)$, $\|\cdot\|_m$ the norm of Sobolev spaces $H^m(\Omega)$, and $\|\cdot\| = \|\cdot\|_0$, $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(\Omega)}$.

For any given positive integer N, let $S_N = \mathrm{Span}\{\sin kx,\ \cos kx :\ |k| \leq N\}$, and denote by $P_N :\ L^2_p(\Omega) \to S_N$ the orthogonal projection operator (see [26]).

Let τ be the mesh size in the variable t, $t_k = k\tau$, $u^k = u(x, t_k)$, $\overline{\partial}_t u^k = \frac{1}{\tau}(u^k - u^{k-1})$. The Fourier spectral scheme for solving (1.1)–(1.3) is to find $u_N^k \in S_N$ such that

$$\begin{aligned} &(\overline{\partial}_t u_N^k - \alpha u_{Nxx}^k + \beta u_{Nxxx}^k,\ \varphi) + B(u_N^{k-1},\ u_N^k,\ \varphi) \\ &= (G(u_N^{k-1})u_N^k + h^k,\ \varphi) \quad \text{for all } \varphi \in S_N, \quad k = 1, 2, \ldots, \end{aligned} \tag{1.4}$$

$$u_N^0 = P_N u_0, \tag{1.5}$$

where

$$B(u_N^{k-1},\ u_N^k,\ \varphi) = (F(u_N^{k-1})u_{Nx}^k,\ \varphi) - (F(u_N^{k-1})u_N^k\ ,\ \varphi_x),$$

$$F(u) = \begin{cases} \dfrac{1}{u^2}\displaystyle\int_0^u sf'(s)ds & \text{for } u \neq 0, \\ \dfrac{1}{2}f'(0) & \text{for } u = 0, \end{cases}$$

$$G(u) = \begin{cases} \dfrac{g(u)}{u} & \text{for } u \neq 0, \\ g'(0) & \text{for } u = 0. \end{cases}$$

It is a linear iteration scheme, and then it needs only to solve a class of linear algebraic equations for every iteration.

The following lemmas are necessary for further discussion.

LEMMA 1.1 (see [26]). *If $u \in H_p^m(\Omega)$, then there exists a constant c independent of u, N such that*

$$\|u - P_N u\|_\mu \le cN^{\mu-m}\|D^m u\| \qquad \text{for all } 0 \le \mu \le m.$$

LEMMA 1.2 (Sobolev interpolation inequality [27]). *Suppose that $u \in L^q(\Omega)$, $D^m u \in L^r(\Omega)$, $\Omega \subset R^n$, $1 \le r \le \infty$, $0 \le j \le m$. Then there exists a constant $c = c(j, m, \Omega, p, q, r)$ independent of u such that*

$$\|D^j u\|_{L^p} \le c\|D^m u\|^a_{W^{m,r}(\Omega)}\|u\|^{1-a}_{L^q},$$

where

$$\frac{1}{p} = \frac{j}{n} + a\Big(\frac{1}{r} - \frac{m}{n}\Big) + (1-a)\frac{1}{q}, \quad \frac{j}{m} < a < 1.$$

LEMMA 1.3 (discrete Gronwall's inequality [17]). *Let y^k, g^k, h^k be three series satisfying*

$$\frac{y^{k+1} - y^k}{\tau} \le g^k y^k + h^k, \quad k = 0, 1, 2, \ldots.$$

Then we have

$$y^n \le y^0 \exp\left(\tau \sum_{k=0}^{n-1} g^k\right) + \tau \sum_{k=0}^{n-1} h^k \exp\left(\tau \sum_{i=k}^{n-1} g^i\right) \qquad \text{for all } n \ge 1.$$

LEMMA 1.4 (discrete uniform Gronwall's inequality [17]). *Let y^k, g^k, h^k be three series satisfying*

$$\frac{y^{k+1} - y^k}{\tau} \le g^k y^k + h^k \qquad \text{for all } k \ge k_0$$

and

$$\tau \sum_{k=k_1}^{n_0+k_1} g^k \le \alpha_1, \quad \tau \sum_{k=k_1}^{n_0+k_1} h^k \le \alpha_2, \quad \tau \sum_{k=k_1}^{n_0+k_1} y^k \le \alpha_3 \qquad \text{for all } k_1 \ge k_0$$

with $\tau n_0 = r$. Then

$$y^k \le \Big(\frac{\alpha_3}{r} + \alpha_2\Big)\mathrm{e}^{\alpha_1} \qquad \text{for all } k \ge\ n_0 + k_0.$$

An outline of this paper is as follows. We consider the autonomous case (i.e., $h = h(x)$) in sections 2 to 4. In section 2 the existence of discrete attractors $\mathcal{A}_N^\tau$ is obtained by the t-independent priori estimates of discrete solutions (Theorem 2.2). In section 3 the convergence of $\mathcal{A}_N^\tau$ is proved by the error estimates in $[0, +\infty)$ of the discrete solutions (Theorem 3.5). In section 4 we prove the stability of the discrete scheme (Theorem 4.1). Finally, we consider the nonautonomous case (i.e., $h = h(x, t)$) in section 5. The long-time stability and the convergence of discrete scheme are proved (Theorem 5.3 and Theorem 5.4).

2. Existence of approximation global attractors $\mathcal{A}_N^\tau$. In sections 2 to 4, we let $h = h(x)$ and $h(x+2\pi) = h(x)$. The main purpose of this section is to prove the existence of the global attractors $\mathcal{A}_N^\tau$ of problem (1.4)–(1.5). To the end, we need the following result (see [1]).

THEOREM 2.1. *Let H be a Banach space, $\{S(t), t \geq 0\}$ a set of continuous semigroup operations, i.e., $S(t) : H \to H$ satisfies*

$$S(t+\tau) = S(t) \cdot S(\tau) \quad \text{for all } t \geq 0,\ \tau \geq 0, \quad S(0) = I,$$

where I is the identity operator. We also assume that

(i) *there exists a bounded absorbing set $\mathcal{B}_0 \subset H$, i.e., for any given bounded set $\mathcal{B} \subset H$, there exists a constant $T(\mathcal{B})$ such that*

$$S(t)\mathcal{B} \subset \mathcal{B}_0 \qquad \text{for all } t \geq T(\mathcal{B});$$

(ii) *the operator $S(t)$ is uniformly compact for t enough large. By this we mean that for every bounded set $\mathcal{B}$ there exists a constant $t_0 = t_0(\mathcal{B})$ such that*

$$\bigcup_{t \geq t_0} S(t)\mathcal{B}$$

is relatively compact in H.

Then the semigroup of operators $\{S(t)\}_{t\geq}$ has a compact global attractor $\mathcal{A} \subset H$. By this we mean that

(a) $S(t)\mathcal{A} = \mathcal{A}$ for all $t \geq 0$;

(b) *for any given bounded set $\mathcal{B} \subset H$, $\lim_{t\to\infty} \operatorname{dist}(S(t)\mathcal{B}, \mathcal{A}) = 0$, where*

$$\operatorname{dist}(X, Y) = \sup_{x\in X} \inf_{y\in Y} \|x - y\|_H.$$

Now we give the main result in this section.

THEOREM 2.2. *If $f \in C^3$, $|f'(u)| \leq A|u|^2$, $g \in C^2$, $g(0) = 0$, $g'(u) \leq b < 0$, $|g(u)| \leq B|u|(|u|^4 + 1)$, $h(x) \in H_p^1(\Omega)$, and $u_0(x) \in H_p^2(\Omega)$, then the semigroup of operator $\{S_N^\tau(n)\}_{n\geq 0}$ generated by problem (1.4), (1.5) has a compact global attractor $\mathcal{A}_N^\tau \subset H_p^2(\Omega) \cap S_N$.*

To prove Theorem 2.2, the following results are necessary.

LEMMA 2.3. *If $f, g \in C^1$, $g(0) = 0$, $g'(s) \leq b < 0$; $h \in L^2(\Omega)$, $u_0 \in L_p^2(\Omega)$. Then for the solution u_N^n of problem (1.4), (1.5), we have the estimates*

$$\|u_N^n\|^2 \leq (1 - b\tau)^{-n}\|u_N^0\|^2 + \frac{\|h\|^2}{b^2} \leq \|u_0\|^2 + \frac{\|h\|^2}{b^2} \triangleq E_0^2 \quad \text{for all } n \geq 1,$$

$$\overline{\lim_{n\to\infty}} \|u_N^k\|^2 \leq \frac{\|h\|^2}{b^2} \triangleq (\rho_0')^2,$$

$$\tau^2 \sum_1^n \|\overline{\partial}_t u_N^k\|^2 \leq C_1(1 + t_n) \qquad \text{for all } n \geq 1,$$

where the constant $C_0 = C_0(\|u_o\|)$ is independent of N, n, and τ.

Proof. Set $\varphi = u_N^k$ in (1.4). Then we have

$$\frac{1}{2}\overline{\partial}_t\|u_N^k\|^2 + \frac{\tau}{2}\|\overline{\partial}_t u_N^k\|^2 + \alpha\|u_{Nx}^k\|^2 + B(u_N^{k-1}, u_N^k, u_N^k) = (G(u_N^{k-1})u_N^k, +h, u_N^k).$$

From the definition of $B(u,v,\varphi)$, $G(u)$ and ε- inequality, we obtain

$$\begin{aligned}
&B(u_N^{k-1},u_N^k,u_N^k)=0,\\
&(G(u_N^{k-1})u_N^k,u_N^k)=(g'(\theta u_N^{k-1})u_N^k,u_N^k)\le b\|u_N^k\|^2,\\
&(h,u_N^k)\le\|h\|\,\|u_N^k\|\le\frac{|b|}{2}\|u_N^k\|^2+\frac{1}{2|b|}\|h\|^2.
\end{aligned}$$

Therefore

$$\overline{\partial}_t\,\|u_N^k\|^2+\tau\,\|\overline{\partial}_t u_N^k\|^2+\alpha\,\|u_{Nx}^k\|^2-b\,\|u_N^k\|^2\le\frac{\|h\|^2}{|b|}. \tag{2.1}$$

Multiplying (2.1) by $(1-b\tau)^{k-1}$ and summing them for k from 1 to n, we have

$$\|u_N^n\|^2\le(1-b\tau)^{-n}\Big(\|u_N^0\|^2-\frac{\|h\|^2}{b^2}\Big)+\frac{\|h\|^2}{b^2}\le\|u_N^0\|^2+\frac{\|h\|^2}{b^2}\triangleq E_0^2,$$

which implies that

$$\overline{\lim_{n\to\infty}}\|u_N^k\|^2\le\frac{\|h\|^2}{b^2}\triangleq(\rho_0')^2.$$

Taking the sum of (2.1) for k from k_0+1 to n, we recover the proof of the lemma. ∎

COROLLARY 2.4. *For any given $\rho_0>\rho_0'$ and $R_0>0$, if $\|u_0\|\le R_0$, then*

$$\|u_N^n\|^2\le\rho_0^2\qquad\text{for all } n\ge n_0=\Big(\ln\frac{R_0^2}{\rho_0^2-(\rho_0')^2}\Big)\Big/\ln(1-b\tau).$$

LEMMA 2.5. *In addition to the conditions of Lemma 2.3, we suppose that $f\in C^2$, $|f'(u)|\le A|u|^2$; $|g(u)|\le B|u|(|u|^4+1)$; $u_0\in H_p^1(\Omega)$, $\|u_0\|\le R_0$. Then for the solution u_N^n of problem (1.4), (1.5), we have the estimates*

$$\begin{aligned}
\|u_{Nx}^k\|^2&\le\rho_1^2\qquad\text{for } n\ge n_0+N_0\triangleq n_1,\\
\|u_{Nx}^k\|^2&\le E_1^2\qquad\text{for } n\ge 1,\\
\tau^2\sum_{k=1}^{n}\|\overline{\partial}_t u_{Nx}^k\|^2&\le C_1(1+t_n)\qquad\text{for all } n\ge 1,
\end{aligned}$$

where n_0 is given by Corollary 2.4, N_0 an arbitrary positive integer, r an arbitrary positive number such that $N_0\tau=r$, the constant ρ_1 independent of N, n, τ, and $\|u_0\|_1$, $C_1=C_1(\|u_o\|_1)$, and $E_1=E_1(\|u_0\|_1)$ independent of N, n, and τ.

Proof. Let $\varphi=-u_{Nxx}^k$ in (1.4). Then we have

$$\begin{aligned}
&\frac{1}{2}\overline{\partial}_t\|u_{Nx}^k\|^2+\frac{\tau}{2}\|\overline{\partial}_t u_{Nx}^k\|^2+\alpha\|u_{Nxx}^k\|^2+B(u_N^{k-1},u_N^k,-u_{Nxx}^k)\\
&\qquad=(G(u_N^{k-1})u_N^k+h,-u_{Nxx}^k).
\end{aligned}\tag{2.2}$$

We now estimate the last two terms in the above equality. First, from the definition of $B(u,v,\varphi)$, we have

$$B(u_N^{k-1},u_N^k,-u_{Nxx}^k)=2(F(u_N^{k-1})u_{Nx}^k,-u_{Nxx}^k)+(F'(u_N^{k-1})u_{Nx}^{k-1}u_N^k,-u_{Nxx}^k),$$

where from the definition of $F(u)$, the mean theorem of integration, and the Sobolev interpolation inequality, we have for two terms in the above equality

$$\begin{aligned}
&2(F(u_N^{k-1})u_{Nx}^k, -u_{Nxx}^k)\\
=\ & 2\left(\frac{u_{Nx}^k}{(u_N^{k-1})^2}\int_0^{u_N^{k-1}} sf'(s)ds, -u_{Nxx}^k\right) = \ (f'(\theta u_N^{k-1})u_{Nx}^k, -u_{Nxx}^k)\\
\le\ & A\|u_N^{k-1}\|_\infty^2\,\|u_{Nx}^k\|\,\|u_{Nxx}^k\| \le c\|u_N^{k-1}\|\,\|u_N^{k-1}\|_1\,\|u_N^k\|^{\frac{1}{2}}\,\|u_{Nxx}^k\|^{\frac{3}{2}}\\
\le\ & \frac{\alpha}{8}\|u_{Nxx}^k\|^2 + c(\|u_N^{k-1}\|^4\,\|u_N^k\|^2\,\|u_{Nx}^{k-1}\|^4 + \|u_N^{k-1}\|^8\,\|u_N^k\|^2)
\end{aligned}$$

and

$$\begin{aligned}
&(F'(u_N^{k-1})u_{Nx}^{k-1}u_N^k, -u_{Nxx}^k)\\
=\ & \left(\left((u_N^{k-1})^2 f'(u_N^{k-1}) - 2\int_0^{u_N^{k-1}} sf'(s)ds\right)\frac{u_{Nx}^{k-1}u_N^k}{(u_N^{k-1})^3}, -u_{Nxx}^k\right)\\
=\ & \left(\left(\frac{f'(u_N^{k-1})}{u_N^{k-1}} - \frac{f'(\theta u_N^{k-1})}{u_N^{k-1}}\right)u_{Nx}^{k-1}u_N^k, -u_{Nxx}^k\right)\\
\le\ & 2A\|u_N^{k-1}\|_\infty\|u_N^k\|_\infty\|u_{Nx}^{k-1}\|\|u_{Nxx}^k\|\\
\le\ & \frac{\alpha}{8}\|u_{Nxx}^k\|^2 + c\|u_N^k\|^2(\|u_N^{k-1}\|^{\frac{4}{3}} + 1)(\|u_{Nx}^{k-1}\|^4 + \|u_N^{k-1}\|^4).
\end{aligned}$$

Therefore

$$B(u_N^{k-1}, u_N^k, -u_{Nxx}^k) \le \frac{\alpha}{4}\|u_{Nxx}^k\|^2 + c\|u_N^k\|^2(\|u_N^{k-1}\|^4 + 1)\,(\|u_{Nx}^{k-1}\|^4 + \|u_N^{k-1}\|^4).$$

Next, from the definition of $G(u)$ we have

$$\begin{aligned}
(G(u_N^{k-1})u_N^k, -u_{Nxx}^k) \le\ & \frac{5B}{4}(\|u_N^{k-1}\|_\infty^4 + 1)\|u_N^k\|\|u_{Nxx}^k\|\\
\le\ & \frac{\alpha}{8}\|u_{Nxx}^k\|^2 + c\|u_N^k\|^2(\|u_N^{k-1}\|^4\|u_{Nx}^{k-1}\|^4 + \|u_N^{k-1}\|^8 + 1).
\end{aligned}$$

In addition, we have

$$(h, -u_{Nxx}^k) \le \frac{\alpha}{8}\|u_{Nxx}^k\|^2 + \frac{4}{\alpha}\|h\|^2.$$

Hence (2.2) can be rewritten as

$$\begin{aligned}
&\overline{\partial}_t\|u_{Nx}^k\|^2 + \tau\|\overline{\partial}_t u_{Nx}^k\|^2 + \alpha\|u_{Nxx}^k\|^2\\
\le\ & c(\|u_N^{k-1}\|^4 + 1)\|u_N^k\|^2\,\|u_{Nx}^{k-1}\|^4 + c\big(\|u_N^k\|^2(\|u_N^{k-1}\|^8 + 1) + \|h\|^2\big).
\end{aligned} \tag{2.3}$$

By using (2.1), Lemma 2.3, and its corollary, we obtain for all $k_0 > n_0$

$$
\begin{aligned}
c\tau \sum_{k=k_0+1}^{k_0+N_0} (\|u_N^{k-1}\|^4+1)\|u_N^k\|^2\|u_{Nx}^{k-1}\|^2 \le\ & c\rho_0^2(1+\rho_0^4)\tau \sum_{k=k_0+1}^{k_0+N_0} \|u_{Nx}^{k-1}\|^2 \\
\le\ & c\rho_0^4(1+\rho_0^4)(1+r) \triangleq \alpha_1,
\end{aligned}
$$

$$
c\tau \sum_{k=k_0+1}^{k_0+N_0} \big(\|u_N^k\|^2(\|u_N^{k-1}\|^8+1)+\|h\|^2\big) \le c\big(\rho_0^2(\rho_0^8+1)+1\big)r \triangleq \alpha_2,
$$

$$
c\tau \sum_{k=k_0+1}^{k_0+N_0} \|u_{Nx}^k\|^2 \le c\Big(\rho_0^2+\frac{\|h\|^2}{|b|}N_0\tau\Big) \le c\rho_0^2(1+r) \triangleq \alpha_3.
$$

By applying the discrete uniform Gronwall's inequality to (2.3), we derive

$$
\|u_{Nx}^k\|^2 \le \Big(\frac{\alpha_3}{r}+\alpha_2\Big)\,\mathrm{e}^{\alpha_1} \triangleq \rho_1^2 \qquad \text{for all } n \ge n_1 = n_0+N_0.
$$

For $n \le n_1$, by applying the discrete Gronwall's inequality to (2.3), we have

$$
\begin{aligned}
\|u_{Nx}^k\|^2 \le\ & \Big(\|u_{0x}\|^2 + c\big(E_0^2(E_0^8+1)+\|h\|^2\big)t_n\Big)\exp\big(cE_0^2(E_0^4+1)(\|u_0\|^2+\rho_0^2 t_{n_1})\big) \\
\triangleq\ & (E_1')^2.
\end{aligned}
$$

Let $E_1^2 = \max\{(\rho_1')^2, (E_1')^2\}$. Then the second relation is obtained. Taking the sum of (2.3), we complete the proof of Lemma 2.5. □

COROLLARY 2.6. *Under the hypotheses of Lemma* 2.5*, we have*

$$
\begin{aligned}
&\|u_N^n\|_\infty \le c(\rho_0,\rho_1) \qquad \text{for all } n \ge n_1, \\
&\|u_N^n\|_\infty \le c(E_0,E_1) \qquad \text{for all } n \ge 1.
\end{aligned}
$$

LEMMA 2.7. *In addition to the conditions of Lemma* 2.5*, we suppose that* $f \in C^3$, $u_0 \in H_p^2(\Omega)$ *satisfying* $\|u_{0xx}\|^2 \le R_0^2$. *Then we have*

$$
\begin{aligned}
&\|u_{Nxx}^n\|^2 \le \rho_2^2 \qquad \text{for all } n \ge n_2 = n_1+N_0, \\
&\|u_{Nxx}^n\|^2 \le E_2^2 \qquad\quad \text{for all } n \ge 1, \\
&\tau^2 \sum_{k=1}^n \|\overline{\partial}_t u_{Nxx}^k\|^2 + \tau \sum_{k=1}^n \|u_{Nxxx}^k\|^2 \le c_2(1+t_n) \qquad \text{for all } n \ge 1,
\end{aligned}
$$

where n_1 *is given by Lemma* 2.5, N_0 *an arbitrary positive integer,* r *an arbitrary positive number such that* $N_0\tau = r$, *the constant* ρ_2 *independent of* N, n, τ, *and* $\|u_0\|_2$, $E_2 = E_2(\|u_0\|_2)$, *and* $C_2 = C_2(\|u_0\|_2)$ *independent of* N, n, *and* τ.

Proof. Let $\varphi = u_{Nx^4}^k$ in (1.4), and then we have

$$
\begin{aligned}
&\frac{1}{2}\overline{\partial}_t\|u_{Nxx}^k\|^2 + \frac{\tau}{2}\|\overline{\partial}_t u_{Nxx}^k\|^2 + \alpha\|u_{Nxxx}^k\|^2 + B(u_N^{k-1}, u_N^k, u_{Nx^4}^k) \\
=\ & (G(u_N^{K-1})u_N^k + h, u_{Nx^4}^k).
\end{aligned}
$$

When $\mathbf{k \geq n_1}$. From Lemma 2.3 to Corollary 2.6, we obtain

$$\begin{aligned}
B(u_N^{k-1}, u_N^k, u_{Nx^4}^k) =& \ (2F(u_N^{k-1})u_{Nx}^k + F'(u_N^{k-1})u_{Nx}^{k-1}u_N^k, u_{Nx^4}^k) \\
=& \ (2F(u_N^{k-1})u_{Nxx}^k + 3F'(u_N^{k-1})u_{Nx}^{k-1}u_{Nx}^k + F'(u_N^{k-1})u_{Nxx}^{k-1}u_N^k \\
& + F''(u_N^{k-1})(u_{Nx}^{k-1})^2 u_N^k, \ -u_{Nx^3}^k) \\
\leq& \ c(\rho_0,\rho_1)\|u_{Nxxx}^k\|^2(\|u_{Nxx}^k\| + \|u_{Nx}^{k-1}\|\|u_{Nx}^k\|_\infty + \|u_{Nxx}^k\| \\
& + \|u_N^k\|_\infty\|u_{Nx}^{k-1}\|_\infty\|u_{Nx}^{k-1}\|) \\
\leq& \ \frac{\alpha}{6}\|u_{Nxxx}^k\|^2 + c(\rho_0,\rho_1)(\|u_{Nx}^k\|^2 + \|u_N^k\|^2\|u_{Nx}^{k-1}\|^2) \\
& + c(\rho_0,\rho_1)(\|u_N^k\|\|u_{Nx}^k\| + \|u_N^k\|^2\|u_N^{k-1}\|^2)\|u_{Nxx}^{k-1}\|^2, \\
(G(u_N^{k-1})u_N^k, u_{Nx^4}^k) =& \ (G(u_N^{k-1})u_{Nx}^k, -u_{Nx^3}^k) + (G'(u_N^{k-1})u_{Nx}^{k-1}u_N^k, -u_{Nx^3}^k) \\
\leq& \ c(\rho_0,\rho_1)\|u_{Nxxx}^k\|(\|u_{Nx}^k\| + \|u_{Nx}^{k-1}\|) \\
\leq& \ \frac{\alpha}{6}\|u_{Nxxx}^k\|^2 + c(\rho_0,\rho_1)(\|u_{Nx}^k\|^2 + \|u_{Nx}^{k-1}\|^2),
\end{aligned}$$

and

$$(h, u_{Nx^4}^k) = (h_x, -u_{Nx^3}^k) \leq \frac{\alpha}{6}\|u_{Nxxx}^k\|^2 + c\|h_x\|^2.$$

Thus

$$\begin{aligned}
& \overline{\partial}_t\|u_{Nxx}^k\|^2 + \tau\|\overline{\partial}_t u_{Nxx}^k\|^2 + \alpha\|u_{Nxxx}^k\|^2 \\
\leq& \ c(\rho_0,\rho_1)(\|u_{Nx}^{k-1}\|^2 + \|u_{Nx}^k\|^2 + \|h_x\|^2).
\end{aligned} \tag{2.4}$$

By using Lemmas 2.3 and 2.5 and (2.3), we derive

$$\tau\sum_{k=k_0+1}^{k_0+N_0}\|u_{Nxx}\|^2 \leq \frac{1}{\alpha}\big(\rho_1^2 + rc(\rho_0,\rho_1)\big) \triangleq \alpha_3,$$

$$c(\rho_0,\rho_1)\tau\sum_{k=k_0+1}^{k_0+N_0}(\|u_{Nx}^k\|^2 + \|u_{Nx}^{k-1}\|^2 + \|h_x\|^2) \leq c(\rho_0,\rho_1)r \triangleq \alpha_2.$$

Then by applying discrete uniform Gronwall's inequality to (2.4), we have

$$\|u_{Nxx}^n\|^2 \leq \frac{\alpha_3}{r} + \alpha_2 \triangleq \rho_2^2 \quad \text{for all } n \geq n_2 = n_1 + N_0. \tag{2.5}$$

For $\mathbf{k \geq 1}$, as in the proof of the inequality (2.4), we have

$$\begin{aligned}
& \overline{\partial}_t\|u_{Nxx}^k\|^2 + \tau\|\overline{\partial}_t u_{Nxx}^k\|^2 + \alpha\|u_{Nxxx}^k\|^2 \\
\leq& \ c(E_0,E_1)(\|u_{Nx}^{k-1}\|^2 + \|u_{Nx}^k\|^2 + \|h_x\|^2).
\end{aligned} \tag{2.6}$$

Taking the sum of (2.6) for k from 1 to n, we obtain

$$\|u_{Nxx}^n\|^2 + \tau^2\|\overline{\partial}_t u_{Nxx}^n\|^2 + \alpha\tau\sum_{k=1}^n\|u_{Nxxx}^k\|^2 \leq C_2(t_n + 1) \quad \text{for all } n \geq 1. \tag{2.7}$$

Combining inequality (2.5) and (2.7), the proof of this lemma is complete. □

Proof of Theorem 2.2. Let $H = H_p^2(\Omega) \cap S_N$, $S_N^\tau(n)$ be a semigroup operator, i.e., the solution operator generated by problem (1.4), (1.5). On account of Theorem 2.1, we shall prove this theorem by checking the conditions (i) and (ii) in Theorem 2.1.

(i) By using the results of Lemmas 2.3–2.7, and assuming that $u_N^0 \in \mathcal{B} = \{u_N^0 \big|\ \|u_N^0\|_2 \le R_0\} \subset H_p^2(\Omega) \cap S_N$, we have

$$\|S_N^\tau(n)u_N^0\|_2 = \|u_N^n\|_2 \le (\rho_0^2 + \rho_1^2 + \rho_2^2)^{\frac{1}{2}} \qquad \text{for all } n \ge n_1(R).$$

Hence

$$\mathcal{B}_0 = \{u_N^n \in H_p^2(\Omega) \cap S_N :\ \|u_N^n\|_2 \le (\rho_0^2 + \rho_1^2 + \rho_2^2)^{\frac{1}{2}}\}$$

is a bounded absorbing set of the semigroup of operator $\{S_N^\tau(n)\}_{n\ge 0}$.

(ii) From Lemmas 2.3–2.7 and their corollaries, we have

$$\|S_N^\tau(n)u_N^0\|_2 \le (E_0^2 + E_1^2 + E_2^2)^{\frac{1}{2}} \qquad \text{for all } n \ge 0.$$

This means that $\{S_N^\tau(n)\}$ is uniformly bounded in $H_p^2(\Omega)\cap S_N$. Since a closed bounded set is a compact set in the finite dimensional space $H_p^2(\Omega) \cap S_N$, the operator $S_N^\tau(n)$ is uniformly compact for any $n \ge 0$.

On the other hand, it is easy to check the continuity of operator $S_N^\tau(n)$ from its boundedness. Thus the proof of the theorem is completed. □

3. Convergence of the global attractors $\mathcal{A}_N^\tau$. In this section, we study the convergence of attractors $\mathcal{A}_N^\tau$. To this end, we first give the following results.

Let $G_N :\ L_p^2(\Omega) \to S_N$ be the integral projection operator, i.e., for any given $u \in L^2(\Omega)$ we have

$$((G_N u)_x, v_x) + (G_N u, v) = (u, v) \quad \text{for all } \ v \in S_N. \tag{3.1}$$

Then for any $u,\ v \in L^2(\Omega)$, we have $(G_N u, v) = (u, G_N v)$.

LEMMA 3.1. *For the integral projection operator G_N, the following results hold:*

(A1) $\|G_N u\|_2 \sim \|P_N u\| \qquad$ for all $u \in L_p^2(\Omega)$;

(A2) $\|G_N u_x\| = \|(G_N u)_x\| \qquad$ for all $u \in H_p^1(\Omega)$,
$\|G_N u_{xx}\| = \|(G_N u_x)_x\| = \|(G_N u)_{xx}\|$ for all $u \in H_p^2(\Omega)$;

(A3) $\|G_N^2 u_{xx}\| = \|(G_N^2 u_x)_x\| = \|(G_N^2 u)_{xx}\|$ for all $u \in H_p^2(\Omega)$.

Proof. (A1). Setting $v = P_N u$ in (3.1) and applying the definition of P_N, it is obtained that

$$\|P_N u\| \le \|G_N u\| + \|(G_N u)_{xx}\|. \tag{3.2}$$

Setting $v = G_N u$ in (3.1), we have

$$\|G_N u\| \le \|P_N u\|. \tag{3.3}$$

Letting $v = (G_N u)_{xx}$ in (3.1), we find

$$\|(G_N u)_{xx}\| \le \|P_N u\|. \tag{3.4}$$

Combining (3.2), (3.3), and (3.4), the equality (A1) is obtained.

(A2). Using the definition of G_N repeatedly, the first equality of (A2) is obtained as follows:

$$\begin{aligned}\|G_N u_x\|^2 =& (G_N u_x, G_N u_x) = -(u, (G_N^2 u_x)x)\\ =& -((G_N u)_x, (G_N^2 u_x)_{xx}) - (G_N u, (G_N^2 u_x)_x)\\ =& ((G_N^2 u_x)_x, (G_N u)_{xx}) + (G_N^2 u_x, (G_N u)_x) = (G_N u_x, (G_N u)_x)\\ =& ((G_N(G_N u)_x, (G_N u)_{xx}) + (G_N(G_N u)_x, (G_N u)_x) = ((G_N u)_x, (G_N u)_x)\\ =& \|(G_N u)_x\|^2.\end{aligned}$$

By using the above equality and the definition of G_N repeatedly, the second equality of (A2) is recovered as follows:

$$\begin{aligned}\|G_N u_{xx}\|^2 &= ((G_N u_x)_x, (G_N u_x)_x) = (u_x, G_N u_x) - (G_N u_x, G_N u_x)\\ &= -(u, (G_N u_x)_x) - \|G_N u_x\|^2\\ &= ((G_N u)_{xx}, (G_N u_x)_x) + ((G_N u)_x, (G_N u_x)) - \|G_N u_x\|^2\\ &= -(u, (G_N u)_{xx}) - \|G_N u_x\|^2\\ &= \|(G_N u)_{xx}\|^2 + \|(G_N u)_x\|^2 - \|G_N u_x\|^2 = \|(G_N u)_{xx}\|^2.\end{aligned}$$

(A3). By using the definition of G_N and results of (A2), the equalities of (A3) are proved as follows:

$$\begin{aligned}\|G_N^2 u_{xx}\|^2 =& -((G_N^4 u_{xx})_x, u_x) = ((G_N u_x)_{xx}, (G_N^4 u_{xx})_x) + ((G_N u_x)_x, G_N^4 u_{xx})\\ =& ((G_N u_x)_x, G_N^3 u_{xx}) = -((G_N^3(G_N u_x)_x)_x, u_x)\\ =& -((G_N u_x)_x, (G_N^3(G_N u_x)_x)_{xx}) - (G_N u_X, (G_N^3(G_N u_x)_x)_x)\\ =& \|G_N(G_N u_x)_x\|^2 = \|(G_N^2 u_x)_x\|^2\end{aligned}$$

and

$$\begin{aligned}\|G_N^2 u_{xx}\|^2 &= -((G_N^4 u_{xx})_x, u_x) = (G_N^4 u_{xx}, u) - (G_N^3 u_{xx}, u)\\ &= -(u_x, (G_N^4 u)_x) + (u_x, (G_N^3 u)_x)\\ &= -(u, G_N^3 u) + (u, G_N^4 u) + (u, G_N^2 u) - (u, G_N^3 u)\\ &= -(G_N u, G_N^2 u) + (G_N^2 u, G_N^2 u) + (u, G_N^2 u) - (G_N u, G_N^2 u)\\ &= -((G_N^2 u)_x, (G_N^2 u)_x) + ((G_N u)_x, (G_N^2 u)_x)\\ &= ((G_N^2 u)_x, (G_N^2 u)_{xx}) - (G_N u, (G_N u, (G_N^2 u)_x)\\ &= -((G_N^2 u)_x, (G_N^2 u)_{xxx}) = \|(G_N^2 u)_{xx}\|^2. \qquad \square\end{aligned}$$

THEOREM 3.2 (see [20]). *If $f \in C^3$, $|f'(u)| \le A|u|^2$; $g \in C^1$, $g(0) = 0$, $g'(s) \le b < 0$; $h(x) \in H_p^1(\Omega)$; $u_0 \in H_p^2(\Omega)$, then there exists a unique global solution $u(x,t) \in L^\infty(R^+;\ H_p^2(\Omega))$ for the problem (1.1)–(1.3) such that*

$$\int_0^t (\|u\|_3^2 + \|u_t\|^2)\,\mathrm{d}t \le c(t+1) \qquad \text{for all } t \in R^+,$$

where the constant c is independent of t.

Furthermore, if $f \in C^3$, $g \in C^2$, then $u(x,t)$ satisfies

$$t\|u_{xxx}\|^2 \le c(t^2+1) \qquad \text{for all } t \in R^+,$$

and there exists a global attractor $\mathcal{A} \subset H_p^2(\Omega)$ of problem (1.1)–(1.3), i.e., there exist a set $\mathcal{A} \subset H_p^2(\Omega)$ such that

(a) $S(t)\mathcal{A} = \mathcal{A}$,

(b) $\operatorname{dist}(S(t)\mathcal{B},\ \mathcal{A}) \to 0$, *as* $t \to +\infty$.

THEOREM 3.3 (see [1]). *Suppose that*

(1) $\{H_\eta\}_{0<\eta\le\eta_0}$ *is a family of closed subspaces of Banach space* H *such that*

$$\bigcup_{0<\eta\le\eta_0} H_\eta$$

is dense in H.

(2) $S_\eta(t)$: $H_\eta \to H_\eta$ *and* $S(t)$: $H \to H$ *are two nonlinear semigroup operators,* $\mathcal{A}_\eta \subset H_\eta$ *and* $\mathcal{A} \subset H$ *are the global attractors of* $S_\eta(t)$ *and* $S(t)$, *respectively.*

(3) *For every compact interval* $I \subset (0, +\infty)$,

$$\delta_\eta(I) = \sup_{u_0\in H_\eta} \sup_{t\in I} \operatorname{dist}\left(S_\eta(t)u_0\,,\, S(t)u_0\right) \to 0 \quad as \quad \eta \to 0.$$

Then $\mathcal{A}_\eta$ *is convergent to* $\mathcal{A}$ *in the sense of the semidistance:*

$$\operatorname{dist}(\mathcal{A}_\eta, \mathcal{A}) \to 0 \quad \text{as} \quad \eta \to 0,$$

where

$$\operatorname{dist}(\mathcal{A}_\eta, \mathcal{A}) = \sup_{u\in\mathcal{A}_\eta} \inf_{v\in\mathcal{A}} \|u - v\|_H.$$

Finally, similar to Lemma 2.7, the following result can be proved easily.

LEMMA 3.4. *Under the hypotheses of Lemma* 2.7, *we have the estimates for the smooth solution* $u(x,t)$ *of problem* (1.1)–(1.3)

$$t\|u_t\|^2 + \int_0^t s\|u_{tx}\|^2 ds \le c(1+t^2),$$

$$t^2\|u_{tx}\|^2 + \int_0^t s^2\|u_{txx}\|^2 ds \le c(1+t^3),$$

$$t^3\|u_{txx}\|^2 + \int_0^t s^3(\|u_{tt}\|^2 + \|u_{txxx}\|^2) ds \le c(1+t^4),$$

$$t^4\|u_{tt}\|^2 + \int_0^t s^4\|u_{ttx}\|^2 ds \le c(1+t^5),$$

where the constant c *is independent of* t.

Now we give the main result of this section.

THEOREM 3.5. *Suppose that the conditions of Theorem* 2.2 *hold. Then*

$$\operatorname{dist}(\mathcal{A}_N^\tau, \mathcal{A}) \to 0 \qquad \text{as } \tau \to 0,\ N \to +\infty.$$

Proof. Let $\|u_0\|_2 \le R_0$. On account of Theorem 3.2, this theorem will be proved by taking the error estimates of the solution u_N^n of discrete problem (1.4), (1.5). Now we accomplish them through two steps.

Step 1. Take the error estimates of the solution v_N^n of the linear scheme as follows:

$$(\overline{\partial}_t v_N^k - \alpha v_{Nxx}^k + \beta v_{Nxxx}^k + \alpha v_N^k + f(u^k)_x, \varphi) = (g(u^k) + \alpha u^k + h, \varphi) \quad \text{for all } \varphi \in S_N,\ \ k = 1, 2, \cdots, \tag{3.5}$$

$$v_N^0 = P_N u_0. \tag{3.6}$$

Let $u^k - v_N^k = (u^k - P_N u^k) + (P_N u^k - v_N^k) = \rho^k + \theta^k$. Then θ^k satisfies

$$(\overline{\partial}_t \theta^k - \alpha\theta_{xx}^k + \beta\theta_{xxx}^k + \alpha\theta^k - (\overline{\partial}_t u^k - u_t^k),\ \ \varphi) = 0 \quad \text{for all } \varphi \in S_N, \tag{3.7}$$

$$\theta^0 = 0. \tag{3.8}$$

Letting $\varphi = \theta^k$ in (3.7), we have

$$\frac{1}{2}\overline{\partial}_t\|\theta^k\|^2 + \frac{\tau}{2}\|\overline{\partial}_t\theta^k\|^2 + \alpha\|\theta^k\|_1^2 = (\overline{\partial}_t\theta^k - u_t^k, \theta^k).$$

From the definition of G_N, we find

$$\begin{aligned}(\overline{\partial}_t u^k - u_t^k, \theta^k) &= ((G_N(\overline{\partial}_t u^k - u_t^k))_x, \theta_x^k) + (G_N(\overline{\partial}_t u^k - u_t^k), \theta^k)\\ &\le \frac{\alpha}{2}\|\theta^k\|_1^2 + \frac{1}{2\alpha}\|G_N(\overline{\partial}_t u^k - u_t^k)\|_1^2,\end{aligned}$$

where

$$\begin{aligned}\|G_N(\overline{\partial}_t u^k - u_t^k)\|_1^2 &= \frac{1}{\tau^2}\Big\|\int_{t_{k-1}}^{t_k}(s - t_{k-1})G_N u_{tt} ds\Big\|_1^2\\ &\le \frac{1}{\tau^2}\int_{t_{k-1}}^{t_k}\frac{(s-t_{k-1})^2}{s^2}\mathrm{d}s\int_{t_{k-1}}^{t_k} s^2\|G_N u_{tt}\|_1^2\mathrm{d}s\\ &\le \frac{\tau}{t_k^2}\int_{t_{k-1}}^{t_k} s^2\|G_N u_{tt}\|_1^2\mathrm{d}s.\end{aligned}$$

Thus

$$\overline{\partial}_t\|\theta^k\|^2 + \tau\|\overline{\partial}_t\theta^k\|^2 + \alpha\|\theta^k\|^2 \le \frac{1}{\alpha}\|G_N(\overline{\partial}_t u^k - u_t^k)\|_1^2 \le \frac{\tau}{\alpha t_k^2}\int_{t_{k-1}}^{t_k} s^2\|G_N u_{tt}\|_1^2\mathrm{d}s.$$

Multiplying the above inequality by t_k^2, taking the sum for k from 1 to n, and using $\|G_N u_{tt}\|_1 \le c\|u_t\|_2$, we have

$$\begin{aligned}&t_n^2\|\theta^n\|^2 + \alpha\tau\sum_{k=1}^n t_k^2\|\theta^k\|_1^2 + \tau^2\sum_{k=1}^n t_k^2\|\overline{\partial}_t\theta^k\|^2\\ \le\ & 3\tau\sum_{k=1}^n t_k\|\theta^k\|^2 + \frac{\tau^2}{\alpha}\sum_{k=1}^n\int_{t_{k-1}}^{t_k} s^2\|G_N u_{tt}\|_1^2\mathrm{d}s\\ \le\ & 3\tau\sum_{k=1}^n t_k\|\theta^k\|^2 + c\tau^2\int_0^{t_n} s^2\|u_t\|_2^2\mathrm{d}s\\ \le\ & 3\tau\sum_{k=1}^n t_k\|\theta^k\|^2 + c\tau^2(1 + t_n^3).\end{aligned} \tag{3.9}$$

Now we estimate $\tau\sum_{k=1}^n t_k\|\theta^k\|^2$ in the above inequality. Let $\varphi = G_N\theta^k$ in (3.7). Then it is obtained that

$$(\overline{\partial}_t\theta^k - \alpha\theta_{xx}^k + \beta\theta_{xxx}^k + \alpha\theta^k - (\overline{\partial}_t u^k - u_t^k), G_N\theta^k) = 0.$$

From the definition of G_N, we find

$$(\overline{\partial}_t\theta^k, G_N\theta^k) = (\overline{\partial}_t(G_N\theta^k)_x, (G_N\theta^k)_x) + (\overline{\partial}_t(G_N\theta^k), G_N\theta^k) = \frac{1}{2}\overline{\partial}_t\|G_N\theta^k\|_1^2,$$

$$(-\alpha\theta^k_{xx} + \alpha\theta^k, G_N\theta^k) = \alpha((G_N\theta^k)_x, \theta^k_x) + \alpha(G_N\theta^k, \theta^k) = \alpha(\theta^k,\ \theta^k) = \alpha\|\theta^k\|^2,$$

$$\begin{aligned}\beta(\theta^k_{xxx}, G_N\theta^k) &= -\beta(\theta^k,\ (G_N\theta^k)_{xxx})\\ &= -\beta((G_N\theta^k)_x,\ (G_N\theta^k)_{x^4}) - \beta(G_N\theta^k, (G_N\theta^k)_{xxx})\\ &= 0\end{aligned}$$

and

$$(\overline{\partial}_t u^k - u^k_t, G_N\theta^k) = (G_N(\overline{\partial}_t u^k - u^k_t), \theta^k) \le \frac{\alpha}{2}\|\theta^k\|^2 + \frac{2}{\alpha}\|G_N(\overline{\partial}_t u^k - u^k_t)\|^2.$$

Therefore

$$\overline{\partial}_t\|G_N\theta^k\|_1^2 + \alpha\|\theta^k\|^2 \le \frac{1}{\alpha}\|G_N(\overline{\partial}_t u^k - u^k_t)\|^2.$$

Multiplying this inequality by τt_k, summing them for k from 1 to n, and using $\|G_N u_{tt}\| \le c\|u_t\|_1$, we have

$$\begin{aligned}& t_n\|G_N\theta^k\|_1^2 + \alpha\tau\sum_{k=1}^n t_k\|\theta^k\|^2\\ \le\ & \frac{\tau^2}{\alpha}\sum_{k=1}^n t_k\|G_N(\overline{\partial}_t u^k - u^k_t)\|^2 + \tau\sum_{k=1}^n\|G_N\theta^k\|_1^2\\ \le\ & \frac{\tau}{\alpha}\int_0^{t_n} s\|G_N u_{tt}\|^2\mathrm{d}s + \tau\sum_{k=1}^n\|G_N\theta^k\|_1^2\\ \le\ & c\tau^2(1+t_n^2) + \tau\sum_{k=1}^n\|G_N\theta^k\|_1^2.\end{aligned} \tag{3.10}$$

To estimate $\tau\sum_{k=1}^n\|G_N\theta^k\|_1^2$ in the above relation, let $\varphi = G_N^2\theta^k$ in (3.7). Then we have

$$(\overline{\partial}_t\theta^k - \alpha\theta^k_{xx} + \beta\theta^k_{xxx} + \alpha\theta^k - (\overline{\partial}_t u^k - u^k_t), G_N^2\theta^k) = 0.$$

From the definition of G_N, we obtain

$$(\overline{\partial}_t\theta^k, G_N^2\theta^k) = \frac{1}{2}\overline{\partial}_t\|G_N\theta^k\|^2 + \frac{\tau}{2}\|G_N(\overline{\partial}_t\theta^k)\|^2,$$

$$(-\alpha\theta^k_{xx} + \alpha\theta^k, G_N^2\theta^k) = \alpha\|G_N\theta^k\|_1^2,$$

$$\beta(\theta^k_{xxx}, G_N^2\theta^k) = 0$$

and

$$\begin{aligned}(\overline{\partial}_t u^k - u^k_t, G_N^2\theta^k) =\ & ((G_N^2(\overline{\partial}_t u^k - u^k_t))_x, (G_N\theta^k)_x) + (G_N^2(\overline{\partial}_t u^k - u^k_t), G_N\theta^k)\\ \le\ & \frac{\alpha}{2}\|G_N\theta^k\|_1^2 + \frac{1}{2\alpha}\|G_N^2(\overline{\partial}_t u^k - u^k_t)\|_1^2.\end{aligned}$$

Therefore

$$\overline{\partial}_t\|G_N\theta^k\|^2 + \alpha\|G_N\theta^k\|_1^2 \le \frac{1}{\alpha}\|G_N^2(\overline{\partial}_t u^k - u^k_t)\|_1^2.$$

Taking the sum of the above relation for k from 1 to n, and applying $\|G_N^2 u_{tt}\|_1 \le c\|u_t\|$, it is derived that

$$\begin{aligned}\|G_N\theta^k\|^2 + \alpha\tau\sum_{k=1}^{n}\|G_N\theta^k\|_1^2 &\le \frac{\tau}{\alpha}\sum_{k=1}^{n}\|G_N^2(\overline{\partial}_t u^k - u_t^k)\|_1^2 \\ &\le \frac{\tau^2}{\alpha}\int_0^{t_n}\|G_N^2 u_{tt}\|_1^2 \mathrm{d}s \le c\tau^2(1+t_n).\end{aligned} \tag{3.11}$$

Combining (3.9)–(3.11), we have

$$\begin{aligned}&t_n^2\|\theta^n\|^2 + \tau^2\sum_{k=1}^{n}t_k^2\|\overline{\partial}_t\theta^k\|^2 + \alpha\tau\sum_{k=1}^{n}t_k^2\|\theta^k\|_1^2 + \alpha\tau\sum_{k=1}^{n}t_k\|\theta^k\|^2 \\ &\quad\le c\tau^2(1+t_n^3).\end{aligned} \tag{3.12}$$

Let $\varphi = -\theta_{xx}^k$ in (3.7), and then we have

$$\begin{aligned}\frac{1}{2}\overline{\partial}_t\|\theta_x^k\|^2 + \frac{\tau}{2}\|\overline{\partial}_t\theta_x^k\|^2 + \alpha\|\theta_{xx}^k\|^2 + \alpha\|\theta_x^k\|^2 &= (\overline{\partial}_t u^k - u_t^k, -\theta_{xx}) \\ &\le \frac{\alpha}{2}\|\theta_{xx}^k\|^2 + \frac{1}{2\alpha}\|\overline{\partial}_t u^k - u_t^k\|^2,\end{aligned}$$

namely,

$$\overline{\partial}_t\|\theta_x^k\|^2 + \tau\|\overline{\partial}_t\theta_x^k\|^2 + \alpha\|\theta_{xx}^k\|^2 + \alpha\|\theta_x^k\|^2 \le \frac{1}{\alpha}\|\overline{\partial}_t u^k - u_t^k\|^2.$$

Multiplying this by τt_k^3, taking the sum for k from 1 to n, and using (3.12), we have

$$\begin{aligned}&t_n^3\|\theta_x^n\|^2 + \tau^2\sum_{k=1}^{n}t_k^3\|\overline{\partial}_t\theta_x^k\|^2 + \alpha\tau\sum_{k=1}^{n}t_k^3(\|\theta_{xx}^k\|^2 + \|\theta_x^k\|^2) \\ \le\ & \frac{\tau}{\alpha}\sum_{k=1}^{n}t_k^3\|\overline{\partial}_t u^k - u_t^k\|^2 + 7\tau\sum_{k=1}^{n}t_k^2\|\theta_x^k\|^2 \\ \le\ & \frac{\tau}{\alpha}\left(\sum_{k=2}^{n}\frac{t_k^3}{\tau^2}\int_{t_{k-1}}^{t_k}\frac{(s-t_{k-1})^2}{s^3}\mathrm{d}s\int_{t_{k-1}}^{t_k}s^3\|u_{tt}\|^2\mathrm{d}s + \tau^3\|\overline{\partial}_t u^1 - u_t^1\|^2\right) \\ &\qquad + c\tau^2(1+t_n^3) \\ \le\ & \frac{2\tau^2}{\alpha}\int_{\tau}^{t_n}s^3\|u_{tt}\|^2\mathrm{d}s + \frac{2\tau^2}{\alpha}\Big\|\int_0^{\tau}u_t ds\Big\|^2 + \frac{2\tau^4}{\alpha}\|u_t^1\|^2 + c\tau^2(1+t_n^3) \\ \le\ & c\tau^2(1+t_n^4) + c\tau^3(1+\tau^2) + \frac{2\tau^2}{\alpha}\int_0^{\tau}\|u_t\|^2\mathrm{d}s \\ \le\ & c\tau^2(1+t_n^4).\end{aligned} \tag{3.13}$$

Multiplying (3.13) by $\dfrac{1}{\alpha\tau}$, we have

$$t_n^3\|\theta_{xx}^n\|^2 \le c\tau(1+t_n^4). \tag{3.14}$$

Step 2. Take the error estimates of solution u_N^n of problem (1.4), (1.5). Let $v_N^k - u_N^k = e^k$. Then e^k satisfies

$$\begin{aligned}&(\overline{\partial}_t e^k - \alpha e_{xx}^k + \beta e_{xxx}^k + f(u^k)_x, \varphi) - B(u_N^{k-1}, u_N^k, \varphi) \\ =\ & (\alpha\,\theta^k + g(u^k) - G(u_N^{k-1})u_N^k, \varphi) \quad \text{for all } \varphi \in S_N,\ \ k = 1, 2, \ldots,\end{aligned} \tag{3.15}$$

$$e^0 = 0. \tag{3.16}$$

Let $\varphi = e^k$ in (3.15), and then we have

$$\frac{1}{2}\overline{\partial}_t\|e^k\|^2+\frac{1}{2}\|\overline{\partial}_t e^k\|^2+\alpha\|e_x^k\|^2+B(u^k,u^k,e^k)-B(u_N^{k-1},u_N^k,e^k)$$
$$=(g(u^k)-G(u_N^{k-1})u_N^k,e^k)+\alpha(\theta^k,e^k).$$

We now estimate every term in the above equality. First,

$$\begin{aligned}&(g(u^k)-G(u_N^{k-1})u_N^k,e^k)\\ =\;&(g(u^k)-g(u_N^k)+G(u_N^k)u_N^k-G(u_N^{k-1})u_N^k,e^k),\end{aligned}$$

where

$$(g(u^k)-g(u_N^k),e^k)=(g'(\eta_1)(\rho^k+\theta^k+e^k),e^k)\le\frac{15b}{16}\|e^k\|^2+c(\|\rho^k\|^2+\|\theta^k\|^2)$$

and

$$(G(u_N^k)u_N^k-G(u_N^{k-1})u_N^k,e^k)\le\frac{|b|}{16}\|e^k\|^2+c\|\overline{\partial}_t u_N^k\|^2.$$

Thus

$$(g(u^k)-G(u_N^{k-1})u_N^k,e^k)\le\frac{7b}{8}\|e^k\|^2+c\,(\|\rho^k\|^2+\|\theta^k\|^2+\tau^2\|\overline{\partial}_t u_N^k\|^2).$$

Second,

$$\begin{aligned}&B(u^k,u^k,e^k)-B(u_N^{k-1},u_N^k,e^k)\\ =\;&(F(u^k)u_x^k-F(u_N^{k-1})u_{Nx}^k,e^k)-(F(u^k)u^k-F(u_N^{k-1})u_N^k,e_x^k),\end{aligned}$$

where

$$\begin{aligned}&(F(u^k)u_x^k-F(u_N^{k-1})u_{Nx}^k,e^k)\\ =\;&(F(u^k)(u_x^k-u_{Nx}^k),e^k)+(F(u^k)-F(u^{k-1}),u_{Nx}^k e^k)+(F(u^{k-1})-F(u_N^{k-1}),u_{Nx}^k e^k)\\ \triangleq\;&I_1+I_2+I_3,\\ &I_1=-(F(u^k)_x\,e^k+F(u^k)e_x^k,\,u^k-u_N^k)\le c\|u^k-u_N^k\|\,(\|e^k\|+\|e_x^k\|)\\ &\le c(\|e^k\|+\|e_x^k\|)\,(\|\rho^{k-1}\|+\|\theta^{k-1}\|+\|e^{k-1}\|+\tau\|\overline{\partial}_t u^k\|+\tau\|\overline{\partial}_t u_N^k\|)\\ &\le c\left(\|e^{k-1}\|^2+\|\rho^{k-1}\|^2+\|\theta^{k-1}\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\mathrm{d}s+\tau^2\|\overline{\partial}_t u_N^k\|^2\right)\\ &\qquad+\Big(\frac{|b|}{16}\|e^k\|^2+\frac{\alpha}{16}\|e_x^k\|^2\Big),\\ &I_2=(F'(\eta)u_{Nx}^k(u^k-u^{k-1}),e^k)\le\frac{|b|}{16}\|e^k\|^2+c\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s,\\ &I_3=\;(F'(\eta)u_{Nx}^k(u^{k-1}-u_N^{k-1}),e^k)\\ &\le\;\frac{|b|}{16}\|e^k\|^2+c\big(\|\rho^{k-1}\|^2+\|e^{k-1}\|^2+\|\theta^{k-1}\|^2\big).\end{aligned}$$

Thus

$$\begin{aligned}(F(u^k)u_x^k-F(u_N^{k-1})u_{Nx}^k,e^k)\le\;&\frac{3|b|}{16}\|\mathrm{e}^k\|^2+\frac{\alpha}{16}\|\mathrm{e}_x^k\|^2\\ &+c\left(\|\mathrm{e}^{k-1}\|^2+\|\rho^{k-1}\|^2+\|\theta^{k-1}\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\mathrm{d}s+\tau^2\|\overline{\partial}_t u_N^k\|^2\right).\end{aligned}$$

Similarly

$$\begin{aligned}
&(F(u^k)u^k - F(u_N^{k-1})u_N^k, e_x^k)\\
=\ &(u^k(F(u^k) - F(u^{k-1}) + (u^k(F(u^{k-1}) - F(u_N^{k-1})) + F(u^k)(u^k - u_N^k), e_x^k)\\
\le\ &\frac{\alpha}{16}\|e_x^k\|^2 + c\left(\|e^{k-1}\|^2 + \|\rho^{k-1}\|^2 + \|\theta^{k-1}\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|^2 + \tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s\right).
\end{aligned}$$

Thus

$$\begin{aligned}
B(u^k, u^k, e^k) - B(u_N^{k-1}, u_N^k, e^k) \le\ &\frac{\alpha}{8}\|e_x^k\|^2 + \frac{3|b|}{16}\|e^k\|^2\\
&+ c\left(\|e^{k-1}\|^2 + \|\rho^{k-1}\|^2 + \|\theta^{k-1}\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|^2 + \tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}ds\right).
\end{aligned}$$

Finally

$$\alpha(\theta^k, e^k) \le \frac{|b|}{16}\|e^k\|^2 + c\|\theta^k\|^2.$$

Therefore

$$\begin{aligned}
\overline{\partial}_t\|e^k\|^2 + \alpha\|e_x^k\|^2 - b\|e^k\|^2 + \tau\|\overline{\partial}_t e^k\|^2 \le\ &c(\|e^{k-1}\|^2 + \|\rho^k\|^2 + \|\theta^k\|^2)\\
&+ c\left(\|\rho^{k-1}\|^2 + \|\theta^{k-1}\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|^2 + \tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\mathrm{d}s\right).
\end{aligned} \tag{3.17}$$

By applying the discrete Gronwall's inequality to (3.17), and using Lemma 1.1, Lemma 2.3, Theorem 3.2, (3.10), and (3.11), we have

$$\begin{aligned}
\|e^n\|^2 &\le\ c\,\mathrm{e}^{ct_n}\tau\sum_{k=1}^{n}\left(\|\rho^k\|^2 + \|\theta^k\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|^2 + \tau\int_{t_{k-1}}^{t_n}\|u_t\|^2\mathrm{d}s\right)\\
&\le\ c(1 + t_n^2)\mathrm{e}^{ct_n}(N^{-4} + \tau).
\end{aligned} \tag{3.18}$$

Taking the sum of (3.17) for k from 1 to n and using (3.18) and (3.9)–(3.11), we have

$$\tau\sum_{k=1}^{n}\|e^k\|_1^2 + \tau^2\sum_{k=1}^{n}\|\overline{\partial}_t e^k\|^2 \le c(1 + t_n^3)\,\mathrm{e}^{ct_n}(N^{-4} + \tau). \tag{3.19}$$

Let $\varphi = -e_{xx}^k$ in (3.15), and then similarly as in (3.17) we have

$$\overline{\partial}_t\|e_x^k\|^2 + \alpha\|e_{xx}^k\|^2 \le\ c\big(\|\rho^k\|_1^2 + \|\theta^k\|_1^2 + \tau^2\|\overline{\partial}_t u_N^k\|_1^2 + \|e^k\|^2\big).$$

Multiplying the above relation by τt_k, summing them for k from 1 to n, and using Lemma 2.5, (3.18), (3.19), (3.9)–(3.11), we have

$$\begin{aligned}
&t_n\|e_x^k\|^2 + \alpha\tau\sum_{k=1}^{n}t_k\|e_{xx}^k\|^2\\
\le\ &\tau\sum_{k=1}^{n}\|e_x^k\|^2 + \tau\sum_{k=1}^{n}t_k(\|e^k\|^2 + \|\theta^k\|_1^2 + \|\rho^k\|_1^2 + \tau^2\|\overline{\partial}_t u_N^k\|_1^2)\\
\le\ &c(1 + t_n^4)\mathrm{e}^{ct_n}(N^{-2} + \tau).
\end{aligned} \tag{3.20}$$

Letting $\varphi = e_{x^4}$ in (3.15), we have

$$\begin{aligned}(3.21)\qquad &\frac{1}{2}\overline{\partial}_t\|e^k_{xx}\|^2+\frac{\tau}{2}\|\overline{\partial}_t e^k_{xx}\|^2+\alpha\|e^k_{xxx}\|^2+B(u^k,u^k,e^k_{x^4})-B(u^{k-1}_N,u^k_N,e^k_{x^4})\\ &\quad=(g(u^k)-G(u^{k-1}_N)u^k_N,e^k_{x^4})+\alpha(\theta^k,e^k_{x^4}).\end{aligned}$$

Now we estimate the last three terms in (3.21). First, from the definition of $B(u,v,\varphi)$, we have

$$\begin{aligned}&B(u^k,u^k,e^k_{x^4})-B(u^{k-1}_N,u^k_N,e^k_{x^4})\\ =\;&2\Big(F(u^k)u^k_{xx}-F(u^{k-1}_N)u^k_{Nxx},-e^k_{xxx}\Big)\\ &+3\Big(F'(u^k)(u^k_x)^2-F'(u^{k-1}_N)u^{k-1}_{Nx}u^k_{Nx},-e^k_{xxx}\Big)\\ &+\Big(F'(u^k)u^ku^k_{xx}-F'(u^{k-1}_N)u^k_Nu^{k-1}_{Nxx},-e^k_{xxx}\Big)\\ &+\Big(F''(u^k)u^k(u^k_x)^2-F''(u^{k-1}_N)u^k_N(u^{k-1}_{Nx})^2,-e^k_{xxx}\Big)\\ =\;&I_7+I_8+I_9+I_{10},\end{aligned}$$

where

$$\begin{aligned}I_7&\le\frac{\alpha}{16}\|e^k_{xxx}\|^2+c(\|\rho^k\|_2^2+\|\theta^k\|_2^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|^2),\\ I_8&\le\frac{\alpha}{16}\|e^k_{xxx}\|^2+c\Big(\|\rho^k\|_1^2+\|\theta^k_x\|^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_1^2\Big),\\ I_9&\le\frac{\alpha}{16}\|e^k_{xxx}\|^2+c\big(\|\rho^k\|_2^2+\|\theta^k\|_2^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_2^2\big)\end{aligned}$$

and

$$I_{10}\le\frac{\alpha}{16}\|e^k_{xxx}\|^2+\big(\|\rho^k\|_1^2+\|\theta^k\|_1^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_1^2\big).$$

Thus

$$\begin{aligned}(3.22)\qquad &B(u^k,u^k,e^k_{x^4})-B(u^{k-1}_n,u^k_N,e^k_{x^4})\\ \le\;&\frac{\alpha}{4}\|e^k_{xxx}\|^2+c\big(\|\rho^k\|_2^2+\|\theta^k\|_2^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_2^2\big).\end{aligned}$$

Second, by the definition of $G(u)$, we have

$$\begin{aligned}(3.23)\qquad &\Big(g(u^k)-G(u^{k-1}_N),e^k_{x^4}\Big)\\ =\;&\Big(g'(u^k)u^k_x-G'(u^{k-1}_N)u^{k-1}_{Nx}u^k_N-G(u^{k-1}_N)u^k_{Nx},-e^k_{xxx}\Big)\\ \le\;&\frac{1}{8}\|e^k_{xxx}\|^2+c\Big(\|\rho^k\|_1^2+\|\theta^k\|_1^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_1^2\Big).\end{aligned}$$

Finally

$$(3.24)\qquad \alpha(\theta^k,e^k_{x^4})\le\frac{\alpha}{8}\|e^k_{xxx}\|^2+c\|\theta^k_x\|^2.$$

Combining (3.21)–(3.24), we have

$$(3.25)\qquad \overline{\partial}_t\|e^k_{xx}\|^2+\alpha\|e^k_{xxx}\|^2\le\;c(\|\rho^k\|_2^2+\|\theta^k\|_2^2+\|e^k\|^2+\tau^2\|\overline{\partial}_t u^k_N\|_1^2).$$

From Lemma 1.1, we have

$$\|\rho^k\|_2^2 \le cN^{-2}\|u_{xxx}^k\|^2.$$

From Theorem 3.2, we have

$$t_k\|u_{xxx}^k\|^2 \le c(1+t_k^2).$$

Hence multiplying (3.25) by τt_k^2, summing them and using (3.9)–(3.14), (3.19), and (3.20), we have

$$\begin{aligned} & t_n^2\|e_{xx}^n\|^2 + \alpha\tau\sum_{k=1}^n t_k^2\|e_{xxx}^k\|^2 \\ \le\ & c\tau\sum_{k=1}^n t_k^2(\|u_{xxx}^k\|^2N^{-2} + \|\theta^k\|_2^2 + \|e^k\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|_1^2) + 2\tau\sum_{k=1}^n t_k\|e_{xx}^k\|^2 \\ \le\ & c(1+t_n^5)\mathrm{e}^{ct_n}(N^{-2}+\tau). \end{aligned}$$

By using the triangle inequality, we have

$$\begin{aligned} \|u^n - u_N^n\|_2^2 &\le 3(\|\rho^n\|_1^2 + \|\theta^n\|_2^2 + \|e^k\|_2^2) \\ &\le c\mathrm{e}^{ct_n}(1+t_n^{-3}+t_n^3)(N^{-2}+\tau) \qquad \text{for all } t\in(0,+\infty). \end{aligned}$$

Finally, applying Theorem 3.3, we complete the proof of this theorem. ∎

4. Stability of the discrete autonomous system. In this section, the stability of the discrete scheme (1.4), (1.5) is proved.

THEOREM 4.1. *Suppose that f, g, h satisfy the conditions of Lemma* 2.5. *Let $\{u_N^n\}$, $\{v_N^n\}$ be the two solutions of the discrete scheme* (1.4), (1.5) *with the initial values $\{u_N^0\}$, $\{v_N^0\}$, respectively, and the initial values satisfy $\|u_N^0\|_1 \le R_0$, $\|v_N^0\|_1 \le R_0$. Then we have*

$$\|u_N^n - v_N^n\|_1 \le c\mathrm{e}^{ct_n}\|u_N^0 - v_N^0\|_1 \qquad \text{for all } n\ge 1.$$

Furthermore, if f, g, h satisfy the conditions of Lemma 2.7, *and $\|u_N^0\|_2 \le R_0$, $\|v_N^0\|_2 \le R_0$, then we have*

$$\|u_N^n - v_N^n\|_2 \le c\mathrm{e}^{ct_n}\|u_N^0 - v_N^0\|_2 \qquad \text{for all } n\ge 1.$$

Proof. Let $w_N^k = u_N^k - v_N^k$, and then w_N^k satisfies

$$\begin{aligned} &(\overline{\partial}_t w_N^k - \alpha w_{Nxx}^k + \beta w_{Nxxx}^k, \varphi) + B(u_N^{k-1}, u_N^k, \varphi) - B(v_N^{k-1}, v_N^k, \varphi) \\ &\qquad = (G(u_N^{k-1})u_N^k - G(v_N^{k-1})v_N^k, \varphi), \qquad \text{for all } \varphi\in S_N, k=1,2,\ldots. \end{aligned} \tag{4.1}$$

First, we assume that the conditions of Lemma 2.5 are satisfied.

Letting $\varphi = w_N^k$ in (4.1), we have

$$\begin{aligned} \frac{1}{2}\overline{\partial}_t\|w_N^k\|^2 + \alpha\|w_{Nx}^k\|^2 \le &-\big(B(u_N^{k-1}, u_N^k, w_N^k) - B(v_N^{k-1}, v_N^k, w_N^k)\big) \\ &+(G(u_N^{k-1})u_N^k - G(v_N^{k-1})v_N^k, w_N^k). \end{aligned} \tag{4.2}$$

Now we estimate the last two terms. First, by applying the definition of $B(u,v,w)$ and $G(u)$, Sobolev interpolation inequality, Young's inequality, and the estimates in

section 2, we obtain

$$\begin{aligned}
&\big|B(u_N^{k-1},u_N^k,w_N^k)-B(v_N^{k-1},v_N^k,w_N^k)\big| = \big|B(u_N^{k-1},u_N^k,w_N^k)-B(v_N^{k-1},u_N^k,w_N^k)\big|\\
&= \big|\big(F(u_N^{k-1})-F(v_N^{k-1})\big)u_{Nx}^k,w_N^k)-((F(u_N^{k-1})-F(v_N^{k-1}))u_N^k,w_{Nx}^k\big)\big|\\
&\le \frac{|b|}{4}\|w_N^k\|^2+\frac{\alpha}{2}\|w_{Nx}^k\|^2+c\|v_N^k\|_1^2\,\|w_N^{k-1}\|^2
\end{aligned}$$

and

$$\begin{aligned}
&(G(u_N^{k-1})u_N^k-G(v_N^{k-1})v_N^k,w_N^k)\\
=\;&\big((G(u_N^{k-1})-G(v_N^{k-1})u_N^k,w_N^k)+(G(v_N^{k-1})w_N^k,w_N^k)\\
\le\;& c\|u_N^k\|_\infty\|w_N^k\|\,\|w_N^{k-1}\|+b\|w_N^k\|^2\le\frac{3b}{4}\|w_N^k\|^2+c\|u_N^k\|_1^2\|w_N^{k-1}\|^2.
\end{aligned}$$

Hence (4.2) can be rewritten as

$$\overline{\partial}_t\|w_N^k\|^2+\alpha\|w_{Nx}^k\|^2-b\|w_N^k\|^2\le c\|u_N^k\|_1^2\,\|w_N^{k-1}\|^2. \tag{4.3}$$

By applying the discrete Gronwall's inequality, we obtain

$$\|w_N^k\|^2\le\|w_N^0\|^2\mathrm{e}^{ct_n}=\mathrm{e}^{ct_n}\|u_N^0-v_N^0\|^2\qquad\text{for all } n\ge 1. \tag{4.4}$$

Taking the sum of (4.3) for k from 1 to n, we have

$$\|w_N^k\|^2+\tau\sum_{k=1}^n\|w_N^k\|_1^2\le(1+c\mathrm{e}^{ct_n})\|u_N^0-v_N^0\|^2. \tag{4.5}$$

Letting $\varphi=-w_{Nxx}^k$ in (4.1), we have

$$\begin{aligned}
&\frac{1}{2}\overline{\partial}_t\|w_{Nx}^k\|^2+\alpha\|w_{Nxx}^k\|^2+B(u_N^{k-1},u_N^k,-w_{Nxx}^k)-B(v_N^{k-1},v_N^k,-w_{Nxx}^k)\\
&\qquad\le(G(u_N^{k-1})u_N^k-G(v_N^{k-1})v_N^k,-w_{Nxx}^k).
\end{aligned} \tag{4.6}$$

Since

$$\begin{aligned}
&B(u_N^{k-1},u_N^k,-w_{Nxx}^k)-B(v_N^{k-1},v_N^k,-w_{Nxx}^k)\\
=\;&B(u_N^{k-1},w_N^k,-w_{Nxx}^k)+B(u_N^{k-1},v_N^k,-w_{Nxx}^k)-B(v_N^{k-1},v_N^k,-w_{Nxx}^k),
\end{aligned}$$

where

$$\begin{aligned}
B(u_N^{k-1},w_N^k,-w_{Nxx}^k)=\;&(2F(u_N^{k-1})w_{Nx}^k+F'(u_N^{k-1})u_{Nx}^{k-1}w_N^k,-w_{Nxx}^k)\\
\le\;&c\|w_{Nxx}^k\|(\|u_N^{k-1}\|_\infty\|w_{Nx}^k\|+\|u_{Nx}^{k-1}\|\,\|w_N^k\|_\infty)\\
\le\;&\frac{\alpha}{6}\|w_{Nxx}^k\|^2+c\|u_N^{k-1}\|_1^2\,\|w_N^k\|^2,\\
&B(u_N^{k-1},v_N^k,-w_{Nxx}^k)-B(v_N^{k-1},v_N^k,-w_{Nxx}^k)\\
=\;&2(F(u_N^{k-1})-F(v_N^{k-1}),-v_{Nx}^kw_{Nxx}^k)+(F'(u_N^{k-1})w_{Nx}^{k-1},-v_N^kw_{Nxx}^k)\\
&\qquad+(F'(u_N^{k-1})-F'(v_N^{k-1}),-v_{Nx}^{k-1}v_N^kw_{Nxx}^k)\\
\le\;&c\|w_{Nxx}^k\|\,(\|v_N^k\|_\infty\|w_{Nx}^{k-1}\|+(\|v_{Nx}^k\|+\|v_{Nx}^{k-1}\|)\,\|w_N^{k-1}\|_\infty)\\
\le\;&\frac{\alpha}{6}\|w_{Nxx}^k\|^2+c(\|v_N^k\|_1^2+\|v_N^{k-1}\|^2)\,\|w_N^{k-1}\|_1^2
\end{aligned}$$

and

$$\begin{aligned}
&(G(u_N^{k-1})u_N^k - G(v_N^{k-1})v_N^k, -w_{Nxx}^k)\\
=\ &(G(u_N^{k-1})w_N^k + (G(u_N^{k-1}) - G(v_N^{k-1})v_N^k, -w_{Nxx}^k)\\
\le\ &c\|w_{Nxx}^k\|\,(\|u_N^{k-1}\|_\infty\|w_N^k\| + \|v_N^k\|_\infty\|w_N^{k-1}\|)\\
\le\ &\frac{\alpha}{6}\|w_{Nxx}^k\|^2 + c(\|u_N^{k-1}\|_1^2 + \|v_N^k\|_1^2)\,(\|w_N^k\|^2 + \|w_N^{k-1}\|^2),
\end{aligned}$$

(4.6) can be rewritten as

$$\overline{\partial}_t\|w_{Nx}^k\|^2 + \alpha\|w_{Nxx}^k\|^2 \le\ c(\|w_N^k\|^2 + \|w_N^{k-1}\|_1^2). \tag{4.7}$$

Taking the sum of (4.7) for k from 1 to n, and applying (4.4) and (4.5), we have

$$\|w_{Nx}\|^2 + \alpha\tau\sum_{k=1}^{n}\|w_{Nxx}^k\|^2 \le \|w_{Nx}^0\|^2 + c\mathrm{e}^{ct_n}\|w_N^0\|^2. \tag{4.8}$$

Now we assume that the conditions of Lemma 2.7 are satisfied.
Letting $\varphi = w_{Nx^4}^k$ in (4.1), we have

$$\begin{aligned}
&\frac{1}{2}\overline{\partial}_t\|w_{Nxx}^k\|^2 + \alpha\|w_{Nxxx}^k\|^2 + B(u_N^{k-1}, u_N^k, w_{Nx^4}^k) - B(v_N^{k-1}, v_N^k, w_{Nx^4}^k)\\
&\qquad \le (G(u_N^{k-1})u_N^k - G(v_N^{k-1})v_N^k, w_{Nx^4}^k).
\end{aligned} \tag{4.9}$$

Since

$$\begin{aligned}
&B(u_N^{k-1}, u_N^k, w_{Nx^4}^k) - B(v_N^{k-1}, v_N^k, w_{Nx^4}^k)\\
=\ &B(u_N^{k-1}, w_N^k, w_{Nx^4}^k) + B(u_N^{k-1}, v_N^k, w_{Nx^4}^k) - B(v_N^{k-1}, v_N^k, w_{Nx^4}^k),
\end{aligned}$$

where

$$\begin{aligned}
&B(u_N^{k-1}, w_N^k, w_{Nx^4}^k)\\
=\ &(2F(u_N^{k-1})w_{Nx}^k + F'(u_N^{k-1})u_{Nx}^{k-1}w_N^k, w_{Nx^4}^k)\\
=\ &(2F(u_N^{k-1})w_{Nxx}^k + 3F'(u_N^{k-1})u_{Nx}^{k-1}w_{Nx}^k + F'(u_N^{k-1})u_{Nxx}^{k-1}w_N^k, -w_{Nxxx}^k)\\
&\quad + (F''(u_N^{k-1})(u_{Nx}^{k-1})^2w_N^k, -w_{Nxxx}^k)\\
\le\ &\|w_{Nxxx}^k\|\,(\|u_N^{k-1}\|\,\|w_{Nxx}^k\| + \|u_{Nx}^{k-1}\|_\infty\|w_{Nx}^k\|\\
&\qquad +\|u_{Nxx}^{k-1}\|\,\|u_{Nx}^{k-1}\|\,\|w_N^k\|_\infty + \|u_{Nx}^{k-1}\|_\infty\|w_N^k\|)\\
\le\ &\frac{\alpha}{8}\|w_{Nxxx}^k\|^2 + c\|u_N^{k-1}\|_2^2\,\|w_N^k\|^2;
\end{aligned}$$

similarly we have

$$B(u_N^{k-1}, v_N^k, w_{Nx^4}^k) - B(v_N^{k-1}, v_N^k, w_{Nx^4}^k) \le \frac{\alpha}{8}\|w_{Nxxx}^k\|^2 + c\|v_N^k\|^2\,\|w_N^{k-1}\|_2^2$$

and

$$\begin{aligned}
&(G(u_N^{k-1})u_N^k - G(v_N^{k-1})v_N^k, w_{Nx^4}^k)\\
\le\ &\frac{\alpha}{6}\|w_{Nxxx}^k\|^2 + c(\|u_N^{k-1}\|_2^2 + \|v_N^k\|_2^2)\,(\|w_N^k\|_1^2 + \|w_N^{k-1}\|_1^2),
\end{aligned}$$

hence

$$\overline{\partial}_t\|w_{Nxx}^k\|^2+\tau\|\overline{\partial}_t w_{Nxx}^k\|^2+\alpha\|w_{Nxxx}^k\|^2\le\ c(\|w_N^k\|^2+\|w_N^{k-1}\|_2^2). \tag{4.10}$$

Taking the sum of (4.10) for k from 1 to n, and using (4.4), (4.5), and (4.8), we have

$$\|w_{Nxx}^n\|^2\le\|w_{Nxx}^0\|^2+c\mathrm{e}^{ct_n}\|w_N^0\|_1^2 \qquad \text{for all } n\ge 1. \tag{4.11}$$

Combining (4.4), (4.8), and (4.11), we complete the proof of this theorem. □

5. Long-time stability and convergence of the discrete scheme. In this section, let $h=h(x,t)$. To prove the long-time stability and the convergence of discrete scheme (1.4), (1.5), the following results are necessary.

LEMMA 5.1. *Suppose that* $f\in C^m, g\in C^{m-1}$, $|f'(s)|\le A(|s|^2+1)$, $g(0)=0$, $g'(s)\le b<0$, $|g'(s)|\le B|s|(|s|^4+1)$; $h,h_t\in L^\infty(R^+;H_p^{m-2}(\Omega))\bigcap L^2(R^+;H_p^{m-3}(\Omega))$; $u_0\in H_p^m(\Omega)$ $(m\ge 5)$. *Then we have the solution* $u(x,t)\in L^\infty(R^+;H_p^m(\Omega))\cap L^2(R^+;H_p^{m+1}(\Omega))$ *of problem* (1.1)–(1.3), *and*

$$u_t\in L^\infty(R^+;H_p^{m-3}(\Omega))\cap L^2(R^+;H_p^{m-2}(\Omega));\quad u_{tt}\in L^2(R^+;H_p^{m-5}(\Omega)).$$

LEMMA 5.2. *Under the conditions of Lemma* 5.1, *we have the estimates for the solution* u_N^k *of discrete problem* (1.4), (1.5),

$$\|u_N^n\|_2^2+\tau^2\sum_{k=1}^n\|\overline{\partial}_t u_N^k\|_2^2+\tau\sum_{k=1}^n\|u_N^k\|_3^2\le c \qquad \text{for all } n\ge 1,$$

where the constant c *is independent of* N, τ, *and* n.

The proofs of Lemmas 5.1 and 5.2 are similar to those of Lemmas 2.3–2.7.

Now we give the main results in this section.

THEOREM 5.3. *Under the conditions of Lemma* 5.1, *we have the following error estimates for the solutions* u_N^k *of discrete scheme* (1.4), (1.5):

$$\|u^n-u_N^n\|_1^2\le c(N^{-2(m-1)}+\tau) \qquad \text{for all } n\ge 1,$$

where the constant $c=c(\|u_0\|_m)$ *is independent of* N, τ, *and* n.

Proof. Let $u^k-u_N^k=(u^k-P_Nu^k)+(P_Nu^k-u_N^k)\triangleq\eta^k+\xi^k$. Then ξ^k satisfies

$$\begin{aligned}&(\overline{\partial}_t\xi^k-\alpha\xi_{xx}^k+\beta\xi_{xxx}^k,\varphi)+B(u^k,u^k,\varphi)-B(u_N^{k-1},u_N^k,\varphi)\\ =\ &(g(u^k)-G(u_N^{k-1})u_N^k,\varphi)+(\overline{\partial}_t u^k-u_t^k,\varphi) \qquad \text{for all } \varphi\in S_N,\ n\ge 1,\end{aligned} \tag{5.1}$$

$$\xi^0=0. \tag{5.2}$$

Letting $\varphi=\xi^k$ in (5.1), we have

$$\begin{aligned}&\frac{1}{2}\overline{\partial}_t\|\xi^k\|^2+\frac{\tau}{2}\|\overline{\partial}_t\xi^k\|^2+\alpha\|\xi_x^k\|^2+B(u^k,u^k,\xi^k)-B(u_N^{k-1},u_N^k,\xi^k)\\ &\quad=(g(u^k)-G(u_N^{k-1})u_N^k,\xi^k)+(\overline{\partial}_t u^k-u_t^k,\xi^k).\end{aligned} \tag{5.3}$$

Now we estimate every term in the above equality. First,

$$(g(u^k)-G(u_N^{k-1})u_N^k,\xi^k)=(g(u^k)-g(u_N^k),\xi^k)+(g(u_N^k)-G(u_N^{k-1})u_N^k,\xi^k),$$

where from the definition of $G(u)$, we have

$$\begin{aligned}
&(g(u^k)-g(u_N^k),\xi^k)\\
&=(g'(0)(u^k-u_N^k)+g''(\theta_2)u_N^k(u^k-u_N^k)+\frac{1}{2}g''(\theta_1)(u^k-u_N^k)^2,\xi^k)\\
&\le b\|\xi^k\|^2+c\|\xi^k\|\|u^k-u_N^k\|(\|u^k\|_\infty+\|u_N^k\|_\infty)\\
&\le \frac{15|b|}{16}\|\xi^k\|^2+c(\|u^k\|_1^2+\|u_N^k\|_1^2)\left(\|\eta^{k-1}\|^2+\|\xi^{k-1}\|^2+\tau^2\|\overline{\partial}_t u_N^k\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s\right)
\end{aligned}$$

and

$$(g(u_N^k)-G(u_N^{k-1})u_N^k,\xi^k)\le\frac{|b|}{8}\|\xi^k\|^2+c\tau^2\|\overline{\partial}_t u_N^k\|^2.$$

This implies

$$\begin{aligned}
&(g(u^k)-G(u_N^{k-1})u_N^k,\xi^k)\\
\le\ & c(\|u^k\|^2+\|u_N^k\|^2)\,(\|\xi^{k-1}\|^2+\|\eta^{k-1}\|^2)+\frac{7b}{8}\|\xi^k\|^2\\
&+c\Big(\tau^2\|\overline{\partial}_t u_N^k\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s\Big).
\end{aligned}\tag{5.4}$$

Second,

$$\begin{aligned}
&B(u^k,u^k,\xi^k)-B(u_N^{k-1},u_N^k,\xi^k)\\
=\ &(F(u^k)u_x^k-F(u_N^{k-1})u_{Nx}^k,\xi^k)-((F(u^k)u^k-F(u_N^{k-1})u_N^k,\xi_x^k),
\end{aligned}$$

where from the definition of $F(u)$, we have

$$\begin{aligned}
&(F(u^k)u_x^k-F(u_N^{k-1})u_{Nx}^k,\xi^k)\\
=\ &(F(u^k)(u_x^k-u_{Nx}^k,\xi^k)+(u_{Nx}^k(F(u^k)-F(u_N^{k-1})),\xi^k)\\
=\ &(F(0)(u_x^k-u_{Nx}^k,\xi^k)+(F'(\theta u^k)u^k(u_x^k-u_{Nx}^k),\xi^k)+(u_{Nx}^k(F(u^k)-F(u_N^{k-1})),\xi^k)\\
\le\ & c\|\xi^k\|\big(\|u^k\|_\infty\|u_x^k-u_{Nx}^k\|+\|u_{Nx}^k\|_\infty(\|u^k-u^{k-1}\|+\|u_N^k-u_N^{k-1}\|)\big)\\
\le\ & c(\|u^k\|_1^2+\|u_N^k\|_2^2)(\|\xi^{k-1}\|^2+\|\eta^{k-1}\|^2+\|\eta_x^k\|^2)+\frac{|b|}{8}\|\xi^k\|^2+\frac{\alpha}{16}\|\xi_x^k\|^2\\
&+c\left(\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s+\tau^2\|\overline{\partial}_t u_N^k\|^2\right)
\end{aligned}$$

and

$$\begin{aligned}
((F(u^k)u^k-F(u_N^{k-1})u_N^k,\xi_x^k)\le\ &\frac{\alpha}{16}\|\xi_x^k\|^2+c(\|u^k\|_1^2+\|u_N^k\|_1^2)\|\xi^{k-1}\|^2\\
&+c(\|u^k\|_1^2+\|u_N^k\|_1^2)\Big(\|\eta^{k-1}\|^2+\tau^2\|\overline{\partial}_t u_N^k\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s\Big).
\end{aligned}$$

Hence

$$\begin{aligned}
&B(u^k,u^k,\xi^k)-B(u_N^{k-1},u_N^k,\xi^k)\\
\le\ & c(\|u^k\|_1^2+\|u_N^k\|_2^2)(\|\xi^{k-1}\|^2+\|\eta^{k-1}\|^2+\|\eta^k\|_1^2)+\frac{\alpha}{8}\|\xi_x^k\|^2\\
&+\frac{|b|}{8}\|\xi^k\|^2+c\left(\tau^2\|\overline{\partial}_t u_N^k\|^2+\tau\int_{t_{k-1}}^{t_k}\|u_t\|^2\,\mathrm{d}s\right).
\end{aligned}\tag{5.5}$$

Finally

$$(\overline{\partial}_t u^k - u_t^k, \xi^k) \le \frac{|b|}{16}\|\xi^k\|^2 + c\tau \int_{t_{k-1}}^{t_k} \|u_{tt}\|^2 \,\mathrm{d}s. \tag{5.6}$$

Then combining (5.3) to (5.6), we obtain

$$\begin{aligned} &\overline{\partial}_t \|\xi^k\|^2 + \tau\|\overline{\partial}_t \xi^k\|^2 + \|\xi^k\|_1^2 \\ &\le\ c(\|u^k\|_1^2 + \|u_N^k\|_2^2)\,\|\xi^{k-1}\|^2 + c(\|u^k\|_1^2 + \|u_N^k\|_2^2)\,(\|\eta^{k-1}\|^2 + \|\eta^k\|_1^2) \\ &\qquad + c\left(\tau^2\|\overline{\partial}_t u_N^k\|^2 + \tau\int_{t_{k-1}}^{t_k} (\|u_{tt}\|^2 + \|u_t\|^2)\mathrm{d}s\right). \end{aligned} \tag{5.7}$$

By using the integrating formula by parts, it is easily proved that

$$\tau\sum_{k=0}^{\infty} \|u^k\|_1^2 \le \tau\int_0^{\infty} \|u_t\|_1^2 \,\mathrm{d}t + (1+\tau)\int_0^{\infty} \|u\|_1^2 \,\mathrm{d}t;$$

then by applying the discrete Gronwall's inequality to (5.7) and using Lemma 1.1 and Lemmas 5.1–5.2, we derive

$$\|\xi^n\|^2 \le c(N^{-2(m-1)} + \tau). \tag{5.8}$$

Taking the sum of (5.7) for k from 1 to n and using the above inequality, it is obtained that

$$\tau^2\sum_{k=1}^{n} \|\overline{\partial}_t \xi^k\|^2 + \tau\sum_{k=1}^{n} \|\xi^k\|_1^2 \le c(N^{-2(m-1)} + \tau).$$

Let $\varphi = \xi_{xx}^k$ in (5.1). Then similar to (5.7) we have

$$\begin{aligned} \overline{\partial}_t \|\xi_x^k\|^2 + \alpha\|\xi_{xx}^k\|^2 &\le c(\|u^k\|_1^2 + \|u_N^k\|_2^2)(\|\eta^k\|_1^2 + \|\eta^{k-1}\|^2) \\ &\quad + \left(\|\xi^k\|^2 + \tau^2\|\overline{\partial}_t u_N^k\|_1^2 + \tau\int_{t_{k-1}}^{t_k} \|u_{tt}\|^2 ds\right). \end{aligned}$$

Summing them for k from 1 to n, and using the previous estimates, we have

$$\|\xi_x^n\|^2 + \tau\sum_{k=1}^{n} \|\xi_{xx}^k\|^2 \le c(N^{-2(m-1)} + \tau). \tag{5.9}$$

Now combining (5.8) and (5.9) and using the triangle inequality, we complete the proof of the theorem. □

THEOREM 5.4. *Suppose that f, g satisfy the conditions of Theorem* 5.3*; h, $h_t \in L^2(R^+;\ H_p^1(\Omega))$; $\{u_N^n\}$ and $\{v_N^n\}$ are the two solutions of the discrete scheme* (1.4), (1.5) *with initial values $\{u_N^0\}$ and $\{v_N^0\}$, respectively; and the initial values satisfy $\|u_N^0\|_2 \le R_0$, $\|v_N^0\|_2 \le R_0$. Then we have*

$$\|u_N^n - v_N^n\|_2 \le c\|u_N^0 - v_N^0\|_2 \qquad \text{for all } n \ge 0,$$

where the constant c is independent of N, τ, and t_n.

Proof. The proof of this theorem is similar to that of Theorem 4.1. Let $u_N^k - v_N^k = w_N^k$. Then by applying Lemma 5.2 and the discrete Gronwall's inequality to (4.3), we have

$$\|w_N^n\|^2 \le \|w_N^0\|^2 \exp\left(c\tau \sum_{k=0}^{n} \|u_N^k\|_1^2\right) \le c\,\|w_N^0\|^2. \tag{5.10}$$

Taking the sum of (4.3) and using (5.10) and Lemma 5.2, we obtain

$$\tau^2 \sum_{k=1}^{n} \|\overline{\partial}_t w_N^k\|^2 + \alpha\tau \sum_{k=1}^{n} \|w_{Nx}^k\|^2 - b \sum_{k=1}^{n} \|w_N^k\|^2 \le c\|w_N^0\|^2. \tag{5.11}$$

Taking the sum of (4.7) and using (5.11), we have

$$\|w_{Nx}^n\|^2 + \tau^2 \sum_{k=1}^{n} \|\overline{\partial}_t w_{Nx}^k\|^2 + \alpha \sum_{k=1}^{n} \|w_{Nxx}^k\|^2 \le \|w_{Nx}^0\|^2 + c\|w_N^0\|^2. \tag{5.12}$$

Taking the sum of (4.10) and using (5.11), (5.12), we find

$$\|w_{Nxx}^n\|^2 \le \|w_{Nxx}^0\|^2 + c\|w_N^0\|_1^2. \tag{5.13}$$

Now combining (5.10), (5.12), and (5.13), we recover the results of this theorem. ∎

REFERENCES

[1] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

[2] C. Foias, G. Sell, and R. Temam, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309–353.

[3] G. R. Sell and Y. You, *Inertial manifolds: the nonself adjoint case*, J. Differential Equations, 96 (1992), pp. 203–255.

[4] J. M. Ghidaglia and R. Temam, *Lower bound on the dimension of the attractor for the Navier-Stokes equations in space* 3, in Mechanics, Analysis and Geometry: 200 Years After Lagrange, M. Francariglia, ed., Elsevier, Amsterdam, 1990.

[5] R. Rosa and R. Temam, *Inertial manifolds and normal hyperbolicity*, Acta. Appl. Math., 45 (1996), pp. 1–50.

[6] G. Boling and W. Bixiang, *Finite dimensional behavior for the derivative Ginzburg-Landau equation in two spatial dimensions*, Phys. D, 89 (1995), pp. 83–90.

[7] G. Boling, *Finite dimensional behavior for weakly damped generalized KdV-Burgers equations*, Northeast. Math. J., 10 (1994), pp. 309–319.

[8] A. Debussche and R. Temam, *Convergence families of approximate inertial manifolds*, J. Math. Pures. Appl., 73 (1994), pp. 489–522.

[9] J. M. Chidaglia, *A note on the strong convergence towards attractors for damped KdV equation*, J. Differential Equations, 110 (1994), pp. 356–359.

[10] Y. Li, D. W. Mclaughlin, J. Shatah, and S. Wiggins, *Persistent homoclinic orbits for a perturbed nonlinear Schrödinger equation*, Comm. Pure Appl. Math., 49 (1996), pp. 1175–1255.

[11] J. K. Hale, X. B. Lin, and G. Raugel, *Upper semi-continuity of attractors for approximations of semi-groups and partial differential equations*, Math. Comp., 50 (1988), pp. 89–123.

[12] C. M. Elliott and S. Larsson, *Error estimates with smooth and nonosmooth data for a finite element method for the Cahn-Hilliard equation*, Math. Comp., 58 (1992), pp. 603–630.

[13] R. Marion, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1988), pp. 1139–1157.

[14] M. Marion and R. Temam, *Nonlinear Galerkin methods: The finite elements case*, Numer. Math., 57 (1990), pp. 205–226.

[15] R. Temam, *Inertial manifolds and multigrid methods*, SIAM J. Math. Anal., 21 (1990), pp. 154–178.

[16] C. Min and R. Temam, *Nonlinear Galerkin methods in the finite difference case and wavelet-like incremental unknowns*, Numer. Math., 64 (1993), pp. 271–294.
[17] S. Jie, *Long-time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–209.
[18] D. Qiang, G. Benyu, and S. Jie, *Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 39 (2001), pp. 735–762.
[19] L. Shujuan, *The spectral method for long-time behavior of generalized KdV-Burgers equation*, Mathematica Numerica Sinica (in Chinese), 21 (1999), pp. 129–138.
[20] Z. Fayong, *Spectral approximations of attractors of generalized KdV-Burgers equations*, Numer. Math. J. Chinese Univ., 22 (2000), pp. 32–47.
[21] C. H. Su and C. S. Gardner, *KdV equation and generalization* IV. *Derivation of the KdV equation and Burgers equations*, J. Math. Phys., 19 (1969), pp. 536–539.
[22] J. C. Saut, *Sur quelques generalizations de* I *equation de Korteweg-de Vries*, J. Math. Pures Appl., 58 (1979), pp. 21–61.
[23] G. Boling, *The global solution for one class of generalized KdV equation*, Acta Math. Sinica., 25 (1982), pp. 641–656.
[24] Z. Yulin and G. Boling, *The periodic boundary problems and the initial value problems for the system of generalized Kortew-de Vries type of high order*, Acta Math. Sinica, 27 (1984), pp. 154–176.
[25] K. Yoshimura and S. Watanabe, *Chaotic behavior of nonlinear evolution equation with fifth-order dispersion*, J. Phys. Soc. Japan, 51 (1982), pp. 3028–3035.
[26] C. Canuto and A. Qarteroni, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., 38 (1982), pp. 201–229.
[27] V. G. Maeja, *Sobolev Space*, Springer-Verlag, New York, 1985.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 586–612

A MATRIX ANALYSIS OF OPERATOR-BASED UPSCALING FOR THE WAVE EQUATION*

OKSANA KOROSTYSHEVSKAYA† AND SUSAN E. MINKOFF†

Abstract. Scientists and engineers who wish to understand the earth's subsurface are faced with a daunting challenge. Features of interest range from the microscale (centimeters) to the macroscale (hundreds of kilometers). It is unlikely that computational power limitations will ever allow modeling of this level of detail. Numerical upscaling is one technique intended to reduce this computational burden. The operator-based algorithm (developed originally for elliptic flow problems) is modified for the acoustic wave equation. With the wave equation written as a first-order system in space, we solve for pressure and its gradient (acceleration). The upscaling technique relies on decomposing the solution space into coarse and fine components. Operator-based upscaling applied to the acoustic wave equation proceeds in two steps. Step one involves solving for fine-grid features internal to coarse blocks. This stage can be solved quickly via a well-chosen set of coarse-grid boundary conditions. Each coarse problem is solved independently of its neighbors. In step two we augment the coarse-scale problem via this internal subgrid information. Unfortunately, the complexity of the numerical upscaling algorithm has always obscured the physical meaning of the resulting solution. Via a detailed matrix analysis, the coarse-scale acceleration is shown to be the solution of the original constitutive equation with input density field corresponding to an averaged density along coarse block edges. The pressure equation corresponds to the standard acoustic wave equation at nodes internal to coarse blocks. However, along coarse cell boundaries, the upscaled solution solves a modified wave equation which includes a mixed second-derivative term.

Key words. upscaling, multiscale methods, acoustic wave propagation, matrix analysis, seismology

AMS subject classifications. 35L05, 74Q15, 86-08, 86A15, 65M06

DOI. 10.1137/050625369

1. Introduction.

Many problems in physics and engineering result in models involving multiple scales. Often, one is forced to solve partial differential equations with highly oscillatory coefficients over very large spatial domains. The size of the resulting discrete problems makes a direct numerical simulation extremely difficult. In reservoir simulation, for example, it is common to require tens of millions of grid blocks to capture the fluctuations in the permeability of the medium [10]. Upscaling (or multiscale) techniques provide a way to solve the problem on a coarser scale while still capturing some of the effects of the fine scale.

There are two main approaches to upscaling. The first idea involves averaging the input data and thereby forming effective (or upscaled) parameters. A new problem corresponding to this upscaled data is solved on a coarse grid [7], [10]. The second approach allows the problem to be solved on the coarse scale without explicitly forming effective coefficients. Instead, some kind of operator-based technique is used to incorporate the fine-grid information into a coarse solution [4], [14].

Averaging techniques [16], [10] and the methods based on homogenization [7], [8] are examples of the first approach. Averaging is one of the simplest methods for

*Received by the editors February 25, 2005; accepted for publication (in revised form) November 2, 2005; published electronically March 17, 2006. This research was performed with funding from both the Collaborative Math-Geoscience Program at NSF (grant EAR-0222181) and a GAANN grant from the U.S. Department of Education (award P200A030097).

http://www.siam.org/journals/sinum/44-2/62536.html

†Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 (oksana1@math.umbc.edu, sminkoff@math.umbc.edu).

calculating effective parameters. The general idea is to obtain an effective value on each coarse block as some kind of average (arithmetic, harmonic, or geometric) of the original input parameter field. The averaging methods have a limited range of application. They only give correct results for certain types of media [16].

The methods based on homogenization theory [7], [8], [1] use asymptotic analysis to replace the given problem by a macroscopic problem with simple effective coefficients. (In many cases the effective coefficients are constants.) Homogenization is based on two main hypotheses:

- the medium under study is periodic, and
- the period is small compared to the size of the domain.

In homogenization, the effective parameters are constructed analytically. This construction requires the solution of so-called cell problems—boundary value problems within a single period cell. The main drawback of such methods is that they are, typically, not applicable to realistic nonperiodic structures.

More recently, a number of methods of the second, operator-based, type have been introduced and developed. The multiscale finite element method for elliptic problems with rapidly oscillating coefficients developed by Hou and Wu is an example of this approach [14], [12]. The idea behind the multiscale finite element method is to construct special coarse-scale finite element basis functions which capture the small-scale information within each element. These basis functions are obtained by solving homogeneous elliptic equations in each element subject to specified boundary conditions. The effect of the small scale is then incorporated into the coarse solution through the global stiffness matrix. The main difficulty is that large errors may result from "resonance" between the grid scale and the scales of the continuous problem. An oversampling technique is used to overcome this limitation [9], [18], [13].

Another example of the operator-based approach to upscaling is the mortar method [15], [19]. This method is based on domain decomposition. The physical domain is decomposed into a series of blocks in which different numerical grids, physical models, and discretization techniques can be used. Mortar finite element spaces are then used to allow for nonmatching grids across the block interfaces and to impose physically meaningful interface continuity conditions. One of the advantages of the method is that one may vary the number of mortar degrees of freedom and thus achieve improvements in accuracy. The disadvantage of the mortar technique is that it can be expensive, especially when a large number of degrees of freedom are used in the interface problems.

The operator-based upscaling technique was introduced for elliptic equations by Arbogast, Minkoff, and Keenan in [4]. The method was further developed by Arbogast et al. in [3], [6], [5]. The idea of the method is to decompose the solution into two parts: a coarse grid representation and a subgrid component. The subgrids are the portions of the domain contained within each of the coarse-grid cells. The problem is solved in two steps. First, we solve for the fine-scale information internal to each coarse cell. Due to a simplifying assumption imposed on the boundaries of the coarse cells, these subgrid problems decouple and can be solved independently. The second step involves using the subgrid solutions to modify the coarse-scale operator. The resulting coarse-scale solution includes some of the fine-scale features of the problem under study. A significant advantage of this method is that we use data provided on the fine scale directly in our computations (no averaging). Further, we need not assume that the medium is periodic. Perhaps most importantly, however, the operator upscaling method does not assume a separation of scales (a fundamental underlying assumption of asymptotic techniques). Geologic materials contain heterogeneities on

a continuum of scales which for realistic problems negates the basic separation of scales hypothesis.

In the companion paper by Vdovina, Minkoff, and Korostyshevskaya [17] we adapt operator-based upscaling for use with the acoustic wave equation. We focus on the second-order in space and time acoustic wave equation which we write as a system of two first-order equations (in space). Thus we solve for both pressure and its derivative (acceleration). The practical details of both the serial and parallel implementations of the algorithm are described in that paper. Specifically we detail the numerical implementation of the system described in section 2.1 of this paper (the original upscaled system). The two-stage algorithm requires solving the subgrid equations (2.5–2.6) and the coarse equation (2.7) for each time step. In Vdovina et al. [17] we describe the relative costs of the serial and parallel implementations as well. Upscaling by definition implies that we are not solving the full fine-grid problem. Some simplifications must be made to speed up the computation. With this method our primary simplification is that we impose zero flux conditions between coarse blocks at the subgrid stage of the algorithm (when we solve for internal subwavelength scale information). This assumption means that the most costly part of the algorithm (the subgrid solve) is trivial to parallelize. Each coarse cell is solved independently of its neighbors. Thus we achieve near-optimal speedup for the parallel upscaling algorithm. Because acceleration is upscaled but pressure is defined only on the fine grid in our current implementation, we find that standard fine grid stability and dispersion results (CFL and number of gridpoints per wavelength) hold. (Extensions of the method for upscaling both pressure and acceleration are also discussed in [17].) Finally, three realistic numerical experiments are described in that paper. We compare the upscaled solution to a full finite difference solution of the wave equation for a periodic (checkerboard) velocity, a finely layered medium, and a stochastic velocity field describing a two-component mixture of materials taken from a von Karman distribution. These three velocity fields were chosen primarily for their geologic relevance. (They contain basic components one might encounter in subsurface regions such as the Gulf of Mexico or deep crust.) The upscaled solution qualitatively captures even the subwavelength-scale heterogeneity of the full solution.

Our focus in this work is on addressing the question of what the upscaled solution means. What physics does the solution model? Is our model still the acoustic wave equation or an attenuated version of the wave equation? We provide the first answers to these questions in this paper via a linear algebra analysis of the method. What this analysis highlights is that the numerical upscaling process solves a constitutive equation similar in form to the original equation. The constitutive equation relates acceleration to the gradient of pressure. For the coarse (upscaled) problem, however, the parameter field (density) reduces to an averaged density along coarse block edges. Similarly, when analyzing the pressure equation, we find the upscaled solution solves the original wave equation at nodes internal to the coarse blocks, but a modified equation at coarse block edges. Specifically, a cross-derivative (involving differentiation with respect to both x and y) enters the standard wave equation. This analysis not only simplifies the coding of the algorithm but illuminates the physical meaning of the upscaled solution. In this paper, we focus on the theory. (We do not include numerical experiments based on the system of equations which result from the analysis given in this paper.) However, the new formulation should simplify the simulations described in [17] considerably and is the subject of future work. The standard implementation of operator upscaling involves technicalities (such as the use of numerical Green's functions) that are no longer necessary as a result of this analysis.

It is important to point out that the equations which result from this analysis (specifically the conclusions of Theorems 1, 10, and 11) are to our knowledge entirely new. While other papers exist which discuss convergence of operator upscaling for elliptic problems in the context of finite elements, no other work (outside this paper and the related work by Vdovina et al. [17]) discusses operator upscaling for the wave equation. More importantly, the physics modeled by operator upscaling has not been illuminated (for any PDE) prior to this work.

In the remainder of the paper we describe the mathematics behind the upscaling algorithm. Then we begin the matrix analysis of the two-equation system. First we analyze the constitutive equation for coarse acceleration. The result is both a finite difference stencil and a continuous differential equation corresponding to the upscaled acceleration equation. Finally, we apply a similar analysis to the pressure equation. We define the finite difference stencils for pressure (which depend on the location of the pressure unknown within the coarse block) and corresponding continuous differential equations coming from the upscaling algorithm.

2. Upscaled acoustic wave equation.

2.1. Subgrid upscaling for the acoustic wave equation: Variational formulation. Let Ω be a two-dimensional domain with boundary Γ. We consider the acoustic wave equation in Ω written as a first-order system for acceleration $\vec{v}$ and pressure p:

$$\vec{v} = -\frac{1}{\rho}\nabla p, \tag{2.1}$$

$$\frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} = -\nabla \cdot \vec{v} + f. \tag{2.2}$$

Here c is the sound velocity, ρ is the density, and f is the source of acoustic energy. Both c and ρ are functions of spatial location only *and are assumed to be heterogeneous.* To simplify the presentation of the method, we will assume zero boundary conditions:

$$\vec{v} \cdot \vec{\nu} = 0 \quad \text{on} \quad \Gamma,$$

where $\vec{\nu}$ is the unit outward normal vector.

The subgrid upscaling technique is developed in the context of a mixed finite element method. Let

$$H_0(\text{div}, \Omega) = \{\vec{v} \in (L^2(\Omega))^2 : \nabla \cdot \vec{v} \in L^2(\Omega), \text{ and } \vec{v} \cdot \vec{\nu} = 0 \text{ on } \Gamma\}.$$

We rewrite (2.1)–(2.2) in weak form as follows: find $\vec{v} \in H_0(\text{div}; \Omega)$ and $p \in L^2(\Omega)$ such that

$$\langle \rho \vec{v}, \vec{u} \rangle = \langle p, \nabla \cdot \vec{u} \rangle \,, \tag{2.3}$$

$$\left\langle \frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2}, w \right\rangle = -\langle \nabla \cdot \vec{v}, w \rangle + \langle f, w \rangle \tag{2.4}$$

for all $\vec{u} \in H_0(\text{div}; \Omega)$ and $w \in L^2(\Omega)$.

The idea of the upscaling method is that while the original problem is posed on a fine grid (specifically the input acceleration and density are measured fields on the fine grid), our goal is to solve the problem on a coarse mesh. We wish to capture some of the fine-scale information internal to each coarse cell and then to use this

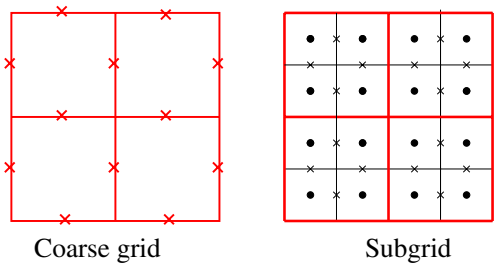


FIG. 2.1. *A 2×2 coarse grid with coarse acceleration unknowns and a corresponding 4×4 fine grid with subgrid unknowns. Pressures live at the centers of the cells. Acceleration lives on cell edges.*

information to determine the best solution possible on the coarse scale. We break the acceleration into two parts—the coarse representation (coarse-grid unknowns) and the subgrid part (fine-grid unknowns internal to each coarse-grid cell):

$$\vec{v} = \vec{v}^c + \delta\vec{v}.$$

In earlier work the pressure space was also decomposed. (See [4] for an example of this decomposition for the elliptic pressure equation.) Unfortunately, the basis functions for the pressure space are computationally clumsy. Since no additional work is required to keep the pressure on the fine grid, we have chosen to only decompose acceleration in this paper. Pressure can be projected onto the coarse grid as a post-processing step if so desired.

In the mixed finite element method, we use a rectangular two-scale mesh and approximate the pressure and the acceleration in finite element spaces W and V, respectively. We take the pressure space W to be the space of piecewise discontinuous constant functions on the fine grid with nodes at the centers of the cells. To define the acceleration space V for the upscaling method, we introduce two finite element spaces δV and V^c associated with the fine and coarse computational grids. Both of these spaces consist of piecewise linear vector functions of the form (a_1x+b_1, a_2y+b_2) living on the edges of the grid blocks (see Figure 2.1). We impose an important simplifying condition on the space δV, namely,

$$\delta\vec{u} \cdot \vec{\nu} = 0 \quad \text{on the boundary of each coarse element.}$$

This is the only simplifying assumption in the definition of our method. It allows us to decouple the subgrid problems coming from different coarse-grid cells. Note that this simplifying assumption only applies to the solution of the subgrid problems which were never intended to be solved exactly but merely approximated. Exact solution of the subgrid problems would lead us back to a full finite difference solution of the wave equation.

The upscaling process consists of two steps. First, we restrict to the subgrid test functions in (2.3)–(2.4) and use the above decomposition to obtain a series of subgrid problems, one for each coarse element E_c:

$$\langle \rho(\vec{v}^c + \delta\vec{v}), \delta\vec{u}\rangle = \langle p, \nabla \cdot \delta\vec{u}\rangle, \tag{2.5}$$

$$\left\langle \frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2}, w \right\rangle = -\langle \nabla \cdot (\vec{v}^c + \delta\vec{v}), w\rangle + \langle f, w\rangle \tag{2.6}$$

for all $\delta\vec{u} \in \delta V$, $w \in W$. The values of $\vec{v}^c$ are unknown at this stage, so we find the solution to the subgrid problem as a function of the coarse unknowns. Note, that the pressure is completely determined by (2.5–2.6).

The second step of the upscaling process uses $\delta\vec{v}$ and p to determine $\vec{v}^c \in V^c$ through solution of the upscaled coarse equation ((2.3) with coarse-grid test functions):

$$\langle \rho(\vec{v}^c + \delta\vec{v}(\vec{v}^c)), \vec{u}^c \rangle = \langle p, \nabla \cdot \vec{u}^c \rangle \tag{2.7}$$

for all $\vec{u}^c \in V^c$.

The problem is solved sequentially in time. We use second-order finite differences to approximate the time derivative in (2.6). First, we find the pressure on the current time level using the velocities and pressure from the previous time levels. Then we solve (2.5) and (2.7) for the subgrid and coarse velocities. The process then repeats for the next time step.

2.2. A matrix analysis discussion of subgrid upscaling. One of the main purposes of this paper is to derive the explicit coarse-scale equation solved by the upscaling algorithm using the discrete form of the subgrid and coarse problems. The advantage of using the discrete formulation is that it allows us to obtain both finite difference and continuous differential equations for coarse acceleration. The finite difference equation yields an explicit stencil for pressure and coarse acceleration that could be implemented directly, thus giving an alternate, simpler way to code the algorithm over that which is detailed in section 2.1 of this paper and in [17]. This analysis shows that coarse acceleration is defined using the average of density values on the boundaries of coarse cells (in a sense defining an upscaled density). From the finite difference equation we can derive the continuous differential equation for the coarse problem which is similar to the original equation (2.1) with density replaced by the upscaled density. We see that the upscaled acceleration approximates the gradient of pressure on the *boundaries* of the coarse cells compensating for the simplifying zero boundary conditions used by the algorithm. In this section, we derive the matrix-vector form of the subgrid and coarse problems. In terms of linear algebra, we see that the subgrid problems can easily be solved for the subgrid acceleration in terms of the pressure and the coarse acceleration. Thus, we can eliminate the subgrid unknowns from the coarse problem to obtain a matrix equation for the coarse acceleration and pressure only. The analysis of the resulting system allows us to determine the differential equation solved by the upscaling algorithm and to understand the physical meaning of the coarse part of the solution.

Homogenization is an alternate technique for analyzing equations with parameters and unknowns that contain multiple scales. The operator upscaling analysis given in this paper is complete in itself and is distinct from a homogenization analysis. Future work may include comparison of operator-based upscaling results to those obtained from asymptotic series solutions. Arbogast and Boyd [2] examine the connection between the multiscale finite element method (a numerical technique based on homogenization) and operator upscaling in the context of elliptic problems. We do not discuss connections between methods here.

We begin by describing the finite element context in which the method is developed. In order to discretize the subgrid and the coarse problems, we use finite element approximations of the unknown functions. In this paper, we restrict our attention to the x component v_x of the acceleration vector, since the equations for v_y are similar.

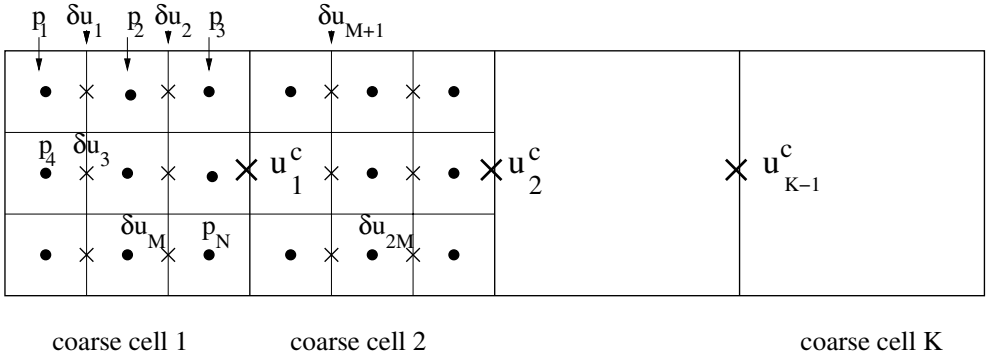


FIG. 2.2. *Subgrid and coarse unknowns. The grid consists of K total coarse cells.*

To simplify notation, consider a domain consisting of one row of coarse cells, each subdivided into a number of fine cells (Figure 2.2). If K is the number of coarse cells, we divide each coarse cell into n_x fine cells in the x direction and n_y fine cells in the y direction. Then the number of fine cells in one coarse block is $N = n_x n_y$, and the total number of fine cells is KN.

In the subgrid upscaling method we will describe here, both the coarse and fine components of the horizontal acceleration are approximated by the space of piecewise continuous linear functions in x and piecewise constants in y. Obviously, other choices of finite element spaces are also possible. However, this choice is the simplest. The coarse and subgrid basis functions $(u_x^c)_i$, $(\delta u_x)_i$ for the acceleration space have nodes at the midpoints of vertical edges of the coarse cells for the coarse acceleration and at the midpoints of vertical edges of the fine cells for the subgrid acceleration (Figure 2.2). At time t,

$$v_x^c(t,x,y) = \sum_{i=1}^{K-1} (\mathbf{v_x^c})_i(t)(u_x^c)_i(x,y), \tag{2.8}$$

$$\delta v_x(t,x,y) = \sum_{i=1}^{MK} (\delta \mathbf{v_x})_i(t)(\delta u_x)_i(x,y). \tag{2.9}$$

Here, $K-1$ is the number of the coarse acceleration unknowns, M is the number of the subgrid acceleration unknowns inside one coarse cell so that MK is the total number of the subgrid acceleration unknowns in the domain, and the coefficients $(\mathbf{v_x^c})_i$, $(\delta \mathbf{v_x})_i$ are to be determined. Since the subgrid acceleration lives on the vertical edges of the fine cells and there are no nodes on the boundaries of the coarse cells, we have that $M = N - n_y$.

The pressure p is approximated by piecewise constants defined on the fine grid. Thus, we have the expansion

$$p(t,x,y) = \sum_{i=1}^{NK} \mathbf{p}_i(t) w_i(x,y). \tag{2.10}$$

Here NK is the total number of pressure unknowns, and w_m is the basis function for pressure, which takes the value 1 on the mth subgrid cell and 0 everywhere else.

We will now derive the matrix forms of the subgrid and coarse problems and then use these equations to eliminate the subgrid unknowns from the coarse problem. To

obtain a linear system from subgrid equations (2.5)–(2.6), we use the finite element expansions given in (2.8)–(2.10):

(2.11)
$$\sum_{j=1}^{K-1}(\mathbf{v_x^c})_j\langle\rho(u_x^c)_j,(\delta u_x)_i\rangle+\sum_{j=1}^{MK}(\delta\mathbf{v_x})_j\langle\rho(\delta u_x)_j,(\delta u_x)_i\rangle=\sum_{j=1}^{NK}\mathbf{p}_j\left\langle w_j,\frac{\partial(\delta u_x)_i}{\partial x}\right\rangle,\quad i=1,2,\ldots,MK,$$

and

(2.12)
$$\sum_{j=1}^{NK}\frac{\partial^2\mathbf{p}_j}{\partial t^2}\left\langle\frac{1}{\rho c^2}w_j,w_i\right\rangle=-\sum_{j=1}^{K-1}(\mathbf{v_x^c})_j\left\langle\frac{\partial(u_x^c)_j}{\partial x},w_i\right\rangle-\sum_{j=1}^{MK}(\delta\mathbf{v_x})_j\left\langle\frac{\partial(\delta u_x)_j}{\partial x},w_i\right\rangle-(v_y\text{ terms})+\langle f,w_i\rangle,\quad i=1,2,\ldots,NK.$$

These equations can be written in matrix form by defining the following matrix entries:

$$a_{i,j}^{ff}=\langle\rho(\delta u_x)_j,(\delta u_x)_i\rangle,\quad a_{i,j}^{cf}=\langle\rho(u_x^c)_j,(\delta u_x)_i\rangle,\quad b_{i,j}^{f}=\left\langle w_j,\frac{\partial(\delta u_x)_i}{\partial x}\right\rangle,\tag{2.13}$$

$$w_{i,j}=\left\langle\frac{1}{\rho c^2}w_j,w_i\right\rangle,\qquad f_i=\langle f,w_i\rangle,\qquad b_{i,j}^{c}=\left\langle w_j,\frac{\partial(u_x^c)_i}{\partial x}\right\rangle.\tag{2.14}$$

Then the subgrid system reduces to the following linear system:

$$A^{cf}\mathbf{v_x^c}+A^{ff}\delta\mathbf{v_x}=B^f\mathbf{p},\tag{2.15}$$

$$W\frac{\partial^2\mathbf{p}}{\partial t^2}=-(B^c)^T\mathbf{v_x^c}-(B^f)^T\delta\mathbf{v_x}-(v_y\text{ terms})+F.\tag{2.16}$$

Similar steps lead to a discretization of the coarse problem (2.7). We obtain the matrix equation:

$$A^{cc}\mathbf{v_x^c}+(A^{cf})^T\delta\mathbf{v_x}=B^c\mathbf{p},\tag{2.17}$$

where the matrix A^{cc} has entries

$$a_{i,j}^{cc}=\langle\rho(u_x^c)_j,(u_x^c)_i\rangle.\tag{2.18}$$

We will now use the matrix form of the subgrid and coarse problems to eliminate the subgrid unknowns from the coarse equation. The subgrid equations are coupled to the coarse scale unknowns. In particular, (2.15) involves coarse acceleration unknowns $\mathbf{v_x^c}$, which are not known at the subgrid step of the algorithm. This equation can easily be solved for $\delta\mathbf{v_x}$ in terms of $\mathbf{v_x^c}$

$$\delta\mathbf{v_x}=-(A^{ff})^{-1}A^{cf}\mathbf{v_x^c}+(A^{ff})^{-1}B^f\mathbf{p}.\tag{2.19}$$

Substituting (2.19) into the coarse problem (2.17), we obtain the matrix equation for the coarse acceleration and pressure only,

$$(A^{cc}-(A^{cf})^T(A^{ff})^{-1}A^{cf})\mathbf{v_x^c}=(B^c-(A^{cf})^T(A^{ff})^{-1}B^f)\mathbf{p}\tag{2.20}$$

or

$$U\mathbf{v}_{\mathbf{x}}^{\mathbf{c}} = D\mathbf{p}. \tag{2.21}$$

Formula (2.21) gives us the matrix equation for the coarse acceleration and pressure. We will use (2.21) later to derive a difference and then differential equations for the coarse acceleration. Therefore, it is worthwhile to study the structure of the matrices U and D, and to explicitly define their entries. In Theorem 1, we show that the choice of bases for the finite element spaces results in matrices which are sparse and have special structure. Moreover, we obtain simple explicit formulas for the entries by computing them with special quadrature rules defined on the fine grid. The choice to use fine grid quadratures is based on the assumption that the parameter fields ρ and c vary on the fine scale. We do not want to introduce averaging errors by requiring these parameters to live on the coarse grid.

Let us begin with some notation that will be used throughout the rest of this paper. Let h_x and h_y be the lengths of a single fine cell in the x and y directions respectively, and H_x, H_y the lengths of a single coarse cell in these two directions. We denote by ρ_l and $(u_x^c)_j|_l$ the values of the density and coarse acceleration basis functions at the lth node of subgrid acceleration, and by ρ_l^j the values of the density on the boundary of the coarse cell corresponding to the jth coarse acceleration node.

THEOREM 1. *Consider the set of subgrid problems* (2.5)–(2.6) *and the coarse problem* (2.7). *The upscaling technique results in the linear system for the coarse acceleration and pressure of the form*

$$U\mathbf{v}_{\mathbf{x}}^{\mathbf{c}} = D\mathbf{p},$$

where U and D are given by (2.20)–(2.21). *U is of size $(K-1)\times(K-1)$ and D is of size $(K-1)\times NK$.*

Further, assume the density ρ and sound velocity c are smooth enough for the inner products in the matrix entries U and D to be computed using fine-grid midpoint and trapezoidal rules. Then the matrix U is diagonal and the matrix D is block upper bidiagonal with blocks of size $1\times N$. The entries of U are given by the sum of density values on the corresponding boundaries of the coarse cells

$$U_{i,i} = (h_x h_y)\sum_{l=1}^{n_y}\rho_l^i. \tag{2.22}$$

The blocks of D are given by

$$D_{i,i} = \frac{h_x h_y}{h_x}\left[0,\quad 0,\quad \ldots,\quad 1,\quad \ldots,\quad 0,\quad 0,\quad \ldots\quad 1\right], \tag{2.23}$$

$$D_{i,i+1} = \frac{h_x h_y}{h_x}\left[-1,\quad 0,\quad \ldots,\quad 0,\quad \ldots,\quad -1,\quad 0,\quad \ldots,\quad 0\right], \tag{2.24}$$

where the nonzero entries correspond to the pressure nodes located along the boundaries of the coarse cells.

Note. In this Theorem, we are using the trapezoidal rule on the fine grid in the x direction and the midpoint rule in the y direction to compute the integrals. When these rules are used, the quadrature node points coincide with the nodes of the subgrid basis functions. This fact simplifies the formulas for the matrix entries. Even more

importantly, both the midpoint rule and the trapezoidal rule are accurate enough that no additional error is incurred for our choice of interpolating polynomials.

The results of Theorem 1 will allow us to determine the explicit difference equation for coarse acceleration solved by the upscaling algorithm and to understand the physical meaning of the resulting problem solution.

Proof. First, let us discuss the coefficient matrix U of coarse acceleration. Recall that

$$U = A^{cc} - (A^{cf})^T (A^{ff})^{-1} A^{cf}. \tag{2.25}$$

In order to explicitly define the entries of U, we need to study the matrices A^{cc}, A^{cf}, and A^{ff}. The following lemmas describe the structure of these matrices.

LEMMA 2. *The matrix A^{cf} is a lower bidiagonal $K \times (K-1)$ block matrix with blocks of size $M \times 1$. If the entries of the matrix are evaluated by the trapezoidal rule in x and the midpoint rule in y on each fine cell, then*

$$a^{cf}_{l,j} = (h_x h_y) \rho_l (u^c_x)_j|_l. \tag{2.26}$$

Proof. The entries of the matrix A^{cf} are the inner products of the coarse and subgrid acceleration basis functions (2.13). Since the total number of subgrid acceleration basis functions is MK and the total number of coarse acceleration basis functions is $K-1$, the matrix A^{cf} is of size $MK \times (K-1)$.

Each coarse acceleration basis function is supported on two coarse cells and, therefore, has nonzero inner products with the subgrid functions from these coarse cells only. For example, the first coarse acceleration basis function is nonzero on the coarse cells 1 and 2 (Figure 2.2). Thus we can partition the matrix into blocks of size $M \times 1$ (where M is the number of subgrid acceleration nodes inside a single coarse cell). Each block A^{cf}_{ij} represents the interaction of the jth coarse basis function with the ith coarse cell. Then the matrix A^{cf} can be written as a $K \times (K-1)$ lower bidiagonal block matrix:

$$A^{cf} = \begin{bmatrix} A^{cf}_{1,1} & & & 0 \\ A^{cf}_{2,1} & A^{cf}_{2,2} & & \\ & \ddots & \ddots & \\ & & & A^{cf}_{K-1,K-1} \\ 0 & & & A^{cf}_{K,K-1} \end{bmatrix}. \tag{2.27}$$

We can evaluate the matrix entries by the trapezoidal rule in x and the midpoint rule in y on each fine cell, so that the quadrature nodal points coincide with the fine acceleration nodes (Figure 2.3). Applying these rules to each entry gives a product of the area of the cell $(h_x h_y)$ with the value of the integrand at the nodal points of the fine acceleration. Since any subgrid acceleration basis function is 1 at the corresponding node and 0 at all other nodes, we obtain

$$a^{cf}_{l,j} = \langle \rho (u^c_x)_j, (\delta u_x)_l \rangle = (h_x h_y) \rho_l (u^c_x)_j \big|_l. \qquad \square$$

LEMMA 3. *If the entries of the matrix A^{ff} are evaluated by the trapezoidal rule in x and the midpoint rule in y on each fine cell, then A^{ff} is an $MK \times MK$ diagonal matrix with entries*

$$a^{ff}_{ll} = (h_x h_y) \rho_l.$$

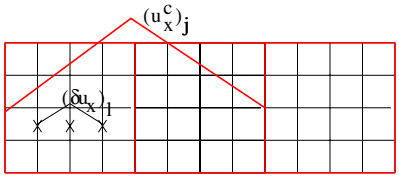


FIG. 2.3. *Coarse and subgrid acceleration basis functions* $(u_x^c)_j$ *and* $(\delta u_x)_l$. *The crosses represent the subgrid acceleration nodes which are used as quadrature nodal points.*

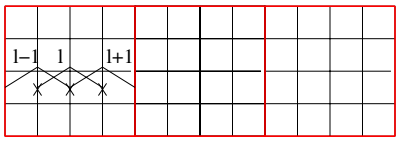


FIG. 2.4. *Subgrid acceleration basis functions* $(\delta u_x)_{l-1}$, $(\delta u_x)_l$, *and* $(\delta u_x)_{l+1}$. *The crosses represent the subgrid acceleration nodes which are used as quadrature nodal points.*

Note. A^{ff} is the coefficient matrix of subgrid unknowns in (2.15). The fact that we are able to reduce A^{ff} to a diagonal matrix makes the elimination of the subgrid unknowns cheap and easy.

Proof. The matrix A^{ff} is the coefficient matrix of subgrid unknowns in (2.15) and its entries are the inner products of subgrid acceleration basis functions (2.13). The total number of subgrid acceleration basis functions is MK, so the matrix A^{ff} has size $MK \times MK$. Each subgrid acceleration basis function interacts with itself and its neighbors to the left and to the right. As in Lemma 2, we use the trapezoidal rule in x and the midpoint rule in y on each fine cell to compute the entries. The quadrature nodal points coincide with the fine acceleration nodes (Figure 2.4). We obtain

$$a^{ff}_{l,l+1} = \langle \rho(\delta u_x)_l, (\delta u_x)_{l+1} \rangle = 0, \qquad a^{ff}_{l-1,l} = \langle \rho(\delta u_x)_{l-1}, (\delta u_x)_l \rangle = 0, \tag{2.28}$$

$$a^{ff}_{l,l} = \langle \rho(\delta u_x)_l, (\delta u_x)_l \rangle = (h_x h_y)\rho_l. \tag{2.29}$$

Here, we made use of the fact that any subgrid acceleration basis function is 1 at the corresponding node and 0 at all other nodes. We see that if the entries are computed as above, the matrix A^{ff} is diagonal. ∎

LEMMA 4. *The matrix* A^{cc} *is a* $(K-1)\times(K-1)$ *tridiagonal matrix. If the entries of* A^{cc} *are evaluated by the composite trapezoidal rule in the* x *direction and the midpoint rule in the* y *direction, then*

$$a^{cc}_{j,j+1} = (h_x h_y) \sum_{l=Mj+1}^{M(j+1)} \rho_l (u_x^c)_{j+1} (u_x^c)_j \big|_l,$$

$$a^{cc}_{j,j-1} = (h_x h_y) \sum_{l=M(j-1)+1}^{Mj} \rho_l (u_x^c)_{j-1} (u_x^c)_j \big|_l,$$

$$a^{cc}_{j,j} = (h_x h_y) \sum_{l=M(j-1)+1}^{M(j+1)} \rho_l (u_x^c)_j^2 \big|_l + (h_x h_y) \sum_{l=1}^{n_y} \rho_l^j.$$

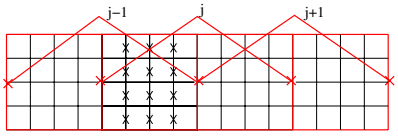


FIG. 2.5. *Coarse acceleration basis functions* $(u_x^c)_{j-1}$, $(u_x^c)_j$, *and* $(u_x^c)_{j+1}$. *The crosses represent the subgrid acceleration nodes* $l = M(j-1)+1, \dots Mj$, *which are used as quadrature nodal points in the computation of* $a^{cc}_{j,j-1}$.

Proof. The entries of the matrix A^{cc} are the inner products of coarse acceleration basis functions (2.18). The total number of the coarse acceleration nodes in the domain is $K-1$, therefore, the matrix has size $(K-1)\times(K-1)$.

Since each coarse acceleration basis function is supported on two coarse cells and interacts with both its neighbors (Figure 2.5), the matrix A^{cc} is tridiagonal. The entries can be computed by the composite trapezoidal rule in the x direction and composite midpoint rule in y on each coarse cell so that the quadrature nodal points coincide with the fine acceleration nodes. Fine grid quadrature is used because the parameter fields ρ and c live on the fine grid, and we do not want to average these parameters. We obtain

$$a^{cc}_{j,j-1} = \langle \rho(u_x^c)_{j-1}, (u_x^c)_j \rangle = (h_x h_y) \sum_{l=M(j-1)+1}^{Mj} \rho_l (u_x^c)_{j-1}\big|_l (u_x^c)_j\big|_l, \tag{2.30}$$

where M is the number of subgrid acceleration nodes inside one coarse cell, and the indices $M(j-1)+1,\dots,Mj$ represent the subgrid acceleration nodes inside coarse cell j (Figure 2.5). Notice that only the nodes internal to the coarse cell are involved in the computation, since either $(u_x^c)_j$ or $(u_x^c)_{j-1}$ is zero on its boundary. Similarly, for $\langle \rho(u_x^c)_{j+1}, (u_x^c)_j \rangle$ we have

$$a^{cc}_{j,j+1} = \langle \rho(u_x^c)_{j+1}, (u_x^c)_j \rangle = (h_x h_y) \sum_{l=Mj+1}^{M(j+1)} \rho_l (u_x^c)_{j+1}\big|_l (u_x^c)_j\big|_l, \tag{2.31}$$

where we sum over the subgrid acceleration nodes inside coarse cell $j+1$. The entry $\langle \rho(u_x^c)_j, (u_x^c)_j \rangle$ is the inner product of the coarse basis function with itself and, therefore, is supported on two coarse cells. We compute the diagonal entry using the composite trapezoidal rule in x and midpoint in y

$$a^{cc}_{j,j} = \langle \rho(u_x^c)_j, (u_x^c)_j \rangle = (h_x h_y) \sum_{l=M(j-1)+1}^{M(j+1)} \rho_l (u_x^c)_j^2\big|_l + (h_x h_y) \sum_{l=1}^{n_y} \rho_l^j. \tag{2.32}$$

The indices $M(j-1)+1,\dots,M(j+1)$ in the first term represent the subgrid acceleration nodes inside coarse cells j and $j+1$. The last term on the right corresponds to the boundary between the two cells at which the basis function $(u_x^c)_i$ has value 1, and ρ_l^j are the values of the density on this boundary. ∎

Proof of formula (2.22) *for* U. Let us now use the above results to study the matrix $U = A^{cc} - (A^{cf})^T (A^{ff})^{-1} A^{cf}$, the coefficient matrix of coarse acceleration in (2.21). Performing block matrix multiplications in $(A^{cf})^T (A^{ff})^{-1} A^{cf}$ and using Lemmas 2 and 3, we see that the product $(A^{cf})^T (A^{ff})^{-1} A^{cf}$ is a tridiagonal matrix.

Let us derive a formula for its upper diagonal entries. For convenience of notation, we present the derivation of the first upper diagonal entry $[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{1,2}$. The derivation of the rest of the entries is similar. We use the block notation for A^{cf} introduced in Lemma 2 and the fact that A^{ff} is diagonal (Lemma 3) to obtain

$$[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{1,2} = (A^{cf}_{2,1})^T \begin{bmatrix} (a^{ff}_{M+1,M+1})^{-1} & \dots & 0 \\ \vdots & & \\ 0 & \dots & (a^{ff}_{2M,2M})^{-1} \end{bmatrix} A^{cf}_{2,2}, \tag{2.33}$$

where the blocks $A^{cf}_{2,2}$ and $(A^{cf}_{2,1})^T$ have size $M \times 1$ and $1 \times M$, respectively. The block $A^{cf}_{2,1}$ represents the interaction of the first coarse acceleration basis function $(u^c_x)_1$ with the subgrid acceleration basis functions from the second coarse cell. The explicit formula (2.26) for the entries of A^{cf} gives that

$$(A^{cf}_{2,1})^T = (h_x h_y) \left[\rho_{M+1}(u^c_x)_1\big|_{M+1} \quad \dots \quad \rho_{2M}(u^c_x)_1\big|_{2M} \right], \tag{2.34}$$

where the indices $M+1, \dots, 2M$ are the subgrid acceleration nodes inside the second coarse cell (Figure 2.2). Similarly, the block $A^{cf}_{2,2}$ consists of the inner products of the second coarse acceleration basis function $(u^c_x)_2$ with the subgrid basis functions from the second coarse cell and is given by

$$A^{cf}_{2,2} = (h_x h_y) \begin{bmatrix} \rho_{M+1}(u^c_x)_2\big|_{M+1} \\ \vdots \\ \rho_{2M}(u^c_x)_2\big|_{2M} \end{bmatrix}. \tag{2.35}$$

The explicit formula (2.29) for the entries of A^{ff} and (2.34)–(2.35) result in

$$[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{1,2} = (h_x h_y) \sum_{l=M+1}^{2M} \rho_l (u^c_x)_2\big|_l (u^c_x)_1\big|_l. \tag{2.36}$$

We can obtain a general formula for the $(j, j+1)$-upper diagonal entry:

$$[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{j,j+1} = (h_x h_y) \sum_{l=Mj+1}^{M(j+1)} \rho_l (u^c_x)_{j+1}\big|_l (u^c_x)_j\big|_l. \tag{2.37}$$

Following the same manipulations, we derive the lower diagonal and diagonal entries

$$[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{j,j-1} = (h_x h_y) \sum_{l=M(j-1)+1}^{Mj} \rho_l (u^c_x)_{j-1}\big|_l (u^c_x)_j\big|_l, \tag{2.38}$$

$$[(A^{cf})^T(A^{ff})^{-1}A^{cf}]_{j,j} = (h_x h_y) \sum_{l=M(j-1)+1}^{M(j+1)} \rho_l (u^c_x)^2_j\big|_l. \tag{2.39}$$

Thus, summing the result for entries of A^{cc} from Lemma 4 ((2.30)–(2.32)) with results (2.37)–(2.39) gives us explicit formulas for the entries of U:

$$U_{j,j+1} = 0, \qquad U_{j,j-1} = 0, \qquad U_{j,j} = (h_x h_y) \sum_{l=1}^{n_y} \rho_l^j. \tag{2.40}$$

We see that the matrix U is diagonal and its entries depend only on those values of density which lie on the corresponding boundary of the coarse cell. ∎

Now let us turn our attention to the coefficient matrix of pressure, namely, D. This matrix is defined by

$$D = B^c - (A^{cf})^T (A^{ff})^{-1} B^f. \tag{2.41}$$

In order to write explicit formulas for the entries of D, we need to study the matrices A^{cf}, A^{ff}, B^f, and B^c. We have already discussed the matrices A^{cf} and A^{ff}. The following two lemmas describe the structure of the matrices B^c and B^f.

LEMMA 5. *The matrix B^f is a block diagonal $Kn_y \times Kn_y$ matrix with blocks of size $(n_x - 1) \times n_x$. The blocks of the matrix are given by*

$$T = -\frac{h_x h_y}{h_x} \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ & \dots & \dots & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{2.42}$$

Proof. The entries of the matrix B^f are the inner products of the x derivative of subgrid acceleration basis functions with pressure basis functions, $b^f_{l,m} = \langle w_m, \frac{\partial(\delta u_x)_l}{\partial x} \rangle$. Since the total number of subgrid acceleration basis functions is $Kn_y(n_x - 1)$ and the total number of pressure basis functions is $Kn_y n_x$, the matrix B^f is of size $Kn_y(n_x - 1) \times Kn_y n_x$. The entries of the matrix can be evaluated easily since the subgrid acceleration basis function is piecewise linear in x and constant in y, so its derivative is easily computed

$$\frac{\partial(\delta u_x)_l}{\partial x} = \begin{cases} \dfrac{1}{h_x} & \text{left branch,} \\ -\dfrac{1}{h_x} & \text{right branch.} \end{cases} \tag{2.43}$$

Then $b^f_{l,m} = \pm(h_x h_y)\frac{1}{h_x}$ whenever the pressure basis function and the subgrid acceleration basis function have the same support. Since each subgrid acceleration basis function interacts with only two pressure nodes, one block of the matrix B^f has the form (2.42). Each block corresponds to the interactions of one row of the pressure and acceleration basis functions within a single coarse cell. Since one row of fine cells inside a single coarse cell has $n_x - 1$ subgrid acceleration nodes and n_x pressure nodes, each block T is of size $(n_x - 1) \times n_x$. Finally B^f can be written as the block diagonal matrix with blocks given by T. ∎

LEMMA 6. *The matrix B^c is an upper bidiagonal $(K - 1) \times K$ block matrix. The blocks are of size $1 \times N$ and are given by*

$$B^c_{l,l} = \frac{h_x h_y}{H_x} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}, \qquad B^c_{l,l+1} = -\frac{h_x h_y}{H_x} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}. \tag{2.44}$$

Proof. The entries of the matrix B^c are the inner products of the x derivative of coarse acceleration basis functions and pressure basis functions (2.14). The total number of the coarse acceleration nodes is $(K-1)$ and the total number of the pressure nodes is NK, so B^c has size $(K-1)\times NK$. Each coarse acceleration basis function is supported on two coarse cells and, therefore, has a nonzero inner product only with the pressure basis functions internal to those cells. If we introduce the blocks B^c_{ij} of size $1\times N$ which represent the interaction of the ith coarse basis function with the pressure basis functions from the jth coarse cell, then the matrix B^c can be written as a bidiagonal matrix:

$$B^c = \begin{bmatrix} B^c_{1,1} & B^c_{1,2} & & & 0 \\ & B^c_{2,2} & B^c_{2,3} & & \\ & & \ddots & \ddots & \\ 0 & & & B^c_{K-1,K-1} & B^c_{K-1,K} \end{bmatrix}. \tag{2.45}$$

The entries of B^c can be computed exactly. To evaluate $(b^c)_{l,m} = \langle w_m, \frac{\partial (u^c_x)_l}{\partial x}\rangle$, we use the fact that $(u^c_x)_l$ is a linear function in x and a constant in y. Thus, its derivative is easily computed and is given by

$$\frac{\partial (u^c_x)_l}{\partial x} = \begin{cases} \dfrac{1}{H_x} & l\text{th coarse cell,} \\ -\dfrac{1}{H_x} & (l+1)\text{th coarse cell.} \end{cases} \tag{2.46}$$

The pressure basis function w_m is 1 on the corresponding fine cell and 0 everywhere else. Therefore, we have

$$b^c_{l,m} = \left\langle w_m, \frac{\partial (u^c_x)_l}{\partial x} \right\rangle = \begin{cases} (h_x h_y)\dfrac{1}{H_x} & l\text{th coarse cell,} \\ -(h_x h_x)\dfrac{1}{H_x} & (l+1)\text{th coarse cell.} \end{cases} \tag{2.47}$$

Thus,

$$B^c_{l,l} = \frac{h_x h_y}{H_x}\begin{bmatrix} 1 & \dots & 1 \end{bmatrix}, \qquad B^c_{l,l+1} = -\frac{h_x h_y}{H_x}\begin{bmatrix} 1 & \dots & 1 \end{bmatrix}. \tag{2.48}$$ ∎

Proof of formula (2.23)–(2.24) *for* D. We can use Lemmas 2, 3, 5, and 6 to study the matrix D, which is given by formula (2.41). Performing block multiplication in $(A^{cf})^T(A^{ff})^{-1}B^f$, we see that the resulting matrix is an upper bidiagonal block matrix. Let us derive explicit formulas for the diagonal blocks first. The results obtained in Lemmas 2, 3, and 5 give that

$$\begin{aligned} &[(A^{cf})^T(A^{ff})^{-1}B^f]_{ii} \\ &\quad = \begin{bmatrix} (u^c_x)_i\big|_{(M(i-1)+1}, & (u^c_x)_i\big|_{M(i-1)+2}, & \dots & (u^c_x)_i\big|_{Mi} \end{bmatrix} \begin{bmatrix} T & & & 0 \\ & T & & \\ & & \ddots & \\ 0 & & & T \end{bmatrix}, \end{aligned} \tag{2.49}$$

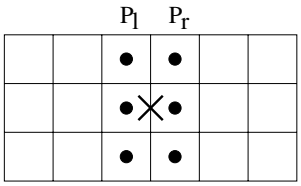


FIG. 2.6. *The nodes of coarse acceleration and pressure involved in the difference equation.*

where $(u_x^c)_i\big|_l$, $l = M(i-1)+1, M(i-1)+2, \ldots, Mi$, are the values of the coarse basis function at the nodal points of subgrid acceleration inside coarse cell i. Note that these values can be easily computed since the function $(u_x^c)_i$ is known (it is linear in x and constant in y, and it takes value 1 at the corresponding node and 0 at all other nodes).

After further matrix multiplications and application of Lemma 6, we see that D is block upper bidiagonal with the blocks given by (2.23)–(2.24). The nonzero entries correspond to those pressure nodes that are located along the boundary of the coarse cell i. □

2.3. Difference and differential equations for coarse acceleration. The results of Theorem 1 allow us to write the difference equation for the coarse acceleration explicitly as

$$(h_x h_y)\sum_{l=1}^{n_y}\rho_l^i(\mathbf{v_x^c})_i = -(h_x h_y)\left[\sum_{l=1}^{n_y}\frac{\mathbf{p}_{Ni+n_x(l-1)+1}}{h_x} - \sum_{l=1}^{n_y}\frac{\mathbf{p}_{N(i-1)+n_x l}}{h_x}\right]. \tag{2.50}$$

Notice that the two sums on the right-hand side of (2.50) are the sums of the pressure values to the right and to the left of the coarse cell boundary on which coarse acceleration $(\mathbf{v_x^c})_i$ is located (Figure 2.6). If we denote the average of pressure along the right and left boundary, respectively, by

$$\bar{\mathbf{p}}_r \equiv \frac{\sum_{l=1}^{n_y}\mathbf{p}_{Ni+n_x(l-1)+1}}{n_y}, \qquad \bar{\mathbf{p}}_l \equiv \frac{\sum_{l=1}^{n_y}\mathbf{p}_{N(i-1)+n_x l}}{n_y}, \tag{2.51}$$

then (2.50) reduces to

$$\bar{\rho}(\mathbf{v_x^c})_i = -\frac{\bar{\mathbf{p}}_r - \bar{\mathbf{p}}_l}{h_x}, \tag{2.52}$$

where $\bar{\rho}$ is the average of the density values on the boundary of the coarse cell, i.e.,

$$\bar{\rho} = \frac{\sum_{l=1}^{n_y}\rho_l^i}{n_y}.$$

Let us define a new function ρ^{ups}, which we will call upscaled density. This function is defined to be the original density values at node points interior to the coarse cells and is an average of density values $\bar{\rho}$ at node points along coarse block edges. In the following theorem, we derive the continuous differential equation for v_x^c and p.

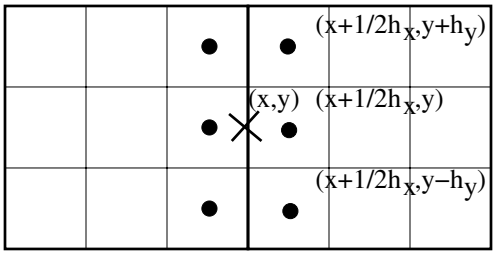


FIG. 2.7. *The coordinates of coarse acceleration and pressure involved in the difference equation in the case of* 3×3 *fine cells inside a single coarse cell.*

THEOREM 7. *Assume pressure* p *is at least four times continuously differentiable. Then the difference equation* (2.52) *is a discretization of the following differential equation*

$$\rho^{ups} v_x^c = -\frac{\partial p}{\partial x},$$

with order of approximation $O(h_x^2 + h_y^2)$.

Note. Theorem 7 gives a continuous differential equation for coarse acceleration which is essentially (2.1) from the original first-order system with density replaced by an upscaled density. We see that the upscaled acceleration approximates the gradient of pressure on the boundaries of coarse cells and, thus, compensates for the simplifying zero boundary conditions used in the algorithm.

Proof. Let acceleration node $(\mathbf{v_x^c})_\mathbf{i}$ be associated with the point (x, y). Replace $(\mathbf{v_x^c})_\mathbf{i}$, $\mathbf{p}_{Ni+n_x(l-1)+1}$, and $\mathbf{p}_{N(i-1)+n_x l}$ in the difference equation by the smooth functions v_x^c and p at the corresponding points and expand each pressure term in a Taylor series about the point (x, y). Notice that the pressure terms in $\bar{\mathbf{p}}_r$ are located at $(x + \frac{1}{2}h_x, y \pm jh_y)$, where jh_y is the distance between the corresponding pressure node and the acceleration node in the y direction. For example, in the case of 3×3 fine cells inside a single coarse cell, we see in Figure 2.7 that the terms in the sum $\bar{\mathbf{p}}_r$ are located at $(x + \frac{1}{2}h_x, y + h_y)$, $(x + \frac{1}{2}h_x, y)$, and $(x + \frac{1}{2}h_x, y - h_y)$. Thus, the subscript j takes on values $j = 0, 1$. Similarly, the pressure terms in $\bar{\mathbf{p}}_l$ are located at $(x - \frac{1}{2}h_x, y \pm jh_y)$. We expand each pressure term in (2.51) in a fourth-order Taylor series about (x, y). Since the pressure nodes are symmetric about the point (x, y), the linear and cubic terms in h_y cancel when computing the sums in (2.51), and we obtain for $\bar{\mathbf{p}}_r$ and $\bar{\mathbf{p}}_l$

(2.53)
$$\bar{\mathbf{p}}_r = \frac{\displaystyle\sum_{l=1}^{n_y} \left[p + \frac{\partial p}{\partial x}\left(\frac{1}{2}h_x\right) + \frac{1}{2}\left[\frac{\partial^2 p}{\partial x^2}\left(\frac{1}{2}h_x\right)^2 + \frac{\partial^2 p}{\partial y^2}(jh_y)^2\right]\right]}{n_y} + O(h_x^3) + O(h_x h_y^2)$$

and

(2.54)
$$\bar{\mathbf{p}}_l = \frac{\displaystyle\sum_{l=1}^{n_y} \left[p - \frac{\partial p}{\partial x}\left(\frac{1}{2}h_x\right) + \frac{1}{2}\left[\frac{\partial^2 p}{\partial x^2}\left(\frac{1}{2}h_x\right)^2 + \frac{\partial^2 p}{\partial y^2}(jh_y)^2\right]\right]}{n_y} + O(h_x^3) + O(h_x h_y^2),$$

where $p \equiv p(x, y)$. Substituting (2.53) and (2.54) into the difference equation (2.52) results in the following differential equation:

$$\rho^{ups} v_x^c = -\frac{\partial p}{\partial x}, \tag{2.55}$$

up to order $O(h_x^2 + h_y^2)$. We conclude that the upscaled acceleration approximates the x derivative of pressure on the boundary of the coarse cell with the upscaled density given by $\bar{\rho}$ at each nodal point. ∎

2.4. Matrix analysis of the pressure equation. In the next two sections, we will explain how coarse acceleration determines pressure in the time-dependent equation ((2.6) of the subgrid problem). The upscaling algorithm modifies the original wave equation on the boundaries of the coarse cells by using the coarse acceleration and the averaged density values on these boundaries. The equation which is solved includes an additional cross-derivative term not seen in the original wave equation. We follow the same basic outline as in sections 2.2–2.3. In other words, we start by analyzing the matrix form of (2.6) and then make use of results from the previous sections to derive the matrix equation for coarse acceleration and pressure. These results are then used to obtain a corresponding difference and finally differential equation for pressure resulting from the upscaling algorithm.

THEOREM 8. *Consider the set of subgrid problems* (2.5)–(2.6). *The elimination of the subgrid unknowns from the pressure equation* (2.6) *results in the following linear system for the coarse acceleration and pressure*

$$W \frac{\partial^2 \mathbf{p}}{\partial t^2} = -D^T \mathbf{v_x^c} - C\mathbf{p} - (v_y \text{ terms}) + F.$$

The matrix D and vector F are defined in Theorem 1 *by* (2.23), (2.24), *and* (2.14). *The matrix W is diagonal with entries having values of $\frac{1}{\rho c^2}$ at the corresponding pressure nodes. The matrix C is a block diagonal matrix. The blocks of C are of size $n_x \times n_x$ and are given by*

$$C_{i,i} = \frac{h_x h_y}{h_x^2} \begin{bmatrix} \rho_{I+1}^{-1} & -\rho_{I+1}^{-1} & & & 0 \\ -\rho_{I+1}^{-1} & \rho_{I+1}^{-1} + \rho_{I+2}^{-1} & -\rho_{I+2}^{-1} & & \\ \ddots & \ddots & \ddots & & \\ & & & & \\ & -\rho_{I+n_x-2}^{-1} & \rho_{I+n_x-2}^{-1} + \rho_{I+n_x-1}^{-1} & -\rho_{I+n_x-1}^{-1} & \\ 0 & & -\rho_{I+n_x-1}^{-1} & \rho_{I+n_x-1}^{-1} & \end{bmatrix}, \tag{2.56}$$

where $I = (i-1)(n_x - 1)$.

Proof. The idea of the proof is similar to that of Theorem 1. First, we use the matrix form of the subgrid problem (2.15)–(2.16) to eliminate the subgrid unknowns from the pressure equation (2.16). We obtain the time-dependent matrix equation for pressure and coarse acceleration only and derive explicit formulas for the matrix entries.

Recall that the matrix form of the subgrid problem is given by (2.15)–(2.16). We see that the first equation can be solved easily for $\delta \mathbf{v_x}$ in terms of $\mathbf{v_x^c}$. Thus, we can eliminate the subgrid unknowns from the time-dependent pressure equation. We

obtain

(2.57)
$$W\frac{\partial^2 \mathbf{p}}{\partial t^2} = -[(B^c)^T - (B^f)^T(A^{ff})^{-1}A^{cf}]\mathbf{v}_\mathbf{x}^\mathbf{c} - (B^f)^T(A^{ff})^{-1}B^f\mathbf{p} - (v_y \text{ terms}) + F,$$

or

$$W\frac{\partial^2 \mathbf{p}}{\partial t^2} = -D^T\mathbf{v}_\mathbf{x}^\mathbf{c} - C\mathbf{p} - (v_y \text{ terms}) + F. \tag{2.58}$$

In the rest of the proof of the theorem, we discuss the structure and entries of the matrices W, D^T, and C. In what follows, if i is the ith pressure node, then $\frac{1}{\rho c^2}\big|_i$ denotes the value of $\frac{1}{\rho c^2}$ at that node.

LEMMA 9. *The matrix W is a diagonal $NK \times NK$ matrix with nonzero entries given by*

$$w_{i,i} = (h_x h_y)\frac{1}{\rho c^2}\Big|_i. \tag{2.59}$$

Proof. The entries of the matrix W are the inner products of pressure basis functions with themselves, $\langle \frac{1}{\rho c^2} w_i, w_j \rangle$. Computing these entries using the midpoint rule and noting that each pressure basis function is supported on one fine cell only gives the desired result. ∎

Completion of the proof of Theorem 8. Let us now consider the matrix for coarse acceleration D^T. Notice that the matrix D was defined in Theorem 1 (2.23)–(2.24). Using those results, we see that D^T is a lower block bidiagonal matrix with blocks given by $D^T_{i,i+1}$ and $D^T_{i,i}$. The nonzero entries in D^T are associated with pressure unknowns located along the vertical boundaries of the coarse cells.

Let us now turn our attention to the coefficient matrix for pressure

$$C = (B^f)^T(A^{ff})^{-1}B^f.$$

The structure and the entries of A^{ff} and B^f were discussed in Lemmas 3 and 5. We have shown that A^{ff} is a diagonal matrix, and B^f is a block diagonal matrix with blocks of size $(n_x - 1) \times n_x$. Thus, C is also a block diagonal matrix with blocks of size $n_x \times n_x$.

We can derive explicit formulas for the entries of C. For simplicity of notation, let us discuss the derivation of the first block $C_{1,1}$. Using the explicit formulas for the entries of A^{ff} and the blocks of B^f in (2.29) and (2.42), we obtain

(2.60) $C_{1,1}$

$$= \frac{h_x h_y}{h_x^2}\begin{bmatrix} -1 & 0 & \dots & 0 \\ 1 & -1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & \dots & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix}\begin{bmatrix} \rho_1^{-1} & & & 0 \\ & \ddots & & \\ 0 & & & \rho_{n_x-1}^{-1} \end{bmatrix}\begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ & & \dots & \dots & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

$$= \frac{h_x h_y}{h_x^2}\begin{bmatrix} \rho_1^{-1} & -\rho_1^{-1} & 0 & \dots & 0 & 0 & 0 \\ -\rho_1^{-1} & \rho_1^{-1} + \rho_2^{-1} & -\rho_2^{-1} & \dots & 0 & 0 & 0 \\ & & & \dots & & & \\ 0 & 0 & 0 & \dots & -\rho_{n_x-2}^{-1} & \rho_{n_x-2}^{-1} + \rho_{n_x-1}^{-1} & -\rho_{n_x-1}^{-1} \\ 0 & 0 & 0 & \dots & 0 & -\rho_{n_x-1}^{-1} & \rho_{n_x-1}^{-1} \end{bmatrix}.$$

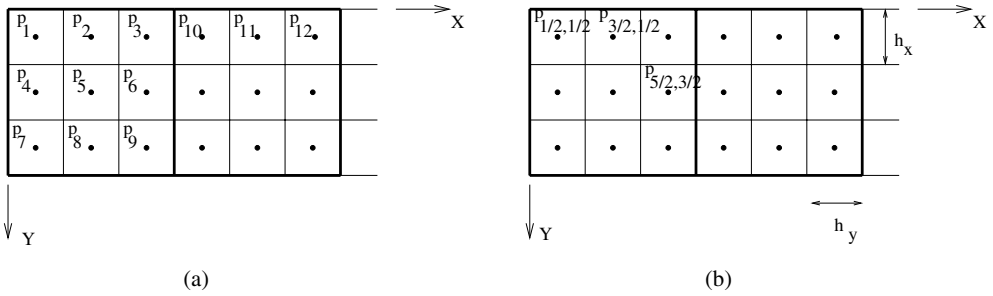


FIG. 2.8. *Pressure unknowns in the case of* 3×3 *fine cells inside a single coarse cell:* (a) *vector notation,* (b) *coordinate notation.*

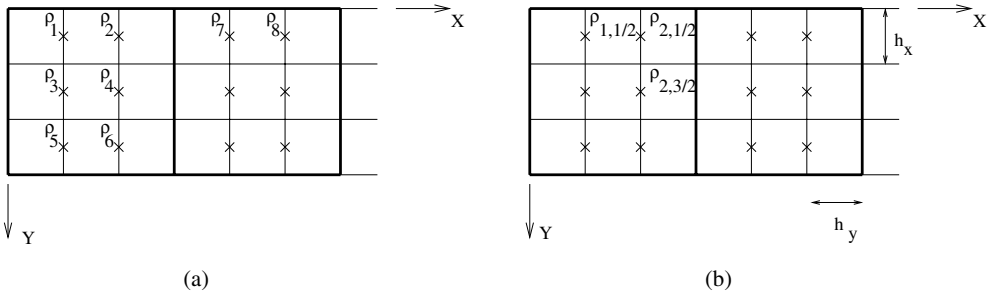


FIG. 2.9. *Density values at subgrid velocity nodes in the case of* 3×3 *fine cells inside a single coarse cell:* (a) *vector notation,* (b) *coordinate notation.*

Similarly, we can compute the rest of the blocks $C_{i,i}$ to get (2.56). Recall that blocks of C have size $n_x \times n_x$, so each block is associated with one row of pressure unknowns inside a single coarse cell. □

2.5. Difference and differential equations for pressure. To better understand the meaning of the matrix problem (2.58), we interpret it as a difference equation first and then as a continuous differential equation. In the following theorems, it is beneficial to change from vector notation (Figures 2.8(a) and 2.9(a)), used in the previous sections, to a spatial coordinate notation (Figures 2.8(b) and 2.9(b)). This notation change simplifies the discussion of the difference equation. In coordinate notation, $p_{i+1/2,j+1/2}$ will denote the value of pressure at the grid point $(x_{i+1/2} = (i + 1/2)h_x$, $y_{j+1/2} = (j + 1/2)h_y)$, and $\rho_{i,j+1/2}$ will be the value of density at the grid point $(x_i = ih_x,\ y_{j+1/2} = (j + 1/2)h_y)$, which is associated with the subgrid acceleration node (Figure 2.9). In the following theorems, we see that the matrix problem gives rise to different difference equations and hence different differential equations at different points in the spatial grid. The three pressure node locations we need to consider are (a) nodes internal to the coarse cell, (b) nodes along the right boundary of the coarse cell, and (c) nodes along the left boundary of the coarse cell (Figure 2.10).

In Theorem 10, we derive the difference and differential equations for the internal pressure nodes (Figure 2.10(a)). We then derive the difference and differential equations for the pressure nodes along the right boundary in Theorem 11. The result for pressure nodes along the left boundary of the coarse cells is similar.

THEOREM 10. *Let* $p_{i+1/2,j+1/2}$ *be a pressure unknown internal to a particular coarse cell (Figure* 2.10(a)*). Then the difference equation corresponding to matrix*

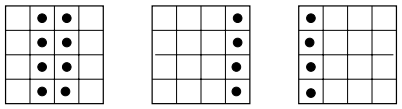


FIG. 2.10. *Three pressure node locations inside one coarse cell:* (a) *internal pressure nodes,* (b) *pressure nodes along the right boundary of the coarse cell,* (c) *pressure nodes along the left boundary of the coarse cell.*

problem (2.58) *has the following form:*

$$\begin{aligned}(2.61)\quad &\frac{1}{\rho c^2}\frac{\partial^2 p_{i+1/2,j+1/2}}{\partial t^2}\\ &= -\frac{1}{h_x^2}\Big(- p_{i+3/2,j+1/2}\rho^{-1}_{i+1,j+1/2} + p_{i+1/2,j+1/2}\left(\rho^{-1}_{i+1,j+1/2} + \rho^{-1}_{i,j+1/2}\right)\\ &\qquad\qquad - p_{i-1/2,j+1/2}\rho^{-1}_{i,j+1/2}\Big)\\ &\quad - \frac{1}{h_y^2}\Big(- p_{i+1/2,j+3/2}\rho^{-1}_{i+1/2,j+1} + p_{i+1/2,j+1/2}\left(\rho^{-1}_{i+1/2,j+1} + \rho^{-1}_{i+1/2,j}\right)\\ &\qquad\qquad - p_{i+1/2,j-1/2}\rho^{-1}_{i+1/2,j}\Big) + f_{i+1/2,j+1/2}.\end{aligned}$$

Further, assume pressure p *and density* ρ *are at least four times continuously differentiable. Then difference equation* (2.61) *is a discretization of the following continuous differential equation:*

$$(2.62)\qquad \frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial}{\partial x}\left(\rho^{-1}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\rho^{-1}\frac{\partial p}{\partial y}\right) + f,$$

with order of approximation $O(h_x^2 + h_y^2)$.

Proof. Recall that the matrix equation for the pressure and coarse acceleration is given by (2.58). For simplicity, let us consider a particular case of 3×3 fine cells inside a single coarse cell (that is, $n_x = n_y = 3$), and we will focus initially on a particular pressure unknown, $p_{3/2,1/2}$, internal to the first coarse cell (Figure 2.8). Notice that in this case each coarse cell contains only three internal pressure unknowns. We chose $p_{3/2,1/2}$, since it is the only internal unknown in the first row of the first coarse cell. We first derive a difference equation for this unknown and then generalize the resulting formula. Since W is diagonal and its entries are given by (2.59) (Lemma 9), the product $W\frac{\partial^2 \mathbf{p}}{\partial t^2}$ yields

$$(2.63)\qquad (h_x h_y)\frac{1}{\rho c^2}\Big|_{3/2,1/2}\frac{\partial^2 p_{3/2,1/2}}{\partial t^2}.$$

As was shown in Theorem 8, the corresponding row of the matrix D^T contains zeroes. Therefore, the difference equation does not involve the coarse acceleration nodes. Let us now consider the term $C\mathbf{p}$. Since the unknown of interest is located in the first row of the first coarse cell, we need to consider the block $C_{1,1}$ (2.60). In the case where

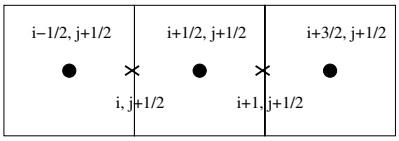


FIG. 2.11. *Finite-difference stencil for the internal pressure unknowns. Note that pressure nodes are denoted by circles and acceleration nodes by x's. Density lives at acceleration nodes, sound velocity at pressure nodes.*

$n_x = n_y = 3$, $C_{1,1}$ becomes

$$C_{1,1} = \frac{h_x h_y}{h_x^2} \begin{bmatrix} \rho_{1,1/2}^{-1} & -\rho_{1,1/2}^{-1} & 0 \\ -\rho_{1,1/2}^{-1} & \rho_{1,1/2}^{-1} + \rho_{2,1/2}^{-1} & -\rho_{2,1/2}^{-1} \\ 0 & -\rho_{2,1/2}^{-1} & \rho_{2,1/2}^{-1} \end{bmatrix}, \tag{2.64}$$

where we have used coordinate notation (Figure 2.9). The pressure unknown $p_{3/2,1/2}$ corresponds to the second entry in the vector $\mathbf{p}$. Thus, multiplying the second row of the matrix C by $\mathbf{p}$, we obtain a sum of three nonzero terms

$$\frac{h_x h_y}{h_x^2} \left(-p_{1/2,1/2} \rho_{1,1/2}^{-1} + p_{3/2,1/2} \left(\rho_{2,1/2}^{-1} + \rho_{1,1/2}^{-1} \right) - p_{5/2,1/2} \rho_{2,1/2}^{-1} \right). \tag{2.65}$$

Putting (2.63) and (2.65) together gives

$$\begin{aligned} \frac{1}{\rho c^2}\Big|_{3/2,1/2} \frac{\partial^2 p_{3/2,1/2}}{\partial t^2} = \frac{1}{h_x^2} \Big(& -p_{1/2,1/2} \rho_{1,1/2}^{-1} + p_{3/2,1/2} \left(\rho_{2,1/2}^{-1} + \rho_{1,1/2}^{-1} \right) - p_{5/2,1/2} \rho_{2,1/2}^{-1} \Big) \\ & - (v_y \text{ terms}) + f_{3/2,1/2}. \end{aligned} \tag{2.66}$$

The above formula can be generalized to the case of $n_x \times n_y$ fine cells inside a coarse cell. Since the structure of the matrices does not change, the same steps will lead us to the following difference equation for internal pressure unknowns (see Figure 2.11 for the x-derivative finite difference stencil):

$$\begin{aligned} & \frac{1}{\rho c^2} \frac{\partial^2 p_{i+1/2,j+1/2}}{\partial t^2} \\ & = -\frac{1}{h_x^2} \Big(- p_{i+3/2,j+1/2} \rho_{i+1,j+1/2}^{-1} + p_{i+1/2,j+1/2} \left(\rho_{i+1,j+1/2}^{-1} + \rho_{i,j+1/2}^{-1} \right) \\ & \qquad - p_{i-1/2,j+1/2} \rho_{i,j+1/2}^{-1} \Big) - (v_y \text{ terms}) + f_{i+1/2,j+1/2}. \end{aligned} \tag{2.67}$$

The difference expression for the v_y terms is similar.

Notice that the first term on the right-hand side of (2.67) is the standard second-order centered finite difference approximation of $\frac{\partial}{\partial x}(\rho^{-1} \frac{\partial p}{\partial x})$ [11]. Expanding the pressure and density terms in Taylor series around the point $(x_{i+1/2}, y_{j+1/2})$, we obtain from (2.67)

$$\frac{1}{\rho c^2} \frac{\partial^2 p}{\partial t^2} = \frac{\partial}{\partial x} \left(\rho^{-1} \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial y} \left(\rho^{-1} \frac{\partial p}{\partial y} \right) + f + O(h_x^2 + h_y^2). \tag{2.68}$$

This shows that the difference equation for the pressure unknowns internal to the coarse cell, which results from the upscaling algorithm, approximates the standard continuous acoustic wave equation up to order $O(h_x^2 + h_y^2)$. □

THEOREM 11. *Let $p_{i+1/2,j+1/2}$ be a pressure unknown located along the right boundary of a particular coarse cell (Figure* 2.10(b)*). Then the difference equation corresponding to* (2.58) *has the following form:*

(2.69)
$$\begin{aligned}
&\frac{1}{\rho c^2}\frac{\partial^2 p_{i+1/2,j+1/2}}{\partial t^2} \\
&= -\frac{1}{h_x}v_x^c - \frac{1}{h_x^2}\left(p_{i+1/2,j+1/2}\rho^{-1}_{i,j+1/2} - p_{i-1/2,j+1/2}\rho^{-1}_{i,j+1/2}\right) \\
&\quad -\frac{1}{h_y^2}\Big(- p_{i+1/2,j+3/2}\rho^{-1}_{i+1/2,j+1} + p_{i+1/2,j+1/2}\left(\rho^{-1}_{i+1/2,j+1} + \rho^{-1}_{i+1/2,j}\right) \\
&\qquad\qquad -p_{i+1/2,j-1/2}\rho^{-1}_{i+1/2,j}\Big) + f_{i+1/2,j+1/2},
\end{aligned}$$

where v_x^c is the value of the coarse acceleration on the given boundary.

Further, assume pressure p and density ρ are at least four times continuously differentiable. Then difference equation (2.69) *is a discretization of the following continuous differential equation with order of approximation $O(h_x + h_y)$:*

(2.70)
$$\frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial}{\partial x}\left((\rho^{ups})^{-1}\frac{\partial p}{\partial x}\right) + \left((\rho^{ups})^{-1}\frac{\partial^2 p}{\partial x \partial y}\right)K + \frac{\partial}{\partial y}\left((\rho^{ups})^{-1}\frac{\partial p}{\partial y}\right) + f,$$

where K is a constant which depends on the location of the pressure node within a single coarse cell.

Proof. The proof of the theorem is similar to that of Theorem 10. We first consider the case of 3×3 fine cells inside a single coarse cell and derive the difference equation for pressure unknown, $p_{5/2,1/2}$ located along the right boundary of the coarse cell (Figure 2.8(b)). The pressure unknown $p_{5/2,1/2}$ corresponds to the third entry in the vector $\mathbf{p}$ (Figure 2.8). Therefore, we need to look at the third rows of the matrices D^T and C. Only the first entry of the third row of D^T is nonzero (Theorems 1 and 8). The third row of the block $C_{1,1}$ in coordinate notation is given by (2.64). Thus, performing matrix-vector multiplications in matrix problem (2.58), we obtain

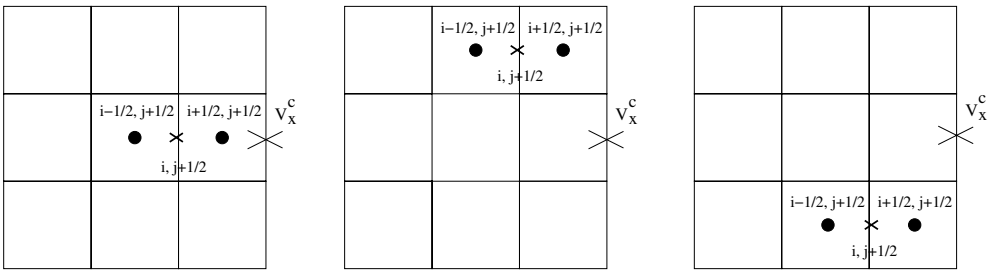


FIG. 2.12. *Locations of pressure nodes relative to coarse velocity for the case of 3×3 fine cells inside a single coarse cell. Note that pressure nodes are denoted by circles and subgrid and coarse acceleration nodes by x's and X's, respectively. Density lives at the same location as the subgrid acceleration nodes.*

the difference equation for $p_{5/2,1/2}$:

$$\frac{1}{\rho c^2}\frac{\partial^2 p_{5/2,1/2}}{\partial t^2} = -\frac{1}{h_x}v_x^c - \frac{1}{h_x^2}\left(p_{5/2,1/2}\rho_{2,1/2}^{-1} - p_{3/2,1/2}\rho_{2,1/2}^{-1}\right) - (v_y \text{ terms}) + f_{5/2,1/2}. \tag{2.71}$$

The above formula can be generalized to the case of $n_x \times n_y$ fine cells inside a coarse cell. We obtain the following difference equation for pressure unknowns along the boundary of the coarse cell (see Figure 2.12)

$$\frac{1}{\rho c^2}\frac{\partial^2 p_{i+1/2,j+1/2}}{\partial t^2} = -\frac{1}{h_x}v_x^c - \frac{1}{h_x^2}\left(p_{i+1/2,j+1/2}\rho_{i,j+1/2}^{-1} - p_{i-1/2,j+1/2}\rho_{i,j+1/2}^{-1}\right) - (v_y \text{ terms}) + f_{i+1/2,j+1/2}, \tag{2.72}$$

where v_x^c is the coarse acceleration unknown on the given boundary.

Let us now derive the differential equation. The idea is to use Taylor expansions around the point $(x_{i+1/2}, y_{j+1/2})$. First, consider the term v_x^c on the right-hand side of (2.72). We have shown in the previous sections that coarse acceleration is related to pressure through the difference equation

$$v_x^c = -\frac{1}{\bar\rho}\frac{\bar p_r - \bar p_l}{h_x}. \tag{2.73}$$

The terms $\bar p_r$, $\bar p_l$ are the averaged sums of pressure unknowns to the right and left of a particular boundary of the coarse cell, and $\bar\rho$ is the average of density values on the same boundary. We can write $\bar p_r$, $\bar p_l$, and $\bar\rho$ using coordinate notation as

$$\bar p_r = \frac{\sum_k p_{i+3/2,j+1/2+k}}{n_y}, \qquad \bar p_l = \frac{\sum_k p_{i+1/2,j+1/2+k}}{n_y}, \tag{2.74}$$

$$\bar\rho = \frac{\sum_k \rho_{i+1,j+1/2+k}}{n_y}, \tag{2.75}$$

where $p_{i+3/2,j+1/2+k} = p(h_x(i+3/2), h_y(j+1/2+k))$, $\rho_{i+1,j+1/2+k} = \rho(h_x(i+1), h_y(j+1/2+k))$, and kh_y represents the distance between the unknown and the point $(x_{i+1/2}, y_{j+1/2})$ in the y direction. The number of terms in each sum is equal to n_y, the number of fine cells inside a single coarse cell in the y direction. The values that k takes will depend on the number of fine cells inside a single coarse cell and the location of a particular pressure unknown in that cell. For example, Figure 2.13 shows that in the 3×3 case, if the pressure unknowns of interest are in bold, then k may take values -1, 0, 1 (Figure 2.13(a)); or the values 0, 1, 2 (Figure 2.13(b)); or 0, -1, -2 (Figure 2.13(c)). Expanding all the pressure unknowns in $\bar p_r$, $\bar p_l$ in a fourth-order Taylor series about the point $(x_{i+1/2}, y_{j+1/2})$ and using (2.73), we obtain for coarse acceleration

$$v_x^c = -\frac{1}{\bar\rho}\left[\frac{\partial p}{\partial x} + \frac{h_x}{2}\frac{\partial^2 p}{\partial x^2} + \frac{\sum_k k}{n_y}h_y\frac{\partial^2 p}{\partial x \partial y}\right] + O(h_x^2 + h_x h_y + h_y^2), \tag{2.76}$$

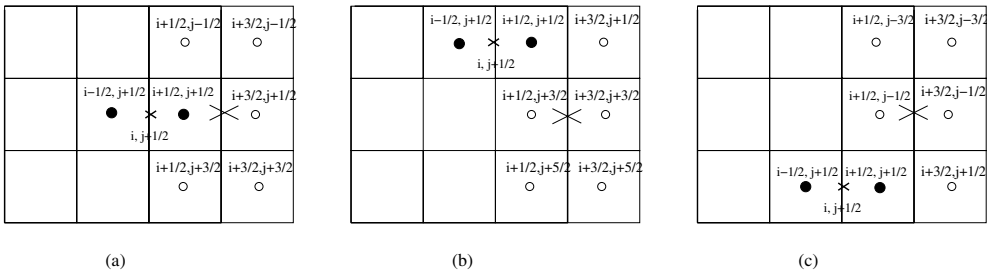


FIG. 2.13. *Examples of finite-difference stencils for different positions of pressure unknown* $p_{i+1/2,j+1/2}$ *in the case of* 3×3 *fine cells inside a single coarse cell. The open circles denote the pressure unknowns used in the calculation of* $\bar{p}_r$ *and* $\bar{p}_l$.

where $p \equiv p(x_{i+1/2}, y_{j+1/2})$. Expanding the rest of the terms on the right-hand side of (2.72) around the same point, we obtain

$$
\begin{aligned}
(2.77)\quad & -\frac{1}{h_x} v_x^c - \frac{1}{h_x^2}\left(p_{i+1/2,j+1/2}\rho^{-1}_{i,j+1/2} - p_{i-1/2,j+1/2}\rho^{-1}_{i,j+1/2}\right) \\
&\qquad = \frac{1}{\bar{\rho}}\left(\frac{1}{h_x}\frac{\partial p}{\partial x} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2} + \frac{\sum\limits_k k}{n_y}\frac{h_y}{h_x}\frac{\partial^2 p}{\partial x \partial y}\right) \\
&\qquad\quad - \left(\frac{1}{h_x}\frac{\partial p}{\partial x} - \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\right) \times \left(\rho^{-1} - \frac{h_x}{2}\frac{\partial \rho^{-1}}{\partial x}\right) + O(h_x + h_y),
\end{aligned}
$$

where $\rho \equiv \rho(x_{i+1/2}, y_{j+1/2})$. We can use the definition of the new function ρ^{ups} to say that

$$
(2.78)\qquad \frac{1}{\bar{\rho}} = \left(\rho^{ups}_{i+1,j+1/2}\right)^{-1}.
$$

Suppose the function ρ^{ups} can be constructed in such a way that it is smooth enough for Taylor series expansion. Expanding $(\rho^{ups}_{i+1,j+1/2})^{-1}$ around the point $(x_{i+1/2}, y_{j+1/2})$, we obtain from (2.77)

$$
\begin{aligned}
(2.79)\quad & -\frac{1}{h_x} v_x^c - \frac{1}{h_x^2}\left(p_{i+1/2,j+1/2}\rho^{-1}_{i,j+1/2} - p_{i-1/2,j+1/2}\rho^{-1}_{i,j+1/2}\right) \\
&\qquad = \frac{\partial p}{\partial x}\frac{\partial (\rho^{ups})^{-1}}{\partial x} + \frac{\partial^2 p}{\partial x^2}(\rho^{ups})^{-1} + \frac{\sum\limits_k k}{n_y}\frac{h_y}{h_x}\frac{\partial^2 p}{\partial x \partial y}(\rho^{ups})^{-1} + O(h_x + h_y).
\end{aligned}
$$

We use (2.79) in (2.72) to obtain the following differential equation

$$
(2.80)\qquad \frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} \;=\; \frac{\partial}{\partial x}\left((\rho^{ups})^{-1}\frac{\partial p}{\partial x}\right) + \left((\rho^{ups})^{-1}\frac{\partial^2 p}{\partial x \partial y}\right) K + (v_y \text{ terms}) + f,
$$

$$
\text{where } K = \frac{\sum\limits_k k}{n_y}\frac{h_y}{h_x}.
$$

Note. The constant K depends on the size of the fine mesh, the size of the coarse mesh, and the position of the pressure node under consideration. In particular, $K = 0$ when the pressure node has the same y coordinate as the coarse acceleration node. In this situation, k takes values $-1, 0, 1$, so that $\sum_k k = 0$ (see Figure 2.13(a)). ∎

3. Conclusions. To model subsurface phenomena ranging from the centimeter to the kilometer scale (micro- to macroscale), requires enormous amounts of computing power. Resolving all fine-scale features over large sections of the earth (at depths ranging from the near-surface down to the deep crust) is computationally prohibitive. Upscaling techniques allow us to perform these simulations on a coarser scale while capturing some of this fine-scale subwavelength information. There are a variety of upscaling methods. However, most of these techniques have been developed in the context of elliptic equations. We have adapted the operator-based upscaling technique, previously developed for elliptic flow, to the variable density, variable sound velocity acoustic wave equation. The upscaling method relies on decomposing the space of unknowns into coarse and subgrid subspaces. The problem is then naturally solved in two steps. First, we solve the subgrid problems for fine-scale information internal to each coarse cell. Then we use the subgrid solutions to augment the coarse-scale operator. A simplifying zero boundary condition imposed on each coarse cell decouples the subgrid problems from one coarse cell to the next. The fine-grid input parameters (density and sound velocity) are used throughout the computations. The algorithm does not explicitly average these input parameters. Further, separation of scales is not assumed with this technique. The numerical implementation of the upscaling algorithm for the wave equation is discussed in detail in Vdovina et al. [17]. Numerical experiments presented in that paper indicate that operator-based upscaling models wave propagation (even at the subwavelength scale) quite accurately relative to full finite difference solutions.

In this paper we convert the second-order acoustic wave equation into a system of two first-order equations (first-order in space) which involve solving for both pressure and its gradient (acceleration). The algorithm is based conceptually on the mixed finite element method. However, the pressure equation is solved via finite differences due to an equivalence between finite elements and finite differences. The first practical result from this analysis is that the system matrix for coarse acceleration is diagonal which greatly simplifies the implementation of the method.

Even more importantly, the analysis presented in this paper gives the first explanation of exactly which physical equations are solved by the upscaling algorithm. What we have shown is that the upscaling algorithm produces a coarse solution to the original constitutive equation for acceleration with the input density field redefined as an averaged density along coarse-block edges. This result indicates that the algorithm compensates for the simplifying zero boundary conditions on coarse block edges. Similarly, the upscaling algorithm leaves the wave equation for pressure untouched at nodes internal to coarse blocks. However, the pressure equation solved on coarse cell edges is modified to include a cross-derivative term for pressure (a second derivative involving both x and y) — a form of diffusion. This analysis allows us to simplify the algorithmic implementation of the method and to gain an understanding of what the solution produced by this technique models physically.

Acknowledgments. We thank Alan Levander and Bill Symes of Rice University for asking the extremely pertinent question, "What does the upscaled solution mean?" We further thank Bill Symes for suggesting that we look at the upscaling algorithm as a linear algebra problem. His insight led to the work described in this

paper. We thank Tetyana Vdovina of UMBC for her help with the derivation of the continuous differential equations for pressure from the difference equations. Finally, we are grateful to John Zweck of UMBC for reading an early draft of the manuscript. His suggestions greatly improved the overall structure of the paper.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.
[2] T. Arbogast and K. Boyd, *Subgrid upscaling and mixed multiscale finite elements*, SIAM J. Numer. Anal., to appear.
[3] T. Arbogast and S. Bryant, *A two-scale numerical subgrid technique for waterflood simulations*, SPE J., 27 (2002), pp. 446–457.
[4] T. Arbogast, S. Minkoff, and P. Keenan, *An operator-based approach to upscaling the pressure equation*, Comput. Methods in Water Resources XII, 1 (1998), pp. 405–412.
[5] T. Arbogast, *Numerical Subgrid Upscaling of Two-Phase Flow in Porous Media*, Lect. Notes Phys. 552, Springer, Berlin, 2000.
[6] T. Arbogast, *Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase flow*, Comput. Geosci., 6 (2002), pp. 453–48.
[7] A. Bensoussan, J. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, North-Holland, Amsterdam, 1979.
[8] D. Bergman, J. Lions, G. Papanicolaou, F. Murat, L. Tartar, and E. Sanchez-Palencia, *Les Méthodes de L'homogénéisation: Théorie et Applications en Physique*, Editions Eyrolles, Paris, 1985.
[9] Z. Chen and T. Hou, *A mixed multiscale finite element method for elliptic problems with oscillating coefficients*, Math. Comp., 72 (2002), pp. 541–576.
[10] M. Christie, *Upscaling for reservoir simulation*, J. Pet. Tech., 48 (1996), pp. 1004–1010.
[11] G. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer-Verlag, New York, 2002.
[12] T. Y. Hou, X. H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[13] T. Y. Hou, X. H. Wu, and Y. Zhang, *Removing the cell resonance error in the multiscale finite element method via a petrov-galerkin formulation*, Commun. Math. Science, 2 (2004), pp. 185–205.
[14] T. Hou and X. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[15] M. Peszynska, M. Wheeler, and I. Yotov, *Mortar upscaling for multiphase flow in multiblock domains*, Comput. Geosci., 6 (2002), pp. 315–332.
[16] P. Renard and G. de Marsily, *Calculating equivalent permeability: A review*, Adv. in Water Resources, 20 (1997), pp. 253–278.
[17] T. Vdovina, S. Minkoff, and O. Korostyshevskaya, *Operator upscaling for the acoustic wave equation*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 1305–1338.
[18] X. Wu, Y. Efendiev, and T. Hou, *Analysis of upscaling absolute permeability*, Discrete Contin. Dyn. Syst., 2 (2002), pp. 185–204.
[19] I. Yotov, *Mortar mixed finite element methods on irregular multiblock domains*, in Iterative Methods in Scientific Computation, IMACS Ser. Comp. Appl. Math. 4, J. Wang, M. B. Allen, B. Chen, and T. Mathew, eds., International Association for Mathematics and Computers in Simulation, Brussels 1998, pp. 239–244.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 613–635

ANALYSIS OF PROJECTION METHODS FOR RATIONAL FUNCTION APPROXIMATION TO THE MATRIX EXPONENTIAL*

L. LOPEZ† AND V. SIMONCINI‡

Abstract. Krylov subspace methods for approximating the action of the matrix exponential $\exp(A)$ on a vector v are analyzed with A Hermitian and negative semidefinite. Our approach is based on approximating the exponential with the commonly employed diagonal Padé and Chebyshev rational functions, which yield a system of equations with a polynomial coefficient matrix. We derive optimality properties and error bounds for the convergence of a Galerkin-type approximation and of a computationally feasible and extensively used alternative. As complementary results, we theoretically justify the use of a popular a posteriori error estimate, and we provide upper bounds for the components of the solution vector. Our theoretical and numerical results show that this methodology may provide an appropriate framework to devise new strategies such as more powerful acceleration schemes.

Key words. matrix, exponential, Krylov subspaces, rational functions, iterative methods

AMS subject classifications. 65F10

DOI. 10.1137/05062590

1. Introduction.

The problem of numerically approximating the action of the matrix exponential $\exp(A)$ on v for a given matrix A and vector v is of great importance in a wide range of applications. In fact, it is the core of many exponential integrators for solving systems of ordinary differential equations (see [26, 25]) or time-dependent partial differential equations [17, 19]. Over the years, several methods have been proposed for approximating the exponential of a matrix; we refer to [33] for a recent survey. For A of large dimension, Krylov subspace methods for approximating $\exp(A)v$ have been successfully used for a long time; see, e.g., [36, 39], the more recent publications [19, 25, 9], and references therein. In the past few years, important contributions have appeared that have significantly increased the theoretical understanding of this approach [48, 11, 12, 24, 41]. In this paper, we restrict our attention to the case of A Hermitian and negative semidefinite, as it is often the case in real applications, although the approach can be used even for non-Hermitian A. Given an $n \times n$ matrix A, the Krylov subspace $K_m(A, v) = \text{span}\{v, Av, \ldots, A^{m-1}v\}$ is characterized by the key relation

$$AV_m = V_{m+1}H_{m+1,m}, \qquad v = V_m e_1 \beta_0, \tag{1.1}$$

with $H_{m+1,m} \in \mathbb{R}^{(m+1)\times m}$ tridiagonal and $\beta_0 = \|v\|$, where $\|v\|$ is the 2-norm of v. Here and in the following, e_k denotes the kth vector of the canonical basis, whose dimension is clear from the context. In later sections, we use e_m^G and e_m^K to denote the *error* vectors for the analyzed methods. Relation (1.1), also known as the Lanczos recurrence, allows one to compute a matrix V_m whose orthonormal columns span $K_m(A, v)$, while $H_{m+1,m}$ contains the coefficients of the orthogonalization process. A

*Received by the editors March 3, 2005; accepted for publication (in revised form) November 8, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/sinum/44-2/62590.html
†Dipartimento di Matematica, Università di Bari, Via E. Orabona 4, I-70125 Bari, Italy (lopezl@dm.uniba.it).
‡Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy and CIRSA, Ravenna, Italy (valeria@dm.unibo.it).

common approximation in $K_m(A, v)$ is

$$\exp(A)v \approx V_m \exp(H_m)e_1\beta_0; \tag{1.2}$$

here H_m is the (Hermitian) $m\times m$ principal part of $H_{m+1,m}$, i.e., $H_m = V_m^* A V_m$, where V_m^* is the transpose of V_m. This approximation was analyzed in [41], where it was also shown that the vector $V_m \exp(H_m)e_1\beta_0$ represents a polynomial approximation to $\exp(A)v$, in which the polynomial of degree $m-1$ interpolates the exponential function in the Hermite sense on the set of eigenvalues of H_m [41, Theorem 3.3]. Our analysis aims to explore this polynomial approach but from a different perspective.

The computation of $\exp(A)$, and of $\exp(H_m)$ cannot be carried out exactly, even assuming exact arithmetic [33]; we refer to [43] for a description of current software for computing the exponential of small matrices. In practice, $\exp(H_m)$ is often very accurately approximated by means of rational functions, such as Padé or Chebyshev functions; see, e.g., [1, 20, 21, 23, 19]. Therefore, given the rational function $\mathcal{R}_{\mu,\nu}(\lambda) := \Phi_\mu(\lambda)/\Psi_\nu(\lambda)$ for some specifically chosen polynomials Ψ_μ, Φ_ν of degree μ and ν, respectively, the approximate solution $V_m \exp(H_m)e_1\beta_0$ is replaced by the vector $V_m\mathcal{R}_{\mu,\nu}(H_m)e_1\beta_0$; see, e.g., the Matlab routine `expm` [32]. In the following we restrict our analysis to the case $\mu = \nu$, and thus we use $\mathcal{R}_\nu \equiv \mathcal{R}_{\nu,\nu}$; see section 3.

The aim of this paper is to increase our understanding of Krylov-subspace-based approximations to the exponential operator by exploiting rational functions and their approximation properties of the exponential. Rational functions may provide an appropriate framework to devise more powerful techniques, as well as to justify currently proposed approaches such as those in [6, 49]; see also section 6. In particular, we wish to set up the stage for the development of new acceleration strategies to computationally enhance the approximation process.

General results on approximation of matrix rational functions within Krylov subspaces are very limited; see, for instance, [51]. In practice, the theoretical as well as computational aspects associated with such approximation have not been completely addressed. We aim to contribute in filling this gap, as very general hypotheses on the rational functions are employed. Therefore, in this paper we first derive new error estimates for projection-type minimization methods used to approximate the action of matrix rational functions. Our results are very general and can be applied in contexts other than the approximation of the exponential. We then derive new insightful relations for the approximation of the exponential, with the commonly employed technique in (1.2).

We start with the preliminary consideration that the approximation

$$\exp(A)v \approx \mathcal{R}_\nu(A)v = (\Psi_\nu(A))^{-1}\Phi_\nu(A)v$$

entails solving the following system of equations:

$$\Psi_\nu(A)x = \Phi_\nu(A)v. \tag{1.3}$$

Note that as an alternative to solving (1.3), one could first approximately solve the system $\Psi_\nu(A)\hat{x} = v$ and then compute $x = \Phi_\nu(A)\hat{x}$. Since for $m > \nu$ it holds that $\Phi_\nu(A)v \in K_m(A, v)$, the approach (1.3) should be preferred in practice. Following [51], we analyze two procedures for solving (1.3). The first approach determines an approximation x_m by imposing a classical Galerkin condition on the residual. The second one, which we call the Krylov approximation,[1] is computationally more

[1] Also called Arnoldi or Lanczos approximation.

appealing and turns out to be equivalent to the standard procedure, namely,

$$x_m = V_m \mathcal{R}_\nu(H_m) e_1 \beta_0 \approx V_m \exp(H_m) e_1 \beta_0. \tag{1.4}$$

For the sake of simplicity, and without loss of generality, in the following we assume that $\|v\| = 1$ so that $\beta_0 = 1$. We analyze the optimality properties of the first method and show that the second method, although nonoptimal, provides an approximate solution that is significantly close to the optimal one. By using the partial fraction expansion of $\mathcal{R}_\nu$ we derive convergence bounds for both approaches that depend on the spectrum of A. In addition, we examine the role of the degree and the poles of both Padé and Chebyshev rational approximants in the convergence behavior as well as in the obtained error bounds.

Our convergence estimates predict linear, and not superlinear, convergence, and in this sense they are weaker than available error bounds; however, we also show that the superlinear behavior can be recovered by varying the degree ν. We stress here that our aim is not to derive better bounds than those in the literature. Instead, we wish to show that rational functions may represent a new numerical tool with no loss in convergence properties if the degree ν is taken into account.

Throughout the paper we assume exact precision arithmetic. We refer to [10] for a detailed analysis of the behavior of Krylov subspace approximations of matrix functions in finite precision computation.

In section 2 we review some basic facts on Krylov approximation of the exponential, while in section 3 we review several important properties of rational functions that will be used extensively in the paper. In section 4 we show an optimality property associated with the Galerkin method and provide bounds for the approximation error. In section 5 we analyze in detail the Krylov method. We first relate its approximate solution with that obtained with the optimal Galerkin approximation. Then, we derive new error estimates and compare them with those obtained for the Galerkin procedure. In section 6 we discuss some computational properties derived by using rational functions. In section 7 we analyze the Padé rational function approximation when the scaling and squaring procedure is employed to handle a matrix whose norm is significantly larger than one, while in section 8 we discuss the role of ν in the occurrence of superlinear convergence. Finally, section 9 discusses some related issues that our analysis brings to light.

2. Krylov subspace approximation to the matrix exponential. Krylov subspace approximations to the exponential have been analyzed in several papers; see, e.g., [41, 19]. However, the most significant error bounds were given in [48, 11] and later with a different approach in [24]. The authors of these papers were able to capture the so-called *superlinear* convergence of the approximation. As a reference, we recall here one of the results stated in [24] in our notation; see [48, 11] for qualitatively similar, although asymptotic bounds. We call these bounds *ideal* bounds, for reasons that will be clear in the following.

THEOREM 2.1 (see [24]). *Let A be a Hermitian negative semidefinite matrix with eigenvalues in the interval $[-4\rho, 0]$. Then the error in the Lanczos approximation of $\exp(A)v$ is bounded as follows:*

$$\|\exp(A)v - V_m \exp(H_m)e_1\| \le 10 e^{-m^2/(5\rho)}, \qquad \sqrt{4\rho} \le m \le 2\rho,$$
$$\|\exp(A)v - V_m \exp(H_m)e_1\| \le \frac{10}{\rho} e^{-\rho} \left(\frac{e\rho}{m}\right)^m, \qquad m \ge 2\rho.$$

Different bounds that also emphasize the superlinear character of the approximation have also been proposed in [47]; we found these latter bounds less sharp than those in [48, 11, 24], at least experimentally.

The Krylov approximation devised in (1.2) is not naturally equipped with a stopping criterion. Since the error norm $\| \exp(A)v - V_m \exp(H_m)e_1\beta_0\|$ cannot be computed explicitly as m increases, a criterion based on the quantity

$$h_{m+1,m}|e_m^* \exp(H_m)e_1\beta_0| \tag{2.1}$$

was proposed in [41, section 5.2]. This criterion works well in many cases, especially when $\|A\|$ is moderate, and qualitative arguments were discussed in [41] to justify its use; a higher order estimate can also be employed, which can be easily derived from (2.1) [41]. In general, an insightful interpretation of the quantity in (2.1) in not always immediate, usually due to the lack of a definition of residual, unlike in equation-based problems. An exception is the situation when the computation is related to the following initial value problem

$$\begin{cases} -Ax(t) + x'(t) = 0, \\ x(0) = v, \end{cases}$$

in which case it holds $h_{m+1,m}|e_m^* \exp(tH_m)e_1\beta_0| = \| - Ax_m(t) + x'_m(t)\|$, that is, the a posteriori estimate is indeed the residual associated with the approximate solution $x_m(t)$; see, e.g., [7, 10]. An alternative viewpoint was proposed in [26], where the authors introduced a new concept of residual norm by generalizing that of error norm in a functional setting. The new residual norm was shown to be equal to the estimate (2.1). With our derivation, we suggest a general role for the stopping criterion (2.1) in terms of residual associated with a matrix equation.

3. Rational function approximation. Rational functions are commonly used to accurately approximate analytic functions such as the exponential [1]. Here we review some characteristics of Chebyshev and Padé rational functions that are used in our analysis. However, several of the results in later sections apply to general rational functions and to the approximation of other smooth functions.

Let us assume that $\exp(\lambda)$ is approximated by the rational function $\mathcal{R}_\nu(\lambda)$. In this case, the quality of the approximation when using the Krylov subspace only affects part of the overall approximation. The bound

$$\begin{aligned} \| \exp(A)v - V_m \exp(H_m)e_1\| \le \| \exp(A)v - \mathcal{R}_\nu(A)v\| + \|\mathcal{R}_\nu(A)v - V_m\mathcal{R}_\nu(H_m)e_1\| \\ +\|\mathcal{R}_\nu(H_m)e_1 - \exp(H_m)e_1\| \end{aligned}$$

emphasizes that there are two components in the error estimate: the second term on the right is the "Krylov subspace error," and it can be monitored as the chosen approximation in $K_m(A, v)$ takes place, whereas the size of the second component, corresponding to the other two terms in the bound, depends on the accuracy of the rational function approximation employed.

The first component is related to the numerical solution of the system (1.3) in the Krylov subspace. The solution of algebraic systems having a matrix function as coefficient matrix by means of Krylov subspaces has been analyzed in detail by van der Vorst in [51]; see also the recent presentation in [50]. Two distinct methods are studied in [51] for special cases of Ψ_ν and for $\Phi_\nu = 1$. In the first approach, the problem is projected onto the smaller dimension space, whereas the second approach

is characterized by a sequential projection. In this paper we generalize these two methods to our framework and analyze their properties.

Several approaches have been considered for choosing the rational function approximation. In the context of one-step methods for initial value differential problems, a stable way to approximate $\exp(A)$ consists of employing diagonal Padé approximants $\mathcal{R}_\nu = \Phi_\nu/\Psi_\nu$, where Φ_ν, Ψ_ν are polynomials of degree ν; see, e.g., [26, 25] and references therein. These two polynomials satisfy $\Phi_\nu(\lambda) = \Psi_\nu(-\lambda)$ so that we can write

$$\Psi_\nu(\lambda) = \psi_\nu \cdot (\lambda - \xi_1) \cdots (\lambda - \xi_\nu), \qquad \Phi_\nu(\lambda) = \phi_\nu \cdot (\lambda + \xi_1) \cdots (\lambda + \xi_\nu). \tag{3.1}$$

Here ψ_ν and ϕ_ν are the leading term coefficients. The two polynomials are uniquely defined apart from a scaling factor. We shall assume in the following that this scaling factor is such that $\Psi_\nu(0) = \Phi_\nu(0) = 1$. The roots of Ψ_ν all have positive real part, so that those of Φ_ν have negative real part; in addition, they come in complex conjugate if their imaginary part is nonzero and their absolute value is larger than one, and increasing with ν. The leading coefficient ψ_ν satisfies

$$|\psi_\nu| = \frac{1}{|\xi_1 \cdot \cdots \cdot \xi_\nu|} \ll 1;$$

it is positive if ν is even, and negative if ν is odd. In addition, $|\Phi_\nu(\lambda)/\Psi_\nu(\lambda)| \le 1$ for $\lambda \le 0$ [52]. Finally, for any nonpositive real λ we have $\Psi_\nu(\lambda) > 0$. In our context, this property ensures that $\Psi_\nu(A)$ is Hermitian and positive definite for any Hermitian negative semidefinite matrix A, that is, $x^*\Psi_\nu(A)x > 0$ for any nonzero vector x.

In the context of parabolic partial differential equations, rational Chebyshev approximations have also been considered; see, e.g., [52, 19], which provide best rational approximations to $\exp(x)$ for $x \in (-\infty, 0]$ in the Chebyshev sense. If $\Phi_\nu(-\lambda)/\Psi_\nu(-\lambda)$ is the Chebyshev approximant for $\lambda \le 0$, then it is known that $\sup_{\lambda \le 0} |\exp(\lambda) - \Phi_\nu(-\lambda)/\Psi_\nu(-\lambda)| \approx 10^{-\nu}$; see [8, Table II]. In addition, since Ψ_ν has all strictly positive coefficients (cf. [8, Table III]), then $\Psi_\nu(x) > 0$ for $x \ge 0$, implying that $\Psi_\nu(-A)$ is positive definite for A Hermitian and negative semidefinite. The roots ξ_j of Ψ_ν appear with positive and negative real part, therefore $M_j = -A - \Re(\xi_j)I$ may be indefinite for some ξ_j, $j = 1, \ldots, \nu$. We note that the Chebyshev rational approximation uses $\Phi_\nu(-\lambda)/\Psi_\nu(-\lambda)$ with $\lambda \le 0$, whereas the Padé approximation employs $\Phi_\nu(\lambda)/\Psi_\nu(\lambda)$ with $\lambda \le 0$. In this paper we do not distinguish between the sign in the two cases, using $\Phi_\nu(\lambda)/\Psi_\nu(\lambda)$ with $\lambda \le 0$, while warning the reader that depending on the strategy used, the variable sign should be changed accordingly.

Padé approximants of degree up to $\nu = 14$ are commonly employed [23], and 10^{-14} is often considered a sufficiently good accuracy for the Chebyshev approximation. Unless otherwise specified, we thus restrict our experiments to the case $\nu \le 14$. For the sake of simplicity, we only consider equal degree approximants, whereas it is known that in the Padé approximation, $\Phi_{\nu-1}, \Psi_\nu$ also provide stable approximations in the context of stiff ordinary differential equations; see, e.g., [20]. In our analysis we use the fact that both the Padé and the Chebyshev approximants have simple poles, although the presence of multiple poles is addressed in section 7 in the context of the scaling and squaring method.

We also recall that the rational function $\mathcal{R}_\nu = \Phi_\nu/\Psi_\nu$ can be written by means of a partial fraction expansion as

$$\mathcal{R}_\nu(\lambda) = \tau_0 + \sum_{j=1}^{\nu} \frac{\tau_j}{(\lambda - \xi_j)}, \tag{3.2}$$

where $\xi_1, \dots, \xi_\nu$ are the distinct roots of Ψ_ν, $\tau_1, \dots, \tau_\nu$ are the coefficients (appearing in complex conjugates) of the expansion, and τ_0 is the remainder.

4. Convergence analysis of the Galerkin method. In this section we analyze the convergence properties of the Galerkin approximation. A Galerkin approach based on the Krylov subspace $K_m(A, v)$ approximates x in (1.3) as $x_m^G = V_m y_m^G$ by imposing that the residual $\Phi_\nu(A)v - \Psi_\nu(A)x_m^G$ be orthogonal to the Krylov subspace, namely, $V_m^*(\Phi_\nu(A)v - \Psi_\nu(A)x_m^G) = 0$. Therefore, y_m^G is computed as the solution to the system

$$V_m^* \Psi_\nu(A) V_m y = V_m^* \Phi_\nu(A) v. \tag{4.1}$$

The method is of interest from a theoretical point of view, because of its optimality properties. From a computational standpoint, the explicit computation of $V_m^*\Psi_\nu(A)V_m$ requires ν evaluations with A at each iteration, making the approach not appealing. The Krylov approximation thus represents a valuable competitive alternative, and we show that the convergence properties are indeed comparable.

We first show that the Galerkin approximate solution has a minimization property, ensuring that the error is nonincreasing with m in the considered norm; then we derive upper bounds for this error norm. All these results appear to be new.

PROPOSITION 4.1. *Let $x_\star = \mathcal{R}_\nu(A)v$ and let x_m^G be the Galerkin approximation to $x_\star$ in $K_m(A, v)$ and assume that $\Psi_\nu(A)$ is Hermitian and positive definite. Then*

$$\min_{x \in K_m(A,v)} \|x_\star - x\|_{\Psi_\nu(A)} = \|x_\star - x_m^G\|_{\Psi_\nu(A)}.$$

Proof. The result follows from imposing the Galerkin condition on the residual $\Psi_\nu(A)(x_\star - x_m)$; cf. [42, Proposition 5.2]. ∎

PROPOSITION 4.2. *Let $[\alpha, \beta]$ be the interval containing all eigenvalues of A and assume that the hypotheses of Proposition* 4.1 *hold. Then*

$$\begin{aligned}\min_{x \in K_m(A,v)} \|x_\star - x\|^2_{\Psi_\nu(A)} &= \min_{q \in \mathbb{P}_{m-1}} \|x_\star - q(A)v\|^2_{\Psi_\nu(A)} \\ &\le \min_{q \in \mathbb{P}_{m-1}} \max_{\lambda \in [\alpha,\beta]} \left|1 - \mathcal{R}_\nu(\lambda)^{-1} q(\lambda)\right|^2 \|v\|^2_{\Psi_\nu(A)}.\end{aligned}$$

Proof. Let $u_1, \dots, u_n$ be the unit norm eigenvectors of A associated with the eigenvalues $\lambda_1, \dots, \lambda_n$. Define $\chi_i := u_i^* x_\star$ and $s(\lambda) := 1 - q(\lambda)(\mathcal{R}_\nu(\lambda))^{-1}$. We have $x_\star - q(A)v = \left(I - q(A)(\mathcal{R}_\nu(A))^{-1}\right) x_\star = \sum_{i=1}^n u_i s(\lambda_i)\chi_i$ so that

$$\begin{aligned}\|x_\star - q(A)v\|^2_{\Psi_\nu(A)} &= \langle x_\star - q(A)v, \Psi_\nu(A)(x_\star - q(A)v)\rangle \\ &= \sum_{i=1}^n \Psi_\nu(\lambda_i)(s(\lambda_i))^2 \chi_i^2 \le \max_{\lambda \in [\alpha,\beta]} |s(\lambda)|^2 \|v\|^2_{\Psi_\nu(A)}.\end{aligned}$$ ∎

We next provide a bound for the polynomial min-max problem in Proposition 4.2; we use the following definitions. For each pole ξ_j, we set $M_j = A - \Re(\xi_j)I$, and we let α_j be the eigenvalue of M_j with largest absolute value and β_j be the eigenvalue of M_j with smallest absolute value. If $\Re(\xi_j) > 0$, then $\alpha_j < \beta_j < 0$. Moreover, we let

$$\rho_j = \gamma_j + \sqrt{\gamma_j^2 - 1} \quad \text{with} \quad \gamma_j = \frac{|\alpha_j - i\Im(\xi_j)| + |\beta_j - i\Im(\xi_j)|}{|\alpha_j - \beta_j|}. \tag{4.2}$$

The following bound holds.

THEOREM 4.3. *Assume that the spectrum of A is contained in the negative interval $[\alpha, \beta]$, and that Ψ_ν has distinct roots. Then, with the notation above,*

$$\min_{q\in\mathbb{P}_{m-1}} \max_{\lambda\in[\alpha,\beta]} \left|1 - (\mathcal{R}_\nu(\lambda))^{-1} q(\lambda)\right| \le 2 \sum_{j=1}^{\nu} \left(\max_{\lambda\in[\alpha,\beta]} \frac{|\tau_j|}{\mathcal{R}_\nu(\lambda)\,|\lambda - \xi_j|} \right) \frac{1}{\rho_j^m + 1/\rho_j^m}.$$

Proof. In the proof we omit the polynomial subscripts. We first rewrite the problem as

$$\min_{q\in\mathbb{P}_{m-1}} \max_{\lambda\in[\alpha,\beta]} \left|1 - (\mathcal{R}(\lambda))^{-1} q(\lambda)\right| = \min_{q\in\mathbb{P}_{m-1}} \max_{\lambda\in[\alpha,\beta]} \left| \frac{1}{\mathcal{R}(\lambda)} (\mathcal{R}(\lambda) - q(\lambda)) \right|. \tag{4.3}$$

Let q^* be the polynomial of degree at most $m-1$ that attains the minimum in (4.3). Using the partial fraction expansion in (3.2), for any $q \in \mathbb{P}_{m-1}$ we have

$$|\mathcal{R}(\lambda) - q^*(\lambda)| = \left| \tau_0 + \sum_{j=1}^{\nu} \frac{\tau_j}{\lambda - \xi_j} - q^*(\lambda) \right| \le \max_{\lambda\in[\alpha,\beta]} \left| \tau_0 + \sum_{j=1}^{\nu} \frac{\tau_j}{\lambda - \xi_j} - q(\lambda) \right|.$$

We choose $q \in \mathbb{P}_{m-1}$ defined as $q(\lambda) = \tau_0 + \sum_{j=1}^{\nu} \tau_j q^{(j)}(\lambda - \xi_j)$, with $\lambda \in [\alpha, \beta]$, while $q^{(j)} \in \mathbb{P}_{m-1}$, $j = 1, \ldots \nu$ are polynomials yet to be determined. We set $p^{(j)}(\lambda - \xi_j) = 1 - (\lambda - \xi_j) q^{(j)}(\lambda - \xi_j)$, with $p^{(j)} \in \mathbb{P}_m$ and $p^{(j)}(\xi_j) = 1$. Thus

$$\begin{aligned} \left| \tau_0 + \sum_{j=1}^{\nu} \frac{\tau_j}{\lambda - \xi_j} - q(\lambda) \right| &= \left| \sum_{j=1}^{\nu} \tau_j \left(\frac{1}{\lambda - \xi_j} - q^{(j)}(\lambda - \xi_j) \right) \right| \\ &= \left| \sum_{j=1}^{\nu} \tau_j \frac{1}{\lambda - \xi_j} p^{(j)}(\lambda - \xi_j) \right|. \end{aligned}$$

It was shown in [15, formula (38)] that the polynomial $p^{(j)}$ can be constructed so that $\max_{\zeta\in[\alpha-\xi_j,\beta-\xi_j]} |p^{(j)}(\zeta)| = 2/(\rho_j^m + 1/\rho_j^m)$. Hence, we can write

$$\begin{aligned} \min_{q\in\mathbb{P}_{m-1}} \max_{\lambda\in[\alpha,\beta]} \left| \frac{1}{\mathcal{R}(\lambda)} (\mathcal{R}(\lambda) - q(\lambda)) \right| &\le \max_{\lambda\in[\alpha,\beta]} \left| \sum_{j=1}^{\nu} \frac{\tau_j}{\mathcal{R}(\lambda)(\lambda - \xi_j)} p^{(j)}(\lambda - \xi_j) \right| \\ &\le \sum_{j=1}^{\nu} \left(\max_{\lambda\in[\alpha,\beta]} \frac{|\tau_j|}{\mathcal{R}(\lambda)\,|\lambda - \xi_j|} \right) \frac{2}{(\rho_j^m + 1/\rho_j^m)}. \end{aligned}$$

∎

The denominator in the bound always satisfies $|\lambda - \xi_j| \neq 0$ if $\Im(\xi_j) \neq 0$. For ν odd and real ξ_j, we have $\xi_j > 0$ for Padé so that $\lambda - \xi_j < 0$ for $\lambda \in [\alpha, \beta]$, while for Chebyshev, we have $\xi_j < 0$ and $-\lambda \in [\alpha, \beta]$ so that $\lambda - \xi_j > 0$. In both cases, $|\lambda - \xi_j| \neq 0$.

Remark 4.4. When $\beta = 0$, so that the eigenvalues of A are in $[\alpha, 0]$, then roughly $\rho_j \approx 2(1 + 2|\xi_j/\alpha|)$. Therefore, we obtain

$$\sum_{j=1}^{\nu} \left(\max_{\lambda\in[\alpha,\beta]} \frac{|\tau_j|}{\mathcal{R}_\nu(\lambda)\,|\lambda - \xi_j|} \right) \frac{2}{(\rho_j^m + 1/\rho_j^m)} \approx e^{-\alpha} \sum_{j=1}^{\nu} \max_{\lambda\in[\alpha,\beta]} \frac{|\tau_j|}{|\lambda - \xi_j|} \cdot \frac{2}{(2 + 4|\frac{\xi_j}{\alpha}|)^m},$$

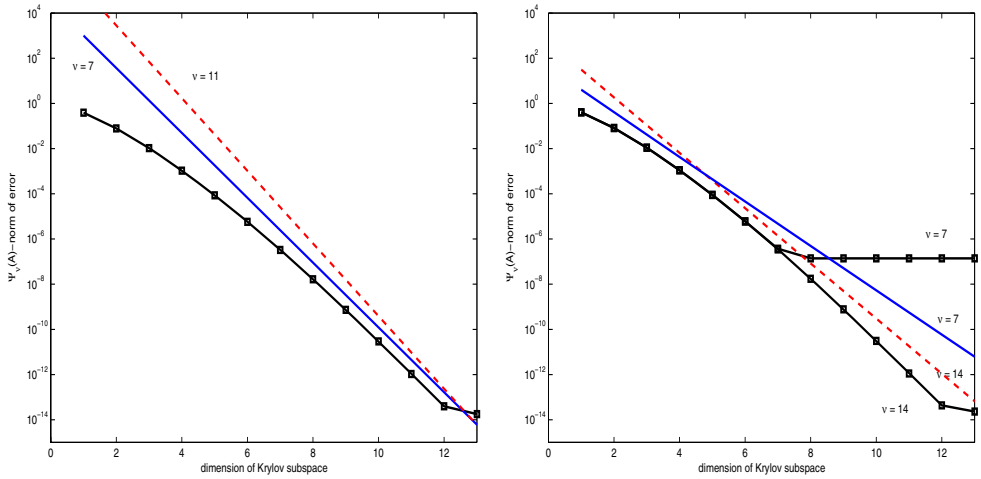


FIG. 4.1. *Example* 4.5*: Galerkin approximation. Left: Padé rational functions (black solid with squares) and upper bound for* $\nu = 7, 11$*. Right: Chebyshev rational functions (black solid with squares) and upper bounds for* $\nu = 7, 14$*.*

which shows the role of ξ_j as $|\alpha|$ gets large. See also section 8 for additional remarks on the role of the poles. ∎

To show the sharpness of the bound in Theorem 4.3, we next report a numerical experiment with both Padé and Chebyshev rational functions. Below and later in the paper, we show convergence curves for $\|\exp(A)v - x_m\|_*$, where $\|\cdot\|_*$ is the norm of interest. The bound $\|\exp(A)v - x_m\|_* \le \|\exp(A)v - \mathcal{R}_\nu(A)v\|_* + \|\mathcal{R}_\nu(A)v - x_m\|_*$ allows us to use estimates such as that in Theorem 4.3 to bound the second term in the right-hand side, whereas the first term depends on the accuracy of the rational function approximation. We will see that the error $\|\exp(A)v - x_m\|_*$ stagnates at the final accuracy level of $\|\exp(A)v - \mathcal{R}_\nu(A)v\|_*$.

Example 4.5. This is a contrived example but here and later in the text, it serves as a simple platform for describing the principal properties of the two rational approximations. We consider the 100×100 diagonal matrix A of the logarithm of equispaced values between 0.2 and 0.99. In Matlab notation, this can be defined as `A = diag(log(linspace(0.2,0.99,100)))`. The spectrum of A is contained in the interval $[-1.61, -0.0101]$. The vector v is chosen to be the vector of all ones, scaled so as to have unit norm. The Padé polynomials Ψ_ν, Φ_ν are computed for $\nu = 7$ and $\nu = 11$, and no prescaling of A is employed (see section 7). The approximate solution $x_m^G = V_m y_m^G$ is obtained at each step m by explicitly computing $V_m^* \Psi_\nu(A) V_m$ and $V_m^* \Phi_\nu(A) v$, and then solving (1.3).

In the left plot of Figure 4.1 we report the relative $\Psi_\nu(A)$-norm of the error, $\|\exp(A)v - x_m^G\|_{\Psi_\nu(A)}$ (solid black curve with squares), together with the upper bound in Theorem 4.3 corresponding to Padé rational functions with $\nu = 7, 11$. Note that the $\Psi_\nu(A)$-norm of the error is indistinguishable for the two values of ν. The right plot shows the same curves associated with Chebyshev rational functions with $\nu = 7, 14$. As predicted by the theory, the final level of accuracy is reached at about $10^{-\nu}$.

We remark that the bound is significantly accurate for the Padé approximation. The deterioration due to the high degree $\nu = 11$ is mostly noticed during the first phase of the convergence; but see also section 8. A very satisfactory bound is obtained for the Chebyshev approximation as well, especially for $\nu = 14$.

5. Layers of Galerkin approximations. The Krylov approximation in (1.4) requires the solution of the system $\Psi_\nu(H_m)y_m = \Phi_\nu(H_m)e_1$. This is interpreted in [51] as a sequential Galerkin projection onto the Krylov subspace $K_m(A, v)$ of the linear systems

$$(A - \xi_j I)w_m^{(j)} = w_m^{(j-1)}, \quad j = 1, \ldots, \nu,$$

with $w_m^{(0)} = V_m y_m^{(0)}$ and $y_m^{(0)} = \Phi_\nu(H_m)e_1$. The solution of each system in the Krylov subspace then corresponds to employing the full orthogonalization method (FOM) [42]. Indeed, writing $w_m^{(j)} = V_m y_m^{(j)}$ for $j = 1, \ldots, \nu$, we have

$$V_m^*(A - \xi_j I)V_m y_m^{(j)} = y_m^{(j-1)}, \quad j = 1, \ldots, \nu,$$

that is, $(H_m - \xi_j I)y_m^{(j)} = y_m^{(j-1)}$, $j = 1, \ldots, \nu$, so that, using (3.1),

$$\begin{aligned} y_m \equiv \frac{1}{\psi_\nu} y_m^{(\nu)} &= \frac{1}{\psi_\nu}(H_m - \xi_\nu I)^{-1} \cdots (H_m - \xi_1 I)^{-1}\Phi_\nu(H_m)e_1 \\ &\equiv (\Psi_\nu(H_m))^{-1}\Phi_\nu(H_m)e_1. \end{aligned}$$

In this section we try to increase our understanding of the Krylov approximation. The next result states an explicit relation between the Galerkin and Krylov solutions when the approximation subspace has dimension larger than ν. We show that the two matrices $V_m^*\Psi_\nu(A)V_m, \Psi_\nu(H_m)$ coincide except for the bottom $\nu - 1$ diagonal block. This result also allows us to conclude that the two solutions tend to coalesce as convergence takes place.

PROPOSITION 5.1. *For $m > \nu$, let $y_m^K = (\Psi_\nu(H_m))^{-1}\Phi_\nu(H_m)e_1$ and $y_m^G = (V_m^*\Psi_\nu(A)V_m)^{-1}V_m^*\Phi_\nu(A)v$ be the Krylov and Galerkin approximations to the vector $x_\star = (\Psi(A))^{-1}\Phi_\nu(A)v$, respectively. Then, there exists a $(\nu - 1) \times (\nu - 1)$ matrix S_ν, such that $V_m^*\Psi_\nu(A)V_m = \Psi_\nu(H_m) + S_{m,\nu}$, with $S_{m,\nu} = E_{m,\nu-1}S_\nu E_{m,\nu-1}^*$, and $E_{m,\nu-1}^* = [0, I_{\nu-1}] \in \mathbb{R}^{(\nu-1)\times m}$. As a consequence, $\Phi_\nu(A)v = V_m\Phi_\nu(H_m)e_1$, and*

$$y_m^K = y_m^G + (V_m^*\Psi_\nu(A)V_m)^{-1}E_{m,\nu-1}S_\nu E_{m,\nu-1}^* y_m^K$$

so that $\|y_m^G - y_m^K\| \le \|(V_m^*\Psi_\nu(A))^{-1}E_{m,\nu-1}S_\nu\| \, \|E_{m,\nu-1}^* y_m^K\|$.

Proof. We eliminate the polynomial subscripts in the proof. For any polynomial Ψ of degree at most ν, it can be explicitly shown that for $j \le m - \nu + 1$, $V_m^*\Psi(A)V_m e_j = \Psi(H_m)e_j$ and because of symmetry, it also holds $e_j^*V_m^*\Psi(A)V_m = e_j^*\Psi(H_m)$. Therefore, we can write $S_{m,\nu} = E_{m,\nu-1}S_\nu E_{m,\nu-1}^*$, $E_{m,\nu-1}^* = [0, I_{\nu-1}]$, for some $(\nu - 1) \times (\nu - 1)$ matrix S_ν, from which the first result follows.

The fact that the two vectors $\Phi(A)v$ and $V_m\Phi(H_m)e_1$ are equal is an immediate consequence of the matrix relation above. Using the definition of y_m^K, y_m^G, we have

$$\begin{aligned} (V_m^*\Psi(A)V_m - S_{m,\nu})y_m^K &= \Phi(H_m)e_1, \\ y_m^K - (V_m^*\Psi(A)V_m)^{-1}S_{m,\nu}y_m^K &= (V_m^*\Psi(A)V_m)^{-1}\Phi(H_m)e_1 \equiv y_m^G, \\ y_m^K - y_m^G &= (V_m^*\Psi(A)V_m)^{-1}S_{m,\nu}y_m^K, \end{aligned}$$

from which the relation and the final bound follow. ∎

Direct inspection shows that $\|S_\nu\| = O(h_{m+1,m}^2)$. In Table 5.1, we report some numerical results that highlight the relation between the two methods. Data in Example 4.5 with the Padé rational function of degree $\nu = 7$ are considered. The columns display the norm of the error for the Krylov method as the subspace dimension increases and the 2-norm and Ψ_ν-norm of the error for the Galerkin method. The two

TABLE 5.1
Example 4.5*: Convergence of Galerkin and Krylov approaches using Padé approximation with* $\nu = 7$*. The last column shows the difference between the two solutions.*

m	$\|x_\star - x_m^K\|$	$\|x_\star - x_m^G\|$	$\|x_\star - x_m^G\|_{\Psi_\nu}$	$\|x_m^K - x_m^G\|$
1	2.3574e-01	2.3565e-01	2.7175e-01	4.0544e-04
2	4.6261e-02	4.6749e-02	5.4550e-02	2.9431e-03
3	6.1459e-03	6.2329e-03	7.3193e-03	5.8970e-04
4	6.1599e-04	6.2576e-04	7.3762e-04	7.1273e-05
5	4.9501e-05	5.0333e-05	5.9474e-05	6.3820e-06
6	3.3163e-06	3.3738e-06	3.9931e-06	4.5871e-07
7	1.9031e-07	1.9368e-07	2.2948e-07	2.7646e-08
8	9.5430e-09	9.7134e-09	1.1518e-08	1.4367e-09
9	4.2452e-10	4.3215e-10	5.1269e-10	6.5655e-11
10	1.6955e-11	1.7261e-11	2.0484e-11	2.6769e-12
11	6.1394e-13	6.2500e-13	7.4186e-13	9.8467e-14
12	2.2013e-14	2.2331e-14	2.7469e-14	3.3028e-15
13	8.4927e-15	8.4682e-15	1.2372e-14	1.8598e-16

approaches show very close, although not identical, approximation (cf. $\|x_m^K - x_m^G\|$). We also notice that the error between the two solutions decreases with m, the difference being one order of magnitude smaller than the approximation error.

Although the result of Proposition 5.1 sheds light on the similarities of the two methods for large m, the two approaches behave very similarly even for $m \le \nu$. The analysis for $m \le \nu$ is of great interest, since for $\|A\|$ not much greater than unit, high convergence rate can be observed for the two methods, and final accuracy is obtained for m possibly smaller than ν, as is the case in Example 4.5. In the following section we show that the error in the Krylov approximation can be bounded in a way similar to what we derived for the Galerkin approximation.

5.1. The Krylov approximation. The proof of Theorem 4.3 inspires an alternative way to justify the use of the Krylov approach in (1.4) to approximate the rational function $\mathcal{R}_\nu(A)v$. Using the partial fraction expansion in (3.2) we can write

$$x_\star = \mathcal{R}_\nu(A)v = \tau_0 v + \sum_{j=1}^{\nu} \tau_j (A - \xi_j I)^{-1} v. \tag{5.1}$$

Therefore, an approximation to $x_\star$ may be obtained by approximating the solution $d^{(j)}$ to each system $(A - \xi_j I)d = v$, $j = 1, \dots, \nu$; this type of approach has been explored, for instance, in [18, 5, 2, 4, 31]. Thanks to the shift invariance property of Krylov subspaces, i.e., $K_m(A, v) = K_m(A - \xi_j I, v)$ for any $\xi_j \in \mathbb{C}$, approximations to $d^{(j)}$ can be obtained in the same subspace $K_m(A, v)$ as $d_m^{(j)} = V_m y_m^{(j)}$ for some $y_m^{(j)}$ using the FOM method. More precisely, if $y_m^{(j)}$ is determined by imposing a Galerkin (orthogonality) condition on the residual $v - (A - \xi_j I)V_m y_m^{(j)}$, then we obtain $y_m^{(j)} = (H_m - \xi_j I)^{-1} e_1$. Therefore, substituting $y_m^{(j)}$ in the expansion yields

$$\begin{aligned} x_\star &= \tau_0 v + \sum_{j=1}^{\nu} \tau_j (A - \xi_j I)^{-1} v \approx \tau_0 v + \sum_{j=1}^{\nu} \tau_j V_m y_m^{(j)} \\ &= V_m \left(\tau_0 e_1 + \sum_{j=1}^{\nu} \tau_j (H_m - \xi_j I)^{-1} e_1 \right) = V_m \mathcal{R}_\nu(H_m) e_1. \end{aligned} \tag{5.2}$$

The last term is precisely the Krylov approximation (1.4).

Theorem 4.3 shows that the approximation obtained by a Galerkin projection minimizes the error in the $\Psi(A)$-norm over all approximations in the subspace $K_m(A, v)$, and thus we expect a larger error with the Krylov approximation. The derivation above shows that the Krylov approximation yields a Galerkin solution on each system with $(A - \xi_j I)$; however, since $(A - \xi_j I)$ is not Hermitian for ξ_j complex, the Galerkin solution does not yield an error minimizing process. Note, however, that if for some j, ξ_j is real (and positive), then the matrix $A - \xi_j I$ is negative definite. Therefore the Krylov approach does provide an error minimizing solution for that term in the expansion.

If one abandons the idea of using the Krylov approximation (5.2), the expansion in (5.1) suggests that one could use any available method for solving $(A - \xi_j I)d = v$ for each j. In particular, one could exploit the fact that $A - \xi_j I$ is normal and complex symmetric to devise an efficient minimum residual approach (cf. [28], [16, Theorem 3.4]) that would yield a termwise (with respect to the partial fraction expansion) optimal method for approximating $x_\star$. We refer to [35] for a discussion of a closely related approach from a polynomial point of view.

5.2. Residual and error in the Krylov approximation. We next provide a direct bound for the error of the Krylov approach in (5.2), defined as

$$e_m^K := \mathcal{R}_\nu(A)v - x_m^K = \sum_{j=1}^{\nu} \tau_j((A - \xi_j I)^{-1}v - V_m y_m^{(j)}). \tag{5.3}$$

We can also introduce the residual vector

$$r_m^K = \sum_{j=1}^{\nu} \tau_j(v - (A - \xi_j I)V_m y_m^{(j)}),$$

which is a linear combination of the ν residuals of the partial fraction expansion. It is remarkable that the error e_m^K and the residual r_m^K written in the expansion form are a fully algebraic representation of the error ϵ_m and of the generalized residual ρ_m defined using the Cauchy integral form in [26, section 6.3, p. 1566]; see also [22] where a similar connection is made. Denote with $r_m^{(j)}$ the residual of the jth term in the expansion. Since $r_m^{(i)} = v - (A - \xi_j I)V_m y_m^{(i)} = v - V_{m+1}H_{m+1,m}y_m^{(i)} = -v_{m+1}h_{m+1,m}e_m^* y_m^{(i)}$, we have

$$r_m^K = -v_{m+1}h_{m+1,m}\sum_{j=1}^{\nu} \tau_j e_m^* y_m^{(j)}, \qquad \text{with} \quad V_m^* r_m^K = 0, \tag{5.4}$$

and $e_m^K = -h_{m+1,m}\sum_{j=1}^{\nu}(A - \xi_j I)^{-1}v_{m+1}\tau_j e_m^* y_m^{(j)}$. Note that the Galerkin residual $r_m^G = \Phi_\nu(A)v - \Psi_\nu(A)x_m^G$ also satisfies $V_m^* r_m^G = 0$, however, r_m^G does not belong to the subspace generated by v_{m+1}.

The quantity $h_{m+1,m}|e_m^* y_m^K|$ is a well established a posteriori estimate of the error of the Krylov approximation [26, 41, 43]. As shown in (5.2), the Krylov approximation is given by $V_m y_m^K = \tau_0 V_m e_1 + V_m \sum_{j=1}^{\nu} \tau_j y_m^{(j)}$ so that $e_m^* y_m^K = \sum_{j=1}^{\nu} \tau_j e_m^* y_m^{(j)}$ for $m > 1$, and

$$h_{m+1,m}|e_m^* y_m^K| = \left| h_{m+1,m}\sum_{j=1}^{\nu} \tau_j e_m^* y_m^{(j)} \right| = \|r_m^K\|.$$

Hence, the a posteriori convergence estimate is precisely the norm of the residual r_m^K; see [7] for similar considerations.

We next bound the error of the Krylov approximation. We first need a lemma that bounds the error obtained when solving each system in the partial fraction expansion; see [29, formula (1.3)] for a qualitatively similar result.

LEMMA 5.2. *Let $M_j = A - \Re(\xi_j)I$, and let α_j, β_j be the largest and smallest eigenvalues of M_j in absolute value, respectively. Moreover, let $\widehat{\beta}_j = \beta_j$ if M_j is definite, which includes the case $\Im(\xi_j) = 0$, otherwise $\widehat{\beta}_j = 0$. Then, with the notation above, the error $e_m^{(j)} = (A - \xi_j I)^{-1}v - V_m(H_m - \xi_j I)^{-1}e_1$ satisfies*

$$\|e_m^{(j)}\| \le \widehat{\kappa}_j \|(A - \xi_j I)^{-1}v\| \frac{2}{\rho_j^m + 1/\rho_j^m},$$

where $\widehat{\kappa}_j = |\alpha_j - i\Im(\xi_j)|/|\widehat{\beta}_j - i\Im(\xi_j)|$ and ρ_j is the solution to the problem in (4.2).

Proof. The proof is inspired by that of [15, Theorem 4]. Let $d_\star = (A - \xi_j I)^{-1}v$, $d_m = V_m(H_m - \xi_j I)^{-1}e_1$, $e_m^{(j)} = d_\star - d_m$, $e_0^{(j)} = d_\star$, and $r_m = (A - \xi_j I)e_m^{(j)} = M_j e_m^{(j)} - i\Im(\xi_j)e_m^{(j)}$. In the following we will omit the superscript in the error, and we will use $\langle u, v\rangle = u^*v$. We have $\langle r_m, e_m\rangle = \langle M_j e_m, e_m\rangle - \langle i\Im(\xi_j)e_m, e_m\rangle$ so that

$$|\langle r_m, e_m\rangle|^2 = (\langle M_j e_m, e_m\rangle)^2 + \Im(\xi_j)^2(\langle e_m, e_m\rangle)^2 \ge (\widehat{\beta}_j^2 + \Im(\xi_j)^2)\|e_m\|^4.$$

We also have $\|r_m\|^2 = \|M_j e_m\|^2 + \Im(\xi_j)^2\|e_m\|^2 \le (\alpha_j^2 + \Im(\xi_j)^2)\|e_m\|^2$. Recalling that $r_m \perp K_m(M_j, v)$, for any $u \in K_m(M_j, v)$ we have

$$\begin{aligned} |\langle r_m, e_m\rangle|^2 &= |\langle r_m, d_\star - u\rangle|^2 \\ &\le \|r_m\|^2 \, \|d_\star - u\|^2 \le (\alpha_j^2 + \Im(\xi_j)^2)\|e_m\|^2 \, \|d_\star - u\|^2. \end{aligned}$$

Collecting all bounds, we obtain $(\widehat{\beta}_j^2 + \Im(\xi_j)^2)\|e_m\|^4 \le (\alpha_j^2 + \Im(\xi_j)^2)\|e_m\|^2 \, \|d_\star - u\|^2$, that is,

$$\|e_m\|^2 \le \frac{\alpha_j^2 + \Im(\xi_j)^2}{\widehat{\beta}_j^2 + \Im(\xi_j)^2} \, \|d_\star - u\|^2.$$

Since u is arbitrary, we take as u the solution to the problem

$$\min_{u \in K_m} \|d_\star - u\| = \min_{p \in \mathbb{P}_m, p(0)=1} \|p(A - \xi_j I)e_0\|.$$

Therefore, arguing as in the proof of Theorem 4.3, we have

$$\min_{p \in \mathbb{P}_m, p(0)=1} \|p(A - \xi_j I)e_0\| \le \|e_0\| \, \frac{2}{\rho_j^m + 1/\rho_j^m},$$

and the final result follows. ∎

In the lemma above, for ν odd, $M_j = A - \Re(\xi_j)I$ is negative definite for the real (positive) Padé pole, whereas in the case of the real (negative) Chebyshev pole, $M_j = -A - \Re(\xi_j)I$ is positive definite (we recall here the change of sign in the case of Chebyshev function). This ensures that the lemma holds for both Chebyshev and Padé rational functions. Moreover, for definite M_j, a sharper bound can be obtained; see Lemma 7.1.

THEOREM 5.3. *Assume the previous notation holds and that the poles ξ_j are distinct. Let $M_j = A - \Re(\xi_j)I$ and α_j, β_j be the largest and smallest eigenvalues of M_j in absolute value, respectively. Then, with the notation of Lemma* 5.2*,*

$$\|e_m^K\| \le 2\sum_{j=1}^{\nu} \left(|\tau_j|\widehat{\kappa}_j\|(A-\xi_j I)^{-1}v\|\right) \frac{1}{\rho_j^m + 1/\rho_j^m}.$$

Proof. Using the definition of e_m^K in (5.3) and its partial fraction representation, we have

$$\|e_m^K\| \le \sum_{j=1}^{\nu} |\tau_j|\|e_m^{(j)}\|, \qquad e_m^{(j)} = (M_j - i\Im(\xi_j)I)^{-1}v - V_m y_m^{(j)}. \tag{5.5}$$

From Lemma 5.2, it follows $\|e_m^{(j)}\| \le 2\widehat{\kappa}_j\|(A-\xi_j I)^{-1}v\|/(\rho_j^m + 1/\rho_j^m)$. Substituting in (5.5), the result follows. ∎

For $\widehat{\kappa}_j \approx 1$, we have

$$|\tau_j|\widehat{\kappa}_j\|(A-\xi_j I)^{-1}v\| \approx |\tau_j|\|(A-\xi_j I)^{-1}v\| \le \max_{\lambda\in[\alpha,\beta]} \frac{|\tau_j|}{|\lambda - \xi_j|},$$

where the last term is precisely the factor in the bound of the Galerkin approach; see Remark 4.4. The condition $\widehat{\kappa}_j \approx 1$ is met, for instance, when $[\alpha, \beta] \approx [-1, 0]$. Indeed, in this case, the poles in the partial fraction expansion are significantly larger in absolute value than the values in $[-1, 0]$ so that $\widehat{\kappa}_j = |(\alpha - \xi_j)/(\beta - \xi_j)| \approx |\xi_j|/|\xi_j| = 1$.

Roughly speaking, the result of Theorem 5.3 tells us that the convergence rate of the bound is driven by the convergence of the single systems $(A - \xi_j I)d = v$, $j = 1, \ldots, \nu$ with the FOM method. We also explicitly observe that the convergence rate of systems with matrix $A - \xi_j I$ is significantly different from that observed with A. This is due to the fact that the spectrum of $A - \xi_j I$ is in a complex line segment near ξ_j, sufficiently far away from the origin to ensure fast convergence. These observations should be compared to those in [24] and earlier literature, where the rate of convergence of iterative solvers for systems with A was observed to be an unsatisfactory tool for describing the convergence rate in the exponential approximation.

Finally, we have experimented with the bounds for the Krylov approximation as in Example 4.5, and numerical results almost identical to those of Figure 4.1 were obtained.

6. Some computational advantages of rational function approximation. The use of rational functions and their partial fraction expansion allows us to exploit and generalize known properties of Krylov subspace methods for the solution of algebraic linear systems. In particular, in this section we show that our formulation makes it easy to interpret recently proposed strategies.

As an immediate consequence of our formulation, in the following proposition we provide a bound for the components of the solution y_m^K. Analyzing the pattern of the solution components is of interest when using some recently developed preconditioning strategies [49]. In these promising techniques, the Krylov subspace with respect to the matrix $(I - \gamma A)^{-1}$ is generated for a conveniently chosen scalar γ. An approximation to $\exp(A)$ is then obtained in this rational space. Since $(I - \gamma A)^{-1}$ cannot be applied exactly, it is shown that the accuracy with which the inverse needs to be employed can be tied to the magnitude of the solution components in the generated Krylov subspace; we refer to [49, section 5] for a more comprehensive discussion.

PROPOSITION 6.1. *Assume the notation of the previous sections holds. Let* $x_m^K = V_m y_m^K$ *be the Krylov approximation to the exponential operator and assume that for* $k \le m$, *the matrix* $H_k - \xi_j I$ *is nonsingular for all poles* ξ_j, $j = 1, \dots, \nu$. *Then*

$$|e_k^* y_m^K| \le |e_k^* e_1 \tau_0| + \sum_{j=1}^{\nu} \frac{|\tau_j|}{\sigma_{\min}(H_m - \xi_j I)} \|r_{k-1}^{(j)}\|, \qquad k \le m,$$

where $\sigma_{\min}(H_m - \xi_j I)$ *is the smallest singular value of* $H_m - \xi_j I$, *and* $\|r_{k-1}^{(j)}\|$ *is the residual norm associated with the* j*th partial fraction expansion system in* $K_{k-1}(A, v)$.

Proof. Using the partial fraction expansion of $\mathcal{R}_\nu(H_m)$ we have

$$y_m^K = \tau_0 e_1 + \sum_{j=1}^{\nu} \tau_j (H_m - \xi_j I)^{-1} e_1 = \tau_0 e_1 + \sum_{j=1}^{\nu} \tau_j y_m^{(j)}, \tag{6.1}$$

from which

$$e_k^* y_m^K = e_k^* e_1 \tau_0 + \sum_{j=1}^{\nu} \tau_j e_k^* y_m^{(j)}. \tag{6.2}$$

It was shown in [46, Lemma 5.2] that if the matrix $H_k - \xi_j I$ is nonsingular for all $k = 1, \dots, m$ so that the residuals $r_{k-1}^{(j)}$ are well defined, then

$$|e_k^* y_m^{(j)}| \le |e_k^* e_1 \tau_0| + \frac{1}{\sigma_{\min}(H_m - \xi_j I)} \|r_{k-1}^{(j)}\|.$$

Substituting in (6.2), the result follows. ∎

The result of Proposition 6.1 shows that the components of y_m^K approximately decrease with the residuals of the shifted systems in the expansion. This fact can be exploited to rigorously show that the preconditioning technique proposed in [49] can be successfully implemented with an *inexact* application of $(I - \gamma A)^{-1}$ and a *relaxed* tolerance [40].

If convergence is not fast, a problem encountered with Krylov subspace approximation is that the maximum allowed approximation space dimension is limited by memory restrictions, and thus some form of restarting must be devised. We next show that our rational function framework provides a simple though effective way to describe a recently proposed restarting strategy. Once the vector $x_m^K = \tau_0 v + \sum_{j=1}^{\nu} V_m y_m^{(j)}$ is determined, a new approximation space can be obtained as $K_m(A, v_{m+1})$ with $v_{m+1} = V_{m+1} e_{m+1}$ and associated matrix $V_m^{(1)}$ so that the approximation can be updated as

$$x_m^K = \left(\tau_0 v + \sum_{j=1}^{\nu} V_m y_m^{(j)} \right) + \sum_{j=1}^{\nu} V_m^{(1)} (y_m^{(j)})^{(1)}, \tag{6.3}$$

with obvious notation for $(y_m^{(j)})^{(1)}$. We rewrite (6.3) as $x_m^K = \tau_0 v + \sum_{j=1}^{\nu} (V_m y_m^{(j)} + V_m^{(1)} (y_m^{(j)})^{(1)})$. Recalling that this approach approximates $\tau_0 v + \sum_{j=1}^{\nu} (A - \xi_j I)^{-1} v$, this relation shows that each term in the expansion formula is nothing but the approximate solution obtained by restarted FOM applied to each shifted system separately. This last statement can be verified by recalling that all system residuals $r_m^{(j)}$ are collinear

with v_{m+1} (cf. (5.4)) so that the method can be restarted with the same approximation space for all systems [45]. Further use of the properties of the restarted FOM method, see, e.g., [44], shows that this restarting procedure corresponds to Algorithm 2 in [14] when used with rational functions; see also [27] for more results on the derivation of the restarted approximation method.

Finally, appropriate Krylov subspace approaches and rational function approximations can be used to preserve geometric properties of the exponential of skew-symmetric matrices; this is the subject of current investigation [30].

7. The role of $\|A\|$ in the Padé approximation. In the previous sections we assumed that $\|A\|$ was not much greater than unit. For the case of Chebyshev rational functions, this is an unnecessary constraint, as good approximations to e^{-x} can be obtained for $x \in [0, +\infty)$; see, e.g., [8]. In the right plot of Figure 7.1 we report the convergence curve and its bound for the Krylov approximation using Chebyshev rational functions with $\nu = 14$ for the matrix in Example 7.3. In the plot, the ideal bound (2.1) is also reported.

Padé rational function approximation is effective for $\|A\|$ close to the origin. Otherwise, a procedure called scaling and squaring is commonly employed in conjunction with Padé functions that allows one to compute an approximation to the exponential of a conveniently scaled matrix; see, e.g., [21]. The procedure amounts to finding the smallest integer $s \geq 0$ such that $\|A\|_\infty/2^s$ is less than a prescribed value, a common value being $1/2$. Recent work by Higham has shown that this latter value can be significantly relaxed, depending on the rational function degree used in the approximation [23]. For the sake of simplicity, here we limit ourselves to the case $\|A\|_\infty 2^{-s} \leq \frac{1}{2}$. Setting $\widetilde{A} = A/2^s$, the scaling and squaring method produces the matrix $B = (\mathcal{R}_\nu(\widetilde{A}))^{-1}$ so that the sought after approximation is obtained as $\exp(A) \approx B^{2^s}$, where the operation is performed by repeated squaring of B. Within the Krylov approximation (1.2), scaling and squaring can be conveniently performed at each step m as $\exp(H_m)e_1 \approx (\mathcal{R}_\nu(H_m/2^s))^{2^s} e_1$, where s may vary with m [43].

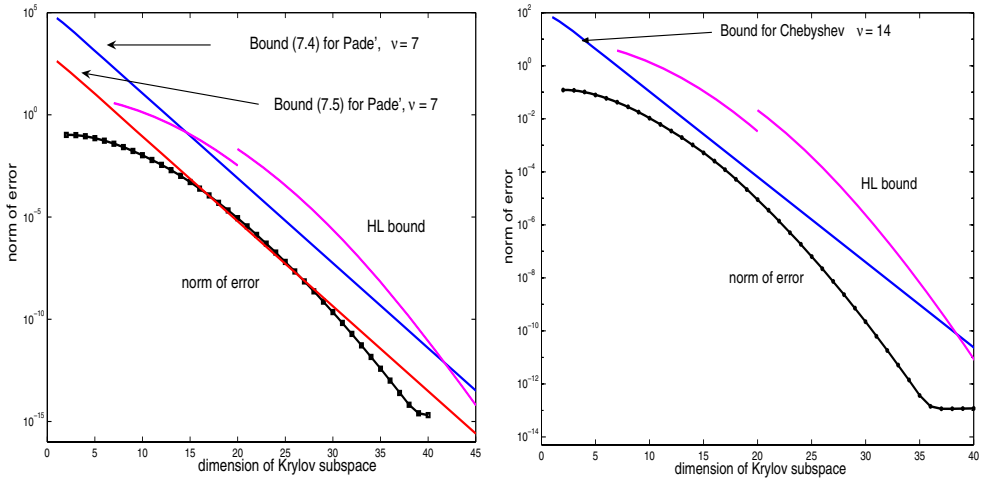


FIG. 7.1. *Example 7.3 with $\|A\| \gg 1$, Krylov approximation. Left. Padé rational function with scaling and squaring: convergence curve and upper bound for $\nu = 7$, and ideal bound. Right. Chebyshev rational function: convergence curve and upper bound for $\nu = 14$, and ideal bound.*

We next show that this corrected scheme can be included in our theoretical analysis. Let $s \geq 0$ be the smallest integer such that $\widetilde{A} := A/2^s$ satisfies $\|\widetilde{A}\| \leq 1/2$. The construction of the Krylov subspace $K_m(\widetilde{A}, v)$ corresponds to scaling H_m in (1.1) by the quantity 2^s, independent of m but dependent on $\|A\|$. Since $\|H_m\| \leq \|A\|$ for all $m \geq 0$, this approach is more conservative than the one that scales H_m by a different quantity at each iteration. We then approximate $\exp(A)v$ as

$$\exp(A)v \approx \left(\mathcal{R}_\nu(\widetilde{A})\right)^{2^s} v \approx V_m \left(\mathcal{R}_\nu(\widetilde{H}_m)\right)^{2^s} e_1 \equiv V_m y_m^K.$$

In the scaling and squaring method, the Padé approximant has poles of multiplicity 2^s so that the partial fraction expansion is given by

$$\mathcal{R}_\nu(\zeta)^{2^s} = \widetilde{\tau}_0 + \sum_{j=1}^{\nu} \sum_{\ell=1}^{2^s} \frac{\widetilde{\tau}_{j,\ell}}{(\zeta - \xi_j)^\ell}, \tag{7.1}$$

where $\xi_1, \ldots, \xi_\nu$ are the roots of Ψ_ν and

$$\widetilde{\tau}_{j,\ell} = \frac{1}{(\ell-1)!} \frac{d^{\ell-1}}{(d\zeta)^{\ell-1}} \frac{(\zeta - \xi_j)^{2^s} \Phi_\nu(\zeta)^{2^s}}{\Psi_\nu(\zeta)^{2^s}} \bigg|_{\zeta=\xi_j}.$$

Therefore, we can write

$$\left(\mathcal{R}_\nu(\widetilde{A})\right)^{2^s} v = \widetilde{\tau}_0 v + \sum_{j=1}^{\nu} \sum_{\ell=1}^{2^s} \widetilde{\tau}_{j,\ell} (\widetilde{A} - \xi_j I)^{-\ell} v.$$

Because of the multiple poles, the computation of $\widetilde{\tau}_{j,\ell}$ may be very ill conditioned, therefore, we do not advocate implementing this procedure. Instead, it provides a way to theoretically justify this computational strategy within our polynomial framework. Once again, setting $x_\star = \mathcal{R}_\nu(\widetilde{A})^{2^s} v$, we use the bound

$$\|\exp(A)v - V_m y_m^K\| \leq \|\exp(A)v - x_\star\| + \|x_\star - V_m y_m^K\|,$$

in which the magnitude of the first term depends on the accuracy of the rational approximation, whereas only the second term depends on the accuracy in the Krylov subspace. By approximating each inverse power in the Krylov subspace, we have

$$\begin{aligned} x_\star &= \widetilde{\tau}_0 v + \sum_{j=1}^{\nu} \sum_{\ell=1}^{2^s} \widetilde{\tau}_{j,\ell} (\widetilde{A} - \xi_j I)^{-\ell} v \\ &\approx \widetilde{\tau}_0 v + \sum_{j=1}^{\nu} \sum_{\ell=1}^{2^s} \widetilde{\tau}_{j,\ell} V_m (\widetilde{H}_m - \xi_j)^{-\ell} e_1 = V_m \left(\mathcal{R}_\nu(\widetilde{H}_m)\right)^{2^s} e_1. \end{aligned}$$

In spite of the presence of multiple poles, we next show that the error can be bounded similarly to what we did for the simple poles. In fact, the bound for the error in practice is not influenced by the presence of the scaling and squaring method.

We first derive a bound for the errors in the partial fraction expansion, which holds for symmetric and definite matrices.

LEMMA 7.1. *Let $M_j = A - \Re(\xi_j) I$ be a symmetric definite matrix, and let α_j, β_j be the largest and smallest eigenvalues of M_j in absolute value, respectively,*

so that $\kappa_j = \alpha_j/\beta_j > 0$ *is the condition number of* M_j. *Then, the error* $e_m^{(j)} = (A-\xi_j I)^{-1}v - V_m(H_m-\xi_j I)^{-1}e_1$ *satisfies*

$$\|e_m^{(j)}\| \le \kappa_j^{\frac{1}{2}}\eta_j\|(A-\xi_j I)^{-1}v\|\frac{2}{\rho_j^m + 1/\rho_j^m},$$

where $\eta_j = \left(1+\frac{\Im(\xi_j)^2(\kappa_j-1)^2}{4\kappa_j\Im(\xi_j)^2+\alpha_j^2(\kappa_j+1)^2}\right)^{\frac{1}{2}}$, *and* ρ_j *is the solution to the problem in* (4.2).

Proof. We assume that M_j is positive definite, otherwise we can work with $-M_j$. For each $j = 1,\dots,\nu$, $y_m^{(j)}$ is computed by imposing a Galerkin condition on the corresponding residual $v-(M_j - i\Im(\xi_j)I)V_m y_m^{(j)}$. It was shown in [15, Theorem 4] that if M_j is definite, then

$$\|e_m^{(j)}\|_{M_j} \le 2\frac{\eta_j}{\rho_j^m + \frac{1}{\rho_j^m}}\|e_0^{(j)}\|_{M_j}.$$

Then $\|e_m^{(j)}\| \le \|M_j^{-1/2}\|\|e_m^{(j)}\|_{M_j}$, $\|e_0^{(j)}\|_{M_j} \le \|M_j^{1/2}\|\|e_0^{(j)}\|$, where $\|M_j^{1/2}\|$ denotes the norm of $M_j^{1/2}$, induced by the vector 2-norm, and $\|e_0^{(j)}\| = \|(A-\xi_j I)^{-1}v\|$. ∎

The bound of Lemma 7.1 is analogous to that in Lemma 5.2 for $\widehat{\kappa}_j \approx 1$, which is the case when $\|A\| \le 1$. The new bound of Lemma 7.1 provides a sharper estimate for $\|A\| > 1$ and M_j definite; therefore it can be used with Padé rational functions.

THEOREM 7.2. *With the notation and definitions of Lemma* 7.1, *let* ξ_j, $j = 1,\dots,\nu$ *be the roots of* $\Psi_\nu(\lambda)$. *Assume that the eigenvalues of* A *are contained in* $[\alpha,\beta]$, *and that* $\frac{1}{2^s}\frac{|\lambda|}{|\xi_j|} \ll 1$ *for* $\lambda \in [\alpha,\beta]$. *Let* $x_\star = \left(\mathcal{R}_\nu(\widetilde{A})\right)^{2^s} v$ *and* $y_m^K = \left(\mathcal{R}_\nu(\widetilde{H}_m)\right)^{2^s} e_1$. *Finally, for* $j = 1,\dots,\nu$, *let*

$$\widehat{\tau}_j := \max_{\ell=1,\dots,2^s}\frac{|\widetilde{\tau}_{j,\ell}|}{|(-\xi_j)^{\ell-1}|}.$$

Then,

$$\|x_\star - V_m y_m^K\| \lesssim 2^{s+1}\sum_{j=1}^{\nu}|\widehat{\tau}_j|\kappa_j^{\frac{1}{2}}\frac{\eta_j}{\rho_j^m+\frac{1}{\rho_j^m}}\|(A-\xi_j I)^{-1}\|,$$

where $\lesssim$ *means that higher order terms are omitted.*

Note that the bound is in terms of A and not of its scaled counterpart $\widetilde{A}$.

Proof. Recalling (7.1), we have

$$\begin{aligned} R_{2^s}(\lambda) := \left(\mathcal{R}_\nu(\frac{\lambda}{2^s})\right)^{2^s} &= \widetilde{\tau}_0 + \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s}\frac{\widetilde{\tau}_{j,\ell}}{(\frac{\lambda}{2^s}-\xi_j)^\ell} \\ &= \widetilde{\tau}_0 + \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s}\frac{\widetilde{\tau}_{j,\ell}}{\xi_j^\ell(\frac{\lambda}{2^s\xi_j}-1)^\ell}. \end{aligned}$$

For $\frac{|\lambda|}{2^s|\xi_j|} \ll 1$, we have $(\frac{\lambda}{2^s\xi_j}-1)^\ell \approx (-1)^\ell(\frac{\ell}{2^s\xi_j}\lambda - 1)$ so that

$$R_{2^s}(\lambda) \approx \widetilde{\tau}_0 + \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s}\frac{\widetilde{\tau}_{j,\ell}}{(-\xi_j)^\ell(\frac{\ell\lambda}{2^s\xi_j}-1)} = \widetilde{\tau}_0 - \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s}\frac{\widetilde{\tau}_{j,\ell}}{(-\xi_j)^{\ell-1}(\frac{\ell}{2^s}\lambda-\xi_j)}.$$

Let $\omega_{j,\ell} := \widetilde{\tau}_{j,\ell}/(-\xi_j)^{\ell-1}$. Then,

$$\begin{aligned}\|x_\star - V_m y_m^K\| &= \|R_{2^s}(A)v - V_m R_{2^s}(H_m)e_1\| \\ &\approx \left\| \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s} \omega_{j,\ell} \left(\left(\frac{\ell}{2^s}A - \xi_j I\right)^{-1} v - V_m \left(\frac{\ell}{2^s}H_m - \xi_j I\right)^{-1} e_1 \right) \right\| \\ &\le \sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s} |\omega_{j,\ell}| \|\epsilon_{j,\ell}\|,\end{aligned}$$

where $\epsilon_{j,\ell} = (\frac{\ell_j}{2^s}A - \xi_j I)^{-1}v - V_m(\frac{\ell_j}{2^s}H_m - \xi_j I)^{-1}e_1$. Note that the approximation above is ensured to hold for H_m because the eigenvalues of $H_m = V_m^T A V_m$ are also contained in $[\alpha, \beta]$ so that $\frac{|\theta|}{2^s|\xi_j|} \ll 1$ for any eigenvalue θ of H_m. Using Lemma 7.1 and appropriately modifying the notation, we have

$$\|\epsilon_{j,\ell}\| \le 2\left(\kappa\left(\frac{\ell}{2^s}A - \Re(\xi_j)I\right)\right)^{\frac{1}{2}} \frac{\eta_{j,\ell}}{\rho_{j,\ell}^m + \frac{1}{\rho_{j,\ell}^m}} \left\| \left(\frac{\ell}{2^s}A - \xi_j I\right)^{-1} v \right\|.$$

Since $\ell/2^s \le 1$ for $\ell \in \{1, \dots, 2^s\}$, we have $\|(\ell_j/2^s A - \xi_j I)^{-1}\| \le \|(A - \xi_j I)^{-1}\|$ and

$$\kappa\left(\frac{\ell_j}{2^s}A - \Re(\xi_j)I\right)\eta_{j,\ell}^2 \le \kappa(A - \Re(\xi_j)I))\eta_{j,2^s}^2 \equiv \kappa_j \eta_j^2; \tag{7.2}$$

see the appendix for a proof of this inequality. From the definition of $\rho_{j,\ell}$, it follows that

$$\begin{aligned}\frac{|\frac{\ell}{2^s}\alpha - \Re(\xi_j) - i\Im(\xi_j)| + |\frac{\ell}{2^s}\beta - \Re(\xi_j) - i\Im(\xi_j)|}{\frac{\ell}{2^s}|\alpha - \beta|} &= \frac{|\alpha - \frac{2^s}{\ell}\xi_j| + |\beta - \frac{2^s}{\ell}\xi_j|}{|\alpha - \beta|} \\ &\ge \frac{|\alpha - \xi_j| + |\beta - \xi_j|}{|\alpha - \beta|},\end{aligned}$$

from which it follows that $\rho_{j,\ell} \ge \rho_{j,2^s} > 1$, and because of monotonicity, we obtain

$$\frac{1}{\rho_{j,\ell}^m + \frac{1}{\rho_{j,\ell}^m}} \le \frac{1}{\rho_{j,2^s}^m + \frac{1}{\rho_{j,2^s}^m}}.$$

Since $\rho_{j,2^s}$ corresponds to the convergence with $(A - \xi_j I)$, we can set $\rho_{j,2^s} = \rho_j$. Collecting all bounds, we obtain

$$\|\epsilon_{j,\ell}\| \le 2\kappa_j^{\frac{1}{2}} \frac{\eta_j}{\rho_j^m + \frac{1}{\rho_j^m}} \|(A - \xi_j I)^{-1}\|.$$

Since $|\omega_{j,\ell}| \le |\widehat{\tau}_j|$, we finally have

$$\begin{aligned}\sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s} |\omega_{j,\ell}|\|\epsilon_{j,\ell}\| &\le 2\sum_{j=1}^{\nu}\sum_{\ell=1}^{2^s} |\widehat{\tau}_j|\kappa_j^{\frac{1}{2}} \frac{\eta_j}{\rho_j^m + \frac{1}{\rho_j^m}} \|(A - \xi_j I)^{-1}\| \\ &\le 2 \cdot 2^s \sum_{j=1}^{\nu} |\widehat{\tau}_j|\kappa_j^{\frac{1}{2}} \frac{\eta_j}{\rho_j^m + \frac{1}{\rho_j^m}} \|(A - \xi_j I)^{-1}\|. \qquad \square\end{aligned}$$

A few remarks are in order. We first notice that the bound in Theorem 7.2 fully mimics the bound in Theorem 5.3. The only relevant difference is the presence of

the additional scaling factor 2^s. The two bounds also differ for the terms $\widehat{\tau}_j$. In our experiments, however, replacing $\widehat{\tau}_j$ with τ_j did not seem to affect the approximation.

Although the approximation $V_m y_m^K$ with Padé does require scaling and squaring, the convergence rate does not seem to depend on it. The proof suggests that one could use the following estimate:

$$\|x_\star - V_m y_m^K\| \approx \|\mathcal{R}_\nu(A)v - V_m \mathcal{R}_\nu(H_m)e_1\|, \tag{7.3}$$

where, however, for $\|A\| \gg 1$, neither $\mathcal{R}_\nu(A)v$ is a good approximation to $x_\star$, nor $V_m\mathcal{R}_\nu(H_m)e_1$ is a good approximation of $V_m y_m^K$.

Example 7.3. We consider the 1001×1001 diagonal matrix A in [24] with entries uniformly distributed in $[-40, 0]$ and the random vector v with uniformly distributed values in $[0, 1]$ (Matlab function `rand`) and unit norm. In Figure 7.1 we report the convergence of the error norm for the Krylov method, together with the ideal bound (2.1) from [24] (referred to as HL bound). In the plot, we also report the following estimate, derived from Theorem 7.2 by replacing $\widehat{\tau}_j$ with τ_j, the latter being the coefficients in the partial fraction expansion of $\mathcal{R}_\nu(\lambda)$,

$$\|x_\star - V_m y_m^K\| \lesssim 2^{s+1} \sum_{j=1}^{\nu} |\tau_j| \frac{\kappa_j}{\rho_j^m + \frac{1}{\rho_j^m}} \|(A - \xi_j I)^{-1}\|, \tag{7.4}$$

$$\|x_\star - V_m y_m^K\| \lesssim 2 \sum_{j=1}^{\nu} |\tau_j| \frac{\kappa_j}{\rho_j^m + \frac{1}{\rho_j^m}} \|(A - \xi_j I)^{-1}\|. \tag{7.5}$$

The new bound for the error norm does not significantly differ from those observed in previous sections, although in this case the actual Padé approximation is performed with scaling and squaring. In other words, convergence seems to be only driven by the spectral properties of the matrix, as stressed by the bound on the convergence rate that is based on polynomial estimates of the shifted spectrum of A.

8. Recovering superlinear convergence. Our bounds for both Galerkin and Krylov approximations with rational functions of degree ν predict *linear* convergence as the Krylov subspace dimension m increases. The superlinear convergence expressed, say, in the first bound of Theorem 2.1, can be recovered as a function of the degree of the rational function. To explain this approach, we first recall that the bounds in Theorem 2.1 were determined in two steps in [24, Theorem 2]. In the first step, the error is bounded by means of the Cauchy integral on a curve Γ, which is the boundary of a piecewise smooth bounded region containing the numerical range of A. The second step amounts to conveniently choosing Γ so as to appropriately bound the Cauchy integral. In particular, the selected curve is a parabola whose right-most point turns out to be equal to $\gamma = m^2/(4\rho)$ in the notation of Theorem 2.1 (cf. [24, pp. 1916-1917]). Therefore, the chosen curve moves away from the spectrum with m^2 as the Krylov subspace dimension m increases.

The partial fraction expansion of $\mathcal{R}_\nu$ may be viewed as a way to approximate the Cauchy integral representation of $\exp(\lambda)$ for a fixed integration curve passing through the expansion poles. It is known that in both cases of Padé and Chebyshev approximations $\mathcal{R}_\nu$, the values ξ_j increase as ν grows. Therefore, a larger degree ν corresponds to selecting a curve Γ that is farther away from the spectrum. In other words, we expect better approximation of the bounds with small ν at an early convergence stage, whereas larger ν are required to appropriately bound the convergence curve as the dimension of the Krylov subspace increases. For the Chebyshev approximation, superlinear convergence as ν increases may be observed in the right plot of

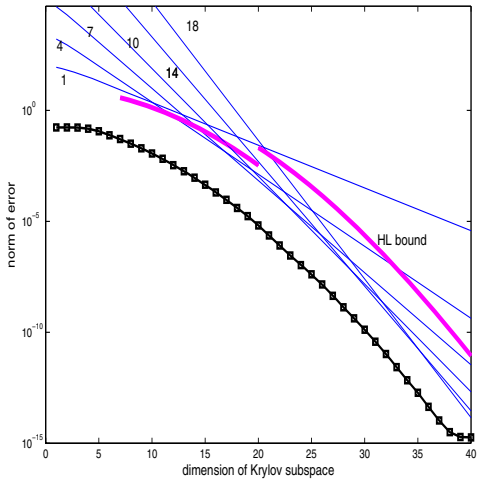


FIG. 8.1. *Example* 7.3*: Krylov approximation with Padé rational function; bound in* (7.4) *for* $\nu = 1, 4, 7, 10, 14, 18$. *The ideal bounds of Theorem 2.1 are also reported (labeled "HL bound").*

Figure 4.1, where the straight line for $\nu = 7$ better represents the early convergence stage, whereas the line for $\nu = 14$ sharply bounds the convergence curve for larger m.

In the case of Padé approximation, this behavior can be better formalized, yielding a relation between the approximation degree and the dimension of the Krylov subspace. Using [1, Theorem 5.7.3], we know that the poles of the Padé approximant $\mathcal{R}_\nu$ of degree ν are located in the complex annulus[2]

$$2 \cdot 0.27\nu \le |z| \le 2\nu + \frac{4}{3}.$$

Therefore, if we wish to employ poles on a curve that intersects the positive real semi-axis close to the "optimal value" $m^2/(4\rho)$, we require that ν satisfies $2 \cdot 0.27\nu \le \frac{m^2}{4\rho} \le 2\nu + \frac{4}{3}$, that is,

$$\frac{m^2}{8\rho} - \frac{2}{3} \le \nu \le \frac{1}{0.27}\frac{m^2}{8\rho}. \tag{8.1}$$

Hence, if the degree ν is chosen within the bounds above, we expect that the Padé rational function approximates the Cauchy integral on a quasi-optimal curve, with respect to m, in the sense of [24]. For instance, below are the values of ν satisfying the lower bound in (8.1) for $m \in \{1, \ldots, 40\}$ and $\rho = 10$.

m	4	8	12	16	20	24	28	32	36	40
ν	0	1	2	3	5	7	10	13	16	20

The fidelity of this correspondence can be fully appreciated in Figure 8.1 for the data in Example 7.3. Reported are the actual error curve (solid with squares), and the ideal bounds of Theorem 2.1. We also display the upper bounds in (7.4) for

[2]It is also possible to detect a parabolic region that is pole-free; see [1, Theorem 5.7.4]. This would more closely mimic the choice of the curve Γ in [24]. However, for the sake of simplicity we limit our presentation to this more intuitive case.

the Padé approximation with ρ_j associated with the poles for $\nu = 1, 4, 7, 10, 14, 18$. Remarkably, the envelope of the reported straight lines completely reproduces the true convergence history, while the single lines well represent the local convergence behavior at the corresponding subspace dimension m (e.g., the line for $\nu = 7$ well approximates the convergence slope around iteration $m = 24$).

Finally, we stress that all reported bounds, including the ideal bounds, only depend on the spectral interval of the given matrix, and not on the location of the eigenvalues within this interval. Particular eigenvalue distributions may cause severe overestimations of the true convergence behavior; see, e.g., [10, section 3.2].

9. Conclusions and outlook. We have proposed a new analysis of Krylov subspace methods for approximating the action of matrix rational functions with specific application to the matrix exponential operator. We have shown a minimization property of one of the methods, we have provided new upper bounds for the approximation error by using partial fraction expansion, and we have theoretically justified some computational strategies.

Our analysis may be generalized to cases in which the rational function employed to approximate the exponential is different from the widely used Padé or Chebyshev, as in [38, 3]. Among these generalizations, the case of polynomial approximation of the exponential is particularly appealing for its simplicity; see, e.g., [11]. We also refer to [14, 35] for examples of polynomial interpolation at points different from the eigenvalues of H_m. Without great differences, one could extend our results to the case of *restricted denominator* (RD) rational forms, which are rational functions $R_{j,k} = \frac{q_j(x)}{(1+\rho x)^k}$, where $\rho \in \mathbb{R}$ and q_j is a polynomial of degree not greater than j. Such RD-rational forms have been introduced in [37] and were recently used to approximate the exponential of a matrix in [34]. Our theoretical results may provide some insight into the selection of the parameters involved in the definition of the RD-rational form.

Some of our results may be naturally extended to the case of non-Hermitian A, although optimality results of the Galerkin method do not carry over. Finally, the techniques and the analysis used in this paper could be adapted to the numerical approximation of other analytic functions in the Krylov subspace, such as $A^{\frac{1}{2}}$; see [13, 53] and references therein.

Appendix. We prove inequality (7.2) in the proof of Theorem 7.2.

Proof. Let $\xi_j = \xi_R + i\xi_I$ and $\kappa_{j,\ell} = \kappa(\frac{\ell}{2^s}A - \xi_j I)$. We have

$$\kappa_{j,\ell}\eta_{j,\ell}^2 = \kappa_{j,\ell} + \kappa_{j,\ell}\frac{\xi_I^2(\kappa_{j,\ell}-1)^2}{4\kappa_{j,\ell}\xi_I^2 + \left(\frac{\ell}{2^s}\alpha - \xi_R\right)(\kappa_{j,\ell}+1)^2}.$$

Let $\chi = \ell/2^s$ and $\kappa_{j,\ell} = (\chi\alpha - \xi_R)/(\chi\beta - \xi_R) =: f(\chi)$. Since $\alpha < \beta < 0$ and $\xi_R > 0$, we have $f'(\chi) = \xi_R(-\alpha+\beta)/(\chi\beta - \xi_R)^2 > 0$, that is, f is a monotonically increasing function so that $\kappa_{j,\ell} = f(\chi) \le f(1) = \kappa_{j,2^s}$ for $\chi \in (0, 1]$. With analogous reasoning, we have that $\sqrt{f(\chi)} + 1/\sqrt{f(\chi)} \ge \sqrt{f(1)} + 1/\sqrt{f(1)}$ and that $(\chi\alpha - \xi_R) \ge (\alpha - \xi_R)$ for $\chi \in (0, 1]$. Therefore,

$$\begin{aligned}\kappa_{j,\ell}\eta_{j,\ell}^2 &= \kappa_{j,\ell} + \kappa_{j,\ell}\frac{\xi_I^2(\kappa_{j,\ell}-1)^2}{\kappa_{j,\ell}\left(4\xi_I^2 + \left(\frac{\ell}{2^s}\alpha - \xi_R\right)\left(\sqrt{\kappa_{j,\ell}} + \frac{1}{\sqrt{\kappa_{j,\ell}}}\right)^2\right)} \\ &\le \kappa_{j,2^s} + \frac{\xi_I^2(\kappa_{j,2^s}-1)^2}{4\xi_I^2 + (\alpha - \xi_R)\left(\sqrt{\kappa_{j,2^s}} + \frac{1}{\sqrt{\kappa_{j,2^s}}}\right)^2} = \kappa_j\eta_j^2. \qquad \square\end{aligned}$$

Acknowledgments. We thank Oliver Ernst for making some early notes of [14] available, and Leonid Knizhnerman, whose comments helped improve the presentation of this paper and also led to the development of section 8. We also acknowledge insightful conversations with Marlis Hochbruck.

REFERENCES

[1] G. A. Baker and P. Graves-Morris, *Padé Approximants*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1996.

[2] C. Baldwin, R. Freund, and E. Gallopoulos, *A parallel iterative method for exponential propagation*, in Proceedings of the Seventh SIAM Conference on Parallel Processing for Scientific Computing, D. Bailey et al., eds., SIAM, Philadelphia, 1995, pp. 534–539.

[3] L. Bergamaschi and M. Vianello, *Efficient computation of the exponential operator for large, sparse, symmetric matrices*, Numer. Linear Algebra Appl., 7 (2000), pp. 27–45.

[4] A. Borici, *Computational methods for UV suppressed fermions*, J. Comput. Phys., 189 (2003), pp. 454–462.

[5] D. Calvetti, E. Gallopoulos, and L. Reichel, *Incomplete partial fractions for parallel evaluation of rational matrix functions*, J. Comput. Appl. Math., 59 (1995), pp. 349–380.

[6] P. Castillo and Y. Saad, *Preconditioning the matrix exponential operator with applications*, J. Sci. Comput., 13 (1999), pp. 225–302.

[7] E. Celledoni and I. Moret, *A Krylov projection method for systems of ODEs*, Appl. Numer. Math., 24 (1997), pp. 365–378.

[8] W. J. Cody, G. Meinardus, and R. S. Varga, *Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems*, J. Approx. Theory, 2 (March 1969), pp. 50–65.

[9] N. del Buono, L. Lopez, and R. Peluso, *Computation of exponentials of large sparse skew symmetric matrices*, SIAM J. Sci. Comput., 27 (2005), pp. 278–293.

[10] V. Druskin, A. Greenbaum, and L. Knizhnerman, *Using nonorthogonal Lanczos vectors in the computation of matrix functions*, SIAM J. Sci. Comput., 19 (1998), pp. 38–54.

[11] V. Druskin and L. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, USSR Comput. Math. Math. Phys., 29 (1989), pp. 112–121.

[12] V. Druskin and L. Knizhnerman, *Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.

[13] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.

[14] M. Eiermann and O. Ernst, *A Restarted Krylov Subspace Method for the Evaluation of Matrix Functions*, technical report, Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Germany, 2005.

[15] R. W. Freund, *On conjugate gradient type methods and polynomial preconditioners for a class of complex non-Hermitian matrices*, Numer. Math., 57 (1990), pp. 285–312.

[16] R. W. Freund, *Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 425–448.

[17] R. Friesner, L. Tuckerman, B. Dornblaser, and T. Russo, *A method for exponential propagation of large systems of stiff nonlinear differential equations*, J. Sci. Comput., 4 (1989), pp. 327–354.

[18] E. Gallopoulos and Y. Saad, *A parallel block cyclic reduction algorithm for the fast solution of elliptic equations*, Parallel Comput., 10 (1989), pp. 143–159.

[19] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by polynomial approximation methods*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 1236–1264.

[20] W. Gautschi, *Numerical Analysis. An Introduction*, Birkhäuser, Boston, 1997.

[21] G. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd ed., Baltimore, MD, 1996.

[22] A. Greenbaum, *Using the Cauchy Integral Formula and the Partial Fractions Decomposition of the Resolvent to Estimate $\|f(A)\|$*, technical report, Mathematics Department, University of Washington, Seattle, WA, 2000.

[23] N. J. Higham, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1179–1193.

[24] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[25] M. Hochbruck and C. Lubich, *Exponential integrators for quantum-classical molecular dynamics*, BIT, 39 (1999), pp. 620–645.
[26] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.
[27] M. Hochbruck and J. Niehoff, *private communication*, 2005.
[28] M. Huhtanen, *A Hermitian Lanczos method for normal matrices*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1092–1108.
[29] L. Knizhnerman, *On Adaptation of the Lanczos Method to the Spectrum*, Technical report #EMG-001-95-12, Schlumberger-Doll Research, Ridgefield, CT, 1995.
[30] L. Lopez and V. Simoncini, *Preserving geometric properties of the exponential matrix by block Krylov subspace methods*, Technical report, Dipartimento di Matematica, Università di Bologna, December, 2005.
[31] Y. Y. Lu, *Computing a matrix function for exponential integrators*, J. Comput. Appl. Math., 161 (2003), pp. 203–216.
[32] The MathWorks, Inc., *MATLAB 7*, September 2004.
[33] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.
[34] I. Moret and P. Novati, *RD-rational approximations of the matrix exponential*, BIT, 44 (2004), pp. 595–615.
[35] I. Moret and P. Novati, *Interpolating functions of matrices on zeros of quasi-kernel polynomials*, Numer. Linear Algebra Appl., 11 (2005), pp. 337–353.
[36] A. Nauts and R. Wyatt, *New approach to many state quantum dynamics: The recurisive residue generation method*, Phys. Rev. Lett., 51 (1983), pp. 2238–2241.
[37] S. P. Nørsett, *Restricted Padé approximations to the exponential function*, SIAM J. Numer. Anal., 15 (Oct. 1978), pp. 1008–1029.
[38] S. P. Nørsett and A. Wolfbrandt, *Attainable order of rational approximation to the exponential function with only real poles*, BIT, 17 (1977), pp. 200–208.
[39] T. Park and J. Light, *Unitary quantum time evolution by iterative Lanczos reduction*, J. Chem. Phys., 85 (1986), pp. 5870–5876.
[40] M. Popolizio and V. Simoncini, *Acceleration techniques for the approximation of the matrix exponential with Krylov subspace methods*, in preparation.
[41] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[42] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, 2003.
[43] R. B. Sidje, *Expokit: A software package for computing matrix exponentials*, ACM Trans. Math. Software, 24 (1998), pp. 130–156.
[44] V. Simoncini, *On the convergence of restarted Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 430–452.
[45] V. Simoncini, *Restarted full orthogonalization method for shifted linear systems*, BIT, 43 (2003), pp. 459–466.
[46] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. Sci. Comput., 25 (2003), pp. 454–477.
[47] D. E. Stewart and T. S. Leyk, *Error estimates for Krylov subspace approximations of matrix exponentials*, J. Comput. Appl. Math., 72 (1996), pp. 359–369.
[48] H. Tal-Ezer, *Spectral methods in time for parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1–11.
[49] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos Approximations to the Matrix Exponential*, Technical report, Mathematical Institute, Heinrich Heine Universität, Düsseldorf, Germany, 2004. To appear in SIAM J. Sci. Comput.
[50] H. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press, Cambridge, 2003.
[51] H. A. van der Vorst, *An iterative solution method for solving f(A)x=b, using Krylov subspace information obtained for the symmetric positive definite matrix A*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.
[52] R. S. Varga, *On higher order stable implicit methods for solving parabolic partial differential equations*, J. Math. Phys., XL (1961), pp. 220–231.
[53] G. Walz, *Computing the matrix exponential and other matrix functions*, J. Comput. Appl. Math., 21 (1988), pp. 119–123.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 636–654

TWO FINITE ELEMENT APPROXIMATIONS OF NAGHDI'S SHELL MODEL IN CARTESIAN COORDINATES*

ADEL BLOUZA†, FRÉDÉRIC HECHT‡, AND HERVÉ LE DRET‡

Abstract. We present a penalized version of Naghdi's model and a mixed formulation of the same model, in Cartesian coordinates for linearly elastic shells with little regularity, and finite element approximations thereof. Numerical tests are given that validate and illustrate our approach.

Key words. Naghdi's shell model, finite element approximation, penalty method, mixed formulation

AMS subject classifications. 74K25, 74S05, 65N30

DOI. 10.1137/050624339

1. Introduction. The purpose of this work is to approximate the solution of a formulation of Naghdi's shell model in Cartesian coordinates that is appropriate for linearly elastic shells that present curvature discontinuities. Our intent is to use finite elements of class C^0 and implement the approximation scheme as simply as possible using the general purpose, open source, two-dimensional finite element package FreeFem++ (http://www.freefem.org).

The formulation of Naghdi's model used here was introduced in Blouza [5] and Blouza and Le Dret [7]. This formulation is based on the idea of using a local basis-free formulation in which the unknowns are described in Cartesian coordinates instead of with covariant or contravariant components as is usually done in shell theory; see, for example, [4]. This formulation is able to accommodate shells with a $W^{2,\infty}$-midsurface, thus allowing for curvature discontinuities, as opposed to C^3 in the classical formalism, and makes for much simpler expressions. Although it was proven to be well-posed and to be the natural limit of the classical formulation when a sequence of regular midsurfaces converges to a $W^{2,\infty}$-midsurface in [7], the new formulation has not been used in a numerical setting to the best of our knowledge.

The literature on finite element approximation of two-dimensional shell models is huge. Let us just mention a few different approaches. Concerning conforming methods, the Ganev and Argyris triangles provide P_4 and P_5 interpolation with high order convergence in $O(h^4)$ when the solution is smooth enough. These elements were used, for example, to study the linear Koiter model for a C^3-shell in the classical covariant formulation; see [3]. This method was applied to approximate geometrically exact shell models in [10]. The Argyris element was also used in [15] for the numerical analysis of Koiter's model for shells with little regularity in the Cartesian formulation proposed in [6]. Let us also mention the three-dimensional shell element approach; see [11].

Still in the context of shells with little regularity, a nonconforming DKT element was used in [16] to approximate a Koiter model similar to the one introduced in [6].

*Received by the editors February 14, 2005; accepted for publication (in revised form) November 10, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/sinum/44-2/62433.html
†Laboratoire de mathématiques Raphaël Salem, Université de Rouen, B.P. 12, 76801 Saint-Étienne-du-Rouvray, France (Adel.Blouza@univ-rouen.fr).
‡Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 75252 Paris Cedex 05, France (hecht@ann.jussieu.fr, ledret@ccr.jussieu.fr).

This article is organized as follows. We first briefly recall the geometry of the midsurface and Naghdi shell formulation given in [5] and [7]. This formulation involves the infinitesimal rotation vector, a vector unknown that is tangent to the midsurface. Such tangency cannot be implemented in a conforming way in finite element spaces (a problem that does not occur in the classical covariant formulation).

Therefore, in section 3, we introduce a penalized version of Naghdi's model intended to approximate the above mentioned tangency. We prove the existence and uniqueness of the solution of the penalized model and establish its convergence to the solution of the original Naghdi problem when the penalization parameter tends to 0.

In section 4, we present a mixed formulation of Naghdi's model in which the tangency condition is enforced by a Lagrange multiplier. We prove that the inf-sup condition is satisfied and that the mixed problem is well-posed and solves the original Naghdi problem.

Section 5 is devoted to the finite element discretization of both formulations. The numerical analysis of the penalized version is rather standard. On the contrary, the discrete inf-sup condition for the mixed formulation does not follow from usual arguments, in the sense of those found in the discussion of approximations of the Stokes problem for instance.

Finally, we present a few numerical tests in section 6. The method was implemented in FreeFem++, a high level, free software package that manages mesh generation and adaption, matrix assembly, and linear system resolution automatically and generally uses as input the resolution domain, boundary conditions, and bilinear and linear forms. In addition to the resolution domain, boundary conditions, and applied loads, the only specific input required from the user by our code is the definition of the covariant vectors and of the partial derivatives of the normal vector. All the other geometrical and mechanical quantities, including the bilinear form, are code-generated. We present results for the standard hyperbolic paraboloid benchmark and for the planar-cylindrical $W^{2,\infty}$ shell considered in [15]. We also show results for a $W^{2,\infty}$ roof constructed on a basket-handle arch profile.

2. Notation. Greek indices and exponents take their values in the set $\{1, 2\}$ and Latin indices and exponents take their values in the set $\{1, 2, 3\}$. Unless otherwise specified, the summation convention for indices and exponents is assumed.

Let (e_1, e_2, e_3) be the canonical orthonormal basis of the Euclidean space $\mathbb{R}^3$. We note $u \cdot v$ the inner product of $\mathbb{R}^3$, $|u| = \sqrt{u \cdot u}$ the associated Euclidean norm and $u \wedge v$ the vector product of u and v.

Let ω be a domain of $\mathbb{R}^2$. We consider a shell whose midsurface is given by $S = \varphi(\bar{\omega})$, where $\varphi \in W^{2,\infty}(\omega; \mathbb{R}^3)$ is one-to-one mapping such that the two vectors

$$a_\alpha = \partial_\alpha \varphi$$

are linearly independent at each point $x \in \bar{\omega}$. We let

$$a_3 = \frac{a_1 \wedge a_2}{|a_1 \wedge a_2|}$$

be the unit normal vector on the midsurface at point $\varphi(x)$. The vectors a_i define the local covariant basis at point $\varphi(x)$. The contravariant basis a^i is defined by the relations $a_i \cdot a^j = \delta_i^j$, where δ_i^j is the Kronecker symbol. In particular $a_3(x) = a^3(x)$. Note that all these vectors are of class $W^{1,\infty}$. We let $a(x) = |a_1(x) \wedge a_2(x)|^2$ so that $\sqrt{a(x)}$ is the area element of the midsurface in the chart φ.

The first fundamental form of the surface is given in covariant components by

$$a_{\alpha\beta} = a_\alpha \cdot a_\beta.$$

Let $u \in H^1(\omega;\mathbb{R}^3)$ be a midsurface displacement and $r \in H^1(\omega;\mathbb{R}^3)$ a rotation of the normal vector (which is related to the actual infinitesimal rotation vector; see formula (15) below), i.e., H^1-regular mappings from ω into $\mathbb{R}^3$ such that r is tangent to the midsurface, given in covariant and Cartesian components by

$$u(x) = u_i(x)a^i(x) = u_i^c(x)e_i, \text{ where } u_i = u \cdot a_i \text{ and } u_i^c = u \cdot e_i,$$

and

$$r(x) = r_\alpha(x)a^\alpha(x) = r_i^c(x)e_i \text{ with the same meaning.}$$

Note that the tangency requirement is easily expressed in covariant coordinates, as it simply reads $r_3 = 0$, whereas it becomes

$$r_i^c(x)a_{3,i}^c(x) = 0 \text{ in } \omega \tag{1}$$

in Cartesian coordinates.

Let $a^{\alpha\beta\rho\sigma} \in L^\infty(\omega)$ be the elasticity tensor, which we assume to satisfy the usual symmetries and to be uniformly strictly positive. In the case of a homogeneous, isotropic material with Young modulus $E > 0$ and Poisson coefficient $0 \le \nu < 1/2$, we have

$$a^{\alpha\beta\rho\sigma} = \frac{E}{2(1+\nu)}(a^{\alpha\rho}a^{\beta\sigma} + a^{\alpha\sigma}a^{\beta\rho}) + \frac{E\nu}{1-\nu^2}a^{\alpha\beta}a^{\rho\sigma},$$

where $a^{\alpha\beta} = a^\alpha \cdot a^\beta$ are the contravariant components of the first fundamental form. In this context, the covariant components of the change of metric tensor read

$$\gamma_{\alpha\beta}(u) = \frac{1}{2}(\partial_\alpha u \cdot a_\beta + \partial_\beta u \cdot a_\alpha), \tag{2}$$

the covariant components of the change of transverse shear tensor read

$$\delta_{\alpha 3}(u,r) = \frac{1}{2}(\partial_\alpha u \cdot a_3 + r \cdot a_\alpha), \tag{3}$$

and the covariant components of the change of curvature tensor read

$$\chi_{\alpha\beta}(u,r) = \frac{1}{2}(\partial_\alpha u \cdot \partial_\beta a_3 + \partial_\beta u \cdot \partial_\alpha a_3 + \partial_\alpha r \cdot a_\beta + \partial_\beta r \cdot a_\alpha); \tag{4}$$

see [5] and [7]. Note that all these quantities make sense for shells with little regularity and are easily expressed with the Cartesian coordinates of the unknowns and geometrical data. For instance, we have

$$\partial_\alpha u \cdot a_\beta = \partial_\alpha u_i^c a_{\beta,i}^c \text{ and so on.}$$

We assume that the boundary $\partial\omega$ of the chart domain is divided into two parts: γ_0 of strictly positive one-dimensional measure on which the shell is clamped and a complementary part γ_1 on which the shell is subjected to applied tractions and moments.

Let us consider the function space, introduced in [5] and [7], which is appropriate in the context of shells with little regularity,

$$\mathcal{V} = \big\{(v,s) \in H^1(\omega;\mathbb{R}^3)^2;\ s\cdot a_3 = 0 \text{ in } \omega, v = s = 0 \text{ on } \gamma_0\big\}. \tag{5}$$

This space is endowed with the natural Hilbert norm

$$\|(v,s)\|_{\mathcal{V}} = \big(\|v\|^2_{H^1(\omega;\mathbb{R}^3)} + \|s\|^2_{H^1(\omega;\mathbb{R}^3)}\big)^{1/2}. \tag{6}$$

The boundary conditions considered are hard clamping conditions on part of the boundary. Soft clamping or simple support conditions correspond to $v = 0$ on γ_0. These conditions also work provided that $\varphi(\gamma_0)$ is not included in a straight line; see [7].

Let us now recall the problem formulation and the existence and uniqueness result in the space $\mathcal{V}$ for the linear Naghdi model for shells with little regularity.

THEOREM 2.1. *Let $f \in L^2(\omega;\mathbb{R}^3)$ be a given resultant force density, $N \in L^2(\gamma_1;\mathbb{R}^3)$ an applied traction density, $M \in L^2(\gamma_1,\mathbb{R}^3)$ an applied moment density such that $M\cdot a_3 = 0$ almost everywhere on γ_1, and $e > 0$ the thickness of the shell. Then there exists a unique solution to the following problem: find $(u,r) \in \mathcal{V}$ such that*

$$\forall (v,s) \in \mathcal{V},\ a\big((u,r);(v,s)\big) = L((v,s)), \tag{7}$$

where

$$\begin{aligned} a\big((u,r);(v,s)\big) = \int_\omega \Big\{ & e a^{\alpha\beta\rho\sigma}\left[\gamma_{\alpha\beta}(u)\gamma_{\rho\sigma}(v) + \frac{e^2}{12}\chi_{\alpha\beta}(u,r)\chi_{\rho\sigma}(v,s)\right] \\ & + e\frac{E}{1+\nu}a^{\alpha\beta}\delta_{\alpha 3}(u,r)\delta_{\beta 3}(v,s)\Big\}\sqrt{a}\,dx \end{aligned} \tag{8}$$

and

$$L((v,s)) = \int_\omega f\cdot v\sqrt{a}\,dx + \int_{\gamma_1}(N\cdot v + M\cdot s)\,d\gamma. \tag{9}$$

Proof. See [5] and [7]. □

Here and in the sequel, we make use of the following notational device: arguments in a bilinear form are separated by a semicolon, whereas members of a couple are separated by a comma. This will help keep track of who does what, since our bilinear forms often apply to couples.

3. A penalized version of Naghdi's model. The purpose of the present work is to approximate the solution of (7) with a finite element method and to proceed in the simplest possible way. (Note that we do not concern ourselves with locking in the present paper; in this respect see [2], [12].) As the solution is in H^1, C^0-Lagrange P_1 elements should be sufficient. However, we immediately encounter a problem since the tangency constraint $s\cdot a_3 = 0$ in ω clearly cannot be implemented in a conforming way for a general shell.

We thus introduce a penalized Naghdi problem in which the unknowns still are the displacement u and rotation r, elements of the space $H^1(\omega;\mathbb{R}^3)$ without any orthogonality constraint on r.

Let us introduce the relaxed function space

$$\mathcal{X} = \big\{(v,s) \in H^1(\omega;\mathbb{R}^3)^2; v = s = 0 \text{ on } \gamma_0\big\} \tag{10}$$

and equip it with the standard H^1 norm.

THEOREM 3.1. *Let $p \in \mathbb{R}$ such that $0 < p \leq 1$. Let $f \in L^2(\omega;\mathbb{R}^3)$, $N \in L^2(\gamma_1;\mathbb{R}^3)$, and $M \in L^2(\gamma_1,\mathbb{R}^3)$. Then there exists a unique solution to the following problem: find $(u_p, r_p) \in \mathcal{X}$ such that*

$$\forall (v,s) \in \mathcal{X}, \quad a\big((u_p,r_p);(v,s)\big) + \frac{1}{p} b(r_p \cdot a_3; s \cdot a_3) = L((v,s)), \tag{11}$$

where

$$b(\lambda;\mu) = \int_\omega \partial_\alpha \lambda \partial_\alpha \mu \, dx. \tag{12}$$

The proof is based on the following version of the infinitesimal rigid displacement lemma.

LEMMA 3.2. *Let $(u,r) \in H^1(\omega;\mathbb{R}^3)^2$ and assume that $\varphi \in W^{2,\infty}(\omega;\mathbb{R}^3)$.*

(i) *If $\gamma_{\alpha\beta}(u) = 0$, then there exists $\psi \in L^2(\omega;\mathbb{R}^3)$ such that $\partial_\alpha u = \psi \wedge a_\alpha$.*
(ii) *If $\delta_{\alpha 3}(u,r) = 0$, then $\partial_\alpha u \cdot a_3 = -r \cdot a_\alpha \in H^1(\omega)$.*
(iii) *If, in addition to* (i) *and* (ii), *$\chi_{\alpha\beta}(u,r) = 0$, then ψ is a constant vector in $\mathbb{R}^3$ and there exists $c \in \mathbb{R}^3$ such that*

$$u(x) = c + \psi \wedge \varphi(x), \tag{13}$$

and

$$r(x) = \psi \wedge a_3(x) + (r(x) \cdot a_3(x)) a_3(x). \tag{14}$$

Proof. The argument is exactly the same as in [7], except that we do not assume $r \cdot a_3 = 0$, hence the extra term in formula (14). ∎

Remarks. 1. Note that the infinitesimal rotation vector ψ is given by

$$\begin{aligned} \psi &= \varepsilon^{\alpha\beta}(\partial_\beta u \cdot a_3) a_\alpha + \varepsilon^{\alpha\beta}(\partial_\alpha u \cdot a_\beta) a_3 \\ &= \varepsilon^{\beta\alpha}(r \cdot a_\beta) a_\alpha + \varepsilon^{\alpha\beta}(\partial_\alpha u \cdot a_\beta) a_3, \end{aligned} \tag{15}$$

where $\varepsilon^{11} = \varepsilon^{22} = 0$ and $\varepsilon^{12} = -\varepsilon^{21} = 1/|a_1 \wedge a_2|$.

2. If $(v,s) \in \mathcal{X}$ are such that (13) and (14) are verified, then we have

$$u = 0 \text{ and } r = (r \cdot a_3) a_3 \text{ a.e. in } \omega,$$

due to the boundary conditions.

We now are in a position to prove the ellipticity of the penalized bilinear form.

LEMMA 3.3. *The bilinear form in* (11) *is $\mathcal{X}$-elliptic, uniformly with respect to p for $0 < p \leq 1$.*

Proof. The proof follows along lines similar to those found in [7] and we omit it for brevity. ∎

Proof of Theorem 3.1. Apply the Lax–Milgram lemma. ∎

Remark. It is important to note that the original bilinear form a is not $\mathcal{X}$-elliptic; indeed it does not even define a norm on the relaxed space. It is therefore necessary to add such terms as the extra terms $\|\partial_\alpha(s \cdot a_3)\|_{L^2}^2$ to recover ellipticity over the larger space. In the case of soft clamping, these extra terms are not sufficient, since $(0, a_3)$ still belongs to the kernel of the penalized bilinear form. In this case, one should add the full H^1 norm of $s \cdot a_3$, i.e., use a penalization term of the form $b(r \cdot a_3; s \cdot a_3) = \int_\omega \big[(r \cdot a_3)(s \cdot a_3) + \partial_\alpha(r \cdot a_3)\partial_\alpha(s \cdot a_3)\big] \, dx$.

It is now fairly classical that the penalization provides an approximation of the constrained problem.

THEOREM 3.4. *Let $U = (u, r)$ and $U_p = (u_p, r_p)$, respectively, be the unique solutions of problems* (7) *and* (11). *Then*

$$\|r_p \cdot a_3\|_{H^1(\omega)} \leq Cp \tag{16}$$

and

$$\|U_p - U\|_{\mathcal{X}} \leq Cp. \tag{17}$$

Proof. Let $\mathcal{L} = L^2(\omega; \mathbb{R}^2)$ and $\Psi \colon \mathcal{X} \to \mathcal{L}$ defined by $U \mapsto \nabla(r \cdot a_3)$. Now, we have $\mathcal{V} = \ker \Psi$ and $b(r \cdot a_3; r \cdot a_3) = (\Psi(U), \Psi(U))_{\mathcal{L}}$. It is known that if Ψ has closed range, then the following estimates hold true (see [9]):

$$\|\Psi(U_p)\|_{\mathcal{L}} \leq Cp \text{ and } \|U_p - U\|_{\mathcal{X}} \leq Cp.$$

The first estimate gives estimate (16) and the second estimate is just estimate (17).

Let us thus check that Ψ has closed range. Consider a sequence $U_n \in \mathcal{X}$ such that $\Psi(U_n) \to Z$ in $\mathcal{L}$. By the Poincaré inequality, it follows that $r_n \cdot a_3$ is bounded in $H^1(\omega)$ and we can extract a weakly convergent subsequence such that $r_n \cdot a_3 \rightharpoonup \zeta$ in $H^1(\omega)$. Moreover, since $r_n \cdot a_3 = 0$ on γ_0 in the sense of traces, it follows that $\zeta = 0$ on γ_0 as well. In addition, clearly $Z = \nabla\zeta$. We thus set $U = (0, \zeta a_3) \in \mathcal{X}$ and we see that $\Psi(U) = Z$. ∎

Remark. Since we are aiming for simplicity of implementation, we have made no attempt to make the penalization term intrinsic. In fact, it does depend on the chart, whereas the other terms do not. This could arguably be considered to be a poor choice, especially if a chart was used that gave much more weight to one part of the shell compared to the rest. An intrinsic choice that obviously works is

$$b'(r \cdot a_3; s \cdot a_3) = \int_\omega a^{\alpha\beta} \partial_\alpha (r \cdot a_3) \partial_\beta (s \cdot a_3) \sqrt{a}\, dx.$$

This penalization term has the same properties as our simple penalization term and does not suffer from the above mentioned drawback.

4. A mixed formulation of Naghdi's model. Another way of imposing a constraint in a variational problem is to use a mixed formulation. We follow this route in this section. Naturally, mixed formulations for Naghdi's model already exist of the displacement/stress type, but in the context of attempting to write nonlocking formulations; see, for instance, [2]. In the present article, we are not concerned with locking issues but only with imposing the tangency of the rotation vector in Cartesian coordinates. Hence the mixed formulation will be relatively simple. In particular, it involves the same bilinear forms as those used in the penalization approach. Let us set $\mathcal{M} = H^1_{\gamma_0}(\omega)$.

THEOREM 4.1. *For all $\rho \geq 0$, the variational problem of finding $(U, \lambda) \in \mathcal{X} \times \mathcal{M}$ such that*

$$\forall (V, \mu) \in \mathcal{X} \times \mathcal{M}, \begin{cases} a(U; V) + \rho b((r \cdot a_3); (s \cdot a_3)) + b((s \cdot a_3); \lambda) = L(V), \\ b((r \cdot a_3); \mu) = 0, \end{cases} \tag{18}$$

has a unique solution, which is such that $U \in \mathcal{V}$ is the solution of Naghdi's problem (7).

Proof. The bilinear form $a + \rho b$ is $\mathcal{V}$-elliptic (and even $\mathcal{X}$-elliptic for $\rho > 0$ by Lemma 3.3). In order to prove that problem (18) has a unique solution, we therefore

just need to prove that b satisfies the inf-sup condition; see [14], [9]. Let thus

$$\beta = \inf_{\mu\in\mathcal{M}} \sup_{V\in\mathcal{X}} \frac{b((s\cdot a_3);\mu)}{\|V\|_{\mathcal{X}}\|\mu\|_{\mathcal{M}}},$$

and we want to show that $\beta > 0$. Let $\mu \in \mathcal{M}\setminus\{0\}$ be arbitrary. Since μ vanishes on γ_0 and since $a_3 \in W^{1,\infty}(\omega;\mathbb{R}^3)$, we clearly have $V = (0,\mu a_3) \in \mathcal{X}$ and $\mu a_3 \cdot a_3 = \mu$. Therefore,

$$\sup_{V\in\mathcal{X}} \frac{b((s\cdot a_3);\mu)}{\|V\|_{\mathcal{X}}} \geq \frac{\|\nabla\mu\|^2_{L^2(\omega;\mathbb{R}^2)}}{\|a_3\otimes\nabla\mu + \mu\nabla a_3\|_{L^2(\omega;\mathbb{M}_{32})}}$$

so that

$$\beta \geq \inf_{\mu\in\mathcal{M}} \frac{\|\nabla\mu\|_{L^2(\omega;\mathbb{R}^2)}}{\|a_3\otimes\nabla\mu + \mu\nabla a_3\|_{L^2(\omega;\mathbb{M}_{32})}}.$$

It is quite clear that the left-hand side of the above inequality is strictly positive, since the denominator is basically a lower order perturbation of the numerator. Let us quickly show this by a contradiction argument. Assume thus that we are given a sequence $\mu_n \in \mathcal{M}$ such that

$$\|\nabla\mu_n\|_{L^2(\omega;\mathbb{R}^2)} \to 0 \text{ but } \|a_3\otimes\nabla\mu_n + \mu_n\nabla a_3\|_{L^2(\omega;\mathbb{M}_{32})} = 1.$$

Obviously, due to the boundary conditions and the Poincaré inequality, $\mu_n \to 0$ in $H^1(\omega)$, hence $a_3\otimes\nabla\mu_n \to 0$ in L^2 and $\mu_n\nabla a_3 \to 0$ in L^2 (recall that $a_3 \in W^{1,\infty}$), a contradiction. Hence the inf-sup condition holds true and the mixed formulation has one and only one solution.

Let us now check that this solution corresponds to the usual Naghdi problem. Taking $\mu = r\cdot a_3$ in the second equation, we see that $U \in \mathcal{V}$. Then, taking $V \in \mathcal{V}$ cancels all terms involving b in the first equation, hence the result. ∎

Remarks. 1. We can also replace b by any scalar multiple of itself, and in the case of soft clamping, we must replace it by the full H^1 scalar product between $s\cdot a_3$ and μ. The lack of intrinsic character can be cured in the same way as for the penalization.

2. The Lagrange multiplier λ that enforces the tangency constraint $r\cdot a_3 = 0$ does not have a specific mechanical meaning, since the bilinear forms are pretty arbitrary. Note that when nonzero, $r\cdot a_3$ is sometimes called the pinching component or pinching strain; see [11]. Indeed, it corresponds to a change in length of the deformed normal fiber in the three-dimensional Kirchhoff–Love displacement constructed from u and r. It is thus conceivable that a mechanical meaning could be ascribed to such a Lagrange multiplier, but we do not pursue this line of reasoning here.

3. We may choose $\rho = 0$ or $\rho > 0$. In the latter case, we are adding a penalization term in the spirit of augmented Lagrangian methods, which can be tuned for the best numerical results.

5. The discrete formulations.

5.1. Finite element discretization of the penalized problem. The penalized problem is a standard variational problem formulated in H^1. We thus propose to use a standard conforming finite element approximation.

Let thus T_h be a regular affine family of triangulations which covers the domain ω. The discrete space of admissible displacements and rotations is given by

$$\mathcal{X}_h = \{(v,s) \in C^0(\omega;\mathbb{R}^3)^2, (v,s)_{|K} \in P_1(K), v = s = 0 \text{ on } \gamma_0\}, \tag{19}$$

which is obviously contained in the continuous space $\mathcal{X}$.

The discrete problem thus reads as follows: find $(u_{p,h}, r_{p,h}) \in \mathcal{X}_h$ such that

$$\forall (v,s) \in \mathcal{X}_h,\ a\big((u_{p,h}, r_{p,h}); (v,s)\big) + \frac{1}{p} b(r_{p,h} \cdot a_3; s \cdot a_3) = L(v,s). \tag{20}$$

Naturally, this problem has a unique solution.

5.2. Convergence. By virtue of the classical properties of Galerkin approximation, we have the following convergence result.

THEOREM 5.1. *There exists a sequence $h_p \to 0$ such that*

$$\|(u,r) - (u_{p,h_p}, r_{p,h_p})\|_{\mathcal{X}} \longrightarrow 0 \quad \textit{when} \quad p \to 0. \tag{21}$$

Proof. For each p, we have $u_{p,h} \to u_p$ when $h \to 0$ because this is a Galerkin approximation of a classical variational problem. We then appeal to Theorem 3.4 to construct a converging diagonal sequence. ∎

If the solution is assumed to have some regularity, the second step of the approximation may of course be controlled via an error estimate.

PROPOSITION 5.2. *Assume that the solution $(u_p; r_p)$ of problem* (11) *belongs to $H^2(\omega, \mathbb{R}^3)^2$ for all p. Then there exists a constant C_p independent of h, such that*

$$\|(u_{p,h}, r_{p,h}) - (u_p, r_p)\|_{\mathcal{X}} \le C_p h \|(u_p, r_p)\|_{H^2}. \tag{22}$$

Proof. See [13], for example. ∎

Remarks. 1. Since we are mostly interested in shells with little regularity—otherwise classical formulations would apply—it is presumably not useful to look for higher order elements in the hope of improving the rate of convergence. Indeed, even without taking into account the penalization term, in the case of such a shell, the underlying system of PDEs has nonsmooth coefficients. It is therefore unclear whether elliptic regularity can be applied to yield even an H^2 regularity, let alone H^{k+1} regularity with $k \ge 1$. Note, however, that if the midsurface chart is smooth and we want to use our formulation nonetheless for simplicity as compared to the classical approach, then elliptic regularity will apply.

2. We could also combine estimates (17) and (22) to obtain a global error estimate for the whole penalization/discretization process. To achieve this goal, we would need to estimate the constant C_p in terms of p, which would probably include terms of the order of p^{-1} due to the continuity constant of the bilinear form a_p, and the term $\|(u_p, r_p)\|_{H^2}$. The latter term could be evaluated by using Nirenberg's translations method, but the technical aspects involved hardly seem worth the effort in this particular case, in view of the previous remark. In any case, it is reasonable to expect locking due to the penalization term.

5.3. Finite element discretization of the mixed problem. The mixed problem is also a standard variational problem formulated in H^1. In order to prove the convergence of conforming finite element approximations, we only need to establish the uniform discrete inf-sup condition. As is often the case, the uniform discrete inf-sup condition turns out to be harder to prove than its continuous counterpart and in our particular case, some of the arguments are rather nonstandard. Let us treat the P_1 case, with zero boundary condition for the multiplier, for simplicity. In this case, we have

$$\mathcal{M}_h = \{\mu_h \in C^0(\bar{\omega}), \mu_{h|K} \in P_1(K), \mu_h = 0 \text{ on } \partial\omega\}.$$

THEOREM 5.3. *For all $\rho \geq 0$, the variational problem of finding $(U_h, \lambda_h) \in \mathcal{X}_h \times \mathcal{M}_h$ such that*

(23)
$$\forall (V_h, \mu_h) \in \mathcal{X}_h \times \mathcal{M}_h, \ \begin{cases} a(U_h; V_h) + \rho b((r_h \cdot a_3); (s_h \cdot a_3)) + b((s_h \cdot a_3); \lambda_h) = L(V_h), \\ b((r_h \cdot a_3); \mu_h) = 0, \end{cases}$$

has a unique solution for h small enough. Moreover

$$\|U - U_h\|_{\mathcal{X}} + \|\lambda - \lambda_h\|_{\mathcal{M}} \to 0 \text{ when } h \to 0.$$

We first need a couple of geometrical results.

LEMMA 5.4. *Let φ be a $W^{2,\infty}$ chart. There exists a constant $C > 0$ such that for all x, y in ω,*

$$|a_3(x) \cdot (a_3(x) - a_3(y))| \leq C\|x - y\|^2. \tag{24}$$

Proof. We adapt an argument of [1, Lemma 3.5]. By our regularity hypothesis, the normal vector a_3 is Lipschitz on $\bar{\omega}$. Hence, for all $x_0 \in \bar{\omega}$, the function

$$Z(x) = (a_3(x) - a_3(x_0)) \cdot a_3(x_0)$$

is also Lipschitz. Therefore, by Rademacher's theorem it is almost everywhere differentiable and we have

$$\nabla Z(x) = \nabla a_3(x)^T a_3(x_0)$$

for almost all $x \in \omega$. Therefore, due to the identification between Lipschitz and $W^{1,\infty}$ functions in a Lipschitz domain (see [1] for a proof), there exists a constant C_ω depending only on ω such that

$$|Z(x)| = |Z(x) - Z(x_0)| \leq C_\omega \|\nabla a_3^T a_3(x_0)\|_{L^\infty(\bar{B}(x_0, \|x - x_0\|) \cap \omega; \mathbb{R}^2)} \|x - x_0\|.$$

Now, a_3 is a unit vector. Hence, at any point y of differentiability of a_3, $a_3(y)$ is orthogonal to the image of $\nabla a_3(y)$, that is to say, $\nabla a_3(y)^T a_3(y) = 0$. Consequently, we have that almost everywhere in $\bar{B}(x_0, \|x - x_0\|) \cap \omega$,

$$\nabla a_3(y)^T a_3(x_0) = \nabla a_3(y)^T a_3(x_0) - \nabla a_3(y)^T a_3(y)$$

so that

$$\begin{aligned} \|\nabla a_3(y)^T a_3(x_0)\| &\leq \|\nabla a_3(y)^T\| \|a_3(x_0) - a_3(y)\| \\ &\leq C_\omega \|\nabla a_3\|^2_{L^\infty(\omega; \mathbb{M}_{32})} \|y - x_0\| \end{aligned}$$

almost everywhere. Therefore,

$$\|\nabla a_3^T a_3(x_0)\|_{L^\infty(\bar{B}(x_0, \|x - x_0\|) \cap \omega; \mathbb{R}^2)} \leq C_\omega \|\nabla a_3\|^2_{L^\infty(\omega; \mathbb{M}_{32})} \|x - x_0\|,$$

hence the result with $C = C_\omega^2 \|\nabla a_3\|^3_{L^\infty(\omega; \mathbb{M}_{32})}$. ∎

Remark. Note that the above geometrical result holds true under the weaker, "minimal" regularity hypotheses advocated in [1] for a shell midsurface, namely, φ bilipschitz and such that a_3 is Lipschitz.

LEMMA 5.5. *Under the same hypotheses, there exists a constant $C > 0$ such that for all x and almost all y in ω,*

$$|a_3(x) \cdot \partial_\alpha a_3(y)| \leq C\|x - y\|. \tag{25}$$

Proof. Let y be a point of differentiability of a_3. We have

$$a_3(x) \cdot \partial_\alpha a_3(y) = (a_3(x) - a_3(y)) \cdot \partial_\alpha a_3(y)$$

so that

$$|a_3(x) \cdot \partial_\alpha a_3(y)| \leq C_\omega \|\nabla a_3\|^2_{L^\infty(\omega;\mathbb{M}_{32})}\|x - y\|$$

for all $x \in \omega$. ∎

We now turn to the inf-sup condition per se. Let Π_h denote either the vector-valued Lagrange interpolation operator from $C^0_0(\bar{\omega};\mathbb{R}^3)$ into $\mathcal{X}_h$ or the scalar-valued Lagrange interpolation operator from $C^0_0(\bar{\omega})$ into $\mathcal{M}_h$, depending on the context, and ψ^h_j the shape function associated with vertex S_j of the triangulation.

LEMMA 5.6. *For all $\mu_h \in \mathcal{M}_h$, we let $R_h(\mu_h) = \Pi_h(\mu_h a_3)$. There exists a constant $C > 0$ independent of h such that*

$$b(R_h(\mu_h) \cdot a_3; \mu_h) \geq C\|\mu_h\|^2_{\mathcal{M}}. \tag{26}$$

Proof. Note that while μ_h is scalar piecewise P_1, $\mu_h a_3$ is vector-valued and $R_h(\mu_h)$ is vector-valued piecewise P_1. Let us set

$$\delta_h = R_h(\mu_h) \cdot a_3 - \mu_h$$

so that

$$b(R_h(\mu_h) \cdot a_3; \mu_h) = \|\mu_h\|^2_{\mathcal{M}} + b(\delta_h; \mu_h)$$

with

$$|b(\delta_h; \mu_h)| \leq \|\mu_h\|_{\mathcal{M}}\|\delta_h\|_{\mathcal{M}}.$$

We thus just need to estimate δ_h in the norm of $\mathcal{M}$. By Lagrange interpolation, we have

$$\mu_h(x) = \sum_{S_j} \mu_h(S_j)\psi^h_j(x)$$

and

$$R_h(\mu_h)(x) = \sum_{S_j} \mu_h(S_j)\psi^h_j(x)a_3(S_j).$$

Therefore

$$R_h(\mu_h) \cdot a_3(x) = \sum_{S_j} \mu_h(S_j)[a_3(S_j) \cdot a_3(x)]\psi^h_j(x),$$

and

$$\delta_h(x) = \sum_{S_j} \mu_h(S_j)[(a_3(S_j) - a_3(x)) \cdot a_3(x)]\psi^h_j(x),$$

since $a_3(x)$ is a unit vector. Consequently, we arrive at the formula

$$\partial_\alpha \delta_h(x) = \sum_{S_j} \mu_h(S_j)[a_3(S_j)\cdot \partial_\alpha a_3(x)]\psi_j^h(x) + \sum_{S_j} \mu_h(S_j)[(a_3(S_j)-a_3(x))\cdot a_3(x)]\partial_\alpha \psi_j^h(x)$$

almost everywhere (namely inside the triangles). At every point of differentiability in ω, at most three terms in the sums are nonzero; therefore we can estimate

$$\|\partial_\alpha \delta_h\|_{L^\infty(\omega)} \le 3\|\mu_h\|_{L^\infty(\omega)} \max_j \max_{K_{k,j}} \left[|a_3(S_j)\cdot \partial_\alpha a_3(x)| + \frac{C}{h}|(a_3(S_j)-a_3(x))\cdot a_3(x)|\right],$$

where $K_{k,j}$ stand for all the triangles having S_j as vertex. Since all triangles have diameter bounded by a constant times h, we deduce with the help of Lemmas 5.4 and 5.5 that

$$\|\partial_\alpha \delta_h\|_{L^\infty(\omega)} \le Ch\|\mu_h\|_{L^\infty(\omega)},$$

where C does not depend on h nor on μ_h.

We now appeal to the classical discrete Sobolev estimate (see [8]) and deduce that

$$\|\nabla \delta_h\|_{L^2(\omega;\mathbb{R}^2)} \le C\|\nabla \delta_h\|_{L^\infty(\omega;\mathbb{R}^2)} \le Ch\|\mu_h\|_{L^\infty(\omega)} \le Ch(\ln h)^{1/2}\|\nabla \mu_h\|_{L^2(\omega;\mathbb{R}^2)}.$$

Taking h small enough so that $Ch(\ln h)^{1/2} \le \frac{1}{2}$, we obtain estimate (26). ∎

We now are in a position to prove the crucial uniform discrete inf-sup condition which guarantees the convergence of the finite element scheme applied to the mixed formulation.

THEOREM 5.7. *There exists $\beta^* > 0$ independent of h such that*

$$\inf_{\mu_h \in \mathcal{M}_h} \sup_{V_h \in \mathcal{X}_h} \frac{b((s_h \cdot a_3);\mu_h)}{\|V_h\|_{\mathcal{X}}\|\mu_h\|_{\mathcal{M}}} \ge \beta^*. \tag{27}$$

Proof. Let thus

$$\beta_h = \inf_{\mu_h \in \mathcal{M}_h} \sup_{V_h \in \mathcal{X}_h} \frac{b((s_h \cdot a_3);\mu_h)}{\|V_h\|_{\mathcal{X}}\|\mu_h\|_{\mathcal{M}}}.$$

By construction, since μ_h vanishes on $\partial\omega$, we see that $V_h = (0, R_h(\mu_h)) \in \mathcal{X}_h$ and that $\|V_h\|_{\mathcal{X}} = \|\nabla R_h(\mu_h)\|_{L^2(\omega;\mathbb{M}_{32})}$. Therefore, by Lemma 5.6,

$$\beta_h \ge C \inf_{\mu_h \in \mathcal{M}_h} \frac{\|\mu_h\|_{\mathcal{M}}}{\|\nabla R_h(\mu_h)\|_{L^2(\omega;\mathbb{M}_{32})}},$$

and it suffices to estimate the denominator from above independently of h. We can write

$$R_h(\mu_h) = R_h(\mu_h) - \mu_h a_3 + \mu_h a_3,$$

and since $\mu_h a_3 \in H^1(\omega;\mathbb{R}^3) \cap C^0(\bar{\omega};\mathbb{R}^3)$ and $R_h(\mu_h)$ is the P_1-interpolate of $\mu_h a_3$ on the triangulation, which is regular by assumption, a classical scaling argument shows that

$$\|\nabla R_h(\mu_h) - \mu_h a_3\|_{L^2(\omega;\mathbb{M}_{32})} \le C\|\nabla(\mu_h a_3)\|_{L^2(\omega;\mathbb{M}_{32})}.$$

Therefore

$$\|\nabla R_h(\mu_h)\|_{L^2(\omega;\mathbb{M}_{32})} \leq (C+1)\|\nabla(\mu_h a_3)\|_{L^2(\omega;\mathbb{M}_{32})}.$$

But we have already seen in the continuous case that

$$\|\nabla(\mu_h a_3)\|_{L^2(\omega;\mathbb{M}_{32})} \leq C\|\mu_h\|_{\mathcal{M}}.$$

We have thus obtained that

$$\|\nabla R_h(\mu_h)\|_{L^2(\omega;\mathbb{M}_{32})} \leq C\|\mu_h\|_{\mathcal{M}},$$

which completes the proof of the Theorem. □

Remark. It is fairly clear that the proof works the same if we replace P_1 interpolation by another Lagrange interpolation, for example, P_2, which is also available in FreeFem++.

The proof of Theorem 5.3 follows as in [9] or [14].

Naturally, if we assume some regularity of the solution, we obtain error estimates.

PROPOSITION 5.8. *Assume that the solution* $((u,r),\lambda)$ *of problem* (4.1) *belongs to* $H^2(\omega,\mathbb{R}^3)^3$. *Then there exists a constant* C *independent of* h *such that*

$$\|(u_h,r_h)-(u,r)\|_{\mathcal{X}} + \|\lambda_h-\lambda\|_{\mathcal{M}} \leq Ch\|((u,r),\lambda)\|_{H^2}. \tag{28}$$

Proof. See [14], for example. □

6. Numerical tests. In this section, we implement the discretization of both penalized and mixed approaches, compare them on a literature benchmark, and apply them to genuinely $W^{2,\infty}$ shells.

6.1. Implementation details. Both model formulations only require the knowledge of a_α, a_3, and $\partial_\alpha a_3$. All other quantities, either geometrical like the elasticity tensor or kinematical like the strain tensors, can be expressed by means of dot products involving these quantities. It is convenient to define these vectors as FreeFem++ functions. The dot products are expressed as FreeFem++ macros, which are then combined into other macros that eventually expand to all the other quantities of interest. The net result is that our code automatically constructs the bilinear forms, with minimal user input, typically between 10 and 20 lines of code. This works well if an analytic description of the midsurface is available. In the case of midsurfaces implicitly defined via interpolation of nodal values, as in [11], the same approach should be possible, provided the interpolated surface chart retains $W^{2,\infty}$ regularity. We plan to address this issue in a further iteration of the code.

Let us note that with respect to user input and code complexity, our approach compares favorably with classical formulations which require the computation of the covariant and mixed components of the second fundamental form and of the Christoffel symbols of the chart; see, for example, [3].

Three-dimensional visualization of the undeformed and deformed shells uses Medit,[1] a free mesh visualization software available at http://www.ann.jussieu.fr/~frey/logiciels/medit.html.

All the tests were run on 1.5GHz Apple PowerBook G4 laptops and 2GHz single-processor Apple Xserve G5.

[1] This software was designed and developed at the Laboratoire Jacques-Louis Lions of the University Pierre et Marie Curie.

6.2. The hyperbolic paraboloid shell. This test is a literature benchmark for shell elements. We use this example, in which the midsurface of the shell is represented by a chart of class C^∞, mainly to validate our code. It does not constitute a relevant test for the $W^{2,\infty}$ case.

The reference domain of the midsurface is given by

$$\omega = \{|x| + |y| < \sqrt{2}b\},$$

and the chart is defined by

$$\varphi(x, y) = \left(x, y, \frac{c}{2b^2}(x^2 - y^2)\right)^T,$$

where $b = 50$ cm and $c = 10$ cm.

The shell is clamped on $\partial\omega$ and subjected to a uniform pressure $q = 0.01$ kp/cm^2. The mechanical data are

$$E = 2.85\ 10^4 \text{ kp/cm}^2, \nu = 0.4,$$

The thickness of the shell is $e = 0.8$ cm.

The reference value for this test is the normal displacement at the center A of the shell. Its value computed by various methods is of -0.024 cm; see [3].

Due to the symmetries of the problem, we use the computational domain

$$\omega' = \{0 < x, 0 < y, x + y < \sqrt{2}b\}$$

and enforce the symmetry conditions

$$u_2 = 0, r_2 = 0 \text{ on } y = 0$$

and

$$u_1 = 0, r_1 = 0 \text{ on } x = 0.$$

These conditions are obtained by expressing the continuity of the three-dimensional Kirchhoff–Love displacement $U = u + x_3 r$ along these edges.

The following are results for both methods using mesh adaption and P_2 elements:

	Degrees of freedom	$(u \cdot a_3)(A)$	Range of values for $r \cdot a_3$
Penalized	7005	-0.0241419	$-3.46456\text{e}{-08}$ $3.52366\text{e}{-08}$
Mixed	7279	-0.0241416	$-3.52081\text{e}{-08}$ $3.51064\text{e}{-08}$

In the penalized test, the penalization parameter was $10^3 \frac{E}{2(1+\nu)}$. Both methods achieve excellent tangency for the rotation vector and similar performance in terms of the reference value, see also Figures 1 and 2.

Remark. We also performed tests on another benchmark, the Scordelis–Lo roof. Unfortunately, in this case, P_1 and P_2 elements present significant locking. (We obtain a maximum normal displacement that is only 65% of the reference value.) Until more sophisticated elements become integrated into FreeFem++, such tests cannot yield satisfactory results. We were, however, able to confirm that both Cartesian formulations provide the same (locked) result as the classical covariant formulation using P_1 and P_2 elements.

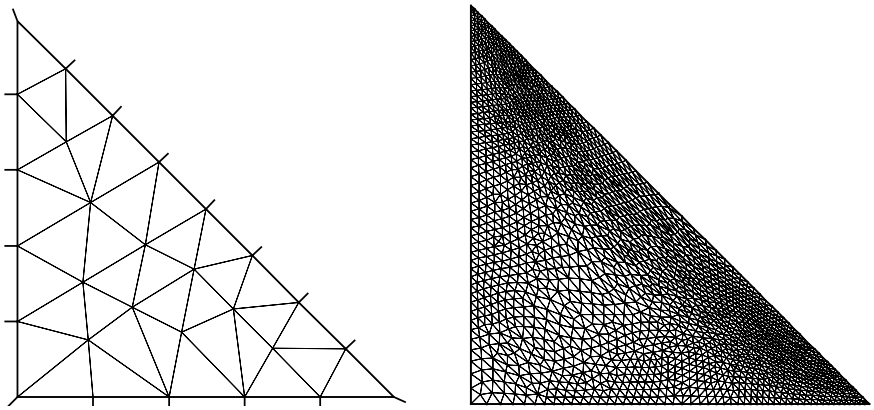


FIG. 1. *The initial and final meshes.*

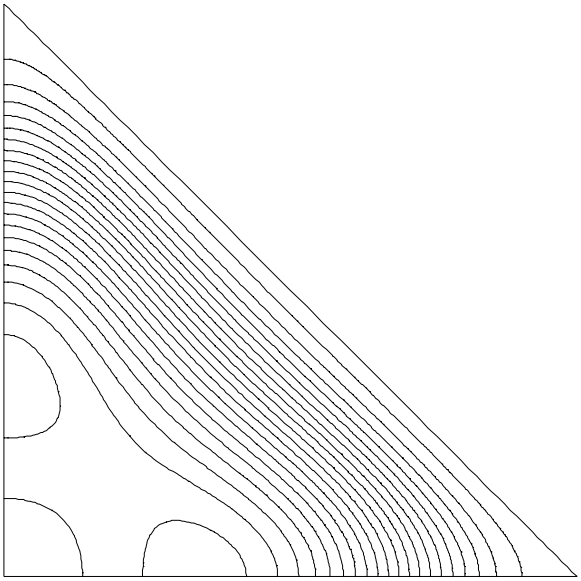


FIG. 2. *Isovalues of the normal displacement* $u \cdot a_3$*, mixed formulation.*

6.3. A plane-cylinder $W^{2,\infty}$-shell. Our next test is a genuine $W^{2,\infty}$ test with curvature discontinuities. The shell consists of a plane part and a cylindrical part with a C^1-join; see Figure 3. The reference domain of the midsurface is given by

$$\omega =]-R, R[\times]-L/2, L/2[,$$

and the chart is defined by

$$\varphi(x,y) = \begin{cases} (x, y, 0)^T & \text{if } x < 0, \\ (R\sin(x/R), y, R(1-\cos(x/R)))^T & \text{if } x \geq 0 \end{cases}$$

with $R = 300$ in. and $L = 600$ in. (These values come from the Scordelis–Lo test.) The thickness of the shell is $e = 3$ in.

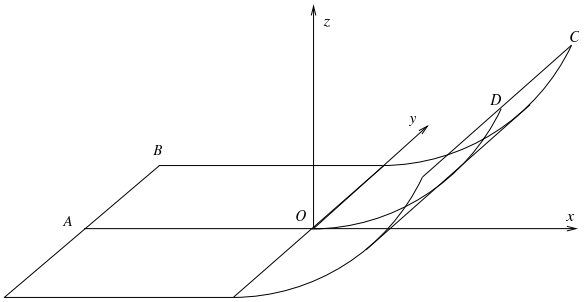


FIG. 3. *The plate-cylinder shell.*

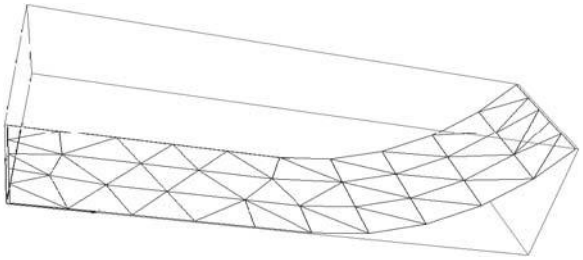


FIG. 4. *The initial mesh on the midsurface.*

The mechanical data are

$$E = 3.0 \times 10^6 \text{ psi}, \nu = 0.0.$$

The shell is submitted to a uniform downward pressure of 0.625 lb/in^2.

Concerning boundary conditions, we consider the case of hard clamping on lines AB and DC

$$u_1 = u_2 = u_3 = 0 \quad \text{and} \quad r_1 = r_2 = r_3 = 0,$$

and the shell is free on its remaining edges. Thanks to the symmetry, we only consider half of the midsurface, $y > 0$. The corresponding symmetry conditions on AD are

$$u_2 = r_2 = 0.$$

Note that the initial mesh ignores the curvature discontinuity at $x = 0$.

Note that in Medit, the coordinate axes are attached to the bounding box: although it seems that the clamped left side of the shell has moved up in Figure 5 compared to Figure 4, this is not actually the case.

It is interesting to note that mesh adaption concentrates around the curvature discontinuity, thus indicating the lack of regularity of the solution across this line. In the isovalues of Figures 6, 7, and 8, the leftmost half of the domain corresponds to the planar part of the shell for $x_1 < 0$ and the other half to the cylindrical part of the shell. The line AD is represented by the bottom side of the domain.

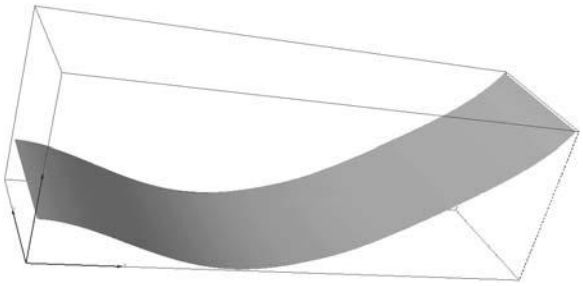


FIG. 5. *The deformed half-shell (displacement magnified by a factor of* 7*).*

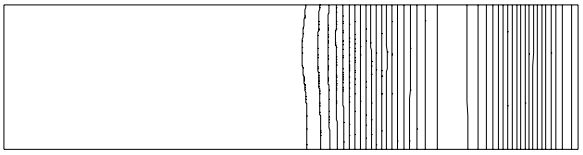


FIG. 6. *Isovalues of* u_1*. The range of values is* $[-0.967042, 0.0130627]$.

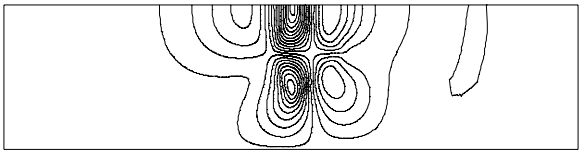


FIG. 7. *Isovalues of* u_2*. The range of values is* $[-0.00181261, 0.00234357]$.

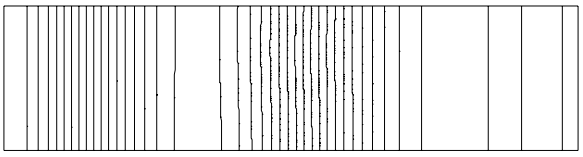


FIG. 8. *Isovalues of* u_3*. The range of values is* $[-6.31062, 0.949939]$*. (Isovalues for* $u \cdot a_3$ *are practically identical and those for* r_1 *and* λ *show similar features.)*

Note that although the pressure acts downward, the cylindrical part of the shell lifts a little bit to compensate for the large deflection of its planar part.

Concerning the rotation vector, we have the following isovalues in Figures 9 and 10. (r_1 is not represented; see Figure 8.)

To see how the mixed formulation manages to enforce the tangency constraint, we also plot the isovalues of the normal rotation $r \cdot a_3$ (Figure 11). We see that the curvature discontinuity makes it harder to capture this constraint than in the C^∞ case of the hyperbolic paraboloid.

Finally, we compare our results with those of [15] for the same geometry, but for the Koiter model, using the Argyris element on a structured mesh that respects the curvature discontinuity. The vertical displacement of point O is found to be

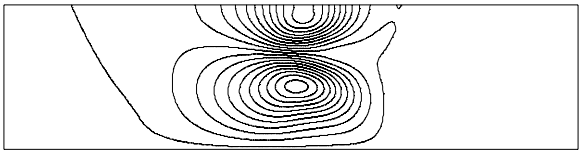


FIG. 9. *Isovalues of* r_2. *The range of values is* $[-0.0011471, 0.000972632]$.

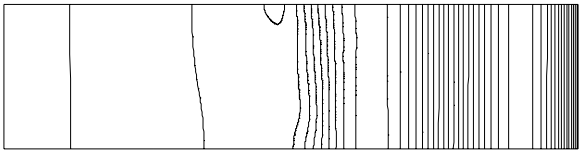


FIG. 10. *Isovalues of* r_3. *The range of values is* $[-0.00547512, 0.0128569]$.

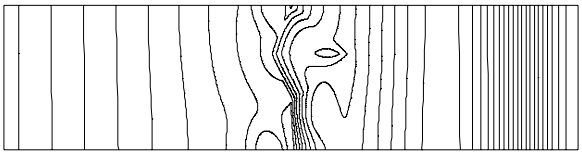


FIG. 11. *Isovalues of* $r \cdot a_3$. *The range of values is* $[-9.18161\text{e}{-06}, 0.00474402]$.

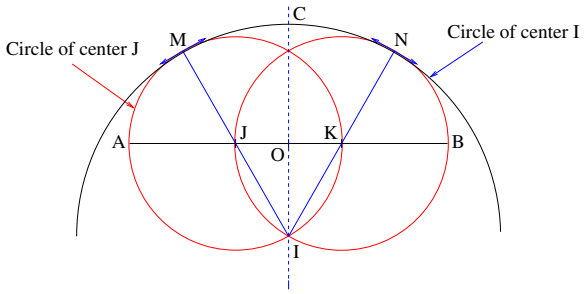


FIG. 12. *AMNB is basket handle or three-centered arch.*

approximately −4.0 in. (value based on a graph in [15]). We find $u_3(0) = -3.83631$ in., which is in good agreement.

6.4. A basket-handle tunnel. A basket-handle is a classical approximation of an arc of ellipse, and a very good one, constructed with three circles (Figure 12). It has long been used in architecture as a replacement for an ellipse. Clearly this arc presents two curvature discontinuities and the same will be true for arches based on it.

We present numerical results for a long, tunnel-like shell based on a slightly extended basket-handle arc.

We use the same mechanical data as for the plate-cylinder shell. Clamping is assumed on both rectilinear sides of the shell. These sides are of length 3000 in. The large circle radius is 400 in. and the small circle radius 200 in.

The natural chart for this shell is of class $W^{2,\infty}$. It is obtained by parametrizing

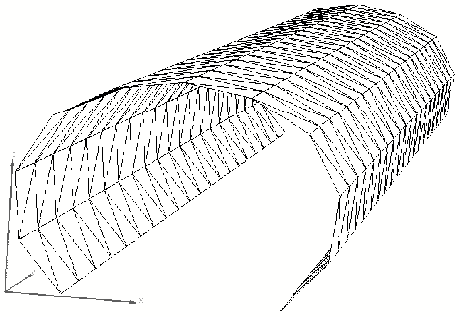


FIG. 13. *The initial mesh on the basket-handle midsurface.*

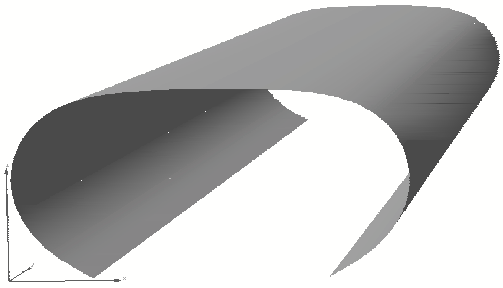


FIG. 14. *The deformed shell (displacement magnified by a factor of* 3*).*

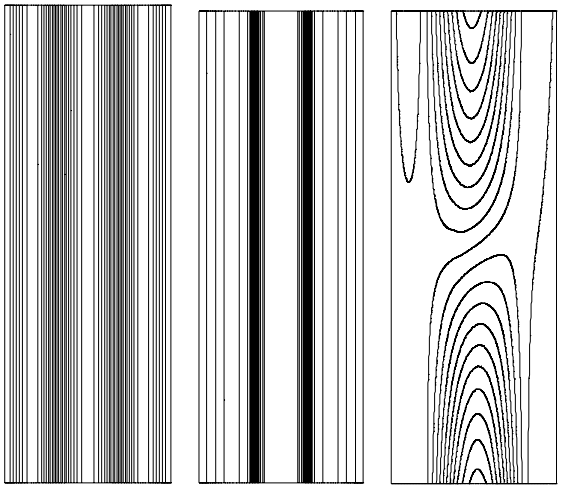


FIG. 15. *Isovalues of* $u \cdot a_3$ *(left),* $r \cdot a_3$ *(middle), and* u_2 *(right).*

the basket-handle by arclength. The computational domain is a rectangle $]-628.32, 628.32[\times]-1800, 1800[$. (We compute the whole shell without using the symmetries for better visualization.)

The vertical displacement of the center of the shell is $u_3(0,0) = -27.3815$ in. Figure 13 shows the initial mesh, Figure 14 the deformed shell, and Figure 15 various isovalues.

Remark. Naturally, the isovalues for u_2 should respect the shell symmetries. However, since the range of values for u_2 is of the order of $[-2e{-}5, 2e{-}5]$, the shape of the isovalue lines is very sensitive to errors. It nonetheless becomes more symmetrical when the mesh is further refined.

Acknowledgments. The authors wish to thank Christine Bernardi and Vivette Girault for illuminating discussions about finite element theory, Marina Vidrascu for insights on the numerics of shells, and the referees for suggestions resulting in significant simplification of some proofs.

REFERENCES

[1] S. Anicic, H. Le Dret, and A. Raoult, *The infinitesimal rigid displacement lemma in Lipschitz coordinates and application to shells with minimal regularity*, Math. Methods Appl. Sci., 27 (2004), pp. 1283–1299.

[2] D. N. Arnold and F. Brezzi, *Locking-free finite element methods for shells*, Math. Comp., 66 (1997), pp. 1–14.

[3] M. Bernadou, *Méthodes d'éléments finis pour les problèmes de coques minces*, RMA 33, Masson, 1994.

[4] M. Bernadou, P. G. Ciarlet, and B. Miara, *Existence theorems for two-dimensional linear shell theories*, J. Elasticity, 34 (1994), pp. 111–138.

[5] A. Blouza, *Existence et unicité pour le modèle de Naghdi pour une coque peu régulière*, C.R. Acad. Sci. Paris Ser. I Math., 324 (1997), pp. 839–844.

[6] A. Blouza and H. Le Dret, *Existence and uniqueness for the linear Koiter model for shells with little regularity*, Quart. Appl. Math., 57 (1999), pp. 317–337.

[7] A. Blouza and H. Le Dret, *Naghdi's shell model: Existence, uniqueness and continuous dependence on the midsurface*, J. Elasticity, 64 (2001), pp. 199–216.

[8] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring*, I, Math. Comp., 47 (1986), pp. 103–134.

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[10] M. Carrive, P. Le Tallec, and J. Mouro, *Approximation par éléments finis d'un modèle de coques géométriquement exact*, Rev. Europ. Elements Finis, 4 (1995), pp. 633–662.

[11] D. Chapelle, A. Ferent, and K. J. Bathe, *3D-shell elements and their underlying mathematical model*, Math. Models Methods Appl. Sci., 14 (2004), pp. 105–142.

[12] D. Chapelle and R. Stenberg, *Stabilized finite element formulations for shells in a bending dominated state*, SIAM J. Numer. Anal., 36 (1999), pp. 32–73.

[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl., Vol. 4, North–Holland, Amsterdam, 1978.

[14] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, 1979.

[15] N. Kerdid and P. Mato Eiroa, *Conforming finite element approximation for shells with little regularity*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 95–107.

[16] P. Le Tallec and S. Mani, *Analyse numérique d'un modèle de coque de Koiter discrétisé en base cartésienne par éléments finis DKT*, Model. Math. Anal. Numer., 32 (1998), pp. 433–450.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 655–676

GENERALIZED CUBIC SPLINE FRACTAL INTERPOLATION FUNCTIONS*

A. K. B. CHAND† AND G. P. KAPOOR†

Abstract. We construct a generalized C^r-Fractal Interpolation Function (C^r-FIF) f by prescribing any combination of r values of the derivatives $f^{(k)}$, $k = 1, 2, \ldots, r$, at boundary points of the interval $I = [x_0, x_N]$. Our approach to construction settles several questions of Barnsley and Harrington [*J. Approx Theory*, 57 (1989), pp. 14–34] when construction is not restricted to prescribing the values of $f^{(k)}$ at only the initial endpoint of the interval I. In general, even in the case when r equations involving $f^{(k)}(x_0)$ and $f^{(k)}(x_N)$, $k = 1, 2, \ldots, r$, are prescribed, our method of construction of the C^r-FIF works equally well. In view of wide ranging applications of the classical cubic splines in several mathematical and engineering problems, the explicit construction of cubic spline FIF $f_\Delta(x)$ through *moments* is developed. It is shown that the sequence $\{f_{\Delta_k}(x)\}$ converges to the defining data function $\Phi(x)$ on two classes of sequences of meshes at least as rapidly as the square of the mesh norm $\|\Delta_k\|$ approaches to zero, provided that $\Phi^{(r)}(x)$ is continuous on I for $r = 2, 3$, or 4.

Key words. fractal, iterated function system, fractal interpolation function, spline, cubic spline fractal interpolation function, convergence

AMS subject classifications. 26A18, 37N30, 41A30, 65D05, 65D07, 65D10

DOI. 10.1137/040611070

1. Introduction.

With the advent of fractal geometry [2], the use of stochastic or deterministic fractal models [3, 4, 5] has significantly enhanced the understanding of complexities in nature and different scientific experiments. Hutchinson [6] has studied the deterministic fractal model based on the theory of Iterated Function System (IFS). Using IFS, Barnsley [3, 7] has introduced the concept of Fractal Interpolation Function (FIF) for approximation of naturally occurring functions showing some sort of self-similarity under magnification. A FIF is the fixed point of the Read–Bajraktarević operator acting on different function spaces. Generally, affine FIFs are nondifferentiable functions and the fractal dimensions of their graphs are nonintegers. The generation of FIF codes provides a powerful technique for compression of images, speeches, time series, and other data; see, e.g., [8, 9, 10].

If the experimental data are approximated by a C^r-FIF f, then one can use the fractal dimension of $f^{(r)}$ as a quantitative parameter for the analysis of experimental data. The differentiable C^r-FIF differs from the classical spline interpolation by a functional relation that gives self-similarity on small scales. Barnsley and Harrington [1] have introduced an algebraic method for the construction of a restricted class of C^r-FIF f, which interpolates the prescribed data by providing values of $f^{(k)}$, $k = 1, 2, \ldots, r$, at the initial endpoint of the interval. However, in their method of construction, specifying boundary conditions similar to those for classical splines has been found to be quite difficult to handle. Massopust [11] has attempted to generalize work in [1] by constructing smooth fractal surfaces via integration.

*Received by the editors July 5, 2004; accepted for publication (in revised form) November 22, 2005; published electronically March 17, 2006. This work was partially supported by the Council of Scientific and Industrial Research, India, grant 9/92(160)/98-EMR-I.

http://www.siam.org/journals/sinum/44-2/61107.html

†Department of Mathematics, Indian Institute of Technology Kanpur, Kanpur 208016, India (akbchand@yahoo.com, gp@iitk.ac.in). The first author is presently at BITS Pilani - Goa Campus, Goa, India.

In the present paper, a method of construction of a C^r-fractal function is developed by removing the requirement of prescribing the values of integrals of the given FIF only at the initial endpoint x_0. Thus, a C^r-fractal function is constructed when successive r values of integrals of a FIF are prescribed in any combination at boundary points of the interval. Further, a general method is proposed to construct an interpolating C^r-FIF for the prescribed data with all possible boundary conditions. The complex algebraic method proposed in [1] uses complicated matrices and particular types of end conditions. Using the functional relations present between the values of the C^r-FIF that involve endpoints of the interval, our approach does not need the complex algebraic method in [1]. Our construction settles several queries of Barnsley and Harrington [1] such as (i) which boundary point conditions lead to uniqueness of a C^r-FIF, (ii) what happens if horizontal scalings are in reverse direction and (iii) how to build up the moment integrals theory in this case. The advantage of such a spline FIF construction is that, for prescribed data and given boundary conditions, one can have an infinite number of spline FIFs depending on the vertical scaling factors, giving thereby a large flexibility in the choice of differentiable C^r-FIFs according to the need of an experiment.

Due to the importance of the cubic splines in computer graphics, CAGD, FEM, differential equations, and several engineering applications [12, 13, 14, 15], cubic spline FIF $f_\Delta(x)$ on a mesh Δ is constructed through *moments* $M_n = f''_\Delta(x_n), n = 0, 1, 2, \ldots, N$. These cubic spline FIFs may have any types of boundary conditions as in classical splines. It is shown that the sequence $\{f_{\Delta_k}(x)\}$ converges to the defining data function $\Phi(x)$ on two classes of sequences of meshes at least as rapidly as the square of the mesh norm $\|\Delta_k\|$ converging to zero, provided that $\Phi^{(r)}(x)$ is continuous on $[x_0, x_N]$ for $r = 2, 3$, or 4.

In section 2, some basic results for FIFs are given and a general method for construction of a C^r-FIF with different boundary conditions is enunciated after developing a basic calculus of C^1-FIFs. The construction of a generalized cubic spline FIF through moments is described in section 3 with all possible boundary conditions, as in the classical splines. In section 4, two classes of sequences of meshes are defined and the convergence of suitable sequence of cubic spline FIFs $\{f_{\Delta_k}\}$ to $\Phi \in C^r[x_0, x_N]$, $r = 2, 3$, or 4, is established. Finally, in section 5, the results obtained in section 3 are illustrated by generating certain examples of cubic spline FIFs for a given data and two different sets of vertical scaling factors.

2. A general method for construction of C^r-FIF. We give the basics of the general theory of FIFs and develop the calculus of C^1-FIFs in section 2.1. The principle of construction of a C^r-FIF that interpolates the given data is described in section 2.2.

2.1. Preliminaries and calculus of C^1-FIFs. Barnsley et al. [1, 3, 8, 16, 17] have developed the theory of FIF and its extensive applications. In the following, some of the notations and results of FIF theory, which we will later need, are described.

Let K be a complete metric space with metric d and $\mathcal{H}$ be the set of nonempty compact subsets of K. Then, $\{K; \omega_n, n = 1, 2, \ldots, N\}$ is an iterated function system (IFS) if $\omega_n : K \to K$ is continuous for $n = 1, 2, \ldots, N$. An IFS is called hyperbolic if $d(\omega_n(x), \omega_n(y)) \leq sd(x, y)$ for all $x, y \in K, n = 1, 2, \ldots, N$ and $0 \leq s < 1$. Set $W(A) = \bigcup_{n=1}^{N} \omega_n(A)$ for $A \in \mathcal{H}$. The following proposition gives a condition on an IFS to have a unique attractor.

PROPOSITION 2.1 (see [3]). *Let $\{K; \omega_n, n = 1, 2, \ldots, N\}$ be a hyperbolic IFS.*

Then, it has an unique attractor G such that $h(W^m(A), G) \to 0$ as $m \to \infty$, where $h(.,.)$ is the Hausdorff metric.

Suppose a set of data points $\{(x_i, y_i) \in I \times \mathbb{R} : i = 0, 1, 2, \ldots, N\}$ is given, where $x_0 < x_1 < \cdots < x_N$ and $I = [x_0, x_N]$. Set $K = I \times D$, where D is a suitable compact set in $\mathbb{R}$. Let $L_n : I \longrightarrow I_n = [x_{n-1}, x_n]$ be the affine map satisfying

$$L_n(x_0) = x_{n-1}, \quad L_n(x_N) = x_n \tag{2.1}$$

and $F_n : K \longrightarrow D$ be a continuous function such that

$$\left.\begin{aligned} F_n(x_0, y_0) = y_{n-1},\ F_n(x_N, y_N) = y_n \\ |F_n(x, y) - F_n(x, y^*)| \le \alpha_n |y - y^*| \end{aligned}\right\}, \tag{2.2}$$

where, $(x, y),\ (x, y^*) \in K$, and $0 \le\ \alpha_n < 1$ for all $n = 1, 2, \ldots, N$. Define $\omega_n(x, y) = (L_n(x), F_n(x, y))$ for all $n = 1, 2, \ldots, N$. The definition of a FIF originates from the following proposition.

PROPOSITION 2.2 (see [3]). *The IFS $\{K; \omega_n, n = 1, 2, \ldots, N\}$ has a unique attractor G such that G is the graph of a continuous function $f : I \to \mathbb{R}$ (called FIF associated with IFS $\{K; \omega_n, n = 1, 2, \ldots, N\}$) satisfying $f(x_n) = y_n$ for $n = 0, 1, 2, \ldots, N$.*

The following observations based on Proposition 2.2 are needed in the sequel. Let $\mathcal{F} = \{f : I \to \mathbb{R} \mid f \text{ is continuous}, f(x_0) = y_0 \text{ and } f(x_N) = y_N\}$ and ρ be the sup-norm on $\mathcal{F}$. Then, $(\mathcal{F}, \rho)$ is a complete metric space. The FIF f is the unique fixed point of the Read–Bajraktarević operator T on $(\mathcal{F}, \rho)$ so that

$$Tf(x) \equiv F_n(L_n^{-1}(x), f(L_n^{-1}(x))) = f(x),\ x \in I_n, \quad n = 1, 2, \ldots, N. \tag{2.3}$$

For an affine FIF, L_n and F_n are given by

$$\left.\begin{aligned} L_n(x) = a_n x + b_n \\ F_n(x, y) = \alpha_n y + q_n(x) \end{aligned}\right\}, \quad n = 1, 2, \ldots, N, \tag{2.4}$$

where $q_n(x)$ is an affine map and $|\alpha_n| < 1$.

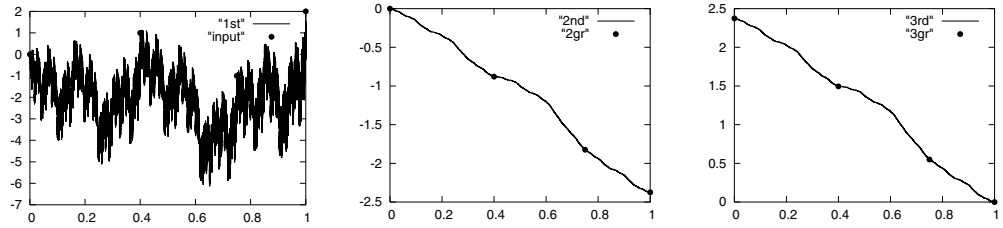
Barnsley and Harrington [1] have observed that the integral of a FIF is also a FIF, although for a different set of interpolation data, provided the value of the integral of the FIF at the initial endpoint of the interval is known. This observation is needed for developing the calculus of C^1-FIFs. Thus, let f be the FIF associated with $\{(L_n(x), F_n(x, y)),\ n = 1, 2, \ldots, N\}$, where F_n is defined by (2.4) and let the value of integral of this FIF be known at x_0. If

$$\hat{f}(x) = \hat{y}_0 + \int_{x_0}^{x} f(\tau) d\tau, \tag{2.5}$$

the function $\hat{f}$ is the FIF associated with IFS $\{(L_n(x), \hat{F}_n(x, y)),\ n = 1, 2, \ldots, N\}$, where $\hat{F}_n(x, y) = a_n \alpha_n y + \hat{q}_n(x)$, $\hat{q}_n(x) = \hat{y}_{n-1} - a_n \alpha_n \hat{y}_0 + a_n \int_{x_0}^{x} q_n(\tau) d\tau$,

$$\hat{y}_n = \hat{y}_0 + \sum_{i=1}^{n} a_i \left\{ \alpha_i (\hat{y}_N - \hat{y}_0) + \int_{x_0}^{x_N} q_i(\tau) d\tau \right\}, \quad n = 1, 2, \ldots, N-1,$$

and $\hat{y}_N = \hat{y}_0 + \sum_{i=1}^{N} a_i \int_{x_0}^{x_N} q_i(\tau) d\tau / 1 - \sum_{i=1} N a_i \alpha_i$. Here, $(x_n, \hat{y}_n), n = 0, 1, 2, \ldots, N$ are interpolation points of FIF $\hat{f}$.



(a) FIF f. (b) $\hat{f}$ with $\hat{y}_0 = 0$. (c) $\hat{f}$ with $\hat{y}_N = 0$.

FIG. 1. *FIF and its integrals.*

Remarks. 1. If the value of the integral of a FIF is known at the final endpoint x_N instead of the initial endpoint x_0, an analogue of the above result can be found by defining

$$\hat{f}(x) = \hat{y}_N - \int_x^{x_N} f(\tau)d\tau. \tag{2.6}$$

The function $\hat{f}$ is the FIF associated with $\{(L_n(x), \hat{F}_n(x,y)),\ n = 1, 2, \ldots, N\}$, where $\hat{F}_n(x,y) = a_n\alpha_n y + \hat{q}_n(x)$, $\hat{q}_n(x) = \hat{y}_n - a_n\alpha_n\hat{y}_N - a_n\int_x^{x_N} q_n(\tau)d\tau$ and the interpolation points of $\hat{f}$ are given by $\hat{y}_n = \hat{y}_N - \sum_{i=n+1}^{N} a_i\{\alpha_i(\hat{y}_N - \hat{y}_0) + \int_{x_0}^{x_N} q_i(\tau)d\tau\}$, $n = 1, 2, \ldots, N-1$ with $\hat{y}_0 = \hat{y}_N - \frac{\sum_{i=1}^{N} a_i \int_{x_0}^{x_N} q_i(\tau)d\tau}{1-\sum_{i=1}^{N} a_i\alpha_i}$. In general, a C^r-FIF interpolating a certain different set of data can be constructed when values of r successive integrals of the FIF are provided at any combination of endpoints.

2. The functional values of FIF $\hat{f}$ are, in general, different for the same set of vertical scaling factors even if $\hat{y}_0$ and $\hat{y}_N$ occurring, respectively, in (2.5) and (2.6) are the same. However, since $\hat{y}_n - \hat{y}_{n-1}$ remains the same for each n in both the cases, the nature of $\hat{f}$ remains the same in both the cases as illustrated by the following example.

Example. Let f be a FIF associated with the data $\{(0,0),\ (\frac{2}{5}, 1),\ (\frac{3}{4}, -1),\ (1,2)\}$ with vertical scaling factor $\alpha_n = 0.8$ for $n = 1, 2, 3$ (Figure 1(a)). Choosing $\hat{y}_0 = 0$, $\hat{f}(x) = \int_{x_0}^{x} f$ interpolates the set of points $\{(0,0), (\frac{2}{5}, \frac{-22}{25}), (\frac{3}{4}, \frac{-73}{40}), (1, \frac{-19}{8})\}$. FIF $\hat{f}$ is associated with the IFS generated by $L_1(x) = \frac{2}{5}x,\ L_2(x) = \frac{7}{20}x + \frac{2}{5},\ L_3(x) = \frac{1}{4}x + \frac{3}{4}$ and $\hat{F}_1(x,y) = \frac{8}{25}y - \frac{3}{25}x^2,\ \hat{F}_2(x,y) = \frac{7}{25}y - \frac{63}{100}x^2 + \frac{7}{20}x - \frac{22}{25},\ \hat{F}_3(x,y) = \frac{1}{5}y + \frac{7}{40}x^2 - \frac{1}{4}x + \frac{73}{40}$. The graph of FIF $\hat{f}$ is shown in Figure 1(b). Next, choosing $\hat{y}_N = 0$, $\hat{f}(x) = -\int_x^{x_N} f$ interpolates the set of points $\{(0, \frac{19}{8}), (\frac{2}{5}, \frac{299}{200}), (\frac{3}{4}, \frac{11}{20}), (1,0)\}$ (Figure 1(c)). In this case, the corresponding IFS contains the same $L_n(x)$ for $n = 1, 2, 3$ and $\hat{F}_1(x,y) = \frac{8}{25}y - \frac{3}{25}x^2 + \frac{323}{100},\ \hat{F}_2(x,y) = \frac{7}{25}y - \frac{63}{100}x^2 + \frac{7}{20}x + \frac{83}{100},\ \hat{F}_3(x,y) = \frac{1}{5}y + \frac{7}{40}x^2 - \frac{1}{4}x + \frac{3}{40}$. The nature of FIFs $\hat{f}$ in Figure 1(b)–(c) remains the same, since the functional values of FIF $\hat{f}$ in Figure 1(c) are shifted by $\frac{19}{8}$ from the functional values of $\hat{f}$ in Figure 1(b) so that $\hat{y}_n - \hat{y}_{n-1}$ remains the same. It is interesting to note that the corresponding functions $\hat{F}_n(x,y)$ for IFS of Figure 1(b)–(c) are not shifted by equal amount although the function $\hat{f}$ is shifted by the fixed amount $\frac{19}{8}$.

In general, the relation between the IFS of FIF f and the IFS of its integral $\hat{f}$ is given as follows [18].

PROPOSITION 2.3. *Let $\hat{f}$ be the FIF defined by* (2.5) *or* (2.6) *for a FIF f with $L_n(x)$ and $F_n(x,y)$ given by* (2.4)*. Then, f is primitive of $\hat{f}$ if and only if $\hat{f}$ is the FIF associated with the IFS $\{\mathbb{R}^2; \hat{w}_n(x,y) = (L_n(x), \hat{F}_n(x,y)), n = 1, 2, \ldots, N\}$, where*

$\hat{F}_n(x,y) = \hat{\alpha}_n y + \hat{q}_n(x)$, $\hat{\alpha}_n = a_n\alpha_n$, *and the polynomial* $\hat{q}_n(x)$ *satisfies* $\hat{q}_n' = a_n q_n$ *for* $n = 1, 2, \ldots, N$.

2.2. Principle of construction of a C^r-FIF. Our approach for the construction of a C^r-FIF that interpolates the given data is based on finding the solution of a system of equations in which any type of boundary conditions are admissible. Such a construction is more general than that of Barnsley and Harrington [1] wherein all the relevant derivatives of the FIF are restricted to be known at the initial endpoint only. The C^r-FIF interpolating prescribed set of data is found as the fixed point of a suitably chosen IFS by using the following procedure.

Let $\{(x_0, y_0), (x_1, y_1), \ldots, (x_N, y_N)\}$, $x_0 < x_1 < \cdots < x_N$, be the given data points and $\mathcal{F}^r = \{g \in C^r(I, \mathbb{R}) \mid g(x_0) = y_0 \text{ and } g(x_N) = y_N\}$, where r is some nonnegative integer and σ is the C^r-norm on $\mathcal{F}^r$. Define the Read–Bajraktarević operator T on $(\mathcal{F}^r, \sigma)$ as

$$Tg(x) = \alpha_n g(L_n^{-1}(x)) + q_n(L_n^{-1}(x)),\ x \in I_n, \quad n = 1, 2, \ldots, N,$$

where $L_n(x) = a_n x + b_n$ satisfies (2.1), $q_n(x)$ is a suitably chosen polynomial, and $|\alpha_n| < a_n^r$ for $n = 1, 2, \ldots, N$. The condition $|\alpha_n| < a_n^r < 1$ gives that T is a contractive operator on $(\mathcal{F}^r, \sigma)$. The fixed point f of T is a FIF that satisfies the functional relation, $f(L_n(x)) = \alpha_n f(x) + q_n(x)$ for $n = 1, 2, \ldots, N$. Using Proposition 2.3, it follows that f' satisfies the functional relation

$$f'(L_n(x)) = \frac{\alpha_n f'(x) + q_n'(x)}{a_n}, \quad n = 1, 2, \ldots, N.$$

Since $\frac{|\alpha_n|}{a_n} \le \frac{|\alpha_n|}{a_n^r} < 1$, f' is a fractal function. Inductively, using the above arguments, the following relations are obtained:

$$f^{(k)}(L_n(x)) = \frac{\alpha_n f^{(k)}(x) + q_n^{(k)}(x)}{a_n^k}, \quad n = 1, 2, \ldots, N, \quad k = 0, 1, 2, \ldots, r, \tag{2.7}$$

where $f^{(0)} = f$ and $q^{(0)} = q$. Since $\frac{|\alpha_n|}{a_n^k} \le \frac{|\alpha_n|}{a_n^r} < 1$, the derivatives $f^{(k)}$, $k = 2, 3, \ldots, r$ are fractal functions. In general $f^{(k)}$, $k = 1, 2, 3, \ldots, r$, interpolates a data different than the given data. In particular, $f^{(r)}$ is an affine FIF if the polynomial $q_n^{(r)}$ occurring in (2.7) with $k = r$ is affine. Thus, $q_n(x)$ is chosen as a polynomial of degree $(r+1)$. Let $q_n(x) = \sum_{k=0}^{r+1} q_{kn} x^k$, $n = 1, 2, \ldots, N$, where the coefficients q_{kn} are chosen suitably such that f interpolates the prescribed data. The continuity of $f^{(k)}$ on I implies

$$f^{(k)}(L_{n+1}(x_0)) = f^{(k)}(L_n(x_N)), \quad k = 0, 1, \ldots, r, \quad n = 1, 2, \ldots, N-1.$$

Therefore, (2.7) results in the following $(r+1)(N-1)$ join-up conditions for $k = 0, 1, \ldots, r$, $n = 1, 2, \ldots, N-1$:

$$\frac{\alpha_{n+1} f^{(k)}(x_0) + q_{n+1}^{(k)}(x_0)}{a_{n+1}^k} = \frac{\alpha_n f^{(k)}(x_N) + q_n^{(k)}(x_N)}{a_n^k}. \tag{2.8}$$

In addition, at the endpoints of the interval, (2.7) implies that the values of $f^{(k)}$ satisfy the following $2r$-conditions:

$$f^{(k)}(x_0) = \frac{\alpha_1 f^{(k)}(x_0) + q_1^{(k)}(x_0)}{a_1^k}, \quad k = 1, 2, \ldots, r, \tag{2.9}$$

and

$$f^{(k)}(x_N) = \frac{\alpha_N f^{(k)}(x_N) + q_N^{(k)}(x_N)}{a_N^k}, \quad k = 1, 2, \ldots, r. \tag{2.10}$$

Let the prescribed interpolation conditions be

$$f(x_n) = y_n, \quad n = 0, 1, \ldots, N. \tag{2.11}$$

In view of (2.8)–(2.11), the total number of conditions for f to interpolate the given data are $(r+1)(N-1) + 2r + (N+1) = (r+2)N + r$. In (2.8)–(2.10), $f^{(k)}(x_0)$ and $f^{(k)}(x_N)$ for $k = 1, 2, \ldots, r$ are $2r$ unknowns and q_{kn}, $k = 0, 1, \ldots, r+1$, $n = 1, 2, \ldots, N$, in the polynomials $q_n(x)$ are additional $(r+2)N$ unknowns. Consequently, in total $(r+2)N+2r$ number of unknowns are to be determined. The principle of construction of a C^r-FIF is to determine these unknowns by choosing additional suitable r conditions in the form of restrictions on the values of the C^r-FIF or the values of its derivatives at the boundary points of $[x_0, x_N]$ such that (2.8)–(2.11) together with these additional conditions are linearly independent. The above unknowns are determined uniquely as the solution of these linear independent system of equations. Thus, the desired C^r-FIF f interpolating the given data is constructed as the attractor of the following IFS:

$$\{\mathbb{R}^2; \omega_n(x, y) = (L_n(x), F_n(x, y) = \alpha_n y + q_n(x)), n = 1, 2, \ldots, N\},$$

where $|\alpha_n| < a_n^r$ and $q_n(x)$, $n = 1, 2, \ldots, N$, are the polynomials with coefficients q_{kn} computed by solving the linear independent system of equations, given by the above procedure. The flexibility of these choices of boundary conditions allows for the construction of a wide range of spline FIFs. Even for a given choice of boundary conditions, depending upon the nature of the problem or simply at the discretion of the user, an infinite number of suitable spline FIFs may be constructed due to the freedom of choices for vertical scaling factors in our construction.

Remarks. 1. Barnsley and Harrington's construction [1] of a C^r-FIF f is done by restricting the choice of boundary values $f^{(k)}(x)$ for $k = 1, 2, \ldots, r$, at the initial endpoint. In our above construction of C^r-FIFs, all kinds of boundary conditions are admissible.

2. It seems that Barnsley and Harrington's question—"whether there exists a FIF as a fixed point of an IFS wherein horizontal scalings are allowed in the reverse direction"—is raised [1], since the construction of a C^r-FIF is based upon restricting boundary values of $f^{(k)}$ at only initial end point of I. Such a question does not arise in our construction since the boundary values of $f^{(k)}$ for C^r-FIF f are admissible at any combination of boundary points of I.

Since the classical cubic splines play a significant role in CAGD, surface analysis, differential equation, FEM, and other applications (see, e.g., [13, 14, 15]), in the sequel a detailed construction for such cubic spline FIFs based on the above approach is given in the following section.

3. Construction of cubic spline FIFs through moments. In the present section, cubic spline FIFs f_Δ are constructed through the *moments* $M_n = f_\Delta''(x_n)$ for $n = 0, 1, 2, \ldots, N$.

DEFINITION 3.1. *A function $f_\Delta(x) \equiv f_\Delta(Y; x)$ is called a cubic spline FIF interpolating a set of ordinates $Y : y_0, y_1, y_2, \ldots, y_N$ with respect to the mesh Δ :*

$x_0 < x_1 < x_2 < \cdots < x_N$ *if* (i) $f_\Delta \in C^2[x_0, x_N]$, (ii) f_Δ *satisfies the interpolation conditions* $f_\Delta(x_n) = y_n, \quad n = 0, 1, \ldots, N$ *and* (iii) *the graph of* f_Δ *is fixed point of a IFS,* $\{\mathbb{R}^2; \omega_n(x, y), n = 1, 2, \ldots, N\}$, *where for* $n = 1, 2, \ldots, N$, $\omega_n(x, y) = (L_n(x), F_n(x, y))$, $L_n(x)$ *is defined by* (2.4), $F_n(x, y) = a_n^2\alpha_n y + a_n^2 q_n(x)$, $0 < |\alpha_n| < 1$, *and* $q_n(x)$ *is a suitable cubic polynomial.*

Using the moments M_n, $n = 0, 1, 2, \ldots, N$, a rectangular system of equations is formed for determining the polynomial $q_n(x)$ by employing the following procedure.

Using property (iii) and (2.3), it follows that $f_\Delta^{''}$ satisfies the functional equation

$$f_\Delta^{''}(L_n(x)) = \alpha_n f_\Delta^{''}(x) + \frac{c_n(x - x_0)}{x_N - x_0} + d_n, \quad n = 1, 2, \ldots, N. \tag{3.1}$$

By (2.1) and (3.1), $c_n = M_n - M_{n-1} - \alpha_n(M_N - M_0)$ and $d_n = M_{n-1} - \alpha_n M_0$. Thus, for $n = 1, 2, \ldots, N$, (3.1) can be rewritten as

$$f_\Delta^{''}(L_n(x)) = \alpha_n f_\Delta{''}(x) + \frac{(M_n - \alpha_n M_N)(x - x_0)}{x_N - x_0} + \frac{(M_{n-1} - \alpha_n M_0)(x_N - x)}{x_N - x_0}. \tag{3.2}$$

The function $f_\Delta^{''}$ being continuous on I could be twice integrated to obtain

$$\begin{aligned} f_\Delta(L_n(x)) =& a_n^2\Bigg\{\alpha_n f_\Delta(x) + \frac{(M_n - \alpha_n M_N)(x - x_0)^3}{6(x_N - x_0)} + \frac{(M_{n-1} - \alpha_n M_0)(x_N - x)^3}{6(x_N - x_0)} \\ &+ c_n^*(x_N - x) + d_n^*(x - x_0)\Bigg\}, \quad n = 1, 2, \ldots, N. \end{aligned} \tag{3.3}$$

Now using interpolation conditions and (2.1), the constants c_n^* and d_n^* are determined as

$$\begin{aligned} c_n^* &= \frac{1}{x_N - x_0}\left(\frac{y_{n-1}}{a_n^2} - \alpha_n y_0\right) - \frac{(M_{n-1} - \alpha_n M_0)(x_N - x_0)}{6}, \\ d_n^* &= \frac{1}{x_N - x_0}\left(\frac{y_n}{a_n^2} - \alpha_n y_N\right) - \frac{(M_n - \alpha_n M_N)(x_N - x_0)}{6}. \end{aligned}$$

Thus, the functional equation (3.3) for the cubic spline FIF in terms of moments can be written as

$$\begin{aligned} f_\Delta(L_n(x)) = a_n^2\Bigg\{&\alpha_n f_\Delta(x) + \frac{(M_n - \alpha_n M_N)(x - x_0)^3}{6(x_N - x_0)} + \frac{(M_{n-1} - \alpha_n M_0)(x_N - x)^3}{6(x_N - x_0)} \\ &- \frac{(M_{n-1} - \alpha_n M_0)(x_N - x_0)(x_N - x)}{6} - \frac{(M_n - \alpha_n M_N)(x_N - x_0)(x - x_0)}{6} \\ &+ \left(\frac{y_{n-1}}{a_n^2} - \alpha_n y_0\right)\frac{x_N - x}{x_N - x_0} + \left(\frac{y_n}{a_n^2} - \alpha_n y_N\right)\frac{x - x_0}{x_N - x_0}\Bigg\}, \quad n = 1, 2, \ldots, N. \end{aligned} \tag{3.4}$$

It follows by (3.4) that $f_\Delta(x)$ is continuous on $[x_0, x_N]$ and satisfies the interpolating conditions $f_\Delta(x_n) = y_n$, $n = 0, 1, 2, \ldots, N$. Further, (3.4) gives that, on $[x_{i-1}, x_i]$,

$i = 1, 2, \ldots, N,$

(3.5)
$$f'_\Delta(L_i(x)) = a_i\Bigg\{\alpha_i f'_\Delta(x) + \frac{(M_i - \alpha_i M_N)(x - x_0)^2}{2(x_N - x_0)} - \frac{(M_{i-1} - \alpha_i M_0)(x_N - x)^2}{2(x_N - x_0)} - \frac{[M_i - M_{i-1} - \alpha_i(M_N - M_0)](x_N - x_0)}{6} + \left[\frac{y_i - y_{i-1}}{a_i^2} - \alpha_i(y_N - y_0)\right]\frac{1}{x_N - x_0}\Bigg\}.$$

Denote $x_n - x_{n-1}$ by h_n = for $n = 1, 2, \ldots, N$. Since, by property (i), $f'_\Delta(x)$ is continuous at $x_1, x_2, \ldots, x_{N-1}$, $\lim_{x\to x_n^-} f'_\Delta(x) = \lim_{x\to x_n^+} f'_\Delta(x),\ n = 1, 2, \ldots, N-1$. Thus, using (3.5) for $i = n$ and $i = n + 1$, we have

(3.6)
$$\begin{aligned} &-a_{n+1}\alpha_{n+1}f'_\Delta(x_0) - \frac{\alpha_n h_n + 2\alpha_{n+1}h_{n+1}}{6}M_0 + \frac{h_n}{6}M_{n-1} + \frac{h_n + h_{n+1}}{3}M_n \\ &\quad + \frac{h_{n+1}}{6}M_{n+1} - \frac{2\alpha_n h_n + \alpha_{n+1}h_{n+1}}{6}M_N + a_n\alpha_n f'_\Delta(x_N) \\ &= \frac{y_{n+1} - y_n}{h_{n+1}} - \frac{y_n - y_{n-1}}{h_n} - (a_{n+1}\alpha_{n+1} - a_n\alpha_n)\frac{y_N - y_0}{x_N - x_0}, \quad n = 1, 2, \ldots, N-1. \end{aligned}$$

Introducing the notations,

$$\begin{aligned} &A_n^* = \frac{-6a_{n+1}\alpha_{n+1}}{h_n + h_{n+1}},\ A_n = \frac{-(\alpha_n h_n + 2\alpha_{n+1}h_{n+1})}{h_n + h_{n+1}},\ \lambda_n = \frac{h_{n+1}}{h_n + h_{n+1}}, \\ &\mu_n = 1 - \lambda_n,\ B_n = \frac{-(2\alpha_n h_n + \alpha_{n+1}h_{n+1})}{h_n + h_{n+1}},\ B_n^* = \frac{6a_n\alpha_n}{h_n + h_{n+1}}, \end{aligned}$$

for $n = 1, 2, \ldots, N-1$, the continuity relation (3.6) reduces to

(3.7)
$$\begin{aligned} &A_n^* f'_\Delta(x_0) + A_n M_0 + \mu_n M_{n-1} + 2M_n + \lambda_n M_{n+1} + B_n M_N + B_n^* f'_\Delta(x_N) \\ &\quad = \frac{6[(y_{n+1} - y_n)/h_{n+1} - (y_n - y_{n-1})/h_n]}{h_n + h_{n+1}} - \frac{6(a_{n+1}\alpha_{n+1} - a_n\alpha_n)}{h_n + h_{n+1}}\frac{y_N - y_0}{x_N - x_0}. \end{aligned}$$

Next, (3.5) with $x = x_0$ and $i = 1$ gives the following functional relation for $f'_\Delta(x_0)$:

(3.8)
$$\begin{aligned} &6(1 - a_1\alpha_1)f'_\Delta(x_0) + 2(1 - \alpha_1)h_1 M_0 + h_1 M_1 - \alpha_1 h_1 M_N \\ &\quad = 6/h_1[y_1 - y_0 - \alpha_1 a_1^2(y_N - y_0)]. \end{aligned}$$

Similarly, (3.5) with $x = x_N$ and $i = N$ gives

(3.9)
$$\begin{aligned} &-\alpha_N h_N M_0 + h_N M_{N-1} + 2(1 - \alpha_N)h_N M_N - 6(1 - a_N\alpha_N)f'_\Delta(x_N) \\ &\quad = -6/h_N[y_N - y_{N-1} - \alpha_N a_N^2(y_N - y_0)]. \end{aligned}$$

To write the system of equations given by (3.7)–(3.9) in matrix from, we introduce the following notations:

$$\begin{aligned} &A_0^* = 6(1 - a_1\alpha_1),\ A_0 = 2(1 - \alpha_1)h_1,\ \lambda_0 = h_1,\ B_0 = -\alpha_1 h_1, \\ &A_N = -\alpha_N h_N,\ \mu_N = h_N,\ B_N = 2(1 - \alpha_N)h_N,\ B_N^* = -6(1 - a_N\alpha_N), \\ &d_0 = 6/h_1[y_1 - y_0 - \alpha_1 a_1^2(y_N - y_0)],\ d_N = -6/h_N[y_N - y_{N-1} - \alpha_N a_N^2(y_N - y_0)]. \end{aligned}$$

Thus, the matrix form of defining (3.7)–(3.9) is

(3.10)
$$\begin{bmatrix} A_0^* & A_0 & \lambda_0 & 0 & 0 & \dots & 0 & 0 & 0 & B_0 & 0 \\ A_1^* & A_1+\mu_1 & 2 & \lambda_1 & 0 & \dots & 0 & 0 & 0 & B_1 & B_1^* \\ A_2^* & A_2 & \mu_2 & 2 & \lambda_2 & \dots & 0 & 0 & 0 & B_2 & B_2^* \\ A_3^* & A_3 & 0 & \mu_3 & 2 & \dots & 0 & 0 & 0 & B_3 & B_3^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{N-3}^* & A_{N-3} & 0 & 0 & 0 & \dots & 2 & \lambda_{N-3} & 0 & B_{N-3} & B_{N-3}^* \\ A_{N-2}^* & A_{N-2} & 0 & 0 & 0 & \dots & \mu_{N-2} & 2 & \lambda_{N-2} & B_{N-2} & B_{N-2}^* \\ A_{N-1}^* & A_{N-1} & 0 & 0 & 0 & \dots & 0 & \mu_{N-1} & 2 & \lambda_{N-1}+B_{N-1} & B_{N-1}^* \\ 0 & A_N & 0 & 0 & 0 & \dots & 0 & 0 & \mu_N & B_N & B_N^* \end{bmatrix} \begin{bmatrix} f'_\Delta(x_0) \\ M_0 \\ M_1 \\ M_2 \\ \\ \vdots \\ M_{N-2} \\ M_{N-1} \\ M_N \\ f'_\Delta(x_N) \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ \\ \vdots \\ d_{N-3} \\ d_{N-2} \\ d_{N-1} \\ d_N \end{bmatrix},$$

where d_n, $n = 1, 2, \dots, N-1$, is given by the right side expression of (3.7) and $f'_\Delta(x_0), M_0, M_1, \dots, M_N, f'_\Delta(x_N)$ are unknowns. Equation (3.10), consisting of a coefficient matrix of order $(N+1)\times(N+3)$, is the desired rectangular matrix equation for computing the unknowns coefficients q_{kn} of the polynomial $q_n(x)$.

Boundary Conditions. By prescribing suitable boundary conditions as in the case of classical cubic splines, the rectangular matrix system of equations (3.10) reduces to a square matrix system of equations. Let the data $\{(x_n, y_n) : n = 0, 1, 2, \dots, N\}$ be generated by a continuous function Φ that is to be approximated by cubic spline FIF f_Δ. The following kinds of boundary conditions are admissible.

Boundary conditions of Type-I: In this case, the values of the first derivative are prescribed at the endpoints of the interval $[x_0, x_N]$, i.e., $f'_\Delta(x_0) = \Phi'(x_0)$, $\quad f'_\Delta(x_N) = \Phi'(x_N)$. So, (3.10) reduces to the following system of equations:

(3.11)
$$\begin{bmatrix} A_0 & \lambda_0 & 0 & 0 & \dots & 0 & 0 & 0 & B_0 \\ A_1+\mu_1 & 2 & \lambda_1 & 0 & \dots & 0 & 0 & 0 & B_1 \\ A_2 & \mu_2 & 2 & \lambda_2 & \dots & 0 & 0 & 0 & B_2 \\ A_3 & 0 & \mu_3 & 2 & \dots & 0 & 0 & 0 & B_3 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ A_{N-3} & 0 & 0 & 0 & \dots & 2 & \lambda_{N-3} & 0 & B_{N-3} \\ A_{N-2} & 0 & 0 & 0 & \dots & \mu_{N-2} & 2 & \lambda_{N-2} & B_{N-2} \\ A_{N-1} & 0 & 0 & 0 & \dots & 0 & \mu_{N-1} & 2 & \lambda_{N-1}+B_{N-1} \\ A_N & 0 & 0 & 0 & \dots & 0 & 0 & \mu_N & B_N \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \\ \\ \vdots \\ M_{N-3} \\ M_{N-2} \\ M_{N-1} \\ M_N \end{bmatrix} = \begin{bmatrix} d_0^1 \\ d_1^1 \\ d_2^1 \\ \\ \vdots \\ d_{N-2}^1 \\ d_{N-1}^1 \\ d_N^1 \end{bmatrix},$$

where $d_0^1 = d_0 - A_0^* f'_\Delta(x_0)$, $d_n^1 = d_n - A_n^* f'_\Delta(x_0) - B_n^* f'_\Delta(x_N)$ for $n = 1, 2, \dots, N-1$, and $d_N^1 = d_N - B_N^* f'_\Delta(x_N)$. Thus, boundary conditions of Type-I result in determination of the *complete cubic spline FIF* by using (3.11).

Boundary conditions of Type-II: In this case, the values of the second derivative given at the endpoints of the segment $[x_0, x_N]$ are prescribed as $f''_\Delta(x_0) = \Phi''(x_0) = M_0, f''_\Delta(x_N) = \Phi''(x_N) = M_N$. With these boundary conditions, (3.10) reduces to

(3.12)
$$\begin{bmatrix} A_0^* & \lambda_0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ A_1^* & 2 & \lambda_1 & 0 & \dots & 0 & 0 & 0 & B_1^* \\ A_2^* & \mu_2 & 2 & \lambda_2 & \dots & 0 & 0 & 0 & B_2^* \\ A_3^* & 0 & \mu_3 & 2 & \dots & 0 & 0 & 0 & B_3^* \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ A_{N-3}^* & 0 & 0 & 0 & \dots & 2 & \lambda_{N-3} & 0 & B_{N-3}^* \\ A_{N-2}^* & 0 & 0 & 0 & \dots & \mu_{N-2} & 2 & \lambda_{N-2} & B_{N-2}^* \\ A_{N-1}^* & 0 & 0 & 0 & \dots & 0 & \mu_{N-1} & 2 & B_{N-1}^* \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & \mu_N & B_N^* \end{bmatrix} \begin{bmatrix} f'_\Delta(x_0) \\ M_1 \\ M_2 \\ M_3 \\ \\ \vdots \\ M_{N-3} \\ M_{N-2} \\ M_{N-1} \\ f'_\Delta(x_N) \end{bmatrix} = \begin{bmatrix} d_0^2 \\ d_1^2 \\ d_2^2 \\ \\ \vdots \\ d_{N-2}^2 \\ d_{N-1}^2 \\ d_N^2 \end{bmatrix},$$

where $d_1^2 = d_1 - (A_1+\mu_1)M_0 - B_1\mu_N$, $d_{N-1}^2 = d_{N-1} - A_{N-1}M_0 - (B_{N-1}+\lambda_{N-1})M_N$, and $d_n^2 = d_n - A_nM_0 - B_nM_N$ for $n = 0, 2, 3, \dots, N-2, N$. Taking free end conditions $M_0 = 0$ and $M_N = 0$, the *natural cubic spline FIF* is computed by using (3.12).

Boundary conditions of Type-III: In this case, the boundary conditions involve the functional values, the values of first and second derivatives of the cubic splines at both endpoints, i.e., $f_\Delta(x_0) = f_\Delta(x_N), \quad f'_\Delta(x_0) = f'_\Delta(x_N), \quad f''_\Delta(x_0) = f''_\Delta(x_N)$. With these boundary conditions, (3.10) takes the following form:

(3.13)
$$\begin{bmatrix} A_0^* & \lambda_0 & 0 & 0 & \dots & 0 & 0 & 0 & A_0+B_0 \\ A_1^*+B_1^* & 2 & \lambda_1 & 0 & \dots & 0 & 0 & 0 & A_1+B_1+\mu_1 \\ A_2^*+B_2^* & \mu_2 & 2 & \lambda_2 & \dots & 0 & 0 & 0 & A_2+B_2 \\ A_3^*+B_3^* & 0 & \mu_3 & 2 & \dots & 0 & 0 & 0 & A_3+B_3 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ A_{N-3}^*+B_{N-3}^* & 0 & 0 & 0 & \dots & 2 & \lambda_{N-3} & 0 & A_{N-3}+B_{N-3} \\ A_{N-2}^*+B_{N-2}^* & 0 & 0 & 0 & \dots & \mu_{N-2} & 2 & \lambda_{N-2} & A_{N-2}+B_{N-2} \\ A_{N-1}^*+B_{N-1}^* & 0 & 0 & 0 & \dots & 0 & \mu_{N-1} & 2 & A_{N-1}+B_{N-1}+\lambda_{N-1} \\ B_N^* & 0 & 0 & 0 & \dots & 0 & 0 & \mu_N & A_N+B_N \end{bmatrix} \begin{bmatrix} f'_\Delta(x_0) \\ M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{N-3} \\ M_{N-2} \\ M_{N-1} \\ M_N \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{N-3} \\ d_{N-2} \\ d_{N-1} \\ d_N \end{bmatrix}.$$

The *periodic cubic spline FIF* is computed by using (3.13).

We confine ourselves to the boundary conditions of Type-I, Type-II, and Type-III only for the convergence results in section 4 although, in addition to the above kinds of boundary conditions, the following types of boundary conditions are also admissible in our approach.

Boundary conditions of Type-IV: In this case, the values of derivatives of given function are known at either initial or final endpoint of the interval, i.e., $f'_\Delta(x_0) = \Phi'(x_0),\ f''_\Delta(x_0) = \Phi''(x_0) = M_0$ or $f'_\Delta(x_N) = \Phi'(x_N),\ f''_\Delta(x_N) = \Phi''(x_N) = M_N$. Barnsley and Harrington [1] used the former set of conditions to obtain the cubic spline FIF by employing an involved algebraic method.

Boundary conditions of Type-V: In this type of boundary condition, two sets of conditions are possible depending on the values of different order of the derivatives at both endpoints, i.e., $f'_\Delta(x_0) = \Phi'(x_0),\ f''_\Delta(x_N) = \Phi''(x_N) = M_N$ or $f'_\Delta(x_N) = \Phi'(x_N),\ f_\Delta{}''(x_0) = \Phi''(x_0) = M_0$. In order to find the respective unknowns, the square matrix of order $(N+1)$ for the boundary conditions of Type-IV and Type-V can be obtained from (3.10).

Boundary conditions of Type-VI: Two linear equations involving $M_0,\ f'_\Delta(x_0)$, $f'_\Delta(x_N)$, and M_N are considered in this case such that these and (3.10) form a linearly independent system of equations. The resulting square matrix of order $(N+3)$ can be solved to find all $(N+3)$ unknowns simultaneously.

Using one of the above types of boundary conditions and solving the corresponding system of equations, the values $f'_\Delta(x_0), M_0, M_1, \dots, M_N$ and $f'_\Delta(x_N)$ are determined. These values of $M_n,\ n = 0, 1, 2, \dots, N$, are used in the construction of an associated IFS given by

$$\{\mathbb{R}^2; \omega_n(x,y) = (L_n(x), F_n(x,y)), n = 1, 2, \dots, N\}, \tag{3.14}$$

where $L_n(x) = a_n x + b_n$ and

(3.15)
$$\begin{aligned} F_n(x,y) = a_n^2 \Bigg\{ & \alpha_n f_\Delta(x) + \frac{(M_n - \alpha_n M_N)(x-x_0)^3}{6(x_N - x_0)} + \frac{(M_{n-1} - \alpha_n M_0)(x_N - x)^3}{6(x_N - x_0)} \\ & - \frac{(M_{n-1} - \alpha_n M_0)(x_N - x_0)(x_N - x)}{6} - \frac{(M_n - \alpha_n M_N)(x_N - x_0)(x - x_0)}{6} \\ & + \left(\frac{y_{n-1}}{a_n^2} - \alpha_n y_0\right)\frac{x_N - x}{x_N - x_0} + \left(\frac{y_n}{a_n^2} - \alpha_n y_N\right)\frac{x - x_0}{x_N - x_0} \Bigg\}. \end{aligned}$$

The graph of the desired cubic spline is the fixed point of the IFS given by (3.14).

Remarks. 1. If the vertical scaling factor $\alpha_n = 0$ for $n = 1, 2, \ldots, N$, $F_n(x, y)$ reduces to a cubic polynomial in each subinterval of I so that in this case the resulting FIF is a classical cubic spline.

2. By the fixed point theorem, with prescribed ordinates at mesh points, the nonperiodic spline FIF always exists and is unique for a given choice of vertical scaling factors. This spline FIF has simple end supports ($M_0 = 0, M_N = 0$), prescribed end moments or simple supports at points beyond mesh extremities. Similarly, the periodic spline FIF exists and is unique for a given data and a given choice of vertical scaling factors. Since the moments depend upon the vertical scaling factors α_n, by changing α_n, infinitely many nonperiodic splines or periodic splines having the same boundary conditions can be constructed. This gives an additional advantage for the applications of the cubic spline FIF over the applications of the classical cubic spline since there is no flexibility in choosing the latter once the boundary conditions are fixed.

3. Clearly, the replacement of y_n by $y_n + c$ does not affect the right-hand sides of (3.7)–(3.9). Thus, $f_\Delta(Y; x) + \eta = f_\Delta(\bar{Y}; x)$, where $\bar{Y} : \bar{y}_0, \bar{y}_1, \ldots, \bar{y}_N$ and $\bar{y}_n = y_n + \eta$, $n = 0, 1, 2, \ldots, N$, with η being a constant. Since the moments M_n do not change by the translation of the ordinates by a constant η, it follows that it is possible to associate more than one cubic spline FIF for a given set of moments M_n. This property of cubic spline FIF f_Δ is analogous to the corresponding property of the periodic classical spline [19].

4. The existence of spline FIF f_Δ gives (3.7)–(3.9). Further, if spline FIF f_Δ is periodic, adding (3.7) to (3.9) gives

$$\sum_{n=1}^{N} [(h_n + h_{n+1})M_n - 2\alpha_n h_n M_N] = 0. \tag{3.16}$$

The condition (3.16) is therefore a necessary condition for the existence of the periodic cubic spline FIF for prescribed moments M_n. With $\alpha_n = 0$ for $n = 1, 2, \ldots, N$, the condition (3.16) reduces to the necessary condition for the existence of periodic classical cubic spline associated with M_n [19, p. 17].

5. For a prescribed set of data and a suitable choice of α_n satisfying $0 \le |\alpha_n| < 1$, it follows from (3.15) that, on the space $\mathcal{F}^* = \{f \in C^2(I, \mathbb{R}) \mid f(x_0) = y_0 \text{ and } f(x_N) = y_N\}$, cubic spline FIF f_Δ is the fixed point of Read–Bajraktarević operator T^* defined by

(3.17)

$$\begin{aligned} T^* f(x) = a_n^2 \Bigg\{ & \alpha_n f(L_n^{-1}(x)) + \frac{(M_n - \alpha_n M_N)(L_n^{-1}(x) - x_0)^3}{6(x_N - x_0)} \\ & + \frac{(M_{n-1} - \alpha_n M_0)(x_N - L_n^{-1}(x))^3}{6(x_N - x_0)} - \frac{(M_{n-1} - \alpha_n M_0)(x_N - x_0)(x_N - L_n^{-1}(x))}{6} \\ & - \frac{(M_n - \alpha_n M_N)(x_N - x_0)(L_n^{-1}(x) - x_0)}{6} \\ & + \left(\frac{y_{n-1}}{a_n^2} - \alpha_n y_0\right)\frac{x_N - L_n^{-1}(x)}{x_N - x_0} + \left(\frac{y_n}{a_n^2} - \alpha_n y_N\right)\frac{L_n^{-1}(x) - x_0}{x_N - x_0} \Bigg\}, \end{aligned}$$

where $x \in I_n$ for $n = 1, 2, \ldots, N$. Since (3.10) is derived from the fixed point relation $T^* f_\Delta = f_\Delta$, the solution of each of the equations (3.11)–(3.13) is unique due to uniqueness of the fixed point. Hence, the coefficient matrices in the systems (3.11)–(3.13) are invertible.

6. The moment integral $\Phi_m = \int_I x^m \Phi(x)\ dx,\ m = 0, 1, 2, \ldots,$ of the data generating function Φ can be approximately calculated by integral moments $f_\Delta^m \equiv \int_I x^m f_\Delta(x)\ dx$ of the cubic spline FIF. One can evaluate explicitly the moment integral f_Δ^m in terms of $f_\Delta^{m-1}, f_\Delta^{m-2}, \ldots, f_\Delta^0$, the data points, the vertical scaling factors $\alpha_n, n = 1, 2, \ldots, N,$ and $Q_m = \int_I x^m Q(x)\ dx$, where $Q(x) = q_n \circ L_n^{-1}(x),\ x \in I_n$. Thus, Barnsley and Harrington's query [1] regarding the moment integrals in case of reverse horizontal scaling is already taken into account in our construction.

4. Convergence of cubic spline FIFs. Define a sequence $\{\Delta_k\}$ of meshes on $[x_0, x_N]$ as $\Delta_k : x_0 = x_{k,0} < x_{k,1} < \cdots < x_{k,N_k} = x_N$, then set $h_{k,n} = x_{k,n} - x_{k,n-1}$ and $\|\Delta_k\| = \max_{1 \le n \le N_k} h_{k,n}$.

We establish that sequences of cubic spline FIFs $\{f_{\Delta_k}(x)\}$ converge to $\Phi(x)$ on suitable sequences of meshes $\{\Delta_k\}$ at the rate of square of the mesh norm $\|\Delta_k\|$, where $\Phi \in C^r(I)$, $r = 2, 3$, or 4, is the data generating function. Since the matrices associated with the cubic spline FIF, satisfying the boundary conditions of Type-I, Type-II, or Type-III (periodic), are not, in general, diagonally dominant and $f_\Delta''(x)$ is not piecewise linear, the convergence procedure for classical cubic spline [19] cannot be adopted for establishing the convergence of the cubic spline FIF. Our convergence results for cubic spline FIFs are in fact derived by using the convergence results for classical splines.

Let $\mathcal{F}^*$ be the set of cubic spline FIFs on the given mesh Δ, interpolating the values y_n at the mesh points. From (3.17), it is clear that for $x \in I = [x_0, x_N]$,

$$f_\Delta(L_n(x)) = a_n^2 \alpha_n f_\Delta(x) + a_n^2 q_n(x), \tag{4.1}$$

where $q_n(x)$ is a cubic polynomial for $n = 1, 2, \ldots, \mathrm{N}$. Throughout the sequel, we assume $|\alpha_n| \le s < 1$ for a fixed s and denote $q_n(\alpha_n, x) \equiv q_n(x)$ for $n = 1, 2, \ldots, N$.

LEMMA 4.1. *Let $f_\Delta(x)$ and $S_\Delta(x)$, respectively, be the cubic spline FIF and the classical cubic spline with respect to the mesh $\Delta : x_0 < x_1 < \cdots < x_N$, interpolating a set of ordinates $\{y_0, y_1, \ldots, y_N\}$ at the mesh points. Let the cubic polynomial $q_n(\alpha_n, x)$ associated with the IFS for FIF $f_\Delta(x)$ satisfy*

$$\left| \frac{\partial^{1+r} q_n(\tau_n, x)}{\partial \alpha_n \partial x^r} \right| \le K_r \tag{4.2}$$

for $|\tau_n| \in (0, s a_n^r),\ x \in I_n,\ r = 0, 1, 2,$ and $n = 1, 2, \ldots, N$. Then,

$$\|f_\Delta^{(r)} - S_\Delta^{(r)}\|_\infty \le \frac{\|\Delta\|^{2-r} \max_{1 \le n \le N} |\alpha_n|}{|I|^{2-r} - \|\Delta\|^{2-r} \max_{1 \le n \le N} |\alpha_n|} (\|S_\Delta^{(r)}\|_\infty + K_r), \quad r = 0, 1, 2, \tag{4.3}$$

where $|I| = |x_N - x_0|$.

Proof. Denote $\mathcal{B}_r = [-s a_1^r, s a_1^r] \times [-s a_2^r, s a_2^r] \times \cdots \times [-s a_N^r, s a_N^r] \equiv \bigotimes_{n=1}^N [-s a_n^r, s a_n^r]$.

Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathcal{B}_0$ and $r = 0$. Since cubic spline FIF f_Δ is unique for a set of scaling factors $\alpha \in \mathcal{B}_0$ and a prescribed boundary condition, using (3.17) the Read–Bajraktarević operator $T_\alpha^* : \mathcal{F}^* \to \mathcal{F}^*$ can be rewritten as

$$T_\alpha^* f^*(x) = a_n^2 \alpha_n f^*(L_n^{-1}(x)) + a_n^2 q_n(\alpha_n, L_n^{-1}(x)), \quad x \in I_n, n = 1, 2, \ldots, N. \tag{4.4}$$

For a given $\alpha \in \mathcal{B}_0$ and at least one $\alpha_n \ne 0$ in (4.4), cubic spline FIF f_Δ is the fixed point of T_α^*. For $\alpha^* = (0, 0, \ldots, 0) \in \mathcal{B}_0$, the classical cubic spline S_Δ is the fixed

point of $T^*_{\alpha^*}$, since in this case $q_n(\alpha_n, x)$ is a polynomial only in x for $n = 1, 2, \ldots, N$. Therefore, using (4.4), for $x \in I_n$,

$$\begin{aligned}|T^*_\alpha f_\Delta(x) - T^*_\alpha S_\Delta(x)| &= |a_n^2\alpha_n f_\Delta(L_n^{-1}(x)) + a_n^2 q_n(\alpha_n, L_n^{-1}(x)) \\ &\quad -[a_n^2\alpha_n S_\Delta(L_n^{-1}(x)) + a_n^2 q_n(\alpha_n, L_n^{-1}(x))]| \\ &\le \frac{\|\Delta\|^2}{|I|^2} \max_{1\le n\le N}|\alpha_n|\, \|f_\Delta - S_\Delta\|_\infty.\end{aligned}$$

Since the above inequality holds for $n = 1, 2, \ldots, N$, it follows that

$$\|T^*_\alpha f_\Delta - T^*_\alpha S_\Delta\|_\infty \le \frac{\|\Delta\|^2}{|I|^2} \max_{1\le n\le N}|\alpha_n|\, \|f_\Delta - S_\Delta\|_\infty. \tag{4.5}$$

Further, for $x \in I_n$, using (4.4) and Mean Value Theorem,

$$\begin{aligned}|T^*_\alpha S_\Delta(x) - T^*_{\alpha^*} S_\Delta(x)| &= |a_n^2\alpha_n S_\Delta(L_n^{-1}(x)) + a_n^2 q_n(\alpha_n, L_n^{-1}(x)) - a_n^2 q_n(0, L_n^{-1}(x))| \\ &\le a_n^2|\alpha_n|\|S_\Delta\|_\infty + a_n^2|\alpha_n|\left|\frac{\partial q_n(\tau_n, L_n^{-1}(x))}{\partial\alpha_n}\right| \\ &\le \frac{\|\Delta\|^2}{|I|^2} \max_{1\le n\le N}|\alpha_n|\, (\|S_\Delta\|_\infty + K_0).\end{aligned}$$

Since the above inequality holds for $n = 1, 2, \ldots, N$,

$$\|T^*_\alpha S_\Delta - T^*_{\alpha^*} S_\Delta\|_\infty \le \frac{\|\Delta\|^2}{|I|^2} \max_{1\le n\le N}|\alpha_n|\, (\|S_\Delta\|_\infty + K_0). \tag{4.6}$$

Using (4.5)–(4.6) together with the inequality

$$\|f_\Delta - S_\Delta\|_\infty = \|T^*_\alpha f_\Delta - T^*_{\alpha^*} S_\Delta\|_\infty \le \|T^*_\alpha f_\Delta - T^*_\alpha S_\Delta\|_\infty + \|T^*_\alpha S_\Delta - T^*_{\alpha^*} S_\Delta\|_\infty$$

gives that

$$\|f_\Delta - S_\Delta\|_\infty \le \frac{\|\Delta\|^2 \max_{1\le n\le N}|\alpha_n|}{|I|^2 - \|\Delta\|^2 \max_{1\le n\le N}|\alpha_n|}(\|S_\Delta\|_\infty + K_0).$$

This proves Lemma 4.1 for $r = 0$. For $r = 1, 2$, the proof of the lemma is analogous to the proof given above for $r = 0$, by taking $\mathcal{B}_1$, $\mathcal{B}_2$, respectively, in place of $\mathcal{B}_0$ and defining Read–Bajraktarević operator on $\mathcal{F}^*_r = \{f \in C^{2-r}(I, \mathbb{R}) \mid f(x_0) = y_0 \text{ and } f(x_N) = y_N\}$ by

$$T^* f^{(r)}(x) = a_n^{2-r} f^{(r)}(L_n^{-1}(x)) + a_n^{2-r} q_n^{(r)}(\alpha_n, L_n^{-1}(x)), \quad r = 1, 2,$$

in place of (4.4). □

For studying the convergence of cubic spline FIFs to a data generating function through sequences of meshes $\{\Delta_k\}$ on $[x_0, x_N]$, define the following types of meshes depending upon vertical scaling factors $\alpha_{k,n}$.

Class A $\{\{\Delta_k\} : \text{For each } k, \max_{1\le n\le N_k}|\alpha_{k,n}| \le \|\Delta_k\| < 1\}$.

Class B $\{\{\Delta_k\} : \text{For each } k, |\alpha_{k,i}| > \|\Delta_k\| \text{ for some } i, 1 \le i \le N_k\}$.

The convergence of a suitable sequence of cubic spline FIFs to the function Φ in $C^2[x_0, x_N]$ generating the interpolation data is described by the following theorem.

THEOREM 4.2. *Let $\Phi \in C^2[x_0, x_N]$ and cubic spline FIFs $f_{\Delta_k}(x)$ satisfy boundary conditions of Type-*I*, Type-*II*, or Type-*III *(periodic) on a sequence of meshes $\{\Delta_k\}$ on $[x_0, x_N]$ with $\lim_{k\to\infty} \|\Delta_k\| = 0$. If $\{\Delta_k\}$ is in Class A, then*

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \circ(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.7}$$

Further, if $\{\Delta_k\}$ is in Class B, then

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \bigcirc(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.8}$$

Proof. By Lemma 4.1, each element of the sequence $\{\Delta_k\}$ satisfies

$$\|f^{(r)}_{\Delta_k} - S^{(r)}_{\Delta_k}\|_\infty \le \frac{\|\Delta_k\|^{2-r} \max_{1\le n\le N_k} |\alpha_{k,n}|}{|I|^{2-r} - \|\Delta_k\|^{2-r} \max_{1\le n\le N_k} |\alpha_{k,n}|}(\|S^{(r)}_{\Delta_k}\|_\infty + K_r), \quad r = 0, 1, 2. \tag{4.9}$$

Further, it is known that [19, 20]

$$\|\Phi^{(r)} - S^{(r)}_{\Delta_k}\|_\infty \le 5\|\Delta_k\|^{2-r}\omega(\Phi^{(r)}; \|\Delta_k\|) \quad (r = 0, 1, 2), \tag{4.10}$$

where $\omega(\Phi; x)$ is the modulus of continuity of $\Phi(x)$. By using the triangle inequality and (4.10), it follows that

$$\|S^{(r)}_{\Delta_k}\|_\infty \le \|\Phi^{(r)}\|_\infty + 5\|\Delta_k\|^{2-r}\omega(\Phi^{(r)}; \|\Delta_k\|). \tag{4.11}$$

The inequality

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty \le \|\Phi^{(r)} - S^{(r)}_{\Delta_k}\|_\infty + \|S^{(r)}_{\Delta_k} - f^{(r)}_{\Delta_k}\|_\infty \tag{4.12}$$

together with (4.9)–(4.11) gives

$$\begin{aligned}\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty \le{}& \|\Delta_k\|^{2-r}\Bigg\{5\omega(\Phi^{(r)}; \|\Delta_k\|) \\ &+ \frac{(\|\Phi^{(r)}\|_\infty + 5\|\Delta_k\|^{2-r}\omega(\Phi^{(r)}; \|\Delta_k\|) + K_r) \max_{1\le n\le N_k} |\alpha_{k,n}|}{|I|^{2-r} - \|\Delta_k\|^{2-r} \max_{1\le n\le N_k} |\alpha_{k,n}|}\Bigg\}.\end{aligned} \tag{4.13}$$

Since $\Phi \in C^2(I)$ and $\max_{1\le n\le N_k} |\alpha_{k,n}| \le \|\Delta_k\| < 1$, the right-hand side of (4.13) tends to zero as $k \to \infty$. The convergence result (4.7) for Class A therefore follows from the error estimate (4.13).

Next, we obtain the convergence result (4.8) for Class B. Since $\max_{1\le n_k\le N_k} |\alpha_{n_k}| \le s < 1$ (cf. definition (4.1)), (4.9) reduces to

$$\|f^{(r)}_{\Delta_k} - S^{(r)}_{\Delta_k}\|_\infty \le \frac{\|\Delta_k\|^{2-r} s}{|I|^{2-r} - \|\Delta_k\|^{2-r} s}(\|S^{(r)}_\Delta\|_\infty + K_r), \quad r = 0, 1, 2. \tag{4.14}$$

The inequalities (4.10), (4.11), and (4.14) together with (4.12) give

$$\begin{aligned}\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty \le{}& \|\Delta_k\|^{2-r}\Bigg\{5\omega(\Phi^{(r)}; \|\Delta_k\|) \\ &+ \frac{(\|\Phi^{(r)}\|_\infty + 5\|\Delta_k\|^{2-r}\omega(\Phi^{(r)}; \|\Delta_k\|) + K_r)s}{|I|^{2-r} - \|\Delta_k\|^{2-r} s}\Bigg\}.\end{aligned} \tag{4.15}$$

The convergence result (4.8) for Class B now follows from (4.15). □

The convergence of a suitable sequence of cubic spline FIFs to the function Φ in $C^3[x_0, x_N]$ generating the interpolation data is given by the following theorem.

THEOREM 4.3. *Let $\Phi \in C^3[x_0, x_N]$ and cubic spline FIFs $f_{\Delta_k}(x)$ satisfy boundary conditions of Type-*I*, Type-*II*, or Type-*III*(periodic) on a sequence of meshes $\{\Delta_k\}$ on $[x_0, x_N]$ with $\lim_{k\to\infty} \|\Delta_k\| = 0$ and $\frac{\|\Delta_k\|}{\min_{1\le n\le N_k} h_{k,n}} \le \beta < \infty$. If $\{\Delta_k\}$ is in Class A, then*

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \circ(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.16}$$

Further, if $\{\Delta_k\}$ is in Class B, then

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \bigcirc(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.17}$$

Proof. It is known that [19, 21], for $r = 0, 1, 2$,

$$\|\Phi^{(r)} - S^{(r)}_{\Delta_k}\|_\infty \le \frac{5}{3}\|\Delta_k\|^{3-r}(3 + \bar{K})\omega(\Phi^{(3)}; \|\Delta_k\|), \tag{4.18}$$

where $\bar{K} = 8\beta^2(1 + 2\beta)(1 + 3\beta)$.

Now, (4.9) and (4.18) together with (4.12) give

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty \le \|\Delta_k\|^{2-r}\left\{\frac{5}{3}\|\Delta_k\|(3 + \bar{K})\omega(\Phi^{(3)}; \|\Delta_k\|) + \frac{(\|\Phi^{(r)}\|_\infty + \frac{5}{3}\|\Delta_k\|(3 + \bar{K})\omega(\Phi^{(3)}; \|\Delta_k\|) + K_r)\max_{1\le n\le N_k} |\alpha_{k,n}|}{|I|^{2-r} - \|\Delta_k\|^{2-r}\max_{1\le n\le N_k} |\alpha_{k,n}|}\right\}.$$

For the sequence of meshes in Class A or Class B, the relations (4.16)–(4.17) now follow immediately from the above error estimate. □

The convergence of a suitable sequence of cubic spline FIFs to the function Φ in $C^4[x_0, x_N]$ generating the interpolation data is described by the following theorem.

THEOREM 4.4. *Let $\Phi \in C^4[x_0, x_N]$ and cubic spline FIFs $f_{\Delta_k}(x)$ satisfy boundary conditions of Type-*I *or Type-*II *on a sequence of meshes $\{\Delta_k\}$ on $[x_0, x_N]$ with $\lim_{k\to\infty} \|\Delta_k\| = 0$ and $\frac{\|\Delta_k\|}{\min_{1\le n\le N_k} h_{k,n}} \le \beta < \infty$. If $\{\Delta_k\}$ is in Class A, then*

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \circ(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.19}$$

Further, if $\{\Delta_k\}$ is in Class B, then

$$\|\Phi^{(r)} - f^{(r)}_{\Delta_k}\|_\infty = \bigcirc(\|\Delta_k\|^{2-r}), \quad r = 0, 1, 2. \tag{4.20}$$

Proof. It is known that [22]

$$\|\Phi^{(r)} - S^{(r)}_{\Delta_k}\|_\infty \le L_r\|\Phi^{(4)}\|_\infty\|\Delta_k\|^{4-r}, \quad r = 0, 1, 2, 3, \tag{4.21}$$

where $L_0 = 5/384$, $L_1 = 1/24$, $L_2 = 3/8$, and $L_3 = (\beta + \beta^{-1})/2$. The inequalities (4.9) and (4.21) together with (4.12) give the error estimate

(4.22)

$$\|\Phi^{(r)} - f_{\Delta_k}^{(r)}\|_\infty \leq \|\Delta_k\|^{2-r} \Bigg\{ L_r \|\Phi^{(4)}\|_\infty \|\Delta_k\|^2 + \frac{(\|\Phi^{(r)}\|_\infty + L_r \|\Phi^{(4)}\|_\infty \|\Delta_k\|^{4-r} + K_r) \max_{1\leq n\leq N_k} |\alpha_{k,n}|}{|I|^{2-r} - \|\Delta_k\|^{2-r} \max_{1\leq n\leq N_k} |\alpha_{k,n}|} \Bigg\}.$$

The convergence results (4.19) and (4.20) now follow from (4.22). □

Remarks. 1. Theorem 4.4 generalizes a result of Navascués and Sebastián [23] proved only for uniform meshes with fixed vertical scaling factors.

2. If $\Phi^{(2)}$ satisfies a Hölder condition of order τ, $0 < \tau \leq 1$, Theorem 4.2 gives that, for $r = 0, 1, 2$, $\|\Phi^{(r)} - f_{\Delta_k}^{(r)}\|_\infty = \circ(\|\Delta_k\|^{2-r})$ if Δ_k is in Class A and $\|\Phi^{(r)} - f_{\Delta_k}^{(r)}\|_\infty = \bigcirc(\|\Delta_k\|^{2-r})$ if Δ_k is in Class B. This provides an analogue of the corresponding result for classical cubic splines [19, Theorem 2.3.3]. The same estimates on $\|\Phi^{(r)} - f_{\Delta_k}^{(r)}\|_\infty$ follow from Theorem 4.3 or Theorem 4.4 if $\Phi^{(3)}$ or $\Phi^{(4)}$, respectively, satisfies the Hölder condition of order τ, $0 < \tau \leq 1$.

3. It follows from Theorems 4.2–4.4 that the sequence of cubic spline FIFs $f_{\Delta_k}^{(r)}$ converges uniformly to $\Phi^{(r)}$ for $r = 0, 1$ and if Δ_k is in Class A, $f_{\Delta_k}^{(2)}(x)$ converges uniformly to $\Phi^{(2)}(x)$, since, for $r = 2$, the vertical scaling factors can be chosen suitably depending on the mesh norm.

5. Examples of cubic spline FIFs. Using the IFS given by (3.14), we first computationally generate examples of cubic spline FIFs with the set of vertical scaling factors as $\alpha_n = 0.8$, $n = 1, 2, 3$, and the interpolation data as $\{(0,0), (\frac{2}{5}, 1), (\frac{3}{4}, -1), (1, 2)\}$ for the nonperiodic splines and as $\{(0,0), (\frac{2}{5}, 1), (\frac{3}{4}, -1), (1, 0)\}$ for the periodic splines. These interpolation data give $L_1(x) = \frac{2}{5}x$, $L_2(x) = \frac{7}{20}x + \frac{2}{5}$, and $L_3(x) = \frac{1}{4}x + \frac{3}{4}$ in the IFS (3.14) for all our examples of cubic spline FIFs. For constructing an example of the cubic spline FIF with a boundary condition of Type-I, we choose $f'_\Delta(x_0) = 2$ and $f'_\Delta(x_N) = 5$. With these choices, the system of equations (3.11) is solved to get the values of moments M_0, M_1, M_2, M_3 (Table 1). These moments are now used in (3.15) for the construction of $F_n(x, y)$ (Table 2). Iterations of this IFS code generate the desired cubic spline FIF (Figure 2(a)) with a boundary condition of Type-I. Again, to construct an example of the cubic spline FIF with a boundary condition of Type-II, we choose $M_0 = 2$ and $M_3 = 5$. The values of M_1 and M_2 (Table 1) are computed by solving the system (3.12). Using (3.15), the coefficients of $F_n(x, y)$, $n = 1, 2, 3$, are computed (Table 2). The iterations of the resulting IFS code generate the cubic spline FIF (Figure 2(c)) with a boundary condition of Type-II. An example of the cubic spline FIF with a boundary condition of Type-III (periodic), i.e., $f'_\Delta(x_0) = f'_\Delta(x_3)$ is constructed and $M_0 = M_3$. The values of moments M_0, M_1, M_2, M_3 (Table 1) are computed by solving the system (3.13). The associated IFS code for the periodic cubic spline is obtained from the resulting (3.14). The desired example of the periodic cubic spline FIF (Figure 2(e)) is generated through iterations of this IFS. Similarly, with a 2nd set of vertical scaling factors as $\alpha_1 = \alpha_3 = -0.9$ and $\alpha_2 = 0.9$, the examples of cubic spline FIFs (Figure 2(b), (d), (f)) with boundary conditions of Type-I, Type-II, and Type-III are generated. We note that cubic spline FIFs given by Figure 2(a)–(b) have completely different shapes though they are generated with the same boundary conditions of Type-I, whereas the same boundary conditions give

TABLE 1
Data for cubic spline FIFs with different boundary conditions.

Figures	α_1	α_2	α_3	$f'_\Delta(x_0)$	M_0	M_1	M_2	M_3	$f'_\Delta(x_3)$
2(a)	0.8	0.8	0.8	2	−77.8748	−331.3818	−59.6840	−462.5397	5
2(b)	−0.9	0.9	−0.9	2	26.2835	−31.5521	81.3627	−67.5836	5
2(c)	0.8	0.8	0.8	9.4232	2	−65.0164	93.8441	5	19.4085
2(d)	−0.9	0.9	−0.9	3.4589	2	−34.3620	79.1610	5	13.5633
2(e)	0.8	0.8	0.8	8.1939	5.4523	−43.8970	63.5040	5.4523	8.1939
2(f)	−0.9	0.9	−0.9	4.2258	−3.7995	−30.8481	46.0958	−3.7995	4.2258
2(g)	0.8	0.8	0.8	2	5	−219.5278	25.0565	−281.2847	9.7366
2(h)	−0.9	0.9	−0.9	2	5	−38.5155	79.6443	30.0172	16.9051
2(i)	0.8	0.8	0.8	−49.6	1066.0	111.1	610.8	5	2
2(j)	−0.9	0.9	−0.9	61.5792	−334.8459	59.9983	42.3613	5	2
2(k)	0.8	0.8	0.8	2	129.4060	−44.2624	155.5007	5	17.3112
2(l)	−0.9	0.9	−0.9	2	11.6444	−36.7487	79.9084	5	13.8840
2(m)	0.8	0.8	0.8	−1.4427	2	−297.1132	−11.3357	−423.6607	5
2(n)	−0.9	0.9	−0.9	5.9477	2	−25.6023	78.7498	−61.1447	5
2(o)	0.8	0.8	0.8	9.7621	−14.1432	−79.5646	80.6184	−16.9354	18.9354
2(p)	−0.9	0.9	−0.9	5.7448	−8.1171	−29.6265	77.9665	−9.3573	11.3573

just one interpolating classical cubic spline. Thus, in our approach, an added flexibility is offered to an experimenter depending upon the need of a problem for the choice of a suitable cubic spline FIF. Similarly, Figure 2(c)–(d) gives a comparison of shape and nature of cubic spline FIFs with a boundary condition of Type-II and Figure 2(e)–(f) gives such a comparison for periodic cubic spline FIFs to see the effect of vertical scaling factors on their shapes.

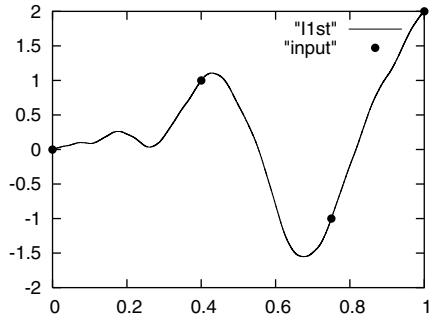
For construction of IFS for cubic spline FIFs (Figure 2(g) and 2(i)) with boundary conditions of Type-IV with first set of vertical scaling factors as $\alpha_n = 0.8, n = 1, 2, 3$, we choose $f'_\Delta(x_0) = 2$, $M_0 = 5$, and $f'_\Delta(x_3) = 2$, $M_3 = 5$, respectively. The examples of cubic spline FIFs (Figure 2(k) and 2(m)) with boundary conditions of Type-V are constructed with $\alpha_n = 0.8, n = 1, 2, 3$, by choosing $f'_\Delta(x_0) = 2$, $M_3 = 5$, and $M_0 = 2$, $f'_\Delta(x_3) = 5$, respectively. Finally, for constructing the cubic spline FIF (Figure 2(o)) with a boundary condition of Type-VI, the associated IFS is generated by choosing $\alpha_n = 0.8, n = 1, 2, 3$, and $f'_\Delta(x_0), M_0, M_3$, and $f'_\Delta(x_3)$ are chosen such that $3f'_\Delta(x_0) + 2M_0 = 1$ and $f'_\Delta(x_3) + M_3 = 2$. The examples of cubic spline FIFs (Figure 5.1(h), (j), (l), (n), (p)) with boundary conditions of Type-IV, V, or VI are analogously constructed by computing the associated IFS with $\alpha_1 = \alpha_3 = -0.9$ and $\alpha_2 = 0.9$. The effect of vertical scaling factors on the shape and nature of cubic spline FIFs with boundary conditions of Type-IV, V, or VI is demonstrated in Figure 2(g)–(p). Thus, infinitely many cubic spline FIFs with different shapes can be generated by varying scaling factor sets for any prescribed boundary conditions. This gives a vast flexibility in the choice of cubic spline FIF according to the need of the problem.

A normal-size font entry in Table 1 is for the value assumed for a parameter in a particular example. An entry in script-size font in Table 1 is for the value of the parameters that are computed by using (3.10). The entries for the coefficients of $F_n(x, y)$ in Table 2 are computed by using (3.15). All the entries in these tables are rounded off up to four decimal places.

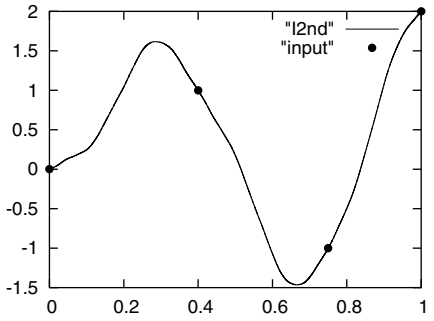
6. Conclusion. A new method is introduced in the present work for the construction of C^r-FIFs so that the complex algebraic method in [1] for construction of C^r-FIFs using complicated matrices is no longer needed. Our method allows admissibility of any kind of boundary conditions while the boundary conditions in [1] are restricted to only at the initial endpoint x_0 of the interval $[x_0, x_N]$. In our approach, r equations involving the spline values or the values of its derivatives at the boundary points are chosen such that the resulting $(r + 2)N + 2r$ equations are linearly independent. This results in generation of a unique C^r-FIF for a prescribed data and a suitable set of vertical scaling factors. This answers a query of Barnsley and Harrington [1, p. 33], regarding uniqueness of the C^r-FIF for a suitable set of vertical scaling

TABLE 2
$\mathbf{F_n(x, y)}$ *for cubic spline FIFs with different boundary conditions.*

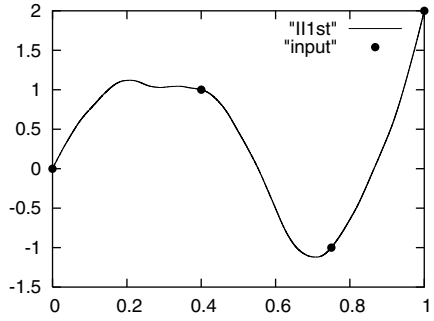
Figures	$\mathbf{F_1(x, y)}$	$\mathbf{F_2(x, y)}$	$\mathbf{F_3(x, y)}$
2(a)	$0.128y + 1.446x^3 - 1.246x^2 + 0.544x$	$0.098y + 11.83x^3 - 16.4813x^2 + 2.4552x + 1$	$0.05y - 0.9909x^3 + 0.0817x^2 + 3.8091x - 1$
2(b)	$0.128y - 3.7951x^3 + 3.9951x^2 + 1.088x$	$0.098y + 4.0302x^3 - 3.3814x^2 - 2.8692x + 1$	$0.05y - 2.4315x^3 + 3.2818x^2 + 2.2622x - 1$
2(c)	$0.128y - 2.1981x^3 + 0.1508x^2 + 2.7913x$	$0.098y + 3.0803x^3 - 4.444x^2 - 0.8323x + 1$	$0.05y - 0.8586x^3 + 2.697x^2 + 1.0615x - 1$
2(d)	$0.128y - 0.7661x^3 + 0.5258x^2 + 1.5283x$	$0.098y + 2.1321x^3 - 2.2953x^2 - 2.0572x + 1$	$0.05y - 0.5886x^3 + 2.5711x^2 + 1.13x - 1$
2(e)	$0.128y - 1.3306x^3 + 0.1311x^2 + 2.1995x$	$0.098y + 2.2375x^3 - 3.0902x^2 - 1.1474x + 1$	$0.05y - 0.5819x^3 + 1.7797x^2 - 0.1978x - 1$
2(f)	$0.128y - 1.1279x^3 + 0.6423x^2 + 1.4856x$	$0.098y + 1.7184x^3 - 2.1224x^2 - 1.596x + 1$	$0.05y - 0.595x^3 + 1.5593x^2 + 0.0356x - 1$
2(g)	$0.128y + 1.6313x^3 - 3.5124x^2 + 2.6252x$	$0.098y - 2.0316x^3 + 7.0014x^2 - 7.1903x + 1$	$0.05y + 1.1209x^3 - 3.302x^2 + 5.081x - 1$
2(h)	$0.128y + 4.481x^3 - 5.8543x^2 + 2.6613x$	$0.098y - 2.0316x^3 + 7.0014x^2 - 7.1903x + 1$	$0.05y + 0.383x^3 - 0.1452x^2 + 2.8747x - 1$
2(i)	$0.128y + 15.6608x^3 - 0.793x^2 - 14.1238x$	$0.098y - 30.2834x^3 + 67.7226x^2 - 39.6352x + 1$	$0.05y - 0.2974x^3 + 4.7097x^2 - 1.5124x - 1$
2(j)	$0.128y - 11.0326x^3 + 9.36x^2 + 2.9605x$	$0.098y + 8.4145x^3 - 23.9039x^2 + 13.2688x + 1$	$0.05y - 0.3639x^3 + 3.6069x^2 - 0.1305x - 1$
2(k)	$0.128y - 2.2398x^3 + 2.0705x^2 + 0.9133x$	$0.098y + 5.9094x^3 - 9.052x^2 + 0.9466x + 1$	$0.05y - 0.5053x^3 + 1.6242x^2 + 1.7811x - 1$
2(l)	$0.128y - 1.2367x^3 + 1.7699x^2 + 0.7548x$	$0.098y + 2.3406x^3 - 2.8928x^2 - 1.6683x + 1$	$0.05y - 0.6668x^3 + 2.8246x^2 + 0.9546x - 1$
2(m)	$0.128y + 1.1228x^3 - 0.0231x^2 - 0.3557x$	$0.098y + 12.7309x^3 - 18.1275x^2 + 3.2006x + 1$	$0.05y - 0.7766x^3 - 0.3182x^2 + 3.9947x - 1$
2(n)	$0.128y - 2.4515x^3 + 0.904x^2 + 2.8355x$	$0.098y + 3.3633x^3 - 1.896x^2 - 3.6878x + 1$	$0.05y - 2.0862x^3 + 2.6282x^2 + 2.5705x - 1$
2(o)	$0.128y - 2.1491x^3 + 0.1562x^2 + 2.7369x$	$0.098y - 2.493x^3 - 1.3446x^2 + 1.6416x + 1$	$0.05y + 1.0781x^3 - 2.7304x^2 + 4.5523x - 1$
2(p)	$0.128y + 1.3637x^3 + 0.8732x^2 - 0.9489x$	$0.098y - 1.7662x^3 - 0.8139x^2 + 0.3596x + 1$	$0.05y + 1.7978x^3 - 0.7643x^2 + 2.0789x - 1$



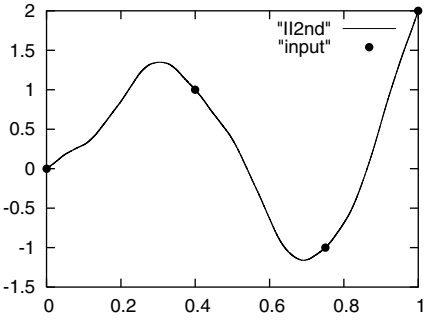
(a) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $f'_\Delta(x_0) = 2$, and $f'_\Delta(x_3) = 5$.



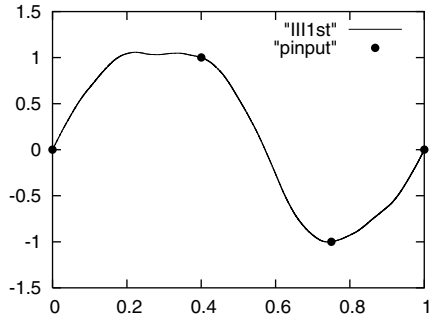
(b) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $f'_\Delta(x_0) = 2$, and $f'_\Delta(x_3) = 5$.



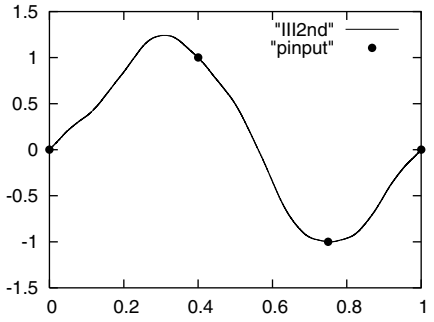
(c) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $M_0 = 2$, and $M_3 = 5$.



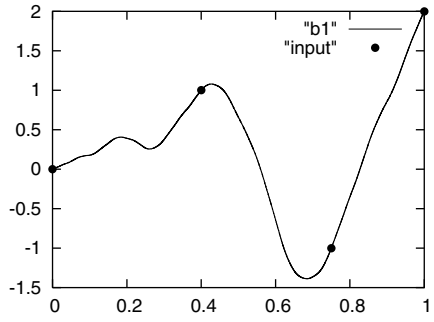
(d) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $M_0 = 2$, and $M_3 = 5$.



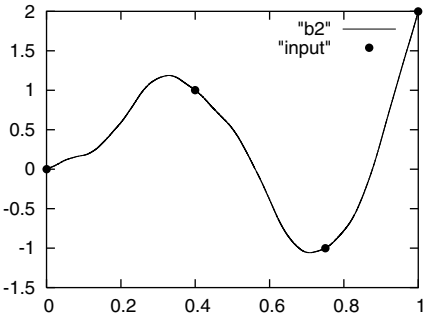
(e) Periodic cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$.



(f) Periodic cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9, \alpha_2 = 0.9$.

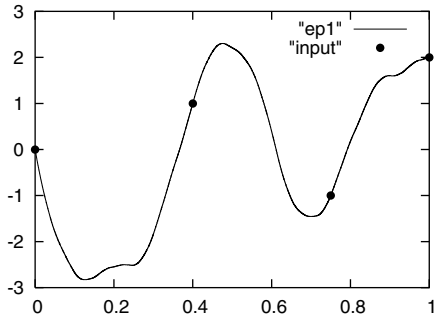


(g) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $f'_\Delta(x_0) = 2$, and $M_0 = 5$.

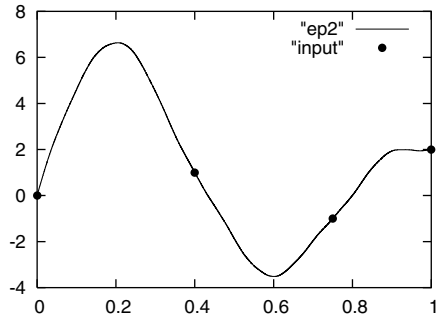


(h) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $f'_\Delta(x_0) = 2$, and $M_0 = 5$.

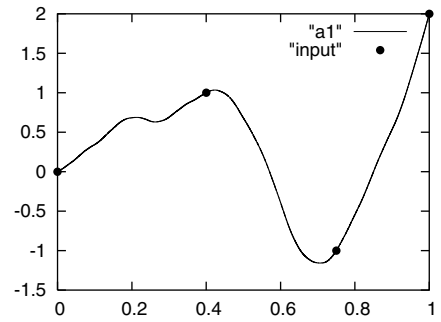
FIG. 2. *Cubic spline FIFs with different boundary conditions.*



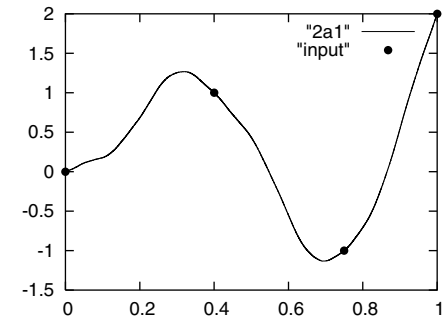
(i) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $f'_\Delta(x_3) = 2$, and $M_3 = 5$.



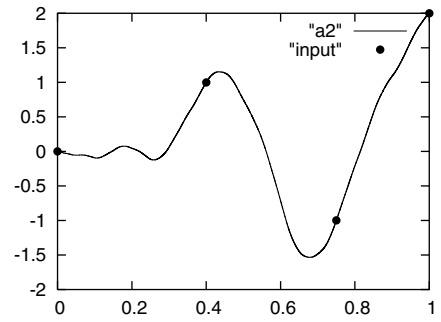
(j) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $f'_\Delta(x_3) = 2$, and $M_3 = 5$.



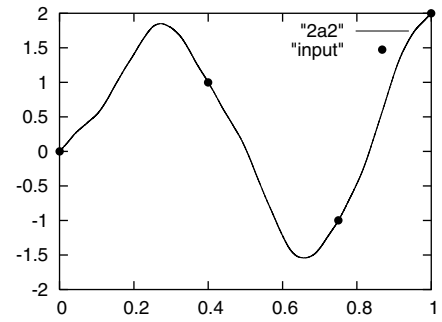
(k) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $f'_\Delta(x_0) = 2$, and $M_3 = 5$.



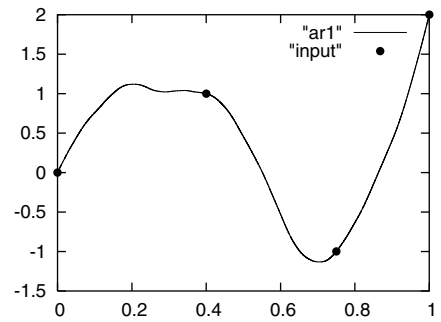
(l) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $f'_\Delta(x_0) = 2$, and $M_3 = 5$.



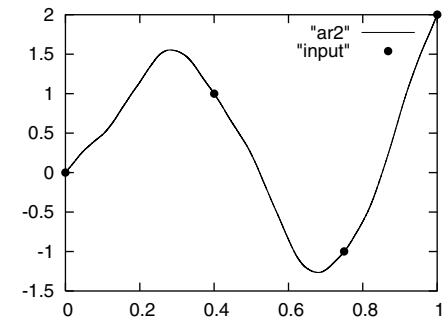
(m) Cubic spline FIF with $\alpha_n = 0.8$, $n = 1, 2, 3$, $f'_\Delta(x_3) = 2$, and $M_0 = 5$.



(n) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9$, $f'_\Delta(x_3) = 2$, and $M_0 = 5$.



(o) Cubic spline FIF with $\alpha_n = 0.8, n = 1, 2, 3$, $3f'_\Delta(x_0) + 2M_0 = 1$, and $f'_\Delta(x_3) + M_3 = 2$.



(p) Cubic spline FIF with $\alpha_1 = \alpha_3 = -0.9$, $\alpha_2 = 0.9, 3f'_\Delta(x_0) + 2M_0 = 1, and\, f'_\Delta(x_3) + M_3 = 2$.

FIG. 2. *Cont.*

factors. The construction of cubic spline FIFs, using the moments $M_n = f_\Delta^{''}(x_n)$, is initiated for the first time in the present work, resulting in a satisfactory generalization of the classical cubic spline theory.

For the data generating function $\Phi \in C^r[x_0, x_n]$, $r = 2, 3$, or 4, it is proved that (cf. Theorems 4.2–4.4), the sequence of cubic spline FIFs $\{f_{\Delta_k}\}$ converges to Φ with arbitrary degree of accuracy for the sequences of meshes in Class A or Class B for boundary conditions of Type-I, Type-II, or Type-III. Our convergence results in section 4 are obtained with more general conditions than those in [23] wherein only uniform meshes are considered in the case $\Phi \in C^{(4)}[x_0, x_n]$. The upper bounds on error in approximation of Φ and its derivatives by cubic spline FIFs f_Δ and its derivatives, respectively, with different boundary conditions are also obtained by results in section 4. As a consequence of our results, the data generating function Φ that satisfies $\Phi^{(2)} \in Lip\ \tau$, $0 < \tau < 1$, can be approximated satisfactorily by a fractal function f_Δ by choosing vertical scaling factors suitably such that $f_\Delta^{(2)} \in Lip\ \tau$.

The vertical scaling factors α_n are important parameters in the construction of C^r-FIFs or cubic spline FIFs. For given boundary conditions, in our approach an infinite number of C^r-FIFs or cubic spline FIFs can be constructed interpolating the same data by choosing different sets of vertical scaling factors. Thus, according to the need of an experiment for simulating objects with smooth geometrical shapes, a large flexibility in the choice of a suitable interpolating smooth fractal spline is offered by our approach. As in the case of vast applications of classical splines in CAM, CAGD, and other mathematical, engineering applications [12, 13, 14, 15], it is felt that cubic spline FIFs generated in the present work can find rich applications in some of these areas. Since the cubic spline FIF is invariant in all scales, it can also be applied to image compression and zooming problems in image processing. Further, as classical cubic splines are a special case of cubic spline FIFs, it should be possible to use cubic spline FIFs for mathematical and engineering problems where the classical spline interpolation approach does not work satisfactorily.

Acknowledgment. The authors are thankful to Dr. Ian Sloan for his valuable suggestions.

REFERENCES

[1] M. F. Barnsley and A. N. Harrington, *The calculus of fractal interpolation functions*, J. Approx. Theory, 57 (1989), pp. 14–34.
[2] B. B. Mandelbrot, *Fractals: Form, Chance and Dimension*, W.H. Freeman, San Francisco, 1977.
[3] M. F. Barnsley, *Fractals Everywhere*, Academic Press, Boston, 1988.
[4] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley, Chichester, UK, 1990.
[5] H. M. Hastings and G. Sugihara, *Fractals: A User's Guide for the Natural Sciences*, Oxford University Press, New York, 1993.
[6] J. Hutchinson, *Fractal and self-similarity*, Indiana Univ. Math. J., 30 (1981), pp. 713–747.
[7] M. F. Barnsley, *Fractal functions and interpolations*, Constr. Approx., 2 (1986), pp. 303–329.
[8] M. F. Barnsley and L. P. Hurd, *Fractal Image Compression*, AK Peters, Wellesley, UK, 1992.
[9] J. L. Véhel, K. Daoudi, and E. Lutton, *Fractal modeling of speech signals*, Fractals, 2 (1994), pp. 379–382.
[10] D. S. Mazel and M. H. Hayes, *Using iterated function systems to model discrete sequences*, IEEE Trans. Signal Process, 40 (1992), pp. 1724–1734.
[11] P. R. Massopust, *Fractal Functions, Fractal Surfaces, and Wavelets*, Academic Press, San Diego, 1994.
[12] M. P. Groover and E. W. Zimmers Jr., *CAD/CAM: Computer-Aided Design and Manufacturing*, Pearson Education, Upper Saddle River, NJ, 1997.

[13] G. Farin, *Curves and Surfaces for Computer Aided Geometric Design: A Practical Guide*, Academic Press, San Diego, 1990.
[14] G. D. Knott, *Interpolating Cubic Splines*, Birkhäuser, Boston, 2000.
[15] G. Micula and S. Micula, *Handbook of Splines*, Kluwer, Dordrecht, The Netherlands, 1999.
[16] M. F. Barnsley, J. H. Elton, D. Hardin, and P. R. Massopust, *Hidden variable fractal interpolation functions*, SIAM J. Math. Anal., 20 (1989), pp. 1218–1242.
[17] M. F. Barnsley, J. H. Elton, and D. Hardin, *Recurrent iterated function systems*, Constr. Approx., 5 (1989), pp. 3–31.
[18] A. K. B. Chand, *A Study on Coalescence and Spline Fractal Interpolation Functions*, Ph.D. thesis, Indian Institute of Technology, Kanpur, India, 2004.
[19] J. Ahlberg, E. Nilson, and J. Walsh, *The Theory of Splines and Their Applications*, Academic Press, New York, 1967.
[20] A. Sharma and A. Meir, *Degree of approximation of spline interpolation*, J. Math. Mech., 15 (1966), pp. 759–767.
[21] G. Birkhoff and C. De Boor, *Error bounds for spline interpolation,* J. Math. Mech., 13 (1964), pp. 827–835.
[22] C. A. Hall and W. W. Meyer, *Optimal error bounds for cubic spline interpolation*, J. Approximation Theory, 16 (1976), pp. 105–122.
[23] M. A. Navascués and M. V. Sebastián, *Some results of convergence of cubic spline fractal interpolation functions*, Fractals, 11 (2003), pp. 1–7.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 677–698

DISCRETE MAXIMAL L_p REGULARITY FOR FINITE ELEMENT OPERATORS*

MATTHIAS GEISSERT†

Abstract. Let $\{A_h\}_{h>0}$ be a family of elliptic finite element operators. Let $I = [0,T]$ and consider the problem $u_h'(t) - A_h u_h(t) = f_h(t)$, $t \in I$, $u_h(0) = 0$. In this paper, we show that for $1 < p < \infty$ the solution of that problem satisfies the estimate

$$\|u_h'\|_{L_p(I;L_p(\Omega))} + \|A_h u_h\|_{L_p(I;L_p(\Omega))} \le C\|f_h\|_{L_p(I;L_p(\Omega))},$$

where C is independent of the parameter h and f_h. In this case $\{A_h\}_{h>0}$ is said to have *discrete* maximal L_p regularity.

Key words. finite elements, maximal regularity

AMS subject classifications. 65M60, 34G10

DOI. 10.1137/040616553

1. Introduction. For $T > 0$, set $I = [0,T]$ and consider the Galerkin finite element method for the approximate solution of

$$\begin{aligned} u'(t) - \Delta^D u(t) &= f(t), \quad t \in I, \\ u(0) &= 0, \end{aligned} \tag{1}$$

where Δ^D is the Dirichlet–Laplacian in $L_2(\Omega)$. Here, $\Omega \subset \mathbb{R}^N$ is a bounded, convex domain with C^2 boundary. Then the corresponding problem for the discrete Dirichlet–Laplacian Δ_h^D acting on some finite element space S_h, i.e., continuous, piecewise linear functions defined on a family of quasi-uniform triangulations with parameter h, reads as:

$$\begin{aligned} u_h'(t) - \Delta_h^D u_h(t) &= f_h(t), \quad t \in I, \\ u_h(0) &= 0. \end{aligned} \tag{2}$$

The solution u_h of (2) is given by the variation of constants formula, i.e.,

$$u_h(t) := \int_0^t T_h(t-s) f_h(s)\, \mathrm{d}s, \tag{3}$$

where $(T_h(t))_{t\ge 0} := (\mathrm{e}^{t\Delta_h^D})_{t\ge 0}$ denotes the semigroup generated by Δ_h^D. In [Tho97] stability estimates for u_h together with stability estimates for the corresponding Ritz projections are used to derive error estimates for $u - u_h$ in $L_2(\Omega)$ and $L_\infty(\Omega)$. In order to prove this stability estimates for u_h representation (3) and uniform estimates in h for $(T_h(t))_{t\ge 0}$, i.e.,

$$\|T_h(t)\|_{\mathcal{L}(S_{h,q})} \le C, \quad \|t\Delta_h^D T_h(t)\|_{\mathcal{L}(S_{h,q})} \le C, \quad t \in I,\ h > 0, \tag{4}$$

*Received by the editors October 7, 2004; accepted for publication (in revised form) August 24, 2005; published electronically March 31, 2006. This work was supported by the DFG-Graduiertenkolleg 853.

http://www.siam.org/journals/sinum/44-2/61655.html

†Technische Universität Darmstadt, Fachbereich Mathematik, Schlossgartenstr. 7, D-64289 Darmstadt, Germany (geissert@mathematik.tu-darmstadt.de).

were very useful. Here, $S_{h,q}$ denotes the space S_h equipped with the $L_q(\Omega)$-norm and $1 \le q \le \infty$.

In this paper, we consider the property of maximal L_p regularity for finite element operators. Although our approach is not at all restricted to the Dirichlet–Laplacian (see Definition 2.1), for notational reasons we restrict our considerations for the time being to this case. Since Δ_h^D is a bounded operator on $S_{h,q}$, it immediately follows from (3) that Δ_h^D has *maximal L_p regularity on $S_{h,q}$*, i.e., for $h > 0$ there exists $C_h > 0$ such that

$$\|u_h'\|_{L_p(I;S_{h,q})} + \|\Delta_h^D u_h\|_{L_p(I;S_{h,q})} \le C_h \|f_h\|_{L_p(I;S_{h,q})}, \quad f_h \in L_p(I;S_{h,q}), \tag{5}$$

where u_h denotes the unique solution of (2) and $\|u_h\|^q_{L_p(I;S_{h,q})} := \int_0^T \|u_h(t)\|^p_{S_{h,q}}\,\mathrm{d}t$. If the family $\{C_h\}_{h>0}$ is uniformly bounded, we say that $\{\Delta_h^D\}_{h>0}$ has *discrete maximal L_p regularity on $S_{h,q}$*. In terms of numerical analysis, discrete maximal L_p regularity means that the numerical scheme is stable. Tracing constants in [DHP03, Remark 3.2(3)], [DHP03, Theorem 4.4], and [HP97, Proposition 2.4], it follows that $\{\Delta_h\}_{h>0}$ has discrete maximal L_p regularity on $S_{h,q}$ for $1 < p < \infty$ and $q = 2$. In this paper, we show that various families of finite element operators (see Definition 2.1) have discrete maximal L_p regularity on $S_{h,q}$ for $1 < p = q < \infty$.

In [Gei04] it is shown that discrete maximal L_p regularity on $S_{h,q}$ and the stability of the Ritz projections lead to an elegant proof for error estimates of the form

$$\|u - u_h\|_{L_p(I;L_q(\Omega))} \le Ch^2 \|f\|_{L_p(I;L_q(\Omega))}, \quad f \in L_p(I;L_q(\Omega)),\ h > 0, \tag{6}$$

where u_h denotes the solution of (2) with $f_h = P_h f$. Here, P_h denotes the orthogonal projection from $L_2(\Omega)$ onto S_h. The approach therein also allows one to treat initial values in certain interpolation spaces. Furthermore, applications of discrete maximal L_p regularity on $S_{h,q}$ to various nonlinear problems are considered. More precisely, if the solution of the linearized problem satisfies (5), there exists a strong solution to the nonlinear problem on a small interval. Since the length of the interval is reciprocally proportional to the constant C_h appearing in (5), uniform estimates for C_h, i.e., discrete maximal L_p regularity, are extremely useful. Finally, error estimates of the type $\|u - u_h\|_{L_p(I;L_q(\Omega))} \le Ch^s$ for such problems are proven. Here, s depends on the nonlinearity.

Whereas discrete maximal L_p regularity has not been considered previously, quite a few authors have dealt with estimates of the form (4) for $q = \infty$. For instance, in [CLT94] it is shown that (4) holds for elliptic operators subject to Dirichlet boundary conditions in one dimension. In fact, the authors gave resolvent estimates which imply (4). A completely different approach was used in [STW98], where (4) is shown for self-adjoint operators subject to Neumann boundary conditions in space dimension $N \ge 2$ by establishing bounds on the semigroup $(T_h(t))_{t\ge0}$ itself. A similar approach is used in [TW00] in order to show (4) for the Dirichlet–Laplacian in space dimension $N \ge 2$. This result is also proven in [BTW03] with a much simpler proof. By interpolation of the results for $q = 2$, which is trivial, and $q = \infty$, and by duality arguments (see [STW98, Corollary 2.2]), we obtain (4) for $1 \le q \le \infty$.

By [Dor93, Theorem 2.2], it follows that discrete maximal L_p regularity on $S_{h,q}$ implies (4). But the converse is not true except for $q = 2$. In this sense the present paper is a generalization and extension of the results mentioned above. However, observe that we cannot expect discrete maximal L_p regularity on $S_{h,\infty}$, since even the Dirichlet–Laplacian does not have maximal L_p regularity on $L_\infty(\Omega)$. Nevertheless we show that (5) holds with $C_h = C|\log(h)|$ for $q = \infty$.

2. Preliminaries. Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$ be a bounded, convex domain with C^2 boundary. For a parameter $h > 0$ we introduce finite element operators A_h^D or A_h^N subject to Dirichlet or Neumann boundary conditions. To do this we construct finite element spaces $S_h^D \subset H_0^1(\Omega)$ and $S_h^N \subset H^1(\Omega)$.

We start with S_h^D. Let $\mathcal{T}_h^D = \{\tau_1^h, \dots, \tau_n^h\}$ be a finite set of disjoint, face-to-face, open simplexes of diameter less than h such that the set $\Omega_h^D := \overset{\circ}{\cup_{i=1}^n \tau_i}$ is convex and any vertex belonging to the boundary of Ω_h^D belongs to $\partial\Omega$. In the following, we always assume that $(\mathcal{T}_h^D)_{h>0}$ is quasi-uniform. We set $S_h^D := \{u \in C(\overline{\Omega}) : u|_\tau \text{ is linear for all } \tau \in \mathcal{T}_h^D,\ u|_{\overline{\Omega}\setminus\Omega_h} = 0\}$.

Next, we construct S_h^N. Contrary to above we construct a triangulation that fits Ω exactly. In order to do so, we introduce *wedge shaped pieces*, i.e., $C \cap \Omega$ where C is an open infinite cone with vertex $e \in \Omega$. For wedge-shape pieces, we call e and the elements of the intersection of the edges of C and $\partial\Omega$ *vertices*. Clearly all vertices but e meet the boundary $\partial\Omega$. Let $\mathcal{T}_h^N = \{\tau_1^h, \dots, \tau_n^h\}$ be a finite set of disjoint, face-to-face, open simplexes or wedge-shaped pieces of diameter less than h such that $\Omega = \Omega_h^N := \overset{\circ}{\cup_{i=1}^n \tau_i}$. Similar to above, we assume that $(\mathcal{T}_h^N)_{h>0}$ is quasi-uniform and we set $S_h^N := \{u \in C(\overline{\Omega}) : u|_\tau \text{ is linear for all } \tau \in \mathcal{T}_h^N\}$.

In what follows we write S_h instead of S_h^D or S_h^N, where $S_h \subset H$ for $H = H_0^1(\Omega)$ or $H = H^1(\Omega)$, respectively. Furthermore, we always denote the quasi-uniform family of triangulations associated to the family $(S_h)_{h>0}$ of finite element spaces by $(\mathcal{T}_h)_{h>0}$. Moreover, we write Ω_h for Ω_h^D and Ω_h^N.

Next, we introduce the finite element operators A_h associated to a sesquilinear form defined as follows: Let $H = H_0^1(\Omega)$ or $H^1(\Omega)$ and denote by $a : H \times H \to \mathbb{C}$ a sesquilinear form given by

$$a(u,v) := \sum_{i,j=1}^N \big(a_{ij}\partial_i u, \partial_j v\big)_\Omega + \sum_{i=1}^N \big(b_i \partial_i u, v\big)_\Omega + \sum_{i=1}^N \big(c_i u, \partial_i v\big)_\Omega + \big(c_0 u, v\big)_\Omega \tag{7}$$

with $a_{ij}, b_i, c_i, c_0 \in L_\infty(\Omega)$, $i,j = 1,\dots,N$. Here, $(\cdot,\cdot)_\Omega$ denotes the usual $L_2(\Omega)$ scalar product. We say that a is sectorial of angle $\varphi \in (0, \frac{\pi}{2})$ if $|\arg a(u,u)| \leq \varphi$ for $u \in H$. A form a satisfying

$$\sum_{i,j=1}^N a_{ij}(x)\xi_i\overline{\xi_j} \geq \mu|\xi|^2, \quad \text{a.a. } x \in \Omega,\ \xi \in \mathbb{C}^N, \tag{8}$$

for some $\mu > 0$ is called *elliptic*.

DEFINITION 2.1. *Let $a : H \times H \to \mathbb{C}$ be a sesquilinear form of the form (7). Assume that a is sectorial of angle $\varphi < \frac{\pi}{2}$, $a_{ij}(x) = \overline{a_{ji}(x)}$ for $x \in \Omega$, $i,j = 1,\dots,N$ and that a is elliptic.*

(a) *For $H = H_0^1(\Omega)$ let $(S_h^D)_{h>0}$ be a family of finite element spaces. We then define the finite element operators A_h^D by*

$$a(u_h, v_h) = -\big(A_h^D u_h, v_h\big)_\Omega, \quad u_h, v_h \in S_h^D.$$

(b) *For $H = H^1(\Omega)$ let $(S_h^N)_{h>0}$ be a family of finite element spaces. We then define the finite element operators A_h^N by*

$$a(u_h, v_h) = -\big(A_h^N u_h, v_h\big)_\Omega, \quad u_h, v_h \in S_h^N.$$

In the following we write A_h instead of A_h^D or A_h^N. The operators A_h are called the *finite element operators* associated to a.

3. Main result. Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$ be a bounded, convex domain with C^2 boundary and let $A : D(A) \to L_2(\Omega)$ be the operator associated to the form a, where a is as in Definition 2.1, and denote the semigroup generated by A by $(T(t))_{t\geq 0}$. The following definition states assumptions on $(T(t))_{t\geq 0}$ needed for the proof of our main theorem. Let $T > 0$ and set for a multi-index α

$$\partial_1^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_N}}{\partial x_N^{\alpha_N}} \text{ and } \partial_2^\alpha := \frac{\partial^{\alpha_1}}{\partial y_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_N}}{\partial y_N^{\alpha_N}}.$$

DEFINITION 3.1. *Let $T > 0$ and let $(T(t))_{t\geq 0}$ be a semigroup on $L_2(\Omega)$. We say that the semigroup $(T(t))_{t\geq 0}$ satisfies the kernel estimate (KE_{Max}) if $T(t)$ is an integral operator, the derivatives $\partial_1^\alpha \partial_2^\beta K$ of its kernel $K(t,\cdot,\cdot)$ exist in $\Omega \times \Omega$ for $0 < t \leq 2T$ and there exist $C, c > 0$ such that*

$$|\partial_1^\alpha \partial_2^\beta K(t,x,y)| \leq C t^{-\frac{N+|\alpha|+|\beta|}{2}} \exp\left(-c\frac{|x-y|^2}{t}\right), \quad 0 < t \leq 2T,\ x, y \in \Omega,$$

whenever $|\alpha| = 0, |\beta| \leq 2$ or $|\alpha| \leq 2, |\beta| = 0$.

Consider the problem

$$\begin{aligned} u_h'(t) - A_h u_h(t) &= f_h(t), \quad t \in I, \\ u_h(0) &= 0 \end{aligned} \tag{9}$$

for $f_h \in L_p(I; S_{h,q})$. The main result of this paper is the following theorem.

THEOREM 3.2. *Let $N \geq 2$, $2 \leq p < \infty$ and let a, $(A_h)_{h>0}$ be as in Definition 2.1. Assume that the domain $D(A)$ of the operator A associated to a satisfies $D(A) \hookrightarrow H^2(\Omega)$ and that the semigroup $(T(t))_{t\geq 0}$ on $L_2(\Omega)$ generated by A satisfies the kernel estimate (KE_{Max}). Then there exists $C > 0$ such that*

$$\|u_h'\|_{L_p(I;S_{h,p})} + \|A_h u_h\|_{L_p(I;S_{h,p})} \leq C\|f_h\|_{L_p(I;S_{h,p})}, \quad h > 0,\ f_h \in L_p(I; S_{h,p}),$$

where u_h denotes the solution of problem (9).

Before we start to prove Theorem 3.2 we state several consequences.

Remark 3.3. The symmetry of the principal part, i.e., $a_{ij}(x) = \overline{a_{ji}}(x)$ for $x \in \Omega$, $i, j = 1, \ldots, N$, is only used for the proof of Lemma 5.7.

By [ÈdI70, Theorem 3], [DHP03, Theorem 8.2], and duality, the following corollary follows easily.

COROLLARY 3.4. *Assume that $\Omega \subset \mathbb{R}^N$, $N \geq 2$, is a bounded, convex domain of class $C^{3+\gamma}$ and that the coefficients of the form a satisfy*

$$a_{ij}, c_i \in \mathrm{C}^{2+\gamma}(\overline{\Omega}) \text{ and } b_i, c_0 \in \mathrm{C}^{1+\gamma}(\overline{\Omega}), \quad i, j = 1, \ldots, N,$$

for some $0 < \gamma < 1$. Then for $1 < p < \infty$ there exists $C > 0$ such that

$$\|u_h'\|_{L_p(I;S_{h,p})} + \|A_h u_h\|_{L_p(I;S_{h,p})} \leq C\|f_h\|_{L_p(I;S_{h,p})}, \quad h > 0,\ f_h \in L_p(I; S_{h,p}),$$

where u_h denotes the solution of (9).

Here, C^γ denotes the space of all γ-Hölder continuous functions. Furthermore, the proof of Theorem 3.2 implies the following corollary.

COROLLARY 3.5. *Let $\{A_h\}_{h>0}$ be as in Theorem 3.2. Then*

$$\|u_h'\|_{L_\infty(I;S_{h,\infty})} + \|A_h u_h\|_{L_\infty(I;S_{h,\infty})} \leq C|\log(h)|\|f_h\|_{L_\infty(I;S_{h,\infty})}, \quad h > 0,\ f_h \in L_\infty(I; S_{h,\infty}),$$

where u_h denotes the solution of (9).

4. Proof of Theorem 3.2. We start this section with the idea of the proof of Theorem 3.2. For $h > 0$ set $Q_h := I \times \Omega_h$. Let $K_h(t,\cdot,\cdot) : \Omega \times \Omega \to \mathbb{C}$ denote the kernel of $T_h(t)$ for $t > 0$. Note that $T_h(t)$ is an integral operator for $t > 0$, since S_h is a finite dimensional space. Moreover, $K_h(t,x,\cdot) \in S_h$ and $K_h(t,\cdot,y) \in S_h$ for $t > 0$, $x, y \in \Omega$. We thus have the representation

$$\overline{\big(T_h^*(t)P_h\tilde{\delta}_{x_0}\big)(y)} = \int_\Omega K_h(t,z,y)\overline{\big(P_h\tilde{\delta}_{x_0}\big)(z)}\,\mathrm{d}z = K_h(t,x_0,y), \quad t > 0,\ x_0 \in \Omega_h,\ y \in \Omega,$$

where $\tilde{\delta}_{x_0}$ is the discrete delta function, introduced in Lemma 5.4. Furthermore, the solution u_h of (9), is given by the variation of constants formula. Indeed,

$$\begin{aligned} A_h u_h(t) &= \int_0^t A_h T_h(t-s) f_h(s)\,\mathrm{d}s = \int_0^t \int_{\Omega_h} \partial_t K_h(t-s,\cdot,y) f_h(s,y)\,\mathrm{d}y\,\mathrm{d}s \\ &=: \big(\partial_t K_h * f\big)(t,\cdot), \quad t \in I,\ f_h \in L_p(I; S_{h,p}). \end{aligned}$$

In order to estimate $\|A_h u_h\|_{L_p(I;L_p(\Omega))}$, we compare the kernels $\partial_t K_h$ with a suitable truncation $\partial_t k_{\mathrm{Tr}}$ of $\partial_t K$, where $K(t,\cdot,\cdot) : \Omega \times \Omega \to \mathbb{C}$ is the kernel of $T(t)$ for $t > 0$. The precise meaning of truncation will follow in Step 2 of the proof. For the time being, we merely stress that the truncation has to be related to h. In [TW00] and [STW98] the authors introduced the *approximate kernel* $k_a : \mathbb{R}_+ \times \Omega_h \times \Omega$ defined by

$$k_a(t,x_0,y) := k_a^h(t,x_0,y) := \overline{\big(T^*(t)\tilde{\delta}_{x_0}\big)(y)}, \quad t > 0,\ x_0 \in \Omega_h,\ y \in \Omega.$$

Our approach to the proof of Theorem 3.2 relies on the representation of $\partial_t K_h$ as

$$\partial_t K_h = \partial_t K_h - \partial_t k_a + \partial_t k_a - \partial_t k_{\mathrm{Tr}} + \partial_t k_{\mathrm{Tr}}.$$

Thomée et al. showed in [TW00] and [STW98] that there exists $C > 0$ such that

$$\|\partial_t k_a(\cdot,x_0,\cdot) - \partial_t K_h(\cdot,x_0,\cdot)\|_{L_1(Q_h)} \le C, \quad h > 0,\ x_0 \in \Omega_h,$$

for the Dirichlet–Laplacian and self-adjoint operators subject to Neumann boundary conditions. We will extend these results to elliptic operators A given by a form a as above. Furthermore, we will show that

$$\|\partial_t k_a(\cdot,x_0,\cdot) - \partial_t k_{\mathrm{Tr}}(\cdot,x_0,\cdot)\|_{L_1(Q_h)} \le C, \quad h > 0,\ x_0 \in \Omega_h.$$

In this way, it is possible to compare $\partial_t K_h$ with $\partial_t k_{\mathrm{Tr}}$. Finally, we show that for $2 \le p < \infty$ there exists $C > 0$ such that

$$\|\partial_t k_{\mathrm{Tr}} * f\|_{L_p(I;L_p(\Omega_h))} \le C\|f\|_{L_p(I;L_p(\Omega_h))}, \quad h > 0,\ f \in L_p(I;L_p(\Omega_h)).$$

This will complete the proof of Theorem 3.2.

Since an easy calculation shows that

$$\|\partial_t k_{\mathrm{Tr}}\|_{L_1(Q_h)} \le C|\log(h)|, \quad h > 0,$$

Corollary 3.5 will also follow.

Proof of Theorem 3.2. Let K, k_a and K_h be as above. Adding sufficiently large λ_0 to the coefficient c_0 of the form a, we obtain

(10)
$$\mathrm{Re}\left(\sum_{i,j=1}^N a_{ij}(x)\xi_i\overline{\xi_j} + \sum_{i=1}^N b_i(x)\xi_i\overline{\eta} + \sum_{i=1}^N c_i(x)\eta\overline{\xi_i} + (c_0(x)+\lambda_0)|\eta|^2\right) \ge \mu_0\big(|\xi|^2 + |\eta|^2\big),$$
$$\text{a.a. } x \in \Omega,\ \xi \in \mathbb{C}^N,\ \eta \in \mathbb{C},$$

for some $\mu_0 > 0$. Since shifting the operator does not affect the property of maximal regularity, we may assume that the form a satisfies (10).

We firstly fix some notation. For simplicity, we identify $u \in L_p(I; L_p(\Omega))$ for some $1 < p < \infty$ with a function $\tilde{u} : Q \to \mathbb{C}$ by $\tilde{u}(t,x) = \big(u(t)\big)(x)$. For convenience, we also denote this function $\tilde{u}$ by u. If $u \in L_p(I; L_p(\Omega))$ is differentiable, we analogously write $u'(t,x)$ for $\big(u'(t)\big)(x)$. Moreover, for $s \in \mathbb{N}$ we set $\|D^s u\|_{L_q(\Omega)} = \sum_{|\alpha|=s} \|D^\alpha u\|_{L_q(\Omega)}$, where $D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_N}}{\partial x_N^{\alpha_N}}$ and we write $D = D^1$. We need two different interpolation operators I_h and $\tilde{I}_h$. As I_h acts on continuous functions $\tilde{I}_h$ acts on Sobolev spaces. Details and error estimates for the interpolation operators are given in Lemmas 5.1, 5.2, and 5.3.

Step 1. $\|\partial_t(K_h - k_a) * f\|_{L_\infty(Q_h)} \le C\|f\|_{L_\infty(Q_h)}, \quad h > 0,\ f_h \in L_\infty(I; S_{h,\infty})$.

Let $h > 0$ and $x_0 \in \Omega_h$. Throughout this step we use the notation

$$e(t,y) = K_h(t,x_0,y) - k_a(t,x_0,y), \quad (t,y) \in Q.$$

By Hölder's inequality, it clearly suffices to show that $\|e'\|_{L_1(Q)} \le C$, where C is independent of x_0 and h. We therefore decompose Q in parabolic anulli. More precisely, for $\mu_* \ge 2(N+4)$ we choose J_h such that $\mu_* h/2 \le 2^{-J_h} \le \mu_* h$ and J_0 such that $2^{-J_0} \ge \operatorname{diam} Q$. We then set

$$Q_* := \Big\{(s,y) \in Q :\ \max\Big\{|y - x_0|,\ s^{\frac{1}{2}}\Big\} \le d_*\Big\}, \quad d_* := 2^{-J_h},$$
$$Q_j := \Big\{(s,y) \in Q :\ d_j \le \max\Big\{|y - x_0|,\ s^{\frac{1}{2}}\Big\} \le 2d_j\Big\}, \quad j \in Z,\ d_j = 2^{-j}.$$

Obviously, $Q = \bigcup_{j=J_0}^{J_h} Q_j \cup Q_*$. Analogously, we define $\Omega_j := \{y \in \Omega : d_j \le |y - x_0| \le 2d_j\}$ and $\Omega_* := \{y \in \Omega : |y - x_0| \le d_*\}$. Furthermore, we set $Q_j' := Q_{j-1} \cup Q_j \cup Q_{j+1}$, $Q_j'' = (Q_j')'$, $Q_*' = Q_* \cup Q_{J_h}$, and Ω_j', Ω_j'' and Ω_j''', analogously. We also need some weighted, local norms, i.e.,

$$\|u\|_{Q_j,s_1,s_2} := \sum_{s=0}^{s_1} d_j^{-2+s} \|D^s u\|_{Q_j} + \sum_{s=0}^{s_2} d_j^s \|D^s u'\|_{Q_j}, \quad j = * \text{ or } j \in \mathbb{Z},\ 0 \le s_1, s_2 \le 1,$$

for u smooth enough. Here, $\|\cdot\|_{Q_j}$ denotes the usual $L_2(Q_j)$-norm. In order to simplify notation, throughout this step, C and c denote constants which are independet of h, μ_*, and x_0.

Hölder's inequality yields

$$\|e'\|_{L_1(Q)} \le C \sum_{j=*, j=J_0}^{J_h} d_j^{\frac{N}{2}+1} \|e\|_{Q_j,1,0}.$$

We will show

$$\sum_{j=*, j=J_0}^{J_h} d_j^{\frac{N}{2}+1} \|e\|_{Q_j,1,0} \le C\mu_*^{\frac{N}{2}+1} + C + C\mu_*^{-1} \sum_{j=*, j=J_0}^{J_h} d_j^{\frac{N}{2}+1} \|e\|_{Q_j,1,0}. \tag{11}$$

Subtracting $C\mu_*^{-1} \sum_{j=*, j=J_0}^{J_h} d_j^{\frac{N}{2}+1} \|e\|_{Q_j,1,0}$ for sufficiently large μ_* on both sides completes Step 1.

The estimate $d_*^{\frac{N}{2}+1}\|e\|_{Q_*,1,0} \le C\mu_*^{\frac{N}{2}+1}$ follows from Lemma 5.5(a). Since $d_j \ge 2(N+2)h$ for $J_0 \le j \le J_h$, we may apply Lemma 5.7 to $\|e\|_{Q_j,1,0}$. Hence there exists $C > 0$ such that

$$\|e\|_{Q_j,1,0} \le C(\|\big(De\big)(0)\|_{\Omega_j'} + d_j^{-1}\|e(0)\|_{\Omega_j'} + d_j^{-2}\|e\|_{Q_j'} \\ +(hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q_j'} + \|k_a - I_h k_a\|_{Q_j',1,1}\big)$$

for $J_0 \le j \le J_h$. We start by showing

(12)
$$\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}+1}\Big(\|\big(De\big)(0)\|_{\Omega_j'} + d_j^{-1}\|e(0)\|_{\Omega_j'} + (hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q_j'} + \|k_a - I_h k_a\|_{Q_j',1,1}\Big) \le C.$$

As $h\sum_{j=J_0}^{J_h} d_j^{-1} \le C$, it suffices to show

$$\|\big(De\big)(0)\|_{\Omega_j'} + d_j^{-1}\|e(0)\|_{\Omega_j'} + (hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q_j'} + \|k_a - I_h k_a\|_{Q_j',1,1} \le Chd_j^{-\frac{N}{2}-2}, \\ J_0 \le j \le J_h.$$

We first consider $\|\big(De\big)(0)\|_{\Omega_j'}$. Note that $e(0) \restriction_{\Omega_j'} = (P_h\tilde{\delta}_{x_0} - \tilde{\delta}_{x_0}) \restriction_{\Omega_j'} = P_h\tilde{\delta} \restriction_{\Omega_j'}$. Therefore, by the inverse estimate (16) in the next section,

$$\|\big(De\big)(0)\|_{\Omega_j'} \le C\|DP_h\tilde{\delta}_{x_0}\|_{\Omega_j'} \le Ch^{-1}\|P_h\tilde{\delta}_{x_0}\|_{\Omega_j'}, \quad J_0 \le j \le J_h.$$

Since, by assumption,

$$cd_j \le \text{dist }(\Omega_j', \text{supp}\,\tilde{\delta}_{x_0}) \le Cd_j, \quad J_0 \le j \le J_h, \tag{13}$$

the exponential decay of P_h on Ω (see [Tho97, Lemma 5.1] or [Gei03, Lemma 2.2.9]) and Lemma 5.4 yields

$$\|\big(De\big)(0)\|_{\Omega_j'} \le Ch^{-1}\exp\left(-c\frac{d_j}{h}\right)\|\tilde{\delta}_{x_0}\|_{\tau_0} \le Ch^{-1}\exp\left(-c\frac{d_j}{h}\right)h^{-N}h^{\frac{N}{2}} \\ \le Chd_j^{-\frac{N}{2}-2}d_j^{\frac{N}{2}+2}h^{-\frac{N}{2}-2}\exp\left(-c\frac{d_j}{h}\right) \le Chd_j^{-\frac{N}{2}-2}, \quad J_0 \le j \le J_h.$$

Analogously, $d_j^{-1}\|e(0)\|_{\Omega_j'} \le Chd_j^{-\frac{N}{2}-2}, \quad J_0 \le j \le J_h.$

Next, by Lemma 5.5(a), we obtain

$$(hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q_j'} \le (hd_j^{-1})^{\frac{N}{2}+2}h^{-\frac{N}{2}-1} \le Chd_j^{-\frac{N}{2}-2}, \quad J_0 \le j \le J_h.$$

Finally, we consider $d_j\|D(k_a' - I_h k_a')\|_{Q_j'}$. By the error estimate for the interpolation operators $\tilde{I}_h$ and Lemma 5.5(b), we obtain

$$d_j\|D(k_a' - I_h k_a')\|_{Q_j'} \le Cd_j h\|D^2 k_a'\|_{Q_j''} \le Chd_j^{-\frac{N}{2}-2}, \quad J_0 \le j \le J_h.$$

The remaining terms of $\|k_a - I_h k_a\|_{Q_j',1,1}$ are similary bounded and the proof of (12) is complete.

Therefore, the proof of inequality (11) will be finished once we have shown

$$\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1}\|e\|_{Q_j',1,0} \le C + C\mu_*^{-1}\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}+1}\|e\|_{Q_j,1,0}. \tag{14}$$

Note that $\|e\|_{Q_j'} = \sup\big\{ \int_Q v(t,x)\overline{e(t,x)}\,\mathrm{d}(t,x) : \ v \in C_c^\infty(Q_j'),\ \|v\|_{Q_j'} = 1\big\}$. Let $v \in C_c^\infty(Q_j')$ with $\|v\|_{Q_j'} = 1$ and set $w(t) := \int_t^T T(s-t)v(s)\,\mathrm{d}s$. A simple calculation shows that w is the solution of the problem

$$\left\{ \begin{array}{rcl} -w'(t) - Aw(t) &=& v(t), \quad 0 \le t < T, \\ w(T) &=& 0. \end{array} \right.$$

Multiplying this equation by e and integrating by parts, we obtain

$$\begin{aligned} \int_0^T \big(v(t),e(t)\big)_\Omega \,\mathrm{d}t &= \big(w(0),e(0)\big)_\Omega + \int_0^T \big(w(t),e'(t)\big)_\Omega \,\mathrm{d}t + \int_0^T a(w(t),e(t))\,\mathrm{d}t \\ &= \big(w(0),e(0)\big)_\Omega + \int_0^T \big(w(t)-\tilde I_h w(t),e'(t)\big)_\Omega + a(w(t) \\ &\qquad\qquad - \tilde I_h w(t),e(t))\,\mathrm{d}t \\ &=: I_1^j + I_2^j, \quad J_0 \le j \le J_h. \end{aligned}$$

Here, we have used that $\big(\chi(t),e'(t)\big)_\Omega + a(\chi(t),e(t)) = 0$ for $t \in I$ and $\chi \in S_h$.

We start by showing $\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1} I_1^j \le C$. Since P_h is the orthogonal projection on S_h in $L_2(\Omega)$, we obtain

$$\begin{aligned} I_1^j &= \big(w(0),P_h\tilde\delta_{x_0} - \tilde\delta_{x_0}\big)_{\Omega_h} = \big(w(0)-\chi,P_h\tilde\delta_{x_0} - \tilde\delta_{x_0}\big)_{\Omega_h} \\ &= \big(w(0)-\chi,P_h\tilde\delta_{x_0}\big)_{\Omega_j'''} + \big(w(0)-\chi,P_h\tilde\delta_{x_0} - \tilde\delta_{x_0}\big)_{\Omega_h\setminus\Omega_j'''} \\ &=: I_{11}^j + I_{12}^j, \quad \chi \in S_h,\ J_0 \le j \le J_h. \end{aligned}$$

We choose $\chi := \tilde I_h(w(0))$. By the error estimate for $\tilde I_h$ and Lemma 5.5(c)(1), we obtain

$$\begin{aligned} I_{11}^j &\le Ch\|\big(Dw\big)(0)\|_\Omega \|P_h\tilde\delta_{x_0}\|_{\Omega_j'''} \\ &\le Ch\|w\|_Q\|P_h\tilde\delta_{x_0}\|_{\Omega_j''} \le Ch\|P_h\tilde\delta_{x_0}\|_{\Omega_j'''}, \quad J_0 \le j \le J_h. \end{aligned}$$

Now, inequality (13), the exponential decay of P_h on Ω, and Lemma 5.4 yield

$$I_{11}^j \le Ch\exp\left(-c\frac{d_j}{h}\right)\|\tilde\delta_{x_0}\|_{\tau_0} \le Ch\exp\left(-c\frac{d_j}{h}\right)h^{-N}h^{\frac{N}{2}} \le Chd_j^{-\frac{N}{2}}, \quad J_0 \le j \le J_h.$$

Since $(P_h)_{h>0}$ is stable in $\mathcal{L}(L_1(\Omega))$, we get $\|P_h\tilde\delta_{x_0} - \tilde\delta_{x_0}\|_{L_1(\Omega_h)} \le C$. By the error estimate for $\tilde I_h$, we thus obtain

$$I_{12}^j \le C\|I_h w(0) - w(0)\|_{L_\infty(\Omega_h\setminus\Omega_j''')} \le Ch\|\big(Dw\big)(0)\|_{L_\infty(\Omega\setminus\Omega_j'')}, \quad J_0 \le j \le J_h.$$

Therefore, $\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1} I_1^j \le C$ follows from $\|Dw(0)\|_{L_\infty(\Omega\setminus\Omega_j'')} \le Cd_j^{-\frac{N}{2}}$, $J_0 \le j \le J_h$, which is a consequence of the kernel estimate (KE_{Max}).

We will show

$$\sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1} I_2^j \le C\mu_*^{-1} \sum_{j=*,j=J_0}^{J_h} d_j^{\frac{N}{2}+1}\|e\|_{Q_j,1,0}. \tag{15}$$

This proves (14) and so (11) which completes Step 1.

By Hölder's inequality and the error estimate for $\tilde{I}_h$, there exists $C > 0$ such that for $J_0 \le j \le J_h$,

$$\begin{aligned} I_2^j &\le C \sum_{*,i=J_0}^{J_h} \|w - \tilde{I}_h w\|_{Q_i} \|e'\|_{Q_i} + C \sum_{*,i=J_0}^{J_h} \|D(w - \tilde{I}_h w)\|_{Q_i} \|De\|_{Q_i} \\ &\quad + C \sum_{*,i=J_0}^{J_h} \|D(w - \tilde{I}_h w)\|_{Q_i} \|e\|_{Q_i} + C \sum_{*,i=J_0}^{J_h} \|w - \tilde{I}_h w\|_{Q_i} \|De\|_{Q_i} \\ &\quad + C \sum_{*,i=J_0}^{J_h} \|w - \tilde{I}_h w\|_{Q_i} \|e\|_{Q_i} \\ &\le C \sum_{*,i=J_0}^{J_h} \big(\|D^2 w\|_{Q_i'} + \|Dw\|_{Q_i'}\big)\big(h^2 \|e'\|_{Q_i} + h\|e\|_{Q_i} + h\|De\|_{Q_i}\big). \end{aligned}$$

Since $h \le \mu_*^{-1} d_i$ and d_i is bounded from above, we have

$$h^2 \|e'\|_{Q_i} + h\|e\|_{Q_i} + h\|De\|_{Q_i} \le C\mu_*^{-1} d_i^2 \|e\|_{Q_i,1,0}, \quad i = *,\ J_0 \le i \le J_h.$$

Now, inequality (15) follows from Lemma 5.5(c)(2). Indeed, we have

$$\begin{aligned} \sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1} I_2^j &\le C\mu_*^{-1} \sum_{i=*,i=J_0}^{J_h} d_i^2 \|e\|_{Q_i,1,0} \sum_{j=J_0}^{J_h} d_j^{\frac{N}{2}-1} \min\left\{d_{i-j}^{\frac{N}{2}+1}, d_{j-i}^{\frac{N}{2}+1}\right\} \\ &\le C\mu_*^{-1} \sum_{i=*,i=J_0}^{J_h} d_i^{\frac{N}{2}+1} \|e\|_{Q_i,1,0} \left(\sum_{i \ge j} d_{i-j}^2 + \sum_{j>i} d_{i-j}^N\right) \\ &\le C\mu_*^{-1} \sum_{i=*,i=J_0}^{J_h} d_i^{\frac{N}{2}+1} \|e\|_{Q_i,1,0}. \end{aligned}$$

Step 2. $\|\partial_t(k_a - k_{\mathrm{Tr}}) * f_h\|_{L_\infty(Q_h)} \le C\|f_h\|_{L_\infty(Q_h)}$, $h > 0$, $f_h \in L_\infty(I; S_{h,\infty})$.

Let $x_0 \in \Omega_h$, $\tau_0 \in \mathcal{T}_h$ with $x_0 \in \overline{\tau_0}$ and μ_*, J_0, J_h, Q_j and Q_* as above. We define the *truncated kernel* $k_{\mathrm{Tr}} : I \times \Omega_h \times \Omega \to \mathbb{C}$ by

$$k_{\mathrm{Tr}}(t, x_0, y) := k_{\mathrm{Tr}}^h(t, x_0, y) := K(t, x_0, y)\chi_{(Q^*)^c}(t, y), \quad (t, y) \in Q,\ x_0 \in \Omega_h.$$

As in Step 1, C denotes some constant which is indepedet of x_0, μ_*, and h. We will show that

$$\|\partial_t k_a(\cdot, x_0, \cdot) - \partial_t k_{\mathrm{Tr}}(\cdot, x_0, \cdot)\|_{L_1(Q_h)} \le C.$$

By the representation $k_a(t, x_0, y) = \int_{\tau_0} K(t, z, y)\overline{\tilde{\delta}_{x_0}(z)}\, \mathrm{d}z$ and $\int_{\Omega_h} \tilde{\delta}_{x_0}(x)\, \mathrm{d}x = 1$ (see Lemma 5.4), we obtain

$$\begin{aligned} |\partial_t k_a(t, x_0, y) - \partial_t k_{\mathrm{Tr}}(t, x_0, y)| &= \left| \int_{\tau_0} \partial_t K(t, z, y)\overline{\tilde{\delta}_{x_0}(z)}\, \mathrm{d}z - \partial_t K(t, x_0, y) \right| \\ &= \left| \int_{\tau_0} \big(\partial_t K(t, z, y) - \partial_t K(t, x_0, y)\big)\overline{\tilde{\delta}_{x_0}(z)}\, \mathrm{d}z \right|, \quad J_0 \le j \le J_h,\ (t, y) \in Q_j. \end{aligned}$$

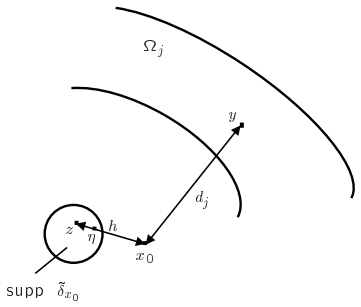


FIG. 1. *Correlation between x_0, y, and z.*

Due to the mean value theorem there exists $\eta(z)$ in the segment $\overline{zx_0}$ joining z and x_0 such that

$$|\partial_t k_a(t,x_0,y) - \partial_t k_{\mathrm{Tr}}(t,x_0,y)| \le \sum_{|\alpha|=1} \left| \int_{\tau_0} |z-x_0| \partial_t \partial_1^\alpha K(t,\eta(z),y) \overline{\tilde{\delta}_{x_0}(z)} \, \mathrm{d}z \right|.$$

Now, by Lemma B.1, there exits $C>0$ such that for $J_0 \le j \le J_h$ and $(t,y) \in Q_j$,

$$|\partial_t k_a(t,x_0,y) - \partial_t k_{\mathrm{Tr}}(t,x_0,y)| \le C \int_{\tau_0} |z-x_0| \left(\max\left\{ \min_{\eta \in \overline{zx_0}} \{|y-\eta|\}, t^{\frac{1}{2}} \right\} \right)^{-N-3} |\tilde{\delta}_{x_0}(z)| \, \mathrm{d}z.$$

Since (see Figure 1) $|z-x_0| \le h$ and $|x_0-y| \le C|y-\eta|$ for $z \in \tau_0$ and $\eta \in \overline{zx_0}$, we obtain

$$|\partial_t k_a(t,x_0,y) - \partial_t k_{\mathrm{Tr}}(t,x_0,y)| \le C h d_j^{-N-3} \int_{\tau_0} |\tilde{\delta}_{x_0}(z)| \, \mathrm{d}z \le C d_j^{-N-2} h d_j^{-1},$$
$$J_0 \le j \le J_h, \ (t,y) \in Q_j.$$

Summing, we get

$$\sum_{j=J_0}^{J_h} \|\partial_t k_a(\cdot,x_0,\cdot) - \partial_t k_{\mathrm{Tr}}(\cdot,x_0,\cdot)\|_{L_1(Q_j)} \le C \sum_{j=J_0}^{J_h} \int_{Q_j} d_j^{-N-2}(h d_j^{-1}) \, \mathrm{d}(s,y) \le C.$$

Finally, by Lemma 5.5(a), we obtain $\|\partial_t k_a(\cdot,x_0,\cdot)\|_{L_1(Q_*)} \le C$. Step 2 is complete.

Step 3. Final Step.

Recall that A_h has discrete maximal L_2 regularity on $S_{h,2}$. By (KE_{Max}) and [HP97, Theorem 3.1], the operator associated to $\partial_t K$ satisfies the assumptions of Lemma A.2. Therefore, the operators associated to $\partial_t k_{\mathrm{Tr}}$ are uniformly bounded in h in $\mathcal{L}(L_2(Q))$, and we obtain

$$\|\big(\partial_t K_h - \partial_t k_{Tr}\big) * f_h\|_{Q_h} \le C \|f_h\|_{Q_h}, \quad h>0, \ f_h \in L_2(I;S_{h,2}).$$

Due to Step 1 and Step 2 the operators associated to $\partial_t K_h - \partial_t k_a + \partial_t k_a - \partial_t k_{\mathrm{Tr}}$ are uniformly bounded in h in $\mathcal{L}(L_\infty(I;S_{h,\infty}))$. Marcinkiewicz' interpolation theorem and the identity $\partial_t K_h - \partial_t k_a + \partial_t k_a - \partial_t k_{\mathrm{Tr}} = \partial_t K_h - \partial_t k_{\mathrm{Tr}}$ yields that the operators associated to the right-hand side are uniformly bounded in h in $\mathcal{L}(L_p(I;S_{h,p}))$, $2 \le p < \infty$. Since $\partial_t K_h = \partial_t K_h - \partial_t k_{\mathrm{Tr}} + \partial_t k_{\mathrm{Tr}}$, Theorem A.2 completes the proof. ∎

5. Technical estimates. In this section we proof technical estimates needed to complete the proof of Theorem 3.2. Throughout this section we use the same notation as therein. We start by some basic properties of the finite element spaces S_h and interpolation operators.

Since we use a quasi-uniform family of triangulations, we have the following inverse inequalities:

$$\|\omega_h D^\alpha u_h\|_\tau \le Ch^{-1}\|\omega_h u_h\|_\tau, \quad h>0,\ \tau\in\mathcal{T}_h,\ \omega_h, u_h\in S_h,\ |\alpha|=1 \tag{16}$$

$$\|\omega_h u_h\|_{L_\infty(\tau)} \le Ch^{-\frac{N}{2}}\|\omega_h u_h\|_\tau, \quad h>0,\ \tau\in\mathcal{T}_h,\ \omega_h,\ u_h\in S_h. \tag{17}$$

Next, we define the *standard interpolant* $I_h : C(\overline{\Omega})\cap H \to S_h$ by $u\mapsto u_h$, where $u_h(\xi)=u(\xi)$ for all vertices ξ of $\mathcal{T}_h$, i.e., ξ is a vertex of τ for some $\tau\in\mathcal{T}_h$. Since we are dealing with piecewise linear functions, the operator I_h is well defined. The following lemma is an easy consequence of Taylor's theorem.

LEMMA 5.1. *There exists $C>0$ such that, for $h>0$ and $\tau\in\mathcal{T}_h$,*

$$\|I_h u-u\|_{L_\infty(\tau)} + h\|D(I_h u-u)\|_{L_\infty(\tau)} \le Ch^2\|D^2u\|_{L_\infty(\tau)},$$
$$u\in C(\overline{\Omega})\cap H \text{ with } u\restriction_{\overline{\tau}}\in C^2(\overline{\tau}).$$

We also need interpolation operators acting on Sobolev spaces $W_1^1(\Omega)$ or $W_{1,0}^1(\Omega)$, where $W_{1,0}^1(\Omega)$ denotes the closure of $C_c^\infty(\Omega)$ in $W_1^1(\Omega)$. We call them $\tilde I_h^D$ and $\tilde I_h^N$, respectively.

For the definition of $\tilde I_h^D$ we refer to [TW00, Lemma 2.1]. Set $U(\tau):=\{x\in\Omega : \operatorname{dist}(x,\tau)\le h\}$ for $\tau\in\mathcal{T}_h$. Then the following lemma is proved therein.

LEMMA 5.2. *Let $1\le q\le\infty$. Then there exists $C>0$ such that, for $h>0$ and $\tau\in\mathcal{T}_h^D$,*

(a) $\|\tilde I_h^D u-u\|_{L_q(\tau)} \le Ch\|Du\|_{L_q(U(\tau))}$, $u\in W_{q,0}^1(\Omega)$,

(b) $\|\tilde I_h^D u-u\|_{L_q(\tau)} \le Ch^2\sum_{s=1}^2\|D^s u\|_{L_q(U(\tau))}$, $u\in W_{q,0}^1(\Omega)$ *with* $u\restriction_{U(\tau)}\in W_q^2(U(\tau))$,

(c) $\|D(\tilde I_h^D u-u)\|_{L_q(\tau)} \le Ch\sum_{s=1}^2\|D^s u\|_{L_q(U(\tau))}$, $u\in W_{q,0}^1(\Omega)$ *with* $u\restriction_{U(\tau)}\in W_q^2(U(\tau))$.

Further, we define $\tilde I_h^N : W_1^1(\Omega)\to S_h^N$. For a boundary vertex ξ choose $\tau\in\mathcal{T}_h$ such that τ has N vertices on the boundary and let F_{bdy} denote the $(N-1)$-simplex given by those vertices. By [SZ90] there exists an affine mapping β_ξ, which is bounded on F_{bdy}, uniformly in h and ξ, such that

$$\frac{1}{|F_{bdy}|}\int_{F_{bdy}}\beta_\xi(x)u(x)\,\mathrm{d}x = u(\xi), \quad u\in C(\overline{\Omega}) \text{ linear}.$$

Moreover, by quasiuniformity, there exists $c>0$, independent of h, such that the ball $B(\xi,ch)$ with radius ch does not intersect the boundary of Ω_h for vertices $\xi\in\Omega_h$. For $u\in W_1^1(\Omega)$ we now define $\tilde I_h^N u:=u_h$, where

$$u_h(\xi)=\begin{cases}\dfrac{1}{|F_{bdy}|}\displaystyle\int_{F_{bdy}}\beta_\xi(x)u(x)\,\mathrm{d}x, & \text{vertices } \xi\in\partial\Omega_h,\\[2ex] \dfrac{1}{|B(\xi,Ch)|}\displaystyle\int_{B(\xi,Ch)}u(x)\,\mathrm{d}x, & \text{vertices } \xi\in\Omega_h.\end{cases}$$

Then, by the Bramble–Hilbert lemma, we obtain the following lemma.

LEMMA 5.3. *Let $1\le q\le\infty$. Then there exists $C>0$ such that for $h>0$ and $\tau\in\mathcal{T}_h^N$*

(a) $\|\tilde{I}_h^N u - u\|_{L_q(\tau)} \le Ch\|Du\|_{L_q(U(\tau))}$, $u \in W_q^1(\Omega)$,
(b) $\|\tilde{I}_h^N u - u\|_{L_q(\tau)} \le Ch^2\|D^2 u\|_{L_q(U(\tau))}$, $u \in W_q^1(\Omega)$ *with* $u \restriction_{U(\tau)} \in W_q^2(U(\tau))$,
(c) $\|D(\tilde{I}_h^N u - u)\|_{L_q(\tau)} \le Ch\|D^2 u\|_{L_q(U(\tau))}$, $u \in W_q^1(\Omega)$ *with* $u \restriction_{U(\tau)} \in W_q^2(U(\tau))$.

We write $\tilde{I}_h$ for $\tilde{I}_h^D$ and $\tilde{I}_h^N$. We next introduce a discrete version of the delta distribution.

LEMMA 5.4. *For $k \in \mathbb{N}$ there exist $C, c > 0$, independent of h, such that for any $x_0 \in \overline{\tau_0}$, where $\tau_0 \in \mathcal{T}_h$, there exists a function $\tilde{\delta}_{x_0} := \tilde{\delta}_{x_0}^h \in C_c^\infty(\Omega)$ with support in τ_0 such that*

(a) $\int_{\tau_0} \chi(x)\overline{\tilde{\delta}_{x_0}(x)}\,\mathrm{d}x = \chi(x_0)$, $\chi \in S_h$,
(b) $\int_{\tau_0} \tilde{\delta}_{x_0}(x)\,\mathrm{d}x = 1$,
(c) diam $\operatorname{supp}\tilde{\delta}_{x_0} \le Ch$,
(d) dist $(\operatorname{supp}\tilde{\delta}_{x_0}, \partial\tau_0) \ge ch$,
(e) $\|\tilde{\delta}_{x_0}\|_{W_\infty^k(\tau_0)} \le C|h|^{-N-k}$.

A proof may be found in [TW00, Lemma 2.2]. The next lemma states a priori estimates for e, k_a, and w.

LEMMA 5.5.

(a) *There exists $C > 0$, independent of x_0 and h, such that*

$$\|e\|_{Q,1,0} \le C\|\tilde{\delta}_{x_0}\|_{H^1(\Omega)} \le Ch^{-\frac{N}{2}-1}.$$

(b) *There exists $C > 0$, independent of x_0 and h, such that*

$$\|\partial_t^k D^\alpha k_a(\cdot, x_0, \cdot)\|_{Q_j} \le C d_j^{-\frac{N}{2}+1-2k-|\alpha|},$$
$$j = J_0, \dots, J_h + 2,\ 0 \le k \le 1,\ |\alpha| \le 2.$$

(c) *For $v \in L_2(Q)$ let $w := \int_t^T T(s-t)v(s)\,\mathrm{d}s$.*
   (1) *Then there exists $C > 0$, independent of v, such that $\|w(0)\|_{H^1(\Omega)} \le C\|v\|_Q$.*
   (2) *Assume that* $\operatorname{supp} v \subset Q_j'$ *with $\|v\|_{Q_j} = 1$. Then there exists $C > 0$, independent of v, such that*

$$\|D^\alpha w\|_{Q_i'} \le C \min\left\{d_{j-i}^{\frac{N}{2}+1}, d_{i-j}^{\frac{N}{2}+1}\right\}, \quad i, j = J_0, \dots, J_h,\ i = *,\ |\alpha| \le 2.$$

Proof.

(a) The estimate $\|\tilde{\delta}_{x_0}\|_{H^1(\Omega)} \le Ch^{-\frac{N}{2}-1}$ follows by Lemma 5.4(c) and (e). It thus remains to prove the first inequality of (a). We will prove $\|K_h(\cdot, x_0, \cdot)\|_{Q,1,0} \le C\|\tilde{\delta}_{x_0}\|_{H^1(\Omega)}$ only since $\|k_a(\cdot, x_0, \cdot)\|_{Q,1,0} \le C\|\tilde{\delta}_{x_0}\|_{H^1(\Omega)}$ follows in a similar way. By assumption, $K_h(\cdot, x_0, \cdot)$ satisfies $K_h'(t, x_0, \cdot) - A^* K_h(t, x_0, \cdot) = 0$ for $0 < t \le T$. Multiplying this equation by $K_h(\cdot, x_0, \cdot)$ and $K_h'(\cdot, x_0, \cdot)$ and integrating by parts yields

$$\|K_h(\cdot, x_0, \cdot)\|_{Q,1,0} + \|K_h(T, x_0, \cdot)\|_\Omega \le C\|P_h\tilde{\delta}_{x_0}\|_{H^1(\Omega)}$$

and

$$\|K_h'(\cdot, x_0, \cdot)\|_Q \le C\|P_h\tilde{\delta}_{x_0}\|_{H^1(\Omega)} + \|K_h(\cdot, x_0, \cdot)\|_{Q,1,0} + \|K_h(T, x_0, \cdot)\|_\Omega,$$

where we have used $K_h(0, x_0, \cdot) = P_h\tilde{\delta}_{x_0}$. Since P_h is stable in $H^1(\Omega)$, assertion (a) follows.

(b) This follows from the representation of the approximate kernel k_a and the kernel estimates given in Lemma B.1.

(c) Let $\tilde{v} : \mathbb{R}_+ \to L_2(\Omega)$ be given by $t \mapsto v(T-t,\cdot)$, $t \in I$, and $t \mapsto 0$, $t > T$. Then the solution $\tilde{w}$ of

$$\begin{aligned} \tilde{w}'(t) - A\tilde{w}(t) &= \tilde{v}(t), \quad t > 0, \\ \tilde{w}(0) &= 0, \end{aligned}$$

satisfies $\tilde{w}(t) = w(T-t)$, $t \in I$. By the assumption $D(A) \hookrightarrow H^2(\Omega)$, there exists $C > 0$, independent of w, such that

$$\|w\|_{H^1(\Omega)} \le C\|w\|_{(L_2(\Omega),H^2(\Omega))_{\frac{1}{2},2}} \le C\|w\|_{(L_2(\Omega),D(A))_{\frac{1}{2},2}}.$$

Therefore, by [Ama95, Theorem 4.10.2] there exists $C > 0$, independent of v, such that

$$\|\tilde{w}\|_{L_\infty(I;(L_2(\Omega),D(A))_{\frac{1}{2},2})} \le C\|\tilde{v}\|_{L_2(\mathbb{R}_+\times\Omega)} = C\|v\|_Q,$$

which proves (c)(1).

By the kernel estimate (KE_{Max}), there exists $C > 0$ such that for $i = *$, $j \le J_h - 3$ or $|i-j| \ge 3$, and $|\alpha| \le 2$,

$$\begin{aligned} \|D^\alpha w\|_{Q_i'} &\le C d_i^{\frac{N}{2}+1} \|D_1^\alpha w\|_{L_\infty(Q_i')} \\ &\le C d_i^{\frac{N}{2}+1} \|v\|_{L_1(Q_j')} \sup_{(t,x)\in Q_i,(s,y)\in Q_j,s>t} |D^\alpha K(s-t,x,y)| \\ &\le C d_{i+j}^{\frac{N}{2}+1} \min\{d_j^{-N-2}, d_i^{-N-2}\} \le C \min\left\{d_{i-j}^{\frac{N}{2}+1}, d_{j-i}^{\frac{N}{2}+1}\right\}. \end{aligned}$$

For $i = *$, $j > J_h - 3$ or $|i-j| < 3$ it suffices to show that $\|D^\alpha w\|_{Q_i}$ is uniformly bounded in i and j. In fact, the assumption $D(A) \hookrightarrow H^2(\Omega)$ and maximal L_2 regularity on $L_2(\Omega)$ of A yields

$$\|D^\alpha w\|_{Q_i} \le \|D^\alpha w\|_Q \le C\|v\|_Q \le C, \quad |\alpha| \le 2.$$

This proves (c)(2). ∎

We conclude this section with local estimates for e. However, for the proof we need superconvergent-type estimates given in the next lemma.

LEMMA 5.6 (superconvergent-type estimate). *There exists some $C > 0$ such that for $h > 0$, $\tau \in \mathcal{T}_h$ $z_h, \omega \in S_h$ satisfying $\|\omega\|_{L_\infty(\Omega)} \le 1$*

$$\|\omega^4 z_h - I_h(\omega^4 z_h)\|_\tau \le C(h^2\|D\omega\|_{L_\infty(\Omega)}^2 + h\|D\omega\|_{L_\infty(\Omega)})\|\omega^2 z_h\|_\tau, \tag{18}$$

$$\|D(\omega^4 z_h - I_h(\omega^4 z_h))\|_\tau \le Ch\|D\omega\|_{L_\infty(\Omega)}^2\|z_h\|_\tau + Ch\|D\omega\|_{L_\infty(\Omega)}\|Dz_h\|_\tau, \tag{19}$$

$$\|\omega^8 z_h - I_h(\omega^8 z_h)\|_\tau \le C(h^2\|D\omega\|_{L_\infty(\Omega)}^2 + h\|D\omega\|_{L_\infty(\Omega)})\|\omega^6 z_h\|_\tau, \tag{20}$$

$$\|D(\omega^8 z_h - I_h(\omega^8 z_h))\|_\tau \le C(h\|D\omega\|_{L_\infty(\Omega)}^2 + \|D\omega\|_{L_\infty(\Omega)})\|\omega^6 z_h\|_\tau. \tag{21}$$

In particular, inequalities (18)–(21) *hold with τ replaced by Ω.*

Proof. For $h > 0$ let $\tau \in \mathcal{T}_h$ and set $\chi := I_h(\omega^4 z_h)$ for $\omega, z_h \in S_h$. Note that $D^2 z_h = 0$. By the error estimate for the standard interpolant I_h, there exists $C > 0$, independent of h, z_h, ω, and τ, such that

$$\begin{aligned} \|\omega^4 z_h - \chi\|_\tau &\le h^{\frac{N}{2}}\|\omega^4 z_h - \chi\|_{L_\infty(\tau)} \le Ch^{\frac{N}{2}+2}\|D^2(\omega^4 z_h)\|_{L_\infty(\tau)} \\ &\le C\left(h^{\frac{N}{2}+2}\|D\omega\|_{L_\infty(\Omega)}^2\|\omega^2 z_h\|_{L_\infty(\tau)} + h^{\frac{N}{2}+2}\|D\omega\|_{L_\infty(\Omega)}\|\omega^2 Dz_h\|_{L_\infty(\tau)}\right). \end{aligned}$$

Now, by the inverse estimates (17) and (16) for z_h, we obtain

$$\begin{aligned}\|\omega^4 z_h - \chi\|_\tau &\le C(h^2\|D\omega\|^2_{L_\infty(\Omega)}\|\omega^2 z_h\|_\tau + h^2\|D\omega\|_{L_\infty(\Omega)}\|\omega^2 Dz_h\|_\tau)\\ &\le C(h^2\|D\omega\|^2_{L_\infty(\Omega)}\|\omega^2 z_h\|_\tau + h\|D\omega\|_{L_\infty(\Omega)}\|\omega^2 z_h\|_\tau).\end{aligned}$$

Similarly, we get (19)–(21). ∎

LEMMA 5.7. *There exists $C > 0$, independent of x_0 and h, such that for $d_j > 2(N+4)h$,*

$$\begin{aligned}\|e\|_{Q_j,1,0} \le C\Big(&\|\big(De\big)(0)\|_{\Omega'_j} + d_j^{-1}\|e(0)\|_{\Omega'_j}\\ &+ d_j^{-2}\|e\|_{Q'_j} + (hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q'_j} + \|k_a - I_h k_a\|_{Q'_j,1,1}\Big).\end{aligned}$$

Proof. Since $e = k_a - I_h k_a = k_a$ on $Q \setminus Q_h$, it suffices to consider the contribution of e in $Q_j \cap Q_h$. Let h and j be fixed and set for $n = 0, \dots, N+4$

$$Q_j^n := \left\{(t,x) \in Q'_j \cap Q_h : d_j - \frac{n}{N+4}d_{j-1} < \max\left\{|x-x_0|, t^{\frac12}\right\} < d_j + \frac{n}{N+4}d_j\right\}.$$

Clearly, $Q_j^0 = Q_j \cap Q_h$ and $Q_j^{N+4} = Q'_j \cap Q_h$. Throughout this proof M, C, C_1, C_2 denote constants independent of j, n, x_0, and h. By quasiuniformity there exist cut-off functions $\omega_n : Q \to [0,1]$, $n = 1, \dots, N+4$, such that $\omega_n(t,\cdot) \in S_h$, $\omega_n(\cdot,x)$ is continuously differentiable, $\omega_n \equiv 1$ on Q_j^{n-1}, $\omega_n \equiv 0$ outside Q_j^n, $\|D\omega_n\|_{L_\infty(Q)} \le Md_j^{-1}$, and $\|\omega'_n\|_{L_\infty(Q)} \le Md_j^{-2}$. Throughout this proof we set $z(t) = k_a(t) - I_h k_a(t)$ and $z_h(t) = K_h(t) - I_h k_a(t)$ for $t \in I$ to shorten notation. Note that $e(t) = z_h(t) - z(t)$. We will use the coercivity of the form a, the equation

$$\big(\chi, e'(t)\big)_\Omega + a(\chi, e(t)) = 0, \quad \chi \in S_h,\ 0 < t \le T, \tag{22}$$

and superconvergent-type estimates to prove

$$\begin{aligned}&d_j^{-1}\|\omega_n^2 De\|_Q + d_j^{-2}\|e\|_{Q_j} + d_j^{-1}\|\big(\omega_n^2 e\big)(T)\|_\Omega \\ &\le C_1\big(d_j^{-1}\|e(0)\|_{\Omega'_j} + d_j^{-2}\|e\|_{Q'_j} + (hd_j^{-1})^{\frac12}(\|e'\|_{Q_j^n} + d_j^{-1}\|De\|_{Q_j^n}) + \epsilon^{-1}\|z\|_{Q'_j,1,1} + \epsilon\|\omega_n^4 e'\|_Q\big),\end{aligned} \tag{23}$$

for $\epsilon > 0$ and

$$\begin{aligned}\|\omega_n^4 e'\|_Q \le C_2\big(&\|(De)(0)\|_{\Omega'_j} + d_j^{-1}\|e(0)\|_{\Omega'_j} + d_j^{-2}\|e\|_{Q'_j} + (hd_j^{-1})^{\frac12}\|e'\|_{Q_j^n}\\ &+ \|z\|_{Q'_j,1,1} + d_j^{-1}\|\omega_n^2 De\|_Q + d_j^{-1}\|\big(\omega_n^2 e\big)(T)\|_\Omega\big).\end{aligned} \tag{24}$$

Then, we multiply inequality (24) by $1/(2C_2)$ and add it to (23) with $\epsilon = 1/(4C_2)$. By subtracting $\epsilon\|\omega^4 e'\|_Q$, $1/2d_j^{-1}\|\big(\omega^2 e\big)(T)\|_\Omega$ and $1/2d_j^{-1}\|\omega^2 De\|_Q$ on both sides, we obtain

$$\begin{aligned}\|e'\|_{Q_j^{n-1}} + d_j^{-1}\|De\|_{Q_j^{n-1}} + d_j^{-2}\|e\|_{Q_j} \le C\big(&\|(De)(0)\|_{\Omega'_j} + d_j^{-1}\|e(0)\|_{\Omega'_j} + d_j^{-2}\|e\|_{Q'_j}\\ &+(hd_j^{-1})^{\frac12}\big(\|e'\|_{Q_j^n} + d_j^{-1}\|De\|_{Q_j^n}\big) + \|z\|_{Q'_j,1,1}\big).\end{aligned} \tag{25}$$

Since $hd_j^{-1} \le 2(N+4)$, by iteration we get

$$\begin{aligned}\|e\|_{Q_j,1,0} \le C\big(&\|\big(De\big)(0)\|_{\Omega'_j} + d_j^{-1}\|e(0)\|_{\Omega'_j} + d_j^{-2}\|e\|_{Q'_j} + (hd_j^{-1})^{\frac{N}{2}+2}\|e'\|_{Q'_j}\\ &+ \|z\|_{Q'_j,1,1} + hd_j^{-2}\|De\|_{Q'_j}\big).\end{aligned} \tag{26}$$

The last term on the right-hand side of inequality (26) may be omitted since

$$hd_j^{-2}\|De\|_{Q_j'} \le hd_j^{-2}\|Dz\|_{Q_j'} + hd_j^{-2}\|Dz_h\|_{Q_j'} \le Cd_j^{-1}\|Dz\|_{Q_j'} + Cd_j^{-2}(\|e\|_{Q_j'} + \|z\|_{Q_j'}).$$

Therefore, it remains to prove inequalities (23) and (24). For readability we suppress the dependency on t and introduce the notation

$$a_\omega^\#(u,v) := \sum_{i,j=1}^N (\omega a_{ij}\partial_i u, \partial_j v)_\Omega, \quad a^\#(u,v) := \sum_{i,j=1}^N (a_{ij}\partial_i u, \partial_j v)_\Omega \quad u,v \in H,$$

where ω is smooth enough. The term $a_\omega(u,v)$ is defined in an analogous way. Note that $a_\omega^\#$ is the principal part of a with some weight ω. For inequality (24), we start with the identity

$$\begin{aligned}
&\int_I \|\omega_n^4 e'\|_\Omega^2 + \frac{1}{2}\frac{d}{dt}a_{\omega_n^8}^\#(e,e)\ \mathrm{d}t \\
&= \int_I \big(\omega_n^8 e', e'\big)_\Omega + \frac{1}{2}a_{(\omega_n^8)'}^\#(e,e) + \mathrm{Re}\ a_{\omega_n^8}^\#(e',e) + \mathrm{Re}\ a(\omega_n^8 z_h', e) - \mathrm{Re}\ a(\omega_n^8 z_h', e)\ \mathrm{d}t \\
&= \int_I \big(- \mathrm{Re}\ (\omega_n^8 z', e')_\Omega + \frac{1}{2}a_{(\omega_n^8)'}^\#(e,e) - \mathrm{Re}\ a_{\omega_n^8}^\#(z',e)\big)\ \mathrm{d}t \\
&\quad + \mathrm{Re}\int_I a_{\omega_n^8}^\#(z_h', e) - a(\omega_n^8 z_h', e)\ \mathrm{d}t + \mathrm{Re}\int_I \big(\omega_n^8 z_h', e'\big)_\Omega + a(\omega_n^8 z_h', e)\ \mathrm{d}t \\
&=: J_1 + J_2 + J_3.
\end{aligned}$$

Noting that $d_j^{-1} > C$, where C depends on Ω only, we obtain

$$\begin{aligned}
|J_1| &\le C\|z\|_{Q_j',1,1}^2 + \frac{1}{16}\|\omega_n^4 e'\|_{Q_j'}^2 + Cd_j^{-2}\|\omega_n^2 De\|_Q^2, \\
|J_2| &\le \left| \int_I a_{z_h'}^\#(\omega_n^8, e) + \sum_{i=1}^N \big(c_i\omega_n^8 z_h', \partial_i e\big)_\Omega + \big(c_0\omega_n^8 z_h', e\big)_\Omega\ \mathrm{d}t \right| \\
&\qquad + \left| \int_I \sum_{i=1}^N \big(b_i \partial_i(\omega_n^8 z_h'), e\big)_\Omega\ \mathrm{d}t \right| \\
&=: J_{21} + J_{22}.
\end{aligned}$$

We start with the first term of J_{21}

$$\begin{aligned}
\left| \int_I a_{z_h'}^\#(\omega_n^8, e)\ \mathrm{d}t \right| &\le Cd_j^{-2}\|\omega_n^2 De\|_Q^2 + \frac{1}{16}\|\omega_n^4 z_h'\|_Q^2 \\
&\le C\|z'\|_{Q_j'}^2 + Cd_j^{-2}\|\omega_n^2 De\|_Q^2 + \frac{1}{16}\|\omega_n^4 e'\|_Q^2.
\end{aligned}$$

By similar estimates for the two remaining terms of J_{21}, we obtain

$$J_{21} \le Cd_j^{-4}\|e\|_{Q_j'}^2 + C\|z'\|_{Q_j'}^2 + Cd_j^{-2}\|\omega_n^2 De\|_Q^2 + \frac{1}{8}\|\omega_n^4 e'\|_Q^2.$$

In order to estimate J_{22}, we use integration by parts in time. Recall that μ denotes

the constant appearing in (8).

$$
\begin{aligned}
J_{22} &= \left| \sum_{i=1}^N \int_I \big(b_i\big[(\partial_i\omega_n^8)(e'+z') + \omega_n^8\partial_i(e'+z')\big](t), e(t)\big)_\Omega \,\mathrm{d}t \right| \\
&= \left| \sum_{i=1}^N \int_I \big(b_i\big((\partial_i\omega_n^8)(e'+z')\big)(t), e(t)\big)_\Omega \,\mathrm{d}t + \sum_{i=1}^N \int_I \big(b_i\big(\omega_n^8\partial_i z'\big)(t), e(t)\big)_\Omega \,\mathrm{d}t \right. \\
&\quad \left. - \sum_{i=1}^N \int_I \big(b_i\big(\partial_i e\big)(t), \big(\omega_n^8 e\big)'(t)\big)_\Omega \,\mathrm{d}t + \sum_{i=1}^N \big(b_i\big(\omega_n^8\partial_i e\big)(t), e(t)\big)_\Omega\big|_0^T \right| \\
&\le Cd_j^{-1}\|\omega_n^4 e'\|_Q\|e\|_{Q_j'} + Cd_j^{-1}\|z'\|_{Q_j'}\|e\|_{Q_j'} + C\|Dz'\|_{Q_j'}\|e\|_{Q_j'} \\
&\quad + Cd_j^{-2}\|\omega_n^2 De\|_Q\|e\|_{Q_j'} + C\|\omega_n^2 De\|_Q\|\omega_n^4 e'\|_Q + C\|\big(\omega_n^4 De\big)(T)\|_\Omega\|\big(\omega_n^2 e\big)(T)\|_\Omega \\
&\quad + C\|(De)(0)\|_{\Omega_j'}\|e(0)\|_{\Omega_j'} \\
&\le \frac{1}{16}\|\omega_n^4 e'\|_Q^2 + Cd_j^{-4}\|e\|_{Q_j'}^2 + C\|z'\|_{Q_j'}^2 + Cd_j^2\|Dz'\|_{Q_j'}^2 + Cd_j^{-2}\|\omega_n^2 De\|_{Q_j'}^2 \\
&\quad + \frac{\mu}{4}\|\big(\omega_n^4 De\big)(T)\|_\Omega^2 + Cd_j^{-2}\|\big(\omega_n^2 e\big)(T)\|_\Omega^2 + C\|(De)(0)\|_{\Omega_j'}^2 + Cd_j^{-2}\|e(0)\|_{\Omega_j'}^2.
\end{aligned}
$$

Next, we apply equality (22) with $\chi = I_h(\omega_n^8 z_h')$ to J_3.

$$
\begin{aligned}
J_3 &= \left| \int_I \big(\omega_n^8 z_h' - \chi, e'\big)_\Omega + a(\omega_n^8 z_h' - \chi, e) \,\mathrm{d}t \right| \\
&\le \big(\|\omega_n^8 z_h' - \chi\|_Q\|e'\|_{Q_j^n} + \|D(\omega_n^8 z_h' - \chi)\|_Q\|e\|_{Q_j'} + \|\omega_n^8 z_h' - \chi\|_Q\|e\|_{Q_j'}\big) \\
&\quad + C\left(\sum_{i=1}^N \left| \int_I (c_i(\omega_n^8 z_h' - \chi), \partial_i e)_\Omega + a^\#(\omega_n^8 z_h' - \chi, e) \,\mathrm{d}t \right|\right) \\
&=: J_{31} + J_{32}.
\end{aligned}
$$

By inequality (20),

$$
\|\omega_n^8 z_h' - \chi\|_Q \le Chd_j^{-1}\|\omega_n^6 z_h'\|_Q \le Chd_j^{-1}\|z'\|_{Q_j'} + Chd_j^{-1}\|e'\|_{Q_j^n}.
$$

Since $\|D(\omega_n^8 z_h' - \chi)\|_Q$ may be estimated in a similar way, we obtain

$$
J_{31} \le Cd_j^{-4}\|e\|_{Q_j'}^2 + C\|z'\|_{Q_j'}^2 + Chd_j^{-1}\|e'\|_{Q_j^n}^2 + \frac{1}{8}\|\omega_n^4 e'\|_Q^2.
$$

Estimating J_{32} is more involved. In fact, since Dz_h is constant on each $\tau \in \mathcal{T}_h$, by inverse estimates we get

$$
\begin{aligned}
\|Dz_h\|_\tau\|\omega_n^2\|_{L_\infty(\tau)} &\le h^{\frac{N}{2}}\|Dz_h\|_{L_\infty(\tau)}\|\omega_n^2\|_{L_\infty(\tau)} = h^{\frac{N}{2}}\big|Dz_h\restriction_\tau\big|\|\omega_n^2\|_{L_\infty(\tau)} \\
&= h^{\frac{N}{2}}\|\omega_n^2 Dz_h\|_{L_\infty(\tau)} \le C\|\omega_n^2 Dz_h\|_\tau, \quad \tau \in \mathcal{T}_h,
\end{aligned}
$$

which implies $|\int_I (c_i(\omega_n^8 z_h' - \chi), \partial_i z_h)_\Omega \,\mathrm{d}t| \le C\|\omega_n^4 z_h'\|_Q\|\omega_n^2 Dz_h\|_Q$ by inequality (20).

Hence, by inequality (20) again and elementary estimates,

$$\begin{aligned}\sum_{i=1}^{N}\left|\int_I (c_i(\omega_n^8 z_h' - \chi), \partial_i e)_\Omega \,\mathrm{d}t\right| &= \sum_{i=1}^{N}\Bigg|\int_I (c_i(\omega_n^8 z_h' - \chi), \partial_i z)_\Omega \\ &\qquad + (c_i(\omega_n^8 z_h' - \chi), \partial_i z_h)_\Omega \,\mathrm{d}t\Bigg| \\ &\le C\|\omega_n^4 z_h'\|_Q \|Dz\|_{Q_j'} + C\|\omega_n^4 z_h'\|_Q \|\omega_n^2 Dz_h\|_Q \\ &\le \frac{1}{16}\|\omega^4 e'\|_Q^2 + C\|z'\|_{Q_j'}^2 + C\|Dz\|_{Q_j'}^2 + C\|\omega^2 Dz_h\|_Q^2 \\ &\le \frac{1}{16}\|\omega_n^4 e'\|_Q^2 + C\|z'\|_{Q_j'}^2 + Cd_j^{-2}\|Dz\|_{Q_j'}^2 \\ &\quad + Cd_j^{-2}\|\omega_n^2 De\|_Q^2.\end{aligned}$$

Moreover, by inequality (21),

$$\begin{aligned}\left|\int_I a^{\#}(\omega^8 z_h' - \chi, z_h)\,\mathrm{d}t\right| &\le Cd_j^{-1}\sum_{\tau\in\mathcal{T}_h}\|\omega_n^6 z_h'\|_\tau \|Dz_h\|_\tau \le Cd_j^{-1}\sum_{\tau\in\mathcal{T}_h}\|\omega_n^4 z_h'\|_\tau \|\omega_n^2 Dz_h\|_\tau \\ &\le Cd_j^{-1}\left(\sum_{\tau\in\mathcal{T}_h}\|\omega_n^4 z_h'\|_\tau^2\right)^{\frac{1}{2}}\left(\sum_{\tau\in\mathcal{T}_h}\|\omega_n^2 Dz_h\|_\tau^2\right)^{\frac{1}{2}} \\ &\le \frac{1}{32}\|\omega_n^4 z_h'\|_Q^2 + Cd_j^{-2}\|\omega_n^2 Dz_h\|_Q^2.\end{aligned}$$

Using inequality (21) again, we thus obtain

$$\begin{aligned}\left|\int_I a^{\#}(\omega^8 z_h' - \chi, e)\,\mathrm{d}t\right| &\le \left|\int_I a^{\#}(\omega_n^8 z_h' - \chi, z_h) + a^{\#}(\omega_n^8 z_h' - \chi, z)\,\mathrm{d}t\right| \\ &\le \frac{1}{32}\|\omega_n^4 z_h'\|_Q^2 + Cd_j^{-2}\|\omega_n^2 Dz_h\|_Q^2 + Cd_j^{-1}\|\omega_n^6 z_h'\|_Q\|Dz\|_{Q_j'} \\ &\le \frac{1}{16}\|\omega_n^4 e'\|_Q^2 + C\|z'\|_{Q_j'}^2 + Cd_j^{-2}\|Dz\|_{Q_j'}^2 + Cd_j^{-2}\|\omega_n^2 De\|_Q^2.\end{aligned}$$

Summing $J_1 + J_2 + J_3$, we have

(27)
$$\begin{aligned}\|\omega_n^4 e'\|_Q + \frac{\mu}{2}\|(\omega_n^4 De)(T)\|_\Omega &\le C\|(De)(0)\|_{\Omega_j'}^2 + Cd_j^{-2}\|e(0)\|_{\Omega_j'}^2 + Cd_j^{-4}\|e\|_{Q_j'}^2 \\ &\quad + Chd^{-1}\|e'\|_{Q_j^n}^2 + C\|z\|_{Q_j',1,1}^2 + Cd_j^{-2}\|\omega_n^2 De\|_Q^2 \\ &\quad + Cd_j^{-2}\|(\omega_n^2 e)(T)\|_\Omega^2 + \frac{1}{2}\|\omega^4 e'\|_Q^2 + \frac{\mu}{4}\|(\omega_n^4 De)(T)\|_\Omega^2.\end{aligned}$$

Finally, subtracting the terms $\frac{1}{2}\|\omega^4 e'\|_Q^2$ and $\frac{\mu}{4}\|\omega^4 De(T)\|_\Omega$ on both sides and taking square roots yields (24).

Starting with the identity

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\omega^2 e\|_\Omega^2 + \operatorname{Re}\, a_{\omega^4}(e,e)\\
&\quad = \operatorname{Re}\, \big(\omega^4 e, e'\big)_\Omega + \frac{1}{2}\big((\omega^4)' e, e\big)_\Omega + \operatorname{Re}\, \big(a_{\omega^4}(z_h,e) - a_{\omega^4}(z,e) + a(\omega^4 z_h, e)\\
&\qquad - a(\omega^4 z_h, e)\big)\\
&\quad = \frac{1}{2}\big((\omega^4)' e, e\big)_\Omega + \operatorname{Re}\, \big(- \big(\omega^4 z, e'\big)_\Omega + a_{\omega^4}(z_h,e) - a_{\omega^4}(z,e) - a(\omega^4 z_h, e)\big)\\
&\qquad + \operatorname{Re}\, \big(\big(\omega^4 z_h, e'\big)_\Omega + a(\omega^4 z_h, e)\big)
\end{aligned}$$

and using (18) and (19), we similarly prove

$$\begin{aligned}
&\|\big(\omega^2 e\big)(T)\|_\Omega^2 - \|e(0)\|_{\Omega_j'}^2 + \|\omega^2 De\|_Q^2 + \|e\|_{Q_j'}^2\\
&\qquad \le C\epsilon^{-2} d_j^{-2}\|z\|_{Q_j',1,1}^2 + \epsilon^2 d_j^2 \|\omega^4 e'\|_Q^2 + C d_j^{-2}\|e\|_{Q_j'}^2\\
&\qquad\qquad + Ch^2\|e'\|_{Q_j^n}^2 + Chd^{-1}\|De\|_{Q_j^n}^2,\ \epsilon > 0.
\end{aligned}$$

Taking square roots and multiplying by d_j^{-1} yields (23). ∎

Appendix A. Truncation of singular integral operators. For $1 < p < \infty$ let $S \in \mathcal{L}(L_p(\mathbb{R}^N))$ be an integral operator with a kernel $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{C}$ satisfying

$$|K(x,y)| \le \frac{C}{|x-y|^N}, \quad x, y \in \mathbb{R}^N,\ x \ne y. \tag{28}$$

Kernel estimates of this form are well known for solutions to elliptic problems. For $\epsilon > 0$ set $\Omega_x^\epsilon := \{y \in \mathbb{R}^N : |x-y| < \epsilon\}$ and define $k_\epsilon : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{C}$ by $k_\epsilon(x,y) := \chi_{(\Omega_x^\epsilon)^c}(y) K(x,y)$ for $x, y \in \mathbb{R}^N$. The function k_ϵ is called *truncated kernel* of S. Since $\|k_\epsilon\|_{L_\infty(\mathbb{R}^N \times \mathbb{R}^N)} \le \frac{C}{\epsilon^N}$, the operator S_ϵ associated to the kernel k_ϵ is well defined on $L_p(\mathbb{R}^N)$. In [Ste93, Chapter I.7] it is shown that the family of operators $(S_\epsilon)_{\epsilon>0}$ is uniformly bounded in $\mathcal{L}(L_p(\mathbb{R}^N))$.

Let $Q := I \times \Omega$, where $\Omega \subset \mathbb{R}^N$ and $I := [0,\tau]$ for some $\tau > 0$. We will adapt this result to parabolic problems in Q. In order to do so, we adapt the kernel estimate stated above to parabolic problems. More precisely, we introduce the following definition.

Definition A.1. *Let $1 < p < \infty$. We say that $T \in \mathcal{L}(L_p(Q))$ satisfies the kernel estimate (KE_{Tr}) if there exists a measurable function $K : Q \times \Omega \to \mathbb{C}$ such that*

$$(Tf)(t,x) = \int_Q K(t-s,x,y) f(s,y)\ \mathrm{d}(s,y), \quad (t,x) \notin \operatorname{supp} f,\ \text{a.a. } (t,x),$$

for $f \in L_\infty(Q)$ and there exist constants $C, c > 0$ such that

$$|K(t,x,y)| \le C t^{-N-2} \exp\left(-c\frac{|x-y|^2}{t}\right), \quad 0 < t \le \tau,\ \text{a.a. } x, y \in \Omega. \tag{29}$$

If T satisfies the kernel estimate (KE_{Tr}) we set

$$Q_{t,x}^\epsilon := \{(s,y) \in Q :\ \max\{|t-s|^{\frac{1}{2}},\ |x-y|\} \le \epsilon\}$$

and

$$k_\epsilon(t,s,x,y) := K(t-s,x,y)\chi_{(Q_{t,x}^\epsilon)^c}(s,y), \quad (s,x),(t,y) \in Q,$$

where K denotes the kernel of T. Then for any $\epsilon > 0$ the truncated operator T_ϵ, given by

$$(T_\epsilon f)(t,x) := \int_Q k_\epsilon(t,s,x,y) f(s,y)\, \mathrm{d}(s,y), \quad (t,x) \in Q, \tag{30}$$

is well defined.

THEOREM A.2. *Let $1 < p < \infty$. Assume that $T \in \mathcal{L}(L_p(Q))$ satisfies the kernel estimate (KE_{Tr}). Then the family of the truncated operators $\left(T_\epsilon\right)_{\epsilon>0}$ is uniformly bounded in $\mathcal{L}(L_p(Q))$.*

Proof. Denote the kernel of T by K. Extending K by 0 outside $Q \times \Omega$, we restrict ourselves to the case $Q = \mathbb{R}_+ \times \mathbb{R}^N$. Since the rational numbers are dense in $\mathbb{R}$, there exists $\{(t_i, x_i)\}_{i\in\mathbb{N}}$ such that

$$Q = \bigcup_{i\in\mathbb{N}} Q^{\epsilon/8}_{t_i,x_i}.$$

We choose a maximal disjoint subset $\{Q^{\epsilon/8}_{t_{i_j},x_{i_j}}\}_{j\in\mathbb{N}}$ of $\{Q^{\epsilon/8}_{t_i,x_i}\}_{i\in\mathbb{N}}$. This implies

$$\bigcup_{j\in\mathbb{N}} Q^{\epsilon/4}_{t_{i_k},x_{i_k}} = Q.$$

For $\tilde{T}_\epsilon := T - T_\epsilon$ we show that there exists $C > 0$ such that

$$\left\| \chi_{Q^{\epsilon/4}_{t_{i_j},x_{i_j}}} \tilde{T}_\epsilon f \right\| L_p(Q) \le C \left\| \chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}} f \right\| L_p(Q), \quad f \in L_p(Q),\ j \in \mathbb{N},\ \epsilon > 0. \tag{31}$$

Let $f \in L_p(Q)$. Note that

$$(\tilde{T}_\epsilon f)(t,x) \equiv 0 \qquad \text{if } \operatorname{supp} f \subset \left(Q^\epsilon_{t,x}\right)^c, \tag{32}$$

$$(\tilde{T}_\epsilon f)(t,x) = (Tf)(t,x) \qquad \text{if } \operatorname{supp} f \subset Q^\epsilon_{t,x}, \tag{33}$$

for a.a. $(t,x) \in Q$. Following [Ste93, Chapter I.7], we write

$$\begin{aligned}\left(\chi_{Q^{\epsilon/4}_{t_{i_j},x_{i_j}}} \tilde{T}_\epsilon f\right)(t,x) &= \left(\chi_{Q^{\epsilon/4}_{t_{i_j},x_{i_j}}} \tilde{T}_\epsilon \chi_{Q^{\epsilon/2}_{t_{i_j},x_{i_j}}} f\right)(t,x) + \left(\chi_{Q^{\epsilon/4}_{t_{i_j},x_{i_j}}} \tilde{T}_\epsilon \chi_{(Q^{2\epsilon}_{t_{i_j},x_{i_j}})^c} f\right)(t,x) \\ &\quad + \left(\chi_{Q^{\epsilon/4}_{t_{i_j},x_{i_j}}} \tilde{T}_\epsilon \left(\chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}} - \chi_{Q^{\epsilon/2}_{t_{i_j},x_{i_j}}}\right) f\right)(t,x) \\ &=: I^\epsilon_{j,1}(t,x) + I^\epsilon_{j,2}(t,x) + I^\epsilon_{j,3}(t,x), \quad \text{a.a. } (t,x) \in Q,\ j \in \mathbb{N},\ \epsilon > 0.\end{aligned}$$

By (33), we obtain

$$\begin{aligned}\|I^\epsilon_{j,1}\|_{L_p(Q)} &\le \left\| \tilde{T} \chi_{Q^{\epsilon/2}_{t_{i_j},x_{i_j}}} f \right\| L_p(Q) \le \|T\|_{\mathcal{L}(L_p(Q))} \left\| \chi_{Q^{\epsilon/2}_{t_{i_j},x_{i_j}}} f \right\| L_p(Q) \\ &\le C \left\| \chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}} f \right\| L_p(Q), \quad f \in L_p(Q),\ j \in \mathbb{N},\ \epsilon > 0.\end{aligned}$$

Since $\operatorname{supp} \chi_{(Q^{2\epsilon}_{t_{i_j},x_{i_j}})^c} f \subset (Q^{2\epsilon}_{t_{i_j},x_{i_j}})^c \subset (Q^\epsilon_{t_{i_j},x_{i_j}})^c$, it follows by (32) that $I^\epsilon_{j,2}(t,x) = 0,\ (t,x) \in Q,\ j \in \mathbb{N},\ \epsilon > 0$.

We use the representation (30) and the kernel estimate (KE_{Tr}) to estimate $\|I^\epsilon_{j,3}\|_{L_p(Q)}$. By definition of $Q^\epsilon_{t,x}$, there exists $C>0$, independent of $j\in\mathbb{N}$ and ϵ, such that

$$|K(t-s,x,y)|\le C\left(\frac{\epsilon}{4}\right)^{-N-2},\quad (s,y)\in Q^{2\epsilon}_{t_{i_j},x_{i_j}}\setminus Q^{\epsilon/2}_{t_{i_j},x_{i_j}}, (t,x)\in Q^{\epsilon/4}_{t_{i_j},x_{i_j}}.$$

Since $\operatorname{supp}\chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}\setminus Q^{\epsilon/2}_{t_{i_j},x_{i_j}}} f\subset Q^{2\epsilon}_{t_{i_j},x_{i_j}}$, we obtain

$$\begin{aligned}\|I^\epsilon_{j,3}\|_{L_p(Q)}&\le \left|Q^{\epsilon/4}_{t_{i_j},x_{i_j}}\right|^{\frac1p}\|I^\epsilon_{j,3}\|_{L_\infty(Q)}\\&\le C\left|Q^{\epsilon/4}_{t_{i_j},x_{i_j}}\right|^{\frac1p}\left|Q^{2\epsilon}_{t_{i_j},x_{i_j}}\right|^{1-\frac1p}\epsilon^{-N-2}\left\|\chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}}f\right\|L_p(Q)\\&\le C\left\|\chi_{Q^{2\epsilon}_{t_{i_j},x_{i_j}}}f\right\|L_p(Q),\quad f\in L_p(Q),\ j\in\mathbb{N},\ \epsilon>0.\end{aligned}$$

Therefore, inequality (31) is proved.

We finally obtain

$$\|\tilde T_\epsilon f\|^p_{L_p(Q)}\le C\sum_{j\in\mathbb{N}}\|\tilde T_\epsilon f\|^p_{L_p(Q^{\epsilon/4}_{t_{i_j},x_{i_j}})}\le C\sum_{j\in\mathbb{N}}\|f\|^p_{L_p(Q^{2\epsilon}_{t_{i_j},x_{i_j}})}.$$

To conclude the proof, we will show that no $(t,x)\in Q$ belongs to more than $M\in\mathbb{N}$ of the $Q^{2\epsilon}_{t_{i_j},x_{i_j}}$, where M does not depend on ϵ. This would imply

$$\sum_{j\in\mathbb{N}}\|f\|^p_{L_p(Q^{2\epsilon}_{t_{i_j},x_{i_j}})}\le M\|f\|^p_{L_p(Q)}.$$

Let $(t,x)\in Q$. Since $Q^{2\epsilon}_{t_{i_j},x_{i_j}}\subset Q^{4\epsilon}_{t,x}$ for all (t_{i_j},x_{i_j}) with $(t,x)\in Q^{2\epsilon}_{t_{i_j},x_{i_j}}$, it suffices to estimate the maximal number of $Q^{\epsilon/8}_{t_{i_j},x_{i_j}}$ in $Q^{4\epsilon}_{t,x}$, independent of $(t,x)\in Q$ and ϵ. Now, the maximal number of $Q^{\epsilon/8}_{t_{i_j},x_{i_j}}$ in $Q^{4\epsilon}_{t,x}$ is bounded by

$$\frac{|Q^{4\epsilon}_{t,x}|}{|Q^{\epsilon/8}_{t,x}|}=\frac{(4\epsilon)^{N+2}}{(\frac{\epsilon}{8})^{N+2}}=32^{N+2},\quad (t,x)\in Q,\ \epsilon>0,$$

since the $Q^{\epsilon/8}_{t_{i_j},x_{i_j}}$ are mutually disjoint. The proof is complete. ∎

Appendix B. Kernel estimates of analytic semigroups. Analytic semigroups satisfying a heat kernel estimate allow us to prove estimates for time derivatives of their kernels as well. The proof of the following lemma uses techniques due to E. B. Davies (see [Dav97]).

LEMMA B.1. *Let $(T(t))_{t\ge0}$ satisfy (KE_{Max}). Then*

$$|\partial_t^k\partial_1^\alpha\partial_2^\beta K(t,x,y)|\le Ct^{-\frac{N+2k+|\alpha|+|\beta|}{2}}\exp\left(-c\frac{|x-y|^2}{t}\right),\quad 0<t\le T,\ x,y\in\Omega,$$

whenever $0\le k\le1, |\alpha|=0, |\beta|\le2$ or $0\le k\le1, |\alpha|\le2, |\beta|=0$.

Proof. Writing $D_1^\alpha D_2^\beta T(z) = (D_1^\alpha D_2^\beta T(\operatorname{Re}\frac{z}{2}))T(z - \operatorname{Re}\frac{z}{2})$, we obtain

$$\begin{aligned}\|D_1^\alpha D_2^\beta T(z)\|_{\mathcal{L}(L_1(\Omega),L_\infty(\Omega))} &\le C(\operatorname{Re} z)^{-\frac{N+|\alpha|+|\beta|}{2}} \|T(z - \operatorname{Re}\tfrac{z}{2})\|_{\mathcal{L}(L_1(\Omega))} \\ &\le C(\operatorname{Re} z)^{-\frac{N+|\alpha|+|\beta|}{2}}\end{aligned}$$

for $z \in \Sigma_{\theta,2T} := \{z \in \mathbb{C} : |\arg z| < \theta,\ |z| < 2T\}$ and some $\theta \in (0, \frac{\pi}{2})$. Thus $D_1^\alpha D_2^\beta T(z)$ is an integral operator for $z \in \Sigma_{\theta,2T}$. Since $D_1^\alpha D_2^\beta T : \Sigma_{\theta,2T} \to \mathcal{L}(L_2(\Omega))$ is analytic, by [AB94, Theorem 3.1], there exists $K^{\alpha,\beta}(\cdot, x, y) : \Sigma_{\theta,2T} \to \mathbb{C}$, analytic for $x, y \in \Omega$, such that $K^{\alpha,\beta}(z, \cdot, \cdot)$ is the kernel of $D_1^\alpha D_2^\beta T(z)$. Moreover, by [DR96, Proposition 3.3], which is a variant of [Dav97, Theorem 4], we obtain

$$|K^{\alpha,\beta}(z, x, y)| \le C(\operatorname{Re} z)^{-\frac{N+|\alpha|+|\beta|}{2}} \exp\left(-c\frac{|x-y|^2}{|z|}\right), \quad z \in \Sigma_{\theta,T},\ x, y \in \Omega. \tag{34}$$

Choosing $r = \min\{\frac{1}{2}, \tan\theta\}$, by Cauchy's theorem, we have

$$\partial_t K^{\alpha,\beta}(t, x, y) = \frac{1}{2\pi i} \int_{B(t,rt)} \frac{K^{\alpha,\beta}(z, x, y)}{(z-t)^2}\, \mathrm{d}z, \quad 0 < t < T,$$

which implies (see [Dav97, Theorem 3])

$$\partial_t K^{\alpha,\beta}(t, x, y)| \le Ct^{-\frac{N+|\alpha|+|\beta|+2}{2}} \exp\left(-c\frac{|x-y|^2}{t}\right), \quad z \in \Sigma_{\theta,T},\ x, y \in \Omega.$$

Since $\partial_t K^{\alpha,\beta}(t, \cdot, \cdot)$ is the kernel of $\frac{d}{dt}T(t)$, the proof is complete. ∎

Acknowledgments. The author would like to thank Matthias Hieber for many discussions and suggestions to improve this paper. Very special thanks go to Stig Larsson for the idea for the present paper which was born during a research visit in Gothenburg.

REFERENCES

[AB94] W. Arendt and A. V. Bukhvalov, *Integral representations of resolvents and semigroups*, Forum Math., 6 (1994), pp. 111–135.

[Ama95] H. Amann, *Linear and Quasilinear Parabolic Problems,* Vol. I, Birkhäuser Boston, Boston, MA, 1995.

[BTW03] N. Y. Bakaev, V. Thomée, and L. B. Wahlbin, *Maximum-norm estimates for resolvents of elliptic finite element operators*, Math. Comp., 72 (2003), pp. 1597–1610 (electronic).

[CLT94] M. Crouzeix, S. Larsson, and V. Thomée, *Resolvent estimates for elliptic finite element operators in one dimension*, Math. Comp., 63 (1994), pp. 121–140.

[Dav97] E. B. Davies, *Non-Gaussian aspects of heat kernel behaviour*, J. London Math. Soc. (2), 55 (1997), pp. 105–125.

[DHP03] R. Denk, M. Hieber, and J. Prüss, *$\mathcal{R}$-Boundedness, Fourier Multipliers and Problems of Elliptic and Parabolic Type*, Mem. Amer. Math. Soc. 166, Providence, RI, 2003.

[Dor93] G. Dore, *L^p regularity for abstract differential equations*, in Functional Analysis and Related Topics, 1991 (Kyoto), Springer, Berlin, 1993, pp. 25–38.

[DR96] X. T. Duong and D. W. Robinson, *Semigroup kernels, Poisson bounds, and holomorphic functional calculus*, J. Funct. Anal., 142 (1996), pp. 89–128.

[ÈdI70] S. D. Èĭ del'man and S. D. Ivasišen, *Investigation of the Green's matrix of a homogeneous parabolic boundary value problem*, Trudy Moskov. Mat. Obšč., 23 (1970), pp. 179–234.

[Gei03] M. Geissert, *Maximal L_p-L_q Regularity for Discrete Elliptic Differential Operators*, Ph.D. thesis, TU Darmstadt, Germany, 2003.

[Gei04] M. Geissert, *Application of Discrete Maximal L_p Regularity for Finite Element Operators*, preprint, TU Darmstadt, Germany, 2004.

[HP97] M. Hieber and J. Prüss, *Heat kernels and maximal L^p-L^q estimates for parabolic evolution equations*, Comm. Partial Differential Equations, 22 (1997), pp. 1647–1669.

[Ste93] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[STW98] A. H. Schatz, V. Thomée, and L. B. Wahlbin, *Stability, analyticity, and almost best approximation in maximum norm for parabolic finite element equations*, Comm. Pure Appl. Math., 51 (1998), pp. 1349–1385.

[SZ90] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[Tho97] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997.

[TW00] V. Thomée and L. B. Wahlbin, *Stability and analyticity in maximum-norm for simplicial Lagrange finite element semidiscretizations of parabolic equations with Dirichlet boundary conditions*, Numer. Math., 87 (2000), pp. 373–389.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 699–731

OPTIMIZED SCHWARZ METHODS*

MARTIN J. GANDER†

Abstract. Optimized Schwarz methods are a new class of Schwarz methods with greatly enhanced convergence properties. They converge uniformly faster than classical Schwarz methods and their convergence rates dare asymptotically much better than the convergence rates of classical Schwarz methods if the overlap is of the order of the mesh parameter, which is often the case in practical applications. They achieve this performance by using new transmission conditions between subdomains which greatly enhance the information exchange between subdomains and are motivated by the physics of the underlying problem. We analyze in this paper these new methods for symmetric positive definite problems and show their relation to other modern domain decomposition methods like the new Finite Element Tearing and Interconnect (FETI) variants.

Key words. optimized Schwarz methods, optimized transmission conditions, domain decomposition, parallel preconditioning

AMS subject classifications. 65N55, 65F10, 65N22.

DOI. 10.1137/S0036142903425409

1. Introduction.

The convergence properties of the classical Schwarz methods are well understood for a wide variety of problems; see, for example, the books [37], [35] or the survey articles [4], [40], [41] and references therein. Over the last decade, people have looked at different transmission conditions for the classical Schwarz method. There were three main motivations: the first motivation for different transmission conditions came from the nonoverlapping variant of the Schwarz method proposed by Lions, since without overlap the classical Schwarz method does not converge. Lions proposed to use Robin conditions to obtain a convergent algorithm in [31]. At the end in his paper, we find the following remark: "First of all, it is possible to replace the constants in the Robin conditions by two proportional functions on the interface, or even by local or nonlocal operators." Lions then gives a simple example in one dimension and shows that the optimal choice for the parameters in the Robin transmission conditions of the algorithm are constants in that case. In higher dimensions, however, the optimal choice involves a nonlocal transmission operator, as was shown for a two-dimensional convection diffusion problem by Charton, Nataf, and Rogier in [5], where a parabolic factorization of the operator was used to derive the optimal transmission conditions. Since nonlocal operators are not convenient to implement and costly ("ils se prêtent peu au calcul numérique" [5]), the authors propose for the convection dominated convection-diffusion problem to expand the symbols of the nonlocal operators in the small viscosity parameter to obtain local approximations. A different approximation using a Taylor expansion in the frequency parameter to obtain local transmission conditions for the convection diffusion equation is proposed in [33]; see also [34] and [32]. For symmetric coercive problems, a formulation of the nonoverlapping Schwarz method with Robin transmission conditions which avoids the explicit use of normal derivatives was introduced independently in [7], and convergence of the resulting algorithm was proved using energy estimates. A first optimization of the

*Received by the editors March 31, 2003; accepted for publication (in revised form) November 15, 2005; published electronically March 31, 2006.
http://www.siam.org/journals/sinum/44-2/42540.html
†Section de Mathématiques, Université de Genève, 2-4 rue du Lièvre, CP 64, CH-1211 Genève, Switzlerland (Martin.Gander@math.unige.ch).

transmission conditions for the performance of the algorithm was done by Japhet in [26] for convection diffusion problems, where one coefficient in a second order transmission condition was optimized, which led to the first optimized Schwarz method in this context. This approach was further developped and refined in [29], [27], and [28] for convection diffusion problems.

The second motivation for changing the transmission conditions came from acoustics. For problems of Helmholtz type, the classical Schwarz algorithm is not convergent, even with overlap. Després therefore proposed in [8] to use radiation conditions instead for the Helmholtz equation and proved convergence of a nonoverlapping variant of the Schwarz algorithm with these transmission conditions using energy estimates; see also [9]. The radiation conditions used were again Robin conditions, and the same conditions were also used in an overlapping context in [3]. Higher order local transmission conditions for the Helmholtz equation were introduced in [6] and a first attempt was made to optimize the free parameter in the transmission conditions for the performance of the algorithm, leading to the first optimized Schwarz method without overlap for the Helmholtz equation. The optimization problem for the Robin transmission conditions was then completely solved for this case in [22] and a simple strategy to optimize the second order transmission conditions was also presented. For a complete optimization of the second order transmission conditions for Helmholtz problems, see [14] for the case without overlap and [21] for the case with overlap.

The third motivation was that the convergence rate of the classical Schwarz method is rather slow and very much dependent on the size of the overlap. In a short note on nonlinear problems [24], Hagstrom, Tewarson, and Jazcilevich introduced Robin transmission conditions between subdomains and suggested, "Indeed, we advocate the use of nonlocal conditions." Later and independently, Tang introduced in [39] the generalized Schwarz alternating method, which uses a weighted average of Dirichlet and Neumann conditions at the interfaces, which is equivalent to a Robin condition. Numerically, optimal values for the weighting parameter were determined, and it was shown that a good choice of the parameter leads to a significant speedup of the algorithm. The main difficulty remaining in this approach is the determination of these parameters on the interfaces, like for successive overrelaxation (SOR) methods. Even stronger coupling was proposed in [38], where the authors introduced the overdetermined Schwarz algorithm, which enforces the coupling not only on the interfaces but also in the overlap itself, in so-called artificial boundary layers, and the relaxation parameter is now a function depending on space, as proposed earlier by Lions [31]. But the link with absorbing boundary conditions was only made later in [10], where an overlapping version of the Schwarz algorithm for Laplace's equation was analyzed with Robin and second order transmission conditions and a first attempt was made to determine asymptotically optimal parameters. In the waveform relaxation community, a link made with Schwarz methods in [23] opened up the way for better transmission conditions in the Schwarz waveform relaxation algorithms; see [19]. This led to optimized Schwarz algorithms for evolution problems, where one can clearly see that the optimal transmission conditions are absorbing boundary conditions. For the case of the wave equation with discontinuous coefficients, a nonoverlapping optimized Schwarz method is introduced and analyzed in detail at both the continuous and the discrete level in [20]. For the heat equation, see [17].

Optimized Schwarz methods have several key features:

1. They converge necessarily faster than classical Schwarz methods, at the same cost per iteration.

2. There are simple optimization procedures to determine the best parameters to be used in the transmission conditions, sometimes even closed formulas, depending on the problem solved.
3. Classical Schwarz implementations need only a small change in the implementation, in the information exchange routine, to benefit from the additional performance.
4. Optimized Schwarz methods can be used with or without overlap.

We present here a complete analysis of optimized Schwarz methods for a symmetric positive definite model problem and analyze in detail the optimization problems and the asymptotic performance of different approximations to the optimal transmission conditions. We restrict our analysis to the simple case of two subdomains, because optimized Schwarz method are greatly enhancing the local coupling between subdomains. Once optimized coupling conditions are found, they can be used in the general context of many subdomains, as we show with numerical examples at the end (see also [22]). As for classical Schwarz methods, a coarse grid is necessary as soon as many subdomains are used, if a convergence rate independent of the number of subdomains is desired, but we do not consider this issue here.

2. The classical Schwarz algorithm for a model problem. We use throughout the paper the model problem

$$\mathcal{L}(u) = (\eta - \Delta)(u) = f \quad \text{on } \Omega = \mathbb{R}^2,\ \eta > 0, \tag{2.1}$$

where we require the solution to decay at infinity. To introduce the ideas behind optimized Schwarz methods, we start by analyzing a parallel variant of the classical alternating Schwarz method introduced by Lions [30], applied to the model problem (2.1). We decompose the domain Ω into the two overlapping subdomains

$$\Omega_1 = (-\infty, L) \times \mathbb{R}, \quad \Omega_2 = (0, \infty) \times \mathbb{R}. \tag{2.2}$$

The Jacobi–Schwarz method for the two subdomains and the model problem is then given by

$$\begin{array}{llll} (\eta - \Delta)u_1^n = f & \text{in } \Omega_1, & (\eta - \Delta)u_2^n = f & \text{in } \Omega_2, \\ u_1^n(L, y) = u_2^{n-1}(L, y), & y \in \mathbb{R}, & u_2^n(0, y) = u_1^{n-1}(0, y), & y \in \mathbb{R}, \end{array} \tag{2.3}$$

and we require the iterates to decay at infinity. By linearity it suffices to consider only the case $f = 0$ and analyze convergence to the zero solution. Our analysis is based on the Fourier transform,

$$\hat{f}(k) = \mathcal{F}(f) := \int_{-\infty}^{\infty} e^{-ikx} f(x)dx, \quad f(x) = \mathcal{F}^{-1}(\hat{f}) := \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} \hat{f}(k)dk, \quad k \in \mathbb{R}. \tag{2.4}$$

Taking a Fourier transform of the Schwarz algorithm (2.3) in the y direction, and using the property of the Fourier transform that derivatives in y become multiplications by ik, we obtain

$$\begin{array}{rlrl} (\eta + k^2 - \partial_{xx})\hat{u}_1^n = 0, & x < L,\ k \in \mathbb{R}, & (\eta + k^2 - \partial_{xx})\hat{u}_2^n = 0, & x > 0,\ k \in \mathbb{R}, \\ \hat{u}_1^n(L, k) = \hat{u}_2^{n-1}(L, k), & k \in \mathbb{R}, & \hat{u}_2^n(0, k) = \hat{u}_1^{n-1}(0, k), & k \in \mathbb{R}. \end{array} \tag{2.5}$$

Hence subdomain solutions in the Fourier transformed domain are of the form

$$\hat{u}_j^n(x, k) = A_j(k)e^{\lambda_1(k)x} + B_j(k)e^{\lambda_2(k)x}, \quad j = 1, 2, \tag{2.6}$$

where the $\lambda_j(k)$, $j = 1, 2$ satisfy the characteristic equation $\eta - \lambda_j^2 + k^2 = 0$ and hence $\lambda_1(k) = \sqrt{k^2+\eta}$ and $\lambda_2(k) = -\sqrt{k^2+\eta}$. By the condition on the iterates at infinity, we obtain for the subdomain solutions

$$\hat{u}_1^n(x,k) = \hat{u}_2^{n-1}(L,k)e^{\sqrt{k^2+\eta}\,(x-L)}, \qquad \hat{u}_2^n(x,k) = \hat{u}_1^{n-1}(0,k)e^{-\sqrt{k^2+\eta}\,x}.$$

Inserting these solutions into algorithm (2.5), we obtain by induction

$$\hat{u}_1^{2n}(0,k) = \rho_{cla}^n \hat{u}_1^0(0,k), \qquad \hat{u}_2^{2n}(L,k) = \rho_{cla}^n \hat{u}_2^0(L,k), \tag{2.7}$$

where the convergence factor $\rho_{cla}(k,\eta,L)$ of the classical Schwarz algorithm is given by

$$\rho_{cla} = \rho_{cla}(k,L,\eta) := e^{-2\sqrt{k^2+\eta}L} < 1 \quad \forall k \in \mathbb{R}. \tag{2.8}$$

Note that we have chosen here to define the convergence factor over two iterations, which would correspond to one iteration of the Gauss–Seidel–Schwarz method in this two-subdomain case. From (2.8), we see that the iterates converge to zero on the line $x = 0$ and $x = L$. Since with zero boundary conditions the solution vanishes identically, we have shown that the classical Schwarz method converges for all frequencies, provided $\eta > 0$. The convergence factor depends on the problem parameter η, the size of the overlap L, and the frequency parameter k: the top curve in Figure 4.1 shows the dependence of ρ_{cla} on k for an overlap $L = \frac{1}{100}$ and $\eta = 1$. One can see that the Schwarz algorithm is a smoother; it damps high frequencies effectively, whereas for low frequencies the convergence factor is close to one and hence the algorithm is very slow.

3. The optimal Schwarz algorithm. We now introduce the key modification in the classical Schwarz method: new transmission conditions between the subdomains. The new algorithm is given by

$$\begin{array}{cc} (\eta-\Delta)u_1^n = f \quad \text{in } \Omega_1, & (\eta-\Delta)u_2^n = f \quad \text{in } \Omega_2, \\ (\partial_x + \mathcal{S}_1)(u_1^n)(L,\cdot) = (\partial_x + \mathcal{S}_1)(u_2^{n-1})(L,\cdot), & (\partial_x + \mathcal{S}_2)(u_2^n)(0,\cdot) = (\partial_x + \mathcal{S}_2)(u_1^{n-1})(0,\cdot), \end{array} \tag{3.1}$$

where $\mathcal{S}_j$, $j = 1, 2$, are linear operators along the interface in the y direction which we will determine in what follows to get the best possible performance of the new Schwarz algorithm. As for the classical Schwarz method, taking a Fourier transform in the y direction for $f = 0$, we obtain

$$\begin{array}{c} \eta\hat{u}_1^n - \partial_{xx}\hat{u}_1^n + k^2\hat{u}_1^n = 0, \quad x < L,\ k \in \mathbb{R}, \\ (\partial_x + \sigma_1(k))(\hat{u}_1^n)(L,k) = (\partial_x + \sigma_1(k))(\hat{u}_2^{n-1})(L,k), \quad k \in \mathbb{R}, \end{array} \tag{3.2}$$

where $\sigma_1(k)$ denotes the symbol of the operator $\mathcal{S}_1$, and

$$\begin{array}{c} \eta\hat{u}_2^n - \partial_{xx}\hat{u}_2^n + k^2\hat{u}_2^n = 0, \quad x > 0,\ k \in \mathbb{R}, \\ (\partial_x + \sigma_2(k))(\hat{u}_2^n)(0,k) = (\partial_x + \sigma_2(k))(\hat{u}_1^{n-1})(0,k), \quad k \in \mathbb{R}, \end{array} \tag{3.3}$$

where $\sigma_2(k)$ is the symbol of $\mathcal{S}_2$. The solutions on the subdomains are again of the form (2.6), and using the condition on the iterates at infinity, the transmission conditions, and the fact that

$$\frac{\partial\hat{u}_1^n}{\partial x} = \sqrt{k^2+\eta}\,\hat{u}_1^n, \qquad \frac{\partial\hat{u}_2^n}{\partial x} = -\sqrt{k^2+\eta}\,\hat{u}_2^n,$$

we find the subdomain solution in Fourier space to be

$$\hat{u}_1^n(x,k) = \frac{\sigma_1(k) - \sqrt{k^2+\eta}}{\sigma_1(k) + \sqrt{k^2+\eta}} e^{\sqrt{k^2+\eta}(x-L)} \hat{u}_2^{n-1}(L,k),$$
$$\hat{u}_2^n(x,k) = \frac{\sigma_2(k) + \sqrt{k^2+\eta}}{\sigma_2(k) - \sqrt{k^2+\eta}} e^{-\sqrt{k^2+\eta}\,x} \hat{u}_1^{n-1}(0,k).$$

Inserting these solutions into algorithm (3.1), we obtain by induction

$$\hat{u}_1^{2n}(0,k) = \rho_{opt}^n \hat{u}_1^0(0,k), \qquad \hat{u}_2^{2n}(L,k) = \rho_{opt}^n \hat{u}_2^0(L,k), \tag{3.4}$$

where the new convergence factor ρ_{opt} is given by

$$\rho_{opt} = \rho_{opt}(k, L, \eta, \sigma_1, \sigma_2) := \frac{\sigma_1(k) - \sqrt{k^2+\eta}}{\sigma_1(k) + \sqrt{k^2+\eta}} \cdot \frac{\sigma_2(k) + \sqrt{k^2+\eta}}{\sigma_2(k) - \sqrt{k^2+\eta}} e^{-2\sqrt{k^2+\eta}L}. \tag{3.5}$$

The only difference between the new convergence factor ρ_{opt} and the one of the classical Schwarz method, ρ_{cla} given in (2.8), is the factor in front of the exponential. But this factor has a tremendous influence on the performance of the method: choosing for the symbols

$$\sigma_1(k) := \sqrt{k^2+\eta}, \quad \sigma_2(k) := -\sqrt{k^2+\eta}, \tag{3.6}$$

the new convergence factor vanishes identically, $\rho_{opt} \equiv 0$, and the algorithm converges in two iterations, independently of the initial guess, the overlap L, and the problem parameter η. This is an optimal result, since the solution in one subdomain depends on the forcing function f in the other subdomain and hence a first solve is necessary to incorporate the influence of f into the subdomain solution, then one information exchange is performed to give this information to the neighboring subdomain and a second solve on the subdomains incorporates this information into the new subdomain solution. Convergence in less than two steps is not possible. One can also see from (3.6) that the optimal choice depends on the problem. The optimal convergence result for two subdomains in two iterations can be generalized to $J > 2$ subdomains and convergence in J iterations (see, for example, [33] or [16]), provided the subdomains are arranged in a sequence. In addition, with this choice of σ_j, the exponential factor in the convergence factor becomes irrelevant and we can have Schwarz methods without overlap.

To use the optimal choice of σ_j in practice, we need to back-transform the transmission conditions involving σ_1 and σ_2 from the Fourier domain into the physical domain to obtain the transmission operators $\mathcal{S}_1$ and $\mathcal{S}_2$. Hence we need

$$\mathcal{S}_1(u_1^n) = \mathcal{F}_k^{-1}(\sigma_1 \hat{u}_1^n), \qquad \mathcal{S}_2(u_2^n) = \mathcal{F}_k^{-1}(\sigma_2 \hat{u}_2^n), \tag{3.7}$$

and thus for the optimal choice of σ_j we have to evaluate a convolution in each step of the algorithm, because the σ_j contain a square root and thus the optimal $\mathcal{S}_j$ are nonlocal operators, as advocated in [24]. If the symbols σ_j were, however, polynomials in ik, then the operators $\mathcal{S}_j$ would consist of derivatives in y and thus be local operators. We will therefore approximate the optimal choice of σ_j by polynomials in the following sections, which leads to the new class of optimized Schwarz methods.

4. Optimized Schwarz algorithms. We approximate the symbols of the optimal transmission conditions found in (3.6) by polynomial symbols in ik which corresponds to local approximations. We choose polynomials of degree two here,

$$\sigma_1^{app}(k) = p_1 + q_1 k^2, \quad \sigma_2^{app}(k) = -p_2 - q_2 k^2. \tag{4.1}$$

Note that we do not consider a first order term, because the operator of the underlying problem is self-adjoint. Higher order approximations would be possible as well, as long as the subdomain problems remain well posed. With the approximation (4.1), the convergence factor of the optimized Schwarz algorithm becomes

$$\rho = \rho(k, L, \eta, p_1, p_2, q_1, q_2) = \frac{\sqrt{k^2+\eta} - p_1 - q_1 k^2}{\sqrt{k^2+\eta} + p_1 + q_1 k^2} \cdot \frac{\sqrt{k^2+\eta} - p_2 - q_2 k^2}{\sqrt{k^2+\eta} + p_2 + q_2 k^2} e^{-2\sqrt{k^2+\eta}L}. \tag{4.2}$$

THEOREM 4.1. *The optimized Schwarz method* (3.1) *with transmission conditions defined by the symbols* (4.1) *converges for* $p_j > 0$, $q_j \geq 0$, $j = 1, 2$, *faster than the classical Schwarz method* (2.3), $|\rho_{opt}(k)| < |\rho_{cla}(k)|$ *for all* k.

Proof. The only difference between ρ_{cla} in (2.8) and ρ_{opt} in (4.2) is the additional factor in front of the exponential, which satisfies for $p_j > 0$ and $q_j \geq 0$

$$\left| \frac{\sqrt{k^2+\eta} - p_1 - q_1 k^2}{\sqrt{k^2+\eta} + p_1 + q_1 k^2} \cdot \frac{\sqrt{k^2+\eta} - p_2 - q_2 k^2}{\sqrt{k^2+\eta} + p_2 + q_2 k^2} \right| < 1 \quad \forall k,$$

and hence $|\rho(k)| < |\rho_{cla}(k)|$ for all k. ∎

The goal of optimized Schwarz methods is now to choose the free parameters $p_j, q_j \geq 0$ for $j = 1, 2$ to further improve the performance of the method.

4.1. Low-frequency approximations. As we have seen, the classical Schwarz method is effective, due to the overlap, for high frequencies but ineffective for low frequencies. The low frequencies can, however, be treated in the new Schwarz algorithm with the transmission conditions: expanding the symbols $\sigma_j(k)$ of the optimal operators $\mathcal{S}_j$ in a Taylor series, we find

$$\sigma_1(k) = \sqrt{\eta} + \frac{1}{2\sqrt{\eta}} k^2 + O(k^4), \qquad \sigma_2(k) = -\sqrt{\eta} - \frac{1}{2\sqrt{\eta}} k^2 + O(k^4), \tag{4.3}$$

and hence a second order Taylor approximation would lead to the values $p_1 = p_2 = \sqrt{\eta}$, $q_1 = q_2 = \frac{1}{2\sqrt{\eta}}$, whereas a zeroth order approximation could be obtained by setting $q_1 = q_2 = 0$ for the same values of p_j. The corresponding optimized Schwarz methods have the convergence factors

$$\begin{aligned} \rho_{T0}(k, L, \eta) &= \left(\frac{\sqrt{k^2+\eta} - \sqrt{\eta}}{\sqrt{k^2+\eta} + \sqrt{\eta}} \right)^2 e^{-2\sqrt{k^2+\eta}L}, \\ \rho_{T2}(k, L, \eta) &= \left(\frac{\sqrt{k^2+\eta} - \sqrt{\eta} - \frac{1}{2\sqrt{\eta}} k^2}{\sqrt{k^2+\eta} + \sqrt{\eta} + \frac{1}{2\sqrt{\eta}} k^2} \right)^2 e^{-2\sqrt{k^2+\eta}L}, \end{aligned} \tag{4.4}$$

where we used the index $T0$ to denote a Taylor approximation of order zero and $T2$ to denote a Taylor approximation of order two of the optimal symbol in the transmission condition. Figure 4.1 shows on the left the convergence factors obtained with this

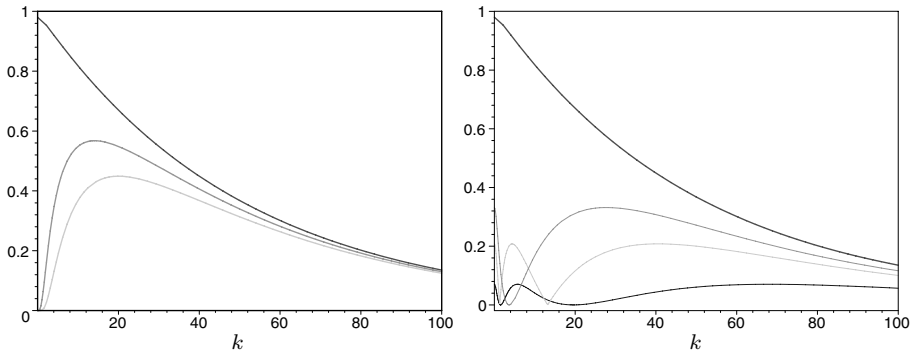


FIG. 4.1. *Convergence factor ρ_{cla} of the classical Schwarz method (top curve) as a function of k, compared on the left to ρ_{T0} (middle curve) and ρ_{T2} (bottom curve) of the optimized Schwarz methods with zeroth and second order transmission conditions, respectively, obtained by Taylor expansion, and on the right compared to the OO0 and OO2 Schwarz methods, and the optimized Schwarz method with two-sided optimized Robin transmission conditions, which lies in between OO0 and OO2.*

choice of transmission conditions for the model problem with two subdomains, overlap $L = \frac{1}{100}$ and problem parameter $\eta = 1$, together with the classical convergence factor ρ_{cla}. First one can clearly see that the optimized Schwarz methods are uniformly better than the classical Schwarz method; in particular the low-frequency behavior is greatly improved. The maximum of the convergence factor of classical Schwarz is about 0.980, whereas the maximum of the convergence factor with zeroth order Taylor condition is 0.568 and the maximum with second order Taylor condition is 0.449 in this example. Hence the classical Schwarz method needs about 28 iterations to obtain the contraction factor of one iteration of the optimized Schwarz method with zeroth order Taylor conditions, and about 40 iterations are needed to obtain the contraction of one iteration of the optimized Schwarz method with second order transmission conditions from Taylor expansion.

As we mentioned earlier, the classical Schwarz method does not converge without overlap: for $L = 0$ we obtain $\rho_{cla}(k, 0, \eta) = 1$ and hence convergence is lost for all modes. Optimized Schwarz methods, however, can be used without overlap, and nonoverlapping Schwarz methods can be of great interest, if the physical properties in the subdomains differ, for example, when there are jumps in the coefficients of the equation as in [20] or the nature of the equations changes, like in the case of coupling of hyperbolic and parabolic problems; see, for example, [18] and references therein. If we set $L = 0$ in the convergence factor (4.4) of the optimized Schwarz method, the exponential term becomes one, but the factor in front remains unchanged, and thus $\rho_{T0}(k, 0, \eta) < 1$ and $\rho_{T2}(k, 0, \eta) < 1$ for all k. In a numerical implementation there is a maximum frequency which can be represented on a grid with grid spacing h. An estimate for this maximum frequency is $k_{\max} = \frac{\pi}{h}$. Hence the slowest convergence for the optimized Schwarz method without overlap and Taylor transmission conditions is obtained for the highest frequency: the method is a rougher as opposed to the smoother the classical Schwarz method is.

In practice, even when using the Schwarz method with overlap, the overlap is often only a few grid cells wide, and thus $L = O(h)$. In that case the convergence factor of the classical Schwarz method deteriorates as well as one refines the mesh and h goes to zero and we have the following comparison theorem.

THEOREM 4.2. *The optimized Schwarz methods with Taylor transmission conditions and overlap $L = h$ have an asymptotically superior performance than the*

classical Schwarz method with the same overlap. As h goes to zero, we have

$$\max_{|k|\le\frac{\pi}{h}} |\rho_{cla}(k,h,\eta)| = 1 - 2\sqrt{\eta}h + O(h^2), \tag{4.5}$$

$$\max_{|k|\le\frac{\pi}{h}} |\rho_{T0}(k,h,\eta)| = 1 - 4\sqrt{2}\eta^{\frac{1}{4}}\sqrt{h} + O(h), \tag{4.6}$$

$$\max_{|k|\le\frac{\pi}{h}} |\rho_{T2}(k,h,\eta)| = 1 - 8\eta^{\frac{1}{4}}\sqrt{h} + O(h). \tag{4.7}$$

Without overlap, the optimized Schwarz methods with Taylor transmission conditions are asymptotically comparable to the classical Schwarz method with overlap $L = h$. As h goes to zero, we have

$$\max_{|k|\le\frac{\pi}{h}} |\rho_{T0}(k,0,\eta)| = 1 - 4\frac{\sqrt{\eta}}{\pi}h + O(h^2), \tag{4.8}$$

$$\max_{|k|\le\frac{\pi}{h}} |\rho_{T2}(k,0,\eta)| = 1 - 8\frac{\sqrt{\eta}}{\pi}h + O(h^2). \tag{4.9}$$

Proof. For the second result it suffices to expand the convergence factors (4.4) for $L = 0$ at $k = k_{\max} = \frac{\pi}{h}$ for h small. Similarly for the classical Schwarz method one expands the convergence factor (2.8) with $L = h$ for h small at $k = 0$. For the optimized Schwarz methods with Taylor transmission conditions and overlap $L = h$, the convergence factors (4.4) attain their maximum in the interior, at

$$\bar{k}_{T0} = \frac{\sqrt{2}\eta^{\frac{1}{4}}}{\sqrt{L}} \quad \text{and} \quad \bar{k}_{T2} = \frac{2\eta^{\frac{1}{4}}}{\sqrt{L}},$$

respectively, as a direct computation shows. Hence with overlap $L = h$, these maxima are in the range of the computational frequencies, since they are smaller than $k_{\max} = \frac{\pi}{h}$ and thus relevant for the convergence factor. Expanding the corresponding convergence factor at these maxima for $L = h$ as h goes to zero leads to the results (4.8) and (4.9). ∎

Hence already for Taylor expansions of the optimal symbols in the transmission conditions the asymptotic performance of the new Schwarz method is better than the one of the classical Schwarz method when the overlap is of the order of the mesh parameter, which is often the case in applications. One can, however, also see that increasing the order of the Taylor approximation does not increase the asymptotic performance further—there is only an initial gain from h to $\sqrt{h}$. This changes with the approach described in the next subsection.

4.2. Uniformly optimized approximations. We now develop an even better choice for the transmission conditions: one can choose the parameters p_j and q_j to optimize the performance of the new Schwarz method, which means minimizing the convergence factor over all frequencies relevant to the problem. For the zeroth order transmission condition we have the min-max problem

$$\min_{p_j\ge 0}\left(\max_{k_{\min}\le k\le k_{\max}}\left|\frac{\sqrt{\eta+k^2}-p_1}{\sqrt{\eta+k^2}+p_1}\right|\left|\frac{\sqrt{\eta+k^2}-p_2}{\sqrt{\eta+k^2}+p_2}\right|e^{-2\sqrt{\eta+k^2}L}\right), \tag{4.10}$$

and for the second order optimized Schwarz method the min-max problem is

$$\min_{p_j,q_j\ge 0}\left(\max_{k_{\min}\le k\le k_{\max}}\left|\frac{\sqrt{\eta+k^2}-p_1-q_1k^2}{\sqrt{\eta+k^2}+p_1+q_1k^2}\right|\left|\frac{\sqrt{\eta+k^2}-p_2-q_2k^2}{\sqrt{\eta+k^2}+p_2+q_2k^2}\right|e^{-2\sqrt{\eta+k^2}L}\right), \tag{4.11}$$

where we have also introduced a lower bound $k_{\min}$ on the frequency range. This is useful for bounded subdomains: if, for example, the subdomains Ω_j were strips of width 1 in the y direction with homogeneous Dirichlet boundary conditions, then the lowest possible frequency on those domains would be $k_{\min} = \pi$, as one can see from a sine expansion. More general, if one uses a coarse grid, which is necessary as soon as one has many subdomains for good performance on elliptic problems, then the highest frequency representable on the coarse grid would be an estimate for $k_{\min}$, since the subdomain iteration does not need to be effective on the coarse grid frequencies.

Since the optimal transmission conditions (3.6) are of the same size with opposite signs, we first analyze the simpler optimization problems when the approximation of the optimal transmission conditions is also of the same size with opposite signs, which means $p_1 = p_2 = p$ and $q_1 = q_2 = q$. In subsection 4.3 we will analyze how much is lost in performance due to this simplifying assumption.

4.2.1. Zeroth order optimized transmission conditions. Using the same zeroth order transmission condition on both sides of the interface, $p_1 = p_2 = p$ and $q_1 = q_2 = 0$, the expression (4.2) of the convergence factor simplifies to

$$\rho_{OO0}(k, L, \eta, p) := \left(\frac{\sqrt{k^2+\eta} - p}{\sqrt{k^2+\eta} + p}\right)^2 e^{-2\sqrt{k^2+\eta}L}. \tag{4.12}$$

To determine the optimal parameter p of the associated optimized Schwarz method (which we call OO0 for "Optimized of Order 0"), we have to solve the min-max problem

$$\min_{p\geq 0}\left(\max_{k_{\min}\leq k\leq k_{\max}} |\rho_{OO0}(k, L, \eta, p)|\right) = \min_{p\geq 0}\left(\max_{k_{\min}\leq k\leq k_{\max}} \left(\frac{\sqrt{\eta+k^2} - p}{\sqrt{\eta+k^2} + p}\right)^2 e^{-2\sqrt{\eta+k^2}L}\right). \tag{4.13}$$

The following Lemma will be needed for several of the results on the min-max problems that arise in the optimization of the new Schwarz methods.

LEMMA 4.3. *Let $f(x,\gamma)$ be a continuously differentiable function, $f : [a,b] \times [c,d] \mapsto \mathbb{R}$, with a unique interior maximum in x at $x^*(\gamma) \in (a,b)$ for each $\gamma \in [c,d]$, $\frac{\partial f}{\partial x}(x^*(\gamma),\gamma) = 0$, and assume that $x^*(\gamma)$ is differentiable and $\frac{\partial f}{\partial \gamma} < 0$ for $x \in [a,b]$, $\gamma \in [c,d]$. Then*

$$\frac{df}{d\gamma}(x^*(\gamma),\gamma) < 0 \quad \forall \gamma \in [c,d].$$

Proof. Since $\frac{\partial f}{\partial x}(x^*(\gamma),\gamma) = 0$ for all $\gamma \in [c,d]$, we have

$$\frac{df}{d\gamma}(x^*(\gamma),\gamma) = \frac{\partial f}{\partial \gamma}(x^*(\gamma),\gamma) + \frac{\partial f}{\partial x}(x^*(\gamma),\gamma)\frac{\partial x^*}{\partial \gamma}(\gamma) = \frac{\partial f}{\partial \gamma}(x^*(\gamma),\gamma) < 0$$

by assumption on the partial derivative with respect to γ. ∎

THEOREM 4.4 (optimal Robin parameter). *For $L > 0$ and $k_{\max} = \infty$, the solution p^* of the min-max problem (4.13) is given by the unique root of the equation*

$$\rho_{OO0}(k_{\min}, L, \eta, p^*) = \rho_{OO0}(\bar{k}(p^*), L, \eta, p^*), \qquad \bar{k}(L,\eta,p) = \frac{\sqrt{L(2p + L(p^2-\eta))}}{L}. \tag{4.14}$$

For $L = 0$ and $k_{\max}$ finite, the optimal parameter p^* is given by

$$p^* = ((k_{\min}^2 + \eta)(k_{\max}^2 + \eta))^{\frac{1}{4}}. \tag{4.15}$$

Proof. The key idea is to use a transformation: the partial derivative of ρ_{OO0} with respect to p is

$$\frac{\partial \rho_{OO0}}{\partial p} = 4\frac{(p - \sqrt{k^2+\eta})\sqrt{k^2+\eta}e^{-2\sqrt{k^2+\eta}L}}{(\sqrt{k^2+\eta}+p)^3},$$

which shows that as long as $p < \sqrt{k_{\min}^2 + \eta}$, increasing p decreases ρ_{OO0} for all $k \in [k_{\min}, \infty)$. Hence one can restrict the range for p in the min-max problem to $p \geq \sqrt{k_{\min}^2 + \eta}$, the solution cannot lie outside this range. This implies that for the new range of p, ρ_{OO0} has a unique zero in $[k_{\min}, \infty)$, namely, at $k = k_1 = \sqrt{p^2 - \eta}$. We can thus transform the min-max problem into a new, equivalent one in the parameter k_1. Defining the function

$$R(k, L, \eta, k_1) := \frac{(\sqrt{k^2+\eta} - \sqrt{k_1^2+\eta})}{\sqrt{k^2+\eta} + \sqrt{k_1^2+\eta}} e^{-\sqrt{k^2+\eta}L}, \tag{4.16}$$

which is negative for $k \in [k_{\min}, k_1)$ and positive for $k > k_1$, the new min-max problem which is equivalent to (4.13) is

$$\min_{k_1 \geq k_{\min}} \left(\max_{k_{\min} \leq k \leq k_{\max}} |R(k, L, \eta, k_1)| \right).$$

Now in the case of overlap, $L > 0$, the derivative with respect to k,

$$\frac{\partial R}{\partial k} = \frac{k e^{-\sqrt{k^2+\eta}L}(2\sqrt{k_1^2+\eta} - Lk^2 + Lk_1^2)}{(\sqrt{k^2+\eta} + \sqrt{k_1^2+\eta})^2\sqrt{k^2+\eta}}$$

shows that the function has a maximum at

$$\bar{k} = \bar{k}(k_1) = \sqrt{\frac{2\sqrt{k_1^2+\eta}}{L} + k_1^2} > k_1.$$

Hence the maximum in the min-max problem can be attained either at $k = k_{\min}$ or at $k = \bar{k}$. Since

$$\frac{\partial R}{\partial k_1} = -2\frac{k_1 e^{-\sqrt{k^2+\eta}L}\sqrt{k^2+\eta}}{(\sqrt{k^2+\eta} + \sqrt{k_1^2+\eta})^2\sqrt{k_1^2+\eta}} < 0, \tag{4.17}$$

the function R decreases monotonically with k_1. For $k_1 = k_{\min}$ we have $0 = |R(k_{\min}, L, \eta, k_{\min})| < R(\bar{k}, L, \eta, k_{\min})$ and for k_1 large, we have $|R(k_{\min}, L, \eta, k_1)| > R(\bar{k}, L, \eta, k_1)$, since in the limit as k_1 goes to infinity, $R(\bar{k}(k_1), L, \eta, k_1)$ goes to zero. By continuity there exists at least one k_1^* such that $-R(k_{\min}, L, \eta, k_1^*) = R(\bar{k}, L, \eta, k_1^*)$. Using now that R decreases monotonically in k_1, we have that $|R(k_{\min}, L, \eta, k_1)|$ increases monotonically with k_1 and by Lemma 4.3 that $R(\bar{k}(k_1), L, \eta, k_1)$ decreases monotonically with k_1. Hence k_1^* is unique and therefore the unique solution of the min-max problem. Back-transforming to the p variable gives the first result of the theorem.

In the case without overlap, $L = 0$, the function R has no interior maximum, hence the maximum can be attained only on the boundary at either $k = k_{\min}$ or at $k = k_{\max}$. Since the sign of the derivative (4.17) remains the same for $L = 0$, the function R decreases monotonically with k_1. For $k_1 = k_{\min}$ we have $0 = |R(k_{\min}, 0, \eta, k_{\min})| < R(k_{\max}, 0, \eta, k_{\min})$ and for $k_1 = k_{\max}$, we have $|R(k_{\min}, L, \eta, k_{\max})| > R(k_{\max}, L, \eta, k_{\max}) = 0$. By continuity there exists at least one k_1^* such that

$$-R(k_{\min}, 0, \eta, k_1^*) = R(k_{\max}, 0, \eta, k_1^*) \tag{4.18}$$

and since R decreases monotonically in k_1, we have that $|R(k_{\min}, L, \eta, k_1)|$ increases monotonically with k_1 and $R(k_{\max}, L, \eta, k_1)$ decreases monotonically with k_1. Hence k_1^* is unique and thus the unique solution of the min-max problem. Solving (4.18) and back-transforming the result to the p variable leads then to the second result of the theorem. ∎

Figure 4.1 shows on the right the convergence factors obtained with the optimized Robin transmission condition for the model problem with overlap $L = \frac{1}{100}$ and $\eta = 1$, comparing the classical Schwarz method and the OO0 Schwarz method. The maximum of the convergence factor of the OO0 Schwarz method is 0.332, which means that about 55 iterations of the classical Schwarz method with convergence factor 0.980 are needed to attain the performance of the OO0 Schwarz method.

THEOREM 4.5 (Robin asymptotics). *The asymptotic performance of the new Schwarz method with optimized Robin transmission conditions and overlap $L = h$, as h goes to zero, is given by*

$$\max_{k_{\min} \le |k| \le \frac{\pi}{h}} |\rho_{OO0}(k, h, \eta, p^*)| = 1 - 4 \cdot 2^{\frac{1}{6}} (k_{\min}^2 + \eta)^{\frac{1}{6}} h^{\frac{1}{3}} + O(h^{\frac{2}{3}}). \tag{4.19}$$

The asymptotic performance without overlap, $L = 0$, is given by

$$\max_{k_{\min} \le |k| \le \frac{\pi}{h}} |\rho_{OO0}(k, 0, \eta, p^*)| = 1 - 4 \frac{(k_{\min}^2 + \eta)^{\frac{1}{4}}}{\sqrt{\pi}} \sqrt{h} + O(h). \tag{4.20}$$

Proof. For the first result, we need to find an asymptotic expansion for the optimal parameter p^* for small h from (4.14). We make the ansatz $p^* = Ch^\alpha$ for $\alpha < 0$, since we know from Theorem 4.4 that the optimal parameter is growing when h diminishes. Inserting this ansatz into (4.14) satisfied by p^* and expanding for small h, we find the leading order terms in the equation to be $4C\sqrt{k_{\min}^2 + \eta} h^\alpha - 4\sqrt{2} C^{\frac{5}{2}} h^{\frac{5}{2}\alpha + \frac{1}{2}}$. Since (4.14) holds for all h, this expression must vanish and hence both the exponents and the coefficients must match, which leads to

$$\alpha = -\frac{1}{3}, \quad C = \frac{(4(k_{\min}^2 + \eta))^{\frac{1}{3}}}{2}$$

and hence the optimal parameter p^* behaves asymptotically like

$$p^* = \frac{(4(k_{\min}^2 + \eta))^{\frac{1}{3}}}{2} h^{-\frac{1}{3}}. \tag{4.21}$$

With this asymptotic behavior of p^*, the interior maximum $\bar{k}$ behaves asymptotically like

$$\bar{k} = (4(k_{\min}^2 + \eta^2)^{\frac{1}{6}} h^{-\frac{2}{3}}, \tag{4.22}$$

which is less than $k_{\max} = \frac{\pi}{h}$ for h small and hence the optimal result given for $k_{\max} = \infty$ in (4.14) is indeed asymptotically the relevant one on the bounded frequency range $|k| \leq k_{\max} = \frac{\pi}{h}$ for $L = O(h)$. Now inserting the asymptotic value of the optimal parameter p^* from (4.21) into the convergence factor (4.12) and expanding at $k = k_{\min}$ leads to (4.19).

For the second result where $L = 0$, the optimal parameter p^* is known in closed form from (4.15) and hence it suffices to insert this p^* into the convergence factor (4.12), to set $k_{\max} = \frac{\pi}{h}$, and to expand the result at $k = k_{\min}$ in a series for small h to find (4.20). ∎

4.2.2. Second order optimized transmission conditions. Using the same second order transmission condition on both sides of the interface, $p_1 = p_2 = p$ and $q_1 = q_2 = q$, the expression (4.2) of the convergence factor simplifies to

$$\rho_{OO2}(k, L, \eta, p, q) := \left(\frac{\sqrt{k^2 + \eta} - p - qk^2}{\sqrt{k^2 + \eta} + p + qk^2} \right)^2 e^{-2\sqrt{k^2+\eta}L}. \tag{4.23}$$

To determine the optimal parameters p and q for the associated Schwarz method (which we call OO2 for "Optimized of Order 2," a term introduced in [25]), we have to solve the min-max problem

$$\begin{aligned} &\min_{p,q \geq 0} \left(\max_{k_{\min} \leq k \leq k_{\max}} |\rho_{OO2}(k, L, \eta, p, q)| \right) \\ &\quad = \min_{p,q \geq 0} \left(\max_{k_{\min} \leq k \leq k_{\max}} \left(\frac{\sqrt{\eta + k^2} - p - qk^2}{\sqrt{\eta + k^2} + p + qk^2} \right)^2 e^{-2\sqrt{\eta+k^2}L} \right). \end{aligned} \tag{4.24}$$

We need a second technical lemma for the analysis of the optimal parameters.

LEMMA 4.6. *Let $R_1(k_1, k_2)$ and $R_2(k_1, k_2)$ be two continuously differentiable functions, $R_j : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$, $j = 1, 2$, such that the partial derivatives satisfy*

$$\frac{\partial R_1}{\partial k_1} < 0, \quad \frac{\partial R_1}{\partial k_2} < 0, \quad \frac{\partial R_2}{\partial k_1} < 0, \quad \frac{\partial R_2}{\partial k_2} > 0 \tag{4.25}$$

and assume that there exists a unique differentiable $k_1^*(k_2)$ such that

$$R_1(k_1^*(k_2), k_2) + R_2(k_1^*(k_2), k_2) = 0. \tag{4.26}$$

Then we must have

$$\frac{dR_2}{dk_2}(k_1^*(k_2), k_2) > 0. \tag{4.27}$$

Proof. Using implicit differentiation of (4.26), we find

$$\frac{dk_1^*}{dk_2}(k_2) = -\frac{\frac{\partial R_1}{\partial k_2}(k_1^*(k_2), k_2) + \frac{\partial R_2}{\partial k_2}(k_1^*(k_2), k_2)}{\frac{\partial R_1}{\partial k_1}(k_1^*(k_2), k_2) + \frac{\partial R_2}{\partial k_1}(k_1^*(k_2), k_2)}$$

and inserting this result into the total derivative, we obtain

$$\begin{aligned} \frac{dR_2}{dk_2}(k_1^*(k_2), k_2) &= \frac{\partial R_2}{\partial k_1}(k_1^*(k_2), k_2) \frac{dk_1^*}{dk_2}(k_2) + \frac{\partial R_2}{\partial k_2}(k_1^*(k_2), k_2) \\ &= \frac{-\frac{\partial R_2}{\partial k_1}(k_1^*(k_2), k_2) \frac{\partial R_1}{\partial k_2}(k_1^*(k_2), k_2) + \frac{\partial R_2}{\partial k_1}(k_1^*(k_2), k_2) \frac{\partial R_1}{\partial k_1}(k_1^*(k_2), k_2)}{\frac{\partial R_1}{\partial k_1}(k_1^*(k_2), k_2) + \frac{\partial R_2}{\partial k_1}(k_1^*(k_2), k_2)} \\ &> 0 \end{aligned}$$

using the assumption on the signs of the partial derivatives. ∎

THEOREM 4.7 (optimal second order parameters). *For $L > 0$ and $k_{\max} = \infty$, the solution p^*, q^* of the min-max problem* (4.24) *is given by the unique root of the system of equations*

$$\rho_{OO2}(k_{\min}, L, \eta, p^*, q^*) = \rho_{OO2}(\bar{k}_1, L, \eta, p^*, q^*) = \rho_{OO2}(\bar{k}_2, L, \eta, p^*, q^*), \tag{4.28}$$

where the locations of the maxima $\bar{k}_1$ and $\bar{k}_2$ are given by

$$\bar{k}_{1,2}(L,\eta,p,q) = \frac{1}{q}\sqrt{\frac{L+2q-2Lpq \mp \sqrt{L^2+4Lq-4L^2pq+4q^2-16Lpq^2+16Lq^3\eta+4L^2q^2\eta}}{2L}}. \tag{4.29}$$

For $L = 0$ and $k_{\max}$ finite, the optimal parameters p^* and q^* are given by

$$p^* = \frac{k_{\max}^2\sqrt{k_{\min}^2+\eta}-k_{\min}^2\sqrt{k_{\max}^2+\eta}}{\sqrt{2(k_{\max}^2-k_{\min}^2)}\left(\left(\sqrt{k_{\max}^2+\eta}-\sqrt{k_{\min}^2+\eta}\right)\left((k_{\max}^2+\eta)\sqrt{k_{\min}^2+\eta}-(k_{\min}^2+\eta)\sqrt{k_{\max}^2+\eta}\right)\right)^{\frac{1}{4}}}, \tag{4.30}$$

$$q^* = \frac{\left(\sqrt{k_{\max}^2+\eta}-\sqrt{k_{\min}^2+\eta}\right)^{\frac{3}{4}}}{\sqrt{2(k_{\max}^2-k_{\min}^2)}\left((k_{\max}^2+\eta)\sqrt{k_{\min}^2+\eta}-(k_{\min}^2+\eta)\sqrt{k_{\max}^2+\eta}\right)^{\frac{1}{4}}}.$$

Proof. The argument is again based on a transformation: the partial derivatives of ρ_{OO2} with respect to p and q are

$$\frac{\partial \rho_{OO2}}{\partial p} = 4\sqrt{k^2+\eta}\,\frac{p+qk^2-\sqrt{k^2+\eta}}{(p+qk^2+\sqrt{k^2+\eta})^3}e^{-2\sqrt{k^2+\eta}L}, \quad \frac{\partial \rho_{OO2}}{\partial q} = k^2\frac{\partial \rho_{OO2}}{\partial p}, \tag{4.31}$$

and hence ρ_{OO2} is monotonically decreasing when p and q are decreased for all $k > k_{\min}$ as long as $p + qk^2 > \sqrt{k^2+\eta}$. This implies that at the solution of the min-max problem ρ_{OO2} must have at least one zero $k_1 > k_{\min}$. Then instead of using the parameter p, we can use equivalently the parameter k_1 in the min-max problem by setting $p := \sqrt{k_1+\eta} - qk_1^2$, which leads to the new form of the convergence factor

$$\rho'_{OO2} = \frac{(\sqrt{k_1^2+\eta}-\sqrt{k^2+\eta}+q(k^2-k_1^2))^2}{(\sqrt{k^2+\eta}+\sqrt{k_1^2+\eta}+q(k^2-k_1^2))^2}e^{-2\sqrt{k^2+\eta}L},$$

which has now necessarily a zero at $k_1 > k_{\min}$. If we suppose that k_1 is the only zero at the optimum, we reach again a contradiction, because a partial derivative with respect to q gives

$$\frac{\partial \rho'_{OO2}}{\partial q} = 4\sqrt{k^2+\eta}(k^2-k_1^2)\frac{\sqrt{k_1^2+\eta}-\sqrt{k^2+\eta}+q(k^2-k_1^2)}{(\sqrt{k^2+\eta}+\sqrt{k_1^2+\eta}+q(k^2-k_1^2))^3}e^{-2\sqrt{k^2+\eta}L},$$

where the denominator is positive, since $\sqrt{k_1+\eta} - qk_1^2 = p \geq 0$, and the numerator changes sign only at $k = k_1$ by assumption, which together with the factor $(k^2 - k_1^2)$ in front makes the sign of the derivative negative for all $q > 0$ as long as there is only one zero at k_1. Thus increasing q the convergence factor ρ'_{OO2} can be decreased for all $k > k_{\min}$ as long as there is no second zero. Hence at the optimum, ρ'_{OO2} must have a second zero, without loss of generality at $k_2 \geq k_1 > k_{\min}$. Thus we can use the parameter k_2 instead of q, which leads to the change of variables

$$p = \frac{\sqrt{k_1^2+\eta}k_2^2 - k_1^2\sqrt{k_2^2+\eta}}{k_2^2-k_1^2}, \quad q = \frac{\sqrt{k_2^2+\eta}-\sqrt{k_1^2+\eta}}{k_2^2-k_1^2} \tag{4.32}$$

and the new min-max problem, which is equivalent to (4.24), is

$$\min_{k_{\min}<k_1\le k_2}\left(\max_{k_{\min}\le k\le k_{\max}}|R(k,L,\eta,k_1,k_2)|\right), \tag{4.33}$$

with the new function $R(k,L,\eta,k_1,k_2)$, representing the square root of the convergence factor, given by

$$R(k,L,\eta,k_1,k_2) = \frac{\sqrt{k^2+\eta}(k_2^2-k_1^2)-(\sqrt{k_1^2+\eta}k_2^2-\sqrt{k_2^2+\eta}k_1^2)-(\sqrt{k_2^2+\eta}-\sqrt{k_1^2+\eta})k^2}{\sqrt{k^2+\eta}(k_2^2-k_1^2)+(\sqrt{k_1^2+\eta}k_2^2-\sqrt{k_2^2+\eta}k_1^2)+(\sqrt{k_2^2+\eta}-\sqrt{k_1^2+\eta})k^2}e^{-\sqrt{k^2+\eta}L}. \tag{4.34}$$

Taking the partial derivatives with respect to k_1 and k_2, we find

$$\frac{\partial R}{\partial k_1} = \frac{2k_1(k^2-k_2^2)\sqrt{k^2+\eta}\left(\sqrt{k_2^2+\eta}-\sqrt{k_1^2+\eta}\right)^2}{\sqrt{k_1^2+\eta}\left(\sqrt{k^2+\eta}(k_1^2-k_2^2)+\sqrt{k_1^2+\eta}(k^2-k_2^2)+(k_1^2-k^2)\sqrt{k_2^2+\eta}\right)^2}e^{-\sqrt{k^2+\eta}L}, \tag{4.35}$$

$$\frac{\partial R}{\partial k_2} = \frac{2k_2(k^2-k_1^2)\sqrt{k^2+\eta}\left(\sqrt{k_2^2+\eta}-\sqrt{k_1^2+\eta}\right)^2}{\sqrt{k_2^2+\eta}\left(\sqrt{k^2+\eta}(k_1^2-k_2^2)+\sqrt{k_1^2+\eta}(k^2-k_2^2)+(k_1^2-k^2)\sqrt{k_2^2+\eta}\right)^2}e^{-\sqrt{k^2+\eta}L}, \tag{4.36}$$

which shows that for $k<k_2$, the function R is decreasing when k_1 is increasing and for $k>k_2$ it is increasing with k_1. Similarly for $k<k_1$, the function R is decreasing when k_2 is increasing, and for $k>k_1$ it is increasing with k_2.

Now for $L>0$, $R=(-1+O(\frac{1}{k}))e^{-Lk}$ as k goes to infinity. Hence the maximum in the min-max problem can be attained, by continuity of R and knowing that there are two zeros at $k_1,k_2>k_{\min}$, either at $k_{\min}$, where R is negative, or at $k=\bar{k}_1$ given in (4.29), where R has a maximum, $k_1\le\bar{k}_1\le k_2$, or at $k=\bar{k}_2$ given in (4.29), where R has a negative minimum, $k_2\le\bar{k}_2$. To show that the solution of the min-max problem is indeed when the three are balanced, we first note that for any fixed k_2, there exists a unique $k_1^*=k_1^*(k_2)\in[k_{\min},k_2]$ such that $|R(k_{\min},L,\eta,k_1^*(k_2),k_2)|=R(\bar{k}_1,L,\eta,k_1^*(k_2),k_2)$, because of continuity and $0=|R(k_{\min},L,\eta,k_{\min},k_2)|<R(\bar{k}_1,L,\eta,k_{\min},k_2)$ and $|R(k_{\min},L,\eta,k_2,k_2)|>R(\bar{k}_1,L,\eta,k_2,k_2)=R(k_2,L,\eta,k_2,k_2)=0$, and $|R(k_{\min},L,\eta,k_1,k_2)|$ is monotonically increasing with k_1 by (4.35) and $R(\bar{k}_1,L,\eta,k_1,k_2)$ is monotonically decreasing in k_1 by (4.35) and Lemma 4.3. Hence denoting by $R_1(k_1,k_2):=R(k_{\min},L,\eta,k_1,k_2)$ and $R_2(k_1,k_2):=R(\bar{k}_1,L,\eta,k_1,k_2)$ Lemma 4.6 applies and therefore $|R(k_{\min},L,\eta,k_1^*(k_2),k_2)|=R(\bar{k}_1,L,\eta,k_1^*(k_2),k_2)$ is monotonically increasing with k_2. Now for $k_2=k_{\min}$, we have $k_1^*(k_{\min})=k_{\min}$ and thus $0=|R(k_{\min},L,\eta,k_{\min},k_{\min})|<|R(\bar{k}_2,L,\eta,k_{\min},k_{\min})|$ and for large k_2 we have $|R(k_{\min},L,\eta,k_1^*(k_2),k_2)|>|R(\bar{k}_2,L,\eta,k_1^*(k_2),k_2)|$ (since the right-hand term goes to zero in the limit). Therefore by continuity, Lemma 4.6 for $|R(k_{\min},L,\eta,k_1^*(k_2),k_2)|$ and Lemma 4.3 for $|R(\bar{k}_2,L,\eta,k_1^*(k_2),k_2)|$ (note that $(k_1^*)'\ge 0$), there exists a unique k_2^* where these two expressions are equal, $|R(k_{\min},L,\eta,$

$k_1^*(k_2^*), k_2^*)| = |R(\bar{k}_2, L, \eta, k_1^*(k_2^*), k_2^*)|$, which is the unique solution of the min-max problem. Back-transforming to the p and q variables using (4.32) we obtain the equations for the solution of the min-max problem given in (4.28).

In the case without overlap, $L = 0$, the function R behaves for large k like $-1+O(\frac{1}{k})$ and hence the maximum in the min-max problem for $L = 0$ can be attained either at $k_{\min}$, $k_{\max}$, or at the interior maximum $\bar{k}_1$ which satisfies $k_1 \le \bar{k}_1 \le k_2$ and is given in the p and q variables by

$$\bar{k}_1 = \frac{\sqrt{q(p - 2q\eta)}}{q}.$$

The same argument used for the case $L > 0$ is still valid, and hence there exists a unique solution p^*, q^* of the min-max problem which is characterized by the system of equations

$$\rho_{OO2}(k_{\min}, 0, \eta, p^*, q^*) = \rho_{OO2}(\bar{k}_1, 0, \eta, p^*, q^*) = \rho_{OO2}(k_{\max}, 0, \eta, p^*, q^*). \tag{4.37}$$

This system can be solved in closed form by first solving $\rho_{OO2}(k_{\min}, 0, \eta, p, q^*) = \rho_{OO2}(k_{\max}, 0, \eta, p, q^*)$ for $q^* = q^*(p)$, which leads to

$$q^*(p) = \frac{p(\sqrt{k_{\max}^2 + \eta} - \sqrt{k_{\min}^2 + \eta})}{\sqrt{k_{\min}^2 + \eta}k_{\max}^2 - k_{\min}^2\sqrt{k_{\max}^2 + \eta}}.$$

Inserting this solution into the remaining equation $\rho_{OO2}(k_{\min}, 0, \eta, p^*, q^*(p^*)) = \rho_{OO2}(\bar{k}_1, L, \eta, p^*, q^*(p^*))$ and solving for p^* leads to the closed form solution (4.30) of the min-max problem for $L = 0$. ∎

Figure 4.1 shows on the right the convergence factor obtained with the second order optimized transmission conditions for our model problem with overlap $L = \frac{1}{100}$ and $\eta = 1$, comparing it to the convergence factor of the classical Schwarz method. The maximum of the convergence factor of the new Schwarz method with optimized second order transmission conditions is 0.0704, which means that about 131 iterations of the classical Schwarz method with convergence factor 0.980 are needed to attain the performance of the second order optimized Schwarz method.

THEOREM 4.8 (OO2 asymptotics). *The asymptotic performance of the new Schwarz method with optimized second order transmission conditions and overlap $L = h$, as h goes to zero, is given by*

$$\max_{k_{\min} \le k \le \frac{\pi}{h}} |\rho_{OO2}(k, h, \eta, p^*, q^*)| = 1 - 4 \cdot 2^{\frac{3}{5}}(k_{\min}^2 + \eta)^{\frac{1}{10}} h^{\frac{1}{5}} + O(h^{\frac{2}{5}}). \tag{4.38}$$

The asymptotic performance without overlap, $L = 0$, is for h small given by

$$\max_{k_{\min} \le k \le \frac{\pi}{h}} |\rho_{OO2}(k, 0, \eta, p^*, q^*)| = 1 - 4\frac{\sqrt{2}(k_{\min}^2 + \eta)^{\frac{1}{8}}}{\pi^{\frac{1}{4}}} h^{\frac{1}{4}} + O(h^{\frac{1}{2}}). \tag{4.39}$$

Proof. To obtain the first result, we need to solve the nonlinear equations (4.28) asymptotically in h for the optimal parameters p^* and q^*. We make the ansatz $p = C_1 h^\alpha$ and $q = C_2 h^\beta$, insert this together with $L = h$ into the nonlinear equations (4.28), and expand for small h. The search for the lowest order terms is simplified by the knowledge that $\alpha < 0$ and $\beta > 0$ since p is growing when h is decaying and q is diminishing with h. Expanding for h small, we find from the equation $\rho_{OO2}(k_{\min}, h, \eta, p^*, q^*) = \rho_{OO2}(\bar{k}_1, h, \eta, p^*, q^*)$ the leading order terms

$$-4\sqrt{2}C_1 h^\alpha \sqrt{h} + 8C_2 h^\beta \sqrt{C_1 h^\alpha} C_1 h^\alpha$$

and from the equation $\rho_{OO2}(k_{\min}, h, \eta, p^*, q^*) = \rho_{OO2}(k_2, h, \eta, p^*, q^*)$ the leading order terms

$$-4\sqrt{2}C_1^2 h^{2\alpha}\sqrt{h} + 4\sqrt{k_{\min}^2+\eta}\sqrt{C_2 h^\beta}C_1 h^\alpha.$$

Since the equations hold at the optimum, the leading order terms must match, which leads to a system of equations for the unknown exponents α and β,

$$\frac{3}{2}\alpha + \beta = \alpha + \frac{1}{2}, \quad 2\alpha + \frac{1}{2} = \alpha + \frac{\beta}{2},$$

whose solution is $\alpha = -\frac{1}{5}$ and $\beta = \frac{3}{5}$, and a system of equations for the constants C_1 and C_2, whose solution is

$$C_1 = 2^{-\frac{3}{5}}(k_{\min}^2+\eta)^{\frac{2}{5}}, \quad C_2 = (2(k_{\min}^2+\eta))^{-\frac{1}{5}}.$$

Hence asymptotically the optimal parameters p^* and q^* are

$$p^* = 2^{-\frac{3}{5}}(k_{\min}^2+\eta)^{\frac{2}{5}}h^{-\frac{1}{5}}, \quad q^* = (2(k_{\min}^2+\eta))^{-\frac{1}{5}}h^{\frac{3}{5}}. \tag{4.40}$$

To see that the min-max solution given in (4.38) on the infinite frequency domain $k \in [k_{\min}, \infty)$ is really the relevant one asymptotically on the bounded frequency domain $|k| < k_{\max} = \frac{\pi}{h}$, we must have that the second maximum $\bar{k}_2$ given in (4.29) satisfies asymptotically $\bar{k}_2 \le k_{\max}$. Inserting the asymptotic expressions of p^* and q^* from (4.40) into the expression of $\bar{k}_2$ in (4.29), setting $L = h$ and expanding for h small, we find

$$\bar{k}_2 = \frac{2^{\frac{3}{5}}(k_{\min}^2+\eta)^{\frac{1}{10}}}{h^{\frac{4}{5}}} + O(h^{-\frac{2}{5}}) \tag{4.41}$$

and hence indeed asymptotically $\bar{k}_2 \le k_{\max} = \frac{\pi}{h}$. Inserting now the asymptotically optimal parameters p^* and q^* from (4.40) into the convergence factor ρ_{OO2} and expanding as h goes to zero, we obtain the result (4.38).

For the second result without overlap, we have the closed formulas (4.30) for the optimal parameters p^* and q^*. It suffices therefore to insert them into the convergence factor and to expand it in h for $k_{\max} = \frac{\pi}{h}$ at $k = k_{\min}$ to find the result (4.39). ∎

4.3. A two-sided optimized Robin transmission condition. We now investigate how the simplifying assumption $p_1 = p_2$ and $q_1 = q_2$ in the min-max problem (4.11) affects the performance of the optimized Schwarz methods. We do this only for the case of Robin transmission conditions to illustrate the change. We thus have $q_1 = q_2 = 0$ and the optimization problem (4.10).

THEOREM 4.9 (optimal two-sided Robin conditions). *If there is overlap, $L > 0$, then the optimal two-sided Robin parameters are given by*

$$p_1^* = \frac{1-\sqrt{1+4\eta(q^*)^2-4p^*q^*}}{2q^*}, \quad p_2^* = \frac{1+\sqrt{1+4\eta(q^*)^2-4p^*q^*}}{2q^*}, \tag{4.42}$$

where p^* and q^* are solutions of (4.28) *with L replaced by $2L$. If there is no overlap, $L = 0$, then the optimal two-sided Robin parameters are* (4.42)*, where p^* and q^* are given by* (4.30)*.*

Proof. Multiplying the two factors in the optimization problem (4.10), we obtain the optimization problem

$$\min_{p_j \geq 0} \left(\max_{k_{\min} < k < k_{\max}} \left| \frac{\sqrt{\eta + k^2} - \frac{\eta + p_1 p_2}{p_1 + p_2} - \frac{k^2}{p_1 + p_2}}{\sqrt{\eta + k^2} + \frac{\eta + p_1 p_2}{p_1 + p_2} + \frac{k^2}{p_1 + p_2}} \right| e^{-2\sqrt{\eta + k^2} L} \right) \tag{4.43}$$

and hence in the new parameters

$$p = \frac{\eta + p_1 p_2}{p_1 + p_2}, \quad q = \frac{1}{p_1 + p_2}, \tag{4.44}$$

this optimization problem is equivalent to the optimization problem (4.24) provided L is replaced by $2L$. The solution for this problem is given for $L > 0$ in (4.28) and for $L = 0$ in (4.30). Back-transforming these results using (4.44) concludes the proof. ∎

The preceding theorem shows that one can generate the performance of higher order transmission conditions using lower order transmission conditions which are not equal on both sides. In the case without overlap, one needs to perform two iterations of the two-sided optimized Robin transmission algorithm to attain an error reduction equivalent to the one from one iteration of the optimized second order transmission conditions algorithm. With overlap, two iterations of the algorithm with optimized two-sided Robin transmission conditions is even a bit better than one iteration of the algorithm with second order transmission conditions, since the overlap has been effective twice.

Figure 4.1 shows on the right the convergence factors obtained with the two-sided optimized Robin conditions for our model problem with overlap $L = \frac{1}{100}$ and $\eta = 1$, comparing it to the convergence factor of the classical and the optimized zeroth and second order Schwarz methods. The maximum of the convergence factor of the new Schwarz method with two-sided optimized Robin conditions is 0.208, which means that about 78 iterations of the classical Schwarz method with convergence factor 0.980 are needed to attain the performance of the two-sided optimized Robin Schwarz method.

COROLLARY 4.10. *The asymptotic performance of the two-sided optimized Schwarz method with $L = h$ is*

$$\max_{k_{\min} \leq k \leq \frac{\pi}{h}} |\rho(k, h, \eta, p_1^*, p_2^*)| = 1 - 2 \cdot 2^{\frac{4}{5}} (k_{\min}^2 + \eta)^{\frac{1}{10}} h^{\frac{1}{5}} + O(h^{\frac{2}{5}}). \tag{4.45}$$

Without overlap, $L = 0$, the asymptotic performance is given by

$$\max_{k_{\min} \leq k \leq \frac{\pi}{h}} |\rho(k, 0, \eta, p_1^*, p_2^*)| = 1 - 2 \frac{\sqrt{2} (k_{\min}^2 + \eta)^{\frac{1}{8}}}{\pi^{\frac{1}{4}}} h^{\frac{1}{4}} + O(h^{\frac{1}{2}}). \tag{4.46}$$

Hence asymptotically, the second order optimized algorithm and the two-sided optimized Robin algorithm are equivalent: one can get the same asymptotic performance from Robin transmission conditions that one gets from second order transmission conditions, provided one uses different parameters in the two transmission conditions.

The idea of not using the same parameters on each side can be generalized by not using the same parameter in each iteration: one uses a sequence of transmission conditions with Robin parameters p_i, $i = 1, 2, \ldots, I$, where I is a number of parameters chosen and one cycles through the transmission conditions from 1 to I in the

Schwarz iteration. This adds more degrees of freedom in the optimization problem and leads to Schwarz algorithms that have an arbitrarily weak dependence of the convergence factor on h, even without overlap (see [15]), but at the cost of having to solve subdomain problems with varying transmission conditions per interation.

5. Optimized Schwarz methods compared to Schur and FETI methods. We now investigate what the relation is between optimized Schwarz methods, which can be used without overlap, to other domain decomposition methods without overlap, like the Schur methods and FETI (Finite Element Tearing and Interconnect [13]). To this end we will address two questions:

1. What conditions can one impose to couple subdomain problems ?
2. Which of these conditions are good to build efficient domain decomposition algorithms ?

Although the ideas in this section hold for general second order elliptic problems, we will use our self-adjoint coercive model problem (2.1) to fix ideas.

5.1. Classical coupling conditions between subdomains. There are two classical ways to couple subdomain problems. For the first one, one uses an overlapping decomposition of Ω, say, $\Omega_1 = (-\infty, L)$ and $\Omega_2 = (0, \infty)$ for $L > 0$, and the coupled subproblems are given by

$$\begin{array}{cccc} \mathcal{L}(u_1) = f & \text{in } \Omega_1, & \mathcal{L}(u_2) = f & \text{in } \Omega_2, \\ u_1(L, y) = u_2(L, y), & y \in \mathbb{R}, & u_2(0, y) = u_1(0, y), & y \in \mathbb{R}. \end{array} \tag{5.1}$$

Note that we do not introduce an algorithm to find the solution of the coupled subproblems here; we only define coupled subdomain problems which are equivalent to the original problem. The equivalence can be seen in this case, for example, by studying the associated Schwarz algorithm.

For the second approach, one uses subdomains without overlap, for example, $\Omega_1 = (-\infty, 0)$ and $\Omega_2 = (0, \infty)$, and the coupled subdomain problems are

$$\begin{array}{cccc} \mathcal{L}(u_1) = f & \text{in } \Omega_1, & \mathcal{L}(u_2) = f & \text{in } \Omega_2, \\ u_1(0, y) = u_2(0, y), & y \in \mathbb{R}, & \partial_x u_2(0, y) = \partial_x u_1(0, y), & y \in \mathbb{R}. \end{array} \tag{5.2}$$

Note the key difference: in the decomposition without overlap, both the solution values as well as the normal derivatives are imposed to agree on the interface for this second order problem, whereas in the approach with overlap, only solution values are imposed to agree, but at two different locations, which implies the agreement of normal derivatives.

The classical algorithm to find a solution for the case of an overlapping decomposition is the one given by Schwarz in [36],

$$\begin{array}{cccc} \mathcal{L}(u_1^n) = f & \text{in } \Omega_1, & \mathcal{L}(u_2^n) = f & \text{in } \Omega_2, \\ u_1^n(L, y) = u_2^{n-1}(L, y), & y \in \mathbb{R}, & u_2^n(0, y) = u_1^{n-1}(0, y), & y \in \mathbb{R}, \end{array} \tag{5.3}$$

and we have derived the linear convergence factor of this algorithm in (2.8).

Can a similar iterative method be used for the nonoverlapping decomposition? This would lead, for example, to

$$\begin{array}{cccc} \mathcal{L}(u_1^n) = f & \text{in } \Omega_1, & \mathcal{L}(u_2^n) = f & \text{in } \Omega_2, \\ u_1^n(0, y) = u_2^{n-1}(0, y), & y \in \mathbb{R}, & \partial_x u_2^n(0, y) = \partial_x u_1^{n-1}(0, y), & y \in \mathbb{R}. \end{array} \tag{5.4}$$

In general not, because this algorithm does not converge, as one can see with Fourier analysis. Setting for the convergence analysis $f = 0$ by linearity and taking a Fourier

transform in y with parameter k of (5.4) leads to the transformed iterates

$$\hat{u}_1^n(x,k) = \hat{u}_2^{n-1}(0,k)e^{\sqrt{\eta+k^2}x}, \quad \hat{u}_2^n(x,k) = -\hat{u}_1^{n-1}(0,k)e^{-\sqrt{\eta+k^2}x}.$$

Thus inserting $\hat{u}_2^{n-1}(0,k)$ from the second equation into the first one and evaluating at $x = 0$, we find

$$\hat{u}_1^n(0,k) = -\hat{u}_1^{n-2}(0,k) \quad \text{and similarly} \quad \hat{u}_2^n(0,k) = -\hat{u}_2^{n-2}(0,k).$$

Hence the convergence factor of this algorithm is $\rho = -1$ and thus it does not converge.

A first remedy consists of introducing relaxation parameters γ_j, $j = 1, 2$, which leads to the transmission conditions

$$\begin{aligned} u_1^n(0,y) &= \gamma_1 u_2^{n-1}(0,y) + (1-\gamma_1)u_1^{n-1}(0,y), \\ \partial_x u_2^n(0,y) &= \gamma_2 \partial_x u_1^{n-1}(0,y) + (1-\gamma_2)\partial_x u_2^{n-1}(0,y), \end{aligned} \tag{5.5}$$

for which convergence results have been established. In [1] we find that for the so-called Dirichlet–Neumann method, $\gamma_2 = 1$, there exist γ_1 for which the algorithm converges, and in [35] we find that for the Neumann–Dirichlet method, $\gamma_1 = 1$, there exist γ_2 for which the algorithm converges. For our model problem, we find for the interface system in the Fourier domain

$$\begin{pmatrix} \hat{u}_1^n(0,k) \\ \partial_x \hat{u}_2^n(0,k) \end{pmatrix} = \begin{bmatrix} 1-\gamma_1 & \frac{-\gamma_1}{\sqrt{\eta+k^2}} \\ \gamma_2\sqrt{\eta+k^2} & 1-\gamma_2 \end{bmatrix} \begin{pmatrix} \hat{u}_1^{n-1}(0,k) \\ \partial_x \hat{u}_2^{n-1}(0,k) \end{pmatrix}. \tag{5.6}$$

The asymptotic convergence factor of this matrix iteration is governed by the spectral radius of the 2×2 matrix, which is given by the larger eigenvalue in modulus,

$$\rho = \left| 1 - \frac{1}{2}(\gamma_1+\gamma_2) + \frac{1}{2}\sqrt{(\gamma_1-\gamma_2)^2 - 4\gamma_1\gamma_2} \right|. \tag{5.7}$$

Note that ρ is independent of the frequency parameter k, which implies that the convergence factor is independent of the mesh parameter h if the algorithm is discretized. In the case of the Dirichlet–Neumann algorithm, where $\gamma_2 = 1$, the asymptotic convergence factor for our model problem is

$$\rho = \frac{1}{2}\left| 1 - \gamma_1 + \sqrt{\gamma_1^2 - 6\gamma_1 + 1} \right|,$$

which is less than 1 for $0 < \gamma_1 < 1$. The optimal value which minimizes the convergence factor is $\gamma_1 = 3 - 2\sqrt{2} \approx 0.1716$, for which the convergence factor becomes $\rho \approx 0.4142$. The same results we find by the symmetry of the parameters γ_i also in the case of the Neumann–Dirichlet algorithm, where $\gamma_1 = 1$. But one could also use both relaxation parameters simultaneously to minimize the convergence factor. With both parameters, we can achieve that both eigenvalues vanish simultaneously by setting the term under the square root and the one outside of the square root in (5.7) equal to zero. We find that for the choice

$$\gamma_1 = 1 \pm \frac{1}{\sqrt{2}}, \quad \gamma_2 = 1 \mp \frac{1}{\sqrt{2}},$$

the spectral radius vanishes identically, $\rho \equiv 0$. Hence this method will converge in at most two iterations for any initial guess. (The matrix is not normal; otherwise convergence would be in one iteration, which we know is not possible.)

In a Gauss–Seidel version of this iteration, subdomain Ω_2 would use directly the newest values at the interface from subdomain Ω_1. In that case the relaxed interface iteration can be found after a short calculation to be

$$\begin{pmatrix} \hat{u}_1^n(0,k) \\ \partial_x \hat{u}_2^n(0,k) \end{pmatrix} = \begin{bmatrix} 1-\gamma_1 & \frac{-\gamma_1}{\sqrt{\eta+k^2}} \\ \gamma_2(1-\gamma_1)\sqrt{\eta+k^2} & 1-\gamma_2-\gamma_1\gamma_2 \end{bmatrix} \begin{pmatrix} \hat{u}_1^{n-1}(0,k) \\ \partial_x \hat{u}_2^{n-1}(0,k) \end{pmatrix}. \tag{5.8}$$

As before the asymptotic convergence of this matrix iteration is governed by the spectral radius of the 2×2 matrix and the term depending on the frequency parameter k cancels; the convergence factor is independent of k. In this case, however, both the Dirichlet–Neumann and the Neumann–Dirichlet algorithm can achieve already a convergence factor $\rho = 0$; one parameter suffices. The optimal choice is $\gamma_1 = \frac{1}{2}$ for the Dirichlet–Neumann case, where $\gamma_2 = 1$, and $\gamma_2 = \frac{1}{2}$ for the Neumann–Dirichlet case, where $\gamma_1 = 1$, results found already in [1] and [35]. Unfortunately all these results depend strongly on the symmetry in the problem; otherwise the two symbols depending on the frequency parameter k and containing the square root would not cancel. Hence for a more general situation with uneven domains or variable coefficients, convergence in two steps will not be possible with this approach. The optimal Schwarz method using the exact Dirichlet-to-Neumann map, however, does still converge in two iterations also in these more general cases.

A second remedy, and this is really the classical approach for subdomain problems coupled without overlap, consists of avoiding an iteration first. One keeps the coupled problem and introduces a name for the quantities at the interface,

$$\begin{array}{cc} \mathcal{L}(u_1) = f \quad \text{in } \Omega_1, & \mathcal{L}(u_2) = f \quad \text{in } \Omega_2, \\ u_1(0,y) = u_2(0,y) =: \lambda(y), & \partial_x u_2(0,y) = \partial_x u_1(0,y) =: \lambda_x(y). \end{array} \tag{5.9}$$

The primal Schur method then works as follows: supposing that $\lambda(y)$ is known, one computes $u_1(x,y,\lambda)$ and $u_2(x,y,\lambda)$ and then sets

$$\partial_x u_1(0,y,f,\lambda) - \partial_x u_2(0,y,f,\lambda) = 0,$$

which is a linear equation to determine the interface function λ. Solving this linear problem with a Krylov method requires at each step two subdomain solves with Dirichlet conditions,

$$\mathcal{A}_p\lambda := \partial_x u_1(0,y,0,\lambda) - \partial_x u_2(0,y,0,\lambda) = -\partial_x u_1(0,y,f,0) + \partial_x u_2(0,y,f,0) =: b_p. \tag{5.10}$$

To learn more about the conditioning of the primal Schur complement system $\mathcal{A}_p\lambda = b_p$, we take a Fourier transform of $\mathcal{A}_p\lambda$ to find the symbol of $\mathcal{A}_p$,

$$\hat{\mathcal{A}}_p\hat{\lambda} = \hat{v}_x(0,y,0,\hat{\lambda}) - \hat{w}_x(0,y,0,\hat{\lambda}) = 2\sqrt{\eta+k^2}\hat{\lambda}. \tag{5.11}$$

This symbol is symmetric in k and hence the condition number of the corresponding operator can be estimated using the ratio of the symbol at the maximum and minimum frequencies occurring in a given computation. Estimating the minimum frequency by 0 and the maximum frequency by $k_{\max} = \frac{\pi}{h}$ as before, where h is the mesh parameter, we find the asymptotic condition number for h small to be

$$\mathcal{K}(\mathcal{A}_p) = \frac{\pi}{\sqrt{\eta}h} + O(h). \tag{5.12}$$

Note that the original operator $(\eta - \Delta)u = f$ had a condition number estimate of $O(\frac{1}{h^2})$ and thus the primal Schur method improves the condition number by a square root. On the negative side the matrix vector product is now more expensive, since it involves subdomain solves.

The dual Schur method, which became famous under the name FETI, is similar, although the key feature of a natural coarse space cannot be seen in this simple setting: supposing that λ_x is known, we compute $u_1(x, y, f, \lambda_x)$ and $u_2(x, y, f, \lambda_x)$ and then set

$$u_1(0, y, f, \lambda_x) - u_2(0, y, f, \lambda_x) = 0,$$

which is now a linear equation for λ_x. Solving this linear problem with a Krylov method requires at each step two subdomain solves with Neumann conditions,

$$\mathcal{A}_d\lambda_x := u_1(0, y, 0, \lambda_x) - u_2(0, y, 0, \lambda_x) = -u_1(0, y, f, 0) + u_2(0, y, f, 0) =: b_d. \tag{5.13}$$

The Fourier transform of the dual Schur complement system $\mathcal{A}_d\lambda_x = b_d$ leads to

$$\hat{\mathcal{A}}_d\hat{\lambda}_x = \hat{v}(0, y, 0, \hat{\lambda}_x) - \hat{w}(0, y, 0, \hat{\lambda}_x) = \frac{2}{\sqrt{\eta + k^2}}\hat{\lambda}_x, \tag{5.14}$$

which shows that the operator $\mathcal{A}_d$ has the symbol $\frac{2}{\sqrt{\eta+k^2}}$. This symbol is also symmetric in k and as in the case of the primal Schur complement, we find the condition number for h small to be

$$\mathcal{K}(\mathcal{A}_d) = \frac{\pi}{\sqrt{\eta}h} + O(h). \tag{5.15}$$

Now note that the dual Schur complement with the symbol $\frac{2}{\sqrt{\eta+k^2}}$ is the inverse of the primal Schur complement that had the symbol $2\sqrt{\eta + k^2}$, up to the constant 4, and hence one is the ideal preconditioner for the other. This led to the famous Neumann–Neumann preconditioner for the primal Schur complement, with condition number independent of the mesh parameter [2]. Similarly, one could use a Dirichlet–Dirichlet preconditioner for the dual Schur complement or FETI to obtain a mesh independent domain decomposition method.

But why should one give preference to either the Dirichlet or the Neumann condition when formulating a Schur method? And why should we impose the same type of interface conditions on each subdomain? In the recent FETI-DP method [11], for some parts of the interfaces continuity of the dual variables is imposed, and for other parts continuity of the primal variables. One could go a step further and first assume that both λ and λ_x are known, then solve for $u_1(x, y, f, \lambda)$ and $u_2(x, y, f, \lambda_x)$, for example, and set

$$\begin{aligned} u_1(0, y, f, \lambda) - u_2(0, y, f, \lambda_x) &= 0, \\ \partial_x u_1(0, y, f, \lambda) - \partial_x u_2(0, y, f, \lambda_x) &= 0, \end{aligned}$$

which is now a two-field formulation for the two unknown fields, λ and λ_x. Solving this linear problem with a Krylov method requires at each step one subdomain solve with Dirichlet and one with Neumann conditions,

$$\begin{aligned} \mathcal{A}_{pd}\begin{pmatrix} \lambda \\ \lambda_x \end{pmatrix} &:= \begin{bmatrix} 1 & -u_2(0, y, 0, \cdot) \\ -\partial_x u_1(0, y, 0, \cdot) & 1 \end{bmatrix}\begin{pmatrix} \lambda \\ \lambda_x \end{pmatrix} \\ &= \begin{pmatrix} u_2(0, y, f, 0) \\ \partial_x u_1(0, y, f, 0) \end{pmatrix} =: b_{pd}. \end{aligned} \tag{5.16}$$

Taking a Fourier transform of the operator $\mathcal{A}_{pd}$, we find

$$\hat{\mathcal{A}}_{pd} = \begin{bmatrix} 1 & \frac{1}{\sqrt{\eta+k^2}} \\ -\sqrt{\eta+k^2} & 1 \end{bmatrix}, \tag{5.17}$$

which is precisely the matrix to which we have applied a Richardson iteration trying simply to relax the interface conditions in (5.4), an iteration which did not converge. By applying a Krylov method to solve the problem directly, however, it would converge in two steps, since the eigenvalues are independent of k, there are only two distinct points in the spectrum.

We can also write the coupled subdomain problems with overlap in substructured form. If we give the unknown functions at the interfaces the names $\lambda_0(y)$ and $\lambda_L(y)$, we get

$$\begin{array}{cc} \mathcal{L}(u_1) = f \text{ in } \Omega_1, & \mathcal{L}(u_2) = f \text{ in } \Omega_2, \\ u_1(L,y) = u_2(L,y) \ =: \ \lambda_L(y) & u_2(0,y) = u_1(0,y) \ =: \ \lambda_0(y). \end{array} \tag{5.18}$$

If we assume that both λ_L and λ_0 are known, then we can compute $u_1(x,y,f,\lambda_L)$ and $u_2(x,y,f,\lambda_0)$ and then set

$$\begin{aligned} u_2(0,y,f,\lambda_0) - u_1(0,y,f,\lambda_L) &= 0, \\ -u_2(L,y,f,\lambda_0) + u_1(L,y,f,\lambda_L) &= 0, \end{aligned}$$

which is a linear system of equations for the unknowns λ_0 and λ_L. Solving this linear problem with a Krylov method requires at each step two subdomain solves with Dirichlet conditions,

$$\mathcal{A}_s \begin{pmatrix} \lambda_0 \\ \lambda_L \end{pmatrix} := \begin{bmatrix} 1 & -u_1(0,y,0,\cdot) \\ -u_2(L,y,0,\cdot) & 1 \end{bmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_L \end{pmatrix} = \begin{pmatrix} u_1(0,y,f,0) \\ u_2(L,y,f,0) \end{pmatrix} =: b_s. \tag{5.19}$$

In Fourier the symbol of the operator $\mathcal{A}_s$ is given by

$$\hat{\mathcal{A}}_s = \begin{bmatrix} 1 & -e^{-\sqrt{\eta+k^2}L} \\ -e^{-\sqrt{\eta+k^2}L} & 1 \end{bmatrix}, \tag{5.20}$$

and we see that the operator is symmetric in this case. If one applies a Richardson iteration to this operator, one recovers the classical Schwarz method for which we have seen that it converges independently of the discretization parameter. The eigenvalues in Fourier are $1 \pm e^{-\sqrt{\eta+k^2}L}$, which shows that the eigenvalues are clustering for large k around 1, a very desirable property when a Krylov method is used to solve the corresponding linear system. The condition number of this symmetric operator can be estimated by the ratio of the largest and smallest eigenvalue,

$$\mathcal{K}(\mathcal{A}_s) = \frac{1+e^{-\sqrt{\eta}L}}{1-e^{-\sqrt{\eta}L}}, \tag{5.21}$$

and it is independent of the mesh parameter h, as long as the overlap L is independent of h. For an overlap which depends on h, $L = h$, we have for h small

$$\mathcal{K}(\mathcal{A}_s) = \frac{2}{\sqrt{\eta}h} + O(h) \tag{5.22}$$

as for the primal and dual Schur methods.

5.2. Coupling conditions optimized for the computation. Optimized Schwarz methods bring the overlapping and nonoverlapping strategies together. They do not use either Dirichlet or Neumann conditions, and they work with or without overlap. The fundamental idea is that the coupled problem can be written with any set of conditions that implies the classical coupling conditions. The coupled problems

$$\begin{array}{ll} \mathcal{L}(u_1) = f \quad \text{in } \Omega_1, & \mathcal{L}(u_2) = f \quad \text{in } \Omega_2, \\ (\partial_x + \mathcal{S}_1)(u_1)(L) = (\partial_x + \mathcal{S}_1)(u_2)(L), & (\partial_x + \mathcal{S}_2)(u_2)(0) = (\partial_x + \mathcal{S}_1)(u_1)(0), \end{array} \tag{5.23}$$

are equivalent to the original, unpartitioned problem, as long as the choice of $\mathcal{S}_j$, $j = 1, 2$, leads to well-posed subdomain problems and implies, for $L > 0$, $u_1(0) = u_2(0)$ and $u_1(L) = u_2(L)$, and for $L = 0$, $u_1(0) = u_2(0)$ and $\partial_x u_1(0) = \partial_x u_2(0)$. To write this system in substructured form, we assume again that the interface functions $\lambda_1(y)$ and $\lambda_2(y)$ are known,

$$\begin{aligned} (\partial_x + \mathcal{S}_1)(u_1)(L, y) = (\partial_x + \mathcal{S}_1)(u_2)(L, y) =: \lambda_1(y), \\ (\partial_x + \mathcal{S}_2)(u_2)(0, y) = (\partial_x + \mathcal{S}_1)(u_1)(0, y) =: \lambda_2(y), \end{aligned}$$

solve the subdomain problems, and then set

$$\begin{aligned} -(\partial_x + \mathcal{S}_1)(u_2(0, y, f, \lambda_2)) + \lambda_1 = 0, \\ \lambda_2 - (\partial_x + \mathcal{S}_2)(u_1(L, y, f, \lambda_1)) = 0. \end{aligned}$$

This is again a linear system to be solved for λ_1 and λ_2. Using a Krylov method, at each iteration two problems with the new transmission conditions need to be solved,

$$\begin{aligned} \mathcal{A}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} &:= \begin{bmatrix} 1 & -(\partial_x + \mathcal{S}_1)(u_2(0, y, 0, \cdot)) \\ -(\partial_x + \mathcal{S}_2)(u_1(L, y, 0, \cdot)) & 1 \end{bmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_L \end{pmatrix} \\ &= \begin{pmatrix} (\partial_x + \mathcal{S}_1)(u_2(0, y, f, 0)) \\ (\partial_x + \mathcal{S}_2)(u_1(L, y, f, 0)) \end{pmatrix} =: b. \end{aligned} \tag{5.24}$$

In the Fourier domain, the symbol of the operator $\mathcal{A}$ becomes for our model problem

$$\hat{\mathcal{A}} = \begin{bmatrix} 1 & -\frac{\sqrt{\eta + k^2} - \sigma_1(k)}{\sqrt{\eta + k^2} - \sigma_2(k)} e^{-\sqrt{\eta + k^2} L} \\ -\frac{\sqrt{\eta + k^2} + \sigma_2(k)}{\sqrt{\eta + k^2} + \sigma_1(k)} e^{-\sqrt{\eta + k^2} L} & 1 \end{bmatrix}. \tag{5.25}$$

For well-posedness of the subdomain problems, we need that $\mathcal{S}_1$ is a positive operator and $\mathcal{S}_2$ a negative one, as one can also see from the denominators in the symbol of the operator $\mathcal{A}$. The iterative optimized Schwarz method is obtained when a Richardson iteration is applied to this system, and we have seen that this iteration converges in two steps, if $= \sigma_2 = -\sigma_1 = \sqrt{\eta + k^2}$, or very fast, if the symbols approximate this choice. If we choose $\sigma_2 = -\sigma_1 > 0$, then the operator becomes symmetric and its condition number equals one for the optimal choice, or it can be made small choosing good approximations. This is the heart of the optimized Schwarz methods: the optimal choice always exists, it is the Dirichlet-to-Neumann map, and good approximations lead to the optimized Schwarz methods with superior performance. The FETI methods have also started to incorporate these ideas; see, for example, the variant FETI-H presented in [12], where the authors state, "The modified Lagrangian formulation presented here can be related to alternative transmission conditions for the subdomain interfaces." FETI-H constructs (5.25) using optimized Robin conditions.

TABLE 6.1
Number of iterations of the classical Schwarz method compared to the different optimized Schwarz methods with fixed small overlap of the size $L = \frac{1}{50}$ between subdomains.

	Classical	Taylor 0	Taylor 2	Optimized 0	Two-sided optimized 0	Optimized 2
h	Schwarz as an iterative solver					
1/50	65	16	11	7	6	4
1/100	77	17	12	7	6	4
1/200	86	16	11	7	6	4
1/400	91	16	12	7	6	4
1/800	93	16	11	7	6	4
	Schwarz use as a preconditioner					
1/50	11	8	7	5	5	3
1/100	12	8	7	5	5	3
1/200	13	8	7	5	5	3
1/400	13	8	7	5	5	3
1/800	13	8	7	5	5	3

6. Numerical experiments. We perform numerical experiments for our model problem on the unit square, $\Omega = (0,1) \times (0,1)$,

$$\begin{aligned} (\eta - \Delta)(u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{6.1}$$

We decompose the unit square Ω into two subdomains $\Omega_1 = (0,\beta) \times (0,1)$ and $\Omega_2 = (\alpha,1) \times (0,1)$, where $0 < \alpha \leq \beta < 1$ and hence the overlap is $L = \beta - \alpha$. Note that we explicitly allow $\alpha = \beta$ such that the method does not have any overlap, $L = 0$. We use a finite difference discretization with the classical five-point discretization for the Laplacian and a uniform mesh with mesh parameter h.

6.1. Overlapping optimized Schwarz methods. Classically the overlap in the Schwarz method is held constant as the mesh is refined to obtain mesh independent convergence factors for the method. The same is true for optimized Schwarz methods because of Theorem 4.1, as iteration counts to reach an error reduction of $1e{-}6$ show in Table 6.1 for a fixed overlap $L = \beta - \alpha = \frac{1}{50}$. We simulate directly the error equations, $f = 0$, and use a random initial guess so that all the frequency components are present. The results show clearly how important transmission conditions are for this algorithm. Note also that while the Krylov method has a big impact on the classical Schwarz method, for the second order optimized Schwarz method the acceleration with the Krylov method does not reduce the iteration count significantly. This situation is well known for multigrid methods, which do not need Krylov acceleration either when applied to a Poisson problem. The Krylov acceleration is then used to improve the performance of the method on more complex problems.

In practical computations, one can often not afford many mesh cells to overlap, so the overlap depends on the mesh parameter h. In the following experiments we choose therefore the overlap $L = \beta - \alpha = h$. Table 6.2 shows the iteration counts for this case. It is interesting to note that the second order optimized Schwarz method without Krylov acceleration is already six times faster than classical Schwarz with Krylov acceleration at high resolution.

In Figure 6.1, we show the number of iterations on a log-log plot so they can be compared to the theoretical asymptotic results. On the top, the Schwarz methods are used as iterative solvers and the numerical results show the asymptotic behavior predicted by the theory. On the bottom, the Schwarz methods are used as preconditioners. This improves the asymptotic performance by a square root, as one can

TABLE 6.2

Number of iterations of the classical Schwarz method compared to the different optimized Schwarz methods with overlap $L = h$ between subdomains.

	Classical	Taylor 0	Taylor 2	Optimized 0	Two-sided optimized 0	Optimized 2
h	Schwarz as an iterative solver					
1/50	65	16	11	7	6	4
1/100	127	22	16	8	7	4
1/200	257	31	21	11	9	5
1/400	510	42	30	13	10	6
1/800	1020	60	41	16	12	7
	Schwarz use as a preconditioner					
1/50	11	8	7	5	5	3
1/100	16	9	8	6	6	4
1/200	21	11	9	6	6	4
1/400	31	13	11	7	7	4
1/800	42	16	13	8	8	5

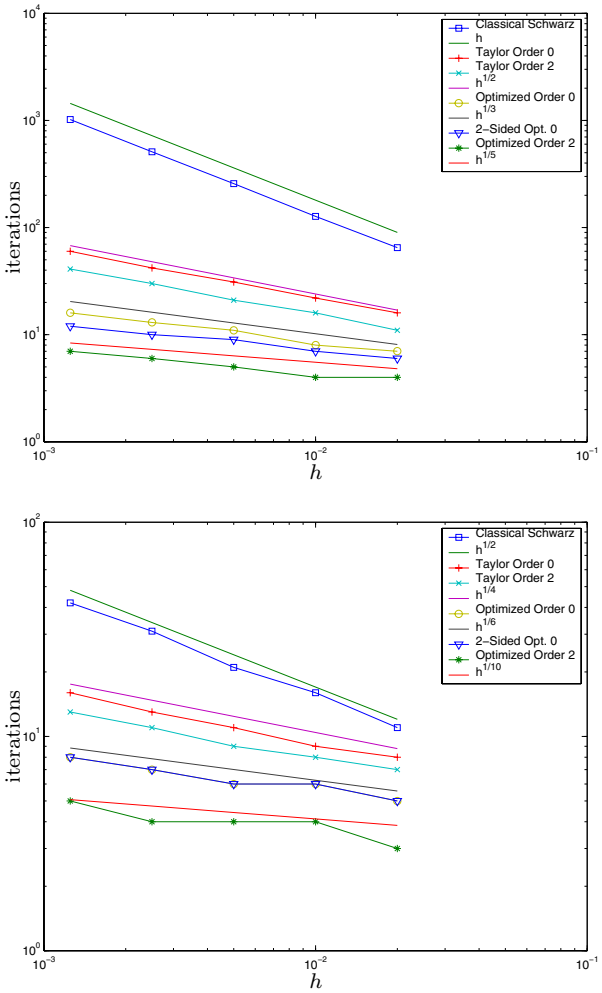


FIG. 6.1. *Number of iterations required by the classical and the optimized Schwarz methods, with overlap $L = h$. On the top the methods are used as iterative solvers, and on the bottom they are used as preconditioners for a Krylov method.*

show in ideal situations, since a square root is taken off the condition number of the preconditioned system. This is also visible in our numerical results.

We now investigate how well the continuous analysis predicts the optimal parameters to be used in the numerical setting. To this end we vary the parameter p in the Robin transmission conditions for a fixed problem of mesh size $h = \frac{1}{100}$ and count for each value of p the number of iterations to reach a residual of $1e-6$. The results for both optimized Schwarz used as an iterative solver and as a preconditioner are shown in Figure 6.2 on the top. The analysis predicts very well the optimal parameter, and when the method is used as a preconditioner, the area where the optimum is attained is widened considerably, which shows that the optimized Schwarz method is robust with respect to the optimal parameter. Similar results hold for the second order optimized Schwarz method, as one can see in Figure 6.2 in the middle when the method is used iteratively and on the bottom when used as a preconditioner.

6.2. Nonoverlapping optimized Schwarz methods. Nonoverlapping Schwarz methods are of interest if the physical properties vary from subdomain to subdomain and one has formulated the subdomain decomposition motivated by this fact; see, for example, [20]. They also facilitate the construction of nonmatching grids per subdomain and the formulation of algorithms in that case. We illustrate the performance of the optimized Schwarz methods without overlap, $\alpha = \beta$ or $L = 0$, by choosing for the mesh parameter diminishing values and counting the number of iterations the methods take to reduce the error by a factor $1e-6$. Table 6.3 shows the performance of the different optimized Schwarz methods in that case. Note that the classical Schwarz method is not shown because classical Schwarz does not converge without overlap. Comparing with the performance of the methods with overlap h, one can see that the number of iterations is by a factor 1.5–1.7 higher for the second order optimized Schwarz method, whereas the cost per subdomain is only slightly higher for the method with overlap; there are m more variables in one subdomain for matrices of size m^2. Hence a physical motivation must outweigh the increased cost of a nonoverlapping Schwarz method.

In Figure 6.3 we show the number of iterations on a log-log plot so they can be compared to the theoretical asymptotic factors.

On the left the methods are used as iterative solvers and one can see that again the numerical results show the asymptotic behavior predicted by the analysis. On the right the results are shown when the Schwarz methods are used as preconditioners, and one can see again that Krylov acceleration improves the performance by about a square root.

We finally show in Figure 6.4 how well the analysis predicts the optimization parameters in the nonoverlapping case.

6.3. An application. We now show how a nonoverlapping optimized Schwarz method can be used to compute the temperature distribution in our apartment on Durocher in Montreal. In Figure 6.5, we show on top the floor plan of our apartment with a finite element discretization and a decomposition into the different rooms: on the left is the living room, connected to the kitchen and with a long hallway to the bathroom and bedroom on the right. Insulated walls are shown in blue, the windows on top are shown in black, where we assume -20 degrees Celsius for a regular Montreal winter day, and the doors at the bottom and on the right are also shown in black. They lead to a heated public hallway, at about 15 degrees Celsius. The interfaces are shown in red, and we introduced curved interfaces and nonrectangular domains, so that the Fourier analysis presented in this paper cannot be applied any more. In

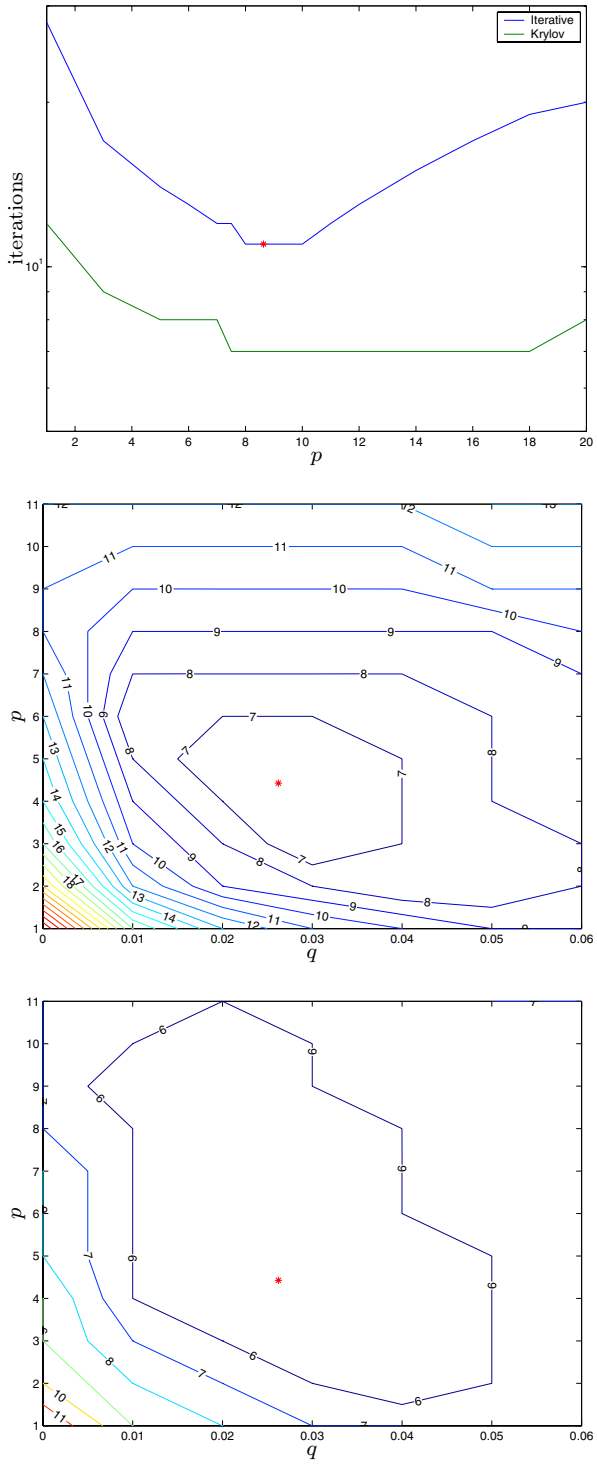


FIG. 6.2. *Optimal parameter (*) found by the analytical optimization compared to the performance of other values of the parameters: on the top for the Robin case, in the middle for the second order case used iteratively, and on the bottom used as a preconditioner.*

TABLE 6.3
Number of iterations of different optimized Schwarz methods without overlap between subdomains.

	Taylor 0	Taylor 2	Optimized 0	Two-sided optimized 0	Optimized 2
h	Optimized Schwarz as an iterative solver				
1/50	425	109	23	13	6
1/100	847	217	31	16	7
1/200	1702	434	44	20	9
1/400	3432	875	62	25	10
1/800	6824	1746	88	30	12
	Optimized Schwarz as a preconditioner				
1/50	21	15	9	8	5
1/100	28	20	11	10	5
1/200	35	26	13	11	6
1/400	46	34	15	12	6
1/800	59	45	18	13	7

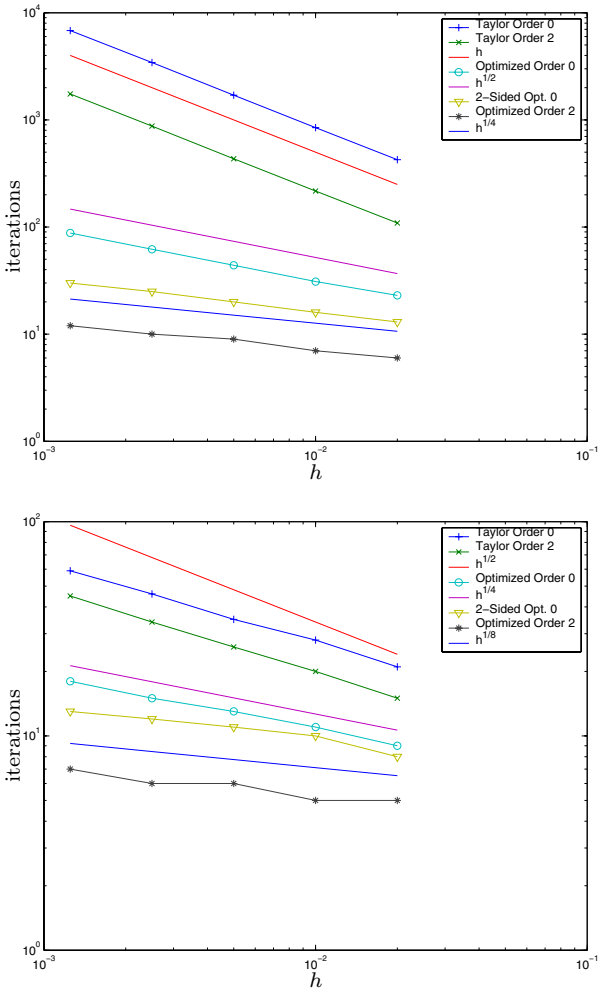


FIG. 6.3. *Asymptotic number of iterations required by the nonoverlapping optimized Schwarz methods: on the top the methods are used as iterative solvers, and on the bottom they are used as preconditioners for a Krylov method.*

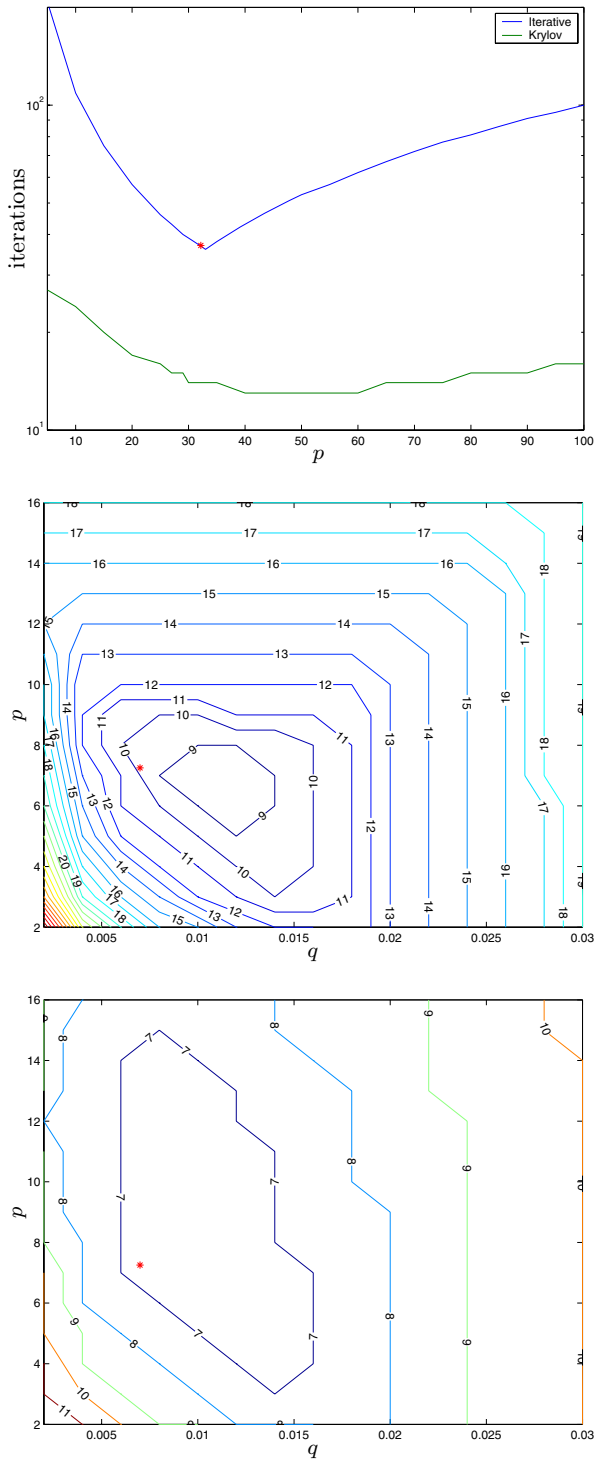


FIG. 6.4. *Comparison of the optimal parameter (*) found in the analysis with the numerical performance of other values of the parameters: for the Robin case on the top, the second order case used iteratively in the middle, and as a preconditioner on the bottom.*

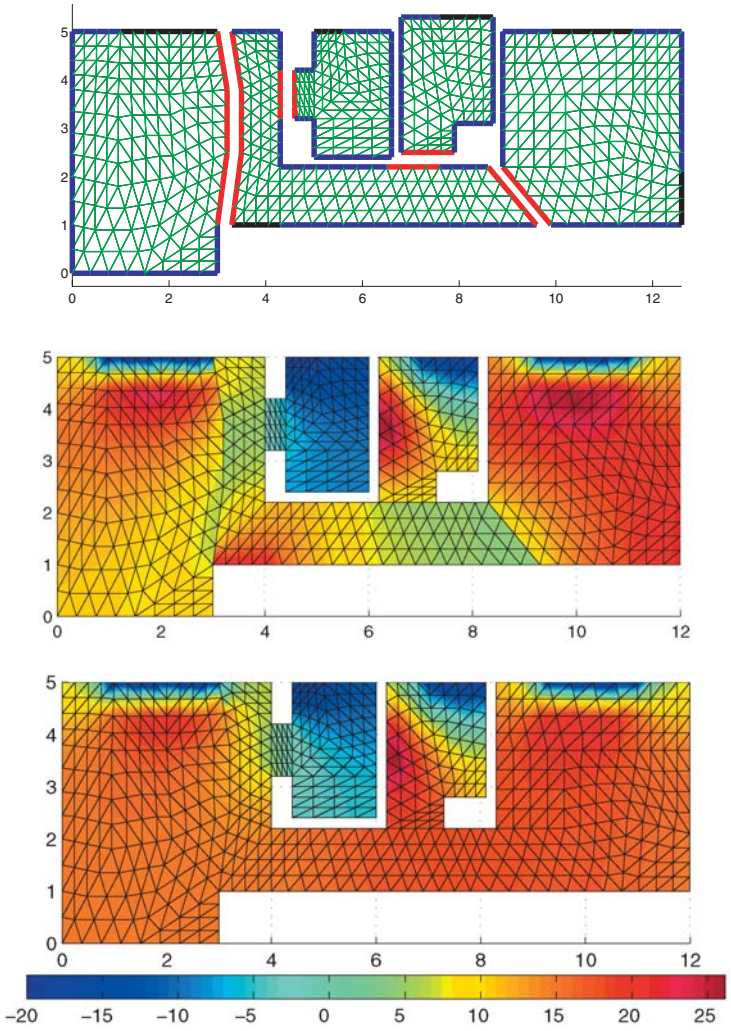


FIG. 6.5. *On top the decomposition of a two-dimensional cross section of an apartment in Montreal, in the middle the first iteration, and at the bottom the final temperature distribution computed in winter with an optimized Schwarz method.*

the middle in Figure 6.5 we show the first iteration of the optimized Schwarz method with Robin transmission conditions, where one can clearly see the isolated effect of the heaters and warm doors in each subdomain: the iterate is discontinuous. In Figure 6.5 at the bottom we show the final result of the simulation, which is now continuous. The method took 25 iterations to converge to a relative residual of $1e-3$ in the iterative case and 12 iterations when used as a preconditioner, using the optimal parameter $p^* = 2.7207$ from the two-subdomain theory. Refining once more, the method took 32 iterations in the iterative case and 13 in the preconditioned case, with the optimal parameter $p^* = 3.8576$ from the two-subdomain theory. The ratio in the iterative case is $32/25 = 1.28 \approx 2^{1/3} = 1.26$, as predicted by the two-subdomain theory for the simple two-subdomain case with straight interfaces, and in the preconditioned case, the ratio is $13/12 = 1.08 \approx 2^{1/6} = 1.12$. This shows that although the Fourier analysis cannot be applied in the more general case, the results predicted by the theory for

the two-subdomain case are also observed in more practical situations.

The results of this simulation were interesting to us: one can see that while the heaters in the living room on the left and the bedroom on the right are well placed to block the cold from the windows, the heater on the left wall in the bathroom is not effective to keep the room warm, a fact we strongly felt in winter. Also, the kitchen is not heated and stays cold, except when cooking and baking.

7. Conclusion. We introduced the reader to a new class of Schwarz methods, the optimized Schwarz methods. The algorithm is the same as for the classical Schwarz method and it can be used either iteratively or as a preconditioner. The difference is a new type of transmission conditions between subdomains, instead of the classical Dirichlet condition. We analyzed for a symmetric positive definite model problem and two subdomains the influence of the transmission conditions on the convergence factor of the Schwarz algorithm. We showed both analytically and numerically that the optimized Schwarz methods have a greatly improved performance compared to the classical Schwarz method. The number of iterations required to achieve a certain accuracy is by a factor smaller, often more than an order of magnitude. This performance is achieved without an increased cost for the subdomain solves, since the same type of matrix problem has to be solved in the subdomains, and the new subdomain matrices have the same bandwidth as the original ones. We also proved that the optimized Schwarz methods are always faster than the classical Schwarz method and since their implementation is not more difficult than the implementation of a classical Schwarz method, they represent a very attractive alternative. We finally showed in numerical experiments that the results derived for the simple two-subdomain configuration with a straight interface also apply in more complicated situations in practice, where Fourier analysis cannot be applied any more.

REFERENCES

[1] P. E. Bjørstad and O. B. Widlund, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1093–1120.

[2] J.-F. Bourgat, R. Glowinski, P. Le Tallec, and M. Vidrascu, *Variational formulation and algorithm for trace operator in domain decomposition calculations*, in Domain Decomposition Methods, T. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, 1989, pp. 3–16.

[3] X.-C. Cai, M. A. Casarin, F. W. Elliott, Jr., and O. B. Widlund, *Overlapping Schwarz algorithms for solving Helmholtz's equation*, in Domain Decomposition Methods 10 (Boulder, CO, 1997), AMS, Providence, RI, 1998, pp. 391–399.

[4] T. F. Chan and T. P. Mathew, *Domain decomposition algorithms*, in Acta Numerica 1994, Cambridge University Press, Cambridge, UK, 1994, pp. 61–143.

[5] P. Charton, F. Nataf, and F. Rogier, *Méthode de décomposition de domaine pour l'équation d'advection-diffusion*, C.R. Acad. Sci., 313 (1991), pp. 623–626.

[6] P. Chevalier and F. Nataf, *Symmetrized method with optimized second-order conditions for the Helmholtz equation*, in Domain Decomposition Methods 10 (Boulder, CO, 1997), AMS, Providence, RI, 1998, pp. 400–407.

[7] Q. Deng, *An analysis for a nonoverlapping domain decomposition iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1525.

[8] B. Després, *Décomposition de domaine et problème de Helmholtz*, C.R. Acad. Sci. Paris, 1 (1990), pp. 313–316.

[9] B. Després, P. Joly, and J. E. Roberts, *A domain decomposition method for the harmonic Maxwell equations*, in Iterative Methods in Linear Algebra (Brussels, 1991), North–Holland, Amsterdam, 1992, pp. 475–484.

[10] B. Engquist and H.-K. Zhao, *Absorbing boundary conditions for domain decomposition*, Appl. Numer. Math., 27 (1998), pp. 341–365.

[11] C. Farhat, M. Lesoinne, P. Le Tallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-*

primal unified FETI method—part I*: A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[12] C. Farhat, A. Macedo, and R. Tezaur, *FETI-H: A scalable domain decomposition method for high frequency exterior Helmholtz problem*, in Proceedings of the 11th International Conference on Domain Decomposition Method, C.-H. Lai, P. Bjørstad, M. Cross, and O. Widlund, eds., DDM.ORG, Augsburg, 1999, pp. 231–241.

[13] C. Farhat and F.-X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.

[14] M. J. Gander, *Optimized Schwarz methods for Helmholtz problems*, in Proceedings of the 13th International Conference on Domain Decomposition, 2001, pp. 245–252.

[15] M. J. Gander and G. H. Golub, *A non-overlapping optimized Schwarz method which converges with an arbitrarily weak dependence on h*, in Proceedings of the 14th International Conference on Domain Decomposition Methods, 2002.

[16] M. J. Gander and L. Halpern, *Méthodes de décomposition de domaines pour l'équation des ondes en dimension* 1, C.R. Acad. Sci. Paris Ser. I, 333 (2001), pp. 589–592.

[17] M. J. Gander and L. Halpern, *Méthodes de relaxation d'ondes pour l'équation de la chaleur en dimension* 1, C.R. Acad. Sci. Paris Ser. I, 336 (2003), pp. 519–524.

[18] M. J. Gander, L. Halpern, and C. Japhet, *Optimized Schwarz algorithms for coupling convection and convection-diffusion problems*, in Proceedings of the 13th International Conference of Domain Decomposition, 2001, pp. 253–260.

[19] M. J. Gander, L. Halpern, and F. Nataf, *Optimal convergence for overlapping and non-overlapping Schwarz waveform relaxation*, in Proceedings of the 11th International Conference of Domain Decomposition Methods, C.-H. Lai, P. Bjørstad, M. Cross, and O. Widlund, eds., 1999.

[20] M. J. Gander, L. Halpern, and F. Nataf, *Optimal Schwarz waveform relaxation for the one dimensional wave equation*, Tech. Rep. 469, CMAP, Ecole Polytechnique, Sept. 2001.

[21] M. J. Gander, L. Halpern, and F. Nataf, *Optimized Schwarz methods*, in Proceedings of the 12th International Conference on Domain Decomposition Methods, Chiba, Japan, T. Chan, T. Kako, H. Kawarada, and O. Pironneau, eds., Domain Decomposition Press, Bergen, 2001, pp. 15–28.

[22] M. J. Gander, F. Magoulès, and F. Nataf, *Optimized Schwarz methods without overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 24 (2002), pp. 38–60.

[23] M. J. Gander and A. M. Stuart, *Space-time continuous analysis of waveform relaxation for the heat equation*, SIAM J. Sci. Comput., 19 (1998), pp. 2014–2031.

[24] T. Hagstrom, R. P. Tewarson, and A. Jazcilevich, *Numerical experiments on a domain decomposition algorithm for nonlinear elliptic boundary value problems*, Appl. Math. Lett., 1 (1988).

[25] C. Japhet, *Conditions aux limites artificielles et décomposition de domaine: Méthode oo*2 *(optimisé d'ordre* 2*). application à la résolution de problèmes en mécanique des fluides*, Tech. Rep. 373, CMAP, Ecole Polytechnique, 1997.

[26] C. Japhet, *Optimized Krylov-Ventcell method. Application to convection-diffusion problems*, in Proceedings of the 9th International Conference on Domain Decomposition Methods, P. E. Bjørstad, M. S. Espedal, and D. E. Keyes, eds., 1998, pp. 382–389.

[27] C. Japhet and F. Nataf, *The best interface conditions for domain decomposition methods: Absorbing boundary conditions*, in Absorbing Boundaries and Layers. Domain Decompsition Methods, L. Tourrette and L. Halpern, eds., Nova Science, 2001, pp. 348-373.

[28] C. Japhet, F. Nataf, and F. Rogier, *The optimized order* 2 *method. Application to convection-diffusion problems*, Future Generation Computer Systems, 18 (2001), pp. 17–30.

[29] C. Japhet, F. Nataf, and F.-X. Roux, *The Optimized Order* 2 *Method with a coarse grid preconditioner. Application to convection-diffusion problems*, in Ninth International Conference on Domain Decompositon Methods in Science and Engineering, P. Bjorstad, M. Espedal, and D. Keyes, eds., John Wiley & Sons, New York, 1998, pp. 382–389.

[30] P.-L. Lions, *On the Schwarz alternating method.* I, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, 1988, pp. 1–42.

[31] P.-L. Lions, *On the Schwarz alternating method.* III: *A variant for nonoverlapping subdomains*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, 1990.

[32] F. Nataf, *Absorbing boundary conditions in block Gauss-Seidel methods for convection problems*, Math. Models Methods Appl. Sci., 6 (1996), pp. 481–502.

[33] F. Nataf and F. Rogier, *Factorization of the convection-diffusion operator and the Schwarz*

algorithm, Math. Models Methods Appl. Sci., 5 (1995), pp. 67–93.
[34] F. Nataf, F. Rogier, and E. de Sturler, *Optimal interface conditions for domain decomposition methods*, Tech. Rep. 301, CMAP, Ecole Polytechnique, 1994.
[35] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, London, 1999.
[36] H. A. Schwarz, *Über einen Grenzübergang durch alternierendes Verfahren*, Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich, 15 (1870), pp. 272–286.
[37] B. F. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[38] H. Sun and W.-P. Tang, *An overdetermined Schwarz alternating method*, SIAM J. Sci. Comput., 17 (1996), pp. 884–905.
[39] W. P. Tang, *Generalized Schwarz splittings*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 573–595.
[40] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[41] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 732–752

L^2-PROJECTED LEAST-SQUARES FINITE ELEMENT METHODS FOR THE STOKES EQUATIONS*

HUO-YUAN DUAN†, PING LIN†, P. SAIKRISHNAN†, AND ROGER C. E. TAN†

Abstract. Two new L^2 least-squares (LS) finite element methods are developed for the velocity-pressure-vorticity first-order system of the Stokes problem with Dirichlet velocity boundary condition. A key feature of these new methods is that a local or almost local L^2 projector is applied to the residual of the momentum equation. Such L^2 projection is always defined onto the linear finite element space, no matter which finite element spaces are used for velocity-pressure-vorticity variables. Consequently, the implementation of this L^2-projected LS method is almost as easy as that of the standard L^2 LS method. More importantly, the former has optimal error estimates in L^2-norm, with respect to both the order of approximation and the required regularity of the exact solution for velocity using equal-order interpolations and for all three variables (velocity, pressure, and vorticity) using unequal-order interpolations. Numerical experiments are given to demonstrate the theoretical results.

Key words. the Stokes equation, velocity-pressure-vorticity least-squares finite element method, L^2 projection, mass-lumping

AMS subject classification. 65N30

DOI. 10.1137/040613573

1. Introduction.

The least-squares (LS) mixed finite element method is widely used in seeking numerical solution of partial differential equations arising from fluid and solid mechanics; cf. [13, 14, 15, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 30, 31, 32]. In a broad sense, the LS method minimizes the residual, measured in some Sobolev norms, of a mixed first-order system of partial differential equations. The mixed first-order system is obtained by introducing one or more additional physically important fields such as stress/pressure/vorticity besides displacement/velocity as unknown variables. There are many advantages to LS methods. The LS method may be viewed as a classical Ritz's method of coercive type [29] and is not subject to the so-called inf-sup condition [2, 9]. Its resulting linear system is symmetric positive definite and can be solved by standard iterative methods such as the conjugate gradient method. In addition, the standard finite element spaces can be employed for each unknown variable. Readers may refer to [12, 13] for more details on LS methods. In this paper we shall introduce and study new LS methods for the Stokes problem written as a system of equations of first order, where velocity, pressure, and vorticity appear as unknown variables. This system involves relatively few unknowns and is widely employed in engineering practice.

Let us first review several LS methods developed in the last decade for the velocity-pressure-vorticity Stokes system. The most widely used LS method is the standard L^2 LS method [13, 15], where the LS functional is the squared L^2-norms of the residual of the first-order system. This method is easy to implement and performs very well in many engineering applications (cf. [13, 19, 20, 30, 31]). However, for the important case of Dirichlet velocity boundary condition, this method is not optimal in the usual

*Received by the editors August 18, 2004; accepted for publication (in revised form) December 1, 2005; published electronically March 31, 2006. This research was supported by NUS Academic Research grant R-146-000-036-592.

http://www.siam.org/journals/sinum/44-2/61357.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (scidhy@nus.edu.sg, matlinp@nus.edu.sg, matsp@nus.edu.sg, scitance@nus.edu.sg).

sense [16, 17]; for example, for equal-order continuous interpolations, the L^2-error bound for velocity is not optimal with respect to both the order of approximation and the required regularity. In the case of convex polygon, no error estimates are available. Also, no improved error estimates are obtained for unequal-order continuous interpolations (see [33] for some numerical results).

The reason that the standard L^2 LS method suffers from suboptimal error estimates in the case of velocity Dirichlet boundary condition may be the following: the coercivity for vorticity and pressure is measured in L^2-norm, whereas their first-order derivatives having appeared in the term resulting from the momentum equation suggests that the continuity condition cannot be obtained in the same measure, which prevents us from obtaining optimal error estimates. To recover the optimal error estimates one has to do some modifications to that term of the momentum equation.

There have been two important methods which can overcome the difficulty from the term of the momentum equation. One is the Bochev–Gunzburger (BG) method, where a factor h^2 is put in front of the term of the momentum equation. Alternatively, the BG method may be scaled as the one with a factor h^{-2} put in front of terms of the incompressibility condition and of the vorticity equation [18]. The other is the H^{-1} LS method, which may be viewed as a modified version of the BG method by introducing an additional term of the momentum equation, where a preconditioner B_h (or an operator of the finite element solution) for the Dirichlet problem of a second-order elliptic equation is applied [14, 21, 32]. In the BG method the effects from the term of the momentum equation can be eliminated because of the factor h^2, and optimal error estimates can be derived with the use of unequal-order continuous interpolations [16, 18]. In the H^{-1} method, when B_h satisfies a spectral equivalence (see equation (2.15) in [21, p. 941]), the coercivity and the optimal error estimates can be established. These are excellent efforts in achieving optimal error estimates of LS methods. However, there is still room for improvement. The BG method does not give optimal L^2-error estimates for the velocity, excludes the use of linear elements [16, 18, 33], and has a condition number $\mathcal{O}(h^{-4})$. This is due to the fact that it lacks a coercivity uniform in mesh sizes or that in its scaled version the scale factor h^{-2} worsens its continuity condition. The H^{-1} method can provide optimal L^2-error bounds for the velocity, but there is a restriction on B_h (see equation (3.24) in [21, p. 947]). There are few examples of B_h known to satisfy that restriction. Also, due to the fact that B_h is defined onto U_h (the approximating space for the velocity), B_h varies with U_h accordingly. This may complicate implementation issues when U_h is of higher-order elements.

Our new idea presented in the paper is to add an L^2-projected term of the momentum equation to the BG method or to use an L^2 projector to replace the preconditioner in the H^{-1} method. With the L^2 projection term, the uniform coercivity holds (see Theorem 3.1), and the error estimates are optimal (see Theorem 3.3 and Theorem 4.1), for velocity using equal-order interpolations and for all three variables (velocity, pressure, and vorticity) using unequal-order interpolations. Also, the condition number is of $\mathcal{O}(h^{-2})$ (see Corollary 3.2) and the implementation of this L^2-projected LS method is almost as easy as that of the standard L^2 LS method, since the L^2 projection is local or almost local and is always defined onto the linear finite element space, no matter which finite element spaces are used for velocity-pressure-vorticity variables. Note that although the L^2 projection is "fixed" onto the linear element space, this does not cause any consistency problem and does not affect the order of the error estimates.

We provide two methods according to the definitions of L^2 projectors applied to the term of the momentum equation. One is called the local L^2 projection method (I) and the other the mass-lumping L^2 projection method (II). The L^2 projection in method (I) is always element-by-element defined onto the discontinuous linear element space; in method (II) the L^2 projection is always defined by using the mass-lumping technique [1] onto the continuous linear element space. Note that the L^2 projection in method (II) is almost local because the resulting matrix of this L^2 projection is diagonal. Standard equal-order or unequal-order finite elements, with lower-order finite elements for pressure and vorticity enriched with element or edge (face) bubbles or both, are employed for approximating velocity, pressure, and vorticity variables. Note that the role of the bubbles for lower-order elements for pressure and vorticity is to make Assumption (A2) (see (3.35) and (3.36)) hold (cf. Remark 3.2 and Theorem 3.4).

Our L^2 projection plays a critical role (see (3.24)) in the derivation of a uniform coercivity (see (3.6)). All the first-order derivatives of pressure and vorticity appear only in L^2-projected and h^2-weighted terms. Due to Assumption (A2), the errors associated with the L^2-projected term can be made zero (see (3.47)), while the h^2-weighted term is obviously consistent in terms of both the order of approximation and the regularity of the exact solution. Therefore, optimal error estimates can be achieved. Also, an $\mathcal{O}(h^{-2})$ condition number is obtained.

Of course, we may project onto a higher-order element space. But obviously the linear element is simpler and the L^2 projection can be easily implemented. For method (I) we could even consider defining the local L^2 projection onto a piecewise constant space and almost all our techniques of analysis might still work. However, an optimal L^2-error bound for velocity cannot be obtained because the interpolation result of the linear element has to be used in the derivation (see (4.20)).

Finally, we remark that in deriving the L^2-error estimates for the velocity we assume that the domain is a convex polygon as usual [5, 29]. For such a domain some known regularity results for Stokes and elasticity problems (cf. [6, 7, 8, 10]) are used.

The outline of this paper is as follows. In section 2, we recall the first-order system of the velocity-pressure-vorticity Stokes problem and formulate L^2-projected methods. In section 3, we establish coercivity and error bounds and verify an important assumption (Assumption (A2)). In section 4, the L^2-error bound for velocity is obtained. In section 5, numerical results are presented to support our theoretical analysis.

2. Problem formulation.

2.1. First-order system of the Stokes problem. Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with boundary Γ and $\mathbf{f} \in (L^2(\Omega))^d$. We consider the following Stokes problem: Find velocity $\mathbf{u}$ and pressure p such that

$$-\Delta \mathbf{u} + \bigtriangledown p = \mathbf{f}, \quad \bigtriangledown \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma. \tag{2.1}$$

Let $\bigtriangledown\times$ denote the curl operator. Introducing the vorticity $\omega = \bigtriangledown \times \mathbf{u} \in (L^2(\Omega))^{2d-3}$ and noting that $-\Delta \mathbf{u} = \bigtriangledown \times \bigtriangledown \times \mathbf{u} - \bigtriangledown \bigtriangledown \cdot \mathbf{u}$ and $\bigtriangledown \cdot \mathbf{u} = 0$, we can write (2.1) as the first-order system

$$\bigtriangledown \times \omega + \bigtriangledown p = \mathbf{f}, \quad \omega = \bigtriangledown \times \mathbf{u}, \quad \bigtriangledown \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{2.2}$$

along with a Dirichlet boundary condition and a pressure mean-zero condition

$$\mathbf{u}_{|\Gamma} = \mathbf{0}, \quad \int_\Omega p = 0. \tag{2.3}$$

Below we shall use the standard Sobolev spaces $H_0^1(D)$ and $H^s(D)$, with norm $\|\cdot\|_{s,D}$ and seminorm $|\cdot|_{s,D}$, where D is some Lipschitz subdomain of Ω; D will be omitted from the notation when $D = \Omega$. We shall use $(\cdot,\cdot)_{0,D}$ for the inner product of $L^2(D)$ $(= H^0(D))$. When $D = \Omega$, $(\cdot,\cdot) := (\cdot,\cdot)_{0,\Omega}$. We shall also define $L_0^2(\Omega) := \{q \in L^2(\Omega); \int_\Omega q = 0\}$. Throughout this paper we always assume that Ω is a Lipschitz polygon (polyhedron in $\mathbb{R}^3$) and that $\mathcal{C}_h$ is a regular triangulation of Ω (tetrahedrons in $\mathbb{R}^3$), with diameters $h_K \le h$ for all triangular elements $K \in \mathcal{C}_h$.

2.2. Local L^2 projection method (I). Introduce

$$Z_h := \{\mathbf{v} \in (L^2(\Omega))^d; \mathbf{v}_{|K} \in (\mathcal{P}_1(K))^d \quad \forall K \in \mathcal{C}_h\}, \tag{2.4}$$

where $\mathcal{P}_1(K)$ denotes the space of linear polynomials on K. For given $\mathbf{g} \in (L^2(\Omega))^d$ we define a function $\breve{R}_h(\mathbf{g}) \in Z_h$ by

$$\int_K \breve{R}_h(\mathbf{g})\,\mathbf{v} = \int_K \mathbf{g}\,\mathbf{v} \quad \forall \mathbf{v} \in (\mathcal{P}_1(K))^d, \quad \forall K \in \mathcal{C}_h. \tag{2.5}$$

Let

$$U_h \subset (H_0^1(\Omega))^d, \quad P_h \subset L_0^2(\Omega), \quad W_h \subset (L^2(\Omega))^{2\,d-3} \tag{2.6}$$

be continuous piecewise polynomial spaces on $\mathcal{C}_h$ for velocity, pressure, and vorticity, respectively. We define an LS functional on $U_h \times P_h \times W_h$ by

$$\begin{aligned}\mathcal{J}_h^I(\mathbf{u},p,\omega;\mathbf{f}) := \|\breve{R}_h(\bigtriangledown\times\omega + \bigtriangledown p - \mathbf{f})\|_0^2 + \sum_{K\in\mathcal{C}_h} h_K^2\,\|\bigtriangledown\times\omega + \bigtriangledown p - \mathbf{f}\|_{0,K}^2 \\ + \|\omega - \bigtriangledown\times\mathbf{u}\|_0^2 + \|\bigtriangledown\cdot\mathbf{u}\|_0^2.\end{aligned} \tag{2.7}$$

We consider a minimization problem: Find $(\mathbf{u}_h, p_h, \omega_h) \in U_h \times P_h \times W_h$ such that

$$\mathcal{J}_h^I(\mathbf{u}_h,p_h,\omega_h;\mathbf{f}) = \inf_{(\mathbf{v}_h,q_h,z_h)\in U_h\times P_h\times W_h} \mathcal{J}_h^I(\mathbf{v}_h,q_h,z_h;\mathbf{f}). \tag{2.8}$$

Taking variations in (2.7) with respect to $(\mathbf{v}_h, q_h, z_h)$, we obtain the weak statement of problem (2.8): Find $(\mathbf{u}_h, p_h, \omega_h) \in U_h \times P_h \times W_h$ such that

(2.9)

$$\left\{\begin{aligned}&\mathcal{L}_h^I((\mathbf{u}_h,p_h,\omega_h);(\mathbf{v}_h,q_h,z_h)) := (\breve{R}_h(\bigtriangledown\times\omega_h + \bigtriangledown p_h), \breve{R}_h(\bigtriangledown\times\ z_h + \bigtriangledown q_h))\\ &\quad + \sum_{K\in\mathcal{C}_h} h_K^2\,(\bigtriangledown\times\omega_h + \bigtriangledown p_h, \bigtriangledown\times z_h + \bigtriangledown q_h)_{0,K}\\ &\quad + (\omega_h - \bigtriangledown\times\mathbf{u}_h, z_h - \bigtriangledown\times\mathbf{v}_h) + (\bigtriangledown\cdot\mathbf{u}_h, \bigtriangledown\cdot\mathbf{v}_h)\\ &= (\mathbf{f}, \breve{R}_h(\bigtriangledown\times z_h + \bigtriangledown q_h)) + \sum_{K\in\mathcal{C}_h} h_K^2\,(\mathbf{f}, \bigtriangledown\times z_h + \bigtriangledown q_h)_{0,K}\end{aligned}\right.$$

holds for all $(\mathbf{v}_h, q_h, z_h) \in U_h \times P_h \times W_h$.

2.3. Mass-lumping L^2 projection method (II). Introduce

$$V_{0,h} := Z_h \cap (H_0^1(\Omega))^d, \tag{2.10}$$

where Z_h is defined in (2.4). Let $(\cdot,\cdot)_h$ denote an inner product in $V_{0,h}$ and the induced norm in $V_{0,h}$ be given by

$$\|\mathbf{v}\|_h := (\mathbf{v},\mathbf{v})_h^{1/2}. \tag{2.11}$$

Remark 2.1. $(\cdot,\cdot)_h$ is usually taken as an approximation of $(\cdot,\cdot)$. For example, when $\mathcal{C}_h$ consists of two-dimensional (2D) triangles, we may take $(\cdot,\cdot)_h$ as the quadrature scheme

$$(u,v)_h := \sum_{K\in\mathcal{C}_h} \frac{\text{area}(K)}{3} \sum_{i=1}^{3} u(i)\, v(i), \tag{2.12}$$

where $i = 1, 2, 3$ denote vertices of the triangle K. In the literature [1], $(\cdot,\cdot)_h$ replacing $(\cdot,\cdot)$ is called *mass-lumping*. The matrix associated with $(\cdot,\cdot)_h$ is diagonal.

For given $w \in (L^2(\Omega))^{2d-3}$ and $p \in L^2(\Omega)$ we define two functions $R_h(\bigtriangledown \times w) \in V_{0,h}$ and $S_h(\bigtriangledown p) \in V_{0,h}$, respectively, by

$$(R_h(\bigtriangledown \times w), \mathbf{v}_h)_h = (w, \bigtriangledown \times \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_{0,h}, \tag{2.13}$$

$$(S_h(\bigtriangledown p), \mathbf{v}_h)_h = -(p, \bigtriangledown \cdot \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_{0,h}. \tag{2.14}$$

For given $\mathbf{g} \in (L^2(\Omega))^d$ we define a function $\bar{R}_h(\mathbf{g}) \in V_{0,h}$ by

$$(\bar{R}_h(\mathbf{g}), \mathbf{v}_h)_h = (\mathbf{g}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_{0,h}. \tag{2.15}$$

Remark 2.2. Clearly, R_h, S_h, and $\bar{R}_h$ are all linear operators. In addition, if $w \in (H^1(\Omega))^{2d-3}$ and $p \in H^1(\Omega)$, we have

$$R_h(\bigtriangledown \times w) + S_h(\bigtriangledown p) = \bar{R}_h(\bigtriangledown \times \omega) + \bar{R}_h(\bigtriangledown p) = \bar{R}_h(\bigtriangledown \times w + \bigtriangledown p). \tag{2.16}$$

We consider the case of P_h and W_h possibly being discontinuous or being linear and quadratic continuous elements, and we define an LS functional on $U_h \times P_h \times W_h$:

$$\begin{aligned} \mathcal{J}_h^{II}(\mathbf{u}, p, \omega; \mathbf{f}) := {} & \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p) - \bar{R}_h(\mathbf{f})\|_h^2 \\ & + \sum_{K\in\mathcal{C}_h} h_K^2 \, \|\bigtriangledown \times \omega + \bigtriangledown p - \mathbf{f}\|_{0,K}^2 \\ & + \sum_{E\in\mathcal{E}_h} h_E \int_E |[p]|^2 + \sum_{E\in\mathcal{E}_h} h_E \int_E |[w]|^2 + \|\omega - \bigtriangledown \times \mathbf{u}\|_0^2 + \|\bigtriangledown \cdot \mathbf{u}\|_0^2, \end{aligned} \tag{2.17}$$

where $\mathcal{E}_h$ denotes the collection of interior edges (faces in $\mathbb{R}^3$), $[p]$ is the jump in p across E, and h_E is the length or diameter of E. With $\mathcal{J}_h^{II}$, in the same way as that for (2.8) and (2.9), we can consider an LS minimization problem and then obtain its weak statement: Find $(\mathbf{u}_h, p_h, \omega_h) \in U_h \times P_h \times W_h$ such that

$$\begin{cases} \mathcal{L}_h^{II}((\mathbf{u}_h, p_h, \omega_h); (\mathbf{v}_h, q_h, z_h)) := (R_h(\bigtriangledown \times \omega_h) + S_h(\bigtriangledown p_h), R_h(\bigtriangledown \times z_h) + S_h(\bigtriangledown q_h))_h \\ \quad + \sum_{K\in\mathcal{C}_h} h_K^2 \, (\bigtriangledown \times \omega_h + \bigtriangledown p_h, \bigtriangledown \times z_h + \bigtriangledown q_h)_{0,K} \\ \quad + \sum_{E\in\mathcal{E}_h} h_E \int_E [\omega_h]\,[z_h] + \sum_{E\in\mathcal{E}_h} h_E \int_E [p_h]\,[q_h] \\ \quad + (\omega_h - \bigtriangledown \times \mathbf{u}_h, z_h - \bigtriangledown \times \mathbf{v}_h) + (\bigtriangledown \cdot \mathbf{u}_h, \bigtriangledown \cdot \mathbf{v}_h) \\ \quad = (\mathbf{f}, R_h(\bigtriangledown \times z_h) + S_h(\bigtriangledown q_h)) + \sum_{K\in\mathcal{C}_h} h_K^2 \, (\mathbf{f}, \bigtriangledown \times z_h + \bigtriangledown q_h)_{0,K} \end{cases} \tag{2.18}$$

holds for all $(\mathbf{v}_h, q_h, z_h) \in U_h \times P_h \times W_h$.

Remark 2.3. Method (I) is simpler than method (II), but the latter applies to lower-order continuous elements and discontinuous elements for pressure and vorticity.

Note that using linear or quadratic elements for pressure and vorticity without the $\breve{R}_h$ term and the h^2 factor, method (I) is the standard L^2 LS method [15] which does not have optimal error estimates. In addition, we remark that without the L^2 projection, our method reduces to the BG method [18], and when replacing the L^2 projector by a preconditioner (or an operator of the finite element solution) for the Dirichlet problem of a second-order elliptic equation, the H^{-1} method [14, 32] is obtained.

3. Coercivity and error bounds in energy norm. We shall give a unified analysis for coercivity and error bounds in energy norm for methods (I) and (II). In what follows, C represents a generic positive constant independent of h and may take different values at different occurrences.

3.1. Coercivity analysis. In this subsection we investigate the coercivity.

PROPOSITION 3.1 (see [2, 11]). *Let Ω be a bounded connected domain with a Lipschitz-continuous boundary Γ. Then*

$$\|\mathbf{v}\|_1^2 \le C\,\{\|\bigtriangledown \times \mathbf{v}\|_0^2 + \|\bigtriangledown \cdot \mathbf{v}\|_0^2\} \quad \forall \mathbf{v} \in (H_0^1(\Omega))^d. \tag{3.1}$$

PROPOSITION 3.2 (see [25]). *Under the assumption on Ω as in Proposition* 3.1*, we have*

$$\inf_{q\in L_0^2(\Omega)} \sup_{\mathbf{0}\neq\mathbf{v}\in(H_0^1(\Omega))^d} \frac{(\bigtriangledown \cdot \mathbf{v}, q)}{\|\mathbf{v}\|_1 \, \|q\|_0} \ge C. \tag{3.2}$$

LEMMA 3.1. *Let X be a given Hilbert space, with inner product $(\cdot,\cdot)_X$ and corresponding norm $\|\cdot\|_X = \sqrt{(\cdot,\cdot)_X}$. For any two elements $A \in X, B \in X$ and for any $\alpha \in \mathbb{R}$, we have*

$$\|A-B\|_X^2 \ge \alpha(1-\alpha/2)\,(\|A\|_X^2 + \|B\|_X^2) - 2\alpha(A,B)_X. \tag{3.3}$$

Proof. Equation (3.3) follows from the sum of the two equations

$$\|A-B\|_X^2 = \|A-\alpha A-B\|_X^2 + \alpha(2-\alpha)\,\|A\|_X^2 - 2\alpha\,(A,B)_X,$$
$$\|A-B\|_X^2 = \|A-B+\alpha B\|_X^2 + \alpha(2-\alpha)\,\|B\|_X^2 - 2\alpha\,(A,B)_X. \qquad \square$$

For the following analysis we recall the well-known Young's inequality

$$|a|\,|b| \le \epsilon\,|a|^2 + \frac{1}{4\,\epsilon}\,|b|^2 \quad \forall a,b \in \mathbb{R}, \quad \forall \epsilon > 0$$

and Green's formulae of integrating by parts

$$(\bigtriangledown \times \mathbf{v}, \phi)_{0,D} - (\mathbf{v}, \bigtriangledown \times \phi)_{0,D} = \int_{\partial\, D} \phi\, \mathbf{v} \times \mathbf{n} \quad \forall \mathbf{v} \in (H^1(D))^d, \phi \in (H^1(D))^{2\,d-3},$$
$$(\bigtriangledown \cdot \mathbf{v}, q)_{0,D} + (\mathbf{v}, \bigtriangledown q)_{0,D} = \int_{\partial D} q\, \mathbf{v} \cdot \mathbf{n} \quad \forall \mathbf{v} \in (H^1(D))^d, q \in H^1(D),$$

where $\mathbf{n}$ denotes the exterior unit normal to $\partial\, D$, and D is a Lipschitz subdomain of Ω. We also introduce a notation

$$|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 := \sum_{K\in\mathcal{C}_h} h_K^2 \|\bigtriangledown \times \omega + \bigtriangledown p\,\|_{0,K}^2 + \sum_{E\in\mathcal{E}_h} h_E \left(\int_E |[p]|^2 + |[\omega]|^2\right). \tag{3.4}$$

Assumption (A1). There exists $C_1 > 0, C_2 > 0$ independent of h such that

$$C_1 \|\mathbf{v}_h\|_0 \le \|\mathbf{v}_h\|_h \le C_2 \|\mathbf{v}_h\|_0 \quad \forall \mathbf{v}_h \in V_{0,h}. \tag{3.5}$$

Remark 3.1. Taking (2.12) as an example. Assumption (A1) can be easily shown by considering each triangle separately (see also [5, p. 157]).

THEOREM 3.1. *Assuming* (A1) *for method* (II), *let* $\mathcal{J}_h$ *stand for* $\mathcal{J}_h^I$ *or* $\mathcal{J}_h^{II}$. *Then, for all* $(\mathbf{u}, p, \omega) \in U_h \times P_h \times W_h$,

$$\mathcal{J}_h(\mathbf{u}, p, \omega; \mathbf{0}) \ge C \{\|\mathbf{u}\|_1^2 + \|p\|_0^2 + \|\omega\|_0^2\}. \tag{3.6}$$

Proof. We consider method (II) here. The argument remains unchanged for method (I). One needs only to note that, in method (I), $P_h \times W_h$ are continuous and $(\cdot,\cdot)$ is in place of $(\cdot,\cdot)_h$.

In the proof we need only to deal with $\|\omega - \bigtriangledown \times \mathbf{u}\|_0^2$ and $\|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2$. The proof is divided into three steps. In the first two steps we find lower bounds for $\|\omega - \bigtriangledown \times \mathbf{u}\|_0^2 + \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2$. In the last step we use the mesh-dependent terms to obtain (3.6).

Step 1. Let $\alpha > 0$ be a constant to be determined. From Lemma 3.1 we have

$$\|\omega - \bigtriangledown \times \mathbf{u}\|_0^2 \ge \alpha\,(1 - \alpha/2)\,\{\|\bigtriangledown \times \mathbf{u}\|_0^2 + \|\omega\|_0^2\} - 2\,\alpha\,(\omega, \bigtriangledown \times\ \mathbf{u}). \tag{3.7}$$

We take $\tilde{\mathbf{u}} \in V_{0,h}$ as the Clément-interpolant [2, 4] of $\mathbf{u} \in U_h$ and have

$$\left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|\tilde{\mathbf{u}} - \mathbf{u}\|_{0,K}^2 + \sum_{E \in \mathcal{E}_h} h_E^{-1} \int_E |\tilde{\mathbf{u}} - \mathbf{u}|^2\right)^{1/2} + \|\tilde{\mathbf{u}}\|_1 \le C \|\mathbf{u}\|_1. \tag{3.8}$$

We also have

$$\begin{aligned} -2\,\alpha\,(\omega, \bigtriangledown \times \mathbf{u}) &= -2\,\alpha\,(\omega, \bigtriangledown \times \tilde{\mathbf{u}}) - 2\,\alpha\,(\omega, \bigtriangledown \times (\mathbf{u} - \tilde{\mathbf{u}})) \\ &= -2\,\alpha\,(R_h(\bigtriangledown \times \omega), \tilde{\mathbf{u}})_h - 2\,\alpha\,(\omega, \bigtriangledown \times (\mathbf{u} - \tilde{\mathbf{u}})), \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} &- 2\,\alpha\,(R_h(\bigtriangledown \times \omega), \tilde{\mathbf{u}})_h + \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2 \\ &\quad = \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p) - \alpha\,\tilde{\mathbf{u}}\|_h^2 - \alpha^2 \|\tilde{\mathbf{u}}\|_h^2 + 2\,\alpha\,(S_h(\bigtriangledown p), \tilde{\mathbf{u}})_h, \end{aligned} \tag{3.10}$$

where

$$\begin{aligned} -\alpha^2 \|\tilde{\mathbf{u}}\|_h^2 &\ge -\alpha^2\,C\,\|\tilde{\mathbf{u}}\|_0^2 \qquad \text{(by Assumption (A1))} \\ &\ge -\alpha^2\,C\,\|\mathbf{u}\|_1^2 \qquad \text{(by (3.8))} \\ &\ge -\alpha^2\,C\,\|\bigtriangledown \times \mathbf{u}\|_0^2 - \alpha^2\,C\,\|\bigtriangledown \cdot \mathbf{u}\|_0^2 \qquad \text{(by Proposition 3.1)}, \end{aligned} \tag{3.11}$$

$$\begin{aligned} 2\,\alpha\,(S_h(\bigtriangledown p), \tilde{\mathbf{u}})_h &= -2\,\alpha\,(p, \bigtriangledown \cdot \tilde{\mathbf{u}}) \\ &= -2\,\alpha\,(p, \bigtriangledown \cdot \mathbf{u}) + 2\,\alpha\,(p, \bigtriangledown \cdot (\mathbf{u} - \tilde{\mathbf{u}})), \end{aligned} \tag{3.12}$$

and

$$-2\alpha(p, \bigtriangledown \cdot \mathbf{u}) \ge -\epsilon_1 \|p\|_0^2 - \frac{\alpha^2\,C}{\epsilon_1} \|\bigtriangledown \cdot \mathbf{u}\|_0^2 \tag{3.13}$$

(by Young's inequality). Here $\epsilon_1 > 0$ is a constant to be determined later.

Therefore, summarizing (3.7) and (3.9)–(3.13), we get

$$\begin{cases} \|\omega - \bigtriangledown \times \mathbf{u}\|_0^2 + \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2 \\ \geq \alpha(1-\alpha/2)\,\|\omega\|_0^2 + \alpha\,[1-\alpha\,(1/2+C)]\;\|\bigtriangledown \times \mathbf{u}\|_0^2 \\ \quad +2\,\alpha\,\{-(\omega, \bigtriangledown \times (\mathbf{u}-\tilde{\mathbf{u}})) + (p, \bigtriangledown \cdot (\mathbf{u}-\tilde{\mathbf{u}}))\} \\ \quad -\epsilon_1\,\|p\|_0^2 - (\frac{\alpha^2\,C}{\epsilon_1} + \alpha^2\,C)\,\|\bigtriangledown \cdot \mathbf{u}\|_0^2, \end{cases} \tag{3.14}$$

where

$$\begin{cases} 2\,\alpha\,\{-(\omega, \bigtriangledown \times (\mathbf{u}-\tilde{\mathbf{u}})) + (p, \bigtriangledown \cdot (\mathbf{u}-\tilde{\mathbf{u}}))\} \\ \quad = -2\,\alpha \sum\limits_{K\in\mathcal{C}_h} (\bigtriangledown \times \omega + \bigtriangledown p, \mathbf{u}-\tilde{\mathbf{u}})_{0,K} \\ \qquad -2\,\alpha \sum\limits_{E\in\mathcal{E}_h} \int_E [\omega]\,(\mathbf{u}-\tilde{\mathbf{u}}) \times \mathbf{n}_E + 2\,\alpha \sum\limits_{E\in\mathcal{E}_h} \int_E [p]\,(\mathbf{u}-\tilde{\mathbf{u}}) \cdot \mathbf{n}_E, \\ \geq -2\,\alpha\,C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)} \\ \times \Big(\sum\limits_{K\in\mathcal{C}_h} h_K^{-2}\,\|\mathbf{u}-\tilde{\mathbf{u}}\|_{0,K}^2 + \sum\limits_{E\in\mathcal{E}_h} h_E^{-1}\,\int_E\,|\mathbf{u}-\tilde{\mathbf{u}}|^2 \Big)^{1/2} \\ \geq -2\,\alpha\,C\,\|\mathbf{u}\|_1\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)} \geq -\alpha^2\,\|\mathbf{u}\|_1^2 - C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 \\ \geq -\alpha^2\,C\,\|\bigtriangledown \times \mathbf{u}\|_0^2 - \alpha^2\,C\,\|\bigtriangledown \cdot \mathbf{u}\|_0^2 - C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2. \end{cases} \tag{3.15}$$

Thus, (3.14) becomes

$$\begin{cases} \|\omega - \bigtriangledown \times \mathbf{u}\|_0^2 + \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2 \\ \geq \alpha(1-\alpha/2)\,\|\omega\|_0^2 + \alpha\,[1-\alpha\,(1/2+2\,C)]\;\|\bigtriangledown \times \mathbf{u}\|_0^2 \\ \quad -\epsilon_1\,\|p\|_0^2 - C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 - \Big(\frac{\alpha^2\,C}{\epsilon_1} + \alpha^2\,2\,C\Big)\;\|\bigtriangledown \cdot \mathbf{u}\|_0^2. \end{cases} \tag{3.16}$$

Step 2. Let $\beta > 0$ be a constant to be determined. From Proposition 3.2 we can find $\mathbf{v}^* \in (H_0^1(\Omega))^d$ such that

$$(\bigtriangledown \cdot \mathbf{v}^*, p) = \|p\|_0^2, \quad \|\mathbf{v}^*\|_1 \leq C\,\|p\|_0. \tag{3.17}$$

We take $\tilde{\mathbf{v}}^* \in V_{0,h}$ as the Clément-interpolant [2, 4] of $\mathbf{v}^*$ and have

$$\left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,\|\tilde{\mathbf{v}}^* - \mathbf{v}^*\|_{0,K}^2 + \sum_{E\in\mathcal{E}_h} h_E^{-1} \int_E\,|\tilde{\mathbf{v}}^* - \mathbf{v}^*|^2 \right)^{1/2} + \|\tilde{\mathbf{v}}^*\|_1 \leq C\,\|\mathbf{v}^*\|_1. \tag{3.18}$$

We can write

$$\begin{aligned} \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p)\|_h^2 &= \|R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p) + \beta\,\tilde{\mathbf{v}}^*\|_h^2 \\ &\quad -\beta^2\,\|\tilde{\mathbf{v}}^*\|_h^2 - 2\,\beta\,\{(R_h(\bigtriangledown \times \omega), \tilde{\mathbf{v}}^*)_h + (S_h(\bigtriangledown p), \tilde{\mathbf{v}}^*)_h\}, \end{aligned} \tag{3.19}$$

where

$$-\beta^2\,\|\tilde{\mathbf{v}}^*\|_h^2 \geq -\beta^2\,C\,\|\tilde{\mathbf{v}}^*\|_0^2 \geq -\beta^2\,C\,\|\mathbf{v}^*\|_1^2 \geq -\beta^2\,C\,\|p\|_0^2, \tag{3.20}$$

$$\begin{cases} -2\,\beta\,\{(R_h(\bigtriangledown \times \omega), \tilde{\mathbf{v}}^*)_h + (S_h(\bigtriangledown p), \tilde{\mathbf{v}}^*)_h\} \\ = -2\,\beta\,\{(\omega, \bigtriangledown \times \tilde{\mathbf{v}}^*) - (p, \bigtriangledown \cdot \tilde{\mathbf{v}}^*)\} \\ = 2\,\beta\,(p, \bigtriangledown \cdot \mathbf{v}^*) - 2\,\beta\,(\omega, \bigtriangledown \times \mathbf{v}^*) \\ \quad +2\,\beta\,\{(\omega, \bigtriangledown \times (\mathbf{v}^* - \tilde{\mathbf{v}}^*)) - (p, \bigtriangledown \cdot (\mathbf{v}^* - \tilde{\mathbf{v}}^*))\} \\ = 2\,\beta\,\|p\|_0^2 - 2\,\beta\,(\omega, \bigtriangledown \times \mathbf{v}^*) + 2\,\beta \sum\limits_{K\in\mathcal{C}_h} (\bigtriangledown \times \omega + \bigtriangledown p, \mathbf{v}^* - \tilde{\mathbf{v}}^*)_{0,K} \\ \quad +2\,\beta \sum\limits_{E\in\mathcal{E}_h} \int_E [\omega]\,(\mathbf{v}^* - \tilde{\mathbf{v}}^*) \times \mathbf{n}_E - 2\,\beta \sum\limits_{E\in\mathcal{E}_h} \int_E [p]\,(\mathbf{v}^* - \tilde{\mathbf{v}}^*) \cdot \mathbf{n}_E \end{cases} \tag{3.21}$$

with

(3.22)
$$\begin{cases} 2\,\beta \sum\limits_{K\in\mathcal{C}_h} (\bigtriangledown\times\omega+\bigtriangledown p, \mathbf{v}^*-\tilde{\mathbf{v}}^*)_{0,K} \\ \quad +2\,\beta \sum\limits_{E\in\mathcal{E}_h} \int_E [\omega]\,(\mathbf{v}^*-\tilde{\mathbf{v}}^*)\times\mathbf{n}_E - 2\,\beta \sum\limits_{E\in\mathcal{E}_h}\int_E [p]\,(\mathbf{v}^*-\tilde{\mathbf{v}}^*)\cdot\mathbf{n}_E \\ \geq -2\,\beta\,C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)} \left(\sum\limits_{K\in\mathcal{C}_h} h_K^{-2}\,\|\tilde{\mathbf{v}}^*-\mathbf{v}^*\|_{0,K}^2 + \sum\limits_{E\in\mathcal{E}_h} h_E^{-1}\int_E |\tilde{\mathbf{v}}^*-\mathbf{v}^*|^2\right)^{1/2} \\ \geq -2\,\beta\,C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}\,\|\mathbf{v}^*\|_1 \\ \geq -2\,\beta\,C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}\,\|p\|_0 \geq -\beta^2\,\|p\|_0^2 - C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2, \end{cases}$$

and

$$-2\,\beta\,(\omega,\bigtriangledown\times\mathbf{v}^*) \geq -2\,\beta\,C\,\|\omega\|_0\,\|\mathbf{v}^*\|_1 \geq -\epsilon_2\,\|\omega\|_0^2 - \frac{C\,\beta^2}{\epsilon_2}\,\|p\|_0^2. \tag{3.23}$$

Here $\epsilon_2>0$ is also a constant to be determined later.

Summarizing (3.19)–(3.23), we get

$$\begin{aligned} \|R_h(\bigtriangledown\times\omega)+S_h(\bigtriangledown p)\|_h^2 \geq \beta\,&[2-\beta\,(C+1+C/\epsilon_2)]\,\|p\|_0^2 \\ &-\epsilon_2\,\|\omega\|_0^2 - C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2. \end{aligned} \tag{3.24}$$

Step 3. From (3.16) and (3.24), taking

$$0<\alpha<\frac{1}{1/2+2\,C},\quad 0<\epsilon_2<\alpha\,(1-\alpha/2), \tag{3.25}$$

$$0<\beta<\frac{2}{C+1+C/\epsilon_2},\quad 0<\epsilon_1<\beta\,[2-\beta\,(C+1+C/\epsilon_2)]\,, \tag{3.26}$$

we have

$$\begin{cases} \|\omega-\bigtriangledown\times\mathbf{u}\|_0^2 + 2\,\|R_h(\bigtriangledown\times\omega)+S_h(\bigtriangledown p)\|_h^2 \\ \geq [\alpha(1-\alpha/2)-\epsilon_2]\,\|\omega\|_0^2 + \alpha\,[1-\alpha\,(1/2+2\,C)]\,\|\bigtriangledown\times\mathbf{u}\|_0^2 \\ \quad +\{\beta\,[2-\beta\,(C+1+C/\epsilon_2)]-\epsilon_1\}\,\|p\|_0^2 \\ \quad -2\,C\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 - (\frac{\alpha^2\,C}{\epsilon_1}+\alpha^2\,2\,C)\,\|\bigtriangledown\cdot\mathbf{u}\,\|_0^2. \end{cases} \tag{3.27}$$

Now we take positive constants L_i, $i=1,2$, such that

$$L_1>2\,C,\quad L_2>\frac{\alpha^2\,C}{\epsilon_1}+\alpha^2\,2\,C, \tag{3.28}$$

and we have

$$\begin{cases} \max(2,L_1,L_2)\,\mathcal{J}_h^{II}(\mathbf{u},p,\omega;\mathbf{0}) \geq \|\omega-\bigtriangledown\times\mathbf{u}\|_0^2 \\ \quad +2\,\|R_h(\bigtriangledown\times\omega)+S_h(\bigtriangledown p)\|_h^2 + L_1\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 + L_2\,\|\bigtriangledown\cdot\mathbf{u}\|_0^2 \\ \geq [\alpha(1-\alpha/2)-\epsilon_2]\,\|\omega\|_0^2 + \alpha\,[1-\alpha\,(1/2+2\,C)]\,\|\bigtriangledown\times\mathbf{u}\|_0^2 \\ \quad +\{\beta\,[2-\beta\,(C+1+C/\epsilon_2)]-\epsilon_1\}\,\|p\|_0^2 \\ \quad +(L_1-2\,C)\,|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2 + [L_2-(\frac{\alpha^2\,C}{\epsilon_1}+\alpha^2\,2\,C)]\,\|\bigtriangledown\cdot\mathbf{u}\|_0^2 \\ \geq C\{\|\omega\|_0^2+\|p\|_0^2+\|\bigtriangledown\times\mathbf{u}\|_0^2+\|\bigtriangledown\cdot\mathbf{u}\|_0^2+|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2\} \\ \geq C\{\|\mathbf{u}\|_1^2+\|p\|_0^2+\|\omega\|_0^2+|(p,\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}^2\}. \end{cases} \tag{3.29}$$

Hence, we obtain (3.6) for $\mathcal{J}_h^{II}$. ∎

Using Theorem 3.1 and the Lax–Milgram lemma [3, 5, 29] we can easily obtain the following.

COROLLARY 3.1. *Under the same assumptions as in Theorem* 3.1*, the finite element problems* (2.9) *and* (2.18) *have a unique solution.*

COROLLARY 3.2. *Under the same assumptions as in Theorem* 3.1 *and assuming a quasi-uniform* $\mathcal{C}_h$*, we can conclude that the condition number of the linear system from the finite element problem is* $\mathcal{O}(h^{-2})$.

Proof. We take method (I) as an example. Since $\breve{R}_h$ is an L^2 projector, we have

$$\|\breve{R}_h(\mathbf{g})\|_0 \le C\,\|\mathbf{g}\|_0;$$

then by the inverse estimation [5, 29] we have, for all $(\mathbf{v}_h, q_h, z_h) \in U_h \times P_h \times W_h$,

$$\|\breve{R}_h(\bigtriangledown \times z_h + \bigtriangledown q_h)\|_0 \le C\,\|\bigtriangledown \times z_h + \bigtriangledown q_h\|_0 \le C\,h^{-1}\,\{\|q_h\|_0 + \|z_h\|_0\},$$

and thus

$$\mathcal{L}_h^I((\mathbf{v}_h, q_h, z_h); (\mathbf{v}_h, q_h, z_h)) \le C\,h^{-2}\,\{\|\mathbf{v}_h\|_0^2 + \|q_h\|_0^2 + \|z_h\|_0^2\},$$

which, together with (3.6) and the symmetry of $\mathcal{L}_h^I$, completes the proof. ∎

3.2. Error bounds in energy norm. Since $\mathcal{L}_h$ is symmetric and positive definite (see Theorem 3.1), we introduce a norm $|||\cdot|||$ on $U_h \times P_h \times W_h$ by

$$|||(\mathbf{v}_h, q_h, z_h)|||^2 := \mathcal{L}_h((\mathbf{v}_h, q_h, z_h); (\mathbf{v}_h, q_h, z_h)), \tag{3.30}$$

where $\mathcal{L}_h$ stands for $\mathcal{L}_h^I$ or $\mathcal{L}_h^{II}$. $|||\cdot|||$ will be referred to as *energy norm*. We can also easily show the generalized Cauchy–Schwarz inequality

$$\mathcal{L}_h((\mathbf{u}, p, \omega); (\mathbf{v}, q, z)) \le |||(\mathbf{u}, p, \omega)|||\,|||(\mathbf{v}, q, z)||| \tag{3.31}$$

for all $(\mathbf{u}, p, \omega), (\mathbf{v}, q, z) \in ((H_0^1(\Omega))^d + U_h) \times (H^1(\Omega) + P_h) \times ((H^1(\Omega))^{2\,d-3} + W_h)$. Here $x \in \mathbb{Y} + \mathbb{T}$ means that $x \in \mathbb{Y}$ or $x \in \mathbb{T}$ or $x = y + t$ with $y \in \mathbb{Y}$ and $t \in \mathbb{T}$.

LEMMA 3.2. *Let* $(\mathbf{u}, p, \omega) \in (H^1(\Omega))^d \times H^1(\Omega) \times (H^1(\Omega))^{2d-3}$ *and* $(\mathbf{u}_h, p_h, \omega_h) \in U_h \times P_h \times W_h$ *be the exact and approximate solutions, respectively. Then, for all* $(\mathbf{v}_h, q_h, z_h) \in U_h \times P_h \times W_h$*,*

$$\mathcal{L}_h((\mathbf{u} - \mathbf{u}_h, p - p_h, \omega - \omega_h); (\mathbf{v}_h, q_h, z_h)) = 0. \tag{3.32}$$

Proof. Equation (3.32) can be easily shown, due to (2.2), (2.16), and

$$\begin{aligned}(\mathbf{f}, \breve{R}_h(\bigtriangledown \times z_h + \bigtriangledown q_h)) &= (\breve{R}_h(\mathbf{f}), \breve{R}_h(\bigtriangledown \times z_h + \bigtriangledown q_h)) \\ &= (\breve{R}_h(\bigtriangledown \times \omega + \bigtriangledown p), \breve{R}_h(\bigtriangledown \times z_h + \bigtriangledown q_h)),\end{aligned} \tag{3.33}$$

$$\begin{aligned}&(\mathbf{f}, R_h(\bigtriangledown \times\ z_h) + S_h(\bigtriangledown q_h)) \\ &\quad = (\bar{R}_h(\mathbf{f}), R_h(\bigtriangledown \times\ z_h) + S_h(\bigtriangledown q_h))_h \\ &\quad = (\bar{R}_h(\bigtriangledown \times \omega + \bigtriangledown p), R_h(\bigtriangledown \times\ z_h) + S_h(\bigtriangledown q_h))_h \\ &\quad = (R_h(\bigtriangledown \times \omega) + S_h(\bigtriangledown p), R_h(\bigtriangledown \times\ z_h) + S_h(\bigtriangledown q_h))_h.\end{aligned} \tag{3.34}$$

∎

Assumption (A2). For $\omega \in (H^{l+1}(\Omega))^{2d-3}$ and $p \in H^{m+1}(\Omega) \cap L_0^2(\Omega)$, with $l, m \ge 0$, there exist $\tilde{\omega} \in W_h$ and $\tilde{p} \in P_h$ such that

$$R_h(\bigtriangledown \times (\omega - \tilde{\omega})) + S_h(\bigtriangledown(p - \tilde{p})) = \mathbf{0} \quad \text{(for method (II))} \tag{3.35}$$

or

$$\breve{R}_h(\bigtriangledown \times (\omega - \tilde{\omega}) + \bigtriangledown(p - \tilde{p})) = \mathbf{0} \quad \text{(for method (I))} \tag{3.36}$$

with

$$\|\tilde{\omega} - \omega\|_0 + |(0, \tilde{\omega} - \omega)|_{(\mathcal{C}_h, \mathcal{E}_h)} \leq C\, h^s\, \|\omega\|_s, \quad 1 \leq s \leq l+1 \tag{3.37}$$

and

$$\|\tilde{p} - p\|_0 + |(\tilde{p} - p, 0)|_{(\mathcal{C}_h, \mathcal{E}_h)} \leq C\, h^s\, \|p\|_s, \quad 1 \leq s \leq m+1. \tag{3.38}$$

Remark 3.2. To make Assumption (A2) hold, the approximation spaces for pressure and vorticity are required to have either interior or edge (face) degrees of freedom or both with respect to each element. All standard approximation spaces satisfy this requirement, with lower-order approximation spaces enriched by suitable artificial element bubbles or edge (face) bubbles or both. Here we briefly consider method (II) for 2D problems. Assumption (A2) holds for method (II), using continuous $\mathcal{P}_r$ elements with $r \geq 3$ or discontinuous $\mathcal{P}_r$ elements with $r \geq 0$ or $\mathcal{P}_1^+ (= \mathcal{P}_1 + b_K)$ or $\mathcal{P}_2^+$, where $b_K \in H_0^1(K)$ is a bubble $\lambda_1\lambda_2\lambda_3$ and λ_i is the ith bary-coordinate on a triangle. See Theorem 3.4 for details of the verification of Assumption (A2).

THEOREM 3.2. *Assuming the assumptions of Theorem* 3.1 *and Assumption* (A2)*, let* $(\mathbf{u}, p, \omega) \in (H^1(\Omega))^d \times H^1(\Omega) \times (H^1(\Omega))^{2\,d-3}$ *and* $(\mathbf{u}_h, p_h, \omega_h) \in U_h \times P_h \times W_h$ *be the exact and approximate solutions. We have*

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h, \omega - \omega_h)||| \leq C\, \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega), \tag{3.39}$$

where $\tilde{\mathbf{u}} \in U_h$ *is any given function,* $\tilde{p}$ *and* $\tilde{\omega}$ *come from Assumption* (A2)*, and*

$$\begin{aligned} \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega) &:= \|\tilde{\mathbf{u}} - \mathbf{u}\|_1 + \|\tilde{p} - p\|_0 \\ &\quad + \|\tilde{\omega} - \omega\|_0 + |(\tilde{p} - p, \tilde{\omega} - \omega)|_{(\mathcal{C}_h, \mathcal{E}_h)}. \end{aligned} \tag{3.40}$$

We further have

$$\|p - p_h\|_0 + \|\omega - \omega_h\|_0 + \|\mathbf{u} - \mathbf{u}_h\|_1 \leq C\, \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega), \tag{3.41}$$

$$|(p - p_h, \omega - \omega_h)|_{(\mathcal{C}_h, \mathcal{E}_h)} \leq C\, \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega), \tag{3.42}$$

$$\|R_h(\bigtriangledown \times (\omega - \omega_h)) + S_h(\bigtriangledown(p - p_h))\|_h \leq C\, \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega) \tag{3.43}$$

or

$$\|\breve{R}_h(\bigtriangledown \times (\omega - \omega_h) + \bigtriangledown(p - p_h))\|_0 \leq C\, \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega). \tag{3.44}$$

Proof. It suffices to show (3.39) with (3.40). From (3.30), Lemma 3.2, and (3.31), we have

$$|||(\tilde{\mathbf{u}} - \mathbf{u}_h, \tilde{p} - p_h, \tilde{\omega} - \omega_h)||| \leq |||(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega)|||. \tag{3.45}$$

Using the triangle inequality, we then get

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h, \omega - \omega_h)||| \leq C\, |||(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega)|||. \tag{3.46}$$

By Assumption (A2) we have

$$\begin{aligned}|||(\tilde{\mathbf{u}}-\mathbf{u},\tilde{p}-p,\tilde{\omega}-\omega)|||^2 &= \mathcal{L}_h((\tilde{\mathbf{u}}-\mathbf{u},\tilde{p}-p,\tilde{\omega}-\omega);(\tilde{\mathbf{u}}-\mathbf{u},\tilde{p}-p,\tilde{\omega}-\omega))\\ &= |(\tilde{p}-p,\tilde{\omega}-\omega)|^2_{(\mathcal{C}_h,\mathcal{E}_h)} + \|\tilde{\omega}-\omega-\bigtriangledown\times(\tilde{\mathbf{u}}-\mathbf{u})\|_0^2\\ &\quad + \|\bigtriangledown\cdot(\tilde{\mathbf{u}}-\mathbf{u})\|_0^2,\end{aligned} \tag{3.47}$$

which completes the proof. ∎

As a concrete application, we consider the 2D case and state the error estimates for methods (I) and (II) in the following.

THEOREM 3.3. *Assuming the same hypotheses as in Theorem* 3.2*, when employing equal-order continuous interpolation* $\mathcal{P}_r-\mathcal{P}_r-\mathcal{P}_r$ *with* $r\geq 3$ *for methods* (I) *and* (II)*, or for method* (II) *employing* $\mathcal{P}_r-\mathcal{P}_r^+-\mathcal{P}_r^+$ *with* $r=1,2$ *or employing* $\mathcal{P}_r$*(continuous)-*$\mathcal{P}_{r-1}$*(discontinuous)-*$\mathcal{P}_{r-1}$*(discontinuous) with* $r\geq 1$ *for* $(\mathbf{u},p,\omega)\in(H^{r+1}(\Omega))^2\times H^r(\Omega)\times H^r(\Omega)$*, we have*

$$\|\mathbf{u}-\mathbf{u}_h\|_1+\|p-p_h\|_0+\|\omega-\omega_h\|_0\leq C\,h^r\ (\|\mathbf{u}\|_{r+1}+\|p\|_r+\|\omega\|_r). \tag{3.48}$$

When employing unequal-order continuous interpolation $\mathcal{P}_{r+1}-\mathcal{P}_r-\mathcal{P}_r$ *with* $r\geq 3$ *for methods* (I) *and* (II)*, or for method* (II) *employing* $\mathcal{P}_{r+1}-\mathcal{P}_r^+-\mathcal{P}_r^+$ *with* $r=1,2$ *for* $(\mathbf{u},p,\omega)\in(H^{r+2}(\Omega))^2\times H^{r+1}(\Omega)\times H^{r+1}(\Omega)$*, we have*

$$\|\mathbf{u}-\mathbf{u}_h\|_1+\|p-p_h\|_0+\|\omega-\omega_h\|_0\leq C\,h^{r+1}\ (\|\mathbf{u}\|_{r+2}+\|p\|_{r+1}+\|\omega\|_{r+1}). \tag{3.49}$$

Proof. Let U_h be the piecewise $\mathcal{P}_l$ continuous element with $l\geq 1$ and let $\tilde{\mathbf{u}}\in U_h$ be an interpolant of $\mathbf{u}\in(H^k(\Omega))^2$ satisfying (see [2, 3, 5, 29])

$$\|\tilde{\mathbf{u}}-\mathbf{u}\|_1\leq C\,h^{s-1}\,\|\mathbf{u}\|_s,\quad 1\leq s\leq\min(k,l+1). \tag{3.50}$$

From Theorem 3.2 and Assumption (A2) we immediately have (3.48) and (3.49). ∎

Remark 3.3. Note that for lower-order approximation spaces ($r=1,2$), the error estimates for method (I) are not optimal and are the same as those of the standard LS method [16], because Assumption (A2) does not hold. However, we can understand $\mathcal{P}_3$ as the enrichment of $\mathcal{P}_1$ or $\mathcal{P}_2$ with suitable element bubbles and edge bubbles. Denoting $\mathcal{P}_3$ as $\mathcal{P}_1^{\#}$ (the enrichment of $\mathcal{P}_1$) or $\mathcal{P}_2^{\square}$ (the enrichment of $\mathcal{P}_2$), we have (3.48) for method (I) with $r=1$ or $r=2$, using $\mathcal{P}_1-\mathcal{P}_1^{\#}-\mathcal{P}_1^{\#}$ or $\mathcal{P}_2-\mathcal{P}_2^{\square}-\mathcal{P}_2^{\square}$ for $(\mathbf{u},p,\omega)$, and we have (3.49) for method (I) with $r=1$ or $r=2$, using $\mathcal{P}_2-\mathcal{P}_1^{\#}-\mathcal{P}_1^{\#}$ or $\mathcal{P}_3-\mathcal{P}_2^{\square}-\mathcal{P}_2^{\square}$ for $(\mathbf{u},p,\omega)$.

Before closing this section we verify Assumption (A2) using triangular finite elements in a 2D domain. For rectangular elements in a 2D or three-dimensional (3D) domain and tetrahedrons in a 3D domain the verification is similar. In the following we use the $\mathcal{P}_3$ element as an example to show how to verify Assumption (A2).

THEOREM 3.4. *If* $\omega\in H^{l+1}(\Omega)$ *and* $p\in H^{m+1}(\Omega)$*, with* $l,m\geq 0$*, then there exist* $\tilde{p}\in P_h$ *and* $\tilde{\omega}\in W_h$ *such that* (3.35)–(3.38) *hold. Here*

$$W_h:=\{q\in H^1(\Omega);q_{|_K}\in\mathcal{P}_3(K)\ \forall K\in\mathcal{C}_h\},\quad P_h:=W_h\cap L_0^2(\Omega). \tag{3.51}$$

Proof. Consider ω as an example. We shall follow the idea of [2, Lemma A.3, p. 100]. In order to include the less regular case $l=m=0$, we first let $\omega^0\in W_h$ be

an interpolation which satisfies [2, p. 111]

$$\|\omega-\omega^0\|_0+\Big(\sum_{K\in\mathcal{C}_h} h_K^2\,|\omega-\omega^0|_{1,K}^2\Big)^{1/2}\le C\,h^s\,\|\omega\|_s \tag{3.52}$$

and

$$\Big(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,\|\omega-\omega^0\|_{0,K}^2\Big)^{1/2}+|\omega-\omega^0|_1\le C\,h^{s-1}\,\|\omega\|_s, \tag{3.53}$$

where $1\le s\le \min(4,l+1)$. We then define $\tilde{\omega}\in W_h$ in (3.54)–(3.56):

$$\tilde{\omega}(i)=\omega^0(i),\quad 1\le i\le 3\quad \forall K\in\mathcal{C}_h, \tag{3.54}$$

$$\int_E(\tilde{\omega}-\omega)\,v=0\quad \forall v\in\mathcal{P}_1(E)\quad \forall E\in\partial K\quad \forall K\in\mathcal{C}_h, \tag{3.55}$$

where $\mathcal{P}_1(E)$ is the space of linear polynomials on E,

$$\int_K(\tilde{\omega}-\omega)=0\quad \forall K\in\mathcal{C}_h. \tag{3.56}$$

We show that

$$\breve{R}_h(\bigtriangledown\times(\tilde{\omega}-\omega))=\mathbf{0},\quad R_h(\bigtriangledown\times(\tilde{\omega}-\omega))=\mathbf{0}. \tag{3.57}$$

In fact, from the definition (2.5) of $\breve{R}_h$, (3.55), and (3.56) we have

$$\begin{aligned}\int_K \breve{R}_h(\bigtriangledown\times(\omega-\tilde{\omega}))\,\mathbf{v}&=\int_K\bigtriangledown\times(\omega-\tilde{\omega})\,\mathbf{v}\\&=\int_K(\omega-\tilde{\omega})\;\bigtriangledown\times\mathbf{v}-\sum_{E\in\partial K}\int_E(\omega-\tilde{\omega})\,\mathbf{v}\times\mathbf{n}_E\\&=0\quad\forall\mathbf{v}\in(\mathcal{P}_1(K))^2,\quad\forall K\in\mathcal{C}_h,\end{aligned} \tag{3.58}$$

where we have used the fact that $\bigtriangledown\times\mathbf{v}_{|_K}\in\mathcal{P}_0(K)$, $\mathbf{v}\times\mathbf{n}_E|_E\in\mathcal{P}_1(E)$, since $\mathbf{v}_{|_K}\in(\mathcal{P}_1(K))^2$. It follows that $\breve{R}_h(\bigtriangledown\times(\tilde{\omega}-\omega))=\mathbf{0}$. Similarly, since $V_{0,h}$ is a piecewise linear polynomial space and

$$(R_h(\bigtriangledown\times(\omega-\tilde{\omega})),\mathbf{v})_h=(\omega-\tilde{\omega},\bigtriangledown\times\mathbf{v})=\sum_{K\in\mathcal{C}_h}(\omega-\tilde{\omega},\bigtriangledown\times\mathbf{v})_{0,K}=0\quad\forall\mathbf{v}\in V_{0,h},$$

we have $R_h(\bigtriangledown\times(\tilde{\omega}-\omega))=\mathbf{0}$. The approximation property (3.37) follows the standard routine and (3.52)–(3.56); cf. [2, Lemma A.4, p. 101 or pp. 136–138] for a similar argument. Similarly, we can construct $\tilde{p}\in P_h$, satisfying similar properties as $\tilde{\omega}$. We proceed as follows: first let $\bar{p}$ be constructed as above, and then let $\tilde{p}=\bar{p}-\frac{1}{|\Omega|}\int_\Omega\bar{p}=\bar{p}-\frac{1}{|\Omega|}\int_\Omega(\bar{p}-p)$, because of $\int_\Omega p=0$. ∎

Remark 3.4. Methods (I) and (II) and their analysis cover nonaffine families of finite elements such as quads and hexes. In the analysis, only the verification of Assumption (A2) may involve the mapping. In the case of nonaffine mapping families, such verification can be done on the reference element through the mapping. We

consider 2D quadrilaterals. Let F_K denote the mapping from the reference element $\hat{K}$ to K, which associates the function q defined on K with the function $\hat{q}$ defined on $\hat{K}$ by $q = \hat{q} \circ F_K^{-1}$. We define $Z_h = \{\mathbf{v} \in (L^2(\Omega))^2; \mathbf{v}|_K \circ F_K \in (Q_1(\hat{K}))^2 \ \forall K \in \mathcal{C}_h\}$, with $Q_r(\hat{K})$, $r \geq 1$, being the standard element [3, 5, 29]. As approximating spaces of the pressure and vorticity, for example, we take $P_h = W_h = \{q \in H^1(\Omega); q|_K \circ F_K \in Q_4(\hat{K}) \ \forall K \in \mathcal{C}_h\}$ for method (I) and $P_h = W_h = \{q \in H^1(\Omega); q|_K \circ F_K \in Q_1(\hat{K}) + \hat{b}\, Q_2^-(\hat{K}) \ \forall K \in \mathcal{C}_h\}$ for method (II), where $\hat{b}$ is a bubble on $\hat{K}$ and $Q_2^-(\hat{K})$ is the standard reduced biquadratic element. We can easily verify Assumption (A2) through F_K by following the argument in Theorem 3.4 and the interpolation theory in [2, p. 108].

4. L^2-error bound.

In this section we establish the L^2-error bound for velocity.

4.1. Additional assumptions.

We first make a few more assumptions.

Assumption (A3). For all $\mathbf{u}_h, \mathbf{v}_h \in V_{0,h}$, there holds

$$|(\mathbf{u}_h, \mathbf{v}_h) - (\mathbf{u}_h, \mathbf{v}_h)_h| \leq C\, h\, \|\mathbf{u}_h\|_1\, \|\mathbf{v}_h\|_0. \tag{4.1}$$

Remark 4.1. Taking (2.12) as an example, Assumption (A3) holds (see [1, 3, 29]).

LEMMA 4.1. *Let Assumptions* (A1) *and* (A3) *hold. For any given* $\mathbf{u} \in V_{0,h}$, *we have*

$$\|\mathbf{u} - \bar{R}_h(\mathbf{u})\|_h \leq C\, h\, \|\mathbf{u}\|_1. \tag{4.2}$$

Proof. By Assumptions (A1) and (A3) and (2.15) we have (4.2), since

$$\begin{aligned} \|\bar{R}_h(\mathbf{u}) - \mathbf{u}\|_h^2 &= (\bar{R}_h(\mathbf{u}) - \mathbf{u}, \bar{R}_h(\mathbf{u}) - \mathbf{u})_h \\ &= (\bar{R}_h(\mathbf{u}), \bar{R}_h(\mathbf{u}) - \mathbf{u})_h - (\mathbf{u}, \bar{R}_h(\mathbf{u}) - \mathbf{u})_h \\ &= (\mathbf{u}, \bar{R}_h(\mathbf{u}) - \mathbf{u}) - (\mathbf{u}, \bar{R}_h(\mathbf{u}) - \mathbf{u})_h \\ &\leq C\, h\, \|\mathbf{u}\|_1\, \|\bar{R}_h(\mathbf{u}) - \mathbf{u}\|_0 \leq C\, h\, \|\mathbf{u}\|_1\, \|\bar{R}_h(\mathbf{u}) - \mathbf{u}\|_h. \end{aligned} \tag{4.3}$$

∎

Assumption (A4). For any given $\mathbf{f} \in (L^2(\Omega))^d$ and $\mathbf{v} \in (H^2(\Omega) \cap H_0^1(\Omega))^d$, the problem

$$-\Delta\, \mathbf{u} + \bigtriangledown\, p = \mathbf{f}, \quad \bigtriangledown \cdot\, \mathbf{u} = \bigtriangledown \cdot\, \mathbf{v} \quad \text{in } \Omega, \quad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma \tag{4.4}$$

has a solution $(\mathbf{u}, p)$ satisfying

$$\|\mathbf{u}\|_2 + \|p\|_1 \leq C\, \{\|\mathbf{f}\|_0 + \|\bigtriangledown \cdot \mathbf{v}\|_1\}. \tag{4.5}$$

PROPOSITION 4.1. *For a convex polygon, Assumption* (A4) *holds.*

Proof. From [6, 8] we know that for $\mathbf{v} \in (H^2(\Omega) \cap H_0^1(\Omega))^2$ we can find a $\mathbf{z} \in (H^2(\Omega) \cap H_0^1(\Omega))^2$ such that

$$\bigtriangledown \cdot\, \mathbf{z} = \bigtriangledown \cdot\, \mathbf{v}, \quad \|\mathbf{z}\|_2 \leq C\, \|\bigtriangledown \cdot \mathbf{v}\|_1. \tag{4.6}$$

We then consider the Stokes problem

$$-\Delta\, \mathbf{w} + \bigtriangledown\, p = \mathbf{f} + \Delta\, \mathbf{z}, \quad \bigtriangledown \cdot\, \mathbf{w} = 0 \quad \text{in } \Omega, \quad \mathbf{w} = \mathbf{0} \quad \text{on } \Gamma. \tag{4.7}$$

From [2, 7] we know that the problem (4.7) has a solution $(\mathbf{w}, p)$ satisfying

$$\|\mathbf{w}\|_2 + \|p\|_1 \leq C\, \{\|\mathbf{f}\|_0 + \|\Delta\, \mathbf{z}\|_0\} \leq C\, \{\|\mathbf{f}\|_0 + \|\bigtriangledown \cdot \mathbf{v}\|_1\}. \tag{4.8}$$

We therefore define

$$\mathbf{u} := \mathbf{z} + \mathbf{w}. \tag{4.9}$$

Clearly, such $(\mathbf{u}, p)$ satisfies (4.4) and (4.5). ∎

Assumption (A5). For any given $\mathbf{f} \in (L^2(\Omega))^d$ and for all $\lambda \geq 0$, the elasticity problem

$$-\Delta\, \mathbf{u} - \lambda\, \bigtriangledown \bigtriangledown \cdot\, \mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \quad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma \tag{4.10}$$

has a solution $\mathbf{u}$ satisfying

$$\|\mathbf{u}\|_2 + \lambda\, \|\bigtriangledown \cdot\, \mathbf{u}\|_1 \leq C\, \|\mathbf{f}\|_0, \tag{4.11}$$

where C is independent of λ.

Remark 4.2. When Ω is a convex polygon in $\mathbb{R}^2$, Assumption (A5) holds (see [6, 10]).

4.2. L^2-error bound for velocity. We can now establish the L^2-error bound.

THEOREM 4.1. *Under the assumptions of Theorem* 3.2 *and Assumptions* (A3)–(A5), *if* $V_{0,h} \subseteq U_h$, *we then have*

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq C\, h\, \Lambda, \tag{4.12}$$

where $\Lambda = \Lambda(\tilde{\mathbf{u}} - \mathbf{u}, \tilde{p} - p, \tilde{\omega} - \omega)$ *is defined by* (3.40).

Proof. We consider only method (II). The same argument for method (I) is straightforward. In principle, we follow the Aubin–Nitsche duality technique [29].

We consider the following auxiliary problem: Given $\mathbf{u} - \mathbf{u}_h \in (L^2(\Omega))^d$, find ρ such that

$$-\Delta\, \rho - \lambda\, \bigtriangledown\, \bigtriangledown \cdot\, \rho = \mathbf{u} - \mathbf{u}_h \quad \text{in } \Omega, \quad \rho = \mathbf{0} \quad \text{on } \Gamma, \tag{4.13}$$

which can also be expressed as the mixed form equivalently:

$$-\Delta\, \rho - \bigtriangledown\, \kappa = \mathbf{u} - \mathbf{u}_h, \quad \bigtriangledown \cdot\, \rho - \lambda^{-1}\, \kappa = 0 \quad \text{in } \Omega, \quad \rho = \mathbf{0} \quad \text{on } \Gamma. \tag{4.14}$$

From Assumption (A5) we have

$$\|\rho\|_2 + \|\kappa\|_1 + \lambda\, \|\bigtriangledown \cdot \rho\|_1 \leq C\, \|\mathbf{u} - \mathbf{u}_h\|_0, \tag{4.15}$$

which holds for any $\lambda \geq 0$. We shall take

$$\lambda := \frac{1}{h}. \tag{4.16}$$

Let

$$e_u := \mathbf{u} - \mathbf{u}_h, \quad e_p := p - p_h, \quad e_\omega := \omega - \omega_h. \tag{4.17}$$

From (4.14) we have

$$\begin{aligned} \|e_u\|_0^2 &= (-\Delta\, \rho - \bigtriangledown\, \kappa, e_u) \\ &= (\bigtriangledown \times \bigtriangledown \times \rho - h\, \bigtriangledown\, \kappa, e_u) + (\kappa, \bigtriangledown \cdot\, e_u) \\ &= (\bigtriangledown \times \rho, \bigtriangledown \times e_u) + h\, (\kappa, \bigtriangledown \cdot\, e_u) + (\kappa, \bigtriangledown \cdot\, e_u) \\ &= (e_\omega - \bigtriangledown \times e_u, - \bigtriangledown \times \rho) + (e_\omega, \bigtriangledown \times \rho) \\ &\quad + h\, (\kappa, \bigtriangledown \cdot\, e_u) + (\kappa, \bigtriangledown \cdot\, e_u) \\ &:= I_1 + I_2 + I_3 + I_4, \end{aligned} \tag{4.18}$$

where, by (3.41),

$$I_3 \le |h\,(\kappa, \bigtriangledown \cdot\, e_u)| \le C\ h \|\kappa\|_0 \|e_u\|_1 \le C\,h\,\|\kappa\|_1\,\Lambda. \tag{4.19}$$

Let $\bar\rho \in V_{0,h}$ be such that (see [2, 3, 4, 29])

$$\Big(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,\|\bar\rho-\rho\|_{0,K}^2 + \sum_{E\in\mathcal{E}_h} h_E^{-1}\int_E |\bar\rho-\rho|^2\Big)^{1/2} + \|\bar\rho-\rho\|_1 \le C\,h\,\|\rho\|_2. \tag{4.20}$$

We have

$$I_2 = (e_\omega, \bigtriangledown\times\bar\rho) + (e_\omega, \bigtriangledown\times(\rho-\bar\rho)), \tag{4.21}$$

where, by (3.41) and (4.20),

$$|(e_\omega, \bigtriangledown\times(\rho-\bar\rho))| \le C\|e_\omega\|_0\|\rho-\bar\rho\|_1 \le Ch\,\|\rho\|_2\Lambda, \tag{4.22}$$

$$\begin{aligned}(e_\omega, \bigtriangledown\times\bar\rho) &= (R_h(\bigtriangledown\times e_\omega), \bar\rho)_h\\ &= (R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar\rho)_h - (S_h(\bigtriangledown e_p), \bar\rho)_h\\ &= (R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar\rho - \bar R_h(\bar\rho))_h\\ &\quad + (R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar R_h(\bar\rho))_h\\ &\quad + (e_p, \bigtriangledown\cdot\,(\bar\rho-\rho) + h\,\kappa),\end{aligned} \tag{4.23}$$

and

(4.24)
$$\begin{cases} |(R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar\rho - \bar R_h(\bar\rho))_h + (e_p, \bigtriangledown\cdot(\bar\rho-\rho) + h\,\kappa)|\\ \le \|R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p)\|_h\,\|\bar\rho - \bar R_h(\bar\rho)\|_h + C\,\|e_p\|_0\,\{\|\bar\rho-\rho\|_1 + h\,\|\kappa\|_0\}\\ \le C\,h\,\{\|\rho\|_2 + \|\kappa\|_1\}\,\Lambda \qquad \text{(by (3.43), Lemma 4.1, (3.41), and (4.20))}.\end{cases}$$

From (4.18), (4.19), and (4.21)–(4.24) we need only to estimate

$$\begin{cases} I_0 := I_1 + I_4 + (R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar R_h(\bar\rho))_h\\ = (e_\omega - \bigtriangledown\times e_u, -\bigtriangledown\times\rho) + (\kappa, \bigtriangledown\cdot\,e_u)\\ \quad + (R_h(\bigtriangledown\times e_\omega) + S_h(\bigtriangledown e_p), \bar R_h(\bar\rho))_h.\end{cases} \tag{4.25}$$

To do so, we consider an auxiliary problem: Find $(\mathbf{u}^*, p^*)$ such that

$$-\Delta\,\mathbf{u}^* + \bigtriangledown\,p^* = \bar\rho + e_u + h\ \bigtriangledown\ \kappa, \quad \bigtriangledown\cdot\,\mathbf{u}^* = \kappa \quad \text{in } \Omega, \quad \mathbf{u}^* = \mathbf{0} \quad \text{on } \Gamma. \tag{4.26}$$

Since $\kappa = h^{-1}\ \bigtriangledown\cdot\rho = \bigtriangledown\cdot\,(h^{-1}\,\rho)$ and $h^{-1}\,\rho \in (H^2(\Omega)\cap H_0^1(\Omega))^d$, from Assumption (A4), (4.20), and (4.15) we have

$$\|\mathbf{u}^*\|_2 + \|p^*\|_1 \le C\,\{\|\bar\rho\|_0 + \|e_u\|_0 + \|\kappa\|_1\} \le C\,\|e_u\|_0. \tag{4.27}$$

Inserting

$$\bigtriangledown\times\bigtriangledown\times\rho - h\ \bigtriangledown\ \kappa - \bigtriangledown\kappa = e_u \qquad \text{(by the first equation of (4.14))} \tag{4.28}$$

into the first equation of (4.26) we get

$$\bigtriangledown \times (\bigtriangledown \times \mathbf{u}^* - \bigtriangledown \times \rho) + \bigtriangledown p^* = \bar{\rho}. \tag{4.29}$$

Let

$$\phi := \bigtriangledown \times \mathbf{u}^* - \bigtriangledown \times \rho, \tag{4.30}$$

and we have

$$\bigtriangledown \times \phi + \bigtriangledown p^* = \bar{\rho}, \quad \|\phi\|_1 \le C \|e_u\|_0. \tag{4.31}$$

Therefore, noting that

$$\bar{R}_h(\bar{\rho}) = \bar{R}_h(\bigtriangledown \times \phi + \bigtriangledown p^*) = R_h(\bigtriangledown \times \phi) + S_h(\bigtriangledown p^*) \qquad \text{(by (2.16))}, \tag{4.32}$$

we have

$$\begin{cases} I_0 = (e_\omega - \bigtriangledown \times e_u, \phi - \bigtriangledown \times \mathbf{u}^*) + (\bigtriangledown \cdot \mathbf{u}^*, \bigtriangledown \cdot e_u) \\ \quad + (R_h(\bigtriangledown \times e_\omega) + S_h(\bigtriangledown e_p), R_h(\bigtriangledown \times \phi) + S_h(\bigtriangledown p^*))_h \\ \quad = (e_\omega - \bigtriangledown \times e_u, \phi - \tilde{\phi} - \bigtriangledown \times (\mathbf{u}^* - \tilde{\mathbf{u}}^*)) + (\bigtriangledown \cdot (\mathbf{u}^* - \tilde{\mathbf{u}}^*), \bigtriangledown \cdot e_u) \\ \quad + (R_h(\bigtriangledown \times e_\omega) + S_h(\bigtriangledown e_p), R_h(\bigtriangledown \times (\phi - \tilde{\phi})) + S_h(\bigtriangledown (p^* - \tilde{p}^*)))_h \\ \quad + (e_\omega - \bigtriangledown \times e_u, \tilde{\phi} - \bigtriangledown \times \tilde{\mathbf{u}}^*) + (\bigtriangledown \cdot \tilde{\mathbf{u}}^*, \bigtriangledown \cdot e_u) \\ \quad + (R_h(\bigtriangledown \times e_\omega) + S_h(\bigtriangledown e_p), R_h(\bigtriangledown \times \tilde{\phi}) + S_h(\bigtriangledown \tilde{p}^*))_h, \end{cases} \tag{4.33}$$

where we have chosen $(\tilde{\mathbf{u}}^*, \tilde{p}^*, \tilde{\phi}) \in U_h \times P_h \times W_h$ such that $\tilde{\mathbf{u}}^* \in V_{0,h} \subseteq U_h$ and Assumption (A2) is satisfied with $s = 1$:

$$\|\mathbf{u}^* - \tilde{\mathbf{u}}^*\|_0 + h \|\mathbf{u}^* - \tilde{\mathbf{u}}^*\|_1 \le C h^2 \|\mathbf{u}^*\|_2, \tag{4.34}$$

$$R_h(\bigtriangledown \times (\phi - \tilde{\phi})) + S_h(\bigtriangledown (p^* - \tilde{p}^*)) = \mathbf{0}, \tag{4.35}$$

$$\begin{aligned} \|\tilde{\phi} - \phi\|_0 + \|\tilde{p}^* - p^*\|_0 + |(\tilde{p}^* - p^*, \tilde{\phi} - \phi)|_{(\mathcal{C}_h, \mathcal{E}_h)} \\ \le C h \{\|p^*\|_1 + \|\phi\|_1\}. \end{aligned} \tag{4.36}$$

We also have

$$\begin{cases} |(e_\omega - \bigtriangledown \times e_u, \phi - \tilde{\phi} - \bigtriangledown \times (\mathbf{u}^* - \tilde{\mathbf{u}}^*)) + (\bigtriangledown \cdot (\mathbf{u}^* - \tilde{\mathbf{u}}^*), \bigtriangledown \cdot e_u)| \\ \le C h \{\|\mathbf{u}^*\|_2 + \|\phi\|_1\} \Lambda \qquad \text{(by (3.41), (4.36), and (4.34))}, \end{cases} \tag{4.37}$$

$$(R_h(\bigtriangledown \times e_\omega) + S_h(\bigtriangledown e_p), R_h(\bigtriangledown \times (\phi - \tilde{\phi})) + S_h(\bigtriangledown (p^* - \tilde{p}^*)))_h = 0 \tag{4.38}$$

(by (4.35)),

$$\begin{cases} (e_\omega - \bigtriangledown \times e_u, \tilde{\phi} - \bigtriangledown \times \tilde{\mathbf{u}}^*) + (\bigtriangledown \cdot \tilde{\mathbf{u}}^*, \bigtriangledown \cdot e_u) \\ \quad + (R_h(\bigtriangledown \times e_\omega) + S_h(\bigtriangledown e_p), R_h(\bigtriangledown \times \tilde{\phi}) + S_h(\bigtriangledown \tilde{p}^*))_h \\ = \mathcal{L}_h((e_u, e_p, e_\omega); (\tilde{\mathbf{u}}^*, \tilde{p}^*, \tilde{\phi})) \\ \quad - \sum_{K \in \mathcal{C}_h} h_K^2 (\bigtriangledown \times e_\omega + \bigtriangledown e_p, \bigtriangledown \times \tilde{\phi} + \bigtriangledown \tilde{p}^*)_{0,K} \\ \quad - \sum_{E \in \mathcal{E}_h} h_E \int_E [e_\omega] [\tilde{\phi} - \phi] - \sum_{E \in \mathcal{E}_h} h_E \int_E [e_p] [\tilde{p}^* - p^*], \end{cases} \tag{4.39}$$

where we have used the regularity of $\phi \in (H^1(\Omega))^{2d-3}$ and $p^* \in H^1(\Omega)$,

$$\mathcal{L}_h((e_u, e_p, e_\omega); (\tilde{\mathbf{u}}^*, \tilde{p}^*, \tilde{\phi})) = 0 \qquad \text{(by Lemma 3.2)}, \tag{4.40}$$

and

$$\begin{cases} \Big| - \sum\limits_{K\in\mathcal{C}_h} h_K^2 \, (\bigtriangledown \times e_\omega + \bigtriangledown e_p, \bigtriangledown \times \tilde{\phi} + \bigtriangledown \tilde{p}^*)_{0,K} \\ \quad - \sum\limits_{E\in\mathcal{E}_h} h_E \int_E [e_\omega]\,[\tilde{\phi} - \phi] - \sum\limits_{E\in\mathcal{E}_h} h_E \int_E [e_p]\,[\tilde{p}^* - p^*] \Big| \\ \le C\,|(e_p, e_\omega)|_{(\mathcal{C}_h,\mathcal{E}_h)}\,\{|(\tilde{p}^* - p^*, \tilde{\phi} - \phi)|_{(\mathcal{C}_h,\mathcal{E}_h)} + h\,(\|\phi\|_1 + \|p^*\|_1)\} \\ \le C\,h\,\{\|\phi\|_1 + \|p^*\|_1\}\,\Lambda \qquad \text{(by (3.42) and (4.36))}. \end{cases} \tag{4.41}$$

Therefore, summarizing (4.18), (4.19), (4.21)–(4.25), (4.33), and (4.37)–(4.41) and (4.15), (4.27), and (4.31), we finally obtain

$$\begin{aligned} \|e_u\|_0^2 &\le C\,h\,\Lambda\,\{\|\rho\|_2 + \|\kappa\|_1 + \|\mathbf{u}^*\|_2 + \|p^*\|_1 + \|\phi\|_1\} \\ &\le C\,h\,\|e_u\|_0\,\Lambda. \end{aligned} \tag{4.42}$$

This completes the proof. ∎

Remark 4.3. We now clarify the assumptions involved for the two methods. For the two methods the common trivial assumptions are that $\mathcal{C}_h$ is regular and Ω is a Lipschitz convex polygon (polyhedron in $\mathbb{R}^3$) and $V_{0,h} \subseteq U_h$ with (3.50) for U_h. For method (I) the additional assumptions are (A2), (A4), and (A5). For method (II) the additional assumptions are (A1)–(A5).

5. Numerical experiments. In this section we report some numerical experiments to illustrate the theoretical error bounds. We shall consider only method (II) in (2.18), employing continuous $U_h \times P_h \times W_h$, with h_K taken as h and $R_h = S_h = \bar{R}_h$. We shall use the two lower-order elements: $\mathcal{P}_1$-$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation and $\mathcal{P}_2$ -$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation. We take $\Omega := [0,1] \times [0,1]$ and partition it into uniform triangles.

We consider a 2D Stokes problem

$$-\Delta\,\mathbf{u} + \bigtriangledown\, p = \mathbf{f}, \quad \bigtriangledown \cdot\, \mathbf{u} = 0 \quad \text{in } \Omega, \quad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma. \tag{5.1}$$

We take the example from [17]. The exact solution $(\mathbf{u}, p, \omega)$ of (5.1) is

$$\begin{pmatrix} u_1 \\ u_2 \\ \omega \\ p \end{pmatrix} = \begin{pmatrix} x^2(1-x)^2(2y - 6y^2 + 4y^3) \\ y^2(1-y)^2(-2x + 6x^2 - 4x^3) \\ x^2(1-x)^2(-2 + 12y - 12y^2) + y^2(1-y)^2(-2 + 12x^2 - 12x^2) \\ x^2 + y^2 - \frac{20}{3}xy + x + y \end{pmatrix}.$$

The right-hand side $\mathbf{f}$ is generated by evaluating the equations on the given exact solution. In our experiment we set $p_h(0,0) = 0$ to replace $\int_\Omega p_h = 0$ to ensure the uniqueness. We employ the conjugate gradient method, with the stopping criterion tolerance 10^{-9} and with a zero initial guess, to solve the resulting linear system. We employ the $\mathcal{P}_r - \mathcal{P}_1^+ - \mathcal{P}_1^+$ approximations, $r = 1, 2$. We list the relative errors in L^2- and H^1-norms in Tables 1 and 2 and Tables 3 and 4, respectively.

Tables 1 and 2 show that the relative errors in L^2-norm and H^1-norm for velocity are $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, respectively, the same as predicted by Theorems 3.3 and 4.1. Tables 1 and 2 show that the relative errors in L^2-norm and H^1-norm for vorticity and pressure are also $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, respectively, higher than predicted.

TABLE 1
Relative errors in L^2-norm for $\mathcal{P}_1$-$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation.

	h=0.25	h=0.125	h=0.0625	h=0.03125
$\frac{\|u_1-u_{1,h}\|_0}{\|u_1\|_0}$	1.169467	0.314805	0.035818	0.007337
$\frac{\|u_2-u_{2,h}\|_0}{\|u_2\|_0}$	1.645415	0.422441	0.051579	0.010812
$\frac{\|\omega-\omega_h\|_0}{\|\omega\|_0}$	2.515323	0.5050733	0.0576332	0.0101290
$\frac{\|p-p_h\|_0}{\|p\|_0}$	0.951209	0.257992	0.007156	0.001472

TABLE 2
Relative errors in H^1-norm for $\mathcal{P}_1$-$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation.

	h=0.25	h=0.125	h=0.0625	h=0.03125
$\frac{\|u_1-u_{1,h}\|_1}{\|u_1\|_1}$	1.681586	0.526141	0.177909	0.006571
$\frac{\|u_2-u_{2,h}\|_1}{\|u_2\|_1}$	1.641017	0.486311	0.175564	0.086430
$\frac{\|\omega-\omega_h\|_1}{\|\omega\|_1}$	2.349043	0.597976	0.176431	0.076122
$\frac{\|p-p_h\|_1}{\|p\|_1}$	0.439165	0.175217	0.080606	0.039919

TABLE 3
Relative errors in L^2-norm for $\mathcal{P}_2$-$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation.

	h=0.25	h=0.125	h=0.0625
$\frac{\|u_1-u_{1,h}\|_0}{\|u_1\|_0}$	2.296299	0.404793	0.029696
$\frac{\|u_2-u_{2,h}\|_0}{\|u_2\|_0}$	2.470892	0.500211	0.061050
$\frac{\|\omega-\omega_h\|_0}{\|\omega\|_0}$	2.967468	0.583148	0.074835
$\frac{\|p-p_h\|_0}{\|p\|_0}$	1.198191	0.364830	0.097487

TABLE 4
Relative errors in H^1-norm for $\mathcal{P}_2$-$\mathcal{P}_1^+$-$\mathcal{P}_1^+$ approximation.

	h = 0.25	h = 0.125	h = 0.0625
$\frac{\|u_1-u_{1,h}\|_1}{\|u_1\|_1}$	2.393610	0.460135	0.057460
$\frac{\|u_2-u_{2,h}\|_1}{\|u_2\|_1}$	2.493563	0.522751	0.073181
$\frac{\|\omega-\omega_h\|_1}{\|\omega\|_1}$	2.717344	0.754938	0.249759
$\frac{\|p-p_h\|_1}{\|p\|_1}$	0.528725	0.198494	0.085175

Tables 3 and 4 show that relative errors in L^2- and H^1-norms for velocity are $\mathcal{O}(h^3)$ and $\mathcal{O}(h^2)$, respectively, and also show that the relative errors in L^2- and H^1-norms for vorticity and pressure are $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, respectively. These are consistent with what was predicted.

Acknowledgement. The authors would like to thank the referees for their valuable suggestions which helped to improve the content and presentation of the paper.

REFERENCES

[1] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

[2] V. Girault and P.-A. Raviart, *Finite Element Methods For Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[3] P. G. Ciarlet, *Basic Error Estimates for Elliptic Problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, New York, 1991, pp. 17–351.

[4] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, 2002.

[6] S. C. Brenner and L.-Y. Sung, *Linear finite element methods for planar linear elasticity*, Math. Comp., 59 (1992), pp. 321–338.

[7] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.

[8] D. N. Arnold, L. R. Scott, and M. Vogelius, *Regular inversion of the divergence operator with Dirichlet boundary conditions on a polygon*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 15 (1988), pp. 169–192.

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[10] C. Bacuta and J. H. Bramble, *Regularity estimates for solutions of the equations of linear elasticity in convex plane polygonal domains*, Z. Angew. Math. Phys., 54 (2003), pp. 874–878.

[11] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[12] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[13] B.-N. Jiang, *The Least-Squares Finite Element Method. Theory and Applications in Computational Dynamics and Electromagnetics*, Springer-Verlag, Heidelberg, 1998.

[14] J. H. Bramble and J. E. Pasciak, *Least-squares method for Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

[15] B.-N. Jiang and C.-L. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Method Appl. Mech. Engrg., 78 (1990), pp. 297–311.

[16] H.-Y. Duan and G.-P. Liang, *On the velocity-pressure-vorticity least-squares mixed finite element method for the* 3*D Stokes equations*, SIAM J. Numer. Anal., 41 (2003), pp. 2114–2130.

[17] C.-L. Chang and S.-Y. Yang, *Analysis of the L^2 least-squares finite element method for the velocity-vorticity-pressure Stokes equations with velocity boundary conditions*, Appl. Math. Comput., 130 (2002), pp. 121–144.

[18] P. B. Bochev and M. D. Gunzberger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[19] B.-N. Jiang, *Least-squares finite elements for incompressible Navier-Stokes problems*, Internat. J. Numer. Methods Fluids, 14 (1992), pp. 843–859.

[20] B.-N. Jiang, *On the least-squares method*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 239–257.

[21] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first-order systems*, Math. Comp., 66 (1997), pp. 935–955.

[22] G. Starke, *Multilevel boundary functionals for least-squares mixed finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1065–1077.

[23] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[24] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[25] J. H. Bramble, *A proof of the inf-sup condition for the Stokes equations on Lipschitz domains*, Math. Models Methods Appl. Sci., 13 (2003), pp. 361–371.

[26] C.-L. Chang, *Finite element method for the solution of Maxwell's equations in multiple media*, Appl. Math. Comput., 25 (1988), pp. 89–99.

[27] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[28] T.-F. Chen, *On least-squares approximations to compressible flow problems*, Numer. Methods Partial Differential Equations, 2 (1986), pp. 207–228.

[29] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, 1978.

[30] D. Lefebvre, J. Peraire, and K. Morgan, *Least-squares finite element solution of compressible and incompressible flows*, Internat. J. Numer. Methods Heat Fluid Flow, 2 (1992), pp. 99–113.

[31] B.-N. Jiang, T. Lin, and L. Povinelli, *Large-scale computation of incompressible viscous flow by least-squares finite element method*, Comput. Methods Appl. Mech. Engrg., 114 (1995), pp. 213–231.

[32] Z.-Q. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order least-squares for velocity-pressure-vorticity form of the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1995), pp. 150–159.

[33] J. M. Deang and M. D. Gunzburger, *Issues related to least-squares finite element methods for the Stokes equations*, SIAM J. Sci. Comput., 20 (1998), pp. 878–906.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 753–778

DISCONTINUOUS GALERKIN METHODS FOR FRIEDRICHS' SYSTEMS. I. GENERAL THEORY*

A. ERN† AND J.-L. GUERMOND‡

Abstract. This paper presents a unified analysis of discontinuous Galerkin methods to approximate Friedrichs' systems. An abstract set of conditions is identified at the continuous level to guarantee existence and uniqueness of the solution in a subspace of the graph of the differential operator. Then a general discontinuous Galerkin method that weakly enforces boundary conditions and mildly penalizes interface jumps is proposed. All the design constraints of the method are fully stated, and an abstract error analysis in the spirit of Strang's Second Lemma is presented. Finally, the method is formulated locally using element fluxes, and links with other formulations are discussed. Details are given for three examples, namely, advection-reaction equations, advection-diffusion-reaction equations, and the Maxwell equations in the so-called elliptic regime.

Key words. Friedrichs' systems, finite elements, partial differential equations, discontinuous Galerkin method

AMS subject classifications. 65N30, 65M60, 35F15

DOI. 10.1137/050624133

1. Introduction.

Discontinuous Galerkin (DG) methods were introduced in the 1970s, and their development has since followed two somewhat parallel routes depending on whether the PDE is hyperbolic or elliptic.

For hyperbolic PDEs, the first DG method was introduced by Reed and Hill in 1973 [28] to simulate neutron transport, and the first analysis of DG methods for hyperbolic equations in an already rather general and abstract form was performed by Lesaint and Raviart in 1974 [23, 24]. The analysis was subsequently improved by Johnson, Nävert, and Pitkäranta who established that the optimal order of convergence in the L^2-norm is $p+\frac{1}{2}$ if polynomials of degree p are used [21]. More recently, DG methods for hyperbolic and nearly hyperbolic equations experienced a significant development based on the ideas of numerical fluxes, approximate Riemann solvers, and slope limiters; see, e.g., Cockburn et al. [9] and references therein for a thorough review. This renewed interest in DG methods is stimulated by several factors including the flexibility offered by the use of nonmatching grids and the possibility to use high-order *hp*-adaptive finite element methods; see, e.g., Süli et al. [30].

For elliptic PDEs, DG methods originated from the early work of Nitsche on boundary-penalty methods [25] and the use of interior penalties (IP) to weakly enforce continuity conditions imposed on the solution or its derivatives across the interfaces between adjoining elements; see, e.g., Babuška [4], Babuška and Zlámal [3], Douglas and Dupont [13], Baker [6], Wheeler [31], and Arnold [2]. DG methods for elliptic problems in mixed form were introduced more recently. Initially, a discontinuous approximation was used solely for the primal variable, the flux being still discretized in a conforming fashion; see, e.g., Dawson [11, 12]. Then, a discontinuous approximation

*Received by the editors February 9, 2005; accepted for publication (in revised form) October 6, 2005; published electronically April 12, 2006.
http://www.siam.org/journals/sinum/44-2/62413.html
†CERMICS, Ecole nationale des ponts et chaussées, Champs sur Marne, 77455 Marne la Vallée Cedex 2, France (ern@cermics.enpc.fr).
‡Department of Mathematics, Texas A&M, College Station, TX 77843-3368 (guermond@math.tamu.edu) and LIMSI (CNRS-UPR 3251), BP 133, 91403, Orsay, France.

of both the primal variable and its flux has been introduced by Bassi and Rebay [7] and further extended by Cockburn and Shu [10] leading to the so-called local discontinuous Galerkin (LDG) method. Around the same time, Baumann and Oden [8] proposed a nonsymmetric variant of DG for elliptic problems. This method was further developed and analyzed by Oden, Babuška, and Baumann [26] and by Rivière, Wheeler, and Girault [29].

The fact that several DG methods (including IP methods) share common features and can be tackled by similar analysis tools called for a unified analysis. A first important step in that direction has been recently accomplished by Arnold et al. [1] for elliptic equations. It is shown in [1] that it is possible to cast many DG methods for the Poisson equation with homogeneous Dirichlet boundary conditions into a single framework amenable to a unified error analysis. The main idea consists of using the mixed formulation of the Poisson equation to define numerical fluxes and to locally eliminate these fluxes so as to derive a method involving only the primal variable.

The goal of the present paper is to propose a unified analysis of DG methods that goes beyond the traditional hyperbolic/elliptic classification of PDEs by making systematic use of the theory of Friedrichs' systems [17] to formulate DG methods and to perform the convergence analysis. This paper, which concentrates on first-order PDEs, is the first part of a more comprehensive study on DG methods for Friedrichs' systems. The forthcoming second part will deal more specifically with Friedrichs' systems associated with second-order PDEs. Some preliminary results on Friedrichs' systems related to this work can be found in [15, p. 227].

The paper is organized as follows. In section 2 we investigate the well posedness of Friedrichs' systems in graph spaces. Originally, Friedrichs addressed the question of the uniqueness of strong solutions in $\mathfrak{C}^1$ and that of the existence of weak solutions in L^2 [17]. The analysis of Friedrichs' systems in graph spaces has been undertaken by Rauch [27] and more recently by Jensen [20]. The main novelty of the present approach is that we avoid invoking traces at the boundary by introducing a bounded linear operator from the graph space to its dual that satisfies sufficient conditions ensuring well posedness. In section 3 we illustrate the abstract results of section 2 on three important examples of Friedrichs' systems, namely, advection-reaction equations, advection-diffusion-reaction equations, and a simplified version of the Maxwell equations in the so-called elliptic regime. Drawing on earlier ideas by Lesaint and Raviart [23, 24] and Johnson et al. [21], we propose in section 4 a general framework for DG methods. This section contains three main contributions. First, the generic DG method is formulated in terms of a boundary operator enforcing boundary conditions weakly and in terms of an interface operator penalizing the jumps of the solution across the mesh interfaces. Second, the convergence analysis is performed in the spirit of Strang's Second Lemma by using two different norms, namely, a stability norm for which a discrete inf-sup condition holds and an approximability norm ensuring the continuity of the DG bilinear form. All the design constraints to be fulfilled by the boundary and the interface operators for the error analysis to hold are clearly stated. Finally, using integration by parts, the DG method is reinterpreted locally by introducing the concept of element fluxes and element adjoint-fluxes, thus providing a direct link with engineering practice where approximation schemes are often designed by specifying such fluxes. Finally, section 5 reviews various DG approximations for the model problems investigated in section 3. In all the cases, the degrees of freedom in the design of the DG method are underlined.

2. Friedrichs' systems. The goal of this section is to reformulate Friedrichs' theory by giving special care to the meaning of the boundary conditions. The main results of this section are Theorems 2.5 and 2.8. Theorem 2.8 will be the starting point of the DG method developed in section 4.

2.1. The setting. Let Ω be a bounded, open, and connected Lipschitz domain in $\mathbb{R}^d$. We denote by $\mathfrak{D}(\Omega)$ the space of $\mathfrak{C}^\infty$ functions that are compactly supported in Ω. Let m be a positive integer. Let $\mathcal{K}$ and $\{\mathcal{A}^k\}_{1\le k\le d}$ be $(d+1)$ functions on Ω with values in $\mathbb{R}^{m,m}$. Following Friedrichs [17], we assume that

$$\text{(A1)} \qquad \mathcal{K} \in [L^\infty(\Omega)]^{m,m},$$

$$\text{(A2)} \qquad \forall k \in \{1,\dots,d\},\ \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m} \quad \text{and} \quad \sum_{k=1}^d \partial_k \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m},$$

$$\text{(A3)} \qquad \forall k \in \{1,\dots,d\},\ \mathcal{A}^k = (\mathcal{A}^k)^t \ \text{ a.e. in } \Omega,$$

$$\text{(A4)} \qquad \mathcal{Z} := \mathcal{K} + \mathcal{K}^t - \sum_{k=1}^d \partial_k \mathcal{A}^k \ge 2\mu_0 \mathcal{I}_m \ \text{ a.e. on } \Omega,$$

where $\mathcal{I}_m$ is the identity matrix in $\mathbb{R}^{m,m}$. Set $L = [L^2(\Omega)]^m$. We say that a function u in L has an A-weak derivative in L if the linear form

$$[\mathfrak{D}(\Omega)]^m \ni \varphi \longmapsto -\int_\Omega \sum_{k=1}^d u^t \partial_k(\mathcal{A}^k \varphi) \in \mathbb{R}, \tag{2.1}$$

is bounded on L, and we denote by Au the function in L that can be associated with the above linear form by means of the Riesz representation theorem. Clearly, if u is smooth enough, e.g., $u \in [\mathfrak{C}^1(\Omega)]^m$,

$$Au = \sum_{k=1}^d \mathcal{A}^k \partial_k u. \tag{2.2}$$

Define the graph space

$$W = \{w \in L;\ Aw \in L\}, \tag{2.3}$$

and equip W with the graph norm

$$\|w\|_W = \|Aw\|_L + \|w\|_L, \tag{2.4}$$

and the associated scalar product. W is a Hilbert space. Indeed, let v_n be a Cauchy sequence in W; i.e., v_n and Av_n are Cauchy sequences in L. Let v and w be the corresponding limits in L. Let $\varphi \in [\mathfrak{D}(\Omega)]^m$. Then, using the symmetry of $\mathcal{A}^k$ and an integration by parts yields

$$\int_\Omega \sum_{k=1}^d v^t \partial_k(\mathcal{A}^k\varphi) \leftarrow \int_\Omega \sum_{k=1}^d v_n^t \partial_k(\mathcal{A}^k\varphi) = -\int_\Omega \varphi^t A v_n \to -\int_\Omega \varphi^t w,$$

which means that v has an A-weak derivative in L and $Av = w$. Since $[\mathcal{D}(\Omega)]^m \subset W$ and $[\mathcal{D}(\Omega)]^m$ is dense in L, W is dense in L; as a result, we shall henceforth use L as a pivot space, i.e., $W \subset L \equiv L' \subset W'$. Note that owing to (A2), $[H^1(\Omega)]^m$ is a subspace of W.

Let $K \in \mathcal{L}(L;L)$ be defined such that $K : L \ni v \mapsto \mathcal{K}v \in L$ and set

$$T = A + K. \tag{2.5}$$

Then, $T \in \mathcal{L}(W;L)$. Let $K^* \in \mathcal{L}(L;L)$ be the adjoint operator of K, i.e., for all $v \in L$, $K^*v = \mathcal{K}^t v$. Let $\tilde{T} \in \mathcal{L}(W;L)$ be the formal adjoint of T,

$$\tilde{T}w = -\sum_{k=1}^{d} \partial_k(\mathcal{A}^k w) + K^* w \quad \forall w \in W. \tag{2.6}$$

In this definition $\sum_{k=1}^{d} \partial_k(\mathcal{A}^k w)$ is understood in the weak sense. It can easily be verified that this weak derivative exists in L whenever w is in W. Moreover, the usual rule for differentiating products applies. In particular, upon introducing the operator $\nabla\cdot A \in \mathcal{L}(L;L)$ such that $(\nabla\cdot A)w = (\sum_{k=1}^{d} \partial_k \mathcal{A}^k)w$ for all $w \in L$, the following holds

$$\forall w \in W, \quad Tw + \tilde{T}w = (K + K^* - \nabla\cdot A)\, w. \tag{2.7}$$

Observe that (A4) means that

$$\forall w \in W, \quad (Tw, w)_L + (w, \tilde{T}w)_L \geq 2\mu_0 \|w\|_L^2. \tag{2.8}$$

DEFINITION 2.1. *Let $D \in \mathcal{L}(W;W')$ be the operator such that*

$$\forall (u,v) \in W \times W, \quad \langle Du, v\rangle_{W',W} = (Tu, v)_L - (u, \tilde{T}v)_L. \tag{2.9}$$

This definition makes sense since both T and $\tilde{T}$ are in $\mathcal{L}(W;L)$. Note that D is a boundary operator in the sense that $[\mathfrak{D}(\Omega)]^m \subset \mathrm{Ker}(D)$; see also Remark 2.1. A more precise result (see [14]) is that $\mathrm{Ker}(D) = W_0$ and $\mathrm{Im}(D) = W_0^\perp$, where W_0 is the closure of $[\mathfrak{D}(\Omega)]^m$ in W and for any subset $E \subset W'$ we denote by $E^\perp$ the polar set of E, i.e., the set of the continuous linear forms in $W'' \equiv W$ that are zero on E.

LEMMA 2.2. *The operator D is self-adjoint.*

Proof. Let $(u,v) \in W \times W$ and set $Z = K + K^* - \nabla\cdot A$. A straightforward calculation yields

$$\begin{aligned}\langle Du, v\rangle_{W',W} - \langle Dv, u\rangle_{W',W} &= (Tu, v)_L - (u, \tilde{T}v)_L - (Tv, u)_L + (v, \tilde{T}u)_L \\ &= (Zu, v)_L - (u, Zv)_L = 0,\end{aligned}$$

since Z is self-adjoint. ∎

Remark 2.1. Let $n = (n_1, \ldots, n_d)^t$ be the unit outward normal to $\partial\Omega$. The usual way of presenting Friedrichs' systems consists of assuming that the fields $\{\mathcal{A}^k\}_{1\leq k\leq d}$ are smooth enough so that the matrix $\mathcal{D} = \sum_{k=1}^{d} n_k \mathcal{A}^k$ is meaningful at the boundary. Then, the operator D can be represented as follows

$$\langle Du, v\rangle_{W',W} = \int_{\partial\Omega} \sum_{k=1}^{d} v^t n_k \mathcal{A}^k u = \int_{\partial\Omega} v^t \mathcal{D} u,$$

whenever u and v are smooth functions. Provided $[\mathfrak{C}^1(\overline{\Omega})]^m$ is dense in $[H^1(\Omega)]^m$ and in W, it can be shown that $\mathcal{D}u \in [H^{-\frac{1}{2}}(\partial\Omega)]^m$. Further characterization and regularity results on $\mathcal{D}u$ can be found in [27] and in [20].

2.2. The well posedness result. Consider the following problem: For f in L, seek $u \in W$ such that $Tu = f$. In general, boundary conditions must be enforced for this problem to be well posed. In other words, one must find a closed subspace V of W such that the restricted operator $T : V \to L$ is an isomorphism.

The key hypothesis introduced by Friedrichs to select boundary conditions consists of assuming that there exists a matrix-valued field at the boundary, say, $\mathcal{M} : \partial\Omega \longrightarrow \mathbb{R}^{m,m}$, such that a.e. on $\partial\Omega$,

$$\mathcal{M} \text{ is positive, i.e., } (\mathcal{M}\xi, \xi)_{\mathbb{R}^m} \geq 0 \text{ for all } \xi \text{ in } \mathbb{R}^m, \tag{2.10}$$

$$\mathbb{R}^m = \mathrm{Ker}(\mathcal{D} - \mathcal{M}) + \mathrm{Ker}(\mathcal{D} + \mathcal{M}), \tag{2.11}$$

where $\mathcal{D}$ is defined in Remark 2.1. Then, it is possible to prove uniqueness of the so-called strong solution $u \in [\mathfrak{C}^1(\overline{\Omega})]^m$ of the PDE system $Tu = f$ supplemented with the boundary condition $(\mathcal{D} - \mathcal{M})u|_{\partial\Omega} = 0$. Moreover, it is also possible to prove existence of a weak solution in L, namely, of a function $u \in L$ such that the relation $(u, \tilde{T}v)_L = (f, v)_L$ holds for all $v \in [\mathfrak{C}^1(\overline{\Omega})]^m$ such that $(\mathcal{D} + \mathcal{M}^t)v|_{\partial\Omega} = 0$; see [27]. In this paper, we want to investigate the bijectivity of T in a subspace V of the graph W, and it is not possible to set $V = \{v \in W;\, (\mathcal{D} - \mathcal{M})v|_{\partial\Omega} = 0\}$ since the meaning of traces is not clear.

To overcome this difficulty, we modify Friedrichs' hypothesis by the following assumption: there exists an operator $M \in \mathcal{L}(W; W')$ such that

$$M \text{ is positive, i.e., } \langle Mw, w\rangle_{W',W} \geq 0 \text{ for all } w \text{ in } W, \tag{M1}$$

$$W = \mathrm{Ker}(D - M) + \mathrm{Ker}(D + M). \tag{M2}$$

Let $M^* \in \mathcal{L}(W; W')$ be the adjoint operator of M, i.e., for all $(u, v) \in W \times W$, $\langle M^*u, v\rangle_{W',W} = \langle Mv, u\rangle_{W',W}$. Then, one can prove (see [14]) that (M1)–(M2) imply that $\mathrm{Ker}(D) = \mathrm{Ker}(M)$, $\mathrm{Im}(D) = \mathrm{Im}(M)$, and

$$W = \mathrm{Ker}(D - M^*) + \mathrm{Ker}(D + M^*). \tag{2.12}$$

Since $\mathrm{Ker}(D) = \mathrm{Ker}(M)$, M is a boundary operator. Set

$$V = \mathrm{Ker}(D - M) \qquad \text{and} \qquad V^* = \mathrm{Ker}(D + M^*), \tag{2.13}$$

and equip V and V^* with the graph norm (2.4). The following result is proven in [14].

LEMMA 2.3. *Assume* (M1)–(M2). *Then,*

$$D(V)^\perp = V^* \qquad \text{and} \qquad D(V^*)^\perp = V. \tag{2.14}$$

LEMMA 2.4. *Assume* (A1)–(A4) *and* (M1)–(M2). *Then, T is L-coercive on V and $\tilde{T}$ is L-coercive on V^*.*

Proof. Using (2.8) and (2.9) yields

$$(Tw, w)_L \geq \mu_0\|w\|_L^2 + \frac{1}{2}\langle Dw, w\rangle_{W',W},$$
$$(\tilde{T}w, w)_L \geq \mu_0\|w\|_L^2 - \frac{1}{2}\langle Dw, w\rangle_{W',W}.$$

Use (2.13) and (M1) to conclude. ∎

THEOREM 2.5. *Assume* (A1)–(A4) *and* (M1)–(M2). *Let V and V^* be defined in* (2.13). *Then,*

(i) *$T : V \to L$ is an isomorphism.*

(ii) *$\tilde{T} : V^* \to L$ is an isomorphism.*

Proof. We only prove (i) since the proof of (ii) is similar.

(1) Owing to (2.13), V is closed in W; hence, V is a Hilbert space. As a result, showing that $T : V \to L$ is an isomorphism amounts to proving statement (ii) in Theorem 2.6 below with $L \equiv L'$.

(2) Proof of (2.15). Let $u \in V$. Observe that $\sup_{v\in L\setminus\{0\}} \frac{(Tu,v)_L}{\|v\|_L} = \|Tu\|_L$. Lemma 2.4 implies $\|Tu\|_L \geq \mu_0\|u\|_L$. Furthermore,

$$\|Tu\|_L \geq \|Au\|_L - \|K\|_{\mathcal{L}(L;L)}\|u\|_L \geq \|Au\|_L - \frac{\|K\|_{\mathcal{L}(L;L)}}{\mu_0}\|Tu\|_L.$$

This readily yields $\|Au\|_L \leq c\|Tu\|_L$ and thus $\|u\|_W \leq c\|Tu\|_L$.

(3) Proof of (2.16). Assume that $v \in L$ is such that $(Tu, v)_L = 0$ for all $u \in V$. Since $[\mathfrak{D}(\Omega)]^m \subset V$, a standard distribution argument shows that $\tilde{T}v = 0$ in $[\mathfrak{D}'(\Omega)]^m$. Still in the distribution sense, this means that $\sum_{k=1}^d \mathcal{A}^k\partial_k v = K^*v - (\nabla\cdot A)v$. Since the right-hand side is a bounded linear functional on L, v has an A-weak derivative in L, i.e., $v \in W$. As a result, $\langle Du, v\rangle_{W',W} = 0$ for all $u \in V$, i.e., $v \in D(V)^\perp$. Owing to Lemma 2.3, $v \in V^*$. Finally, since $(\tilde{T}v, v)_L = 0$ and $v \in V^*$, Lemma 2.4 implies that v is zero. ∎

THEOREM 2.6 (Banach–Nečas–Babuška (BNB)). *Let V, L be two Banach spaces, and denote by $\langle\cdot,\cdot\rangle_{L',L}$ the duality pairing between L' and L. The following statements are equivalent:*

(i) *$T \in \mathcal{L}(V; L)$ is bijective.*

(ii) *There exists a constant $\alpha > 0$ such that*

$$\forall u \in V, \quad \sup_{v\in L'\setminus\{0\}} \frac{\langle v, Tu\rangle_{L',L}}{\|v\|_{L'}} \geq \alpha\|u\|_V, \tag{2.15}$$

$$\forall v \in L', \quad (\langle v, Tu\rangle_{L',L} = 0 \ \forall u \in V) \implies (v = 0). \tag{2.16}$$

As an immediate consequence of Theorem 2.5, the following problems are well posed: For f in L,

$$\text{seek } u \in V \text{ such that } Tu = f, \tag{2.17}$$

$$\text{seek } u^* \in V^* \text{ such that } \tilde{T}u^* = f. \tag{2.18}$$

Remark 2.2. To guarantee that $T : V \to L$ and $\tilde{T} : V^* \to L$ are isomorphisms, it is also possible to specify assumptions on the spaces V and V^* without using the boundary operator M. Introduce the cones $C^\pm = \{w \in W;\ \pm\langle Dw, w\rangle_{W',W} \geq 0\}$. Then, under the following assumptions:

$$V \subset C^+ \text{ and } V^* \subset C^-, \tag{V1}$$

$$V^* = D(V)^\perp \text{ and } V = D(V^*)^\perp, \tag{V2}$$

$T : V \to L$ and $\tilde{T} : V^* \to L$ are isomorphisms [14]. This way of introducing Friedrichs' systems seems to be new. We think that assumptions (V1)–(V2) are more natural than (M1)–(M2) since they do not involve the somewhat ad hoc operator M.

2.3. Boundary conditions weakly enforced. As we have in mind to solve (2.17) by means of DG methods with the boundary conditions weakly enforced, we

now propose alternative formulations of (2.17) and (2.18). Define the bilinear forms

$$a(u,v) = (Tu,v)_L + \frac{1}{2}\langle (M-D)u, v\rangle_{W',W}, \tag{2.19}$$

$$a^*(u,v) = (\tilde{T}u,v)_L + \frac{1}{2}\langle (M^*+D)u, v\rangle_{W',W}. \tag{2.20}$$

It is clear that a and a^* are in $\mathcal{L}(W\times W;\mathbb{R})$. A remarkable property is the following lemma.

LEMMA 2.7. *Under assumption* (A4)*, the following holds for all* $w\in W$*,*

$$a(w,w) \geq \mu_0\|w\|_L^2 + \frac{1}{2}\langle Mw, w\rangle_{W',W}, \tag{2.21}$$

$$a^*(w,w) \geq \mu_0\|w\|_L^2 + \frac{1}{2}\langle Mw, w\rangle_{W',W}. \tag{2.22}$$

As a result, a and a^* are L-coercive on W whenever (A4) *and* (M1) *hold.*

Proof. Let $w\in W$. Owing to (2.9),

$$\begin{aligned} a(w,w) &= (Tw,w)_L - \frac{1}{2}\langle Dw, w\rangle_{W',W} + \frac{1}{2}\langle Mw, w\rangle_{W',W} \\ &= \frac{1}{2}((T+\tilde{T})w, w)_L + \frac{1}{2}\langle Mw, w\rangle_{W',W}. \end{aligned}$$

Hence, (2.21) follows from (2.8). The proof of (2.22) is similar. ∎

Consider the following problems: For $f\in L$,

$$\text{seek } u\in W \text{ such that } a(u,v) = (f,v)_L \ \forall v\in W, \tag{2.23}$$

$$\text{seek } u^*\in W \text{ such that } a^*(u^*,v) = (f,v)_L \ \forall v\in W. \tag{2.24}$$

THEOREM 2.8. *Assume* (A1)–(A4) *and* (M1)–(M2)*. Then,*

(i) *There is a unique solution to* (2.23) *and this solution solves* (2.17)*;*

(ii) *There is a unique solution to* (2.24) *and this solution solves* (2.18).

Owing to Theorem 2.5, there is a unique $u\in V$ solving $Tu=f$. Moreover, since u is in V, $(D-M)u=0$. Hence, $a(u,v)=(f,v)_L$ for all $v\in W$, i.e., u solves (2.23). In addition, since a is L-coercive on W owing to Lemma 2.7, it is clear that the solution to (2.23) is unique. This proves statement (i). The proof of the second statement is similar. ∎

Remark 2.3. Neither the bilinear form a nor the bilinear form a^* induce an isomorphism between W and W'. In particular, there is no guarantee that (2.23) or (2.24) has a solution if the right-hand side is replaced by $\langle f, v\rangle_{W',W}$ whenever $f\in W'$.

3. Examples. This section discusses admissible boundary conditions for three important examples of Friedrichs' systems: advection-reaction equations, advection-diffusion-reaction equations, and a simplified version of the Maxwell equations in the elliptic regime. We stress the fact that the existence of an operator $M\in\mathcal{L}(W;W')$ such that (M1)–(M2) hold provides sufficient conditions for well posedness. Although the existence of $M\in\mathcal{L}(W;W')$ may not be granted in all cases (this is reflected, for instance, in the necessity to make assumption (H2) to treat advection–reaction equations; see section 3.1), the formalism appears to be general enough to treat advection-diffusion-reaction equations, and Maxwell's equations in the elliptic regime; see sections 3.2 and 3.3.

3.1. Advection-reaction. Let β be a vector field in $\mathbb{R}^d$, assume $\beta \in [L^\infty(\Omega)]^d$, $\nabla\cdot\beta \in L^\infty(\Omega)$, and define

$$\partial\Omega^\pm = \{x \in \partial\Omega;\ \pm\beta(x)\cdot n(x) > 0\}, \tag{3.1}$$

as well as $\partial\Omega^0 = \partial\Omega\backslash(\overline{\partial\Omega^-} \cup \overline{\partial\Omega^+})$; $\partial\Omega^-$ is the inflow boundary, $\partial\Omega^+$ the outflow boundary, and $\partial\Omega^0$ the interior of the set $\{x \in \partial\Omega;\ \beta(x)\cdot n(x) = 0\}$.

Let μ be a function in $L^\infty(\Omega)$ such that

$$\mu(x) - \frac{1}{2}\nabla\cdot\beta(x) \geq \mu_0 > 0 \quad \text{a.e. in } \Omega, \tag{3.2}$$

and consider the advection-reaction equation

$$\mu u + \beta\cdot\nabla u = f. \tag{3.3}$$

This PDE falls into the category studied above by setting $Kv = \mu v$ for all $v \in L^2(\Omega)$, and $\mathcal{A}^k = \beta^k$ for $k \in \{1, \ldots, d\}$. It is clear that (A1)–(A4) hold with $m = 1$. The graph space is $W = \{w \in L^2(\Omega);\ \beta\cdot\nabla w \in L^2(\Omega)\}$.

Henceforth, we assume that

(H1) $\mathfrak{C}_0^1(\mathbb{R}^d)$ is dense in W,

(H2) $\partial\Omega^-$ and $\partial\Omega^+$ are well separated, i.e., $\operatorname{dist}(\partial\Omega^-, \partial\Omega^+) > 0$.

Hypothesis (H1) is a regularity assumption on Ω. It can be shown to hold by using Friedrichs' mollifier whenever Ω and β are smooth. Let $L^2(\partial\Omega; |\beta\cdot n|)$ be the space of real-valued functions that are square integrable with respect to the measure $|\beta\cdot n|\,\mathrm{d}x$ where $\mathrm{d}x$ is the Lebesgue measure on $\partial\Omega$.

LEMMA 3.1. *Provided* (H1)–(H2) *hold,*

(i) *The trace operator* $\gamma : \mathfrak{C}_0^1(\mathbb{R}^d) \ni v \longrightarrow v \in L^2(\partial\Omega; |\beta\cdot n|)$ *extends uniquely to a continuous operator on* W*;*

(ii) *The operator* D *has the following representation: for all* $u, v \in W$*,*

$$\langle Du, v\rangle_{W',W} = \int_{\partial\Omega} uv(\beta\cdot n). \tag{3.4}$$

Proof. Since $\partial\Omega^-$ and $\partial\Omega^+$ are well separated, there are two nonnegative functions ψ^- and ψ^+ in $\mathfrak{C}_0^1(\mathbb{R}^d)$ such that

$$\psi^- + \psi^+ = 1 \ \text{ on } \overline{\Omega}, \qquad \psi^-|_{\partial\Omega^+} = 0, \qquad \psi^+|_{\partial\Omega^-} = 0. \tag{3.5}$$

Let u be a function in $\mathfrak{C}_0^1(\mathbb{R}^d)$. Then,

$$\begin{aligned}\int_{\partial\Omega} u^2|\beta\cdot n| &= \int_{\partial\Omega} u^2(\psi^- + \psi^+)|\beta\cdot n| = \int_{\partial\Omega^-\cup\partial\Omega^0} u^2\psi^-|\beta\cdot n| + \int_{\partial\Omega^+\cup\partial\Omega^0} u^2\psi^+|\beta\cdot n| \\ &= -\int_{\partial\Omega} u^2\psi^-(\beta\cdot n) + \int_{\partial\Omega} u^2\psi^+(\beta\cdot n) = -\int_\Omega \nabla\cdot(u^2\psi^-\beta) + \int_\Omega \nabla\cdot(u^2\psi^+\beta).\end{aligned}$$

Hence, $0 \leq \int_{\partial\Omega} u^2|\beta\cdot n| \leq c(\psi^+, \psi^-)\|u\|_W^2$. Statement (i) follows from the density of $\mathfrak{C}_0^1(\mathbb{R}^d)$ in W. The proof of (ii) is an immediate consequence of the existence of traces in $L^2(\partial\Omega; |\beta\cdot n|)$. ∎

To specify boundary conditions, define for $u, v \in W$,

$$\langle Mu, v\rangle_{W',W} = \int_{\partial\Omega} uv|\beta\cdot n|. \tag{3.6}$$

LEMMA 3.2. *Let $M \in \mathcal{L}(W;W')$ be defined in* (3.6). *Then,*
(i) (M1)–(M2) *hold;*
(ii) $V = \{v \in W;\, v|_{\partial\Omega^-} = 0\}$ *and* $V^* = \{v \in W;\, v|_{\partial\Omega^+} = 0\}$.

Proof of (i). (M1) directly results from (3.6). Let ψ^+, ψ^- be the partition of unity introduced in (3.5). Let $w \in W$ and write $w = \psi^+ w + \psi^- w$. It is clear that $\psi^+ w \in \mathrm{Ker}(D-M)$ since for all $v \in W$, $\langle (D-M)\psi^+ w, v\rangle_{W',W} = \int_{\partial\Omega^+} \psi^+ vw(\beta\cdot n - |\beta\cdot n|) = 0$. Similarly, $\psi^- w \in \mathrm{Ker}(D+M)$. Hence, (M2) holds.

Proof of (ii). Let $v \in \mathrm{Ker}(D-M)$. Then, for all $w \in W$, $-2\int_{\partial\Omega^-} |\beta\cdot n| vw = 0$. Take $w = v$ to infer $v|_{\partial\Omega^-} = 0$; thus, $\mathrm{Ker}(D-M) \subset V$. Conversely, if $v|_{\partial\Omega^-} = 0$, it is clear that for all $w \in W$, $\langle (D-M)v, w\rangle_{W',W} = -2\int_{\partial\Omega^-} |\beta\cdot n| vw = 0$, i.e., $v \in \mathrm{Ker}(D-M)$. Proceed similarly to prove that $V^* = \{v \in W;\, v|_{\partial\Omega^+} = 0\}$. ∎

3.2. Advection-diffusion-reaction equations. Let $\beta : \Omega \longrightarrow \mathbb{R}^d$ be a vector field such that $\beta \in [L^\infty(\Omega)]^d$ and $\nabla\cdot\beta \in L^\infty(\Omega)$. Let μ be a function in $L^\infty(\Omega)$ such that (3.2) holds, and consider the advection-diffusion-reaction equation

$$-\Delta u + \beta\cdot\nabla u + \mu u = f. \tag{3.7}$$

This equation can be written as a system of first-order PDEs in the form

$$\begin{cases} \sigma + \nabla u = 0, \\ \mu u + \nabla\cdot\sigma + \beta\cdot\nabla u = f. \end{cases} \tag{3.8}$$

The above differential operator can be cast into the form of a Friedrichs' operator by setting $K(\sigma,u) = (\sigma,\mu u)$ for all $(\sigma,u) \in [L^2(\Omega)]^{d+1}$, and for $k \in \{1,\dots,d\}$,

$$\mathcal{A}^k = \left[\begin{array}{c:c} 0 & e^k \\ \hdashline (e^k)^t & \beta^k \end{array}\right], \tag{3.9}$$

where e^k is the kth vector in the canonical basis of $\mathbb{R}^d$. It is clear that hypotheses (A1)–(A4) hold with $m = d+1$. Upon observing the norm equivalence

$$\begin{aligned} c_1(\|\nabla u\|_{L^2(\Omega)} + \|\nabla\cdot\sigma\|_{L^2(\Omega)}) \le \|\nabla u\|_{L^2(\Omega)} + \|\beta\cdot\nabla u + \nabla\cdot\sigma\|_{L^2(\Omega)} \\ \le c_2(\|\nabla u\|_{L^2(\Omega)} + \|\nabla\cdot\sigma\|_{L^2(\Omega)}), \end{aligned}$$

it is inferred that the graph space is $W = H(\mathrm{div};\Omega)\times H^1(\Omega)$. Moreover, the boundary operator D is such that for all $(\sigma,u),(\tau,v) \in W$,

$$\langle D(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \sigma\cdot n, v\rangle_{-\frac12,\frac12} + \langle \tau\cdot n, u\rangle_{-\frac12,\frac12} + \int_{\partial\Omega} (\beta\cdot n)uv, \tag{3.10}$$

where $\langle , \rangle_{-\frac12,\frac12}$ denotes the duality pairing between $H^{-\frac12}(\partial\Omega)$ and $H^{\frac12}(\partial\Omega)$. Note that (3.10) makes sense since functions in $H^1(\Omega)$ have traces in $H^{\frac12}(\partial\Omega)$ and vector fields in $H(\mathrm{div};\Omega)$ have normal traces in $H^{-\frac12}(\partial\Omega)$.

3.2.1. Dirichlet boundary conditions. A suitable operator M to weakly enforce Dirichlet boundary conditions is such that for all $(\sigma,u),(\tau,v) \in W$,

$$\langle M(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \sigma\cdot n, v\rangle_{-\frac12,\frac12} - \langle \tau\cdot n, u\rangle_{-\frac12,\frac12}. \tag{3.11}$$

LEMMA 3.3. *Let $M \in \mathcal{L}(W;W')$ be defined in* (3.11)*. Then,*
(i) (M1)–(M2) *hold;*
(ii) $V = V^* = \{(\sigma,u) \in W;\, u|_{\partial\Omega} = 0\}$.

Proof of (i). (M1) clearly holds since $M + M^* = 0$. Let $w = (\sigma,u) \in W$ and write $w = w^+ + w^-$ with $w^+ = (-\frac{1}{2}\beta u, u)$ and $w^- = (\sigma + \frac{1}{2}\beta u, 0)$. By assumption on β, the vector-valued field βu is in $H(\mathrm{div};\Omega)$ if $u \in H^1(\Omega)$; hence, $w^\pm$ are in W. Moreover, a straightforward calculation shows that $w^\pm \in \mathrm{Ker}(D \pm M)$. Hence, (M2) holds.

Proof of (ii). The identity $V = V^*$ results from the fact that $M + M^* = 0$. Moreover, observe that for all $(\sigma,u),(\tau,v) \in W$,

$$\langle (D-M)(\sigma,u),(\tau,v)\rangle_{W',W} = 2\langle \tau{\cdot}n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (\beta{\cdot}n)uv.$$

Let $(\sigma,u) \in \mathrm{Ker}(D-M)$. Let $\gamma \in H^{-\frac{1}{2}}(\partial\Omega)$. There exists $\tau \in H(\mathrm{div};\Omega)$ such that $\tau{\cdot}n = \gamma$ in $H^{-\frac{1}{2}}(\partial\Omega)$. Then, using $(\tau,0)$ in the above equation yields $\langle \gamma, u\rangle_{-\frac{1}{2},\frac{1}{2}} = 0$. Since γ is arbitrary, this implies $u|_{\partial\Omega} = 0$. Hence, $V \subset \{(\sigma,u) \in W;\, u|_{\partial\Omega} = 0\}$. Conversely, let $(\sigma,u) \in W$ be such that $u|_{\partial\Omega} = 0$. Then, the above equation shows that $(\sigma,u) \in \mathrm{Ker}(D-M) = V$. ∎

Remark 3.1. The choice of the operator M to enforce homogeneous Dirichlet boundary conditions is not unique. For instance, one can take $\langle M(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \sigma{\cdot}n, v\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle \tau{\cdot}n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} \varsigma uv$, where ς is a nonnegative real number.

3.2.2. Neumann and Robin boundary conditions. Let $\varrho \in L^\infty(\partial\Omega)$ be such that $2\varrho + \beta{\cdot}n \geq 0$ a.e. on $\partial\Omega$. Neumann and Robin boundary conditions are treated simultaneously, the choice $\varrho = 0$ yielding a Neumann boundary condition (in this case, $\beta{\cdot}n \geq 0$ a.e. on $\partial\Omega$ corresponding to an outflow boundary). A suitable operator M to weakly enforce Neumann or Robin boundary conditions is such that for all $(\sigma,u),(\tau,v) \in W$,

$$\langle M(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \tau{\cdot}n, u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle \sigma{\cdot}n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (2\varrho + \beta{\cdot}n)uv. \tag{3.12}$$

LEMMA 3.4. *Let $M \in \mathcal{L}(W;W')$ be defined in* (3.12)*. Then,*
(i) (M1)–(M2) *hold;*
(ii) $V = \{(\sigma,u) \in W;\, \sigma{\cdot}n = \varrho u|_{\partial\Omega}\}$ *and* $V^* = \{(\sigma,u) \in W;\, \sigma{\cdot}n = -(\varrho + \beta{\cdot}n)u|_{\partial\Omega}\}$.

Proof. (M1) holds since $2\varrho + \beta{\cdot}n \geq 0$ a.e. on $\partial\Omega$. Furthermore, observe that

$$\begin{aligned}
\langle (D-M)(\sigma,u),(\tau,v)\rangle_{W',W} &= 2\langle \sigma{\cdot}n, v\rangle_{-\frac{1}{2},\frac{1}{2}} - 2\textstyle\int_{\partial\Omega} \varrho uv,\\
\langle (D+M)(\sigma,u),(\tau,v)\rangle_{W',W} &= 2\langle \tau{\cdot}n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + 2\textstyle\int_{\partial\Omega} (\varrho + \beta{\cdot}n)uv,\\
\langle (D+M^*)(\sigma,u),(\tau,v)\rangle_{W',W} &= 2\langle \sigma{\cdot}n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + 2\textstyle\int_{\partial\Omega} (\varrho + \beta{\cdot}n)uv.
\end{aligned}$$

Let $w = (\sigma,u) \in W$. Since $\varrho u|_{\partial\Omega} \in H^{-\frac{1}{2}}(\partial\Omega)$, there is $\sigma_0 \in H(\mathrm{div};\Omega)$ such that $\sigma_0{\cdot}n = \varrho u|_{\partial\Omega}$. Then, setting $w^+ = (\sigma - \sigma_0, 0)$ and $w^- = (\sigma_0, u)$, it is easily verified that $w^\pm \in \mathrm{Ker}(D \pm M)$ and, hence, (M2) holds. Finally, proceed as in the proof of Lemma 3.3 to prove (ii). ∎

3.3. Maxwell's equations in the elliptic regime. We close this series of examples by considering a simplified form of Maxwell's equations in $\mathbb{R}^3$ in the elliptic regime, i.e., when displacement currents are negligible. Let σ and μ be two positive functions in $L^\infty(\Omega)$ uniformly bounded away from zero. Consider the following

problem

$$\begin{cases} \mu H + \nabla\times E = f, \\ \sigma E - \nabla\times H = g. \end{cases} \tag{3.13}$$

This problem can be cast into the form of a Friedrichs' system by setting $K(H,E) = (\mu H, \sigma E)$ for all $(H,E) \in [L^2(\Omega)]^3 \times [L^2(\Omega)]^3$ and for $k \in \{1,2,3\}$,

$$\mathcal{A}^k = \left[\begin{array}{c:c} 0 & \mathcal{R}^k \\ \hdashline (\mathcal{R}^k)^t & 0 \end{array}\right]. \tag{3.14}$$

The entries of the matrices $\mathcal{R}^k \in \mathbb{R}^{3,3}$ are those of the Levi–Civita permutation tensor, i.e., $\mathcal{R}^k_{ij} = \epsilon_{ikj}$ for $1 \le i,j,k \le 3$. Hypotheses (A1)–(A4) hold with $m = 6$. The graph space is $W = H(\mathrm{curl};\Omega)\times H(\mathrm{curl};\Omega)$, and the boundary operator D is such that for all $(H,E),(h,e) \in W$,

$$\begin{aligned} \langle D(H,E),(h,e)\rangle_{W',W} = {} & (\nabla\times E, h)_{[L^2(\Omega)]^3} - (E, \nabla\times h)_{[L^2(\Omega)]^3} \\ & + (H, \nabla\times e)_{[L^2(\Omega)]^3} - (\nabla\times H, e)_{[L^2(\Omega)]^3}. \end{aligned} \tag{3.15}$$

When H and E are smooth the above duality product can be interpreted as the boundary integral $\int_{\partial\Omega}(n\times E)\cdot h + (n\times e)\cdot H$.

Let us now define acceptable boundary conditions for (3.13). One possibility (among many others) consists of setting for all $(H,E),(h,e) \in W$,

$$\begin{aligned} \langle M(H,E),(h,e)\rangle_{W',W} = {} & -(\nabla\times E, h)_{[L^2(\Omega)]^3} + (E, \nabla\times h)_{[L^2(\Omega)]^3} \\ & + (H, \nabla\times e)_{[L^2(\Omega)]^3} - (\nabla\times H, e)_{[L^2(\Omega)]^3}. \end{aligned} \tag{3.16}$$

LEMMA 3.5. *Let M be defined in* (3.16). *Then,*

(i) (M1)–(M2) *hold;*

(ii) $V = V^* = \{(H,E) \in W;\ (E\times n)|_{\partial\Omega} = 0\}$.

Proof of (i). Observe that $M + M^* = 0$; hence, M is positive. Let $w = (H,E) \in W$. Write $w = w^+ + w^-$ with $w^+ = (0,E)$ and $w^- = (H,0)$. One easily verifies that $w^\pm \in \mathrm{Ker}(D \pm M)$, i.e., (M2) holds.

Proof of (ii). The identity $V = V^*$ results from the fact that $M + M^* = 0$. Let $(H,E) \in \mathrm{Ker}(D - M)$. Then, for all $(h,e) \in W$,

$$\langle (D-M)(H,E),(h,e)\rangle_{W',W} = 2(\nabla\times E, h)_{[L^2(\Omega)]^3} - 2(E, \nabla\times h)_{[L^2(\Omega)]^3} = 0.$$

Since vector fields in $H(\mathrm{curl};\Omega)$ have tangential traces in $[H^{-\frac12}(\partial\Omega)]^3$, we infer that for all $h \in [H^1(\Omega)]^3$, $\langle (E\times n), h\rangle_{-\frac12,\frac12} = 0$. Since h is arbitrary and the traces of vector fields in $[H^1(\Omega)]^3$ span $[H^{\frac12}(\partial\Omega)]^3$, we conclude that $(E\times n)|_{\partial\Omega} = 0$. Conversely, let $(H,E) \in W$ be such that $(E\times n)|_{\partial\Omega} = 0$. Then, it is clear that $\langle (D-M)(H,E),(h,e)\rangle_{W',W} = 0$ for all $h \in [H^1(\Omega)]^3$ and all $e \in H(\mathrm{curl};\Omega)$. Since $[H^1(\Omega)]^3$ is dense in $H(\mathrm{curl};\Omega)$ and both D and M are in $\mathcal{L}(W;W')$, it follows that $(H,E) \in \mathrm{Ker}(D-M)$. ∎

4. Discontinuous Galerkin. The goal of this section is to introduce a generic DG method to approximate the abstract problem (2.23). The fact that the boundary conditions are enforced weakly through the boundary operator M is a key to the theory. The discrete problem is stated in (4.12)–(4.13). The design constraints of the method are (DG1) to (DG8). The main convergence result is stated in Theorem 4.6.

4.1. The discrete setting. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes of Ω. The meshes are assumed to be affine to avoid unnecessary technicalities, i.e., Ω is assumed to be a polyhedron. However, we do not make any assumption on the matching of element interfaces.

Let p be a nonnegative integer. Define

$$W_h = \{v_h \in [L^2(\Omega)]^m;\, \forall K \in \mathcal{T}_h,\, v_h|_K \in [\mathbb{P}_p]^m\}, \tag{4.1}$$

$$W(h) = [H^1(\Omega)]^m + W_h. \tag{4.2}$$

We denote by $\mathcal{F}_h^{\mathrm{i}}$ the set of interior faces (or interfaces), i.e., $F \in \mathcal{F}_h^{\mathrm{i}}$ if F is a $(d-1)$-manifold and there are $K_1(F)$, $K_2(F) \in \mathcal{T}_h$ such that $F = K_1(F) \cap K_2(F)$. We denote by $\mathcal{F}_h^{\partial}$ the set of the faces that separate the mesh from the exterior of Ω, i.e., $F \in \mathcal{F}_h^{\partial}$ if F is a $(d-1)$-manifold and there is $K(F) \in \mathcal{T}_h$ such that $F = K(F) \cap \partial\Omega$. Finally, we set $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\partial}$. Since every function v in $W(h)$ has a (possibly two-valued) trace almost everywhere on $F \in \mathcal{F}_h^{\mathrm{i}}$, it is meaningful to set

$$v^1(x) = \lim_{\substack{y \to x \\ y \in K_1(F)}} v(y), \qquad v^2(x) = \lim_{\substack{y \to x \\ y \in K_2(F)}} v(y), \qquad \text{for a.e. } x \in F, \tag{4.3}$$

$$[\![v]\!] = v^1 - v^2, \qquad \{v\} = \frac{1}{2}(v^1 + v^2), \qquad \text{a.e. on } F. \tag{4.4}$$

The arbitrariness in the choice of $K_1(F)$ and $K_2(F)$ could be avoided by choosing an intrinsic notation that would, however, unnecessarily complicate the presentation. For instance, we could have chosen to set $[\![v]\!] = v^1 \otimes n^1 + v^2 \otimes n^2$ where n^1, n^2 are the unit outward normals of $K_1(F)$ and $K_2(F)$, respectively. Although having to choose $K_1(F)$ and $K_2(F)$ may seem cumbersome, nothing that is said hereafter depends on the choice that is made.

For any measurable subset of Ω or $\mathcal{F}_h$, say E, $(\cdot,\cdot)_{L,E}$ denotes the scalar product induced by $[L^2(\Omega)]^m$ or $[L^2(\mathcal{F}_h)]^m$ on E, respectively, and $\|\cdot\|_{L,E}$ the associated norm. Similarly, $\|\cdot\|_{L^d,E}$ denotes the norm induced by $[L^2(\Omega)]^{m\times d}$ or $[L^2(\mathcal{F}_h)]^{m\times d}$ on E. For $K \in \mathcal{T}_h$ (resp., $F \in \mathcal{F}_h$), h_K (resp., h_F) denotes the diameter of K (resp., F).

The mesh family $\{\mathcal{T}_h\}_{h>0}$ is assumed to be shape-regular so that there is a constant c, independent of $h = \max_{K\in\mathcal{T}_h} h_K$, such that for all $v_h \in W_h$ and for all $K \in \mathcal{T}_h$,

$$\|\nabla v_h\|_{L^d,K} \le c\, h_K^{-1} \|v_h\|_{L,K}, \tag{4.5}$$

$$\|v_h\|_{L,F} \le c\, h_K^{-\frac{1}{2}} \|v_h\|_{L,K} \quad \forall F \subset \partial K. \tag{4.6}$$

4.2. Boundary operators. Henceforth we denote $\mathcal{D}_{\partial\Omega} = \sum_{k=1}^d n_k \mathcal{A}^k$ and we assume that the boundary operator M is associated with a matrix-valued field $\mathcal{M} : \partial\Omega \longrightarrow \mathbb{R}^{m,m}$. Hence, for all functions u, v smooth enough (e.g., $u, v \in [H^1(\Omega)]^m$), the following holds:

$$\langle Du, v\rangle_{W',W} = \int_{\partial\Omega} v^t \mathcal{D}_{\partial\Omega} u, \qquad \langle Mu, v\rangle_{W',W} = \int_{\partial\Omega} v^t \mathcal{M} u. \tag{4.7}$$

To enforce boundary conditions weakly, we introduce for all $F \in \mathcal{F}_h^{\partial}$ a linear operator $M_F \in \mathcal{L}([L^2(F)]^m; [L^2(F)]^m)$. The design of the boundary operators $\{M_F\}_{F\in\mathcal{F}_h^{\partial}}$

must comply with the following conditions: For all $F \in \mathcal{F}_h^\partial$ and for all $v, w \in [L^2(F)]^m$,

(DG1) $(M_F(v), v)_{L,F} \geq 0,$

(DG2) $(\mathcal{M}v = \mathcal{D}_{\partial\Omega}v) \implies (M_F(v) = \mathcal{D}_{\partial\Omega}v),$

(DG3) $|(M_F(v) - \mathcal{D}_{\partial\Omega}v, w)_{L,F}| \leq c|v|_{M,F}\|w\|_{L,F},$

(DG4) $|(M_F(v) + \mathcal{D}_{\partial\Omega}v, w)_{L,F}| \leq c\|v\|_{L,F}|w|_{M,F},$

where c is a mesh-independent constant and where we have introduced for all $v \in W(h)$ the following seminorms:

$$|v|_M^2 = \sum_{F\in\mathcal{F}_h^\partial} |v|_{M,F}^2 \qquad \text{with} \qquad |v|_{M,F}^2 = (M_F(v), v)_{L,F}. \tag{4.8}$$

Remark 4.1.

(i) Examples of boundary operators M_F are presented in section 5 for all the model problems introduced in section 3.

(ii) Assumption (DG2) is a consistency assumption while assumptions (DG3) and (DG4) are related to the stability and continuity of the discrete bilinear form; see the analysis in section 4.5.

4.3. Interface operators. For $K \in \mathcal{T}_h$, define the matrix-valued field $\mathcal{D}_{\partial K} : \partial K \to \mathbb{R}^{m,m}$ as

$$\mathcal{D}_{\partial K}(x) = \sum_{k=1}^{d} n_{K,k}\mathcal{A}^k(x) \quad \text{a.e. on } \partial K, \tag{4.9}$$

where $n_K = (n_{K,1}, \ldots, n_{K,d})^t$ is the unit outward normal to K on ∂K. Note that this definition is compatible with that of $\mathcal{D}_{\partial\Omega}$ in (4.7) if $\partial K \cap \partial\Omega \neq \emptyset$. Moreover, observe that for all u, v in $W(h)$ and for all $K \in \mathcal{T}_h$,

$$(\mathcal{D}_{\partial K}u, v)_{L,\partial K} = (Tu, v)_{L,K} - (u, \tilde{T}v)_{L,K}. \tag{4.10}$$

We denote by $\mathcal{D}$ the matrix-valued field defined on $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^\partial$ as follows. On $\mathcal{F}_h^\partial$, $\mathcal{D}$ is single-valued and coincides with $\mathcal{D}_{\partial\Omega}$. On $\mathcal{F}_h^{\mathrm{i}}$, $\mathcal{D}$ is two-valued and for all $F \in \mathcal{F}_h^{\mathrm{i}}$, its two values are $\mathcal{D}_{\partial K_1(F)}$ and $\mathcal{D}_{\partial K_2(F)}$. Note that $\{\mathcal{D}\} = 0$ a.e. on $\mathcal{F}_h^{\mathrm{i}}$.

To control the jumps of functions in W_h across mesh interfaces, we introduce for all $F \in \mathcal{F}_h^{\mathrm{i}}$ a linear operator $S_F \in \mathcal{L}([L^2(F)]^m; [L^2(F)]^m)$. The analysis below will show that the design of the interface operators $\{S_F\}_{F\in\mathcal{F}_h^{\mathrm{i}}}$ must comply with the following conditions. For all $F \in \mathcal{F}_h^{\mathrm{i}}$ and for all $v, w \in [L^2(F)]^m$,

(DG5) $(S_F(v), v)_{L,F} \geq 0,$

(DG6) $\|S_F(v)\|_{L,F} \leq c\|v\|_{L,F},$

(DG7) $|(S_F(v), w)_{L,F}| \leq c|v|_{S,F}|w|_{S,F},$

(DG8) $|(\mathcal{D}_{\partial K(F)}v, w)_{L,F}| \leq c|v|_{S,F}\|w\|_{L,F},$

where c is a mesh-independent constant, $K(F)$ denotes any of the two elements sharing F and $\partial K(F)$ its boundary, and where we have introduced for all $v \in W(h)$ the following seminorms:

$$|v|_S^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |v|_{S,F}^2 \qquad \text{with} \qquad |v|_{S,F}^2 = (S_F(v), v)_{L,F}. \tag{4.11}$$

Remark 4.2.

(i) Examples of interface operators S_F are presented in section 5 for all the model problems introduced in section 3.

(ii) Since S_F is positive, a sufficient condition for (DG7) to hold with $c = 1$ is S_F be self-adjoint.

4.4. The discrete problem. We now turn our attention to the construction of a discrete counterpart of (2.23). To this end we introduce the bilinear form a_h such that for all v, w in $W(h)$,

$$\begin{aligned} a_h(v,w) = &\sum_{K\in\mathcal{T}_h} (Tv,w)_{L,K} + \sum_{F\in\mathcal{F}_h^\partial} \frac{1}{2}(M_F(v) - \mathcal{D}v, w)_{L,F} \\ &- \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}v\},\{w\})_{L,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F([\![v]\!]),[\![w]\!])_{L,F}. \end{aligned} \tag{4.12}$$

Then, we construct an approximate solution to (2.23) as follows. For $f \in L$,

$$\begin{cases} \text{seek } u_h \in W_h \text{ such that} \\ a_h(u_h, v_h) = (f, v_h)_L \quad \forall v_h \in W_h. \end{cases} \tag{4.13}$$

Remark 4.3. In the definition of a_h, the second term weakly enforces the boundary conditions. The purpose of the third term is to ensure that a coercivity property holds, see Lemma 4.1. The last term controls the jump of the discrete solution across interfaces. Some user-dependent arbitrariness appears in the second and fourth term through the definition of the operators M_F and S_F. The design constraints on M_F and S_F are (DG1)–(DG4) and (DG5)–(DG8), respectively.

4.5. Convergence analysis. To perform the error analysis we introduce the following discrete norms on $W(h)$,

$$\|v\|_{h,A}^2 = \|v\|_L^2 + |v|_J^2 + |v|_M^2 + \sum_{K\in\mathcal{T}_h} h_K \|Av\|_{L,K}^2, \tag{4.14}$$

$$\|v\|_{h,\frac{1}{2}}^2 = \|v\|_{h,A}^2 + \sum_{K\in\mathcal{T}_h} [h_K^{-1}\|v\|_{L,K}^2 + \|v\|_{L,\partial K}^2], \tag{4.15}$$

where we have introduced the jump seminorms

$$|v|_J^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |v|_{J,F}^2 \qquad \text{with} \qquad |v|_{J,F} = |[\![v]\!]|_{S,F}. \tag{4.16}$$

The norm $\|\cdot\|_{h,A}$ is used to measure the approximation error, and the norm $\|\cdot\|_{h,\frac{1}{2}}$ serves to measure the interpolation properties of the discrete space W_h.

Throughout this section, we assume that:

- Problem (2.23) is well-posed.
- The mesh family $\{\mathcal{T}_h\}_{h>0}$ is shape-regular so that (4.5) and (4.6) hold.
- The design assumptions (DG1)–(DG8) on M_F and S_F hold.
- For all $k \in \{1,\dots,d\}$, $\mathcal{A}^k \in [\mathfrak{C}^{0,\frac{1}{2}}(\overline{\Omega})]^{m,m}$.

LEMMA 4.1 (L-coercivity). *For all h and for all v in $W(h)$,*

$$a_h(v,v) \geq \mu_0 \|v\|_L^2 + |v|_J^2 + \frac{1}{2}|v|_M^2. \tag{4.17}$$

Proof. Let v in $W(h)$. Using (4.10) and summing over the mesh elements yields

$$\sum_{F\in\mathcal{F}_h^\partial}\frac{1}{2}(\mathcal{D}v,v)_{L,F}+\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F\left\{v^t\mathcal{D}v\right\}=\frac{1}{2}\sum_{K\in\mathcal{T}_h}[(Tv,v)_{L,K}-(v,\tilde{T}v)_{L,K}].$$

Subtracting this equation from (4.12) and using the fact that $\{v^t\mathcal{D}v\}=2\{v^t\}\{\mathcal{D}v\}$ leads to

$$a_h(v,v)=\frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big[(Tv,v)_{L,K}+(v,\tilde{T}v)_{L,K}\Big]+|v|_J^2+\frac{1}{2}|v|_M^2.$$

Then, the desired result follows using (A4). ∎

LEMMA 4.2. *There is $c>0$, independent of h, such that for all F in $\mathcal{F}_h^{\mathrm{i}}$ and for all $v,w\in W(h)$,*

$$|(S_F([\![v]\!]),[\![w]\!])_{L,F}|+|(\{\mathcal{D}v\},\{w\})_{L,F}|\le c|v|_{J,F}(\|\{w\}\|_{L,F}+\|[\![w]\!]\|_{L,F}). \tag{4.18}$$

Proof.

(1) Owing to (DG7), $(S_F([\![v]\!]),[\![w]\!])_{L,F}\le c|v|_{J,F}|w|_{J,F}$, and owing to (DG6), $|w|_{J,F}\le c\|[\![w]\!]\|_{L,F}$. Hence, $(S_F([\![v]\!]),[\![w]\!])_{L,F}\le c|v|_{J,F}\|[\![w]\!]\|_{L,F}$.

(2) Let $K_1(F)$ and $K_2(F)$ be the two mesh elements such that $F=K_1(F)\cap K_2(F)$. Then, $2\{\mathcal{D}v\}=\mathcal{D}_{K_1(F)}[\![v]\!]$ since $\{\mathcal{D}\}=0$. Using (DG8) yields

$$|(\{\mathcal{D}v\},\{w\})_{L,F}|=|(\mathcal{D}_{K_1(F)}[\![v]\!],\{w\})_{L,F}|\le c|v|_{J,F}\|\{w\}\|_{L,F}.$$

The proof is complete. ∎

LEMMA 4.3 (stability). *There is $c>0$, independent of h, such that*

$$\inf_{v_h\in W_h\setminus\{0\}}\sup_{w_h\in W_h\setminus\{0\}}\frac{a_h(v_h,w_h)}{\|v_h\|_{h,A}\|w_h\|_{h,A}}\ge c. \tag{4.19}$$

Proof.

(1) Let v_h be an arbitrary element in W_h. Let $K\in\mathcal{T}_h$. Denote by $\overline{\mathcal{A}_K^k}$ the mean-value of $\mathcal{A}^k$ on K; then,

$$\|\mathcal{A}^k-\overline{\mathcal{A}_K^k}\|_{[L^\infty(K)]^{m,m}}\le\|\mathcal{A}^k\|_{[\mathfrak{C}^{0,\frac{1}{2}}(\overline{\Omega})]^{m,m}}h_K^{\frac{1}{2}}. \tag{4.20}$$

Set $\overline{A}_Kv_h=\sum_{k=1}^d\overline{\mathcal{A}_K^k}\partial_kv_h$ and $\pi_h=\sum_{K\in\mathcal{T}_h}h_K\overline{A}_Kv_h$. Clearly, $\pi_h\in W_h$. Using (4.20), together with the inverse inequalities (4.5) and (4.6), leads to

$$\begin{cases}\|\overline{A}_Kv_h\|_{L,F}\le c\,h_K^{-\frac{1}{2}}\|\overline{A}_Kv_h\|_{L,K} & \text{if } F\in\mathcal{F}_h^\partial,\\ \|\left\{\overline{A}_Kv_h\right\}\|_{L,F}+\|[\![\overline{A}_Kv_h]\!]\|_{L,F}\le c\,h_K^{-\frac{1}{2}}\|\overline{A}_Kv_h\|_{L,K_1\cup K_2} & \text{if } F\in\mathcal{F}_h^{\mathrm{i}},\end{cases} \tag{4.21}$$

$$\|\overline{A}_Kv_h\|_{L,K}\le c\,\min(\|Av_h\|_{L,K}+h_K^{-\frac{1}{2}}\|v_h\|_{L,K},h_K^{-1}\|v_h\|_{L,K}). \tag{4.22}$$

Note that (4.22) implies $\|\pi_h\|_L \le c\|v_h\|_L$. From the definition of a_h it follows that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}^2 &= a_h(v_h,\pi_h) - (Kv_h,\pi_h)_L - \sum_{F\in\mathcal{F}_h^\partial}\frac{1}{2}(M_F(v_h)-\mathcal{D}v_h,\pi_h)_{L,F}\\ &\quad+\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\left[2(\{\mathcal{D}v_h\},\{\pi_h\})_{L,F}-(S_F([\![v_h]\!]),[\![\pi_h]\!])_{L,F}\right]\\ &\quad+\sum_{K\in\mathcal{T}_h} h_K(Av_h,(A-\overline{A}_K)v_h)_{L,K}\\ &= a_h(v_h,\pi_h)+R_1+R_2+R_3+R_4,\end{aligned}$$

where R_1, R_2, R_3, and R_4 denote the second, third, fourth, and fifth term in the right-hand side of the above equation, respectively. Each of these terms is bounded from above as follows. Using (4.22) yields $\|\pi_h\|_L \le c\|v_h\|_L$ and hence,

$$|R_1| \le c\|v_h\|_L\|\pi_h\|_L \le c\|v_h\|_L^2.$$

Using (DG3) together with (4.21) and (4.22) leads to

$$\begin{aligned}|R_2| &\le \sum_{F\in\mathcal{F}_h^\partial}\left[c_\gamma(M_F(v_h),v_h)_{L,F}+\gamma\|\pi_h\|_{L,F}^2\right]\\ &\le c(\|v_h\|_L^2+|v_h|_M^2)+\gamma\sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}^2,\end{aligned}$$

where $\gamma>0$ can be chosen as small as needed. For the third term, use Lemma 4.2, together with inequalities (4.21) and (4.22), as follows:

$$\begin{aligned}|R_3| &\le \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} c_\gamma|v_h|_{J,F}^2+\gamma\sum_{K\in\mathcal{T}_h} h_K\|\overline{A}_Kv_h\|_{L,K}^2\\ &\le c(\|v_h\|_L^2+|v_h|_J^2)+\gamma\sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}^2.\end{aligned}$$

For the last term, (4.5) and (4.20) yield

$$\begin{aligned}|R_4| &\le \sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}ch_K^{\frac12}\|\nabla v_h\|_{L^d,K}\\ &\le c\sum_{K\in\mathcal{T}_h} h_K^{\frac12}\|Av_h\|_{L,K}\|v_h\|_{L,K} \le c\|v_h\|_L^2+\gamma\sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}^2.\end{aligned}$$

Using the above four bounds, $\gamma=\frac16$, and Lemma 4.1 leads to

$$\frac12\sum_{K\in\mathcal{T}_h} h_K\|Av_h\|_{L,K}^2 \le a_h(v_h,\pi_h)+c\,a_h(v_h,v_h). \tag{4.23}$$

(2) Let us now prove that $\|\pi_h\|_{h,A} \le c\,\|v_h\|_{h,A}$. We have already seen that $\|\pi_h\|_L \le c\|v_h\|_L$. Using (4.5), together with inequalities (4.20) and (4.22), leads to

$$\sum_{K\in\mathcal{T}_h} h_K\|A\pi_h\|_{L,K}^2 \le c\sum_{K\in\mathcal{T}_h} h_K^{-1}\|\pi_h\|_{L,K}^2 \le c\sum_{K\in\mathcal{T}_h}[h_K\|Av_h\|_{L,K}^2+\|v_h\|_{L,K}^2].$$

Moreover, the inverse inequality (4.6), assumption (DG6), and inequalities (4.21) and (4.22) yield

$$|\pi_h|_J^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |\pi_h|_{J,F}^2 \le c \sum_{K\in\mathcal{T}_h} h_K^{-1}\|\pi_h\|_{L,K}^2 \le c \sum_{K\in\mathcal{T}_h} [h_K\|Av_h\|_{L,K}^2 + \|v_h\|_{L,K}^2].$$

Proceed similarly to control $|\pi_h|_M$. In conclusion,

$$\|\pi_h\|_{h,A} \le c\,\|v_h\|_{h,A}. \tag{4.24}$$

(3) Owing to (4.17) and (4.23), there is $c_1 > 0$ such that

$$\|v_h\|_{h,A}^2 \le c_1 a_h(v_h, v_h) + a_h(v_h, \pi_h) = a_h(v_h, \pi_h + c_1 v_h).$$

Then, setting $w_h = \pi_h + c_1 v_h$ and using (4.24) yields

$$\|v_h\|_{h,A}\|w_h\|_{h,A} \le c\,\|v_h\|_{h,A}^2 \le c\,a_h(v_h, w_h).$$

The conclusion is straightforward. ∎

Remark 4.4. Note that (4.5) and (4.17) readily imply coercivity in the weaker norm $\|v\|_{h,A^-}^2 = \|v\|_L^2 + |v|_J^2 + |v|_M^2 + \sum_{K\in\mathcal{T}_h} h_K^2\|Av\|_{L,K}^2$, but this property is not sufficient to prove an optimal convergence rate in the broken graph norm; see (4.32).

LEMMA 4.4 (continuity). *There is c, independent of h, such that*

$$\forall (v,w) \in W(h)\times W(h), \quad a_h(v,w) \le c\,\|v\|_{h,\frac12}\|w\|_{h,A}. \tag{4.25}$$

Proof. The general principle of the proof is to integrate by parts $a_h(v,w)$ by making use of the formal adjoint $\tilde{T}$. Observing that

$$\sum_{K\in\mathcal{T}_h} [(Tv,w)_{L,K} - (v,\tilde{T}w)_{L,K}] = \sum_{F\in\mathcal{F}_h^{\partial}} (\mathcal{D}v,w)_{L,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F 2\left\{w^t\mathcal{D}v\right\},$$

and $2\{w^t\mathcal{D}v\} = 2\{w^t\}\{\mathcal{D}v\} + \frac12[\![w^t]\!][\![\mathcal{D}v]\!]$, it is clear that

$$\begin{aligned} a_h(v,w) = {} & \sum_{K\in\mathcal{T}_h} (v,\tilde{T}w)_{L,K} + \sum_{F\in\mathcal{F}_h^{\partial}} \frac12 (M_F(v) + \mathcal{D}v, w)_{L,F} \\ & + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \frac12 ([\![\mathcal{D}v]\!], [\![w]\!])_{L,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F([\![v]\!]), [\![w]\!])_{L,F}. \end{aligned} \tag{4.26}$$

Let R_1 to R_4 be the four terms in the right-hand side. Using the Cauchy–Schwarz inequality yields

$$|R_1| \le c \sum_{K\in\mathcal{T}_h} \|v\|_{L,K}(\|w\|_{L,K} + \|Aw\|_{L,K}) \le c\|v\|_{h,\frac12}\|w\|_{h,A}.$$

Use (DG4) together with the Cauchy–Schwarz inequality to infer

$$|R_2| \le c \sum_{F\in\mathcal{F}_h^{\partial}} \|v\|_{L,F}|w|_{M,F} \le c\|v\|_{h,\frac12}\|w\|_{h,A}.$$

For the third and fourth term, use (DG6) and (DG7), together with the fact that $[\![\mathcal{D}v]\!] = 2\mathcal{D}_{\partial K_1(F)}\{v\}$, to obtain

$$|R_3| + |R_4| \le c \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (\|\{v\}\|_{L,F} + \|[\![v]\!]\|_{L,F})|w|_{J,F} \le c\|v\|_{h,\frac12}\|w\|_{h,A}.$$

The result follows easily. ∎

LEMMA 4.5 (consistency). *Let u solve* (2.23) *and let u_h solve* (4.13). *If $u \in [H^1(\Omega)]^m$, then,*

$$\forall v_h \in W_h, \quad a_h(u - u_h, v_h) = 0. \tag{4.27}$$

Proof. Since $u \in [H^1(\Omega)]^m$ solves (2.23), $\mathcal{M}u = \mathcal{D}u$ a.e. on $\partial\Omega$ and $Tu = f$ in L. Assumption (DG2) yields $M_F(u|_F) = \mathcal{D}u|_F$ for all $F \in \mathcal{F}_h^{\partial}$. Moreover, $u \in [H^1(\Omega)]^m$ implies that $\{\mathcal{D}u\} = 0$ and $[\![u]\!] = 0$ a.e. on $\mathcal{F}_h^{\mathrm{i}}$. As a result,

$$\forall v_h \in W_h, \quad a_h(u, v_h) = (Tu, v_h)_L = (f, v_h)_L = a_h(u_h, v_h).$$

The conclusion follows readily. ∎

THEOREM 4.6 (convergence). *Let u solve* (2.23) *and let u_h solve* (4.13). *Assume that $u \in [H^1(\Omega)]^m$. Then, there is c, independent of h, such that*

$$\|u - u_h\|_{h,A} \le c \inf_{v_h\in W_h} \|u - v_h\|_{h,\frac12}. \tag{4.28}$$

Proof. Simple application of Strang's Second Lemma; see, e.g., [15, p. 94]. Let $v_h \in W_h$. Owing to Lemmas 4.3, 4.4, and 4.5,

$$\begin{aligned}\|v_h - u_h\|_{h,A} &\le c \sup_{w_h\in W_h\setminus\{0\}} \frac{a_h(v_h - u_h, w_h)}{\|w_h\|_{h,A}} \\ &\le c \sup_{w_h\in W_h\setminus\{0\}} \frac{a_h(v_h - u, w_h)}{\|w_h\|_{h,A}} \le c\,\|u - v_h\|_{h,\frac12}.\end{aligned}$$

Conclude using the triangle inequality. ∎

Owing to the definition of W_h, and the regularity of the mesh family $\{\mathcal{T}_h\}_{h>0}$, the following interpolation property holds. There is c, independent of h, such that for all $v \in [H^{p+1}(\Omega)]^m$, there is $v_h \in W_h$ satisfying

$$\|v - v_h\|_{h,\frac12} \le c h^{p+\frac12}\|v\|_{[H^{p+1}(\Omega)]^m}. \tag{4.29}$$

COROLLARY 4.7. *If u is in $[H^{p+1}(\Omega)]^m$, there is c, independent of h, such that*

$$\|u - u_h\|_{h,A} \le c\,h^{p+\frac12}\|u\|_{[H^{p+1}(\Omega)]^m}. \tag{4.30}$$

In particular,

$$\|u - u_h\|_L \le c\,h^{p+\frac12}\|u\|_{[H^{p+1}(\Omega)]^m}, \tag{4.31}$$

and if the mesh family $\{\mathcal{T}_h\}_{h>0}$ is quasi-uniform,

$$\left(\sum_{K\in\mathcal{T}_h} \|A(u - u_h)\|_{L,K}^2\right)^{\frac12} \le c\,h^p\|u\|_{[H^{p+1}(\Omega)]^m}. \tag{4.32}$$

The above estimates show that, provided the exact solution is smooth enough, the method yields optimal order convergence in the broken graph norm and $(p+\frac12)$-order convergence in the L-norm.

Remark 4.5. The estimates (4.30) to (4.32) are identical to those that can be obtained by other stabilization methods like GaLS [5, 19, 21] or subgrid viscosity [18] and many other methods.

Finally, when the exact solution is not smooth enough to be in $[H^1(\Omega)]^m$ but only in the graph space W, we use a density argument to infer the convergence of the DG approximation in the L-norm.

COROLLARY 4.8. *Let u solve* (2.23) *and let u_h solve* (4.13). *Assume that $[H^1(\Omega)]^m \cap V$ is dense in V. Then,*

$$\lim_{h\to 0} \|u - u_h\|_L = 0. \tag{4.33}$$

Proof. Let $\epsilon > 0$. There is $u_\epsilon \in [H^1(\Omega)]^m \cap V$ such that $\|u - u_\epsilon\|_W \le \frac{\epsilon}{2}$. Let $u_{\epsilon h}$ be the unique solution in W_h such that $a_h(u_{\epsilon h}, v_h) = (Tu_\epsilon, v_h)_L$ for all $v_h \in W_h$. From the regularity of u_ϵ together with Theorem 4.6 and Corollary 4.7 applied with $p = 0$, it is inferred that $\lim_{h\to 0} \|u_{\epsilon h} - u_\epsilon\|_{h,A} = 0$. Furthermore, using the discrete inf-sup condition (4.19) yields

$$\begin{aligned}\|u_{\epsilon h} - u_h\|_L &\le \sup_{v_h\in W_h\setminus\{0\}} \frac{a_h(u_{\epsilon h}, v_h) - a_h(u_h, v_h)}{\|v_h\|_{h,A}} = \sup_{v_h\in W_h\setminus\{0\}} \frac{(T(u_\epsilon - u), v_h)_L}{\|v_h\|_{h,A}} \\ &\le \|T(u_\epsilon - u)\|_L \sup_{v_h\in W_h\setminus\{0\}} \frac{\|v_h\|_L}{\|v_h\|_{h,A}} \le \|u - u_\epsilon\|_W \le \frac{\epsilon}{2},\end{aligned}$$

where we have used the fact that for all $v_h \in W_h$, $a_h(u_h, v_h) = (Tu, v_h)_L$. Finally, using the triangle inequality

$$\|u - u_h\|_L \le \|u - u_\epsilon\|_L + \|u_\epsilon - u_{\epsilon h}\|_L + \|u_{\epsilon h} - u_h\|_L,$$

it is deduced that $\limsup_{h\to 0} \|u - u_h\|_L \le \epsilon$. ∎

4.6. Localization, fluxes, and adjoint-fluxes. The purpose of this section is to discuss briefly some equivalent formulations of the discrete problem (4.13) in order to emphasize the link with other formalisms derived previously for DG methods, namely that of Lesaint and Raviart [23, 24] and Johnson et al. [21, 22] for Friedrichs' systems. To this end, we rewrite the bilinear form (4.12) in various equivalent ways and introduce the concept of element fluxes and that of element adjoint-fluxes.

Let $K \in \mathcal{T}_h$. Define the operator $M^{\mathrm{L}}_{\partial K} \in \mathcal{L}([L^2(\partial K)]^m; [L^2(\partial K)]^m)$ as follows. For $v \in [L^2(\partial K)]^m$ and a face $F \subset \partial K$, set

$$M^{\mathrm{L}}_{\partial K}(v)|_F = \begin{cases} M_F(v|_F) & \text{if } F \in \mathcal{F}_h^\partial, \\ 2S_F(v|_F) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}. \end{cases} \tag{4.34}$$

Furthermore, for $v \in W(h)$ and $x \in \partial K$, set

$$v^i(x) = \lim_{\substack{y\to x \\ y\in K}} v(y), \qquad v^e(x) = \lim_{\substack{y\to x \\ y\notin K}} v(y), \tag{4.35}$$

$$[\![v]\!]_{\partial K}(x) = v^i(x) - v^e(x), \qquad \{v\}_{\partial K}(x) = \frac{1}{2}(v^i(x) + v^e(x)), \tag{4.36}$$

with $v^e(x) = 0$ if $x \in \partial\Omega$. Then, a straightforward calculation shows that the bilinear

form a_h defined in (4.12) can be rewritten as follows:

$$a_h(u,v) = \sum_{K\in\mathcal{T}_h} (Tu,v)_{L,K} + \sum_{K\in\mathcal{T}_h} \frac{1}{2}(M^{\mathrm{L}}_{\partial K}([\![u]\!]_{\partial K}) - \mathcal{D}_{\partial K}[\![u]\!]_{\partial K}, v^i)_{L,\partial K} \tag{4.37}$$

$$= \sum_{K\in\mathcal{T}_h} (u,\tilde{T}v)_{L,K} + \sum_{K\in\mathcal{T}_h} \frac{1}{2}(M^{\mathrm{L}}_{\partial K}([\![u]\!]_{\partial K}) + 2\mathcal{D}_{\partial K}\{u\}_{\partial K}, v^i)_{L,\partial K}. \tag{4.38}$$

The bilinear form (4.37) is that analyzed by Lesaint and Raviart [24, 23] and further investigated by Johnson et al. [21] in the particular case where the operator $M^{\mathrm{L}}_{\partial K}$ is defined pointwise using a matrix-valued field on ∂K; see section 5.1 for further discussion.

DEFINITION 4.9. *Let $K \in \mathcal{T}_h$ and let $v \in W(h)$. The* element flux *of v on ∂K, say $\phi_{\partial K}(v) \in [L^2(\partial K)]^m$, is defined on a face $F \subset \partial K$ by*

$$\phi_{\partial K}(v)|_F = \begin{cases} \frac{1}{2}M_F(v|_F) + \frac{1}{2}\mathcal{D}_{\partial\Omega}v & \text{if } F \subset \partial K^{\partial}, \\ S_F([\![v]\!]_{\partial K}|_F) + \mathcal{D}_{\partial K}\{v\}_{\partial K} & \text{if } F \subset \partial K^{\mathrm{i}}, \end{cases} \tag{4.39}$$

where ∂K^{i} denotes that part of ∂K that lies in Ω and ∂K^{∂} denotes that part of ∂K that lies on $\partial\Omega$. Likewise, the element adjoint-flux *of v on ∂K, say, $\tilde{\phi}_{\partial K}(v) \in [L^2(\partial K)]^m$, is defined on a face $F \subset \partial K$ by*

$$\tilde{\phi}_{\partial K}(v)|_F = \begin{cases} \frac{1}{2}M_F(v|_F) - \frac{1}{2}\mathcal{D}_{\partial\Omega}v & \text{if } F \subset \partial K^{\partial}, \\ S_F([\![v]\!]_{\partial K}|_F) - \frac{1}{2}\mathcal{D}_{\partial K}[\![v]\!]_{\partial K} & \text{if } F \subset \partial K^{\mathrm{i}}. \end{cases} \tag{4.40}$$

The relevance of the notion of flux and adjoint-flux is clarified by the following proposition.

PROPOSITION 4.10. *The discrete problem* (4.13) *is equivalent to each of the following two local formulations.*

$$\begin{cases} \text{Seek } u_h \in W_h \text{ such that } \forall K \in \mathcal{T}_h \text{ and } \forall v_h \in [\mathbb{P}_p(K)]^m, \\ (u_h, \tilde{T}v_h)_{L,K} + (\phi_{\partial K}(u_h), v_h)_{L,\partial K} = (f, v_h)_{L,K}. \end{cases} \tag{4.41}$$

$$\begin{cases} \text{Seek } u_h \in W_h \text{ such that } \forall K \in \mathcal{T}_h \text{ and } \forall v_h \in [\mathbb{P}_p(K)]^m, \\ (Tu_h, v_h)_{L,K} + (\tilde{\phi}_{\partial K}(u_h), v_h)_{L,\partial K} = (f, v_h)_{L,K}. \end{cases} \tag{4.42}$$

Proof. Localize the test functions in (4.13) to the mesh elements and use the fact that $\phi_{\partial K}(v)|_F = \frac{1}{2}M^{\mathrm{L}}_{\partial K}([\![v]\!]_{\partial K}) + \mathcal{D}_{\partial K}\{v\}_{\partial K}$ and $\tilde{\phi}_{\partial K}(v)|_F = \frac{1}{2}M^{\mathrm{L}}_{\partial K}([\![v]\!]_{\partial K}) - \frac{1}{2}\mathcal{D}_{\partial K}[\![v]\!]_{\partial K}$. ∎

Let v be a function in $W(h)$. We define the *interface fluxes* (resp., *interface adjoint-fluxes*) of v, say, $\phi^{\mathrm{i}}(v)$, (resp., say, $\tilde{\phi}^{\mathrm{i}}(v)$), to be the two-valued function defined on $\mathcal{F}^{\mathrm{i}}_h$ that collects all the element fluxes (resp., adjoint-fluxes) of v on the interior faces. Likewise we define the *boundary fluxes* (resp., *boundary adjoint-fluxes*) of v, say, $\phi^{\partial}(v)$, (resp., say, $\tilde{\phi}^{\partial}(v)$), to be the single-valued function defined on $\mathcal{F}^{\partial}_h$ that collects all the element fluxes (resp., adjoint-fluxes) of v on the boundary faces.

Remark 4.6.

(i) The link between DG methods and the concept of element fluxes has been explored recently by Arnold et al. [1] for the Poisson equation (in [1], the terminology "numerical fluxes" is employed instead).

(ii) In engineering practice, approximation schemes such as (4.41) are often designed by a priori specifying the element fluxes. The above analysis then provides a practical means to assess the stability and convergence properties of the scheme. Indeed, once the element fluxes are given, the boundary operators M_F and the interface operators S_F can be directly retrieved from (4.39). Then, properties (DG1)–(DG8) provide sufficient conditions to analyze the scheme.

(iii) The interface fluxes are such that $\{\phi^{\mathrm{i}}(v)\} = 0$ a.e. on $\mathcal{F}_h^{\mathrm{i}}$. Approximation schemes where the interface fluxes satisfy this property are often termed *conservative*. Note that the concept of conservativity as such does not play any role in the present analysis of the method, although it can play a role when deriving improved L^2-error estimates by using the Aubin–Nitsche lemma; see, e.g., Arnold et al. [1] and the second part of this work [16].

(iv) The following relation links the element fluxes and the element adjoint-fluxes

$$\phi_{\partial K}(v) - \tilde{\phi}_{\partial K}(v) = \mathcal{D}_{\partial K} v^i. \tag{4.43}$$

In particular, the element adjoint-fluxes are not conservative.

(v) Both the element fluxes and the element adjoint-fluxes are associated with the operator T, i.e., they are derived from a DG discretization of (2.23). It is also possible to design a DG discretization of the adjoint problem (2.24) involving the operator $\tilde{T}$ and the bilinear form a^*. This would lead to two new families of fluxes, the element fluxes for $\tilde{T}$ and the element adjoint-fluxes for $\tilde{T}$. It should be noted that the element adjoint-fluxes for T are not the element fluxes for $\tilde{T}$. In particular, the former are not conservative whereas the latter are conservative.

5. Applications. This section shows how the conditions (DG1)–(DG8) can be used to design DG approximations of the model problems introduced in section 3.

5.1. Pointwise boundary and interface operators. For ease of presentation, the boundary and interface operators discussed in this section are constructed from matrix-valued fields defined on all the mesh faces. This simpler construction is sufficient to recover several DG methods considered in the literature. Examples where a more general form for the boundary and interface operators is needed will be presented in a forthcoming work [16].

For all $F \in \mathcal{F}_h^{\partial}$, let $\mathcal{M}_F$ be a matrix-valued field such that for all $\xi, \zeta \in \mathbb{R}^m$,

(DG1a) $\mathcal{M}_F$ is positive,

(DG2a) $\mathrm{Ker}(\mathcal{M} - \mathcal{D}_{\partial\Omega}) \subset \mathrm{Ker}(\mathcal{M}_F - \mathcal{D}_{\partial\Omega})$,

(DG3a) $|\zeta^t(\mathcal{M}_F - \mathcal{D}_{\partial\Omega})\xi| \le c(\xi^t \mathcal{M}_F \xi)^{\frac{1}{2}} \|\zeta\|_{\mathbb{R}^m}$,

(DG4a) $|\zeta^t(\mathcal{M}_F + \mathcal{D}_{\partial\Omega})\xi| \le c(\zeta^t \mathcal{M}_F \zeta)^{\frac{1}{2}} \|\xi\|_{\mathbb{R}^m}$,

where $\|\cdot\|_{\mathbb{R}^m}$ denotes the Euclidean norm in $\mathbb{R}^m$. Similarly, for all $F \in \mathcal{F}_h^{\mathrm{i}}$, let $\mathcal{S}_F$ be a matrix-valued field such that for all $\xi, \zeta \in \mathbb{R}^m$,

(DG5a) $\mathcal{S}_F$ is positive,

(DG6a) $\mathcal{S}_F$ is uniformly bounded,

(DG7a) $|\zeta^t \mathcal{S}_F \xi| \le c(\xi^t \mathcal{S}_F \xi)^{\frac{1}{2}} (\zeta^t \mathcal{S}_F \zeta)^{\frac{1}{2}}$,

(DG8a) $|\zeta^t \mathcal{D} \xi| \le c(\xi^t \mathcal{S}_F \xi)^{\frac{1}{2}} \|\zeta\|_{\mathbb{R}^m}$.

A straightforward verification yields the following proposition.

PROPOSITION 5.1. *For all* $F \in \mathcal{F}_h^\partial$, *define* $M_F : [L^2(F)]^m \ni v \mapsto \mathcal{M}_F|_F v \in [L^2(F)]^m$, *and for all* $F \in \mathcal{F}_h^{\mathrm{i}}$, *define* $S_F : [L^2(F)]^m \ni v \mapsto \mathcal{S}_F|_F v \in [L^2(F)]^m$. *Then,* (DG1)–(DG8) *hold.*

Remark 5.1.

(i) Whenever the matrix-valued field $\mathcal{M}$ defined in (4.7) satisfies (DG3a)–(DG4a), one simply sets $\mathcal{M}_F = \mathcal{M}$; otherwise, it is necessary to strengthen $\mathcal{M}$. This last situation occurs, for instance, when approximating advection-diffusion-reaction problems and the Maxwell equations in the elliptic regime; see sections 5.3 and 5.4.

(ii) One possible way of constructing $\mathcal{S}_F$ follows. Since $\mathcal{D}$ is symmetric, $\mathcal{D}$ is diagonalizable; hence, the absolute value of $\mathcal{D}$, say, $|\mathcal{D}|$, can be defined. Moreover, observing that $|\mathcal{D}|$ is single-valued on $\mathcal{F}_h^{\mathrm{i}}$, one can set $\mathcal{S}_F = |\mathcal{D}|$.

5.2. Advection-reaction. Consider the advection-reaction problem introduced in section 3.1. Assume that all the faces in $\mathcal{F}_h^\partial$ are in $\partial\Omega^-$, in $\partial\Omega^+$, or in $\partial\Omega \setminus (\partial\Omega^- \cup \partial\Omega^+)$. The integral representation (4.7) holds with

$$\mathcal{D}_{\partial\Omega} = \beta{\cdot}n \qquad \text{and} \qquad \mathcal{M} = |\beta{\cdot}n|. \tag{5.1}$$

Let $\alpha > 0$ (this design parameter can vary from face to face) and for all $F \in \mathcal{F}_h$, set

$$\mathcal{M}_F = \mathcal{M} = |\beta{\cdot}n| \qquad \text{and} \qquad \mathcal{S}_F = \alpha|\beta{\cdot}n_F|, \tag{5.2}$$

where n_F is a unit normal vector to F (its orientation is clearly irrelevant). It is straightforward to verify the following proposition.

PROPOSITION 5.2. *Properties* (DG1a)–(DG8a) *hold.*

Remark 5.2. The specific value $\alpha = \frac{1}{2}$ has received considerable attention in the literature. When working with the local formulation (4.42), the interface and boundary fluxes are given by

$$\begin{aligned}\tilde{\phi}^{\mathrm{i}}(u_h)|_{\partial K} &= \left(\alpha|\beta{\cdot}n_K| - \frac{1}{2}\beta{\cdot}n_K\right)[\![u_h]\!]_{\partial K},\\ \tilde{\phi}^{\partial}(u_h) &= -|\beta{\cdot}n|u_h 1_{\partial\Omega^-},\end{aligned}$$

where $1_{\partial\Omega^-}$ denotes the characteristic function of $\partial\Omega^-$. Setting $\alpha = \frac{1}{2}$, one obtains the DG method analyzed by Lesaint and Raviart [24, 23]; in this case the interface adjoint-flux $\tilde{\phi}^{\mathrm{i}}$ is nonzero only on that part of the boundary ∂K where $\beta{\cdot}n_K < 0$. Similarly, when working with the local formulation (4.41), the interface and boundary fluxes are given by

$$\begin{aligned}\phi^{\mathrm{i}}(u_h)|_{\partial K} &= (\beta{\cdot}n_K)\{u_h\} + \alpha|\beta{\cdot}n_K|[\![u_h]\!]_{\partial K},\\ \phi^{\partial}(u_h) &= |\beta{\cdot}n|u_h 1_{\partial\Omega^+},\end{aligned}$$

where $1_{\partial\Omega^+}$ denotes the characteristic function of $\partial\Omega^+$. Setting $\alpha = \frac{1}{2}$ leads to $\phi^{\mathrm{i}}(u_h)|_{\partial K} = (\beta{\cdot}n_K)u_h^\uparrow$, where $u_h^\uparrow = u_h^i$ if $\beta{\cdot}n_K > 0$ and $u_h^\uparrow = u_h^e$ otherwise, i.e., the well-known upwind flux is recovered as a particular case of the above analysis which is valid for any $\alpha > 0$. This coincidence has led many authors to believe that DG methods are methods of choice to solve hyperbolic problems. Actually DG methods, as presented herein, are tailored to solve symmetric positive systems of first-order PDEs, and as pointed out by Friedrichs, the notion of symmetric systems goes beyond the hyperbolic/elliptic traditional classification of PDEs. Moreover, the presence of the user-dependent interface operator S_F (see (DG5)–(DG8)) points to

the fact that DG methods are merely stabilization techniques. This fact is even clearer when one realizes that the error estimates (4.30)–(4.32) are identical to those that can be obtained by using other stabilization techniques like GaLS (also sometimes called streamline diffusion) [5, 19, 21] or subgrid viscosity [18] methods.

5.3. Advection-diffusion-reaction. Consider the advection-diffusion-reaction problem introduced in section 3.2. The integral representation (4.7) for D holds with

$$\mathcal{D}_{\partial\Omega} = \left[\begin{array}{c:c} 0 & n \\ \hdashline n^t & \beta\cdot n \end{array}\right]. \tag{5.3}$$

To simplify, we assume that the parameters β and μ are of order 1, i.e., we hide the dependency on these coefficients in the constants. Special cases such as advection-dominated problems go beyond the scope of the present work. We begin with the interface operator since its design is independent of the boundary conditions imposed. Let $\alpha > 0$, $\eta > 0$, and $\delta \in \mathbb{R}^d$. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, define

$$\mathcal{S}_F = \left[\begin{array}{c:c} \alpha n_F\otimes n_F & (\delta\cdot n_F)n_F \\ \hdashline -(\delta\cdot n_F)n_F^t & \eta \end{array}\right]. \tag{5.4}$$

PROPOSITION 5.3. *Properties* (DG5a)–(DG8a) *hold.*

Proof. For $\xi \in \mathbb{R}^{d+1}$, denote by $\xi = (\xi_\sigma, \xi_u)$ its decomposition in $\mathbb{R}^d \times \mathbb{R}$ and use a similar notation for $\zeta = (\zeta_\sigma, \zeta_u) \in \mathbb{R}^{d+1}$. The field $\mathcal{S}_F$ is clearly positive and bounded, i.e., (DG5a) and (DG6a) hold. Moreover, for ξ, $\zeta \in \mathbb{R}^{d+1}$,

$$\zeta^t\mathcal{S}_F\xi = \alpha(\xi_\sigma\cdot n)(\zeta_\sigma\cdot n) + (\delta\cdot n)(\zeta_\sigma\cdot n)\xi_u - (\delta\cdot n)(\xi_\sigma\cdot n)\zeta_u + \eta\xi_u\zeta_u,$$

and $\xi^t\mathcal{S}_F\xi = \alpha(\xi_\sigma\cdot n)^2 + \eta\xi_u^2$, when (DG7a) is readily deduced. Finally, since

$$\zeta^t\mathcal{D}_{\partial K}\xi = (\xi_\sigma\cdot n_K)\zeta_u + (\zeta_\sigma\cdot n_K)\xi_u + (\beta\cdot n_K)\xi_u\zeta_u,$$

(DG8a) holds. ∎

Remark 5.3.

(i) We stress the fact that the above DG method yields $(p+\frac{1}{2})$-order estimates in the L-norm for both u_h and σ_h.

(ii) The σ- and u-component of the interface fluxes are given by

$$\begin{aligned}\phi^{\sigma,\mathrm{i}}(\sigma_h,u_h)|_{\partial K} &= (\{u_h\} + \alpha n_K\cdot[\![\sigma_h]\!]_{\partial K} + (\delta\cdot n_K)[\![u_h]\!]_{\partial K})n_K,\\ \phi^{u,\mathrm{i}}(\sigma_h,u_h)|_{\partial K} &= n_K\cdot\{\sigma_h\} - (\delta\cdot n_K)n_K\cdot[\![\sigma_h]\!]_{\partial K} + \eta[\![u_h]\!]_{\partial K} + \beta\cdot n_K\{u_h\}.\end{aligned}$$

Owing to the fact that $\alpha \neq 0$, the local formulation (4.41) or (4.42) cannot be used to derive a local reconstruction formula where $\sigma_h|_K$ is expressed solely in terms of u_h. To this end, the coefficient α has to be set to zero, and this requires a nontrivial modification of the analysis that will be reported in [16]. With this modification, the DG approximation does no longer yield a $(p+\frac{1}{2})$-order estimate for σ_h in the L-norm.

(iii) The design parameters α, δ, and η can vary from face to face. In particular, one can take δ to be any bounded vector-valued field on $\mathcal{F}_h^{\mathrm{i}}$; $\delta = 0$ is a suitable choice. Other particular choices lead to DG methods already reported in the literature for advection-diffusion-reaction problems. A more detailed discussion, including a comparison with methods where the unknown $\sigma_h|_K$ is eliminated locally, is postponed to [16].

5.3.1. Dirichlet boundary conditions. The integral representation (4.7) of the boundary operator M defined in (3.11) holds with

$$\mathcal{M} = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^t & 0 \end{array}\right]. \tag{5.5}$$

Let $\varsigma > 0$ (this design parameter can vary from face to face). For all $F \in \mathcal{F}_h^\partial$, define

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^t & \varsigma \end{array}\right]. \tag{5.6}$$

It is straightforward to verify the following proposition.

PROPOSITION 5.4. *Properties* (DG1a)–(DG4a) *hold.*

Remark 5.4. Observe that setting $\mathcal{M}_F = \mathcal{M}$ is not adequate here since with this choice (DG3a) does not hold.

5.3.2. Neumann and Robin boundary conditions. The integral representation (4.7) of the boundary operator M defined in (3.12) holds with

$$\mathcal{M} = \left[\begin{array}{c:c} 0 & n \\ \hdashline -n^t & 2\varrho + \beta{\cdot}n \end{array}\right]. \tag{5.7}$$

Consider first Neumann boundary conditions, i.e., $\varrho = 0$. Let $\lambda > 0$ (this design parameter can vary from face to face). For all $F \in \mathcal{F}_h^\partial$, define

$$\mathcal{M}_F = \left[\begin{array}{c:c} \lambda n{\otimes}n & n \\ \hdashline -n^t & 0 \end{array}\right]. \tag{5.8}$$

It is straightforward to verify the following proposition.

PROPOSITION 5.5. *Properties* (DG1a)–(DG4a) *hold.*

Consider next Robin boundary conditions and assume that $\varrho + \min(\beta{\cdot}n, 0) \geq 0$ (this assumption is not restrictive since Robin boundary conditions are often enforced on inflow boundaries by setting $\varrho = -\beta{\cdot}n$). Let $\lambda \in]0, \frac{1}{\rho}[$ ($\lambda \in]0, +\infty)$ if $\varrho = 0$), $\theta = 1 - \lambda\varrho$, and $\alpha = -\lambda\varrho^2$. For all $F \in \mathcal{F}_h^\partial$, define

$$\mathcal{M}_F = \left[\begin{array}{c:c} \lambda n{\otimes}n & \theta n \\ \hdashline -\theta n^t & 2\varrho + \beta{\cdot}n + \alpha \end{array}\right]. \tag{5.9}$$

PROPOSITION 5.6. *Properties* (DG1a)–(DG4a) *hold.*

Proof. Since $\varrho + \beta{\cdot}n \geq 0$ by assumption and since $\varrho + \alpha > 0$ by construction, it is inferred that for all $\xi \in \mathbb{R}^{d+1}$, $\xi^t \mathcal{M}_F \xi \geq c((\xi_\sigma{\cdot}n)^2 + \xi_u^2)$ with $c > 0$. The rest of the proof is straightforward. ∎

Remark 5.5. The bilinear forms $(u, v) \longmapsto \int_{\partial\Omega} v^t \mathcal{M}_F u$ considered above cannot be extended to $W \times W$ due to the presence of the upper-left block in $\mathcal{M}_F$. The difficulty stems from the fact that vector fields in $H(\mathrm{div}; \Omega)$ may not have normal traces in $L^2(\partial\Omega)$. As a consequence, the approximate method is meaningful only if the exact solution is smooth enough; see the definition of $W(h)$ in (4.2).

5.4. Maxwell's equations in the elliptic regime. We close this series of applications with Maxwell's equations in the elliptic regime; see section 3.3. The integral representation (4.7) holds with the $\mathbb{R}^{6,6}$-valued fields

$$\mathcal{D}_{\partial\Omega} = \left[\begin{array}{c:c} 0 & \mathcal{N} \\ \hdashline \mathcal{N}^t & 0 \end{array}\right] \qquad \text{and} \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & -\mathcal{N} \\ \hdashline \mathcal{N}^t & 0 \end{array}\right], \tag{5.10}$$

where $\mathcal{N} = \sum_{k=1}^{3} n_k \mathcal{R}^k \in \mathbb{R}^{3,3}$ and $n = (n_1, n_2, n_3)^t$ is the unit outward normal to Ω on $\partial\Omega$. Observe that $\mathcal{N}\xi = n \times \xi$ for all $\xi \in \mathbb{R}^3$.

Let $\varsigma > 0$, $\alpha_1 > 0$, and $\alpha_2 > 0$ (these design parameters can vary from face to face) and set

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -\mathcal{N} \\ \hdashline \mathcal{N}^t & \varsigma \mathcal{N}^t \mathcal{N} \end{array}\right] \qquad \text{and} \qquad \mathcal{S}_F = \left[\begin{array}{c:c} \alpha_1 \mathcal{N}_F^t \mathcal{N}_F & 0 \\ \hdashline 0 & \alpha_2 \mathcal{N}_F^t \mathcal{N}_F \end{array}\right], \tag{5.11}$$

where $\mathcal{N}_F$ is defined as $\mathcal{N}$ by replacing n by n_F. It is straightforward to verify the following proposition.

PROPOSITION 5.7. *Properties* (DG1a)–(DG8a) *hold.*

REFERENCES

[1] D. ARNOLD, F. BREZZI, B. COCKBURN, AND L. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1749–1779.

[2] D. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] I. BABUŠKA AND M. ZLÁMAL, *Nonconforming elements in the finite element method with penalty*, SIAM J. Numer. Anal., 10 (1973), pp. 863–875.

[4] I. BABUŠKA, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.

[5] C. BAIOCCHI, F. BREZZI, AND L. FRANCA, *Virtual bubbles and Galerkin-Least-Squares type methods (GaLS)*, Comput. Methods Appl. Mech. Eng., 105 (1993), pp. 125–141.

[6] G. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[7] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[8] C. BAUMANN AND J. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[9] B. COCKBURN, G. KARNIADAKIS, AND C. SHU, *Discontinuous Galerkin Methods. Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000.

[10] B. COCKBURN AND C. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[11] C. DAWSON, *Godunov-mixed methods for advection-diffusion equations in multidimensions*, SIAM J. Numer. Anal., 30 (1993), pp. 1315–1332.

[12] C. DAWSON, *Analysis of an upwind-mixed finite element method for nonlinear contaminant transport equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1709–1724.

[13] J. DOUGLAS JR. AND T. DUPONT, *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, Lecture Notes in Phys. 58, Springer, Berlin, 1976.

[14] A. ERN, J.-L. GUERMOND, AND G. CAPLAIN, *An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems*, Comm. Partial Differential Equations, in press, internal report available online from http://cermics.enpc.fr/reports/CERMICS-2005.

[15] A. ERN AND J.-L. GUERMOND, *Theory and Practice of Finite Elements*, Appl. Math. Sci., Springer, New York, 2004.

[16] A. ERN AND J.-L. GUERMOND, *Discontinuous Galerkin methods for Friedrichs' systems.* II. *Second-order PDEs*, SIAM J. Numer. Anal., submitted.

[17] K. FRIEDRICHS, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[18] J.-L. GUERMOND, *Subgrid stabilization of Galerkin approximations of linear monotone operators*, IMA J. Numer. Anal., 21 (2001), pp. 165–197.

[19] T. HUGHES, L. FRANCA, AND G. HULBERT, *A new finite element formulation for computational fluid dynamics:* VIII. *The Galerkin/least-squares method for advection-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[20] M. JENSEN, *Discontinuous Galerkin Methods for Friedrichs Systems with Irregular Solutions*, Ph.D. thesis, University of Oxford, 2004.

[21] C. JOHNSON, U. NÄVERT, AND J. PITKÄRANTA, *Finite element methods for linear hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[22] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[23] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Element Methods in Partial Differential Equations, C. A. deBoor, ed., Academic Press, New York, 1974, pp. 89–123.

[24] P. Lesaint, *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*, Ph.D. thesis, University of Paris VI, 1975.

[25] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.

[26] J. Oden, I. Babuška, and C. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.

[27] J. Rauch, *Symmetric positive systems with boundary characteristic of constant multiplicity*, Trans. Amer. Math. Soc., 291 (1985), pp. 167–187.

[28] W. Reed and T. Hill, *Triangular mesh methods for the neutron transport equation*, Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[29] B. Rivière, M. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems.* I, Comput. Geosci., 8 (1999), pp. 337–360.

[30] E. Süli, C. Schwab, and P. Houston, *hp-DGFEM for partial differential equations with non-negative characteristic form*, in Discontinuous Galerkin Methods. Theory, Computation, and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer, Berlin, 2000, pp. 221–230.

[31] M. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 779–803

MULTISTEP METHODS FOR SDES AND THEIR APPLICATION TO PROBLEMS WITH SMALL NOISE*

EVELYN BUCKWAR† AND RENATE WINKLER†

Abstract. In this article the numerical approximation of solutions of Itô stochastic differential equations is considered, in particular for equations with a small parameter ϵ in the noise coefficient. We construct stochastic linear multistep methods and develop the fundamental numerical analysis concerning their mean-square consistency, numerical stability in the mean-square sense and mean-square convergence. For the special case of two-step Maruyama schemes we derive conditions guaranteeing their mean-square consistency. Further, for the small noise case we obtain expansions of the local error in terms of the step size and the small parameter ϵ. Simulation results using several explicit and implicit stochastic linear k-step schemes, $k = 1,\ 2$, illustrate the theoretical findings.

Key words. stochastic linear multistep method, mean-square convergence, mean-square numerical stability, mean-square consistency, small noise, two-step Maruyama methods

AMS subject classifications. 60H35, 65C30, 65L06, 60H10

DOI. 10.1137/040602857

1. Introduction.

We consider Itô stochastic differential equations (SDEs) of the form

$$X(s)\Big|_{t_0}^{t} = \int_{t_0}^{t} f(X(s),s)\,\mathrm{d}s + \int_{t_0}^{t} G(X(s),s)\,\mathrm{d}W(s), \quad X(t_0) = X_0, \tag{1.1}$$

for $t \in \mathcal{J}$, where $\mathcal{J} = [t_0, T]$. The drift and diffusion functions are given as $f : \mathbb{R}^n \times \mathcal{J} \to \mathbb{R}^n$, $G = (g_1, \ldots, g_m) : \mathbb{R}^n \times \mathcal{J} \to \mathbb{R}^{n\times m}$, respectively. The process W is an m-dimensional Wiener process on the given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_t)_{t\in\mathcal{J}}$, and X_0 is a given $\mathcal{F}_{t_0}$-measurable initial value, independent of the Wiener process and with finite second moment. We assume that there exists a pathwise unique strong solution $X(\cdot)$ of (1.1).

The aim of this article is to analyze the mean-square convergence properties of, in general, drift-implicit linear multistep methods (LMMs) for the approximation of the solution of (1.1). An advantage LMMs have in deterministic numerics is that they require less evaluations of the right-hand side in comparison with Runge–Kutta schemes with the same order of convergence which makes them often preferable for problems with an expensive right-hand side. We recall that in the deterministic case a high order of convergence is always based on sufficient smoothness of the solution of the differential equation. In contrast, the solution of an SDE is not smooth in the ordinary sense and a high order of convergence is achievable only by including more information on the driving Wiener process, i.e., a sufficient number of multiple stochastic integrals. In the case of general systems of SDEs, some of these integrals are difficult to simulate; in several special settings, such as additive or diagonal noise, the situation is less difficult. Our interest in stochastic linear multistep methods (SLMMs)

*Received by the editors January 5, 2004; accepted for publication (in revised form) October 12, 2005; published electronically April 12, 2006. The first author was supported by DFG grant 234499. The second author was supported by the DFG Research Center *Mathematics for Key Technologies* in Berlin.

http://www.siam.org/journals/sinum/44-2/60285.html

†Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (buckwar@mathematik.hu-berlin.de, winkler@mathematik.hu-berlin.de).

stems from applications with small noise in circuit simulation [11, 12, 25, 30, 31], where especially the backward differentiation formulae (BDF) have proven valuable in the deterministic case. An application in the geosciences has been reported in [13].

Stochastic two-step methods already appear in [24] (methods of order 3/2 for equations with additive noise), in [18], and in the recent book [22]. In [2, 3], two-step methods for Itô SDEs are analyzed. Stochastic versions of Adams methods for order up to 5 have been implemented and tested for SDEs with additive noise in [11]. Consistency of SLMMs for Stratonovich SDEs has been considered in [5]; in addition stochastic Adams methods have been implemented as predictor-corrector schemes and tested. A stochastic version of Adams–Bashforth methods, which involves higher order multiple Wiener integrals, is considered in [13]. For small noise SDEs the authors considered SLMMs incorporating mixed classical stochastic integrals in [7].

The current article provides a unified treatment of the mean-square convergence analysis of a general class of SLMMs. This class generalizes all deterministic multistep schemes as well as all Itô schemes considered in the papers cited above. We aim at understanding when and why these schemes converge in the mean-square sense. This includes understanding the relation between local and global errors. To this end we introduce the fundamental notions of mean-square consistency, stability and convergence. The fundamental results in this respect are Theorems 3.2 and 3.3.

As an application of this result, we recapture some of the properties of deterministic LMMs for SDEs with small noise, i.e., SDEs that can be written in the form

$$X(s)\Big|_{t_0}^{t} = \int_{t_0}^{t} f(X(s),s)\,\mathrm{d}s + \int_{t_0}^{t} \epsilon \hat{G}(X(s),s)\,\mathrm{d}W(s), \quad t \in \mathcal{J}, \quad X(t_0) = X_0, \tag{1.2}$$

where $\epsilon \ll 1$ is a small parameter, and $\hat{G} = (\hat{g}_1, \ldots, \hat{g}_m) : \mathbb{R}^n \times \mathcal{J} \to \mathbb{R}^{n\times m}$ and its derivatives are assumed to have moderate values. For this we will derive expansions of the error in terms of the small parameter ϵ and the step size, as has been done for one-step methods in [21]. Then the small diffusion term makes it unnecessary to use high order multiple stochastic integrals, and Wiener increments will be sufficient if the step-size is not too small.

In section 2 we introduce the class of SLMMs considered and provide necessary definitions and useful facts. In section 3 we consider the solvability of the discrete system and the boundedness of the iterates. We then establish two fundamental results, the first one concerns the relation between mean-square numerical stability of the SLMM and Dahlquist's root condition, and the second one concerns the relation between mean-square convergence, mean-square numerical stability, and mean-square consistency of the SLMM. This mirrors the results and the essential role of zero-stability in the deterministic analysis of discretization schemes, see, e.g., [10, 15, 17, 19]. In section 4 we consider two-step Maruyama methods and give conditions for their mean-square consistency. These conditions allow determining the parameters for the stochastic part from the parameters of the deterministic scheme and reduce to those of the underlying deterministic schemes when there is no noise. We then apply the two-step Maruyama methods to the SDE with small noise (1.2) and derive expansions of the local error in a manner similar to that in [21]. In section 5 we provide illustrative examples. The appendix contains the proof of Theorem 3.2.

2. Definitions and preliminary results. We denote by $|\cdot|$ the Euclidian norm in $\mathbb{R}^n$ and by $\|\cdot\|$ the corresponding induced matrix norm. The mean-square norm of a vector-valued square-integrable random variable $Z \in L_2(\Omega, \mathbb{R}^n)$, with $\mathbb{E}$

the expectation with respect to $\mathbb{P}$, will be denoted by

$$\|Z\|_{L_2} := (\mathbb{E}|Z|^2)^{1/2} .$$

We define a deterministic grid on $\mathcal{J}$ as $t_0 < t_1 < \cdots < t_N = T$ with (for simplicity) a constant step-size $h := T/N$ and $t_\ell = t_0 + \ell \cdot h,\ \ell = 0, \ldots, N$. We consider a stochastic linear k-step method, which for $\ell = k, \ldots, N$, takes the form

$$\sum_{j=0}^{k} \alpha_j\ X_{\ell-j} = h \sum_{j=0}^{k} \beta_j\ f(X_{\ell-j}, t_{\ell-j}) + \sum_{j=1}^{k} \Gamma_j(X_{\ell-j}, t_{\ell-j})\ I^{t_{\ell-j}, t_{\ell-j+1}}. \tag{2.1}$$

We set $\alpha_0 = 1$. We require given initial values $X_0, \ldots, X_{k-1} \in L_2(\Omega, \mathbb{R}^n)$ such that X_ℓ is $\mathcal{F}_{t_\ell}$-measurable for $\ell = 0 \ldots, k-1$. As in the deterministic case, usually only $X_0 = X(t_0)$ is given by the initial value problem and the values $X_1, \ldots, X_{k-1}$ need to be computed numerically. This can be done by suitable one-step methods, where one has to be careful to achieve the desired accuracy. Every diffusion term $\Gamma_j(x,t)\ I^{t,t+h}$ is a finite sum of terms, each containing an appropriate function $\mathcal{G}$ of x and t multiplied by a multiple Wiener integral over $[t, t+h]$, i.e., it takes the general form

$$\Gamma_j(x,t)\ I^{t,t+h} = \sum_{r=1}^{m} \mathcal{G}_j^r(x,t)\ I_r^{t,t+h} + \sum_{\substack{r_1, r_2 = 0 \\ r_1 + r_2 > 0}}^{m} \mathcal{G}_j^{r_1, r_2}(x,t)\ I_{r_1,r_2}^{t,t+h} + \ldots\ .$$

A general multiple Wiener integral is given by

$$I_{r_1, r_2, \ldots, r_j}^{t,t+h}(y) = \int_t^{t+h} \int_t^{s_1} \ldots \int_t^{s_{j-1}} y(X(s_j), s_j)\ \mathrm{d}W_{r_1}(s_j) \ldots\ \mathrm{d}W_{r_j}(s_1), \tag{2.2}$$

where $r_i \in \{0, 1, \ldots, m\}$ and $\mathrm{d}W_0(s) = \mathrm{d}s$. If $y \equiv 1$ we write $I_{r_1, r_2, \ldots, r_j}^{t,t+h}$. Note that the integral $I_r^{t,t+h}$ is simply the increment $W_r(t+h) - W_r(t)$ of the scalar Wiener process W_r. The term $I^{t,t+h}$ denotes the collection of multiple Wiener integrals associated with the interval $[t, t+h]$. We emphasize that an explicit discretization is used for the diffusion term. For $\beta_0 = 0$, the SLMM (2.1) is explicit; otherwise it is drift-implicit. We give two examples of two-step methods.

Example 2.1. The first is a stochastic variant of the implicit two-step backward differentiation formula (BDF) method, which we term the BDF2-Maruyama method. For $\ell = 2, \ldots, N$, it takes the form

$$\begin{aligned} X_\ell - \frac{4}{3} X_{\ell-1} + \frac{1}{3} X_{\ell-2} = h\ \frac{2}{3} f(X_\ell, t_\ell) \\ + \sum_{r=1}^{m} g_r(X_{\ell-1}, t_{\ell-1})\ I_r^{t_{\ell-1}, t_\ell} - \frac{1}{3} \sum_{r=1}^{m} g_r(X_{\ell-2}, t_{\ell-2})\ I_r^{t_{\ell-2}, t_{\ell-1}}. \end{aligned}$$

Here one has $\alpha_0 = 1,\ \alpha_1 = -\frac{4}{3},\ \alpha_2 = \frac{1}{3},\ \beta_0 = \frac{2}{3},\ \beta_1 = \beta_2 = 0$, and

$$\Gamma_1(x,t)\ I^{t,t+h} = \sum_{r=1}^{m} g_r(x,t)\ I_r^{t,t+h}, \qquad \Gamma_2(x,t)\ I^{t,t+h} = -\frac{1}{3} \sum_{r=1}^{m} g_r(x,t)\ I_r^{t,t+h}.$$

Example 2.2. The second example is a Milstein variant of the two-step Adams–Bashforth method (we refer to [28, 29] for possibilities to simulate the double Wiener

integrals if the SDE is not of a type that allows simplifications in their simulation):

$$X_\ell - X_{\ell-1} = h\ \left(\frac{3}{2}f(X_{\ell-1},t_{\ell-1}) - \frac{1}{2}f(X_{\ell-2},t_{\ell-2})\right) + \sum_{r=1}^{m} g_r(X_{\ell-1},t_{\ell-1})\ I_r^{t_{\ell-1},t_\ell} + \sum_{q,r=1}^{m} (g_q)'_x g_r(X(t_{\ell-1}),t_{\ell-1})\ I_{q,r}^{t_{\ell-1},t_\ell}.$$

For this method one has $\alpha_0 = 1,\ \alpha_1 = -1,\ \alpha_2 = 0,\ \beta_0 = 0,\ \beta_1 = \frac{3}{2},\ \beta_2 = -\frac{1}{2}$ and

$$\Gamma_2(x,t)\ I^{t,t+h} \equiv 0, \quad \Gamma_1(x,t)\ I^{t,t+h} = \sum_{r=1}^{m} g_r(x,t)\ I_r^{t,t+h} + \sum_{q,r=1}^{m} (g_q)'_x g_r(x,t)\ I_{q,r}^{t,t+h}.$$

A further example of a stochastic Adams–Bashforth method, which is covered by the general form of (2.1), is studied in [13].

We will consider mean-square convergence of SLMMs in the sense discussed by Milstein and others [1, 24, 22, 30]. Note that in the literature the term *strong* convergence is sometimes used synonymously for our expression *mean-square* convergence.

DEFINITION 2.3. *We call the SLMM* (2.1) *for the approximation of the solution of the SDE* (1.1) *mean-square convergent if the global error* $X(t_\ell) - X_\ell$ *satisfies*

$$\max_{\ell=1,\dots,N} \|X(t_\ell) - X_\ell\|_{L_2} \to 0 \ \text{as}\ h \to 0$$

We say it is mean-square convergent with order γ $(\gamma > 0)$ *if the global error satisfies*

$$\max_{\ell=1,\dots,N} \|X(t_\ell) - X_\ell\|_{L_2} \le C \cdot h^\gamma,$$

with constant $C > 0$ *which is independent of the step size* h.

In the following we will define what we understand by local errors. We would like to point out that for the analysis of one-step schemes essentially two different but related concepts are used in the literature. In the first one the local error is defined as the defect that is obtained when the exact solution values are inserted into the numerical scheme. In the second one the local error is defined as the difference after one step of the exact and the numerical solution started at an arbitrary deterministic value. These concepts differ in the way the error is transported to the end of the integration interval, in the first via the numerical method, in the second via the exact solution. The second definition has been used in the fundamental work of Milstein in [23, 24], where for the first time the relation between local and global errors of one-step methods for SDEs has been clarified. However, only the first definition extends easily to multistep methods; hence we will use it here. For comparison of these principles in the deterministic setting see [16, Chapters II.3, III.4].

DEFINITION 2.4. *We define the local error of the SLMM* (2.1) *for the approximation of the solution of the SDE* (1.1)*, for* $\ell = k, \dots, N$*, as*

(2.3)
$$L_\ell := \sum_{j=0}^{k} \alpha_j X(t_{\ell-j}) - h\sum_{j=0}^{k} \beta_j f(X(t_{\ell-j}),t_{\ell-j}) - \sum_{j=1}^{k} \Gamma_j(X(t_{\ell-j}),t_{\ell-j})\ I^{t_{\ell-j},t_{\ell-j+1}}.$$

We aim to conclude mean-square convergence from local properties of the SLMM by means of numerical stability in the mean-square sense. Numerical stability concerns

the influence of perturbations of the right-hand side of the discrete scheme on the global solution of that discrete scheme. Sources of perturbations may be the local error, round-off errors or defects in the approximate solution of implicit schemes. The mean-square stability estimate of the global error is based on the mean-square norm *and* on the conditional mean of the perturbations. In the case of one-step schemes this appears, e.g., in [1, 30]; we refer in particular to the discussion in [24, Chapter 1.4]. We remark that in the case of k-step schemes the conditional mean has to be taken with respect to the σ-algebra $\mathcal{F}_{t_{\ell-k}}$.

In our analysis we thus consider the following discrete system, the perturbed form of (2.1), for $\ell = k, \dots, N$,

$$\sum_{j=0}^{k} \alpha_j \tilde{X}_{\ell-j} = h \sum_{j=0}^{k} \beta_j f(\tilde{X}_{\ell-j}, t_{\ell-j}) + \sum_{j=1}^{k} \Gamma_j(\tilde{X}_{\ell-j}, t_{\ell-j})\, I^{t_{\ell-j},t_{\ell-j+1}} + D_\ell, \tag{2.4}$$

with initial values $\tilde{X}_\ell = X_\ell + D_\ell,\ \ell = 0, \dots, k-1$. We suppose that the perturbations D_ℓ are $\mathcal{F}_{t_\ell}$-measurable and that $D_\ell \in L_2(\Omega, \mathbb{R}^n)$.

Remark 2.5. It is useful to represent the perturbations in the form

$$D_\ell = R_\ell + S_\ell =: R_\ell + \sum_{j=1}^{k} S_{j,\ell-j+1}, \quad \ell = k, \dots, N, \tag{2.5}$$

where each $S_{j,\ell}$ is $\mathcal{F}_{t_\ell}$-measurable with $\mathbb{E}(S_{j,\ell}|\mathcal{F}_{t_{\ell-1}}) = 0$. The representation (2.5) is not unique. One extreme possibility is $R_\ell = D_\ell,\ S_\ell = 0$; another, more useful one, is given by

$$\begin{aligned} R^*_\ell &= \mathbb{E}(D_\ell|\mathcal{F}_{t_{\ell-k}}), \qquad S^*_\ell = D_\ell - R^*_\ell, \\ S^*_{j,\ell-j+1} &= \mathbb{E}(D_\ell - R^*_\ell - \textstyle\sum_{i=j+1}^{k} S^*_{i,\ell-i+1}|\mathcal{F}_{t_{\ell-j+1}}), \quad j = k, k-1, \dots, 1\ . \end{aligned} \tag{2.6}$$

This construction guarantees the required measurability conditions in (2.5). We also note that this decomposition is orthogonal in $L_2(\Omega)$, i.e.,

$$\|D_\ell\|^2_{L_2} = \|R^*_\ell\|^2_{L_2} + \sum_{j=1}^{k} \|S^*_{j,\ell-j+1}\|^2_{L_2}. \tag{2.7}$$

As an example one obtains for $k = 3$,

$$\begin{aligned} R^*_\ell &= \mathbb{E}(D_\ell|\mathcal{F}_{t_{\ell-3}}) \\ S^*_{3,\ell-2} &= \mathbb{E}(D_\ell - R^*_\ell|\mathcal{F}_{t_{\ell-2}}) \\ S^*_{2,\ell-1} &= \mathbb{E}(D_\ell - R^*_\ell - S^*_{3,\ell-2}|\mathcal{F}_{t_{\ell-1}}) \\ S^*_{1,\ell} &= D_\ell - R^*_\ell - S^*_{3,\ell-2} - S^*_{2,\ell-1}. \end{aligned}$$

Here, in the hypothetical case that $D_\ell = c_0 I_r^{t_{\ell-1},t_\ell} + c_1 I_r^{t_{\ell-2},t_{\ell-1}} + c_2 I_r^{t_{\ell-3},t_{\ell-2}} + c_3$, we have $R^*_\ell = c_3,\ S^*_{3,\ell-2} = c_2 I_r^{t_{\ell-3},t_{\ell-2}},\ S^*_{2,\ell-1} = c_1 I_r^{t_{\ell-2},t_{\ell-1}},\ S^*_{1,\ell} = c_0 I_r^{t_{\ell-1},t_\ell}$.

Now we give the precise definition of mean-square stability and consistency that we consider in this paper.

DEFINITION 2.6. *We call the SLMM* (2.1) ***numerically stable in the mean-square sense*** *if there exist constants $h_0 > 0$ and $S > 0$ such that for all step sizes*

$h < h_0$ and for all $\mathcal{F}_{t_\ell}$-measurable perturbations $D_\ell \in L_2(\Omega, \mathbb{R}^n)$ $(\ell = 0, \dots, N)$ and all their representations (2.5)*, the following inequality holds*

(2.8)
$$\max_{\ell=0,\dots,N} \|X_\ell - \tilde{X}_\ell\|_{L_2} \le S \left\{ \max_{\ell=0,\dots,k-1} \|D_\ell\|_{L_2} + \max_{\ell=k,\dots,N} \left(\frac{\|R_\ell\|_{L_2}}{h} + \frac{\|S_\ell\|_{L_2}}{h^{1/2}} \right) \right\},$$

where $(X_\ell)_{\ell=1}^N$ and $(\tilde{X}_\ell)_{\ell=1}^N$ are the solutions of the SLMM (2.1) *and the perturbed discrete system* (2.4)*, respectively.*

We refer to S as the stability constant and to (2.8) as the stability inequality.

DEFINITION 2.7. *We call the SLMM* (2.1) *for the approximation of the solution of the SDE* (1.1) *mean-square consistent if the local error L_ℓ satisfies*

$$h^{-1} \, \|\mathbb{E}(L_\ell | \mathcal{F}_{t_{\ell-k}})\|_{L_2} \to 0 \text{ for } h \to 0 \quad \text{and} \quad h^{-1/2} \, \|L_\ell\|_{L_2} \to 0 \text{ for } h \to 0.$$

We call the SLMM (2.1) *for the approximation of the solution of the SDE* (1.1) *mean-square consistent of order $\gamma (\gamma > 0)$ if the local error L_ℓ satisfies*

$$\|\mathbb{E}(L_\ell | \mathcal{F}_{t_{\ell-k}})\|_{L_2} \le \bar{c} \cdot h^{\gamma+1} \quad \text{and} \quad \|L_\ell\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}, \quad \ell = k, \dots, N,$$

with constants $c\,, \bar{c} > 0$ only depending on the SDE and its solution.

We remind the reader that consistency is only concerned with the local error. In the case that we disregard other sources of errors in (2.4), we only have to deal with perturbations $D_\ell = L_\ell$.

LEMMA 2.8. *The SLMM* (2.1) *is mean-square consistent of order γ if*

$$\|R_\ell\|_{L_2} \le \bar{c} \cdot h^{\gamma+1} \quad \text{and} \quad \|S_\ell\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}, \quad \ell = k, \dots, N$$

for any representation (2.5) *of the local error $D_\ell = L_\ell$. The SLMM* (2.1) *is mean-square consistent of order γ if and only if*

$$\|R_\ell^*\|_{L_2} \le \bar{c} \cdot h^{\gamma+1} \quad \text{and} \quad \|S_\ell^*\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}, \quad \ell = k, \dots, N,$$

where the representation (2.6) *is chosen for the local error $D_\ell = L_\ell$.*

Proof. First, let $L_\ell = R_\ell + S_\ell =: R_\ell + \sum_{j=1}^k S_{j,\ell-j+1}$ be a representation (2.5) with $\|R_\ell\|_{L_2} \le \bar{c} \cdot h^{\gamma+1}$, and $\|S_\ell\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}, \quad \ell = k, \dots, N$. By the conditions $\mathbb{E}(S_{j,\ell} | \mathcal{F}_{t_{\ell-1}}) = 0$ we conclude

$$\|\mathbb{E}(L_\ell | \mathcal{F}_{t_{\ell-k}})\|_{L_2} = \|\mathbb{E}(R_\ell | \mathcal{F}_{t_{\ell-k}})\|_{L_2} \le \|R_\ell\|_{L_2} \le \bar{c} \cdot h^{\gamma+1}.$$

Further, we have, for $h \le 1$,

$$\|L_\ell\|_{L_2} = \|R_\ell + S_\ell\|_{L_2} \le \|R_\ell\|_{L_2} + \|S_\ell\|_{L_2} \le \bar{c} \cdot h^{\gamma+1} + c \cdot h^{\gamma+\frac{1}{2}} \le (\bar{c} + c) \cdot h^{\gamma+\frac{1}{2}}.$$

Second, let the SLMM (2.1) be mean-square consistent of order γ, i.e., $\|\mathbb{E}(L_\ell | \mathcal{F}_{t_{\ell-k}})\|_{L_2}$ $\le \bar{c} \cdot h^{\gamma+1}$, and $\|L_\ell\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}, \quad \ell = k, \dots, N$. Because of $R_\ell^* = \mathbb{E}(L_\ell | \mathcal{F}_{t_{\ell-k}})$, we then, obviously, have $\|R_\ell^*\|_{L_2} \le \bar{c} \cdot h^{\gamma+1}$, and, further

$$\|S_\ell^*\|_{L_2} = \|L_\ell\|_{L_2} - \|R_\ell^*\|_{L_2} \le \|L_\ell\|_{L_2} \le c \cdot h^{\gamma+\frac{1}{2}}. \qquad \square$$

For further reference we state the following definitions and results.

DEFINITION 2.9. *A function $f : \mathbb{R}^n \times \mathcal{J} \to \mathbb{R}^n$ satisfies a uniform Lipschitz condition with respect to x if there exists a positive constant L_f such that*

$$|f(x,t) - f(y,t)| \le L_f|x-y| \quad \forall x, y \in \mathbb{R}^n, t \in \mathcal{J}. \tag{2.9}$$

A function $\Gamma : \mathbb{R}^n \times \mathcal{J} \to \mathbb{R}^{n\times m_\Gamma}$ satisfies a uniform Lipschitz condition with respect to x if there exists a positive constant L_Γ such that

$$\|\Gamma(x,t) - \Gamma(y,t)\| \le L_\Gamma|x-y| \quad \forall x, y \in \mathbb{R}^n, t \in \mathcal{J}. \tag{2.10}$$

Let $C^{s,s-1}$ denote the class of all functions from $\mathbb{R}^n \times \mathcal{J}$ to $\mathbb{R}^n$ having continuous partial derivatives up to order $s-1$ and, in addition, continuous partial derivatives of order s with respect to the first variable.

Let C^K denote the class of functions y from $\mathbb{R}^n \times \mathcal{J}$ to $\mathbb{R}^n$ that satisfy a linear growth condition in the form

$$|y(x,t)| \le K(1+|x|^2)^{\frac{1}{2}} \quad \forall x \in \mathbb{R}^n, t \in \mathcal{J}. \tag{2.11}$$

DEFINITION 2.10. *The characteristic polynomial of* (2.1) *is given by*

$$\rho(\zeta) = \alpha_0\,\zeta^k + \alpha_1\zeta^{k-1} + \cdots + \alpha_k. \tag{2.12}$$

The SLMM (2.1) *is said to fulfill Dahlquist's root condition if*

(i) *the roots of $\rho(\zeta)$ lie on or within the unit circle;*

(ii) *the roots on the unit circle are simple.*

LEMMA 2.11 (A discrete version of Gronwall's lemma). *Let a_ℓ, $\ell = 1, \ldots, N$, and C_1, C_2 be nonnegative real numbers and assume that the inequalities*

$$a_\ell \le C_1 + C_2\frac{1}{N}\sum_{i=1}^{\ell-1} a_i, \quad \ell = 1, \ldots, N,$$

are valid. Then we have $\max_{\ell=1,\ldots,N} a_\ell \le C_1 \exp(C_2)$.

To estimate the multiple integrals (2.2) we will use the following lemma (cf. Lemmas 2.1 and 2.2 in [24]).

LEMMA 2.12. *For any function y belonging to the class C^K, and any $t \in \mathcal{J}$, $h > 0$, such that $t+h \in \mathcal{J}$, we have that*

$$\mathbb{E}(I^{t,t+h}_{r_1\ldots r_j}(y)|\mathcal{F}_t) = 0 \quad \text{if } r_i \ne 0 \quad \text{for some } i \in \{1, \ldots, j\}, \tag{2.13}$$

$$\|\mathbb{E}(I^{t,t+h}_{r_1,\ldots,r_j}(y)|\mathcal{F}_t)\|_{L_2} \le \|I^{t,t+h}_{r_1,\ldots,r_j}(y)\|_{L_2} = \mathcal{O}(h^{l_1+l_2/2}), \tag{2.14}$$

where l_1 is the number of zero indices r_i, and l_2 the number of nonzero indices r_i.

3. Global properties of stochastic LMMs. In this section we will first establish the solvability of the recurrence equations (2.4) (and thus of (2.1)), then we will discuss numerical stability and mean-square convergence of the SLMM (2.1). The former characterizes the robustness of a numerical scheme with respect to small perturbations such as rounding errors. As a property of the numerical scheme alone, it is not a priori giving evidence on the approximation power of the scheme (which may very well approximate a different problem than intended). However, numerical stability and consistency together yield convergence of the numerical solution to the exact solution. In order to distinguish this stability concept from others, it is sometimes called zero-stability or, in honor of Dahlquist, also D-stability. It should not

be mistaken for properties like asymptotic stability, which guarantee that for fixed step sizes (and long or unbounded time intervals) qualitative properties of the exact solutions like damping behavior in dissipative systems are preserved by the discrete approximations. For further discussions we refer the reader to the deterministic literature (see, e.g., [10, 19, 20]). In the stochastic literature mean-square numerical stability for one-step schemes has been considered in [1, 3, 4, 14, 27, 30]. In the first four of these works only perturbations in the initial data have been treated.

We now turn to the solvability of the recurrence equations. If in (2.1) and (2.4) the parameter $\beta_0 = 0$, the discrete systems are explicit and every iterate $\tilde{X}_\ell$, $\ell \geq k$, can be obtained explicitly for given $I^{t_\ell, t_{\ell+1}}$, i.e., the recurrence equations (2.1) and (2.4) obviously have unique solutions. In the case of implicit systems we need to consider the solvability of the systems of nonlinear equations (2.1) and (2.4). In addition, we have to verify that the mean-square norm of the iterates exists. (The straightforward extension to fully implicit systems would serve as an example where the mean-square norm of the iterates does not exist.)

THEOREM 3.1. *Suppose that $\beta_0 \neq 0$ and the drift-coefficient f satisfies* (2.9) *and assume that $2\,h\,\beta_0\,L_f < 1$. Then the perturbed discrete scheme* (2.4) *and, in consequence, the SLMM* (2.1) *have a unique solution. If, in addition, the coefficients Γ_j satisfy* (2.10)*, then the mean-square norm of the iterates exists.*

Proof. The proof of the existence of unique solutions of the perturbed discrete system (2.4) (and thus of (2.1)) follows the line of proofs used in the deterministic analysis of multistep schemes. The idea is to express (2.4) as

$$\tilde{X}_\ell = h\beta_0 f(\tilde{X}_\ell, t_\ell) + \tilde{B}_\ell \,, \tag{3.1}$$

where

$$\tilde{B}_\ell := -\sum_{j=1}^{k} \alpha_j \tilde{X}_{\ell-j} + h\sum_{j=1}^{k} \beta_j f(\tilde{X}_{\ell-j}, t_{\ell-j}) + \sum_{j=1}^{k} \Gamma_j(X_{\ell-j}, t_{\ell-j})\; I^{t_{\ell-j}, t_{\ell-j+1}} + \tilde{D}_\ell$$

is a known $\mathcal{F}_{t_\ell}$ -measurable random variable, when we suppose that $\tilde{X}_{\ell-j}$ are known $\mathcal{F}_{t_{\ell-j}}$-measurable random variables for $j = 1, \ldots, k$. We can then view (2.4) as a fixed-point equation in x,

$$x = h\beta_0 f(x, t_\ell) + b_\ell$$

and apply the contraction mapping principle. We refer, e.g., to [15, Thm. 6.1.1] for more details.

It remains to be shown that the second moments of the iterates exist. We start from the assumption $\mathbb{E}|\tilde{X}_{\ell-j}|^2 < \infty$, $j = 1, \ldots, k$ on the initial values. Recursively, we conclude that $\mathbb{E}|\tilde{X}_\ell|^2 < \infty$, $\ell = k, \ldots, N$ by comparing $\tilde{X}_\ell$ with the solution of the fixed-point equation (3.1) for $\tilde{B}_\ell = 0$, i.e., with the solution of the deterministic implicit equation $x = h\beta_0 f(x, t_\ell)$, and applying Lipschitz continuity arguments. For more details we refer to [30, Thm. 5]. □

We now formulate our main theorem on numerical stability. The proof is given in the appendix.

THEOREM 3.2. *The stochastic linear multistep method* (2.1) *is numerically stable in the mean-square sense for every continuous f and Γ_j satisfying* (2.9) *and* (2.10)*, respectively, if and only if its characteristic polynomial $\rho(\zeta)$* (2.12) *satisfies Dahlquist's root condition given in Definition* 2.10*.*

With the powerful notion of numerical stability in the mean-square sense, together with mean-square consistency, the mean-square convergence follows almost immediately.

THEOREM 3.3. *A mean-square consistent SLMM* (2.1) *for the approximation of the solution of SDE* (1.1) *is mean-square convergent for all continuous* f *and* Γ_j *satisfying* (2.9) *and* (2.10)*, respectively, if and only if it is numerically stable in the mean-square sense. If, in addition, it is mean-square consistent with order* $\gamma > 0$*, then the SLMM* (2.1) *is mean-square convergent with order* γ*.*

Proof. First, let us assume that the mean-square consistent numerical method (2.1) is mean-square convergent. Then the necessity of stability can essentially be proved as in the deterministic case. Set $f \equiv 0,\ \Gamma_j \equiv 0,\ X_0 = 0$. Then (2.1) reduces to $\sum_{j=1}^{k} \alpha_j X_{\ell-j} = 0,\ l = k, k+1, \ldots$, a deterministic homogeneous difference equation, and stability follows by standard arguments, see, e.g., [15].

Second, let us assume that the numerical method (2.1) is mean-square stable and consistent with order $\gamma > 0$. Then mean-square convergence with order γ follows by applying the stability estimate (2.8) to $\{\tilde{X}_\ell := X(t_\ell)\}$ related to the perturbations $\{D_\ell := L_\ell = R^*_\ell + S^*_\ell\}$. ∎

4. Two-step Maruyama schemes. In this section we consider linear two-step Maruyama schemes; thus, we have for $\ell = 2, \ldots, N$

$$\sum_{j=0}^{2} \alpha_j X_{\ell-j} = h \sum_{j=0}^{2} \beta_j f(X_{\ell-j}, t_{\ell-j}) + \sum_{j=1}^{2} \gamma_j \sum_{r=1}^{m} g_r(X_{\ell-j}, t_{\ell-j})\ I_r^{t_{\ell-j}, t_{\ell-j+1}}. \tag{4.1}$$

For drift and diffusion coefficients $f,\ g_1, \ldots, g_m$, which are continuous and satisfy (2.9), Theorem 3.2 applies and the two-step scheme (4.1) is mean-square stable if the coefficients $\alpha_0,\ \alpha_1,\ \alpha_2$ satisfy Dahlquist's root condition. If, additionally, the scheme is mean-square consistent of order γ, which in general requires more smoothness of the coefficient functions, then the scheme (4.1) is mean-square convergent of that order. Thus we will be concerned with mean-square consistency of the above scheme and derive order conditions in terms of the coefficients $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2$. In general, the mean-square order of convergence will be only $\frac{1}{2}$, since the only information about the driving noise process that the Maruyama-type schemes include are the Wiener increments. We note that the simple Euler–Maruyama method would suffice to obtain the same order of convergence. However, convergence is an asymptotic property, i.e., it holds for $h \to 0$ and a result concerning the order of convergence may not provide sufficient information about the size of the actual error that arises for reasonable choices of the step size. In particular when one considers equations with a small noise term as in (1.2), one may find that the influence of the noise is not dominant and properties of the methods in the deterministic setting are recovered to a certain extent.

From the deterministic theory we know that for linear k-step methods

$$\sum_{j=0}^{k} \alpha_j\, x_{\ell-j} = h \sum_{j=0}^{k} \beta_j\, f(x_{\ell-j}, t_{\ell-j}), \quad \text{applied to } x'(t) = f(x(t), t),$$

the local error is of order $p+1$ for sufficiently smooth functions f if

$$\sum_{j=0}^{k} \alpha_j = 0 \quad \text{and} \quad \sum_{j=0}^{k} \alpha_j (k-j)^q = q \sum_{j=0}^{k} \beta_j (k-j)^{q-1} \quad \text{for } q = 1, \ldots, p.$$

In the first part of this section we derive consistency conditions for the two-step scheme (4.1) applied to the general SDE (1.1). We establish a representation of the local error L_ℓ in terms of certain multiple stochastic integrals obtained by the Itô–Taylor expansion. It turns out that consistency is guaranteed under the above conditions for deterministic order 1 and additional conditions that determine the method parameters γ_1 and γ_2.

In the second part of this section we consider the application of the scheme (4.1) in the case that the noise is small. The smallness of the noise is measured by means of the parameter ϵ in the diffusion coefficient $G(x,t) = \epsilon\hat{G}(x,t)$. We emphasize that the numerical schemes include only values of G, the explicit dependence of the diffusion coefficient on the parameter ϵ is used only for a discussion of the errors. We follow the ideas of [21] and develop the local error in powers of the step size h and the small parameter ϵ. The expansion yields the deterministic conditions for order 2, and we discuss for which choices of ϵ and h the stochastic component in the error estimates becomes small compared to these order-2 terms.

4.1. Two-step schemes for general SDEs. To analyze the local error L_ℓ of the scheme (4.1) for the SDE (1.1) and to achieve a suitable representation (2.5) we will derive appropriate Itô–Taylor expansions, where we take special care to separate the stochastic integrals over the different subintervals of integration. We introduce operators Λ_0 and Λ_r, $r = 1, \dots, m$, defined on $C^{2,1}$ and $C^{1,0}$, respectively, by

$$\Lambda_0 y = y'_t + y'_x f + \frac{1}{2}\sum_{r=1}^{m} y''_{xx}[g_r, g_r]\ , \quad \Lambda_r y = y'_x g_r\ ,\ r = 1, \dots, m \tag{4.2}$$

and remind the reader of the notation for multiple Wiener integrals (2.2). Using these operators the Itô formula for a function y in $C^{2,1}$ and the solution X of (1.1) reads

$$y(X(t),t) = y(X(t_0),t_0) + I_0^{t_0,t}(\Lambda_0 y) + \sum_{r=1}^{m} I_r^{t_0,t}(\Lambda_r y), \quad t \in \mathcal{J}. \tag{4.3}$$

Applying the Itô formula (4.3) on the corresponding intervals to the drift coefficient f as well as to the diffusion coefficients g_r yields for $s \in [t_{\ell-j}, t_{\ell-j+1}]$, $j = 1, 2$

$$f(X(s),s) = f(X(t_{\ell-j}),t_{\ell-j}) + I_0^{t_{\ell-j},s}(\Lambda_0 f) + \sum_{r=1}^{m} I_r^{t_{\ell-j},s}(\Lambda_r f), \tag{4.4}$$

$$g_r(X(s),s) = g_r(X(t_{\ell-j}),t_{\ell-j}) + I_0^{t_{\ell-j},s}(\Lambda_0 g_r) + \sum_{q=1}^{m} I_q^{t_{\ell-j},s}(\Lambda_q g_r). \tag{4.5}$$

We trace back the values of the drift coefficient to the point $t_{\ell-2}$ to obtain

$$f(X(t_{\ell-1}),t_{\ell-1}) = f(X(t_{\ell-2}),t_{\ell-2}) + I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + \sum_{r=1}^{m} I_r^{t_{\ell-2},t_{\ell-1}}(\Lambda_r f), \tag{4.6}$$

$$\begin{aligned} f(X(t_\ell),t_\ell) = {} & f(X(t_{\ell-2}),t_{\ell-2}) + I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + I_0^{t_{\ell-1},t_\ell}(\Lambda_0 f) \\ & + \sum_{r=1}^{m} I_r^{t_{\ell-2},t_{\ell-1}}(\Lambda_r f) + \sum_{r=1}^{m} I_r^{t_{\ell-1},t_\ell}(\Lambda_r f). \end{aligned} \tag{4.7}$$

For the general SDE (1.1) we have the following result.

LEMMA 4.1. *Assume that the coefficients* $f,\ g_r, r = 1,\dots,m$ *of the SDE* (1.1) *belong to the class* $C^{2,1}$ *with* $\Lambda_0 f, \Lambda_0 g_r, \Lambda_r f, \Lambda_q g_r \in C^K$ *for* $r,q = 1,\dots,m$. *Then the local error* (2.3) *of the stochastic two-step scheme* (4.1) *allows the representation*

$$L_\ell = R_\ell^\circ + S_{1,\ell}^\circ + S_{2,\ell-1}^\circ, \quad \ell = 2,\dots,N, \tag{4.8}$$

where $R_\ell^\circ, S_{j,\ell}^\circ, j = 1,2$ *are* $\mathcal{F}_{t_\ell}$*-measurable with* $\mathbb{E}(S_{j,\ell}^\circ|\mathcal{F}_{t_{\ell-1}}) = 0$ *and*

$$\begin{aligned}
R_\ell^\circ &= \left[\sum_{j=0}^{2}\alpha_j\right] X(t_{\ell-2}) + \left[2\alpha_0 + \alpha_1 - \sum_{j=0}^{2}\beta_j\right] hf(X(t_{\ell-2}),t_{\ell-2}) + \tilde{R}_\ell^\circ,\\
S_{1,\ell}^\circ &= \Big[\alpha_0 - \gamma_1\Big]\sum_{r=1}^{m} g_r(X(t_{\ell-1}),t_{\ell-1})I_r^{t_{\ell-1},t_\ell} + \tilde{S}_{1,\ell}^\circ,\\
S_{2,\ell-1}^\circ &= \Big[(\alpha_0+\alpha_1) - \gamma_2\Big]\sum_{r=1}^{m} g_r(X(t_{\ell-2}),t_{\ell-2})I_r^{t_{\ell-2},t_{\ell-1}} + \tilde{S}_{2,\ell-1}^\circ
\end{aligned}$$

with

$$\|\tilde{R}_\ell^\circ\|_{L_2} = O(h^2), \quad \|\tilde{S}_{1,\ell}^\circ\|_{L_2} = O(h), \quad \|\tilde{S}_{2,\ell-1}^\circ\|_{L_2} = O(h). \tag{4.9}$$

COROLLARY 4.2. *Let the coefficients* $f,\ g_r,\ r = 1,\dots,m$, *of the SDE* (1.1) *satisfy the assumptions of Lemma* 4.1 *and suppose they are Lipschitz continuous with respect to their first variable. Let the coefficients of the stochastic linear two-step scheme* (4.1) *satisfy Dahlquist's root condition and the consistency conditions*

$$\sum_{j=0}^{2}\alpha_j = 0, \quad 2\alpha_0 + \alpha_1 = \sum_{j=0}^{2}\beta_j, \quad \alpha_0 = \gamma_1, \quad \alpha_0 + \alpha_1 = \gamma_2. \tag{4.10}$$

Then the global error of the scheme (4.1) *applied to* (1.1) *allows the expansion*

$$\max_{\ell=2,\dots,N}\|X(t_\ell) - X_\ell\|_{L_2} = \mathcal{O}(h^{1/2}) + \mathcal{O}(\max_{\ell=0,1}\|X(t_\ell) - X_\ell\|_{L_2})\ .$$

Proof of Corollary 4.2. By Lemma 4.1 we have the representation (4.8) for the local error. Applying the consistency conditions (4.10) yields

$$R_\ell^\circ = \tilde{R}_\ell^\circ, \quad S_{1,\ell}^\circ = \tilde{S}_{1,\ell}^\circ, \quad S_{2,\ell-1}^\circ = \tilde{S}_{2,\ell-1}^\circ, \quad \ell = 2,\dots,N.$$

As the scheme (4.1) satisfies Dahlquist's root condition, it is numerically stable in the mean-square sense. Now the assertion follows from the estimates (4.9) by means of the stability inequality. ∎

Proof of Lemma 4.1. To derive a representation of the local error in the form (4.8) we evaluate and resume the deterministic parts at the point $(X(t_{\ell-2}),t_{\ell-2})$ and separate the stochastic terms carefully over the different subintervals $[t_{\ell-2},t_{\ell-1}]$ and $[t_{\ell-1},t_\ell]$. This ensures the independence of the random variables. It does make the calculations more messy, though. By rewriting

$$\sum_{j=0}^{2}\alpha_j X(t_{\ell-j}) = \alpha_0\big(X(t_\ell) - X(t_{\ell-1})\big) + (\alpha_0+\alpha_1)\big(X(t_{\ell-1}) - X(t_{\ell-2})\big) + \big(\sum_{j=0}^{2}\alpha_j\big)X(t_{\ell-2}),$$

we can express the local error (2.3) as

$$L_\ell = \alpha_0\big(X(t_\ell) - X(t_{\ell-1})\big) + (\alpha_0+\alpha_1)\big(X(t_{\ell-1}) - X(t_{\ell-2})\big) + \sum_{j=0}^{2}\alpha_j X(t_{\ell-2})$$
$$-h\sum_{j=0}^{2}\beta_j f(X(t_{\ell-j}),t_{\ell-j}) - \sum_{j=1}^{2}\gamma_j\sum_{r=1}^{m} g_r(X(t_{\ell-j}),t_{\ell-j})I_r^{t_{\ell-j},t_{\ell-j+1}}.$$

The SDE (1.1) implies the identities

$$X(t_{\ell-1}) - X(t_{\ell-2}) = hf(X(t_{\ell-2}),t_{\ell-2}) + I_{00}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + \sum_{r=1}^{m} I_{r0}^{t_{\ell-2}t_{\ell-1}}(\Lambda_r f)$$
$$+\sum_{r=1}^{m} g_r(X(t_{\ell-2}),t_{\ell-2})I_r^{t_{\ell-2},t_{\ell-1}} + \sum_{r=1}^{m} I_{0r}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 g_r) + \sum_{r,q=1}^{m} I_{qr}^{t_{\ell-2},t_{\ell-1}}(\Lambda_q g_r),$$

and, additionally using (4.6),

$$X(t_\ell) - X(t_{\ell-1}) = h\big\{f(X(t_{\ell-2}),t_{\ell-2}) + I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + \sum_{r=1}^{m} I_r^{t_{\ell-2},t_{\ell-1}}(\Lambda_r f)\big\}$$
$$+I_{00}^{t_{\ell-1},t_\ell}(\Lambda_0 f) + \sum_{r=1}^{m} I_{r0}^{t_{\ell-1}t_\ell}(\Lambda_r f) + \sum_{r=1}^{m} g_r(X(t_{\ell-1}),t_{\ell-1})I_r^{t_{\ell-1},t_\ell}$$
$$+\sum_{r=1}^{m} I_{0r}^{t_{\ell-1},t_\ell}(\Lambda_0 g_r) + \sum_{r,q=1}^{m} I_{qr}^{t_{\ell-1},t_\ell}(\Lambda_q g_r).$$

Inserting this and the expansions (4.6, 4.7) into the local error formula and reordering the terms, yields

$$L_\ell = \left[\sum_{j=0}^{2}\alpha_j\right] X(t_{\ell-2}) + \left[2\alpha_0+\alpha_1-\sum_{j=0}^{2}\beta_j\right] hf(X(t_{\ell-2}),t_{\ell-2}) + \tilde{R}_\ell^\circ$$
$$+\Big[\alpha_0-\gamma_1\Big]\sum_{r=1}^{m} g_r(X(t_{\ell-1}),t_{\ell-1})I_r^{t_{\ell-1},t_\ell} + \tilde{S}_{1,\ell}^\circ$$
$$+\Big[(\alpha_0+\alpha_1)-\gamma_2\Big]\sum_{r=1}^{m} g_r(X(t_{\ell-2}),t_{\ell-2})I_r^{t_{\ell-2},t_{\ell-1}} + \tilde{S}_{2,\ell-1}^\circ,$$

where

$$\tilde{R}_\ell^\circ = \alpha_0\big\{hI_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + I_{00}^{t_{\ell-1},t_\ell}(\Lambda_0 f)\big\} + (\alpha_0+\alpha_1)I_{00}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f)$$
$$-h\beta_0\big\{I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f) + I_0^{t_{\ell-1},t_\ell}(\Lambda_0 f)\big\} - h\beta_1 I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 f), \tag{4.11}$$

$$\tilde{S}_{1,\ell}^\circ = \sum_{r=1}^{m}\Big(\alpha_0 I_{r0}^{t_{\ell-1},t_\ell}(\Lambda_r f) - h\beta_0 I_r^{t_{\ell-1},t_\ell}(\Lambda_r f)\Big) + \alpha_0\sum_{r=1}^{m} I_{0r}^{t_{\ell-1},t_\ell}(\Lambda_0 g_r)$$
$$+\alpha_0\sum_{r,q=1}^{m} I_{qr}^{t_{\ell-1},t_\ell}(\Lambda_q g_r)), \tag{4.12}$$

$$\tilde{S}^{\circ}_{2,\ell-1} = h(\alpha_0-\beta_0-\beta_1)\sum_{r=1}^{m} I_r^{t_{\ell-2},t_{\ell-1}}(\Lambda_r f) + (\alpha_0+\alpha_1)\sum_{r=1}^{m} I_{r0}^{t_{\ell-2},t_{\ell-1}}(\Lambda_r f)$$

$$+(\alpha_0+\alpha_1)\sum_{r=1}^{m} I_{0r}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0 g_r) + (\alpha_0+\alpha_1)\sum_{r,q=1}^{m} I_{qr}^{t_{\ell-2},t_{\ell-1}}(\Lambda_q g_r). \tag{4.13}$$

Finally, the estimates (4.9) are derived by means of Lemma 2.12, where the last terms in (4.12) and (4.13) determine the order $\mathcal{O}(h)$. ∎

4.2. Two-step schemes for small noise SDEs. For the numerical integration of ODEs two-step schemes of order 2 or higher are particularly interesting. They offer a high order of convergence for low computational cost per step. In this section we discuss the special case of small noise SDEs (1.2) where the parameter ϵ in the diffusion coefficients $g_r = \epsilon\hat{g}_r,\ r = 1,\dots,m$ measures the smallness of the noise. Lemma 4.1 provides a representation for the local error (2.3) of the stochastic linear two-step scheme (4.1) applied to (1.2). Starting from this expression we will further analyze the local error by expanding the term $\Lambda_0 f$ appearing in $\tilde{R}^{\circ}_{\ell}$ (4.11). Naturally, this requires more smoothness of the coefficients. A sufficient condition would be $\Lambda_0 f \in C^{2,1}$, for which, in general, one needs the existence of fourth order derivatives of f with respect to x. However, for small noise SDEs, the term $f'_x f + f'_t$ dominates $\Lambda_0 f$. This allows weakening the smoothness assumptions again. The expansion of $\Lambda_0 f$ also yields additional multiple Itô integrals whose conditional expectation vanishes. By moving these terms from $\tilde{R}^{\circ}_{\ell}$ into the stochastic parts of the representation of the local error we achieve better estimates. With this analysis we are able to prove that some of the potential of deterministic two-step schemes can be recovered in the special case of small noise SDEs.

To be able to exploit the effect of the small parameter ϵ in the expansions of the local error we introduce operators $\Lambda_0^f, \hat{\Lambda}_0$ and $\hat{\Lambda}_r,\ r = 1,\dots,m$ defined on $C^{2,1}$ and $C^{1,0}$, respectively, by

$$\Lambda_0^f y := y'_t + y'_x f, \quad \hat{\Lambda}_0 y := \frac{1}{2}\sum_{r=1}^{m} y''_{xx}[\hat{g}_r,\hat{g}_r], \quad \hat{\Lambda}_r y := y'_x \hat{g}_r\ . \tag{4.14}$$

In terms of the original definition (4.2) we have

$$\Lambda_0 y = \Lambda_0^f y + \epsilon^2 \hat{\Lambda}_0 y \quad \text{and} \quad \Lambda_r y = \epsilon\hat{\Lambda}_r y. \tag{4.15}$$

LEMMA 4.3. *Assume that the coefficients $f, \hat{g}_r, r = 1,\dots,m$ of the small noise SDE* (1.2)*, as well as $\Lambda_0^f f = f'_x f + f'_t$ belong to the class $C^{2,1}$ with $\Lambda_0 f, \Lambda_0\hat{g}_r, \hat{\Lambda}_r f$, $\hat{\Lambda}_q\hat{g}_r, \Lambda_0\Lambda_0^f f,\ \hat{\Lambda}_r\Lambda_0^f f \in C^K$ for $r,q = 1,\dots,m$. Let the stochastic two-step scheme* (4.1) *satisfy the consistency conditions* (4.10)*. Then the local error* (2.3) *of the method* (4.1) *for the small noise SDE* (1.2) *allows the representation*

$$L_\ell = R^{\diamond}_{\ell} + S^{\diamond}_{1,\ell} + S^{\diamond}_{2,\ell-1}, \quad \ell = 2,\dots,N, \tag{4.16}$$

where $R^{\diamond}_{\ell}, S^{\diamond}_{j,\ell}, j = 1,2$ are $\mathcal{F}_{t_\ell}$-measurable with $\mathbb{E}(S^{\diamond}_{j,\ell}|\mathcal{F}_{t_{\ell-1}}) = 0$, and

$$R^{\diamond}_{\ell} = \Big[(4\alpha_0+\alpha_1) - (4\beta_0+2\beta_1)\Big]\frac{h^2}{2}(f'_t + f'_x f)(X(t_{\ell-2}),t_{\ell-2}) + \tilde{R}^{\diamond}_{\ell},$$
$$S^{\diamond}_{1,\ell} = \tilde{S}^{\circ}_{1,\ell} + \tilde{S}^{\diamond}_{1,\ell},$$
$$S^{\diamond}_{2,\ell-1} = \tilde{S}^{\circ}_{2,\ell-1} + \tilde{S}^{\diamond}_{2,\ell-1},$$

where

$$\|\tilde{R}_\ell^\diamond\|_{L_2} = O(h^3 + \epsilon^2 h^2), \quad \|\tilde{S}_{1,\ell}^\diamond\|_{L_2} = O(\epsilon h^{5/2}), \quad \|\tilde{S}_{2,\ell-1}^\diamond\|_{L_2} = O(\epsilon h^{5/2}). \tag{4.17}$$

The terms $\tilde{S}_{1,\ell}^\circ, \tilde{S}_{2,\ell-1}^\circ$ *are given by* (4.12, 4.13) *in the proof of Lemma* 4.1 *and satisfy here*

$$\|\tilde{S}_{1,\ell}^\circ\|_{L_2} = \mathcal{O}(\epsilon^2 h + \epsilon h^{3/2}), \quad \|\tilde{S}_{2,\ell}^\circ\|_{L_2} = \mathcal{O}(\epsilon^2 h + \epsilon h^{3/2}). \tag{4.18}$$

Proof. We have from Lemma 4.1, if the consistency conditions (4.10) are satisfied, the representation

$$L_\ell = \tilde{R}_\ell^\circ + \tilde{S}_{1,\ell}^\circ + \tilde{S}_{2,\ell-1}^\circ, \quad \ell = 2, \dots, N,$$

where $\tilde{R}_\ell^\circ$, $\tilde{S}_{1,\ell}^\circ$, $\tilde{S}_{2,\ell-1}^\circ$ are given by (4.11, 4.12, 4.13). Splitting $\Lambda_0 f = \Lambda_0^f f + \epsilon^2 \hat{\Lambda}_0 f$ immediately yields $\tilde{R}_\ell^\circ = \tilde{R}_\ell^{\circ f} + \epsilon^2 \hat{R}_\ell^\circ$ with

$$\begin{aligned} \tilde{R}_\ell^{\circ f} &:= (\alpha_0 - \beta_0 - \beta_1) h I_0^{t_{\ell-2}, t_{\ell-1}}(\Lambda_0^f f) + (\alpha_0 + \alpha_1) I_{00}^{t_{\ell-2}, t_{\ell-1}}(\Lambda_0^f f) \\ &\quad + \alpha_0 I_{00}^{t_{\ell-1}, t_\ell}(\Lambda_0^f f) - h\beta_0 I_0^{t_{\ell-1}, t_\ell}(\Lambda_0^f f) \end{aligned} \tag{4.19}$$

$$\begin{aligned} \hat{R}_\ell^{\circ} &:= (\alpha_0 - \beta_0 - \beta_1) h I_0^{t_{\ell-2}, t_{\ell-1}}(\hat{\Lambda}_0 f) + (\alpha_0 + \alpha_1) I_{00}^{t_{\ell-2}, t_{\ell-1}}(\hat{\Lambda}_0 f) \\ &\quad + \alpha_0 I_{00}^{t_{\ell-1}, t_\ell}(\hat{\Lambda}_0 f) - h\beta_0 I_0^{t_{\ell-1}, t_\ell}(\hat{\Lambda}_0 f). \end{aligned} \tag{4.20}$$

We note that (4.20) appears with the factor ϵ^2 in the local error representation, thus yielding the $\mathcal{O}(\epsilon^2 h^2)$ term in the estimate of $\|\tilde{R}_\ell^\diamond\|_{L_2}$ in (4.17). We concentrate on developing $\tilde{R}_\ell^{\circ f}$ in more detail. Applying the Itô formula (4.3) to $\Lambda_0^f f(X(s), s)$ for $s \in [t_{\ell-2}, t_{\ell-1}]$ and integrating yields $I_0^{t_{\ell-2}, s}(\Lambda_0^f f) = (s - t_{\ell-2})\Lambda_0^f f(X(t_{\ell-2}), t_{\ell-2}) + I_{00}^{t_{\ell-2}, s}(\Lambda_0 \Lambda_0^f f) + \epsilon \sum_{r=1}^m I_{r0}^{t_{\ell-2}, s}(\hat{\Lambda}_r \Lambda_0^f f)$.

For $s = t_{\ell-1}$ we obtain an expression for the first integral in (4.19) . Integrating again we have for the second integral in (4.19) $I_{00}^{t_{\ell-2}, t_{\ell-1}}(\Lambda_0^f f) = \frac{h^2}{2}\Lambda_0^f f(X(t_{\ell-2}), t_{\ell-2}) + I_{000}^{t_{\ell-2}, t_{\ell-1}}(\Lambda_0 \Lambda_0^f f) + \epsilon \sum_{r=1}^m I_{r00}^{t_{\ell-2}, t_{\ell-1}}(\hat{\Lambda}_r \Lambda_0^f f)$.

Both the other integrals are over the interval $[t_{\ell-1}, t_\ell]$. In the analogous expressions for these the term $\Lambda_0^f f(X(t_{\ell-1}), t_{\ell-1})$ must be substituted by $\Lambda_0^f f(X(t_{\ell-1}), t_{\ell-1}) = \Lambda_0^f f(X(t_{\ell-2}), t_{\ell-2}) + I_0^{t_{\ell-2}, t_{\ell-1}}(\Lambda_0 \Lambda_0^f f) + \epsilon \sum_{r=1}^m I_r^{t_{\ell-2}, t_{\ell-1}}(\Lambda_r \Lambda_0^f f)$.

Then we obtain from (4.19)

$$\tilde{R}_\ell^{\circ f} = \Big[(4\alpha_0 + \alpha_1) - (4\beta_0 + 2\beta_1)\Big] \frac{h^2}{2} \Lambda_0^f f(X(t_{\ell-2}), t_{\ell-2}) + \tilde{R}_\ell^{\diamond f} + \tilde{S}_{1,\ell}^\diamond + \tilde{S}_{2,\ell}^\diamond,$$

where

$$\begin{aligned}
\tilde{R}_\ell^{\diamond f} &= (\alpha_0 - 2\beta_0)\frac{h^2}{2} I_0^{t_{\ell-2},t_{\ell-1}}(\Lambda_0\Lambda_0^f f) \\
&\quad + (\alpha_0 - \beta_0 - \beta_1) h I_{00}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0\Lambda_0^f f) - \beta_0 h I_{00}^{t_{\ell-1},t_\ell}(\Lambda_0\Lambda_0^f f) \\
&\quad + (\alpha_0 + \alpha_1) I_{000}^{t_{\ell-2},t_{\ell-1}}(\Lambda_0\Lambda_0^f f) + \alpha_0 I_{000}^{t_{\ell-1},t_\ell}(\Lambda_0\Lambda_0^f f), \\
\tilde{S}_{1,\ell}^{\diamond} &= \alpha_0\epsilon \sum_{r=1}^m I_{r00}^{t_{\ell-1},t_\ell}(\hat{\Lambda}_r\Lambda_0^f f) - h\beta_0\epsilon \sum_{r=1}^m I_{r0}^{t_{\ell-1},t_\ell}(\hat{\Lambda}_r\Lambda_0^f f), \\
\tilde{S}_{2,\ell}^{\diamond} &= (\alpha_0 - 2\beta_0)\frac{h^2}{2}\epsilon \sum_{r=1}^m I_r^{t_{\ell-2},t_{\ell-1}}(\hat{\Lambda}_r\Lambda_0^f f) + (\alpha_0 - \beta_0 - \beta_1) h\epsilon \sum_{r=1}^m I_{r0}^{t_{\ell-2},t_{\ell-1}}(\hat{\Lambda}_r\Lambda_0^f f) \\
&\quad + (\alpha_0 + \alpha_1)\epsilon \sum_{r=1}^m I_{r00}^{t_{\ell-2},t_{\ell-1}}(\hat{\Lambda}_r\Lambda_0^f f).
\end{aligned}$$

We arrive at $\tilde{R}_\ell^{\diamond} = \tilde{R}_\ell^{\diamond f} + \epsilon^2 \hat{R}_\ell^{\circ}$. Finally, the estimates (4.17) are derived by means of Lemma 2.12. ∎

COROLLARY 4.4. *Let the coefficients* f, $\hat{g}_r$, $r = 1, \dots, m$, *of the SDE* (1.2) *satisfy the assumptions of Lemma* 4.3 *and suppose they are Lipschitz continuous with respect to their first variable. Let the coefficients of the stochastic linear two-step scheme* (4.1) *satisfy Dahlquist's root condition and the consistency conditions* (4.10) *and*

$$(4\alpha_0 + \alpha_1) - (4\beta_0 + 2\beta_1) = 0. \tag{4.21}$$

Then the global error of the scheme (4.1) *applied to* (1.2) *allows the expansion*

$$\max_{\ell=2,\dots,N} \|X(t_\ell) - X_\ell\|_{L_2} = \mathcal{O}(h^2 + \epsilon h + \epsilon^2 h^{1/2}) + \mathcal{O}(\max_{\ell=0,1} \|X(t_\ell) - X_\ell\|_{L_2}) .$$

Proof. Lemma 4.3 stated the representation (4.16) for the local error. Applying the consistency condition (4.21) yields $R_\ell^{\diamond} = \tilde{R}_\ell^{\diamond}$ and by (4.17) we have $\|R_\ell^{\diamond}\|_{L_2} = \mathcal{O}(h^3 + \epsilon^2 h^2)$. The stochastic terms $S_{1,\ell}^{\diamond}$, $S_{2,\ell-1}^{\diamond}$ are dominated by $\tilde{S}_{1,\ell}^{\circ}$, $\tilde{S}_{2,\ell-1}^{\circ}$ and thus are of order of magnitude $\mathcal{O}(\epsilon^2 h + \epsilon h^{3/2})$. As the scheme (4.1) satisfies Dahlquist's root condition, it is numerically stable in the mean-square sense. Applying the stability inequality (2.8) to the representation (4.16) of the local error yields the assertion. ∎

Remark 4.5. We note that $R_\ell^{\diamond}$ is responsible for the $\mathcal{O}(h^2)$ term in the expansion of the global error. In the limit $\epsilon \to 0$ this is the only remaining term. It reflects the convergence properties of the scheme in the deterministic setting. On the other hand, for asymptotically small step sizes $h \to 0$ and fixed parameter ϵ, even if it is small, the term of order $\mathcal{O}(\epsilon^2 h^{1/2})$ causes the low order $\frac{1}{2}$ of convergence. The question arises for which choices of ϵ and h the schemes still show the order 2 behavior. Thus we are interested in when the term $\mathcal{O}(h^2)$ dominates the term $\mathcal{O}(\epsilon h + \epsilon^2 h^{1/2})$. Clearly both these terms depend on the actual coefficients $f, \hat{g}_r, r = 1, \dots, m$, of the SDE (1.2) and their derivatives. Assuming moderate function values, the term $\mathcal{O}(h^2)$ dominates $\mathcal{O}(\epsilon h + \epsilon^2 h^{1/2})$, if $h^2 \gg \epsilon^2 h^{1/2}$, i.e., $h \gg \epsilon^{4/3}$, and $h^2 \gg \epsilon h$, i.e., $h \gg \epsilon$. Obviously, in general, the second condition is stronger. Summarizing, we can expect order 2 behavior if $h \gg \epsilon$.

However, even if the chosen step size does not satisfy this condition for a given ϵ, schemes satisfying the consistency condition (4.21) often show a better performance

TABLE 5.1
Coefficients of two-step schemes.

Method	α_0	α_1	α_2	β_0	β_1	β_2	γ_1	γ_2
Unstable method	1	-3	2	0	1/2	$-3/2$	1	-2
Explicit Euler	1	-1	0	0	1	0	1	0
Implicit Euler	1	-1	0	1	0	0	1	0
Trapezoidal rule	1	-1	0	1/2	1/2	0	1	0
BDF 2	1	$-4/3$	1/3	2/3	0	0	1	$-1/3$
Adams–Bashforth 2	1	-1	0	0	3/2	$-1/2$	1	0
Adams–Moulton 2	1	-1	0	5/12	8/12	$-1/12$	1	0
Milne–Simpson	1	0	-1	1/3	4/3	1/3	1	1

than other schemes. The reason is that their error is dominated by $\mathcal{O}(\epsilon h)$ instead of $\mathcal{O}(\epsilon^2 h^{1/2})$, resulting in an order 1 behavior with the small parameter ϵ in the error constant. Again assuming moderate function values, one may expect this for $\epsilon h \gg \epsilon^2 h^{1/2}$, i.e., $h \gg \epsilon^2$.

5. Test results. We report results for several explicit and implicit stochastic linear k-step schemes for $k = 1, 2$, applied to two examples of SDEs. Table 5.1 summarizes the methods we have implemented and tested. All methods, including the one-step schemes, satisfy the consistency conditions (4.10) and all methods, excluding the Euler schemes, satisfy the consistency condition (4.21). Further, all methods, exept the first one, are zero-stable. The Milne–Simpson method is the only linear two-step method with deterministic order of convergence 4. However, its characteristic polynomial possesses the second root $\alpha_2 = -1$ on the unit circle which easily causes weak instabilities in the integration of decaying solutions, such that this scheme is generally not recommended.

The first example is the simple bilinear scalar SDE

$$X(t) = 1 + \int_0^t \alpha X(s)\, \mathrm{d}s + \int_0^t \beta X(s)\, \mathrm{d}W(s), \quad t \in [0,1] \tag{5.1}$$

with coefficients $f(x,t) := \alpha x$, $G(x,t) = (g(x,t)) = (\beta x)$, parameters $\alpha, \beta \in \mathbb{R}$, and a scalar Wiener process W. This example has the advantage that an explicit solution is available to compare with the numerical approximations. This solution is given by the geometric Brownian motion $X(t) = \exp\big((\alpha - \frac{1}{2}\beta^2)t + \beta W(t)\big)$.

The second example is a two-dimensional system of SDEs which is taken from [8, 9, Problem P1], where it is given in Stratonovich notation. For the purpose of using it as a test equation for small noise SDEs in the form (1.2) we have transformed

it into Itô notation and included a parameter ϵ. We arrive at

$$
\begin{aligned}
(5.2)\quad X(t) &= X(0) + \int_0^t FX(s)\,\mathrm{d}s + \int_0^t G_1 X(s)\,\mathrm{d}W_1(s) + \int_0^t G_2 X(s)\,\mathrm{d}W_2(s),\\
F &= \begin{pmatrix} -0.9 & 0 \\ 0.25 & -0.5 \end{pmatrix} + \frac{\epsilon^2}{2}\begin{pmatrix} 0.75^2+0.9^2 & 0 \\ 0 & 0.75^2+0.9^2 \end{pmatrix},\\
G_1 &= \epsilon\begin{pmatrix} 0.75 & 0 \\ 0 & -0.75 \end{pmatrix}, \quad G_2 = \epsilon\begin{pmatrix} 0 & 0.9 \\ 0.9 & 0 \end{pmatrix}, \quad t \in [0,2]\,.
\end{aligned}
$$

The above system is fully noncommutative, i.e., $[F, G_1]$, $[F, G_2]$, $[G_1, G_2]$ are all nonzero, where $[A, B] = AB - BA$ for matrices A, B. Here we cannot use an explicit formula for the solution of the system; thus we have computed a "reference solution" with the trapezoidal rule on a very fine grid by using 262144 steps.

Results for another example, a scalar SDE with polynomial drift and diffusion coefficients, are reported in [6].

In our experiments we have investigated the relation between step size h and achieved accuracy for several choices of parameters for the two test examples. The accuracy is measured as the maximum approximate L_2-norm of the global errors in the time interval $[0, 1]$ and $[0, 2]$, respectively:

$$
\max_{\ell=1,\dots,N}\left(\frac{1}{M}\sum_{j=1}^{M}|X(t_\ell,\omega_j) - X_\ell(\omega_j)|^2\right)^{1/2} \approx \max_{\ell=1,\dots,N}\|X(t_\ell) - X_\ell\|_{L_2},
$$

where N denotes the number of steps and M the number of computed paths. In our computations we used $M = 100$, unless specified otherwise.

The results are presented as figures, where we have plotted the achieved accuracy versus the step sizes in logarithmic scale with base 10. Then the slope of the resulting lines corresponds to the observed order of the schemes. Lines with slopes 0.5, 1, 2, 3 are provided in some figures to enable comparisons with convergence of these orders.

Our first test concerned the effect of applying a method which is not numerically stable in the mean-square sense, i.e., the coefficients of the method do not satisfy Dahlquist's root condition. For the method considered (the parameters are given in the first row of Table 5.1), one of the roots of the characteristic polynomial ρ (2.12) is 2. Figure 1 shows the behavior of the error of this method applied to the SDE (5.1) with $\alpha = -1$, $\beta = 0.01$, when the step size decreases. We note that the scale is logarithmic!

Next we report results for the other methods listed in Table 5.1, applied to the simple linear SDE (5.1). To start off the integration with the two-step schemes we needed a second starting value X_1. In this example we used the exact solution value $X(t_1)$, thus avoiding introducing additional errors. In computational practice, the starting value X_1 could be computed, e.g., by means of the trapezoidal rule. Then the error structure of the two-step Maruyama schemes is maintained. For reasons of space restriction we present only results for one set of parameters, $\alpha = -1$ and $\beta = 0.01$, where β takes the role of the parameter ϵ. The results are given in Figure 2 and illustrate the observations made in Remark 4.5 very well. We see that the error of the other schemes is smaller than that of the Euler schemes. Their error appears to be of the size $\max(c_1\beta^2 h^{\frac{1}{2}}, c_2\beta h, c_3 h^2)$, c_3 being an error constant particular to the scheme. The constants c_1, c_2 appear to coincide for the considered schemes.

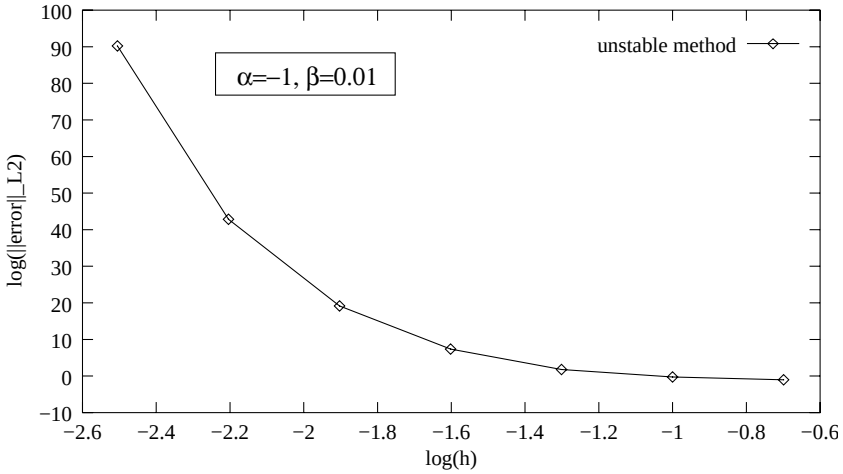


FIG. 1. *Simulation results for the SDE* (5.1) *with the unstable scheme.*

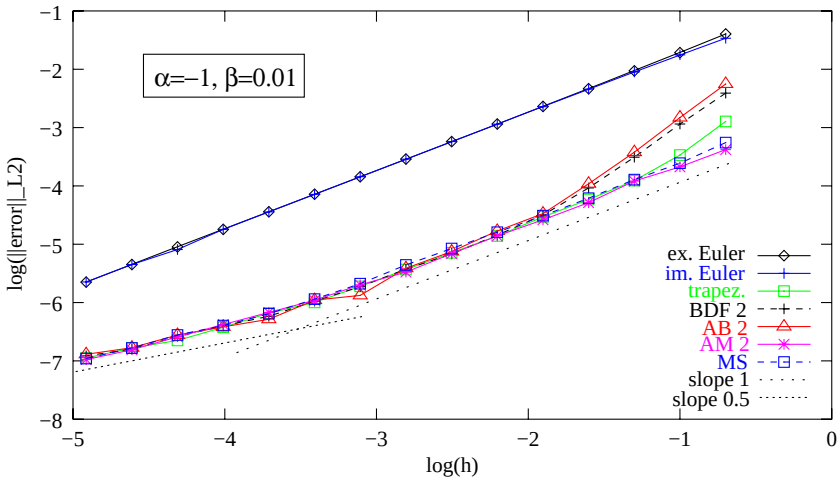


FIG. 2. *Simulation results for the SDE* (5.1) *with the stable schemes of Table* 5.1.

Results of experiments with several methods of Table 5.1 applied to the second example (5.2) are presented in Figures 3 to 5. The second starting value X_1 has been taken as the corresponding value of the reference solution, which amounts to a computation by the trapezoidal rule with very small step sizes. In Figures 3 and 4 we show simulation results with the parameter $\epsilon = 10^{-4}, 10^{-2}$. Again we have plotted the achieved accuracy versus the step size in logarithmic scale. The results are qualitatively the same as for the simple example above. The error of the Euler schemes is larger than that of the other schemes and the error of the schemes with deterministic order 2 (trapezoidal rule, Adams–Bashforth2 and BDF2) appears to be of the size $\max(c_1\beta^2 h^{\frac{1}{2}}, c_2\beta h, c_3 h^2)$, whereas for the Adams–Moulton2 with deterministic order 3 one even observes errors of the size $\max(c_1\beta^2 h^{\frac{1}{2}}, c_2\beta h, c_4 h^3)$.

Figure 5 relates to the case where the parameter in the diffusion coefficient is not small. Here the order $h^{1/2}$ term dominates the error for all chosen step sizes.

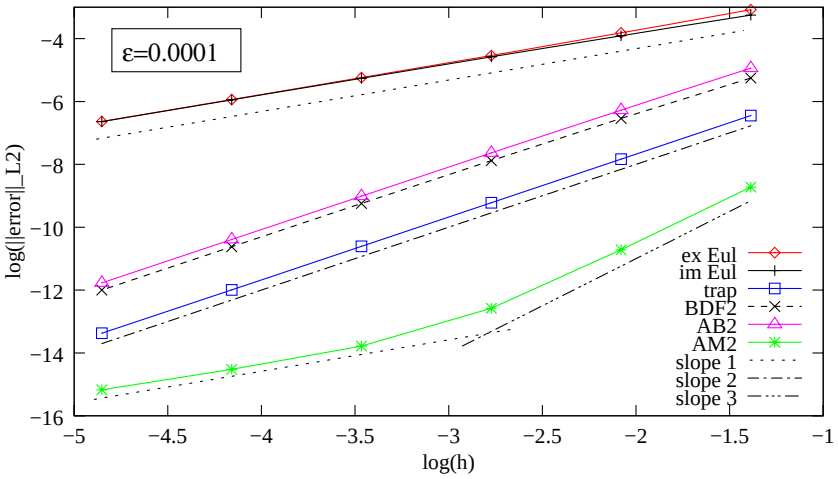


FIG. 3. *Simulation results for the SDE* (5.2), $\epsilon = 10^{-4}$, *with schemes of Table* 5.1.

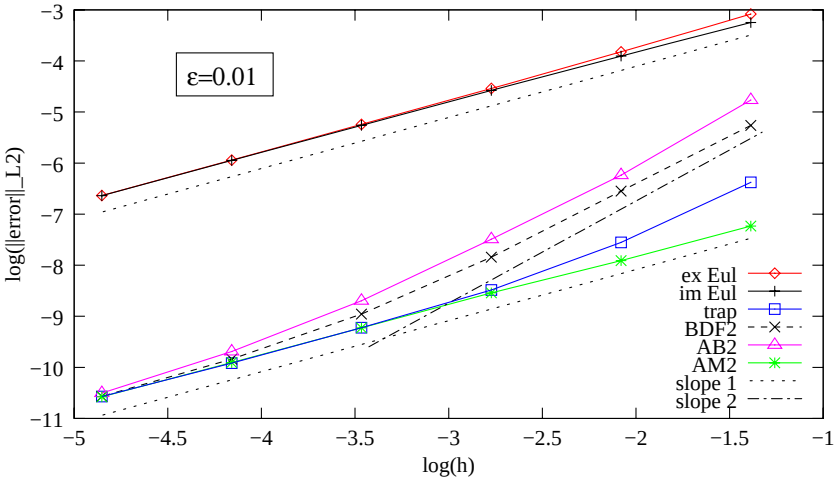


FIG. 4. *Simulation results for the SDE* (5.2), $\epsilon = 10^{-2}$, *with schemes of Table* 5.1.

Furthermore, due to the larger noise the statistical error in the approximation is more visible, which we partly compensated by using 500 paths instead of only 100. In this case one can clearly see that the Euler schemes perform as well as the other schemes and investing the higher effort for the two-step schemes does not pay.

Appendix A. Proof of Theorem 3.2.

Proof. **Necessity:** This part can be proved as in the deterministic case, i.e., we take the equation $X'(t) = 0$, then f and Γ_j satisfy obviously (2.9) and (2.10). We then follow in principle the proof of [15, Thm. 6.3.3].

Sufficiency: Since the SLMM (2.1) contains the stochastic part related to the Γ_j, we cannot rely on the theory of difference equations and the representations of their solutions. Instead, we will follow the route of rewriting the k-step recurrence equation as a one-step recurrence equation in a higher dimensional space (see, e.g., [16, Chap. III.4][26, Chap. 8.2.1]).

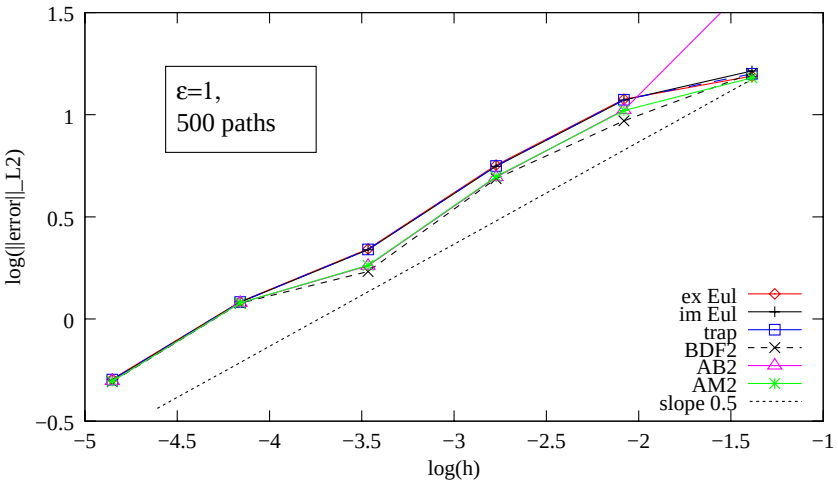


FIG. 5. *Simulation results for the SDE* (5.2), $\epsilon = 1$, *with schemes of Table* 5.1.

For X_ℓ and $\tilde{X}_\ell$ being the solutions of (2.1) and (2.4), respectively, let the n-dimensional vector E_ℓ be defined as the difference $X_\ell - \tilde{X}_\ell$. We have with $E_0, \dots, E_{k-1} \in L_2(\Omega, \mathbb{R}^n)$ for $\ell = k, \dots, N$, the recursion

$$E_\ell = -\sum_{j=1}^{k} \alpha_j \; E_{\ell-j} + \underbrace{h \sum_{j=0}^{k} \beta_j \; \Delta f_{\ell-j}}_{=:\Delta\phi^\ell} + \underbrace{\sum_{j=1}^{k} \Delta\Gamma_{j,\ell-j} \; I^{t_{\ell-j},t_{\ell-j+1}}}_{=:\Delta\psi^\ell} - D_\ell, \tag{A.1}$$

where

$$\Delta f_{\ell-j} := f(X_{\ell-j}, t_{\ell-j}) - f(\tilde{X}_{\ell-j}, t_{\ell-j})$$
$$\Delta\Gamma_{j,\ell-j} := \Gamma_j(X_{\ell-j}, t_{\ell-j}) - \Gamma_j(\tilde{X}_{\ell-j}, t_{\ell-j}).$$

We rearrange this k-step recursion in the space $L_2(\Omega, \mathbb{R}^n)$ to a one-step recursion in $L_2(\Omega, \mathbb{R}^{k\times n})$. Together with the trivial identities $E_{\ell-1} = E_{\ell-1}, \; \dots \; E_{\ell-k+1} = E_{\ell-k+1}$ we obtain

$$\underbrace{\begin{pmatrix} E_\ell \\ E_{\ell-1} \\ \vdots \\ E_{\ell-k+1} \end{pmatrix}}_{=:\,\mathcal{E}_\ell} = \underbrace{\begin{pmatrix} -\alpha_1 I & \cdots & \cdots & -\alpha_k I \\ I & 0 & & \\ & \ddots & \ddots & \\ & & I & 0 \end{pmatrix}}_{=:\,\mathcal{A}} \underbrace{\begin{pmatrix} E_{\ell-1} \\ E_{\ell-2} \\ \vdots \\ E_{\ell-k} \end{pmatrix}}_{=:\,\mathcal{E}_{\ell-1}} + \underbrace{\begin{pmatrix} \Delta\phi^\ell \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{=:\,\Delta\Phi_\ell} + \underbrace{\begin{pmatrix} \Delta\psi^\ell \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{=:\,\Delta\Psi_\ell} + \underbrace{\begin{pmatrix} -D_\ell \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{=:\,\mathcal{D}_\ell}$$

or, in compact form,

$$\mathcal{E}_\ell = \mathcal{A}\mathcal{E}_{\ell-1} + \Delta\Phi_\ell + \Delta\Psi_\ell + \mathcal{D}_\ell \;, \; \ell = k, \dots, N \text{ and } \mathcal{E}_{k-1} = (-D_{k-1}, -D_{k-2}, \dots, -D_0)^T,$$

where $\mathcal{E}_\ell \in L_2(\Omega, \mathbb{R}^{k\times n}), \; \ell = k-1, k, \dots, N$. The vector $\mathcal{E}_{k-1}$ consists of the perturbations to the initial values. We now trace back the recursion in $\mathcal{E}_\ell$ to the initial

vector $\mathcal{E}_{k-1}$. For $\ell = k, \dots, N$ we have

$$\begin{aligned}
\mathcal{E}_\ell &= \mathcal{A}\mathcal{E}_{\ell-1} + \Delta\Phi_\ell + \Delta\Psi_\ell + \mathcal{D}_\ell \\
&= \mathcal{A}(\mathcal{A}\mathcal{E}_{\ell-2} + \Delta\Phi_{\ell-1} + \Delta\Psi_{\ell-1} + \mathcal{D}_{\ell-1}) + \Delta\Phi_\ell + \Delta\Psi_\ell + \mathcal{D}_\ell \\
&= \mathcal{A}^2\mathcal{E}_{\ell-2} + (\Delta\Phi_\ell + \mathcal{A}\Delta\Phi_{\ell-1}) + (\Delta\Psi_\ell + \mathcal{A}\Delta\Psi_{\ell-1}) + (\mathcal{D}_\ell + \mathcal{A}\mathcal{D}_{\ell-1}) \\
&\;\;\vdots \\
&= \mathcal{A}^{\ell-k+1}\mathcal{E}_{k-1} + \sum_{i=0}^{\ell-k} \mathcal{A}^i \Delta\Phi_{\ell-i} + \sum_{i=0}^{\ell-k} \mathcal{A}^i \Delta\Psi_{\ell-i} + \sum_{i=0}^{\ell-k} \mathcal{A}^i \mathcal{D}_{\ell-i} \\
&= \mathcal{A}^{\ell-k+1}\mathcal{E}_{k-1} + \sum_{i=k}^{\ell} \mathcal{A}^{\ell-i} \Delta\Phi_i + \sum_{i=k}^{\ell} \mathcal{A}^{\ell-i} \Delta\Psi_i + \sum_{i=k}^{\ell} \mathcal{A}^{\ell-i} \mathcal{D}_i \,.
\end{aligned}$$

A crucial point for the subsequent calculations is to find a scalar product inducing a matrix norm such that this norm of the matrix $\mathcal{A}$ is less than or equal to 1 (see, e.g., [16, Chap. III.4, Lemma 4.4]. This is possible if the eigenvalues of the Frobenius matrix $\mathcal{A}$ lie inside the unit circle of the complex plane and are simple if their modulus is equal to 1. The eigenvalues of $\mathcal{A}$ are the roots of the characteristic polynomial ρ (2.12) and due to the assumption that Dahlquist's root condition is satisfied they have the required property. Then there exists a nonsingular matrix $\mathcal{C}$ with a block structure like $\mathcal{A}$ such that $\|\mathcal{C}^{-1}\mathcal{A}\mathcal{C}\|_2 \le 1$, where $\|\cdot\|_2$ denotes the spectral matrix norm that is induced by the Euclidian vector norm in $\mathbb{R}^{k\times n}$. We can thus choose a scalar product for $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{k\times n}$ as

$$\langle \mathcal{X}, \mathcal{Y} \rangle_* := \langle \mathcal{C}^{-1}\mathcal{X}, \mathcal{C}^{-1}\mathcal{Y} \rangle_2$$

and then have $|.|_*$ as the induced vector norm on $\mathbb{R}^{k\times n}$ and $\|\cdot\|_*$ as the induced matrix norm with $\|\mathcal{A}\|_* = \|\mathcal{C}^{-1}\mathcal{A}\mathcal{C}\|_2 \le 1$. We also have

$$\langle \mathcal{X}, \mathcal{Y} \rangle_* = \mathcal{X}^T \mathcal{C}^{-T}\mathcal{C}^{-1}\mathcal{Y} = \mathcal{X}^T \mathcal{C}^* \mathcal{Y} \quad \text{with} \quad \mathcal{C}^* = \mathcal{C}^{-T}\mathcal{C}^{-1} = (c^*_{ij} I_n)_{i,j=1,\dots,k}.$$

Due to the norm equivalence there are constants $c^*, c_* > 0$ such that

$$|\mathcal{X}|_2^2 \le c^* |\mathcal{X}|_*^2 \quad \text{and} \quad |\mathcal{X}|_*^2 \le c_* |\mathcal{X}|_\infty^2 \quad \forall \mathcal{X} \in \mathbb{R}^{k\times n} \,,$$

where $|\mathcal{X}|_2^2 = \sum_{j=1,\dots,k} |x_j|^2$, $|\mathcal{X}|_\infty = \max_{j=1,\dots,k} |x_j|$ for $\mathcal{X} = (x_1^T, \dots, x_k^T)^T$. For the special vectors $\mathcal{X} = (x^T, 0, \dots, 0)^T$ and $\mathcal{Y} = (y^T, 0, \dots, 0)^T$ with $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{k\times n}$ and $x, y \in \mathbb{R}^n$, one has $\langle \mathcal{X}, \mathcal{Y} \rangle_* = c^*_{11} \langle x, y \rangle_2 = c^*_{11} x^T y$, where c^*_{11} is given by the matrix $\mathcal{C}^*$.

We now apply $|.|_*^2$ to estimate $|\mathcal{E}_\ell|_*^2$ and, later, $\mathbb{E}|\mathcal{E}_\ell|_*^2$. We start with

$$|\mathcal{E}_\ell|_*^2 \le 4 \left\{ \underbrace{|\mathcal{A}^{\ell-k+1}\mathcal{E}_{k-1}|_*^2}_{(1)} + \underbrace{|\sum_{i=k}^{\ell} \mathcal{A}^{\ell-i}\Delta\Phi_i|_*^2}_{(2)} + \underbrace{|\sum_{i=k}^{\ell} \mathcal{A}^{\ell-i}\Delta\Psi_i|_*^2}_{(3)} + \underbrace{|\sum_{i=k}^{\ell} \mathcal{A}^{\ell-i}\mathcal{D}_i|_*^2}_{(4)} \right\}.$$

For the term labeled (1) we have $|\mathcal{A}^{\ell-k+1}\mathcal{E}_{k-1}|_*^2 \le |\mathcal{E}_{k-1}|_*^2$, and thus

$$\mathbb{E}|\mathcal{A}^{\ell-k+1}\mathcal{E}_{k-1}|_*^2 \le \mathbb{E}|\mathcal{E}_{k-1}|_*^2 \,. \tag{A.2}$$

For the term labeled (2) we have

$$\left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\Delta\Phi_i\right|_*^2 \le (\ell-k+1)\sum_{i=k}^{\ell}|\mathcal{A}^{\ell-i}\Delta\Phi_i|_*^2 \;\le\; N\sum_{i=k}^{\ell}|\Delta\Phi_i|_*^2 = \frac{T}{h}c_{11}^*\sum_{i=k}^{\ell}|\Delta\phi^i|^2$$
$$\le h\ T\ c_{11}^*\ (k{+}1)\sum_{i=k}^{\ell}\sum_{j=0}^{k}|\beta_j\ \Delta f_{i-j}|^2 \le h\ T\ c_{11}^*\ (k{+}1)\ L_f^2\sum_{i=k}^{\ell}\sum_{j=0}^{k}\beta_j^2\ |E_{i-j}|^2$$
$$\le h\ T\ c_{11}^*\ (k{+}1)\ L_f^2\left\{\beta_0^2\ |E_\ell|^2+\sum_{i=k}^{\ell}\left\{\beta_0^2\ |E_{i-1}|^2+\sum_{j=1}^{k}\beta_j^2\ |E_{i-j}|^2\right\}\right\}$$
$$\le h\ T\ c_{11}^*\ (k{+}1)\ L_f^2\left\{c^*\ \beta_0^2\ |\mathcal{E}_\ell|_*^2+C_\beta\ c^*\sum_{i=k-1}^{\ell-1}|\mathcal{E}_i|_*^2\right\},$$

where $C_\beta = 2\ \max_{j=0,\dots,k}\beta_j$. Hence,

$$\text{(A.3)}\quad \mathbb{E}|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\Delta\Phi_i|_*^2 \le h\ T\ c_{11}^*\ (k{+}1)\ L_f^2\ \left\{c^*\ \beta_0^2\ \mathbb{E}|\mathcal{E}_\ell|_*^2+C_\beta\ c^*\sum_{i=k-1}^{\ell-1}\mathbb{E}|\mathcal{E}_i|_*^2\right\}.$$

We will now treat the term labeled (3). For that purpose we introduce the notation $\Delta\Psi_{j,i-j} := ((\Delta\Gamma_{j,i-j}I^{t_{i-j},t_{i-j+1}})^T,0,\dots,0)^T$. Using this we can write

$$\Delta\Psi_i = ((\Delta\psi^i)^T,0,\dots,0)^T = \left(\left(\sum_{j=1}^{k}\Delta\Gamma_{j,i-j}\ I^{t_{i-j},t_{i-j+1}}\right)^T,0,\dots,0\right)^T = \sum_{j=1}^{k}\Delta\Psi_{j,i-j}$$

and

$$\left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\Delta\Psi_i\right|_*^2 \;=\; \left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\sum_{j=1}^{k}\Delta\Psi_{j,i-j}\right|_*^2.$$

Every $\Delta\Psi_{j,i-j}$ is $\mathcal{F}_{t_{i-j+1}}$-measurable and $\mathbb{E}(\Delta\Psi_{j,i-j}|\mathcal{F}_{t_{i-j}}) = 0$. We can now reorder the last term above such that we have a sum of terms where each term contains all multiple Wiener integrals over just one subinterval. The expectation of

products of terms from different subintervals vanishes; hence we obtain

$$
\begin{aligned}
&\mathbb{E}\left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\Delta\Psi_i\right|_*^2\\
&=\mathbb{E}|\mathcal{A}^{\ell-k}\Delta\Psi_{k,0}|_*^2\\
&\quad+\mathbb{E}|\mathcal{A}^{\ell-k-1}\Delta\Psi_{k,1}+\mathcal{A}^{\ell-k}\Delta\Psi_{k-1,1}|_*^2\\
&\quad\vdots\\
&\quad\quad+\mathbb{E}|\mathcal{A}^{\ell-2k+1}\Delta\Psi_{k,k-1}+\mathcal{A}^{\ell-2k+2}\Delta\Psi_{k-1,k-1}+\ldots+\mathcal{A}^{\ell-k}\Delta\Psi_{1,k-1}|_*^2\\
&\quad\quad\vdots\\
&\quad\quad+\mathbb{E}|\mathcal{A}^{0}\Delta\Psi_{k,\ell-k}+\mathcal{A}^{1}\Delta\Psi_{k-1,\ell-k}+\ldots+\mathcal{A}^{k-1}\Delta\Psi_{1,\ell-k}|_*^2\\
&\quad\vdots\\
&\quad+\mathbb{E}|\mathcal{A}^{0}\Delta\Psi_{2,\ell-2}+\mathcal{A}^{1}\Delta\Psi_{1,\ell-2}|_*^2\\
&\quad+\mathbb{E}|\mathcal{A}^{0}\Delta\Psi_{1,\ell-1}|_*^2\\
&\le k\sum_{i=k}^{\ell}\sum_{j=1}^{k}\mathbb{E}|\Delta\Psi_{j,i-j}|_*^2 \qquad \le k\ c_{11}^*\sum_{i=k}^{\ell}\sum_{j=1}^{k}\mathbb{E}\|\Delta\Gamma_{j,i-j}\|^2\ \mathbb{E}|I^{t_{i-j},t_{i-j+1}}|^2\\
&\le h\ k\ c_{11}^*\ L_\Gamma^2\sum_{i=k}^{\ell}\sum_{j=1}^{k}\mathbb{E}|E_{i-j}|^2 \ \le\ h\ k\ c_{11}^*\ L_\Gamma^2\ c^*\sum_{i=k}^{\ell}|\mathcal{E}_{i-1}|_*^2.
\end{aligned}
$$

Thus, for the term labeled (3), we obtain

$$
\mathbb{E}|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\Delta\Psi_i|_*^2\le\ h\ k\ c_{11}^*\ L_\Gamma^2\ c^*\sum_{i=k-1}^{\ell-1}|\mathcal{E}_i|_*^2. \tag{A.4}
$$

We will, for a shorter notation, deal with the term labeled (4), i.e., the perturbations D_i in $\mathcal{D}_i$, after obtaining an intermediate result. Using (A.2), (A.3), and (A.4), and setting $L_0:=L_f^2\ (k+1)\ c_{11}^*\ T\ c^*\ \beta_0^2$ and $L:=L_f^2\ (k+1)\ c_{11}^*\ T\ c_\beta^*\ +\ L_\Gamma^2\ k\ c_{11}^*\ c^*$, we have now arrived at

$$
\mathbb{E}|\mathcal{E}_\ell|_*^2\le 4\left\{\mathbb{E}|\mathcal{E}_{k-1}|_*^2+hL_0\mathbb{E}|\mathcal{E}_\ell|^2+hL\sum_{i=k-1}^{\ell-1}\mathbb{E}|\mathcal{E}_i|_*^2+\mathbb{E}|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\mathcal{D}_i|_*^2\right\},\ \ell=k,\ldots,N.
$$

If necessary, we choose a bound h^0 on the step size such that $4\ h\ L_0<\frac{1}{2}$ holds for all $h<h^0$ and conclude that

$$
\mathbb{E}|\mathcal{E}_\ell|_*^2\le 8\ \mathbb{E}|\mathcal{E}_{k-1}|_*^2+8\ \mathbb{E}\left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\mathcal{D}_i\right|_*^2+8\ L\ T\ \frac{1}{N}\sum_{i=k-1}^{\ell-1}\mathbb{E}|\mathcal{E}_i|_*^2.
$$

We now apply Lemma 2.11 with $a_\ell:=0,\ \ell=1,\ldots,k-2$, and $a_\ell:=\mathbb{E}|\mathcal{E}_\ell|_*^2,\ \ell=k-1,\ldots,N$, and obtain the intermediate result

$$
\max_{\ell=k-1,\ldots,N}\mathbb{E}|\mathcal{E}_\ell|_*^2\le\hat{S}\left\{\mathbb{E}|\mathcal{E}_{k-1}|_*^2+\max_{\ell=k,\ldots,N}\mathbb{E}\left|\sum_{i=k}^{\ell}\mathcal{A}^{\ell-i}\mathcal{D}_i\right|_*^2\right\},\ \hat{S}:=8\exp(8LT)\ . \tag{A.5}
$$

It remains to deal with the term labeled (4), i.e., the perturbations D_i in $\mathcal{D}_i$. We decompose D_i, and, analogously, $\mathcal{D}_i$, into $D_i = R_i + S_i = R_i + \sum_{j=1}^{k} S_{j,i-j+1}$, $\mathcal{D}_i = \mathcal{R}_i + \mathcal{S}_i = \mathcal{R}_i + \sum_{j=1}^{k} \mathcal{S}_{j,i-j+1}$, where $S_{j,i-j+1}$ is $\mathcal{F}_{t_{i-j+1}}$-measurable with $\mathbb{E}(S_{j,i-j+1}|\mathcal{F}_{t_{i-j}}) = 0$ for $i = k, \dots, N$ and $j = 1, \dots, k$. Then $\mathbb{E}\langle \mathcal{A}^{\ell_1}\mathcal{S}_{j_1,i_1}, \mathcal{A}^{\ell_2}\mathcal{S}_{j_2,i_2}\rangle_* = 0$ for $i_1 \neq i_2$, and by similar computations as above, we obtain

$$\mathbb{E}\left|\sum_{i=k}^{\ell} \mathcal{A}^{\ell-i}\mathcal{D}_i\right|_*^2 \leq 2\,(\ell-k+1)\sum_{i=k}^{\ell}\mathbb{E}|\mathcal{A}^{\ell-i}\mathcal{R}_i|_*^2 + 2\,k\,\sum_{i=k}^{\ell}\sum_{j=1}^{k}\mathbb{E}|\mathcal{A}^{\ell-i}\mathcal{S}_{j,i-j+1}|_*^2$$
$$\leq 2\sum_{i=k}^{\ell}\left(\frac{T}{h}\,\mathbb{E}|\mathcal{R}_i|_*^2 + k\sum_{j=1}^{k}\mathbb{E}|\mathcal{S}_{j,i-j+1}|_*^2\right).$$

Inserting this into the intermediate result (A.5) we obtain

$$\max_{\ell=k-1,\dots,N}\mathbb{E}|\mathcal{E}_\ell|_*^2 \leq \hat{S}\left\{\mathbb{E}|\mathcal{E}_{k-1}|_*^2 + 2\sum_{i=k}^{\ell}\left(\frac{T}{h}\,\mathbb{E}|\mathcal{R}_i|_*^2 + k\sum_{j=1}^{k}\mathbb{E}|\mathcal{S}_{j,i-j+1}|_*^2\right)\right\},$$

and

$$\max_{\ell=k-1,\dots,N}\mathbb{E}|E_\ell|^2 \leq c^*\hat{S}\left\{c_*\max_{\ell=0,\dots,k-1}\mathbb{E}|E_\ell|^2 + 2\,c_{11}^*\max_{\ell=k,\dots,N}\left(\frac{T^2}{h^2}\mathbb{E}|R_\ell|^2 + \frac{kT}{h}\sum_{j=1}^{k}\mathbb{E}|S_{j,\ell-j+1}|^2\right)\right\}.$$

Taking the square root yields the final estimate

$$\max_{\ell=k-1,\dots,N}\|E_\ell\|_{L_2}$$
$$\leq \sqrt{c^*\hat{S}}\left\{\sqrt{c_*}\max_{\ell=0,\dots,k-1}\|E_\ell\|_{L_2} + \sqrt{2c_{11}^*}\max_{\ell=1,\dots,N}\left(\frac{T}{h}\|R_\ell\|_{L_2} + \sqrt{\frac{kT}{h}\sum_{j=1}^{k}\|S_{j,\ell-j+1}\|_{L_2}^2}\right)\right\}$$
$$\leq S\left\{\max_{\ell=0,\dots,k-1}\|E_\ell\|_{L_2} + \max_{\ell=1,\dots,N}\left(\frac{\|R_\ell\|_{L_2}}{h} + \frac{\sqrt{\sum_{j=1}^{k}\|S_{j,\ell-j+1}\|_{L_2}^2}}{\sqrt{h}}\right)\right\},$$

which completes the proof. □

Acknowledgments. We thank Tony Shardlow for careful reading of the manuscript and anonymous referees for their comments which helped to improve the presentation of the material.

REFERENCES

[1] C. Baker and E. Buckwar, *Numerical analysis of explicit one-step methods for stochastic delay differential equations*, LMS J. Comput. Math., 3 (2000), pp. 315–335.

[2] R. H. Bokor, *On two-step methods for stochastic differential equations*, Acta Cybernet., 13 (1997), pp. 197–207.

[3] R. H. Bokor, *Convergence and stability properties for numerical approximations of stochastic ordinary differential equations*, Ph.D. thesis, University of Zagreb, Yugoslavia, 2000.

[4] R. H. Bokor, *Stochastically stable one-step approximations of solutions of stochastic ordinary differential equations*, J. Appl. Numer. Math., 44 (2003), pp. 299–312.

[5] L. Brugnano, K. Burrage, and P. Burrage, *Adams-type methods for the numerical solution of stochastic ordinary differential equations*, BIT, 40 (2000), pp. 451–470.

[6] E. Buckwar and R. Winkler, *On two-step schemes for SDEs with small noise*, PAMM, 4 (2004), pp. 15–18.
[7] E. Buckwar and R. Winkler, *Improved linear multi-step methods for stochastic ordinary differential equations*, J. Comput. Appl. Math., to appear.
[8] K. Burrage and P. M. Burrage, *General order conditions for stochastic Runge-Kutta methods for both commuting and non-commuting stochastic ordinary differential equation systems.*, Appl. Numer. Math., 28 (1998), pp. 161–177.
[9] P. Burrage, *Runge-Kutta methods for stochastic differential equations*, Ph.D. thesis, University of Queensland, Brisbane, Australia, 1999.
[10] G. Dahlquist, *33 years of numerical instability, Part I*, BIT, 25 (1985), pp. 188–204.
[11] G. Denk and S. Schäffler, *Adams methods for the efficient solution of stochastic differential equations with additive noise*, Computing, 59 (1996), pp. 153–161.
[12] G. Denk and R. Winkler, *Modeling and simulation of transient noise in circuit simulation*, Math. Comput. Modelling, to appear.
[13] B. D. Ewald and R. Témam, *Numerical analysis of stochastic schemes in geophysics*, SIAM J. Numer. Anal., 42 (2005), pp. 2257–2276.
[14] T. C. Gard, *Introduction to Stochastic Differential Equations*, Marcel-Dekker, New York, 1988.
[15] W. Gautschi, *Numerical Analysis. An Introduction*, Birkhäuser, Boston, 1997.
[16] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations.* I: *Nonstiff Problems,* 2nd rev. ed., Springer Ser. Comput. Math. 8, Springer, New York, 1993.
[17] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.
[18] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1992.
[19] J. Lambert, *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*, John Wiley & Sons, Chichester, 1991.
[20] P. Lax and R. Richtmyer, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), pp. 267–293.
[21] G. Milstein and M. Tretyakov, *Mean-square numerical methods for stochastic differential equations with small noise*, SIAM J. Sci. Comput., 18 (1997), pp. 1067–1087.
[22] G. Milstein and M. Tretyakov, *Stochastic Numerics for Mathematical Physics*, Springer-Verlag, Berlin, 2004.
[23] G. Milstein, *Theorem on the order of convergence for mean-square approximations of solutions of stochastic differential equations*, Theory Probab. Appl., 32 (1987), pp. 738–741.
[24] G. Milstein, *Numerical Integration of Stochastic Differential Equations*, Kluwer, Norwell, MA, 1995.
[25] C. Penski, *A new numerical method for sdes and its application in circuit simulation*, J. Comput. Appl. Math., 115 (2000), pp. 461–470.
[26] R. Plato, *Numerische Mathematik Kompakt. Grundlagenwissen für Studium und Praxis*, Vieweg, Braunschweig, Germany, 2000.
[27] W. Römisch and R. Winkler, *Stochastic DAEs in circuit simulation*, in Modeling, Simulation and Optimization of Integrated Circuits, A. G. K. Antreich, R. Bulirsch and P. Rentrop, eds., Birkhäuser, Basel, Switzerland, 2003, pp. 303–318.
[28] T. Rydén and M. Wiktorsson, *On the simulation of iterated Itô integrals*, Stochastic Process. Appl., 91 (2001), pp. 151–168.
[29] M. Wiktorsson, *Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions*, Ann. Appl. Probab., 11 (2001), pp. 470–487.
[30] R. Winkler, *Stochastic differential algebraic equations of index 1 and applications in circuit simulation*, J. Comput. Appl. Math., 157 (2003), pp. 477–505.
[31] R. Winkler, *Stochastic DAEs in transient noise simulation*, in Proceedings of Scientific Computing in Electrical Engineering, Eindhoven Springer Series Mathematics in Industry, Springer, Berlin, 2004, pp. 408–415.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 804–825

ON A LINEARIZED BACKWARD EULER METHOD FOR THE EQUATIONS OF MOTION OF OLDROYD FLUIDS OF ORDER ONE*

AMIYA K. PANI†, JIN YUN YUAN‡, AND PEDRO D. DAMÁZIO‡§

Abstract. In this paper, a linearized backward Euler method is discussed for the equations of motion arising in the Oldroyd model of viscoelastic fluids. Some new a priori bounds are obtained for the solution under realistically assumed conditions on the data. Further, the exponential decay properties for the exact as well as the discrete solutions are established. Finally, a priori error estimates in $\mathbf{H}^1$ and $\mathbf{L}^2$-norms are derived for the the discrete problem which are valid uniformly for all time $t > 0$.

Key words. viscoelastic fluids, Oldroyd model, a priori bounds, exponential decay, linearized backward Euler method, uniform convergence in time

AMS subject classifications. 35L70, 65M30, 76D05, 78A10

DOI. 10.1137/S0036142903428967

1. Introduction.

The motion of an incompressible fluid in a bounded domain Ω in $\mathbb{R}^2$ is described by the system of partial differential equations

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \nabla \cdot \boldsymbol{\sigma} + \nabla p = \mathbf{F}(x,t), \quad x \in \Omega, \ t > 0,$$

$$\nabla \cdot \mathbf{u} = 0, \quad x \in \Omega, \ t > 0,$$

with appropriate initial and boundary conditions. Here, $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_{ik})$ denotes the stress tensor with $tr\boldsymbol{\sigma} = 0$, $\mathbf{u}$ represents the velocity vector, p is the pressure of the fluid, and $\mathbf{F}$ is the external force. The defining relation between the stress tensor $\boldsymbol{\sigma}$ and the rate of deformation tensor $\mathbf{D} = (\mathbf{D}_{ik}) = \frac{1}{2}(\mathbf{u}_{ix_k} + \mathbf{u}_{kx_i})$, called the equation of state or sometimes the rheological equation, in fact, establishes the type of fluids under consideration. When $\boldsymbol{\sigma} = 2\nu\mathbf{D}$ (using Newton's law) with ν the kinematic coefficient of viscosity, we obtain Newton's model of incompressible viscous fluid and the corresponding system is widely known as Navier–Stokes equations. This has been a basic model for describing the flow at moderate velocities of the majority of the incompressible viscous fluids encountered in practice. However, there are many fluids with complex microstructure, such as biological fluids, polymeric fluids, suspensions, and liquid crystals, which are used in the current industrial processes and show (nonlinear) viscoelastic behavior that cannot be described by the classical linear viscous Newtonian models. The deparature from the Navier–Stokes behavior manifests itself in a variety of ways, such as non-Newtonian viscosity, stress relaxation, and nonlinear creeping. The model of rate type such as Oldroyd fluids (see [4], [23], [32]) can predict the stress relaxation as well as the retardation of deformation and, therefore,

*Received by the editors June 1, 2003; accepted for publication (in revised form) November 15, 2005; published electronically April 12, 2006.
http://www.siam.org/journals/sinum/44-2/42896.html
†Department of Mathematics, Industrial Mathematics Group, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India (akp@math.iitb.ac.in).
‡ Department of Mathematics, Federal University of Paraná, Centro Politécnico, Curitiba, Cx.P: 19081, CEP: 81531-990, PR, Brazil (jin@mat.ufpr.br, pedro@mat.ufpr.br).
§Isaac Newton Institute for Mathematical Sciences Preprint Series NI03028-CPD, Cambridge, UK.

have become popular for describing polymeric suspension. In order to model the behavior of a dilute polymer solution in a Newtonian solvent, the extra stress tensor is often split into two components: a viscoelastic one and a purely viscous one. So the Oldroyd fluids of order one as it is known in the Russian literature (see [23], [2], [18]) are described by the defining relation

$$\left(1+\lambda\frac{\partial}{\partial t}\right)\boldsymbol{\sigma}=2\nu\left(1+\kappa\nu^{-1}\frac{\partial}{\partial t}\right)\mathbf{D},$$

where λ, ν, κ are positive constants with $(\nu-\kappa\lambda^{-1})>0$. Here, ν denotes the kinematic viscocity, λ is the relaxation time, and κ represents the retardation time. In the form of an integral equation, we write the above defining relation as

$$\begin{aligned}\boldsymbol{\sigma}(x,t)&=2\kappa\lambda^{-1}\mathbf{D}(x,t)+2\lambda^{-1}(\nu-\kappa\lambda^{-1})\int_0^t\exp(-\lambda^{-1}(t-\tau))\mathbf{D}(x,\tau)\,d\tau\\&\quad+(\boldsymbol{\sigma}(x,0)-2\kappa\lambda^{-1}\mathbf{D}(x,0))\exp(-\lambda^{-1}t).\end{aligned}$$

Now the equation of motion of the Oldroyd fluids of order one can be described most naturally by the system of integrodifferential equations

$$\frac{\partial\mathbf{u}}{\partial t}+\mathbf{u}\cdot\nabla\mathbf{u}-\mu\Delta\mathbf{u}-\int_0^t\beta(t-\tau)\Delta\mathbf{u}(x,\tau)\,d\tau+\nabla p=\mathbf{f},\ \ x\in\Omega,\ t>0, \tag{1.1}$$

and incompressibility condition

$$\nabla\cdot\mathbf{u}=0,\ \ x\in\Omega,\, t>0, \tag{1.2}$$

with initial and boundary conditions

$$\mathbf{u}(x,0)=\mathbf{u}_0,\quad x\in\Omega,\quad\text{and}\quad\mathbf{u}(x,t)=0,\quad x\in\partial\Omega,\quad t\geq 0. \tag{1.3}$$

Here, Ω is a bounded domain in two-dimensional Euclidean space $\mathbb{R}^2$ with smooth boundary $\partial\Omega$, $\mu=\kappa\lambda^{-1}>0$ and the kernel $\beta(t)=\gamma\exp(-\delta t)$, where $\gamma=\lambda^{-1}(\nu-\kappa\lambda^{-1})$ and $\delta=\lambda^{-1}$. For details of the physical background and its mathematical modeling, see [4], [17], [23], [24], and [32].

Throughout this paper, we shall assume that $\mu=1$ and the nonhomogeneous term $\mathbf{f}=0$. In fact, assuming conservative force, the function $\mathbf{f}$ can be absorbed in the pressure term.

As in Temam [28], we recast the above problem (1.1)–(1.3) as an abstract evolution equation in an appropriate function space setting. Let us denote by $H^m(\Omega)$ the standard Hilbert–Sobolev space and by $\|\cdot\|_m$ the norm defined on it. When $m=0$, we call $H^0(\Omega)$ as the space of square integrable functions $L^2(\Omega)$ with the usual norm $\|\cdot\|$ and inner product $(\cdot,\cdot)$. Further, let $H_0^1(\Omega)$ be the completion of $C_0^\infty(\Omega)$ with respect to $H^1(\Omega)$-norm. In fact, the seminorm $\|\nabla\phi\|$ on $H_0^1(\Omega)$ is a norm and is equivalent to H^1-norm. We also use the following function spaces for the vector valued functions.

Define

$$\mathbf{D}(\Omega):=\{\boldsymbol{\phi}\in(C_0^\infty(\Omega))^2:\ \ \nabla\cdot\boldsymbol{\phi}=0\ \text{ in }\Omega\},$$

$$\mathbf{H}:=\text{the closure of }\mathbf{D}(\Omega)\text{ in }\ (L^2(\Omega))^2-\text{space},$$

and

$$\mathbf{V}:=\text{the closure of }\mathbf{D}(\Omega)\text{ in }\ (H_0^1(\Omega))^2-\text{space}.$$

Note that under some smoothness assumptions on the boundary $\partial\Omega$, it is possible to characterize $\mathbf{V}$ as

$$\mathbf{V} := \{\boldsymbol{\phi} \in (H_0^1)^2 : \quad \nabla \cdot \boldsymbol{\phi} = 0 \ \text{ in } \ \Omega\}.$$

The spaces of vector functions are indicated by boldface letters, for instance, $\mathbf{H}_0^1 = (H_0^1(\Omega))^2$. The inner product on $\mathbf{H}_0^1$ is denoted by

$$(\nabla\boldsymbol{\phi}, \nabla\mathbf{w}) = \sum_{i=1}^{2} (\nabla\phi_i, \nabla w_i)$$

and the norm by

$$\|\nabla\boldsymbol{\phi}\| = \left(\sum_{i=1}^{2} \|\nabla\phi_i\|^2\right)^{\frac{1}{2}}.$$

Using the Poincaré inequality, it can be shown that the norm on $\mathbf{H}_0^1$ is equivalent to $\mathbf{H}^1 = (H^1(\Omega))^2$- norm. Let $\mathbf{P}$ denote the orthogonal projection of $\mathbf{L}^2(\Omega)$ $(= (L^2(\Omega))^2)$ onto $\mathbf{H}$. Now the orthogonal complement $\mathbf{V}^\perp$ of $\mathbf{V}$ in $\mathbf{L}^2(\Omega)$ consists of functions $\boldsymbol{\phi}$ such that $\boldsymbol{\phi} = \nabla p$ for some $p \in H^1(\Omega)/\mathbb{R}$. We define the Stokes operator $A\mathbf{v} = -\mathbf{P}\Delta\mathbf{v}, \quad \mathbf{v} \in D(A) = \mathbf{H}^2 \bigcap \mathbf{V}$. The Stokes operator is a closed linear self-adjoint and positive operator on $\mathbf{H}$ with densely defined domain $D(A)$ in $\mathbf{H}$. Note that its inverse is compact in $\mathbf{H}$; see [28]. Moreover, we set the sth power of A as A^s for every $s \in \mathbb{R}$. For $0 \le s \le 2$, $D(A^{s/2})$ is a Hilbert space with the inner product $(A^{s/2}\mathbf{v}, A^{s/2}\mathbf{w})$ and norm $\|A^{s/2}\mathbf{v}\| := (A^{s/2}\mathbf{v}, A^{s/2}\mathbf{v})^{1/2}$. For $\mathbf{v} \in D(A^{s/2}), \ \ 0 \le s \le 2$, we note that $\|\mathbf{v}\|_s$ and $\|A^{s/2}\mathbf{v}\|$ are equivalent. We also define a bilinear operator $B(\mathbf{u}, \mathbf{v}) = \mathbf{P}((\mathbf{u} \cdot \nabla)\mathbf{v}), \ \mathbf{u}, \mathbf{v} \in \mathbf{V}$.

With the notations described above, we now rewrite the problem (1.1)–(1.3) in its abstract form as follows.

Find $\mathbf{u}(t) \in D(A)$ such that for $t \ge 0$

$$\frac{d\mathbf{u}}{dt}(t) + A\mathbf{u}(t) + B(\mathbf{u}(t), \mathbf{u}(t)) + \int_0^t \beta(t-s)A\mathbf{u}(s)\,ds = 0, \quad t > 0, \tag{1.4}$$
$$\mathbf{u}(0) = \mathbf{u}_0.$$

In an Oldroyd fluid, the stresses after instantaneous cessation of the motion decay like $\exp(-\lambda^{-1}t)$, while the velocities of the flow after instantaneous removal of the stresses die out like $\exp(-\kappa^{-1}t)$. Therefore, it is of interest to discuss the exponential decay property of the solution of (1.4), and we derive these results in section 2. For some related studies in the decay of solution of the linear parabolic equations with memory, see [30] and [3].

The main focus of this paper is to discuss the linearized backward Euler method for time discretization of the system of equations (1.4). For the temporal discretization of the above abstract problem (1.4), let k denote the time step and $t_n = nk$. For smooth function ϕ defined on $[0, \infty)$, set $\phi^n = \phi(t_n)$ and $\bar{\partial}_t\phi^n = (\phi^n - \phi^{n-1})/k$. For the integral term, we apply the right rectangle rule as

$$q^n(\phi) = k\sum_{j=1}^{n} \beta_{n-j}\phi^j \approx \int_0^{t_n} \beta(t_n - s)\phi(s)\,ds, \tag{1.5}$$

where $\beta_{n-j} = \beta(t_n - t_j)$.

Now the linearized version of the backward Euler method applied to the problem (1.4) determines a sequence of functions $\{\mathbf{U}^n\}_{n\geq 0} \subset D(A)$ as solutions of

$$\bar{\partial}_t \mathbf{U}^n + A\mathbf{U}^n + B(\mathbf{U}^{n-1}, \mathbf{U}^n) + q^n(A\mathbf{U}) = 0, \quad n > 0, \tag{1.6}$$
$$\mathbf{U}^0 = \mathbf{u}_0.$$

The main objective of this paper is to derive the following result.

THEOREM 1. *Let $\mathbf{u}_0 \in D(A)$ and let $\mathbf{U}^n$ satisfy* (1.6). *Then there is a positive constant C independent of k but that may depend on $\|\mathbf{u}_0\|_2$ and Ω such that for some $k_0 > 0$ with $0 < k < k_0$ and for positive α with $0 < \alpha < \min(\delta, \lambda_1)$*

$$\|\mathbf{u}(t_n) - \mathbf{U}^n\|_1 \leq C(\|\mathbf{u}_0\|_2) e^{-\alpha t_n} k \left(t_n^{-1/2} + \log \frac{1}{k} \right),$$

where λ_1 is the least eigenvalue of the Stokes operator A.

Once Theorem 1 is proved, the proof of the following theorem becomes routine work. However, we shall indicate only the major steps without proving it in detail in the end of section 3.

THEOREM 2. *Under the assumptions of Theorem* 1, *there is a positive constant C independent of k but that may depend on $\|\mathbf{u}_0\|_2$ and Ω such that for some $k_0 > 0$ with $0 < k < k_0$ and $0 < \alpha < \min(\delta, \lambda_1)$*

$$\|\mathbf{u}(t_n) - \mathbf{U}^n\| \leq C(\|\mathbf{u}_0\|_2) e^{-\alpha t_n} k.$$

Based on the analysis of Ladyzenskaya [20] for the solvability of the Navier–Stokes equations, Oskolkov [24] proved the global existence of unique "almost" classical solutions in finite time interval for the initial and boundary value problem (1.1)–(1.3). The invesigations on solvability were further continued by the coworkers of Oskolkov [19] and Agranovich and Sobolevskii [1] under various sufficient conditions. In these articles, the regularity results are proved which are, in principle, based on some nonlocal compatibility conditions for the data at $t = 0$. Note that these compatibility conditions are either hard to verify or difficult to meet in practice. In case of Navier–Stokes equations, we refer to Heywood and Rannacher [14] for a similar kind of nonlocal conditions. In the present article, we have obtain some new a priori bounds for the solutions of (1.4) under realistically assumed conditions on the initial data. Recently, Sobolevskii [27] discussed the long-time behavior of solution under some stabilizing conditions on the nonhomogeneous forcing function using a combination of energy arguments and semigroup theoretic approach. When the forcing function is zero, we have derived, in sections 2 and 3, the exponential decay properties for the exact solution as well as for the discrete solution using only energy arguments.

For earlier works on the numerical approximations to the solutions of the problem (1.1)–(1.3), see [2] and [5]. While Akhmatov and Oskolkov [2] applied a finite difference scheme to the equation of motion arising in the Oldroyd model, Cannon et al. [5] analyzed a modified nonlinear Galerkin scheme for a periodic problem using spectral Galerkin procedure and discussed the rates of convergence for the semidiscrete approximations. Recently, Pani and Yuan [26] and He et al. [12] applied finite element methods to discretize the spatial variables and derived optimal error estimates for the problems (1.1)–(1.3) without using nonlocal compatibility conditions. In all these pappers [5], [26], [12], only semidiscrete approximations are discussed keeping the time variable continuous. In this article, we have proposed and analyzed a time discretization scheme based on linearized modification of the backward Euler

method. Note that the results on higher order time discretization can easily be proved under the assumption that the exact solutions are sufficiently smooth when t is near 0. These regularity results as we have mentioned earlier entail nonlocal compatibility conditions for the initial data which cannot be verified in practice. Recently, in the context of Oldroyd B fluid, which is a generalization of Oldroyd fluid of order one, a second order Crank–Nicolson scheme [8] is used for the temporal discetization in conjuction with the finite element methods for spatial discretization under regularity requirements on the solutions which cannot be realistically assumed. Therefore, an attempt has been made in this paper to discuss the error estimates for the linearized modified backward Euler scheme (1.6) applied to (1.4) under realistically assumed conditions on the initial data. Finally, in section 4, we conclude with a summary and possible extensions.

The approach of the present article is influenced by the earlier results of Fujita and Mizutani [10], Thomée [29], and references therein on the approximation of semigroups for the parabolic problems; Okamoto [22] on the spatial discretization and Geveci [11] on the time discretization of the Navier–Stokes equations; and Thomée and Zhang [31] for the time discretization of the linear parabolic integrodifferential equations with nonsmooth initial data.

2. Some a priori estimates. For our future use, we make use of the positive definite property (see [21], for a definition) of the kernel β of the integral operator in (1.1). This can be seen as a consequence of the following lemma. For a proof, see Sobolevskii [27, p. 1601] and McLean and Thomeé [21].

LEMMA 3. *For arbitrary $\alpha > 0$, $t^* > 0$, and $\phi \in L^2(0, t^*)$, the following positive definite property holds:*

$$\int_0^{t^*} \left(\int_0^t \exp\left[-\alpha(t-s)\right] \phi(s)\, ds \right) \phi(t)\, dt \geq 0.$$

Since $\beta(t) = \gamma e^{-\delta t}$ with $\gamma > 0$, therefore, the above result is true for $\beta(t)$.

Below, we discuss some a priori bounds for the solution $\mathbf{u}$ of (1.4).

LEMMA 4. *Let $0 < \alpha < \min(\delta, \lambda_1)$ and $\mathbf{u}_0 \in \mathbf{L}^2(\Omega)$. Then, the following estimate holds:*

$$\|\mathbf{u}(t)\| \leq e^{-\alpha t} \|\mathbf{u}_0\|, \ t > 0.$$

Moreover,

$$2\left(1 - \frac{\alpha}{\lambda_1}\right) \int_0^t e^{2\alpha\tau} \|A^{1/2}\mathbf{u}(\tau)\|^2 d\tau \leq \|\mathbf{u}_0\|^2.$$

Proof. Setting $\hat{\mathbf{u}}(t) = e^{\alpha t}\mathbf{u}(t)$ for some $\alpha > 0$, we rewrite (1.4) as

$$\frac{d}{dt}\hat{\mathbf{u}} - \alpha\hat{\mathbf{u}} + e^{-\alpha t} B(\hat{\mathbf{u}}, \hat{\mathbf{u}}) + A\hat{\mathbf{u}} + \int_0^t \beta(t-\tau) e^{\alpha(t-\tau)} A\hat{\mathbf{u}}(\tau)\, d\tau = 0. \tag{2.1}$$

Form L^2-inner product between (2.1) and $\hat{\mathbf{u}}$. Note that $(B(\hat{\mathbf{u}}, \hat{\mathbf{u}}), \hat{\mathbf{u}}) = 0$, $(A\mathbf{u}, \mathbf{v}) = (A^{\frac{1}{2}}\mathbf{u}, A^{1/2}\mathbf{v})$, and $\|\hat{\mathbf{u}}\|^2 \leq \lambda_1^{-1} \|A^{1/2}\hat{\mathbf{u}}\|^2$, where λ_1 is the least eigenvalue of the Stokes operator A. Then

$$\begin{aligned} &\frac{d}{dt}\|\hat{\mathbf{u}}\|^2 + 2\left(1 - \frac{\alpha}{\lambda_1}\right) \|A^{1/2}\hat{\mathbf{u}}\|^2 \\ &\qquad + 2\int_0^t \beta(t-\tau) e^{\alpha(t-\tau)} (A^{1/2}\hat{\mathbf{u}}(\tau), A^{1/2}\hat{\mathbf{u}}(\tau)) d\tau \leq 0. \end{aligned} \tag{2.2}$$

After integrating (2.2) with respect to time, the third term becomes nonnegative, since $\delta > \alpha$, and the second term on the left-hand side of (2.2) is also nonnegative if $\alpha < \lambda_1$. With $0 < \alpha < \min(\delta, \lambda_1)$, we find that

$$\|\hat{\mathbf{u}}\| \le \|\mathbf{u}_0\|.$$

Moreover,

$$2\left(1 - \frac{\alpha}{\lambda_1}\right) \int_0^t e^{2\alpha\tau} \|A^{1/2}\mathbf{u}(\tau)\|^2 d\tau \le \|\mathbf{u}_0\|^2.$$

This completes the rest of the proof. ∎

LEMMA 5. *Under the hypothesis of Lemma* 4*, the solution* $\mathbf{u}$ *of* (1.4) *satisfies*

$$\|A^{1/2}\mathbf{u}(t)\|^2 + e^{-2\alpha t} \int_0^t e^{2\alpha\tau} \|A\mathbf{u}(\tau)\|^2 \, d\tau \le C(\|A^{1/2}u_0\|)e^{-2\alpha t}.$$

Proof. Forming L^2-inner product between (2.1) and $A\hat{\mathbf{u}}$, we obtain

$$(\hat{\mathbf{u}}_t, A\hat{\mathbf{u}}) + \|A\hat{\mathbf{u}}\|^2 + \int_0^t \beta(t-\tau)e^{\alpha(t-\tau)}(A\hat{\mathbf{u}}(\tau), A\hat{\mathbf{u}})\, d\tau = \alpha(\hat{\mathbf{u}}, A\hat{\mathbf{u}}) - e^{-\alpha t}(B(\hat{\mathbf{u}}, \hat{\mathbf{u}}), A\hat{\mathbf{u}}). \tag{2.3}$$

Note that

$$(\hat{\mathbf{u}}_t, A\hat{\mathbf{u}}) = \frac{1}{2}\frac{d}{dt}\|A^{1/2}\hat{\mathbf{u}}\|^2.$$

On integration of (2.3) with respect to time and using Lemma 3 along with the definition of β, it follows for $0 < \alpha \le \delta$ that

$$\begin{aligned} \|A^{1/2}\hat{\mathbf{u}}(t)\|^2 + 2\int_0^t \|A\hat{\mathbf{u}}(\tau)\|^2 \, d\tau &\le \|A^{1/2}\mathbf{u}_0\|^2 + 2\alpha \int_0^t (\hat{\mathbf{u}}, A\hat{\mathbf{u}})\, d\tau \\ &\quad - 2\int_0^t e^{-\alpha\tau}(B(\hat{\mathbf{u}}, \hat{\mathbf{u}}), A\hat{\mathbf{u}})\, d\tau \\ &= \|A^{1/2}\mathbf{u}_0\|^2 + I_1 + I_2. \end{aligned} \tag{2.4}$$

To estimate $|I_1|$, we apply the Poincaré inequality and Cauchy–Schwarz inequality with $ab \le \frac{1}{2\epsilon}a^2 + \frac{\epsilon}{2}b^2$, $a, b \ge 0$, $\epsilon > 0$. Then the use of Lemma 4 yields

$$\begin{aligned} |I_1| &\le C(\alpha, \lambda_1, \epsilon) \int_0^t \|A^{1/2}\hat{\mathbf{u}}(\tau)\|^2 \, d\tau + \epsilon \int_0^t \|A\hat{\mathbf{u}}(\tau)\|^2 \, d\tau \\ &\le C(\alpha, \lambda_1, \epsilon)\|\mathbf{u}_0\|^2 + \epsilon \int_0^t \|A\hat{\mathbf{u}}(\tau)\|^2 \, d\tau. \end{aligned} \tag{2.5}$$

For the estimation of I_2, we apply Hölder's inequality repeatedly with the form of the Sobolev inequality (see Temam [28])

$$\|\boldsymbol{\phi}\|_{L^4(\Omega)} \le C\|\boldsymbol{\phi}\|^{\frac{1}{2}} \, \|A^{1/2}\boldsymbol{\phi}\|^{\frac{1}{2}}, \quad \boldsymbol{\phi} \in \mathbf{H}^1(\Omega),$$

to obtain

$$\begin{aligned} |(B(\hat{\mathbf{u}}, \hat{\mathbf{u}}), A\hat{\mathbf{u}})| &\le \|B(\hat{\mathbf{u}}, \hat{\mathbf{u}})\|\|A\hat{\mathbf{u}}\| \\ &\le C\|\hat{\mathbf{u}}\|^{\frac{1}{2}}\|A^{1/2}\hat{\mathbf{u}}\|\|A\hat{\mathbf{u}}\|^{\frac{3}{2}}. \end{aligned}$$

Thus,

$$|I_2| \le C \int_0^t e^{-\alpha\tau} \|\hat{\mathbf{u}}\|^{\frac{1}{2}} \|A^{1/2}\hat{\mathbf{u}}\| \, \|A\hat{\mathbf{u}}\|^{\frac{3}{2}} \, d\tau.$$

An application of Young's inequality $ab \le \frac{a^p}{\epsilon^{p/q}} + \frac{\epsilon b^q}{q}$, $a, b \ge 0$, $\epsilon > 0$, and $\frac{1}{p} + \frac{1}{q} = 1$ with $p = 4$ and $q = \frac{4}{3}$ yields

$$|I_2| \le C(\epsilon) \int_0^t e^{-4\alpha\tau} \|\hat{\mathbf{u}}\|^2 \|A^{1/2}\hat{\mathbf{u}}\|^4 \, d\tau + \epsilon \int_0^t \|A\hat{\mathbf{u}}\|^2 \, d\tau. \tag{2.6}$$

Substituting (2.5)–(2.6) in (2.4), and using $\epsilon = \frac{1}{2}$, we find that

$$\|A^{1/2}\hat{\mathbf{u}}(t)\|^2 + \int_0^t \|A\hat{\mathbf{u}}(\tau)\|^2 \, d\tau \le C(\alpha, \lambda_1, \|A^{1/2}\mathbf{u}_0\|) + C \int_0^t e^{-4\alpha\tau} \|\hat{\mathbf{u}}\|^2 \|A^{1/2}\hat{\mathbf{u}}\|^4 \, d\tau.$$

An application of Gronwall's lemma yields

$$\|A^{1/2}\hat{\mathbf{u}}(t)\|^2 + \int_0^t \|A\hat{\mathbf{u}}(\tau)\|^2 \, d\tau \le C(\alpha, \lambda_1, \|A^{1/2}\mathbf{u}_0\|) \exp\left\{ C \int_0^t e^{-4\alpha\tau} \|\hat{\mathbf{u}}\|^2 \|A^{1/2}\hat{\mathbf{u}}\|^2 \, d\tau \right\}.$$

Using the a priori bounds in Lemma 4 for $0 < \alpha < \min(\delta, \lambda_1)$, we obtain the desired result. This completes the proof. □

Remark 1. Based on the Faedo–Galerkin method and the a priori bounds derived in the above two lemmas, it is possible to prove the existence of global strong solutions to the problem (1.1)–(1.3). For a similar analysis in the case of Navier–Stokes equations, see Heywood [13], Temam [28], and Ladyzenskaya [20]. Since the analysis is quite standard, we state without proof the global existence theorem [25].

THEOREM 6. *Assume that* $\mathbf{u}_0 \in D(A)$. *Then for any given time* $T > 0$ *with* $0 < T \le \infty$, *there exists a unique strong solution* $\mathbf{u}$ *of* (1.4) *satisfying*

$$\mathbf{u} \in L^2(0, T; D(A)) \cap L^\infty(0, T; \mathbf{V}) \cap H^1(0, T; \mathbf{H}),$$

and the initial condition in the sense that

$$\|A^{1/2}(\mathbf{u}(t) - \mathbf{u}_0)\| \longrightarrow 0, \quad \text{as } t \longrightarrow 0.$$

Recently, Cannon et al. [5] proved existence of a global weak solution $\mathbf{u}$ satisfying

$$\mathbf{u} \in L^\infty(0, T; \mathbf{H}) \cap L^2(0, T; \mathbf{V}), \quad T > 0,$$

for a periodic problem, under the assumption that the forcing function $\mathbf{f} \in L^\infty(0, \infty; \mathbf{L}^2)$ and $\mathbf{u}_0 \in \mathbf{H}$. It is easy to extend our analysis to (1.1)–(1.3) with periodic boundary conditions and $\mathbf{f} = 0$.

Below, we derive some new regularity results without nonlocal assumptions on the data.

LEMMA 7. *Under the assumptions of Lemma* 4, *there is a positive constant* C *such that*

$$\|A\mathbf{u}(t)\| + \|\mathbf{u}_t\| \le C(\|A\mathbf{u}_0\|)e^{-\alpha t}, \quad t > 0, \tag{2.7}$$

and

$$\left(\int_0^t e^{2\alpha s} \|A^{1/2}\mathbf{u}_t(s)\|^2 \, ds \right)^{1/2} \le C(\|A\mathbf{u}_0\|). \tag{2.8}$$

Further, the following estimate holds:

$$\|A^{1/2}\mathbf{u}_t(t)\| + \left(\sigma(t)\int_0^t \sigma(s)\|A\mathbf{u}_t(s)\|^2\,ds\right)^{1/2} \le \frac{C(\|A\mathbf{u}_0\|)}{(\tau^*(t))^{1/2}}e^{-\alpha t}, \quad t>0, \tag{2.9}$$

where $\sigma(t)=\tau^*(t)e^{2\alpha t}$ *and* $\tau^*(t)=\min(t,1)$.

Proof. From (2.1), we obtain

$$e^{\alpha t}\|\mathbf{u}_t\| \le \|A\hat{\mathbf{u}}\| + e^{-\alpha t}\|B(\hat{\mathbf{u}},\hat{\mathbf{u}})\| + \int_0^t \beta(t-s)e^{\alpha(t-s)}\|A\hat{\mathbf{u}}(s)\|\,ds. \tag{2.10}$$

Using the form of B and the Sobolev inequality, it follows that

$$\begin{aligned}\|B(\hat{\mathbf{u}},\hat{\mathbf{u}})\| &\le C\|\hat{\mathbf{u}}\|^{\frac{1}{2}}\|A^{1/2}\hat{\mathbf{u}}\|\|A\hat{\mathbf{u}}\|^{\frac{1}{2}}\\ &\le C\|\hat{\mathbf{u}}\|\|A^{1/2}\hat{\mathbf{u}}\|^2 + C\|A\hat{\mathbf{u}}\|.\end{aligned} \tag{2.11}$$

On squaring (2.10) and integrating with respect to time, we find from (2.11) that

$$\begin{aligned}\int_0^t e^{2\alpha s}\|\mathbf{u}_t\|^2\,ds \le C\Bigg[&\int_0^t \|A\hat{\mathbf{u}}\|^2\,ds + \int_0^t e^{-2\alpha s}\|\hat{\mathbf{u}}\|^2\|A^{1/2}\hat{\mathbf{u}}\|^4\,ds\\ &+ \int_0^t \Big(\int_0^s \beta(s-\tau)e^{\alpha(s-\tau)}\|A\hat{\mathbf{u}}(\tau)\|\,d\tau\Big)^2 ds\Bigg].\end{aligned} \tag{2.12}$$

For the last term on the right-hand side of (2.12), use the form of β and Hölder's inequality to obtain

$$\begin{aligned}I &= \int_0^t \left(\int_0^s \beta(s-\tau)e^{\alpha(s-\tau)}\|A\hat{\mathbf{u}}(\tau)\|\,d\tau\right)^2 ds\\ &= \gamma^2\int_0^t \left(\int_0^s e^{-(\delta-\alpha)(s-\tau)}\|A\hat{\mathbf{u}}(\tau)\|\,d\tau\right)^2 ds\\ &\le \gamma^2\int_0^t \left(\int_0^s e^{-(\delta-\alpha)(s-\tau)}\,d\tau\right)\left(\int_0^s e^{-(\delta-\alpha)(s-\tau)}\|A\hat{\mathbf{u}}(\tau)\|^2\,d\tau\right) ds\\ &\le \frac{\gamma^2}{\delta-\alpha}\int_0^t \left(\int_0^s e^{-(\delta-\alpha)(s-\tau)}\|A\hat{\mathbf{u}}\|^2\,d\tau\right) ds.\end{aligned}$$

Using a change of variable, we find that

$$I \le \frac{\gamma^2}{\delta-\alpha}\int_0^t \left(\int_0^s e^{-(\delta-\alpha)\tau}\|A\hat{\mathbf{u}}(s-\tau)\|^2\,d\tau\right) ds.$$

Now a change in the order of integration yields

$$\begin{aligned}I &\le \frac{\gamma^2}{\delta-\alpha}\int_0^t e^{-(\delta-\alpha)\tau}\left(\int_\tau^t \|A\hat{\mathbf{u}}(s-\tau)\|^2\,ds\right)d\tau\\ &\le \frac{\gamma^2}{(\delta-\alpha)^2}\int_0^t e^{-(\delta-\alpha)(t-\tau)}\left(\int_0^t \|A\hat{\mathbf{u}}\|^2\,ds\right)d\tau,\end{aligned}$$

and hence,

$$I \le \left(\frac{\gamma}{\delta-\alpha}\right)^2\int_0^t \|A\hat{\mathbf{u}}(s)\|^2\,ds. \tag{2.13}$$

Using (2.13) in (2.12), we arrive at

$$\int_0^t e^{2\alpha s}\|\mathbf{u}_t\|^2\,ds \le C\left[\int_0^t \|A\hat{\mathbf{u}}\|^2\,ds + \int_0^t e^{-2\alpha s}\|\hat{\mathbf{u}}\|^2\|A^{1/2}\hat{\mathbf{u}}\|^4\,ds\right] \le C(\|A^{1/2}\mathbf{u}_0\|). \tag{2.14}$$

Differentiate (1.4) with respect to time, and integrate by parts with respect to the temporal variable for the integral term to obtain

$$\mathbf{u}_{tt} + A\mathbf{u}_t + \int_0^t \beta(t-s)A\mathbf{u}_s(s)\,ds = -(B(\mathbf{u}_t,\mathbf{u}) + B(\mathbf{u},\mathbf{u}_t)) - \beta(t)A\mathbf{u}_0. \tag{2.15}$$

Forming an inner product between (2.15) and $e^{2\alpha t}\mathbf{u}_t$, we arrive at

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|e^{\alpha t}\mathbf{u}_t\|^2 + e^{2\alpha t}\|A^{1/2}\mathbf{u}_t\|^2 + \int_0^t \beta(t-s)e^{\alpha(t-s)}(A^{1/2}e^{\alpha s}\mathbf{u}_s, A^{1/2}e^{\alpha t}\mathbf{u}_t)\,ds\\ &\qquad = \alpha\|e^{\alpha t}\mathbf{u}_t\|^2 - e^{2\alpha t}\left((B(\mathbf{u}_t,\mathbf{u}) + B(\mathbf{u},\mathbf{u}_t),\mathbf{u}_t) - \beta(t)(A\mathbf{u}_0,\mathbf{u}_t)\right).\end{aligned} \tag{2.16}$$

Note that $(B(\hat{\mathbf{u}}, e^{\alpha t}\mathbf{u}_t), e^{\alpha t}\mathbf{u}_t) = 0$. Thus, it follows after integration of (2.16) with respect to time and using the positivity property of the kernel, i.e., Lemma 3 that

$$\begin{aligned}&e^{2\alpha t}\|\mathbf{u}_t\|^2 + 2\int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t\|^2\,ds \le \|\mathbf{u}_t(0)\|^2 + 2\alpha\int_0^t e^{2\alpha s}\|\mathbf{u}_t\|^2\,ds\\ &\qquad + 2\int_0^t e^{-\alpha s}|B(e^{\alpha s}\hat{\mathbf{u}}_t,\hat{\mathbf{u}}), e^{\alpha s}\mathbf{u}_t)|\,ds + 2\gamma\|A\mathbf{u}_0\|\int_0^t e^{-(\delta-\alpha)s}\|e^{\alpha s}\mathbf{u}_t\|\,ds.\end{aligned} \tag{2.17}$$

The last term on the right-hand side of (2.17) is bounded by

$$\le C(\alpha,\delta,\gamma)\left[\|A\mathbf{u}_0\|^2 + \int_0^t e^{2\alpha s}\|\mathbf{u}_t\|^2\,ds\right]. \tag{2.18}$$

For the second term on the right-hand side of (2.17), we note with the help of Sobolev inequality that

$$\begin{aligned}2\int_0^t e^{-\alpha s}|(B(e^{\alpha s}\hat{\mathbf{u}}_t,\hat{\mathbf{u}}), e^{\alpha s}\mathbf{u}_t)|\,ds &\le C\sup_{0\le s\le t}\|A^{1/2}\hat{\mathbf{u}}(s)\|^4\int_0^t e^{-4\alpha s}(e^{2\alpha s}\|\mathbf{u}_t\|^2)\,ds\\ &\quad + \int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t\|^2\,ds.\end{aligned} \tag{2.19}$$

On substitution of (2.18)–(2.19) in (2.17) and using Lemmas 4 and 5, we obtain

$$\begin{aligned}&e^{2\alpha t}\|\mathbf{u}_t\|^2 + \int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t\|^2\,ds\\ &\qquad \le C(\delta,\alpha)\left[\|\mathbf{u}_t(0)\|^2 + \|A\mathbf{u}_0\|^2 + \int_0^t e^{2\alpha s}\|\mathbf{u}_t\|^2\,ds\right].\end{aligned} \tag{2.20}$$

From the main equation (1.4), we have at $t = 0$, $\|\mathbf{u}_t(0)\| \le C(\|A\mathbf{u}_0\|)$, and hence, using (2.14) we find that

$$\|\mathbf{u}_t\|^2 + e^{-2\alpha t}\int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t(s)\|^2\,ds \le C(\|A\mathbf{u}_0\|)e^{-2\alpha t}. \tag{2.21}$$

To estimate $\|A\mathbf{u}(t)\|$, we now form an inner product between (2.1) and $A\hat{\mathbf{u}}(t)$ to obtain

$$\|A\hat{\mathbf{u}}\|^2 \le e^{\alpha t}\|\mathbf{u}_t\|\|A\hat{\mathbf{u}}\| + e^{-\alpha t}|(B(\hat{\mathbf{u}},\hat{\mathbf{u}}),A\hat{\mathbf{u}})| + \alpha\|\hat{\mathbf{u}}\|\|A\hat{\mathbf{u}}\| + \int_0^t \beta(t-s)e^{\alpha(t-s)}\|A\hat{\mathbf{u}}(s)\|\|A\hat{\mathbf{u}}(t)\|\,ds. \tag{2.22}$$

The first three terms on the right-hand side of (2.22) are bounded by

$$\le C(\epsilon)[\|\hat{\mathbf{u}}\|^2 + e^{2\alpha t}\|\mathbf{u}_t\|^2 + e^{-4\alpha t}\|\hat{\mathbf{u}}\|^2\|A^{1/2}\hat{\mathbf{u}}\|^4] + \epsilon\|A\hat{\mathbf{u}}\|^2.$$

For the last term on the right-hand side of (2.22), we have applied the Hölder's inequality with Sobolev inequality. Then the last term is bounded by

$$C(\gamma,\delta,\alpha,\epsilon)\int_0^t e^{2\alpha\tau}\|A\mathbf{u}(\tau)\|^2\,d\tau + \epsilon\|A\hat{\mathbf{u}}\|^2.$$

Note that we have used $e^{-2(\delta-\alpha)(t-s)} \le 1$. On substituting in (2.22), we choose $\epsilon = \frac{1}{4}$. An appeal to Lemmas 4 and 5 with the estimate (2.21) yields

$$\|A\hat{\mathbf{u}}\|^2 \le C(\|A\mathbf{u}_0\|),$$

and thus we complete the proof of (2.7)–(2.8).

In order to derive (2.9), we now differentiate (1.4) with respect to time and then form an inner product with $\sigma(t)A\mathbf{u}_t$, where $\sigma(t) = \tau^*(t)e^{2\alpha t}$, to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}(\sigma(t)\|A^{1/2}\mathbf{u}_t\|^2) + \sigma(t)\|A\mathbf{u}_t\|^2 &= -\sigma(t)(A\mathbf{u},A\mathbf{u}_t) + \frac{1}{2}\sigma_t\|A^{1/2}\mathbf{u}_t\|^2\\ -\sigma(t)\int_0^t \beta_t(t-s)(A\mathbf{u}(s),A\mathbf{u}_t(t))\,ds &- \tau^*(t)e^{-\alpha t}\big(B(e^{\alpha t}\mathbf{u}_t,\hat{\mathbf{u}})\\ + B(\hat{\mathbf{u}},e^{\alpha t}\mathbf{u}_t),e^{\alpha t}A\mathbf{u}_t\big) &= I_1 + I_2 + I_3 + I_4.\end{aligned} \tag{2.23}$$

For I_1, we use Young's inequality to arrive at

$$|I_1| \le \frac{\gamma^2}{2\epsilon}\tau^*(t)\|A\hat{\mathbf{u}}\|^2 + \frac{\epsilon}{2}\sigma(t)\|A\mathbf{u}_t\|^2. \tag{2.24}$$

Since $\sigma_t = \tau_t^* e^{2\alpha t} + 2\alpha\tau^* e^{2\alpha t}$ with $\tau^*,\ \tau_t^* \le 1$, we obtain

$$|I_2| \le C(\alpha)e^{2\alpha t}\|A^{1/2}\mathbf{u}_t\|^2. \tag{2.25}$$

To estimate I_4, a use of Sobolev inequality with Young's inequality yields

$$|I_4| \le C(\epsilon)e^{2\alpha t}\|A^{1/2}\mathbf{u}_t\|^2(\|A^{1/2}\hat{\mathbf{u}}\|\|A\hat{\mathbf{u}}\| + \|A^{1/2}\hat{\mathbf{u}}\|^2) + \epsilon\sigma(t)\|A\mathbf{u}_t\|^2. \tag{2.26}$$

Since $\beta_t(t-s) = -\frac{1}{\delta}\beta(t-s)$, we obtain a bound for I_3 as

$$|I_3| \le \frac{\gamma^2}{2\epsilon\delta^2}\tau^*\left(\int_0^t e^{-(\delta-\alpha)(t-s)}\|A\hat{\mathbf{u}}(s)\|\,ds\right)^2 + \frac{\epsilon}{2}\sigma(t)\|A\mathbf{u}_t\|^2, \tag{2.27}$$

and hence, integrating with respect to time and using the estimate (2.3) for I term, we find that

$$\begin{aligned}\int_0^t |I_3|\,ds &\le \frac{\gamma^2}{2\epsilon\delta^2}\tau^* I + \frac{\epsilon}{2}\int_0^t \sigma(s)\|A\mathbf{u}_t\|^2\,ds\\ &\le C(\gamma,\delta,\alpha,\epsilon)\tau^*(t)\int_0^t \|A\hat{\mathbf{u}}(s)\|^2\,ds + \frac{\epsilon}{2}\int_0^t \sigma(s)\|A\mathbf{u}_t(s)\|^2\,ds.\end{aligned} \tag{2.28}$$

Multiply (2.23) by 2 and integrate with respect to time. Substitute (2.24)–(2.28) in (2.23). With $\epsilon = \frac{1}{4}$, it now follows that

$$\sigma(t)\|A^{1/2}\mathbf{u}_t\|^2 + \int_0^t \sigma(s)\|A\mathbf{u}_t(s)\|^2\,ds \le C(\gamma,\delta,\alpha)\left[\tau^* \int_0^t \|A\hat{\mathbf{u}}(s)\|^2\,ds + \int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t\|^2(\|A^{1/2}\hat{\mathbf{u}}\|\|A\hat{\mathbf{u}}\| + \|A^{1/2}\hat{\mathbf{u}}\|^4)\,ds\right] + \int_0^t e^{2\alpha s}\|A^{1/2}\mathbf{u}_t(s)\|^2\,ds. \tag{2.29}$$

Using Lemmas 4 and 5 and the estimates (2.7) and (2.8) in (2.29), we obtain the required result (2.9), and this completes the rest of the proof. □

Remark 2. The estimate for $\|A^{1/2}\mathbf{u}_t\|$ shows the singular behavior near $t = 0$ and also indicates the exponential decay property as $t \longrightarrow \infty$. In Lemma 7, the regularity results are derived without any nonlocal compatibility conditions.

3. Decay properties for the discrete solution and error estimates. In this section, we discuss the decay properties for the solution of the linearized backward Euler method. Finally, we derive a priori bounds for the error in H^1-norm and present briefly the error estimate in L^2-norm.

The right-hand rectangle rule q^n which is used to discretize the integral in (1.4) is positive in the sense that

$$k\sum_{n=1}^{J} q^n(\phi)\phi^n \ge 0 \quad \forall \phi = (\phi^1,\ldots,\phi^J)^T.$$

For a proof, we refer to McLean and Thomée [21, pp. 40–42]. Moreover, the following Lemma is easy to prove using the line of proof of [21].

LEMMA 8. *For any $\alpha \ge 0$, $J > 0$, and sequence $\{\phi^n\}_{n=1}^{\infty}$, the following positivity property holds:*

$$k^2\sum_{n=1}^{J}\left(\sum_{j=1}^{n} e^{-\alpha(t_n - t_j)}\phi^j\right)\phi^n \ge 0.$$

LEMMA 9. *With $0 < \alpha < \min(\delta,\ \lambda_1)$, choose $k_0 > 0$ small so that for $0 < k \le k_0$*

$$(\lambda_1 k + 1) > e^{\alpha k}.$$

Then the discrete solution $\mathbf{U}^J$, $J \ge 1$ of (1.6) is exponentially stable in the following sense:

$$\|\mathbf{U}^J\| + e^{-\alpha t_J}\left(k\sum_{n=1}^{J}\|A^{1/2}\hat{\mathbf{U}}^n\|^2\right)^{1/2} \le C(\lambda_1,\alpha)\,\|\mathbf{U}^0\|e^{-\alpha t_J}, \quad J \ge 1, \tag{3.1}$$

and

$$\|A^{1/2}\mathbf{U}^J\| \le C(\lambda_1,\alpha,\|A^{1/2}\mathbf{U}^0\|)e^{-\alpha t_J}, \quad J \ge 1. \tag{3.2}$$

Proof. Setting $\hat{\mathbf{U}}^n = e^{\alpha t_n}\mathbf{U}^n$, we rewrite (1.6) as

$$e^{\alpha t_n}\bar{\partial}_t\mathbf{U}^n + A\hat{\mathbf{U}}^n + e^{-\alpha t_{n-1}}B(\hat{\mathbf{U}}^{n-1},\hat{\mathbf{U}}^n) + e^{\alpha t_n}q^n(A\mathbf{U}) = 0.$$

Note that

$$e^{\alpha t_n}\bar{\partial}_t \mathbf{U}^n = e^{\alpha k}\bar{\partial}_t \hat{\mathbf{U}}^n - \left(\frac{e^{\alpha k}-1}{k}\right)\hat{\mathbf{U}}^n.$$

On substitution and then multiplying the resulting equation by $e^{-\alpha k}$, we obtain

$$\bar{\partial}_t \hat{\mathbf{U}}^n - \left(\frac{1-e^{-\alpha k}}{k}\right)\hat{\mathbf{U}}^n + e^{-\alpha k}A\hat{\mathbf{U}}^n + e^{-\alpha t_n}B(\hat{\mathbf{U}}^{n-1},\hat{\mathbf{U}}^n) + \gamma e^{-\alpha k}k\sum_{j=1}^{n} e^{-(\delta-\alpha)(t_n-t_j)}A\hat{\mathbf{U}}^j = 0. \tag{3.3}$$

Forming an inner product between (3.3) and $\hat{\mathbf{U}}^n$, use

$$(B(\hat{\mathbf{U}}^{n-1},\hat{\mathbf{U}}^n),\hat{\mathbf{U}}^n) = 0, \quad \|\hat{\mathbf{U}}^n\|^2 \le \frac{1}{\lambda_1}\|A^{1/2}\hat{\mathbf{U}}^n\|^2, \quad \text{and} \quad (\bar{\partial}_t\hat{\mathbf{U}}^n,\hat{\mathbf{U}}^n) \ge \frac{1}{2}\bar{\partial}_t\|\hat{\mathbf{U}}^n\|^2$$

to obtain

$$\frac{1}{2}\bar{\partial}_t\|\hat{\mathbf{U}}^n\|^2 + \left(e^{-\alpha k} - \left(\frac{1-e^{-\alpha k}}{k}\right)\lambda_1^{-1}\right)\|A^{1/2}\hat{\mathbf{U}}^n\|^2 + \gamma e^{-\alpha k}k\sum_{j=1}^{n} e^{-(\delta-\alpha)(t_n-t_j)}(A^{1/2}\hat{\mathbf{U}}^j, A^{1/2}\hat{\mathbf{U}}^n) \le 0. \tag{3.4}$$

With $0<\alpha<\min(\lambda_1,\delta)$, choose $0<k_0$ such that for $0<k<k_0$

$$(\lambda_1 k+1) \ge e^{\alpha k}.$$

Then for $0<k\le k_0$, the coefficient of the second term on the left-hand side of (3.4), $\left(e^{-\alpha k} - (\frac{1-e^{-\alpha k}}{k})\lambda_1^{-1}\right)$, becomes positive. Multiplying (3.4) by $2k$ and summing from $n=1$ to J, the last term becomes nonnegative by Lemma 8 and thus we obtain the estimate (3.1).

For the estimate (3.2), we form an inner product between (3.3) and $A\hat{\mathbf{U}}^n$ and observe that

$$(\bar{\partial}_t\hat{\mathbf{U}}^n, A\hat{\mathbf{U}}^n) = (\bar{\partial}_t A^{1/2}\hat{\mathbf{U}}^n, A^{1/2}\hat{\mathbf{U}}^n) \ge \frac{1}{2}\bar{\partial}_t\|A^{1/2}\hat{\mathbf{U}}^n\|^2.$$

Altogether, we find that

$$\frac{1}{2}\bar{\partial}_t\|A^{1/2}\hat{\mathbf{U}}^n\|^2 + e^{-\alpha k}\|A\hat{\mathbf{U}}^n\|^2 + \gamma e^{-\alpha k}k\sum_{j=1}^{n} e^{-(\delta-\alpha)(t_n-t_j)}(A\hat{\mathbf{U}}^j, A\hat{\mathbf{U}}^n) \le \left(\frac{1-e^{-\alpha k}}{k}\right)(\hat{\mathbf{U}}^n, A\hat{\mathbf{U}}^n) - e^{-\alpha t_n}(B(\hat{\mathbf{U}}^{n-1},\hat{\mathbf{U}}^n), A\hat{\mathbf{U}}^n). \tag{3.5}$$

Multiplying (3.5) by $2k$ and summing from $n=1$ to J, the third term on the left-hand side becomes nonnegative by applying Lemma 8 as $0<\alpha<\delta$. Then, we obtain

$$\begin{aligned}\|A^{1/2}\hat{\mathbf{U}}^J\|^2 + 2ke^{-\alpha k}\sum_{n=1}^{J}\|A\hat{\mathbf{U}}^n\|^2 &\le \|A^{1/2}\mathbf{U}^0\|^2 + 2(1-e^{-\alpha k})k\sum_{n=1}^{J}|(\hat{\mathbf{U}}^n, A\hat{\mathbf{U}}^n)| \\ &\quad + 2e^{-\alpha k}k\sum_{n=1}^{J} e^{-\alpha t_{n-1}}|(B(\hat{\mathbf{U}}^{n-1},\hat{\mathbf{U}}^{n-1}), A\hat{\mathbf{U}}^n)| \\ &\le \|A^{1/2}\mathbf{U}^0\|^2 + I_1 + I_2.\end{aligned} \tag{3.6}$$

To estimate I_1, we have by the mean value theorem $\frac{1-e^{-\alpha k}}{k} = \alpha e^{-\alpha k^*}$ for some $0 < k^* < k$, and hence, using (3.1), we find that

$$|I_1| \le 2\alpha e^{-\alpha k^*} k \sum_{n=1}^{J} \|A^{1/2}\hat{\mathbf{U}}^n\|^2 \le C(\lambda_1, \alpha)\|\mathbf{U}^0\|^2.$$

For I_2, a repeated use of Hölder's inequality with Sobolev inequality yields

$$e^{-\alpha t_{n-1}}|(B(\hat{\mathbf{U}}^{n-1}, \hat{\mathbf{U}}^n), A\hat{\mathbf{U}}^n)| \le Ce^{-\alpha t_{n-1}}\|\hat{\mathbf{U}}^{n-1}\|^{1/2}\|A^{1/2}\hat{\mathbf{U}}^{n-1}\|^{1/2} \|A^{1/2}\hat{\mathbf{U}}^n\|^{1/2}\|A\hat{\mathbf{U}}^n\|^{3/2}.$$

By an application of Young's inequality, it follows that

$$|I_2| \le Cke^{-\alpha k}\sum_{n=1}^{J} e^{-4\alpha t_{n-1}}(\|\hat{\mathbf{U}}^{n-1}\|^2\|A^{1/2}\hat{\mathbf{U}}^{n-1}\|^2)\|A^{1/2}\hat{\mathbf{U}}^n\|^2 + ke^{-\alpha k}\sum_{n=1}^{J}\|A\hat{\mathbf{U}}^n\|^2.$$

Using the estimate $\|\hat{\mathbf{U}}^{n-1}\|$ and

$$k\|A^{1/2}\hat{\mathbf{U}}^{J-1}\|^2 \le k\sum_{n=1}^{J}\|A^{1/2}\hat{\mathbf{U}}^n\|^2,$$

we easily find that from (3.1)

$$|I_2| \le C(\lambda,\alpha)\|\mathbf{U}^0\|^2 ke^{-\alpha k}\sum_{n=1}^{J-1} e^{-4\alpha t_{n-1}}\|A^{1/2}\hat{\mathbf{U}}^{n-1}\|^2\|A^{1/2}\hat{\mathbf{U}}^n\|^2 + C\|\mathbf{U}^0\|^4 e^{-\alpha k}e^{-4\alpha t_{J-1}}\ \|A^{1/2}\hat{\mathbf{U}}^J\|^2 + ke^{-\alpha k}\sum_{n=1}^{J}\|A\hat{\mathbf{U}}^n\|^2.$$

Now substitute the estimates of I_1 and I_2 in (3.6). For small k, we note that $(1 - C\|\mathbf{U}^0\|^4 e^{-4\alpha k})$ can be made positive. Then apply discrete Gronwall's lemma with estimate (3.1) to complete the rest of the proof. ∎

3.1. Error analysis. Now we are ready to discuss the proof of our main result that is the proof of Theorem 1.

Let ε^n be the quadrature error associated with the quadrature rule (1.5) and for $\phi \in C^1[0, t_n]$, let it be given by

$$\varepsilon^n(\phi) := \int_0^{t_n} \beta(t_n - s)\phi(s)\, ds - q^n(\phi).$$

Note that the quadrature error ε^n satisfies

$$|\varepsilon^n(\phi)| \le Ck\int_0^{t_n}\left|\frac{\partial}{\partial s}\left(\beta(t_n - s)\phi(s)\right)\right| ds \le Ck\int_0^{t_n}\left(|\beta_s(t_n - s)|\ |\phi(s)| + |\beta(t_n - s)|\ |\phi_s(s)|\right) ds. \tag{3.7}$$

For the proof of the main Theorem, we appeal to the semigroup theoretic approach; see Thomée [29], Fujita and Kato [9], and Okamoto [22]. It is well known that the Stoke's operator $-A$ generates an analytic semigroup, say, $E(t),\ t>0$ on $\mathbf{H}$; see [28] or [9]. Moreover, the following estimates are also satisfied:

$$\|A^r E(t)\| \le Ct^{-r}e^{-\lambda_1 t}, \quad t>0, \quad r>0, \tag{3.8}$$

and for $r \in (0,1]$, and $\mathbf{v} \in D(A^r)$, the domain of A^r,

$$\|(E(t)-I)\mathbf{v}\| \le C_r t^r \|A^r \mathbf{v}\|, \quad t>0, \tag{3.9}$$

where C_r is a positive constant. For a proof, see [6, p. 383]. Further, we use the discrete semigroup E_k, which is given by

$$E_k = (I+kA)^{-1}.$$

Using spectral representation of A [29], the following estimate is easy to derive:

$$\|A^r E_k^n\| \le Ct_n^{-r}e^{-\lambda_1 t_n}, \quad t_n>0, \quad 0<r\le 1. \tag{3.10}$$

Now, using Duhamel's principle, (1.4) is written in an equivalent form as

$$\mathbf{u}(t) = E(t)\mathbf{u}_0 - \int_0^t E(t-s)\tilde{A}\mathbf{u}(s)\,ds - \int_0^t E(t-s)B(\mathbf{u}(s),\mathbf{u}(s))\,ds,$$

where for simplicity of symbol, we denote

$$\tilde{A}\mathbf{u}(t) = \int_0^t \beta(t-\tau)A\mathbf{u}(\tau)\,d\tau.$$

Similarly, using discrete semigroup $E_k = (I+kA)^{-1}$, we rewrite (1.6) as

$$\mathbf{U}^n = E_k^n\mathbf{u}_0 - \sum_{j=1}^n kE_k^{n-j+1}q^j(A\mathbf{U}) - \sum_{j=1}^n kE_k^{n-j+1}B(\mathbf{U}^{j-1},\mathbf{U}^j).$$

Proof of Theorem 1. Note that the error $\mathbf{e}^n := \mathbf{u}(t_n) - \mathbf{U}^n$ is written in the form

$$\begin{aligned}\mathbf{e}^n &= (E(t_n)-E_k^n)\,\mathbf{u}_0 - \left(\int_0^{t_n} E(t_n-s)\tilde{A}\mathbf{u}(s)\,ds - \sum_{j=1}^n kE_k^{n-j+1}q^j(A\mathbf{U})\right)\\ &\quad - \left(\int_0^{t_n} E(t_n-s)B(\mathbf{u}(s),\mathbf{u}(s))\,ds - \sum_{j=1}^n kE_k^{n-j+1}B(\mathbf{U}^{j-1},\mathbf{U}^j)\right)\\ &= I_1^n - I_2^n - I_3^n.\end{aligned} \tag{3.11}$$

Since $F_k^n := (E(t_n)-E_k^n)$ denotes the error operator for the purely parabolic problem, then following Thomée [29], we estimate $A^{1/2}I_1^n$ as

$$\|A^{1/2}I_1^n\| = \|A^{1/2}F_k^n\mathbf{u}_0\| \le C(\|A\mathbf{u}_0\|,\Omega)\frac{e^{-\alpha t_n}}{t_n^{1/2}}k. \tag{3.12}$$

In order to estimate $\|A^{1/2}I_2^n\|$, i.e., the memory term, we first rewrite I_2^n as

$$
\begin{aligned}
I_2^n &= \left(\int_0^{t_n} E(t_n - s) \left(\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n) \right) ds - \sum_{j=1}^{n} kE_k^{n-j+1} \left(q^j(A\mathbf{u}) - \tilde{A}\mathbf{u}(t_n) \right) \right) \\
&+ \left(\int_0^{t_n} E(t_n - s)\, ds - \sum_{j=1}^{n} kE_k^{n-j+1} \right) \tilde{A}\mathbf{u}(t_n) \qquad (3.13) \\
&+ \sum_{j=1}^{n} kE_k^{n-j+1} q^j(A\mathbf{e}) = I_{2,1}^n + I_{2,2}^n + I_{2,3}^n.
\end{aligned}
$$

For $I_{2,2}^n$, we obtain using the semigroup property

$$
\int_0^{t_n} E(t_n - s) - \sum_{j=1}^{n} kE_k^{n-j+1} = -F_k^n A^{-1},
$$

and hence, using the definition of β, we arrive at

$$
\begin{aligned}
\|A^{1/2}I_{2,2}^n\| &= \|A^{1/2}F_k^n A^{-1}\tilde{A}\mathbf{u}(t_n)\| \\
&\le Ck \frac{e^{-\lambda_1 t_n}}{t_n^{1/2}} e^{-\alpha t_n} \| \int_0^{t_n} e^{-(\delta-\alpha)(t_n-\tau)} A\hat{\mathbf{u}}(\tau)\, d\tau \| \\
&\le Ck \frac{e^{-\lambda_1 t_n}}{t_n^{1/2}} e^{-\alpha t_n} \left(\int_0^{t_n} \|A\hat{\mathbf{u}}(\tau)\|^2\, d\tau \right)^{1/2}.
\end{aligned}
$$

An application of Lemma 5 yields for $0 < \alpha < \min(\lambda_1, \delta,)$

$$
\|A^{1/2}I_{2,2}^n\| \le C(\|A^{1/2}\mathbf{u}_0\|) k \frac{e^{-\alpha t_n}}{t_n^{1/2}}.
$$

For estimating $I_{2,3}^n$, we first use the change of variable and then the change of summation to obtain

$$
\begin{aligned}
A^{1/2}I_{2,3}^n &= \sum_{j=0}^{n-1} kAE_k^{n-j} A^{-1/2} \sum_{i=1}^{j+1} k\beta_{j+1-i} A\mathbf{e}^i = \sum_{j=0}^{n-1} kAE_k^{n-j} \sum_{i=0}^{j} k\beta_{j-i} A^{1/2}\mathbf{e}^{i+1} \\
&= k \sum_{i=0}^{n-1} \left(\sum_{j=i}^{n-1} k\beta_{j-i} AE_k^{n-j} \right) A^{1/2}\mathbf{e}^{i+1} \\
&= k \sum_{i=0}^{n-1} \left(\sum_{j=i}^{n-1} k\beta_{n-i} AE_k^{n-j} \right) A^{1/2}\mathbf{e}^{i+1} \\
&\quad - k \sum_{i=0}^{n-1} \left(\sum_{j=i}^{n-1} k(\beta_{n-i} - \beta_{j-i}) AE_k^{n-j} \right) A^{1/2}\mathbf{e}^{i+1}.
\end{aligned}
$$

For the first term on the right-hand side of $A^{1/2}I_{2,3}^n$, we have from the spectral property of the Stoke's operator and $r(\lambda) = (1+\lambda)^{-1}$:

$$\left\| k\sum_{j=i}^{n-1} AE_k^{n-j} \right\| = \sup_{\lambda\in Sp(A)} \left| \sum_{j=i}^{n-1} k\lambda r(k\lambda)^{n-j} \right| \le \sup_{\lambda>0} \sum_{j=i}^{n-1} \lambda r(\lambda)^{n-j}$$
$$\le \sup_{\lambda>0} \frac{\lambda r(\lambda)}{1-r(\lambda)} = 1,$$

where $Sp(A)$ is the spectrum of the Stokes operator A. For the second term on the right-hand side of $A^{1/2}I_{2,3}^n$, we use the smoothing property (3.8) of E_k^n, and therefore we obtain

$$\begin{aligned}
&\|A^{1/2}I_{2,3}^n\| \\
&\le \gamma k\sum_{i=0}^{n-1} e^{-\delta(t_n-t_i)} \left\| k\sum_{j=i}^{n-1} AE_k^{n-j} \right\| \ \|A^{1/2}\mathbf{e}^{i+1}\| \\
&+\gamma k\sum_{i=0}^{n-1} \left(\sum_{j=i}^{n-1} k|(e^{-\delta t_{n-i}} - e^{-\delta t_{j-i}})| \ \|AE_k^{n-j}\| \right) \|A^{1/2}\mathbf{e}^{i+1}\| \\
&\le Cke^{-\alpha t_n} \sum_{i=0}^{n-1} e^{\alpha t_i} \|A^{1/2}\mathbf{e}^{i+1}\| \\
&+Cke^{-\alpha t_n} \sum_{i=0}^{n-1} e^{\alpha t_i} \left(\sum_{j=i}^{n-1} ke^{-(\delta-\alpha)(t_j-t_i)} \frac{e^{-\delta(t_n-t_j)}-1}{(t_n-t_j)} e^{-(\lambda_1-\alpha)(t_n-t_j)} \right) \|A^{1/2}\mathbf{e}^{i+1}\|.
\end{aligned}$$

Using the meanvalue property of the exponential function, we find that

$$\left(\sum_{j=i}^{n-1} ke^{-(\delta-\alpha)(t_j-t_i)} \frac{e^{-\delta(t_n-t_j)}-1}{(t_n-t_j)} e^{-(\lambda_1-\alpha)(t_n-t_j)} \right) \le C,$$

and hence we arrive at

$$\|A^{1/2}I_{2,3}^n\| \le Ce^{-\alpha t_n}e^{-\alpha k}k\sum_{i=0}^{n} e^{\alpha t_i}\|A^{1/2}\mathbf{e}^i\|.$$

Now for the term $I_{2,1}^n$, we first rewrite it as

$$\begin{aligned}
I_{2,1}^n &= \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} (E(t_n-s) - E(t_{n-j+1})) \left(\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n) \right) ds \\
&+ \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} E(t_{n-j+1}) \left(\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_j) \right) ds \\
&+ \sum_{j=1}^{n} kF_k^{n-j+1} \left(\tilde{A}\mathbf{u}(t_j) - \tilde{A}\mathbf{u}(t_n) \right) + \sum_{j=1}^{n} kE_k^{n-j+1}\varepsilon^j(A\mathbf{u}) \\
&= M_1^n + M_2^n + M_3^n + M_4^n.
\end{aligned}$$

For M_1^n, we write it as

$$A^{1/2}M_1^n = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} A^{3/2}E(t_n - s)A^{-1}\left(I - E(s - t_{j-1})\right)\left(\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\right) ds.$$

Thus, using (3.8)–(3.9), we obtain

$$\begin{aligned}\|A^{1/2}M_1^n\| &\le \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \|A^{3/2}E(t_n - s)\|\,\|A^{-1}\left(I - E(s - t_{j-1})\right)\left(\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\right)\| ds \\ &\le Ck \int_0^{t_n} \frac{e^{-\lambda_1(t_n - s)}}{(t_n - s)^{3/2}} \|\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\| ds.\end{aligned}$$

In order to estimate $\|\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\|$, we note that

$$\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n) = \int_0^s \left(\beta(s - \tau) - \beta(t_n - \tau)\right) A\mathbf{u}(\tau)\, d\tau - \int_s^{t_n} \beta(t_n - \tau)A\mathbf{u}(\tau)\, d\tau,$$

and hence, using the definition of β, the mean value theorem, $0 < \alpha < \min(\lambda_1, \delta)$, and Lemma 7, we now obtain

$$\begin{aligned}\|\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\| &\le \gamma e^{-\delta s}\left(1 - e^{-\delta(t_n - s)}\right) \int_0^s e^{\delta\tau}\|A\mathbf{u}(\tau)\| d\tau \\ &\quad + \gamma \int_s^{t_n} e^{-\delta(t_n - \tau)}\|A\mathbf{u}(\tau)\| d\tau \\ &\le \delta\gamma(t_n - s)e^{-\alpha s}e^{-\delta s^*} \int_0^s e^{-(\delta - \alpha)(s - \tau)}\|e^{\alpha\tau}A\mathbf{u}(\tau)\| d\tau \\ &\quad + C(\|A\mathbf{u}_0\|, \gamma) \int_s^{t_n} e^{-\delta(t_n - \tau)}e^{-\alpha\tau} d\tau \\ &\le \delta\gamma(t_n - s)e^{-\alpha s}\left(\int_0^s e^{-2(\delta - \alpha)(s - \tau)} d\tau\right)^{1/2}\left(\int_0^s e^{2\alpha\tau}\|A\mathbf{u}(\tau)\|^2 d\tau\right)^{1/2} \\ &\quad + C(\|A\mathbf{u}_0\|, \gamma)(t_n - s)e^{-\alpha s}.\end{aligned}$$

Using Lemma 5 and the boundedness of

$$\int_0^s e^{-2(\delta - \alpha)(s - \tau)} d\tau \le \frac{1}{2(\delta - \alpha)},$$

we arrive at

$$\|\tilde{A}\mathbf{u}(s) - \tilde{A}\mathbf{u}(t_n)\| \le C(\|A\mathbf{u}_0\|)(t_n - s)e^{-\alpha s}.$$

Therefore,

$$\begin{aligned}\|A^{1/2}M_1^n\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n} \int_0^{t_n} \frac{e^{-(\lambda_1 - \alpha)(t_n - s)}}{(t_n - s)^{1/2}} ds \\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n} \int_0^{t_n} \frac{e^{-(\lambda_1 - \alpha)\tau}}{\tau^{1/2}} d\tau \\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n} \int_0^{\infty} \frac{e^{-(\lambda_1 - \alpha)\tau}}{\tau^{1/2}} d\tau \le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}.\end{aligned}$$

To estimate M_2^n, we use the definition of $\tilde{A}$ and the property (3.8) to find that

$$\begin{aligned}\|A^{1/2}M_2^n\| &\le \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\|A^{1/2}E(t_{n-j+1})\|\|\tilde{A}\mathbf{u}(s)-\tilde{A}\mathbf{u}(t_j)\|\,ds\\ &\le C\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\frac{e^{-\lambda_1(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\|\tilde{A}\mathbf{u}(s)-\tilde{A}\mathbf{u}(t_j)\|\,ds.\end{aligned}$$

Since

$$\|\tilde{A}\mathbf{u}(s)-\tilde{A}\mathbf{u}(t_j)\| \le C(\|A\mathbf{u}_0\|)(t_j-s)e^{-\alpha s} \le C(\|A\mathbf{u}_0\|)ke^{-\alpha s},$$

we now obtain

$$\begin{aligned}\|A^{1/2}M_2^n\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}\sum_{j=1}^{n}\frac{e^{-(\lambda_1-\alpha)(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\left(e^{\alpha t_{j-1}}\int_{t_{j-1}}^{t_j}e^{-\alpha s}\,ds\right)\\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}\left(k\sum_{j=1}^{n}\frac{e^{-(\lambda_1-\alpha)(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\right)\\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}.\end{aligned}$$

Note that we have used the boundedness of the summation term within the bracket.

In order to estimate M_3^n, we use the property of F_k^n and obtain

$$\|A^{1/2}M_3^n\| \le Ck^2\sum_{j=1}^{n}\frac{e^{-\lambda_1(t_n-t_{j-1})}}{(t_n-t_{j-1})^{3/2}}\|\tilde{A}\mathbf{u}(t_j)-\tilde{A}\mathbf{u}(t_n)\|.$$

As in the estimate of $\|A^{1/2}M_1^n\|$, we now find that

$$\begin{aligned}\|A^{1/2}M_3^n\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}e^{-\alpha k}\left(k\sum_{j=1}^{n}\frac{e^{-(\lambda_1-\alpha)(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\right)\\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}.\end{aligned}$$

Finally for M_4^n, we note that

$$\|A^{1/2}M_4^n\| \le \sum_{j=1}^{n}k\|AE_k^{n-j+1}\|\,\|\varepsilon^j(A^{1/2}\mathbf{u})\|.$$

Using (3.8), we obtain

$$\|A^{1/2}M_4^n\| \le \sum_{j=1}^{n}k\frac{e^{-\lambda_1(t_n-t_{j-1})}}{(t_n-t_{j-1})}\|\varepsilon^j(A\mathbf{u})\|.$$

To complete the estimate, we use (3.7) to compute the quadrature error $\|\varepsilon^j(A\mathbf{u})\|$ as

$$\|\varepsilon^j(A\mathbf{u})\| \le Ck\int_0^{t_j}\left(|\beta_s(t_j-s)\|A^{1/2}\mathbf{u}(s)\| + |\beta(t_j-s)|\|A^{1/2}\mathbf{u}_s(s)\|\right)ds,$$

and hence we find from Lemma 6 that

$$\begin{aligned}\|\varepsilon^j(A\mathbf{u})\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_j}\int_0^{t_j} e^{-(\delta-\alpha)(t_j-s)}\,ds\\ &\quad + Cke^{-\alpha t_j}\left(\int_0^{t_j} e^{-2(\delta-\alpha)(t_j-s)}\,ds\right)^{1/2}\left(\int_0^{t_j} e^{2\alpha s}\|A^{1/2}\mathbf{u}_s(s)\|^2\,ds\right)^{1/2}\\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_j}.\end{aligned}$$

Thus, we arrive at

$$\begin{aligned}\|A^{1/2}M_4^n\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}e^{-\alpha k}\left(k\sum_{j=1}^{n}\frac{e^{-(\lambda_1-\alpha)(t_n-t_{j-1})}}{(t_{n-j+1})}\right)\\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}e^{-\alpha k}\left(k\sum_{j=1}^{n}\frac{1}{(t_{n-j+1})}\right)\\ &\le C(\|A\mathbf{u}_0\|)k\left(\log\frac{1}{k}\right)e^{-\alpha t_n}.\end{aligned}$$

All together, we therefore obtain

$$\begin{aligned}\|A^{1/2}I_2^n\| &\le C(\|A\mathbf{u}_0\|)e^{-\alpha t_n}k\left(1+\log\frac{1}{k}\right)+C(\|A^{1/2}\mathbf{u}_0\|)\frac{e^{-\alpha t_n}}{t_n^{1/2}}k\\ &\quad + Ce^{-\alpha t_n}k\sum_{i=0}^{n-1}e^{\alpha t_i}\|A^{1/2}\mathbf{e}^i\| + Cke^{-\alpha k}\|A^{1/2}\mathbf{e}^n\|.\end{aligned}\tag{3.14}$$

Finally, in order to estimate I_3^n involving the nonlinear term, we may split it as in Geveci [11] and apply Hölder's inequality, Sobolev imbedding theorem with Sobolev inequality. Lastly, with the help of Lemmas 4, 5, 7, and 9, we obtain

$$\begin{aligned}\|A^{1/2}I_3^n\| &\le C(\|A\mathbf{u}_0\|)\frac{e^{-\alpha t_n}}{t_n^{1/2}}k + C(\|A^{1/2}u_0\|)e^{-\alpha t_n}k^{1/4}\|A^{1/2}\mathbf{e}^n\|\\ &\quad + Ce^{-\alpha t_n}k\sum_{i=0}^{n-1}\frac{e^{\alpha t_i}}{(t_n-t_i)^{3/4}}\|A^{1/2}\mathbf{e}^i\|.\end{aligned}\tag{3.15}$$

On substituting (3.12), (3.14), and (3.15) in (3.9), we obtain, for sufficiently small k,

$$\begin{aligned}e^{\alpha t_n}\|A^{1/2}\mathbf{e}^n\| &\le C(\|A\mathbf{u}_0\|)\left[k\left(t_n^{-1/2}+\log\frac{1}{k}\right)\right.\\ &\quad \left. +\ k\sum_{i=0}^{n-1}\left(\frac{1}{(t_n-t_i)^{3/4}}+1\right)e^{\alpha t_i}\|A^{1/2}\mathbf{e}^i\|\right].\end{aligned}\tag{3.16}$$

Using the generalized discrete Gronwall's lemma (see Lemma 7.1 in [7]) and the arguments of Okamoto [22, p. 635], we complete the rest of the proof. □

The convergence in $\mathbf{L}^2$-norm now becomes a routine work. However, we indicate, below, only the major steps in the proof for achieving this result.

Proof of Theorem 2. From (3.9), the error e^n satisfies

$$\mathbf{e}^n = I_1^n - I_2^n - I_3^n.$$

Since a straightforward modification of $\mathbf{H}^1$-estimates of Geveci [11] yields the $\mathbf{L}^2$-estimates of I_1^n and I_3^n, it remains to estimate $\|I_2^n\|$. Note that the $\mathbf{L}^2$-estimates of $I_{2.2}^n$ and $I_{2,3}^n$ in (3.13) follow easily as

$$\begin{aligned}\|I_{2,2}^n\| &= \|F_k^n A^{-1}\tilde{A}\mathbf{u}(t_n)\| \\ &\le Cke^{-\lambda_1 t_n}\|\int_0^{t_n}\beta(t_n-s)A\mathbf{u}(s)\,ds\| \\ &\le Cke^{-\alpha t_n}\left(\int_0^{t_n}\|A\mathbf{u}(s)\|^2\,ds\right)^{1/2} \le C(\|A^{1/2}\mathbf{u}_0\|)ke^{-\alpha t_n}\end{aligned}$$

and

$$\|I_{2,3}^n\| = \|k\sum_{j=0}^{n-1}AE_k^{n-j}\sum_{i=0}^{j}k\beta_{j-i}\mathbf{e}^{i+1}\|.$$

We repeat the analysis for estimating $A^{1/2}I_{2,3}^n$ in Theorem 1, but now $\mathbf{e}^{i+1}$ is made free of $A^{1/2}$. Thus, we obtain

$$\|I_{2,3}^n\| \le Ce^{-\alpha t_n}k\sum_{i=0}^{n-1}e^{\alpha t_i}\|\mathbf{e}^i\| + Ck\|\mathbf{e}^n\|.$$

In order to estimate $I_{2,1}^n$, it is a routine matter to derive the estimates of $\|M_1^n\|$, $\|M_2^n\|$, and $\|M_3^n\|$. To complete the rest of the proof, we therefore need an estimate for $\|M_4^n\|$. Note that

$$\begin{aligned}\|M_4^n\| &\le \sum_{j=1}^{n}k\|A^{1/2}E_k^{n-j+1}\|\,\|\varepsilon^j(A^{1/2}\mathbf{u})\| \\ &\le Ck\sum_{j=1}^{n}\frac{e^{-\lambda_1(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\|\varepsilon^j(A^{1/2}\mathbf{u})\|.\end{aligned}$$

Using the estimate of $\|\varepsilon^j(A^{1/2}\mathbf{u})\|$ as in the proof of Theorem 1, we now obtain

$$\begin{aligned}\|M_4^n\| &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}\left(k\sum_{j=1}^{n}\frac{e^{-(\lambda_1-\alpha)(t_n-t_{j-1})}}{(t_n-t_{j-1})^{1/2}}\right) \\ &\le C(\|A\mathbf{u}_0\|)ke^{-\alpha t_n}.\end{aligned}$$

Note that the summation in the bracket is bounded by a constant which is independent of k. This completes the rest of the proof. □

4. Conclusion. In this paper, we have proved new regularity results for the solutions which are valid for all time $t > 0$ without nonlocal compatibility conditions for the data and established the exponential decay property for the exact solution. Further, we have derived optimal error estimates in $\mathbf{H}^1$ and $\mathbf{L}^2$-norms for the linearized backward Euler scheme under realistically assumed conditions on the initial data. Here, the analysis is not complete as at each time level, we have still to solve an infinite dimensional problem. However, we can easily derive the error estimates for a completely discrete scheme by combining the present analysis with the semidiscrete

results obtained in [26]. Since the problem (1.1)–(1.3) can be thought of as an integral perturbation of the Navier–Stokes equations, we would like to investigate *how far the results on finite element analysis combined with higher order time discretizations of the Navier–Stokes equations* [15], [16], [22] *can be carried over to the present case.* We shall pursue this in future. Finally, we note that we have discussd our results only for the two-dimensional problem and we can easily generalize the analysis of this paper to the problem in three-dimensional bounded domain under smallness conditions on the initial data.

Acknowledgments. The first author acknowledges the financial support provided by the CNPq, Brazil during his visit to the Department of Mathematics, Federal University of Paraná, Curitiba. He also thanks the organizers of the program "Computational Challenges in PDEs" for inviting him to the Isaac Newton Institute for Mathematical Sciences, Cambridge (UK), where this work was finalized. The authors are grateful to the referees for their valuable comments and suggestions.

REFERENCES

[1] Yu. Ya. Agranovich and P. E. Sobolevskii, *Investigation of viscoelastic fluid mathematical model*, RAC Ukranian SSR. Ser. A, 10 (1989), pp. 71–74.
[2] M. M. Akhmatov and A. P. Oskolkov, *On convergent difference schemes for the equations of motion of an Oldroyd fluid*, J. Soviet Math. 47 (1989), pp. 2926–2933.
[3] W. Allegretto and Y. Lin, *Longtime stability of finite element approximations for parabolic equations with memory*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 333–354.
[4] G. Astarita and G. Marrucci, *Principles of Non-Newtonian Fluid Mechanics*, McGraw–Hill, New York, 1974.
[5] J. R. Cannon, R. E. Ewing, Y. He, and Y. Lin, *A modified nonlinear Galerkin method for the viscoelastic fluid motion equations*, Internat. J. Engrg. Sci., 37 (1999), pp. 1643–1662.
[6] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Volume* 5*: Evolution Problems* I, Springer-Verlag, Berlin, 1992.
[7] C. M. Elliott and S. Larsson, *Error estimates with smooth and nonsmooth data for a finite element method for the Cahn-Hilliard equation*, Math. Comp., 58 (1992), pp. 603–630.
[8] V. J. Ervin and N. Heuer, *Approximation of time dependent viscoelastic fluid flow: Crank-Nicolson, finite element approximation*, Numer. Methods Partial Differential Equations, 20 (2003), pp. 248–283.
[9] H. Fujita and T. Kato, *On the Navier–Stokes initial value problem* I, Arch. Ration. Mech. Anal., 16 (1964), pp. 269–315.
[10] H. Fujita and A. Mizutani, *On the finite element method for parabolic equations* I: *Approximation of holomorphic semigroups*, J. Math. Soc. Japan, 28 (1976), pp. 749–771.
[11] T. Geveci, *On the convergence of a time discretization scheme for the Navier–Stokes equations*, Math. Comp., 53 (1989), pp. 43–53.
[12] Y. He, Y. Lin, S. Shen, W. Sun, and R. Tait, *Finite element approximation for the viscoelastic fluid motion problem*, J. Comput. Appl. Math., 155 (2003), pp. 201–222.
[13] J. G. Heywood, *The Navier–Stokes equations: On the existence, regularity and decay of solutions*, Indiana Univ. Math. J., 29 (1980), pp. 639–381.
[14] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem:* I. *Regularity of solutions and second order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[15] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem:* IV. *Error analysis for second order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.
[16] A. T. Hill and E. Süli, *Approximation of global attractor for the incompressible Navier–Stokes equations*, IMA J. Numer. Anal., 20 (2000), pp. 633–667.
[17] D. D. Joseph, *Fluid Dynamics of Viscoelastic Liquids*, Springer-Verlag, New York, 1990.
[18] N. A. Karazeeva, A. A. Kotsiolis, and A. P. Oskolkov, *On the dynamical system generated by the equations of motion equations of Oldroyd fluids of order L*, J. Soviet Math., 47 (1989), pp. 2399–2403.

[19] A. A. Kotsiolis and A. P. Oskolkov, *Solvability of the basic initial boundary value problem for the motion equations of an Oldroyd's fluid on* $(0, \infty)$ *and the behavior of its solutions as* $t \mapsto \infty$, J. Soviet Math., 46 (1989), pp. 1595–1598.
[20] O. A. Ladyzenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.
[21] W. McLean and V. Thomée, *Numerical solution of an evolution equation with a positive type memory term*, J. Austral. Math. Soc. Ser. B, 35 (1993), pp. 23–70.
[22] H. Okamoto, *On the semi-discrete finite element approximation for the nonstationary Navier–Stokes equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Vo., 29 (1982), pp. 613–651.
[23] J. G. Oldroyd, *Non-Newtonian flow of liquids and solids*, Rheology: Theory and Applications, vol. I, F. R. Eirich, ed., Academic Press, New York (1956), pp. 653–682.
[24] A. P. Oskolkov, *Initial boundary value problems for the equations of motion of Kelvin-Voigt fluids and Oldroyd fluids*, Proc. Steklov Inst. Math., 2 (1989), pp. 137–182.
[25] A. K. Pani, *On the Equations of Motions Arising in the Oldroyd Model: Global Existence and Regularity*, Research Report, Department of Mathematics, IIT, Bombay, 1996.
[26] A. K. Pani and J. Y. Yuan, *Semidiscerete finite element Galerkin approximations to the equations of motion arising in the Oldroyd model*, IMA J. Numer. Anal., 25 (2005), pp. 750–782.
[27] P. E. Sobolevskii, *Stabilization of viscoelastic fluid motion (Olderoyd's mathematical model)*, Differential Integral Equations, 7 (1994), pp. 1597–1612.
[28] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, North–Holland, Amsterdam, 1984.
[29] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Series in Comput. Math. 25, Springer-Verlag, Berlin, 1997.
[30] V. Thomée and L. B. Wahlbin, *Longtime numerical solution of a parabolic equation with memory*, Math. Comp., 62 (1994), pp. 477–496.
[31] V. Thomée and N.-Y. Zhang, *Backward Euler type methods for parabolic integro-differential equations with nonsmooth data*, WSSIAA, 2 (1993), pp. 373–388.
[32] W. L. Wilkinson, *Non-Newtonian Fluids*, Pergamon Press, Oxford, UK, 1960.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 826–850

SPECTRAL DISCRETIZATION OF THE VORTICITY, VELOCITY, AND PRESSURE FORMULATION OF THE STOKES PROBLEM*

CHRISTINE BERNARDI† AND NEJMEDDINE CHORFI‡

Abstract. We consider the Stokes problem in a square or a cube provided with nonstandard boundary conditions which involve the normal component of the velocity and the tangential components of the vorticity. We write a variational formulation of this problem with three independent unknowns: the vorticity, the velocity, and the pressure. Next, we propose a discretization by spectral methods which relies on this formulation and, since it leads to an inf-sup condition on the pressure in a natural way, we prove optimal error estimates for the three unknowns. We present numerical experiments which are in perfect coherence with the analysis.

Key words. Stokes problem, vorticity-velocity-pressure formulation, spectral method

AMS subject classifications. 65N35, 35Q30

DOI. 10.1137/050622687

1. Introduction.

We consider the Stokes problem in a two- or three-dimensional bounded domain when provided with boundary conditions on the normal component of the velocity and the vorticity in dimension 2 and on the normal component of the velocity and the tangential components of the vorticity in dimension 3. The well-posedness of this problem is first proved in the pioneering paper [6]; however, the formulation that is considered in this work deals with the velocity and the pressure as only unknowns and requires the convexity or some regularity of the domain. As first proposed in [18] and [26] (see also [19] and [1]), this problem admits an equivalent variational formulation where the unknowns are the vorticity, the velocity, and the pressure. This formulation involves the domains of the divergence and curl operators, as first suggested in [25]. We also refer the reader to [21] for a different formulation where the unknowns are the vorticity, the vector potential, and the pressure and to [20] for a comparison between different formulations. The aim of this paper is to propose and analyze a discrete problem which relies on the vorticity, velocity, and pressure formulation and is constructed by spectral methods.

Indeed, it seems that the numerical analysis of discretizations relying on this formulation has been performed only for finite element methods; see [26], [2], and [11]. We refer the reader to [8] for the analysis of a spectral discretization of the same problem relying on the velocity and pressure formulation. However, the formulation that we consider here leads naturally to a more accurate approximation of the pressure. One of the difficulties in the discretization consists in handling both the two- and three-dimensional cases. Indeed, the vorticity is a scalar function in dimension 2 and can be approximated in a standard polynomial space while it is a vector field in dimension 3: This requires the introduction of appropriate polynomial spaces which are the spectral analogues of Nédélec's finite elements; see [24]. The discretization that we propose takes into account these considerations, and its numerical analysis leads to optimal

*Received by the editors January 14, 2005; accepted for publication (in revised form) November 2, 2005; published electronically April 12, 2006.
http://www.siam.org/journals/sinum/44-2/62268.html
†Laboratoire Jacques-Louis Lions, C.N.R.S. & Université Pierre et Marie Curie, B.C. 187, 4 place Jussieu, 75252 Paris Cedex 05, France (bernardi@ann.jussieu.fr).
‡Département de Mathématiques, Faculté des Sciences de Tunis, Campus Universitaire, 1060 Tunis, Tunisie (nejmeddine.chorfi@fst.rnu.tn).

error estimates on the three unknowns. This is the main advantage of this formulation since, in most usual spectral discretizations of the Stokes problem, a lack of optimality appears in the estimate concerning the pressure; see [13, sects. 24–26], and [14] for a possible but less natural improvement. We present numerical experiments which confirm the optimality of the discretization and its efficiency for the Stokes problem provided with this type of boundary condition, both in the two- and three-dimensional situations.

The extension of this study to the case of the nonlinear Navier–Stokes equations is presently under consideration. The main difficulty here is the choice of variational spaces in order to preserve the compactness of the nonlinear term. We also intend to treat more complex geometries by using a spectral element discretization.

An outline of the paper is as follows.

- In section 2, we write the variational formulation of the problem in the case of homogeneous boundary conditions.
- Section 3 is devoted to the description of the spectral discrete problem. We also prove its well-posedness.
- Optimal error estimates are derived in section 4.
- The extension to the case of nonhomogeneous boundary conditions on the velocity is explained in section 5.
- In section 6, we present some numerical experiments which turn out to be in good agreement with the analysis.

2. The velocity, vorticity, and pressure formulation. Let Ω be a bounded connected domain in $\mathbb{R}^d$, $d = 2$ or 3, with a Lipschitz-continuous boundary $\partial\Omega$. We assume for simplicity that Ω is simply connected and has a connected boundary. The generic point in Ω is denoted by $\boldsymbol{x} = (x, y)$ or $\boldsymbol{x} = (x, y, z)$ according to the dimension d. We introduce the unit outward normal vector $\boldsymbol{n}$ to Ω on $\partial\Omega$ and consider the Stokes problem

$$\begin{cases} -\nu\,\Delta\boldsymbol{u} + \mathbf{grad}\,p = \boldsymbol{f} & \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} = 0 & \text{in } \Omega, \\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega, \\ \gamma_t(\mathbf{curl}\,\boldsymbol{u}) = \boldsymbol{0} & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

To make precise the sense of the operator γ_t, we recall the following.

- In dimension $d = 2$, for any vector field $\boldsymbol{v}$ with components v_x and v_y, $\mathbf{curl}\,\boldsymbol{v}$ stands for the scalar function $\partial_x v_y - \partial_y v_x$, so that the operator γ_t is the trace operator on $\partial\Omega$.
- In dimension $d = 3$, for any vector field $\boldsymbol{v}$ with components v_x, v_y, and v_z, $\mathbf{curl}\,\boldsymbol{v}$ stands for the vector field with components $\partial_y v_z - \partial_z v_y$, $\partial_z v_x - \partial_x v_z$, and $\partial_x v_y - \partial_y v_x$, and the operator γ_t is the tangential trace operator on $\partial\Omega$, defined by $\gamma_t(\boldsymbol{w}) = \boldsymbol{w}\times\boldsymbol{n}$.

Of course, the operator γ_t is only defined on smooth enough functions as will be made precise later on.

In system (2.1), the unknowns are the velocity $\boldsymbol{u}$ and the pressure p, while the data $\boldsymbol{f}$ represents a density of body forces. The viscosity ν is a positive constant. To go further, we introduce the vorticity $\boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u}$ and observe that system (2.1) is

fully equivalent to

$$\text{(2.2)}\qquad \begin{cases} \nu\,\mathbf{curl}\,\boldsymbol{\omega} + \mathbf{grad}\,p = \boldsymbol{f} & \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} = 0 & \text{in } \Omega, \\ \boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u} & \text{in } \Omega, \\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega, \\ \gamma_t(\boldsymbol{\omega}) = \boldsymbol{0} & \text{on } \partial\Omega. \end{cases}$$

Note that the operator **curl** in the first line of this system coincides with the previous one in dimension $d = 3$ while, in dimension $d = 2$, it is applied to scalar functions $\boldsymbol{\varphi}$: $\mathbf{curl}\,\boldsymbol{\varphi}$ here denotes the vector field with components $\partial_y\boldsymbol{\varphi}$ and $-\partial_x\boldsymbol{\varphi}$.

In order to write the variational formulation of problem (2.2), we consider the full scale of Sobolev spaces $H^s(\Omega)$. As usual, we denote by $L^2_0(\Omega)$ the space of functions in $L^2(\Omega)$ with a null integral on Ω. Let also $\mathcal{D}(\Omega)$ be the space of infinitely differentiable functions with a compact support in Ω. We introduce the domain $H(\operatorname{div},\Omega)$ of the divergence operator, namely

$$\text{(2.3)}\qquad H(\operatorname{div},\Omega) = \big\{\boldsymbol{v}\in L^2(\Omega)^d;\ \operatorname{div}\boldsymbol{v}\in L^2(\Omega)\big\}.$$

A consequence of the Stokes formula, valid for smooth enough vector fields $\boldsymbol{v}$ and scalar function φ,

$$\text{(2.4)}\qquad \int_\Omega (\operatorname{div}\boldsymbol{v})(\boldsymbol{x})\varphi(\boldsymbol{x})\,d\boldsymbol{x} = -\int_\Omega \boldsymbol{v}(\boldsymbol{x})\cdot(\mathbf{grad}\,\varphi)(\boldsymbol{x})\,d\boldsymbol{x} + \int_{\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})(\boldsymbol{\tau})\varphi(\boldsymbol{\tau})\,d\boldsymbol{\tau},$$

is that the normal trace operator $\boldsymbol{v}\mapsto\boldsymbol{v}\cdot\boldsymbol{n}$ can be defined from $H(\operatorname{div},\Omega)$ into $H^{-\frac12}(\partial\Omega)$. This leads us to introduce its kernel

$$\text{(2.5)}\qquad H_0(\operatorname{div},\Omega) = \big\{\boldsymbol{v}\in H(\operatorname{div},\Omega);\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\big\}.$$

Similarly, we introduce the domain of the **curl** operator

$$\text{(2.6)}\qquad H(\mathbf{curl},\Omega) = \big\{\boldsymbol{\vartheta}\in L^2(\Omega)^{\frac{d(d-1)}{2}};\ \mathbf{curl}\,\boldsymbol{\vartheta}\in L^2(\Omega)^d\big\}.$$

The Stokes formula here reads, for smooth enough functions $\boldsymbol{\vartheta}$ in $L^2(\Omega)^{\frac{d(d-1)}{2}}$ and $\boldsymbol{v}$ in $L^2(\Omega)^d$,

$$\text{(2.7)}\qquad \int_\Omega (\mathbf{curl}\,\boldsymbol{\vartheta})(\boldsymbol{x})\cdot\boldsymbol{v}(\boldsymbol{x})\,d\boldsymbol{x} = \int_\Omega \boldsymbol{\vartheta}(\boldsymbol{x})\cdot(\mathbf{curl}\,\boldsymbol{v})(\boldsymbol{x})\,d\boldsymbol{x} - \int_{\partial\Omega}\gamma_t(\boldsymbol{\vartheta})(\boldsymbol{\tau})\cdot\tilde\gamma_t(\boldsymbol{v})(\boldsymbol{\tau})\,d\boldsymbol{\tau},$$

where $\tilde\gamma_t(\boldsymbol{v})$ is equal to $\boldsymbol{v}$ in dimension $d = 3$ and to $v_yn_x - v_xn_y$ in dimension $d = 2$. This allows us to define the kernel

$$\text{(2.8)}\qquad H_0(\mathbf{curl},\Omega) = \big\{\boldsymbol{\vartheta}\in H(\mathbf{curl},\Omega);\ \gamma_t(\boldsymbol{\vartheta}) = \boldsymbol{0} \text{ on } \partial\Omega\big\}.$$

Remark 2.1. Note that the spaces $H(\mathbf{curl},\Omega)$ and $H_0(\mathbf{curl},\Omega)$ coincide with the spaces $H^1(\Omega)$ and $H^1_0(\Omega)$ in dimension $d = 2$, so that their approximation relies on more standard discrete spaces than in dimension $d = 3$.

We now consider the following variational problem:
Find $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *in* $H_0(\mathbf{curl}, \Omega) \times H_0(\mathrm{div}, \Omega) \times L^2_0(\Omega)$ *such that*

$$\begin{aligned} &\forall \boldsymbol{v} \in H_0(\mathrm{div}, \Omega), && a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) + b(\boldsymbol{v}, p) = \langle \boldsymbol{f}, \boldsymbol{v} \rangle, \\ &\forall q \in L^2_0(\Omega), && b(\boldsymbol{u}, q) = 0, \\ &\forall \boldsymbol{\vartheta} \in H_0(\mathbf{curl}, \Omega), && c(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{\vartheta}) = 0, \end{aligned} \tag{2.9}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0(\mathrm{div}, \Omega)$ and its dual space. The bilinear forms $a(\cdot, \cdot; \cdot)$, $b(\cdot, \cdot)$, and $c(\cdot, \cdot; \cdot)$ are defined by

$$\begin{aligned} a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) = \nu \int_\Omega (\mathbf{curl}\, \boldsymbol{\omega})(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x})\, d\boldsymbol{x}, \quad b(\boldsymbol{v}, q) = -\int_\Omega (\mathrm{div}\, \boldsymbol{v})(\boldsymbol{x}) q(\boldsymbol{x})\, d\boldsymbol{x}, \\ c(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{\varphi}) = \int_\Omega \boldsymbol{\omega}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x})\, d\boldsymbol{x} - \int_\Omega \boldsymbol{u}(\boldsymbol{x}) \cdot (\mathbf{curl}\, \boldsymbol{\varphi})(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned} \tag{2.10}$$

It can be noted that the boundary conditions that appear in (2.2) are treated as essential ones in (2.9). So a direct consequence of the density of $\mathcal{D}(\Omega)^d$ in $H_0(\mathrm{div}, \Omega)$ and of $\mathcal{D}(\Omega)^{\frac{d(d-1)}{2}}$ in $H_0(\mathbf{curl}, \Omega)$ (see [22, Chap. I, sect. 2]) is the following statement.

PROPOSITION 2.2. *Problems* (2.2) *and* (2.9) *are equivalent, in the sense that any triple* $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *in* $H(\mathbf{curl}, \Omega) \times H(\mathrm{div}, \Omega) \times L^2_0(\Omega)$ *is a solution of problem* (2.2) *if and only if it is a solution of problem* (2.9).

We briefly recall from [26] (see also [11, sect. 3]) the main arguments for the analysis of problem (2.9), in view of their discrete analogues. Let V be the kernel

$$V = \big\{ \boldsymbol{v} \in H_0(\mathrm{div}, \Omega);\ \forall q \in L^2_0(\Omega),\ b(\boldsymbol{v}, q) = 0 \big\}. \tag{2.11}$$

Since the divergence of any function in $H_0(\mathrm{div}, \Omega)$ belongs to $L^2_0(\Omega)$, it is readily checked that V coincides with the space of divergence-free functions in $H_0(\mathrm{div}, \Omega)$. We also introduce the kernel

$$\mathcal{W} = \big\{ (\boldsymbol{\vartheta}, \boldsymbol{v}) \in H_0(\mathbf{curl}, \Omega) \times V;\ \forall \boldsymbol{\varphi} \in H_0(\mathbf{curl}, \Omega),\ c(\boldsymbol{\vartheta}, \boldsymbol{v}; \boldsymbol{\varphi}) = 0 \big\}. \tag{2.12}$$

As can easily be derived from the previously quoted density result, $\mathcal{W}$ coincides with the space of pairs $(\boldsymbol{\vartheta}, \boldsymbol{v})$ in $H_0(\mathbf{curl}, \Omega) \times V$ such that $\boldsymbol{\vartheta}$ is equal to $\mathbf{curl}\, \boldsymbol{v}$ in the distribution sense. Moreover, it follows from the continuity properties of the forms $b(\cdot, \cdot)$ and $c(\cdot, \cdot; \cdot)$ that both V and $\mathcal{W}$ are Hilbert spaces.

We observe that, for any solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ of problem (2.9), the pair $(\boldsymbol{\omega}, \boldsymbol{u})$ is a solution of the following reduced problem:
Find $(\boldsymbol{\omega}, \boldsymbol{u})$ *in* $\mathcal{W}$ *such that*

$$\forall \boldsymbol{v} \in V, \quad a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) = \langle \boldsymbol{f}, \boldsymbol{v} \rangle. \tag{2.13}$$

So we first investigate its well-posedness.

LEMMA 2.3. *There exists a positive constant* α *such that the form* $a(\cdot, \cdot; \cdot)$ *satisfies*

$$\begin{aligned} &\forall \boldsymbol{v} \in V \setminus \{0\}, \quad \sup_{(\boldsymbol{\omega}, \boldsymbol{u}) \in \mathcal{W}} a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) > 0, \\ &\forall (\boldsymbol{\omega}, \boldsymbol{u}) \in \mathcal{W}, \quad \sup_{\boldsymbol{v} \in V} \frac{a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v})}{\|\boldsymbol{v}\|_{L^2(\Omega)^d}} \geq \alpha \big(\|\boldsymbol{\omega}\|_{H(\mathbf{curl}, \Omega)} + \|\boldsymbol{u}\|_{L^2(\Omega)^d} \big). \end{aligned} \tag{2.14}$$

Proof. It is performed in two steps, only in the case $d = 3$ for brevity.

(1) Let $\boldsymbol{v}$ be a function in V such that $a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v})$ cancels for all $(\boldsymbol{\omega}, \boldsymbol{u})$ in $\mathcal{W}$. Since Ω is simply connected, and $\boldsymbol{v}$ is divergence-free and satisfies $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$, it follows from [3, Thm. 3.17] that there exists a divergence-free function $\boldsymbol{\psi}$ in $H_0(\mathbf{curl}, \Omega)$ such that $\boldsymbol{v}$ is equal to $\mathbf{curl}\, \boldsymbol{\psi}$. Similarly, since Ω has a connected boundary and $\boldsymbol{\psi}$ is divergence-free, it follows from [3, Thm. 3.12] that there exists a function $\boldsymbol{z}$ in V such that $\boldsymbol{\psi}$ is equal to $\mathbf{curl}\, \boldsymbol{z}$. So the pair $(\boldsymbol{\psi}, \boldsymbol{z})$ belongs to $\mathcal{W}$. Taking $(\boldsymbol{\omega}, \boldsymbol{u})$ equal to $(\boldsymbol{\psi}, \boldsymbol{z})$ thus yields that $\boldsymbol{v}$ is zero, whence the first part of (2.14).

(2) For any $(\boldsymbol{\omega}, \boldsymbol{u})$ in $\mathcal{W}$, we observe that the function $\boldsymbol{v} = \mathbf{curl}\, \boldsymbol{\omega} + \boldsymbol{u}$ belongs to V. With this choice, we have

$$a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) = \nu \, \|\mathbf{curl}\, \boldsymbol{\omega}\|^2_{L^2(\Omega)^d} + \nu \int_\Omega (\mathbf{curl}\, \boldsymbol{\omega})(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) \, d\boldsymbol{x}.$$

Since $\boldsymbol{\omega}$ is equal to $\mathbf{curl}\, \boldsymbol{u}$, we obtain by integrating by parts

$$a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) = \nu \, \|\mathbf{curl}\, \boldsymbol{\omega}\|^2_{L^2(\Omega)^d} + \frac{\nu}{2} \, \|\boldsymbol{\omega}\|^2_{L^2(\Omega)^{\frac{d(d-1)}{2}}} + \frac{\nu}{2} \, \|\mathbf{curl}\, \boldsymbol{u}\|^2_{L^2(\Omega)^{\frac{d(d-1)}{2}}}.$$

Next, using [3, Cor. 3.16] yields that, since Ω is simply connected,

$$\forall \boldsymbol{w} \in V, \quad \|\boldsymbol{w}\|_{L^2(\Omega)^d} \le c \, \|\mathbf{curl}\, \boldsymbol{w}\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}}. \tag{2.15}$$

Inserting this inequality applied to $\boldsymbol{u}$ into the previous line gives

$$a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) \ge \frac{\nu}{2} \, \|\boldsymbol{\omega}\|^2_{H(\mathbf{curl}, \Omega)} + \frac{\nu}{2c^2} \, \|\boldsymbol{u}\|^2_{L^2(\Omega)^d}.$$

This, combined with the bound

$$\|\boldsymbol{v}\|_{L^2(\Omega)^d} \le \sqrt{2} \, \big(\|\mathbf{curl}\, \boldsymbol{\omega}\|^2_{L^2(\Omega)^d} + \|\boldsymbol{u}\|^2_{L^2(\Omega)^d} \big)^{\frac{1}{2}},$$

leads to the inf-sup condition in (2.14).

The next result is now a direct consequence of (2.14); see [22, Chap. I, Lem. 4.1].

COROLLARY 2.4. *For any data $\boldsymbol{f}$ in the dual space of $H_0(\mathrm{div}, \Omega)$, problem* (2.13) *has a unique solution $(\boldsymbol{\omega}, \boldsymbol{u})$ in $\mathcal{W}$. Moreover, this solution satisfies:*

$$\|\boldsymbol{\omega}\|_{H(\mathbf{curl}, \Omega)} + \|\boldsymbol{u}\|_{L^2(\Omega)^d} \le c \, \|\boldsymbol{f}\|_{H_0(\mathrm{div}, \Omega)'}. \tag{2.16}$$

We recall the inf-sup condition, which is easily derived by taking $\boldsymbol{v}$ equal to $\mathbf{grad}\, \mu$, where μ is the unique solution in $H^1(\Omega) \cap L^2_0(\Omega)$ of the Laplace equation with data q and zero Neumann boundary condition:

$$\forall q \in L^2_0(\Omega), \quad \sup_{\boldsymbol{v} \in H_0(\mathrm{div}, \Omega)} \frac{b(\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{H(\mathrm{div}, \Omega)}} \ge \beta \, \|q\|_{L^2(\Omega)}, \tag{2.17}$$

where β is a positive constant. We are now in a position to prove the main result of this section.

THEOREM 2.5. *For any data $\boldsymbol{f}$ in the dual space of $H_0(\mathrm{div}, \Omega)$, problem* (2.9) *has a unique solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ in $H_0(\mathbf{curl}, \Omega) \times H_0(\mathrm{div}, \Omega) \times L^2_0(\Omega)$. Moreover, this solution satisfies:*

$$\|\boldsymbol{\omega}\|_{H(\mathbf{curl}, \Omega)} + \|\boldsymbol{u}\|_{H(\mathrm{div}, \Omega)} + \|p\|_{L^2(\Omega)} \le c \, \|\boldsymbol{f}\|_{H_0(\mathrm{div}, \Omega)'}. \tag{2.18}$$

Proof. We prove separately the existence and the uniqueness.

(1) With any data $\boldsymbol{f}$ in $H_0(\mathrm{div},\Omega)'$, we associate the unique solution $(\boldsymbol{\omega},\boldsymbol{u})$ of problem (2.13) by applying Corollary 2.4. It follows from the definition of V and $\mathcal{W}$ that the second and third lines in (2.9) are satisfied by this solution. Moreover, since the norms $\|\cdot\|_{L^2(\Omega)^d}$ and $\|\cdot\|_{H(\mathrm{div},\Omega)}$ coincide on V, this solution satisfies the first part of (2.18). On the other hand, the pressure p must now satisfy

$$\forall \boldsymbol{v} \in H_0(\mathrm{div},\Omega), \quad b(\boldsymbol{v},p) = \langle \boldsymbol{f},\boldsymbol{v}\rangle - a(\boldsymbol{\omega},\boldsymbol{u};\boldsymbol{v}).$$

Since the right-hand side of the previous line vanishes for all $\boldsymbol{v}$ in V (see (2.13)), the existence of a solution p of this equation in $L^2_0(\Omega)$, together with the second part of (2.18), is a consequence of condition (2.17); again see [22, Chap. I, Lem. 4.1].

(2) Let $(\boldsymbol{\omega},\boldsymbol{u},p)$ be a solution of (2.9) with data $\boldsymbol{f}$ equal to zero. Then $(\boldsymbol{\omega},\boldsymbol{u})$ is a solution of (2.13) with $\boldsymbol{f}=\boldsymbol{0}$, so that it is zero thanks to Corollary 2.4. Then the pressure p satisfies

$$\forall \boldsymbol{v} \in H_0(\mathrm{div},\Omega), \quad b(\boldsymbol{v},p) = 0,$$

so that it is zero due to the inf-sup condition (2.17). This yields the uniqueness of the solution of (2.9).

We refer the reader to [10] for the characterization of the dual space of $H_0(\mathrm{div},\Omega)$ to which the data $\boldsymbol{f}$ must belong. To conclude, we state some regularity properties of the solution of problem (2.9) which can easily be derived from [3, sect. 2], [16], and [17].

PROPOSITION 2.6. *The mapping $\boldsymbol{f} \mapsto (\boldsymbol{\omega},\boldsymbol{u},p)$, where $(\boldsymbol{\omega},\boldsymbol{u},p)$ is the solution of problem* (2.9) *with data $\boldsymbol{f}$, is continuous from $H^{\max\{0,s-1\}}(\Omega)^d$ into $H^s(\Omega)^{\frac{d(d-1)}{2}} \times H^s(\Omega)^d \times H^s(\Omega)$ for*

(i) *all $s \le \frac{1}{2}$ in the general case,*
(ii) *all $s \le 1$ when Ω is convex,*
(iii) *all $s < \frac{\pi}{\omega}$ in dimension $d=2$ when Ω is a polygon with largest angle equal to ω.*

Moreover, when the data $\boldsymbol{f}$ belongs to $L^2(\Omega)^d$, the pressure p belongs to $H^1(\Omega)$, together with the vorticity $\boldsymbol{\omega}$ in dimension $d=2$.

We refer the reader to [16] and [17] for more details about the previous statement. These properties seem weaker than the corresponding ones for the Stokes problem with Dirichlet boundary conditions on the velocity. But they are the appropriate ones for proving the convergence of the discretization.

3. The spectral discrete problem. From now on, we assume that Ω is the square or cube $]-1,1[^d$, $d=2$ or 3. The discrete spaces are constructed from the finite elements proposed by Nédélec on cubic three-dimensional meshes; see [24, sect. 2]. In order to describe them and for any triple (ℓ,m,n) of nonnegative integers, we introduce

- in dimension $d=2$, the space $\mathbb{P}_{\ell,m}(\Omega)$ of restrictions to Ω of polynomials with degree $\le \ell$ with respect to x and $\le m$ with respect to y,
- in dimension $d=3$, the space $\mathbb{P}_{\ell,m,n}(\Omega)$ of restrictions to Ω of polynomials with degree $\le \ell$ with respect to x, $\le m$ with respect to y, and $\le n$ with respect to z.

When ℓ and m are equal to n, these spaces are simply denoted by $\mathbb{P}_n(\Omega)$.

Let N be an integer ≥ 2. The space $\mathbb{D}_N$ which approximates $H_0(\mathrm{div},\Omega)$ is defined

by

(3.1)
$$\mathbb{D}_N = H_0(\mathrm{div},\Omega) \cap \begin{cases} \mathbb{P}_{N,N-1}(\Omega) \times \mathbb{P}_{N-1,N}(\Omega) & \text{if } d=2, \\ \mathbb{P}_{N,N-1,N-1}(\Omega) \times \mathbb{P}_{N-1,N,N-1}(\Omega) \times \mathbb{P}_{N-1,N-1,N}(\Omega) & \\ & \text{if } d=3. \end{cases}$$

In contrast, the space $\mathbb{C}_N$ which approximates $H_0(\mathbf{curl},\Omega)$ is rather different according to the dimension, for the reasons explained in Remark 2.1; it is defined by

(3.2)
$$\mathbb{C}_N = \begin{cases} H_0^1(\Omega) \cap \mathbb{P}_N(\Omega) & \text{if } d=2, \\ H_0(\mathbf{curl},\Omega) \cap \big(\mathbb{P}_{N-1,N,N}(\Omega) \times \mathbb{P}_{N,N-1,N}(\Omega) \times \mathbb{P}_{N,N,N-1}(\Omega)\big) & \text{if } d=3. \end{cases}$$

Finally, for the approximation of $L_0^2(\Omega)$, we consider the space $\mathbb{M}_N$:

$$\mathbb{M}_N = L_0^2(\Omega) \cap \mathbb{P}_{N-1}(\Omega). \tag{3.3}$$

Setting $\xi_0 = -1$ and $\xi_N = 1$, we introduce the $N-1$ nodes ξ_j, $1 \le j \le N-1$, and the $N+1$ weights ρ_j, $0 \le j \le N$, of the Gauss–Lobatto quadrature formula. Denoting by $\mathbb{P}_n(-1,1)$ the space of restrictions to $[-1,1]$ of polynomials with degree $\le n$, we recall that the following equality holds:

$$\forall \Phi \in \mathbb{P}_{2N-1}(-1,1), \quad \int_{-1}^{1} \Phi(\zeta)\, d\zeta = \sum_{j=0}^{N} \Phi(\xi_j)\, \rho_j. \tag{3.4}$$

We also recall [13, form. (13.20)] the property, which is useful in what follows,

$$\forall \varphi_N \in \mathbb{P}_N(-1,1), \quad \|\varphi_N\|^2_{L^2(-1,1)} \le \sum_{j=0}^{N} \varphi_N^2(\xi_j)\, \rho_j \le 3\, \|\varphi_N\|^2_{L^2(-1,1)}. \tag{3.5}$$

Relying on this formula, we introduce the discrete product, defined on continuous functions u and v by

$$(u,v)_N = \begin{cases} \sum_{i=0}^{N} \sum_{j=0}^{N} u(\xi_i,\xi_j) v(\xi_i,\xi_j)\, \rho_i \rho_j & \text{if } d=2, \\ \sum_{i=0}^{N} \sum_{j=0}^{N} \sum_{k=0}^{N} u(\xi_i,\xi_j,\xi_k) v(\xi_i,\xi_j,\xi_k)\, \rho_i \rho_j \rho_k & \text{if } d=3. \end{cases} \tag{3.6}$$

It follows from (3.5) that it is a scalar product on $\mathbb{P}_N(\Omega)$. Finally, let $\mathcal{I}_N$ denote the Lagrange interpolation operator at the nodes (ξ_i,ξ_j), $0 \le i,j \le N$, in dimension $d=2$ and at the nodes (ξ_i,ξ_j,ξ_k), $0 \le i,j,k \le N$, in dimension $d=3$, with values in $\mathbb{P}_N(\Omega)$.

From now on, we assume that the data $\boldsymbol{f}$ are continuous on $\overline{\Omega}$. The discrete problem is constructed from (2.9) by using the Galerkin method combined with numerical integration. It reads as follows:

Find $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ *in* $\mathbb{C}_N \times \mathbb{D}_N \times \mathbb{M}_N$ *such that*

$$\begin{aligned} &\forall \boldsymbol{v}_N \in \mathbb{D}_N, \quad a_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{v}_N) + b_N(\boldsymbol{v}_N, p_N) = (\boldsymbol{f}, \boldsymbol{v}_N)_N, \\ &\forall q_N \in \mathbb{M}_N, \quad b_N(\boldsymbol{u}_N, q_N) = 0, \\ &\forall \boldsymbol{\vartheta}_N \in \mathbb{C}_N, \quad c_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{\vartheta}_N) = 0, \end{aligned} \tag{3.7}$$

where the bilinear forms $a_N(\cdot,\cdot;\cdot)$, $b_N(\cdot,\cdot)$, and $c_N(\cdot,\cdot;\cdot)$ are defined by

$$
\begin{aligned}
&a_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{v}_N)=\nu\,(\mathbf{curl}\,\boldsymbol{\omega}_N,\boldsymbol{v}_N)_N,\quad b_N(\boldsymbol{v}_N,q_N)=-(\operatorname{div}\boldsymbol{v}_N,q_N)_N,\\
&c_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{\varphi}_N)=(\boldsymbol{\omega}_N,\boldsymbol{\varphi}_N)_N-(\boldsymbol{u}_N,\mathbf{curl}\,\boldsymbol{\varphi}_N)_N.
\end{aligned}
\tag{3.8}
$$

It follows from (3.5) combined with Cauchy–Schwarz inequalities that the forms $a_N(\cdot,\cdot;\cdot)$, $b_N(\cdot,\cdot)$, and $c_N(\cdot,\cdot;\cdot)$ are continuous on $\big(\mathbb{C}_N\times\mathbb{D}_N\big)\times\mathbb{D}_N$, $\mathbb{D}_N\times\mathbb{M}_N$, and $\big(\mathbb{C}_N\times\mathbb{D}_N\big)\times\mathbb{C}_N$, respectively, with norms bounded independently of N. Moreover, as a consequence of the exactness property (3.4), the forms $b(\cdot,\cdot)$ and $b_N(\cdot,\cdot)$ coincide on $\mathbb{D}_N\times\mathbb{M}_N$.

The somewhat complex choice of the discrete spaces is justified by the following lemma (its finite element analogue is well known; see [24]).

LEMMA 3.1. *The range of $\mathbb{D}_N$ by the divergence operator is contained in $\mathbb{M}_N$. The range of $\mathbb{C}_N$ by the curl operator is contained in $\mathbb{D}_N$.*

Proof. For any $\boldsymbol{v}_N$ in $\mathbb{D}_N$, $\operatorname{div}\boldsymbol{v}_N$ belongs to $\mathbb{P}_{N-1}(\Omega)$ and the fact that it has a zero integral is derived from the property $\boldsymbol{v}_N\cdot\boldsymbol{n}=0$ on $\partial\Omega$ together with the Stokes formula. This yields the first part of the lemma. Similarly, for any $\boldsymbol{\vartheta}_N$ in $\mathbb{C}_N$, each component of $\mathbf{curl}\,\boldsymbol{\vartheta}_N$ is a polynomial of the right degree and the boundary conditions $\gamma_t(\boldsymbol{\vartheta}_N)=0$ on $\partial\Omega$ imply that $\mathbf{curl}\,\boldsymbol{\vartheta}_N\cdot\boldsymbol{n}$ vanishes on $\partial\Omega$, which concludes the proof.

In analogy with the continuous case, in order to investigate the properties of problem (3.7), we introduce the kernel

$$
V_N=\big\{\boldsymbol{v}_N\in\mathbb{D}_N;\ \forall q_N\in\mathbb{M}_N,\ b_N(\boldsymbol{v}_N,q_N)=0\big\}.
\tag{3.9}
$$

The following result is easily derived from Lemma 3.1 by taking q_N equal to $\operatorname{div}\boldsymbol{v}_N$ in the previous line.

COROLLARY 3.2. *The kernel V_N is the space of divergence-free polynomials in $\mathbb{D}_N$; i.e., it coincides with $\mathbb{D}_N\cap V$.*

We now introduce the kernel

$$
\mathcal{W}_N=\big\{(\boldsymbol{\vartheta}_N,\boldsymbol{v}_N)\in\mathbb{C}_N\times V_N;\ \forall\boldsymbol{\varphi}_N\in\mathbb{C}_N,\ c_N(\boldsymbol{\vartheta}_N,\boldsymbol{v}_N;\boldsymbol{\varphi}_N)=0\big\},
\tag{3.10}
$$

and we consider the following reduced discrete problem:

Find $(\boldsymbol{\omega}_N,\boldsymbol{u}_N)$ in $\mathcal{W}_N$ such that

$$
\forall\boldsymbol{v}_N\in V_N,\quad a_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{v}_N)=(\boldsymbol{f},\boldsymbol{v}_N)_N.
\tag{3.11}
$$

We must now establish the analogues of (2.14) for the form $a_N(\cdot,\cdot;\cdot)$ on the discrete spaces. This requires several preliminary lemmas. We refer the reader to among others [15, Thm. 2.1] and [23] for analogous results in the finite element case.

LEMMA 3.3. *The kernel of the curl operator in $\mathbb{C}_N$ is reduced to $\{0\}$ in dimension $d=2$ and equal to the range of $H_0^1(\Omega)\cap\mathbb{P}_N(\Omega)$ by the gradient operator in dimension $d=3$.*

Proof. Since the lemma is obvious in dimension $d=2$, we prove it only in dimension $d=3$. Let $\boldsymbol{\vartheta}_N$ be a curl-free polynomial in $\mathbb{C}_N$. Using [22, Chap. I, Thm. 2.9] yields that it is the gradient of a function μ. It follows from the identity $\boldsymbol{\vartheta}_N=\mathbf{grad}\,\mu$ that μ belongs to $\mathbb{P}_N(\Omega)$. Moreover, the two tangential derivatives of μ vanish on all faces of Ω: Indeed, for instance on the faces contained in the planes $x=\pm1$, the second and third components of $\boldsymbol{\vartheta}_N$ are equal to zero thanks to the definition of $\mathbb{C}_N$, so that $\partial_y\mu$ and $\partial_z\mu$ vanish. So μ is constant on $\partial\Omega$ and, since it is

defined up to an additive constant, it can be taken equal to zero on $\partial\Omega$. The converse imbedding is readily checked.

The next lemma makes the second part of Lemma 3.1 more precise.

LEMMA 3.4. *The range of $\mathbb{C}_N$ by the curl operator is equal to V_N.*

Proof. Let $\boldsymbol{v}_N$ be any polynomial in V_N. We treat only the more complex case of dimension $d = 3$. Denoting the components of $\boldsymbol{v}_N$ by v_{Nx}, v_{Ny}, and v_{Nz}, we first define a function $\boldsymbol{\psi}_N = (\psi_{Nx}, \psi_{Ny}, \psi_{Nz})$ by

(3.12)
$$\psi_{Nx}(x,y,z) = \int_{-1}^{z} v_{Ny}(x,y,\zeta)\,d\zeta,$$
$$\psi_{Ny}(x,y,z) = -\int_{-1}^{z} v_{Nx}(x,y,\zeta)\,d\zeta, \quad \psi_{Nz} = 0.$$

The first two components of $\boldsymbol{\psi}_N$ belong to $\mathbb{P}_{N-1,N,N}(\Omega)$ and $\mathbb{P}_{N,N-1,N}(\Omega)$, respectively. This function is such that the first two components of its curl are equal to v_{Nx} and v_{Ny}. Moreover, since $\boldsymbol{v}_N$ belongs to V_N. we have

$$(\partial_x\psi_{Ny} - \partial_y\psi_{Nx})(x,y,z) = -\int_{-1}^{z} (\partial_x v_{Nx} + \partial_y v_{Ny})(x,y,\zeta)\,d\zeta$$
$$= \int_{-1}^{z} (\partial_z v_{Nz})(x,y,\zeta)\,d\zeta = v_{Nz}(x,y,z).$$

So $\mathbf{curl}\,\boldsymbol{\psi}_N$ is equal to $\boldsymbol{v}_N$. Moreover, it is readily checked that $\gamma_t(\boldsymbol{\psi}_N)$ vanishes on all faces of Ω but on the face Γ is contained in the plane $z = 1$. In a second step, we look for a function μ_N in $\mathbb{P}_N(\Omega)$ such that $\gamma_t(\mathbf{grad}\,\mu_N)$ is equal to zero on $\partial\Omega\setminus\Gamma$ and to $\gamma_t(\boldsymbol{\psi}_N)$ on Γ. Denoting by g_{Nx} and g_{Ny} the functions defined on Γ by

$$g_{Nx}(x,y) = \int_{-1}^{1} v_{Ny}(x,y,\zeta)\,d\zeta, \quad g_{Ny}(x,y) = -\int_{-1}^{1} v_{Nx}(x,y,\zeta)\,d\zeta,$$

we observe that the function $\boldsymbol{g}_N = (g_{Nx}, g_{Ny})$ belongs to $\mathbb{P}_{N-1,N}(\Gamma)\times\mathbb{P}_{N,N-1}(\Gamma)$, with obvious notation for these new spaces, has its tangential component equal to zero on the four edges of Γ, and satisfies, for the same reasons as previously,

$$(\partial_x g_{Ny} - \partial_y g_{Nx})(x,y) = \int_{-1}^{1} (\partial_z v_{Nz})(x,y,\zeta)\,d\zeta = 0.$$

Again applying [22, Chap. I, Thm. 2.9] yields that $\boldsymbol{g}_N$ is the tangential gradient of a function k_N^g, which is defined up to an additive constant. When choosing this constant such that $k_N^g(-1,-1)$ is zero, we easily derive that k_N^g belongs to $H_0^1(\Gamma)\cap\mathbb{P}_N(\Gamma)$. Then, using an appropriate lifting operator of polynomial traces as proposed in [9, Chap. II, Thm. 4.1], we derive the existence of a μ_N in $\mathbb{P}_N(\Omega)$ equal to 0 on $\partial\Omega\setminus\Gamma$ and to k_N^g on Γ. The function $\boldsymbol{\psi}_N - \mathbf{grad}\,\mu_N$ now belongs to $\mathbb{C}_N$ and has its curl equal to $\boldsymbol{v}_N$, whence the desired result.

It follows from Lemmas 3.3 and 3.4 that, for any function $\boldsymbol{v}_N$ in V_N, there exists a unique function $\boldsymbol{\psi}_N^*$ in $\mathbb{C}_N$ such that $\mathbf{curl}\,\boldsymbol{\psi}_N^*$ is equal to $\boldsymbol{v}_N$ and which, moreover, satisfies in dimension $d = 3$

$$\forall\mu_N \in H_0^1(\Omega)\cap\mathbb{P}_N(\Omega), \quad (\boldsymbol{\psi}_N^*, \mathbf{grad}\,\mu_N)_N = 0. \tag{3.13}$$

Let A_N be the operator defined from V_N into $\mathbb{C}_N$ by $A_N(\boldsymbol{v}_N) = \boldsymbol{\psi}_N^*$.

LEMMA 3.5. *There exists a constant c independent of N such that the following inequality holds:*

$$\forall \boldsymbol{v}_N \in V_N, \quad \|A_N(\boldsymbol{v}_N)\|_{H(\mathbf{curl},\Omega)} \leq c\,\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}. \tag{3.14}$$

Proof. Since $\mathbf{curl}\, A_N(\boldsymbol{v}_N)$ is equal to $\boldsymbol{v}_N$, it suffices to bound $\|A_N(\boldsymbol{v}_N)\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}}$. There also, we consider only the case $d = 3$. The function $\boldsymbol{\psi}_N$ defined in (3.12) obviously satisfies

$$\|\boldsymbol{\psi}_N\|_{L^2(\Omega)^3} \leq 2\,\|\boldsymbol{v}_N\|_{L^2(\Omega)^3}. \tag{3.15}$$

For the same reason, the function k_N^g introduced in the proof of Lemma 3.4 satisfies, thanks to the Poincaré–Friedrichs inequality,

$$\|k_N^g\|_{H^1(\Gamma)} \leq c\,\|\boldsymbol{g}_N\|_{L^2(\Gamma)^2} \leq c\sqrt{2}\|\boldsymbol{v}_N\|_{L^2(\Omega)^3}.$$

Thus applying [9, Chap. II, Thm. 4.1] leads to the estimate

$$\|\mathbf{grad}\,\mu_N\|_{L^2(\Omega)^3} \leq c\,\|k_N^g\|_{H^1(\Gamma)},$$

whence

$$\|\mathbf{grad}\,\mu_N\|_{L^2(\Omega)^3} \leq c'\,\|\boldsymbol{v}_N\|_{L^2(\Omega)^3}. \tag{3.16}$$

Finally, the Lax–Milgram lemma combined with (3.5) and the Poincaré–Friedrichs inequality yields that there exists a unique $\tilde{\mu}_N$ in $H_0^1(\Omega) \cap \mathbb{P}_N(\Omega)$ such that

$$\forall \rho_N \in H_0^1(\Omega) \cap \mathbb{P}_N(\Omega), \quad (\mathbf{grad}\,\tilde{\mu}_N, \mathbf{grad}\,\rho_N)_N = (\boldsymbol{\psi}_N - \mathbf{grad}\,\mu_N, \mathbf{grad}\,\rho_N)_N.$$

Moreover, this function satisfies

$$\|\mathbf{grad}\,\tilde{\mu}_N\|_{L^2(\Omega)^3} \leq 3^{\frac{3}{2}}\,(\|\boldsymbol{\psi}_N\|_{L^2(\Omega)^3} + \|\mathbf{grad}\,\mu_N\|_{L^2(\Omega)^3}). \tag{3.17}$$

The choice of $\tilde{\mu}_N$ yields that the function $\boldsymbol{\psi}_N - \mathbf{grad}\,\mu_N - \mathbf{grad}\,\tilde{\mu}_N$ is equal to $A_N(\boldsymbol{v}_N)$, so that the desired estimate follows from (3.15) to (3.17).

We are now in a position to prove successively the two analogues of (2.14).

LEMMA 3.6. *The form $a_N(\cdot,\cdot;\cdot)$ satisfies the following positivity property:*

$$\forall \boldsymbol{v}_N \in V_N \setminus \{0\}, \quad \sup_{(\boldsymbol{\omega}_N, \boldsymbol{u}_N) \in \mathcal{W}_N} a_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{v}_N) > 0. \tag{3.18}$$

Proof. Let $\boldsymbol{v}_N$ be a polynomial in V_N such that $a_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{v}_N)$ vanishes for all pairs $(\boldsymbol{\omega}_N, \boldsymbol{u}_N)$ in $\mathcal{W}_N$. We set $\boldsymbol{\vartheta}_N = A_N(\boldsymbol{v}_N)$ and consider the equation:

Find $\boldsymbol{z}_N$ in V_N such that

$$\forall \boldsymbol{w}_N \in V_N, \quad (\boldsymbol{z}_N, \boldsymbol{w}_N)_N = (\boldsymbol{\vartheta}_N, A_N(\boldsymbol{w}_N))_N. \tag{3.19}$$

Since the norms $\|\cdot\|_{H(\mathrm{div},\Omega)}$ and $\|\cdot\|_{L^2(\Omega)^3}$ are equal on V_N, it follows from (3.5) that the bilinear form in the left-hand side is elliptic on V_N, so that this problem has a unique solution $\boldsymbol{z}_N$. Moreover, this function satisfies for any $\boldsymbol{\varphi}_N$ in $\mathbb{C}_N$

$$(\boldsymbol{z}_N, \mathbf{curl}\,\boldsymbol{\varphi}_N)_N = \big(\boldsymbol{\vartheta}_N, A_N(\mathbf{curl}\,\boldsymbol{\varphi}_N)\big)_N.$$

Note that $A_N(\mathbf{curl}\,\boldsymbol{\varphi}_N)$ is the sum of $\boldsymbol{\varphi}_N$ and of the gradient of a function μ_N in $H^1_0(\Omega)\cap\mathbb{P}_N(\Omega)$. Then it follows from the choice of $\boldsymbol{\vartheta}_N$ (see (3.13)) that

$$(\boldsymbol{z}_N,\mathbf{curl}\,\boldsymbol{\varphi}_N)_N=(\boldsymbol{\vartheta}_N,\boldsymbol{\varphi}_N)_N.$$

So the pair $(\boldsymbol{\vartheta}_N,\boldsymbol{z}_N)$ belongs to $\mathcal{W}_N$ and taking $(\boldsymbol{\omega}_N,\boldsymbol{u}_N)$ equal to $(\boldsymbol{\vartheta}_N,\boldsymbol{z}_N)$ yields, thanks to (3.5), that $\boldsymbol{v}_N=\mathbf{curl}\,\boldsymbol{\vartheta}_N$ is zero, which concludes the proof.

LEMMA 3.7. *There exists a positive constant α_* independent of N such that the form $a_N(\cdot,\cdot;\cdot)$ satisfies the following inf-sup condition:*

$$\forall(\boldsymbol{\omega}_N,\boldsymbol{u}_N)\in\mathcal{W}_N,\qquad \sup_{\boldsymbol{v}_N\in V_N}\frac{a_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{v}_N)}{\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}}\geq\alpha_*\left(\|\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)}+\|\boldsymbol{u}_N\|_{L^2(\Omega)^d}\right). \tag{3.20}$$

Proof. For any $(\boldsymbol{\omega}_N,\boldsymbol{u}_N)$ in $\mathcal{W}_N$, we set $\boldsymbol{v}_N=\boldsymbol{u}_N+\mathbf{curl}\,\boldsymbol{\omega}_N$. Thanks to the definition of $\mathcal{W}_N$, this gives

$$a_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{v}_N)\geq\nu\,\|\boldsymbol{\omega}_N\|^2_{H(\mathbf{curl},\Omega)}.$$

On the other hand, again using the definition of $\mathcal{W}_N$ and (3.5), we write

$$\begin{aligned}\|\boldsymbol{u}_N\|^2_{L^2(\Omega)^d}\leq\big(\boldsymbol{u}_N,\mathbf{curl}\,A_N(\boldsymbol{u}_N)\big)_N&=\big(\boldsymbol{\omega}_N,A_N(\boldsymbol{u}_N)\big)_N\\&\leq 3^d\,\|\boldsymbol{\omega}_N\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}}\,\|A_N(\boldsymbol{u}_N)\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}}.\end{aligned}$$

So by using Lemma 3.5 we obtain

$$\|\boldsymbol{u}_N\|_{L^2(\Omega)^d}\leq 3^d\,c\,\|\boldsymbol{\omega}_N\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}}.$$

By combining these two inequalities and noting that

$$\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}\leq\sqrt{2}\left(\|\boldsymbol{\omega}_N\|^2_{H(\mathbf{curl},\Omega)}+\|\boldsymbol{u}_N\|^2_{L^2(\Omega)^d}\right)^{\frac12},$$

we derive the desired inf-sup condition.

The following result is a direct consequence of Lemmas 3.6 and 3.7; see [22, Chap. I, Lem. 4.1]. Note also from (3.5) that if $\mathcal{I}_N$ denotes the Lagrange interpolation operator introduced at the beginning of this section, the following property holds for any $\boldsymbol{v}_N$ in $\mathbb{D}_N$ (note that this requires the continuity of $\boldsymbol{f}$):

$$(\boldsymbol{f},\boldsymbol{v}_N)_N=(\mathcal{I}_N\boldsymbol{f},\boldsymbol{v}_N)_N\leq 3^d\,\|\mathcal{I}_N\boldsymbol{f}\|_{L^2(\Omega)^d}\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}.$$

COROLLARY 3.8. *For any data $\boldsymbol{f}$ continuous on $\overline{\Omega}$, problem* (3.11) *has a unique solution $(\boldsymbol{\omega}_N,\boldsymbol{u}_N)$ in $\mathcal{W}_N$. Moreover, this solution satisfies for a constant c independent of N:*

$$\|\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)}+\|\boldsymbol{u}_N\|_{L^2(\Omega)^d}\leq c\,\|\mathcal{I}_N\boldsymbol{f}\|_{L^2(\Omega)^d}. \tag{3.21}$$

To go further, we now state an inf-sup condition on the form $b_N(\cdot,\cdot)$. We refer the reader to [4, Lem. 3.1] for the main arguments of the proof in a slightly different case (and to [5] and [13, Thm. 24.6] for the basic ideas).

LEMMA 3.9. *There exists a positive constant β_* independent of N such that the form $b_N(\cdot,\cdot)$ satisfies the following inf-sup condition:*

$$\forall q_N \in \mathbb{M}_N, \quad \sup_{\boldsymbol{v}_N \in \mathbb{D}_N} \frac{b_N(\boldsymbol{v}_N, q_N)}{\|\boldsymbol{v}_N\|_{H(\mathrm{div},\Omega)}} \geq \beta_* \, \|q_N\|_{L^2(\Omega)}. \tag{3.22}$$

Note that Lemma 3.9 makes the first part of Lemma 3.1 more precise: Indeed, it implies that the range of $\mathbb{D}_N$ by the divergence operator is equal to $\mathbb{M}_N$. We skip the proof of the next theorem since it relies on exactly the same arguments as Theorem 2.5 with Corollary 2.4 replaced by Corollary 3.8 and (2.17) replaced by (3.22).

THEOREM 3.10. *For any data $\boldsymbol{f}$ continuous on $\overline{\Omega}$, problem* (3.7) *has a unique solution $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ in $\mathbb{C}_N \times \mathbb{D}_N \times \mathbb{M}_N$. Moreover, this solution satisfies for a constant c independent of N:*

$$\|\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} + \|p_N\|_{L^2(\Omega)} \leq c \, \|\mathcal{I}_N \boldsymbol{f}\|_{L^2(\Omega)^d}. \tag{3.23}$$

4. Error estimates. We now wish to derive the error estimates between the solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ of problem (2.9) and the solution $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ of problem (3.7). The proof is rather technical and requires several lemmas. In all that follows, c stands for a generic constant which can vary from one line to the next one but is always independent of N.

LEMMA 4.1. *The following estimate holds for the error between the solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ of problem* (2.9) *and the solution $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ of problem* (3.7)*:*

$$\begin{aligned} &\|\boldsymbol{\omega} - \boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} \\ &\qquad \leq c \inf_{(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \in \mathcal{W}_N} \Big(\|\boldsymbol{\omega} - \boldsymbol{\vartheta}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{w}_N\|_{L^2(\Omega)^d} \\ &\qquad\qquad + E_N^{\boldsymbol{f}} + E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \Big), \end{aligned} \tag{4.1}$$

where the quantities $E_N^{\boldsymbol{f}}$ and $E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ are defined by

$$\begin{aligned} E_N^{\boldsymbol{f}} &= \sup_{\boldsymbol{v}_N \in \mathbb{D}_N} \frac{\langle \boldsymbol{f}, \boldsymbol{v}_N \rangle - (\boldsymbol{f}, \boldsymbol{v}_N)_N}{\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}}, \\ E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) &= \sup_{\boldsymbol{v}_N \in \mathbb{D}_N} \frac{(a - a_N)(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N; \boldsymbol{v}_N)}{\|\boldsymbol{v}_N\|_{L^2(\Omega)^d}}. \end{aligned} \tag{4.2}$$

Proof. Let $(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ be an approximation of $(\boldsymbol{\omega}, \boldsymbol{u})$ in $\mathcal{W}_N$. By using (3.11), we have, for all $\boldsymbol{v}_N$ in V_N,

$$a_N(\boldsymbol{\omega}_N - \boldsymbol{\vartheta}_N, \boldsymbol{u}_N - \boldsymbol{w}_N; \boldsymbol{v}_N) = (\boldsymbol{f}, \boldsymbol{v}_N)_N - a_N(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N; \boldsymbol{v}_N).$$

Then using problem (2.13) (we recall that V_N is contained in V) leads to

$$\begin{aligned} a_N(\boldsymbol{\omega}_N - \boldsymbol{\vartheta}_N, \boldsymbol{u}_N - \boldsymbol{w}_N; \boldsymbol{v}_N) = (\boldsymbol{f}, \boldsymbol{v}_N)_N - \langle \boldsymbol{f}, \boldsymbol{v}_N \rangle + a(\boldsymbol{\omega} - \boldsymbol{\vartheta}_N, \boldsymbol{u} - \boldsymbol{w}_N; \boldsymbol{v}_N) \\ + (a - a_N)(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N; \boldsymbol{v}_N). \end{aligned}$$

By combining this identity with the inf-sup condition (3.20), we derive

$$\begin{aligned} &\|\boldsymbol{\omega}_N - \boldsymbol{\vartheta}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u}_N - \boldsymbol{w}_N\|_{L^2(\Omega)^d} \\ &\qquad \leq c \big(\|\mathbf{curl}\,(\boldsymbol{\omega} - \boldsymbol{\vartheta}_N)\|_{L^2(\Omega)^d} + E_N^{\boldsymbol{f}} + E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \big). \end{aligned}$$

We conclude thanks to a triangle inequality, by noting that both $\boldsymbol{u}$ and $\boldsymbol{u}_N$ are exactly divergence-free.

LEMMA 4.2. *The following estimate holds for the error between the solution* $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *of problem* (2.9) *and the solution* $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ *of problem* (3.7)*:*

(4.3)
$$\begin{aligned}\|p - p_N\|_{L^2(\Omega)} \le\ & c \inf_{q_N \in \mathbb{M}_N} \|p - q_N\|_{L^2(\Omega)} \\ & + c \inf_{(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \in \mathcal{W}_N} \Big(\|\boldsymbol{\omega} - \boldsymbol{\vartheta}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{w}_N\|_{L^2(\Omega)^d} \\ & \qquad + E_N^{\boldsymbol{f}} + E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \Big),\end{aligned}$$

where the quantities $E_N^{\boldsymbol{f}}$ *and* $E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ *are defined in* (4.2).

Proof. It follows from problems (2.9) and (3.7) (note also that $b(\cdot,\cdot)$ and $b_N(\cdot,\cdot)$ coincide on $\mathbb{D}_N \times \mathbb{M}_N$) that, for any $\boldsymbol{v}_N$ in $\mathbb{D}_N$ and q_N in $\mathbb{M}_N$,

$$\begin{aligned}b_N(\boldsymbol{v}_N, p_N - q_N) = (\boldsymbol{f}, \boldsymbol{v}_N)_N - \langle \boldsymbol{f}, \boldsymbol{v}_N \rangle + a(\boldsymbol{\omega} - \boldsymbol{\omega}_N, \boldsymbol{u} - \boldsymbol{u}_N; \boldsymbol{v}_N) \\ + (a - a_N)(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{v}_N) + b(\boldsymbol{v}_N, p - q_N).\end{aligned}$$

Moreover, we use the identity

$$(a - a_N)(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{v}_N) = (a - a_N)(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N; \boldsymbol{v}_N) + (a - a_N)(\boldsymbol{\omega}_N - \boldsymbol{\vartheta}_N, \boldsymbol{u}_N - \boldsymbol{w}_N; \boldsymbol{v}_N).$$

Combining the inf-sup condition (3.22) with Lemma 4.1 and a triangle inequality leads to (4.3).

In order to evaluate the distance from $(\boldsymbol{\omega}, \boldsymbol{u})$ to $\mathcal{W}_N$, we now prove an inf-sup condition on the form $c_N(\cdot,\cdot;\cdot)$.

LEMMA 4.3. *There exists a positive constant* γ_* *independent of* N *such that the form* $c_N(\cdot,\cdot;\cdot)$ *satisfies the following inf-sup condition:*

(4.4)
$$\forall \boldsymbol{\varphi}_N \in \mathbb{C}_N, \qquad \sup_{(\boldsymbol{\omega}_N, \boldsymbol{u}_N) \in \mathbb{C}_N \times V_N} \frac{c_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{\varphi}_N)}{\|\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u}_N\|_{L^2(\Omega)^d}} \ge \gamma_* \|\boldsymbol{\varphi}_N\|_{H(\mathbf{curl},\Omega)}.$$

Proof. For any $\boldsymbol{\varphi}_N$ in $\mathbb{C}_N$, we take $(\boldsymbol{\omega}_N, \boldsymbol{u}_N)$ equal to $(\boldsymbol{\varphi}_N, -\mathbf{curl}\, \boldsymbol{\varphi}_N)$ and note that it belongs to $\mathbb{C}_N \times V_N$; see Lemma 3.1. Next, we derive from (3.5) that

$$c_N(\boldsymbol{\omega}_N, \boldsymbol{u}_N; \boldsymbol{\varphi}_N) = (\boldsymbol{\varphi}_N, \boldsymbol{\varphi}_N)_N + (\mathbf{curl}\, \boldsymbol{\varphi}_N, \mathbf{curl}\, \boldsymbol{\varphi}_N)_N \ge \|\boldsymbol{\varphi}_N\|^2_{H(\mathbf{curl},\Omega)}.$$

On the other hand, we have

$$\|\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u}_N\|_{L^2(\Omega)^d} \le 2 \|\boldsymbol{\varphi}_N\|_{H(\mathbf{curl},\Omega)},$$

which leads to the desired inf-sup condition.

COROLLARY 4.4. *The following estimate holds:*

(4.5)
$$\begin{aligned}&\inf_{(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \in \mathcal{W}_N} \Big(\|\boldsymbol{\omega} - \boldsymbol{\vartheta}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{w}_N\|_{L^2(\Omega)^d} \Big) \\ &\qquad \le c \inf_{\boldsymbol{\zeta}_N \in \mathbb{C}_N} \inf_{\boldsymbol{z}_N \in V_N} \Big(\|\boldsymbol{\omega} - \boldsymbol{\zeta}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{z}_N\|_{L^2(\Omega)^d} + E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N) \Big),\end{aligned}$$

where the quantity $E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N)$ *is defined by*

$$E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N) = \sup_{\boldsymbol{\varphi}_N \in \mathbb{C}_N} \frac{(c - c_N)(\boldsymbol{\zeta}_N, \boldsymbol{z}_N; \boldsymbol{\varphi}_N)}{\|\boldsymbol{\varphi}_N\|_{H(\mathbf{curl},\Omega)}}. \tag{4.6}$$

Proof. For any $(\boldsymbol{\zeta}_N, \boldsymbol{z}_N)$ in $\mathbb{C}_N \times V_N$, we derive from the inf-sup condition (4.4) the existence of a pair $(\tilde{\boldsymbol{\zeta}}_N, \tilde{\boldsymbol{z}}_N)$ also in $\mathbb{C}_N \times V_N$ which satisfies, for all $\boldsymbol{\varphi}_N$ in $\mathbb{C}_N$,

$$c_N(\tilde{\boldsymbol{\zeta}}_N, \tilde{\boldsymbol{z}}_N; \boldsymbol{\varphi}_N) = c_N(\boldsymbol{\zeta}_N, \boldsymbol{z}_N; \boldsymbol{\varphi}_N),$$

and, moreover,

$$\|\tilde{\boldsymbol{\zeta}}_N\|_{H(\mathbf{curl},\Omega)} + \|\tilde{\boldsymbol{z}}_N\|_{L^2(\Omega)^d} \le (\gamma_*)^{-1} \sup_{\boldsymbol{\varphi}_N \in \mathbb{C}_N} \frac{c_N(\boldsymbol{\zeta}_N, \boldsymbol{z}_N; \boldsymbol{\varphi}_N)}{\|\boldsymbol{\varphi}_N\|_{H(\mathbf{curl},\Omega)}}.$$

We also note that

$$c_N(\boldsymbol{\zeta}_N, \boldsymbol{z}_N; \boldsymbol{\varphi}_N) = -c(\boldsymbol{\omega} - \boldsymbol{\zeta}_N, \boldsymbol{u} - \boldsymbol{z}_N; \boldsymbol{\varphi}_N) - (c - c_N)(\boldsymbol{\zeta}_N, \boldsymbol{z}_N; \boldsymbol{\varphi}_N).$$

Since the pair $(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ with $\boldsymbol{\vartheta}_N = \boldsymbol{\zeta}_N - \tilde{\boldsymbol{\zeta}}_N$ and $\boldsymbol{w}_N = \boldsymbol{z}_N - \tilde{\boldsymbol{z}}_N$ belongs to $\mathcal{W}_N$, the desired estimate is easily derived from the two previous lines.

Remark 4.5. The same argument as in the previous proof, combined with the inf-sup condition (3.22), leads to the estimate

$$\inf_{\boldsymbol{z}_N \in V_N} \|\boldsymbol{u} - \boldsymbol{z}_N\|_{L^2(\Omega)^d} \le c \inf_{\boldsymbol{v}_N \in \mathbb{D}_N} \|\boldsymbol{u} - \boldsymbol{v}_N\|_{H(\mathrm{div},\Omega)}. \tag{4.7}$$

However, we prefer to avoid dealing with the approximation error in the $H(\mathrm{div},\Omega)$-norm and directly estimate the distance from $\boldsymbol{u}$ to V_N.

By combining Lemmas 4.1 and 4.2 and Corollary 4.4, we observe that the full error

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} + \|p - p_N\|_{L^2(\Omega)}$$

is bounded by the sum of the three terms of approximation error,

$$\inf_{\boldsymbol{\zeta}_N \in \mathbb{C}_N} \|\boldsymbol{\omega} - \boldsymbol{\zeta}_N\|_{H(\mathbf{curl},\Omega)}, \qquad \inf_{\boldsymbol{z}_N \in V_N} \|\boldsymbol{u} - \boldsymbol{z}_N\|_{L^2(\Omega)^d}, \qquad \inf_{q_N \in \mathbb{M}_N} \|p - q_N\|_{L^2(\Omega)},$$

plus the three quantities $E_N^{\boldsymbol{f}}$, $E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$, and $E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N)$ which are issued from numerical integration.

In order to estimate these last ones, we introduce the orthogonal projection operator Π_{N-1} from $L^2(\Omega)$ onto $\mathbb{P}_{N-1}(\Omega)$. Indeed, we derive from (3.4) that, for any $\boldsymbol{v}_N$ in $\mathbb{D}_N$,

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{v}_N\rangle - (\boldsymbol{f}, \boldsymbol{v}_N)_N &= \langle \boldsymbol{f} - \Pi_{N-1}\boldsymbol{f}, \boldsymbol{v}_N\rangle - (\boldsymbol{f} - \Pi_{N-1}\boldsymbol{f}, \boldsymbol{v}_N)_N \\ &= \langle \boldsymbol{f} - \Pi_{N-1}\boldsymbol{f}, \boldsymbol{v}_N\rangle - (\mathcal{I}_N\boldsymbol{f} - \Pi_{N-1}\boldsymbol{f}, \boldsymbol{v}_N)_N,\end{aligned}$$

so that, owing to (3.5),

$$E_N^{\boldsymbol{f}} \le (1 + 3^d)\,\|\boldsymbol{f} - \Pi_{N-1}\boldsymbol{f}\|_{L^2(\Omega)^d} + 3^d\,\|\boldsymbol{f} - \mathcal{I}_N\boldsymbol{f}\|_{L^2(\Omega)^d}. \tag{4.8}$$

Similarly, we have, for any $\boldsymbol{v}_N$ in $\mathbb{D}_N$,

$$\begin{aligned}(a - a_N)(\boldsymbol{\vartheta}_N, \boldsymbol{z}_N; \boldsymbol{v}_N) = \nu \int_\Omega \big(\mathbf{curl}\,\boldsymbol{\vartheta}_N - \Pi_{N-1}(\mathbf{curl}\,\boldsymbol{\omega})\big)(\boldsymbol{x}) \cdot \boldsymbol{z}_N(\boldsymbol{x})\,d\boldsymbol{x} \\ - \nu\,\big(\mathbf{curl}\,\boldsymbol{\vartheta}_N - \Pi_{N-1}(\mathbf{curl}\,\boldsymbol{\omega}), \boldsymbol{z}_N\big)_N,\end{aligned}$$

so that

(4.9)
$$E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N) \leq (1+3^d)\left(\|\mathbf{curl}\,(\boldsymbol{\omega}-\boldsymbol{\vartheta}_N)\|_{L^2(\Omega)^d} + \|\mathbf{curl}\,\boldsymbol{\omega} - \Pi_{N-1}(\mathbf{curl}\,\boldsymbol{\omega})\|_{L^2(\Omega)^d}\right).$$

Note that a bound for the quantity $\|\mathbf{curl}\,(\boldsymbol{\omega}-\boldsymbol{\vartheta}_N)\|_{L^2(\Omega)^d}$ is provided by Corollary 4.4. Finally, the same arguments lead to

(4.10)
$$E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N) \leq (1+3^d)\left(\|\boldsymbol{\omega}-\boldsymbol{\zeta}_N\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}} + \|\boldsymbol{\omega}-\Pi_{N-1}\boldsymbol{\omega}\|_{L^2(\Omega)^{\frac{d(d-1)}{2}}} + \|\boldsymbol{u}-\boldsymbol{z}_N\|_{L^2(\Omega)^d} + \|\boldsymbol{u}-\Pi_{N-1}\boldsymbol{u}\|_{L^2(\Omega)^d}\right).$$

We now recall from [13, Thms. 7.1 and 14.2] the approximation properties of the operators Π_{N-1} and $\mathcal{I}_N$: For any function g in $H^s(\Omega)$, $s \geq 0$,

$$\|g - \Pi_{N-1} g\|_{L^2(\Omega)} \leq c\, N^{-s}\, \|g\|_{H^s(\Omega)}, \tag{4.11}$$

and, for any function g in $H^s(\Omega)$, $s > \frac{d}{2}$,

$$\|g - \mathcal{I}_N g\|_{L^2(\Omega)} \leq c\, N^{-s}\, \|g\|_{H^s(\Omega)}. \tag{4.12}$$

Estimates (4.11) and (4.12), when applied to each component of $\boldsymbol{f}$ and combined with (4.8), lead to the desired bound for $E_N^{\boldsymbol{f}}$. When combined with (4.9) and (4.10), they allow us to reduce the evaluation of $E_N^a(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ and $E_N^c(\boldsymbol{\zeta}_N, \boldsymbol{z}_N)$ to a bound for the approximation error.

The approximation error for the pressure can also be estimated from (4.11). To go further we recall the following:

• In dimension $d = 2$, the orthogonal projection operator $\Pi_N^{1,0}$ from $H_0^1(\Omega)$ onto $\mathbb{C}_N$ satisfies, for all functions $\boldsymbol{\varphi}$ in $H^s(\Omega) \cap H_0^1(\Omega)$, $s \geq 1$,

$$\|\boldsymbol{\varphi} - \Pi_N^{1,0}\boldsymbol{\varphi}\|_{L^2(\Omega)} + N^{-1}\,|\boldsymbol{\varphi} - \Pi_N^{1,0}\boldsymbol{\varphi}|_{H^1(\Omega)} \leq c\, N^{-s}\, \|\boldsymbol{\varphi}\|_{H^s(\Omega)}. \tag{4.13}$$

• In dimension $d = 3$, a spectral analogue $\mathcal{R}_N$ of the Nédélec operator [24, sect. 2] has been constructed in [7, sect. 4]. It maps smooth functions in $H_0(\mathbf{curl}, \Omega)$ onto $\mathbb{C}_N$ and satisfies, for all functions $\boldsymbol{\varphi}$ in $H^s(\Omega)^3 \cap H_0(\mathbf{curl}, \Omega)$, $s \geq 2$,

$$\|\boldsymbol{\varphi} - \mathcal{R}_N\boldsymbol{\varphi}\|_{L^2(\Omega)^3} \leq c\, N^{-s}\, \|\boldsymbol{\varphi}\|_{H^s(\Omega)^3}, \tag{4.14}$$

and, for all functions $\boldsymbol{\varphi}$ in $H_0(\mathbf{curl}, \Omega)$ such that $\mathbf{curl}\,\boldsymbol{\varphi}$ belongs to $H^s(\Omega)^3$, $s \geq 1$,

$$\|\mathbf{curl}\,(\boldsymbol{\varphi} - \mathcal{R}_N\boldsymbol{\varphi})\|_{L^2(\Omega)^3} \leq c\, N^{-s}\, \|\mathbf{curl}\,\boldsymbol{\varphi}\|_{H^s(\Omega)^3}. \tag{4.15}$$

Applying these estimates leads to a bound for the approximation error on $\boldsymbol{\omega}$. Moreover, since the velocity $\boldsymbol{u}$ is divergence-free and has a zero normal trace on $\partial\Omega$, it is equal to $\mathbf{curl}\,\boldsymbol{\psi}$ for a function $\boldsymbol{\psi}$ in $H_0(\mathbf{curl}, \Omega)$. Thus, thanks to Lemma 3.1, its best approximation in V_N can be bounded from (4.13) or (4.15).

To state the final estimate, we introduce the scale of spaces, for $s \geq 0$,

$$H^s(\mathbf{curl}, \Omega) = \left\{\boldsymbol{\varphi} \in H^s(\Omega)^{\frac{d(d-1)}{2}};\ \mathbf{curl}\,\boldsymbol{\varphi} \in H^s(\Omega)^d\right\}. \tag{4.16}$$

Note that this space coincides with $H^{s+1}(\Omega)$ in dimension $d = 2$.

THEOREM 4.6. *Assume that the data* $\boldsymbol{f}$ *belong to* $H^\sigma(\Omega)^d$ *for a real number* $\sigma > \frac{d}{2}$ *and that the solution* $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *of problem* (2.9) *belongs to* $H^s(\mathbf{curl}, \Omega) \times H^s(\Omega)^d \times H^s(\Omega)$ *for a real number* $s \geq d-1$. *Then the following error estimate holds between this solution and the solution* $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ *of problem* (3.7)*:*

(4.17)
$$\begin{aligned}\|\boldsymbol{\omega} - \boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} + \|p - p_N\|_{L^2(\Omega)} \\ \leq c\left(N^{-s}\left(\|\boldsymbol{\omega}\|_{H^s(\mathbf{curl},\Omega)} + \|\boldsymbol{u}\|_{H^s(\Omega)^d} + \|p\|_{H^s(\Omega)}\right) + N^{-\sigma}\,\|\boldsymbol{f}\|_{H^\sigma(\Omega)^d}\right).\end{aligned}$$

Estimate (4.17) is fully optimal, which is especially interesting as far as the pressure is concerned since this optimality is not obtained for most spectral discretizations of the Stokes problem.

The regularity which is required ($s \geq d-1$) concerns only the vorticity $\boldsymbol{\omega}$ and seems reasonable at least in the case of a square. Moreover, it follows from [16] and [17] that both $\boldsymbol{\omega}$ and $\boldsymbol{u}$ can be written as a sum of a regular part and the gradient of a linear combination of the singular functions associated with the Laplace operator. These two terms can be approximated separately and, as usual in spectral methods [12], the approximation of the singular part is better than can be hoped from the general theory. This leads to the following result, where, in dimension $d = 2$, σ_Ω is equal to $4 - \varepsilon$ for any $\varepsilon > 0$.

COROLLARY 4.7. *Assume that the data* $\boldsymbol{f}$ *belong to* $H^\sigma(\Omega)^d$ *for a real number* $\sigma > \frac{d}{2}$. *Then the following error estimate holds between the solution* $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *of problem* (2.9) *and the solution* $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ *of problem* (3.7)*:*

$$\begin{aligned}\|\boldsymbol{\omega} - \boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} + \|p - p_N\|_{L^2(\Omega)} \\ \leq c\,N^{-\min\{\sigma,\sigma_\Omega\}}\,\|\boldsymbol{f}\|_{H^\sigma(\Omega)^d},\end{aligned} \tag{4.18}$$

where σ_Ω *is a real number* ≥ 1 *depending only on* Ω.

5. Case of nonhomogeneous boundary conditions. We briefly explain how the results of the previous sections can be extended to the problem

$$\begin{cases}\nu\,\mathbf{curl}\,\boldsymbol{\omega} + \mathbf{grad}\,p = \boldsymbol{f} & \text{in } \Omega,\\ \mathrm{div}\,\boldsymbol{u} = 0 & \text{in } \Omega,\\ \boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u} & \text{in } \Omega,\\ \boldsymbol{u}\cdot\boldsymbol{n} = g & \text{on } \partial\Omega,\\ \gamma_t(\boldsymbol{\omega}) = \boldsymbol{0} & \text{on } \partial\Omega,\end{cases} \tag{5.1}$$

where the function g belongs to $H^{-\frac{1}{2}}(\partial\Omega)$ and satisfies the compatibility condition (here $\langle\cdot,\cdot\rangle_{\partial\Omega}$ obviously denotes the duality pairing between $H^{-\frac{1}{2}}(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$)

$$\langle g, 1\rangle_{\partial\Omega} = 0. \tag{5.2}$$

We consider the following variational problem:
Find $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *in* $H_0(\mathbf{curl}, \Omega) \times H(\mathrm{div}, \Omega) \times L^2_0(\Omega)$ *such that*

$$\boldsymbol{u}\cdot\boldsymbol{n} = g \quad \text{on } \partial\Omega \tag{5.3}$$

and that

$$\begin{aligned}&\forall \boldsymbol{v} \in H_0(\mathrm{div}, \Omega), && a(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{v}) + b(\boldsymbol{v}, p) = \langle \boldsymbol{f}, \boldsymbol{v}\rangle,\\ &\forall q \in L^2_0(\Omega), && b(\boldsymbol{u}, q) = 0,\\ &\forall \boldsymbol{\vartheta} \in H_0(\mathbf{curl}, \Omega), && c(\boldsymbol{\omega}, \boldsymbol{u}; \boldsymbol{\vartheta}) = 0.\end{aligned} \tag{5.4}$$

Thanks to the arguments given in section 2, it can be checked that problems (5.1) and (5.3)–(5.4) are equivalent, in the sense made precise in Proposition 2.2. To prove the well-posedness of problem (5.3)–(5.4), we need a lifting of the boundary condition (5.3).

LEMMA 5.1. *For any g in $H^{-\frac{1}{2}}(\partial\Omega)$ satisfying* (5.2)*, there exists a divergence-free and curl-free function $\boldsymbol{u}_b$ in $L^2(\Omega)^d$ such that $\boldsymbol{u}_b \cdot \boldsymbol{n}$ is equal to g on $\partial\Omega$. Moreover, this function satisfies*

$$\|\boldsymbol{u}_b\|_{H(\mathrm{div},\Omega)} \leq c\,\|g\|_{H^{-\frac{1}{2}}(\partial\Omega)}. \tag{5.5}$$

Proof. The following variational problem,
Find μ in $H^1(\Omega) \cap L^2_0(\Omega)$ such that

$$\forall \rho \in H^1(\Omega) \cap L^2_0(\Omega), \quad \int_\Omega \mathbf{grad}\,\mu \cdot \mathbf{grad}\,\rho\, d\boldsymbol{x} = \langle g, \rho\rangle_{\partial\Omega},$$

admits a unique solution μ which satisfies

$$|\mu|_{H^1(\Omega)} \leq c\,\|g\|_{H^{-\frac{1}{2}}(\partial\Omega)}.$$

Then the function $\boldsymbol{u}_b = \mathbf{grad}\,\mu$ satisfies all the properties stated in the lemma.

THEOREM 5.2. *For any data $\boldsymbol{f}$ in the dual space of $H_0(\mathrm{div},\Omega)$ and g in $H^{-\frac{1}{2}}(\partial\Omega)$ satisfying* (5.2)*, problem* (5.3)–(5.4) *has a unique solution $(\boldsymbol{\omega},\boldsymbol{u},p)$ in $H_0(\mathbf{curl},\Omega) \times H(\mathrm{div},\Omega) \times L^2_0(\Omega)$. Moreover, this solution satisfies*

$$\|\boldsymbol{\omega}\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u}\|_{H(\mathrm{div},\Omega)} + \|p\|_{L^2(\Omega)} \leq c\,\big(\|\boldsymbol{f}\|_{H_0(\mathrm{div},\Omega)'} + \|g\|_{H^{-\frac{1}{2}}(\partial\Omega)}\big). \tag{5.6}$$

Proof. Using the function $\boldsymbol{u}_b$ introduced in Lemma 5.1 and setting $\boldsymbol{u}_0 = \boldsymbol{u} - \boldsymbol{u}_b$, we note that $(\boldsymbol{\omega},\boldsymbol{u},p)$ is a solution of problem (5.3)–(5.4) if and only if $(\boldsymbol{\omega},\boldsymbol{u}_0,p)$ is a solution of a problem similar to (2.9). So the existence and uniqueness of $(\boldsymbol{\omega},\boldsymbol{u},p)$ follow from Theorem 2.5, and estimate (5.6) is derived by combining (2.18) and (5.5).

In order to write the discrete problem, we introduce the space

$$\overline{\mathbb{D}}_N = \begin{cases} \mathbb{P}_{N,N-1}(\Omega) \times \mathbb{P}_{N-1,N}(\Omega) & \text{if } d = 2, \\ \mathbb{P}_{N,N-1,N-1}(\Omega) \times \mathbb{P}_{N-1,N,N-1}(\Omega) \times \mathbb{P}_{N-1,N-1,N}(\Omega) & \text{if } d = 3. \end{cases} \tag{5.7}$$

Assuming that the function g belongs to $L^2(\partial\Omega)$, we define an approximation g_N of g as follows: On each edge ($d = 2$) or face ($d = 3$) Γ_r of Ω, $1 \leq r \leq 2d$, $g_{N|\Gamma_r}$ is equal to the image of $g_{|\Gamma_r}$ by the orthogonal projection operator from $L^2(\Gamma_r)$ onto $\mathbb{P}_{N-1}(\Gamma_r)$. Then we consider the following problem:

Find $(\boldsymbol{\omega}_N,\boldsymbol{u}_N,p_N)$ in $\mathbb{C}_N \times \overline{\mathbb{D}}_N \times \mathbb{M}_N$ such that

$$\boldsymbol{u}_N \cdot \boldsymbol{n} = g_N \quad \text{on } \partial\Omega \tag{5.8}$$

and that

$$\begin{aligned} &\forall \boldsymbol{v}_N \in \mathbb{D}_N, \quad a_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;v_N) + b_N(\boldsymbol{v}_N,p_N) = (\boldsymbol{f},\boldsymbol{v}_N)_N, \\ &\forall q \in \mathbb{M}_N, \quad b_N(\boldsymbol{u}_N,q_N) = 0, \\ &\forall \boldsymbol{\vartheta}_N \in \mathbb{C}_N, \quad c_N(\boldsymbol{\omega}_N,\boldsymbol{u}_N;\boldsymbol{\vartheta}_N) = 0. \end{aligned} \tag{5.9}$$

Remark 5.3. The choice of g_N as the discrete boundary condition is justified at least in a first step by the following reasons:

(i) The normal trace operator on each Γ_r, $1 \le r \le 2d$, maps $\overline{\mathbb{D}}_N$ onto $\mathbb{P}_{N-1}(\Gamma_r)$, so that each $g_{N|\Gamma_r}$ belongs to the right space.

(ii) In dimension $d = 2$, on two adjacent edges (i.e., that share a vertex), the normal trace operator involves different components of any function in $H(\mathrm{div}, \Omega)$, so that g_N does not have to satisfy any compatibility conditions at the common vertex. The same remark holds in dimension $d = 3$ for two adjacent faces (i.e., that share an edge).

(iii) Property (5.2) is still satisfied with g replaced by g_N, which is essential since we intend to work with exactly divergence-free discrete velocities.
Moreover, the computation of g_N is not too expensive.

THEOREM 5.4. *For any data $\boldsymbol{f}$ continuous on $\overline{\Omega}$ and g in $L^2(\partial\Omega)$ satisfying* (5.2)*, problem* (5.8)–(5.9) *has a unique solution $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ in $\mathbb{C}_N \times \overline{\mathbb{D}}_N \times \mathbb{M}_N$.*

Proof. It is readily checked that problem (5.8)–(5.9) can be written as a square linear system. Moreover, it follows from Theorem 3.10 that the unique solution of this problem when the data $\boldsymbol{f}$ and g_N are zero is $(\mathbf{0}, \mathbf{0}, 0)$. This yields the existence and uniqueness property.

We briefly recall the arguments that can be used in order to derive the same error estimates as in section 4.

(1) It follows from [7, sect. 4] that, in dimension $d = 3$, an extension $\overline{\mathcal{R}}_N$ of the operator $\mathcal{R}_N$ introduced in section 4 can be constructed such that estimate (4.15) still holds but now for $s \ge \frac{3}{2}$ and that, for any smooth enough function $\boldsymbol{\varphi}$, the normal traces of $\mathbf{curl}\,(\overline{\mathcal{R}}_N\boldsymbol{\varphi})$ on each Γ_r coincides with the images of the normal traces of $\mathbf{curl}\,\boldsymbol{\varphi}$ by the projection operator from $L^2(\Gamma_r)$ onto $\mathbb{P}_{N-1}(\Gamma_r)$. A similar operator can obviously be constructed in dimension $d = 2$.

(2) Since the velocity $\boldsymbol{u}$ is divergence-free and thanks to (5.2), there exists a function $\boldsymbol{\psi}$ such that $\mathbf{curl}\,\boldsymbol{\psi} = \boldsymbol{u}$. Then the function $\boldsymbol{z}_N = \mathbf{curl}\,(\overline{\mathcal{R}}_N\boldsymbol{\psi})$ belongs to $\mathbb{D}_N$, is divergence-free, and has its normal trace equal to g_N on $\partial\Omega$. Moreover, the distance of $\boldsymbol{u}$ to $\boldsymbol{z}_N$ in $L^2(\Omega)^d$ can easily be evaluated from (4.15).

(3) Let $\overline{V}_N$ denote the space of divergence-free functions in $\overline{\mathbb{D}}_N$. Thanks to Lemma 4.3 (see also the proof of Corollary 4.4), for the previous function $\boldsymbol{z}_N$ and any $\boldsymbol{\zeta}_N$ in $\mathbb{C}_N$, there exists a $(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ in $\mathbb{C}_N \times \overline{V}_N$ such that the pair $(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N)$ satisfies (4.5), the normal traces of $\boldsymbol{w}_N$ and $\boldsymbol{z}_N$ coincide on $\partial\Omega$, and, moreover,

$$\forall \boldsymbol{\varphi}_N \in \mathbb{C}_N,\ c_N(\boldsymbol{\vartheta}_N, \boldsymbol{w}_N; \boldsymbol{\varphi}_N) = 0.$$

(4) The pair $(\boldsymbol{\omega}_N - \boldsymbol{\vartheta}_N, \boldsymbol{u}_N - \boldsymbol{z}_N)$ now belongs to $\mathcal{W}_N$. So exactly the same arguments as in the proof of Lemma 4.1 lead to estimate (4.1).

THEOREM 5.5. *If the assumptions of Theorem* 4.6 *hold and the data g satisfies condition* (5.2) *and is such that each $g_{|\Gamma_r}$, $1 \le r \le 2d$, belongs to $H^\tau(\Gamma_r)$ for a nonnegative real number τ, the following error estimate holds between the solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ of problem* (5.3)–(5.4) *and the solution $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ of problem* (5.8)–(5.9)*:*

(5.10)
$$\begin{aligned}\|\boldsymbol{\omega} - \boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)} + \|p - p_N\|_{L^2(\Omega)} \\ \le c\Big(N^{-s}\left(\|\boldsymbol{\omega}\|_{H^s(\mathbf{curl},\Omega)} + \|\boldsymbol{u}\|_{H^s(\Omega)^d} + \|p\|_{H^s(\Omega)}\right) \\ + N^{-\sigma}\,\|\boldsymbol{f}\|_{H^\sigma(\Omega)^d} + N^{-\tau-\frac{1}{2}}\sum_{r=1}^{2d}\|g\|_{H^\tau(\Gamma_r)}\Big).\end{aligned}$$

COROLLARY 5.6. *Assume that the data $(\boldsymbol{f}, g)$ belong to $H^\sigma(\Omega)^d \times H^{\sigma-\frac{1}{2}}(\partial\Omega)$ for a real number $\sigma > \frac{d}{2}$ and that condition* (5.2) *is satisfied. Then the following error*

estimate holds between the solution $(\boldsymbol{\omega}, \boldsymbol{u}, p)$ *of problem* (5.3)–(5.4) *and the solution* $(\boldsymbol{\omega}_N, \boldsymbol{u}_N, p_N)$ *of problem* (5.8)–(5.9) *for the same real number* σ_Ω *as in Corollary* 4.7*:*

$$\begin{aligned}\|\boldsymbol{\omega}-\boldsymbol{\omega}_N\|_{H(\mathbf{curl},\Omega)}+\|\boldsymbol{u}-\boldsymbol{u}_N\|_{H(\mathrm{div},\Omega)}+\|p-p_N\|_{L^2(\Omega)}&\\ \le c\,N^{-\min\{\sigma,\sigma_\Omega\}}\left(\|\boldsymbol{f}\|_{H^\sigma(\Omega)^d}+\|g\|_{H^{\sigma-\frac12}(\partial\Omega)}\right).&\end{aligned}\tag{5.11}$$

6. Some numerical experiments. Before presenting the numerical experiments, we briefly describe how problem (3.7) is implemented. Let φ_j, $0 \le j \le N$, denote the Lagrange polynomials in $\mathbb{P}_N(-1,1)$ associated with the nodes ξ_j. We fix an integer j_* between 0 and N (usually equal to the integer part of $\frac{N}{2}$), define J^* as the set $\{0,\dots,N\}\setminus\{j^*\}$, and set

$$\varphi_j^*(\zeta)=\varphi_j(\zeta)\,\frac{\xi_j-\xi_{j^*}}{\zeta-\xi_{j^*}},\quad j\in J^*.\tag{6.1}$$

Then the unknowns $\boldsymbol{\omega}_N$ and $\boldsymbol{u}_N$ and a pseudopressure $\tilde{p}_N$ admit the expansions, in dimension $d=2$ for simplicity,

$$\begin{aligned}\boldsymbol{\omega}_N(x,y)&=\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\omega_{ij}\,\varphi_i(x)\varphi_j(y),\\ u_{Nx}(x,y)&=\sum_{i=1}^{N-1}\sum_{j\in J^*}u^x_{ij}\,\varphi_i(x)\varphi_j^*(y),\quad u_{Ny}(x,y)=\sum_{i\in J^*}\sum_{j=1}^{N-1}u^y_{ij}\,\varphi_i^*(x)\varphi_j(y),\\ \tilde{p}_N(x,y)&=\sum_{i\in J^*,j\in J^*,(i,j)\neq(0,0)}p_{ij}\,\varphi_i^*(x)\varphi_j^*(y).\end{aligned}$$

The function $\tilde{p}_N$ vanishes in $(-1,-1)$ but no longer belongs to $L^2_0(\Omega)$; however, the real pressure p_N can easily be recovered in a postprocessing step, thanks to the formula

$$p_N(x,y)=\tilde{p}_N(x,y)-\frac{1}{2^d}\,(\tilde{p}_N,1)_N.\tag{6.2}$$

We denote by $\Omega^\diamond$, U, and P the vectors made of these coefficients. Their dimensions are equal to $\frac{d(d-1)}{2}\,N^{d-2}(N-1)^2$, $d\,N^{d-1}(N-1)$, and N^d-1, respectively. Problem (3.7) can thus be written equivalently as the square linear system

$$\begin{pmatrix}A & 0 & B\\ 0 & B^T & 0\\ C_{\boldsymbol{\omega}} & C_{\boldsymbol{u}} & 0\end{pmatrix}\begin{pmatrix}\Omega^\diamond\\ U\\ P\end{pmatrix}=\begin{pmatrix}F\\ 0\\ 0\end{pmatrix},\tag{6.3}$$

where B^T denotes the transposed matrix of B. The global matrix is not symmetric, even if the subblocks $\left(\begin{smallmatrix}0 & B\\ B^T & 0\end{smallmatrix}\right)$ and $C_{\boldsymbol{\omega}}$ are symmetric. Note that, up to the multiplicative constant $-\nu^{-1}$, the matrix $C_{\boldsymbol{u}}$ coincides with A^T. The system is very similar in the case of nonhomogeneous boundary conditions, except that a further vector $-\tilde{B}^TG$ appears in the second line of the right-hand side.

In what follows, system (6.3) is solved via the GMRES method, so that it has not to be assembled. As a preconditioner, we use the matrix issued from an incomplete LU factorization of the global matrix in (6.3). Moreover, as standard in spectral methods, it follows from the tensorization properties of the polynomial spaces that each product of this matrix by a vector is realized with $c\,N^{d+1}$ operations, which highly reduces the cost of the inversion.

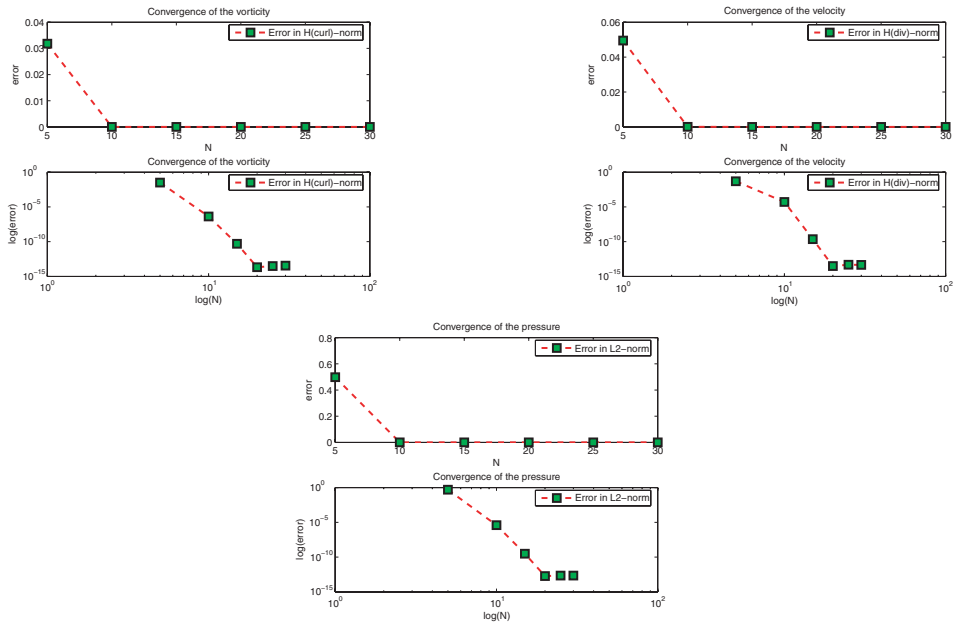


FIG. 1. *Error curves for the solution defined from* (6.4).

Two-dimensional experiments. We first work in the square $\Omega =]-1,1[^2$, with $\nu = 1$. We consider a given solution constructed thanks to the formulas $\boldsymbol{u} = \mathbf{curl}\,\psi$ and $\boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u}$ in the two situations

(i) of functions ψ and p of class $\mathcal{C}^\infty$, defined by

$$\psi(x,y) = \sin(\pi x)\sin(\pi y), \quad p(x,y) = xy, \tag{6.4}$$

and

(ii) of functions ψ and p of limited regularity, defined by

$$\psi(x,y) = (1-x^2)^3(1-y^2)^{\frac{7}{2}}, \quad p(x,y) = x(1-x^2)^{\frac{3}{2}}(1+y^2)^{-\frac{1}{2}}. \tag{6.5}$$

Figure 1 for the solution issued from (6.4) and Figure 2 for the solution issued from (6.5) present the convergence curves of the relative errors on $\boldsymbol{\omega}$, $\boldsymbol{u}$, and p in the corresponding norms, both in standard and logarithmic scales, for N varying from 5 to 30.

In Figure 1, the convergence is exponential and the three errors are smaller than 10^{-10} from $N = 15$. The convergence is, of course, slower in Figure 2. It can be noted that the vorticity and the pressure have the same regularity near the edges of Ω contained in the lines $x = \pm 1$ (they behave like $(1-x^2)^{\frac{3}{2}}$) and that the error slopes are the same.

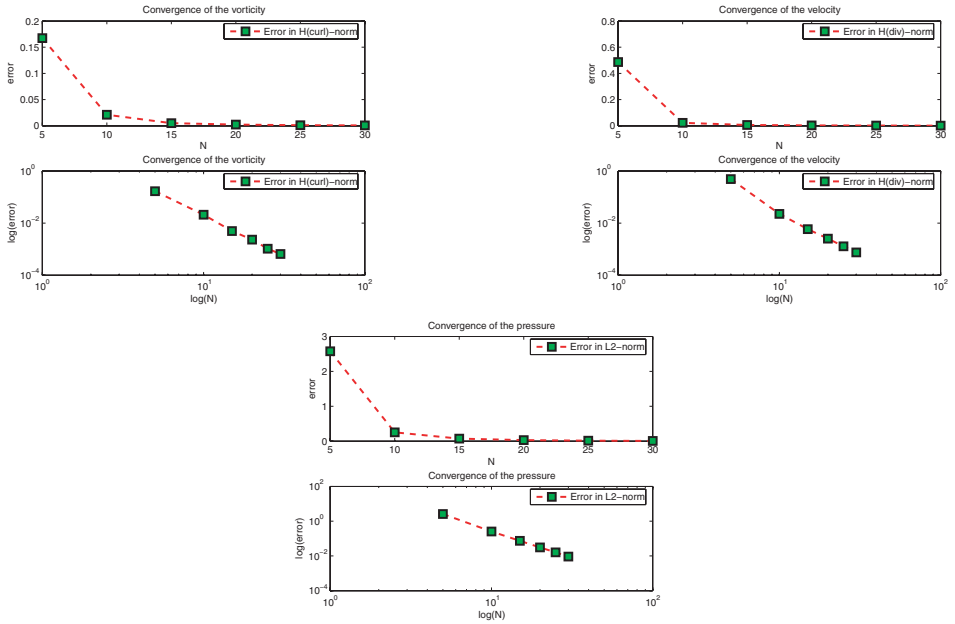


FIG. 2. *Error curves for the solution defined from* (6.5).

Figure 3 presents, from left to right and top to bottom, the values of the vorticity of the two components of the velocity and of the pressure corresponding to the data $\boldsymbol{f} = (f_x, f_y)$, with

$$f_x = 0, \quad f_y = xy^2, \tag{6.6}$$

in the case $g = 0$ of homogeneous boundary conditions, obtained with $N = 40$.

Figure 4 presents, from left to right and top to bottom, the values of the vorticity of the two components of the velocity and of the pressure corresponding to the data $\boldsymbol{f} = (f_x, f_y)$, given in (6.6) and with g given by

$$g(-1, y) = -(1 - y^2)^{\frac{3}{2}}, \qquad g(1, y) = (1 - y^2)^{\frac{3}{2}}, \qquad g(x, \pm 1) = 0, \tag{6.7}$$

obtained with $N = 40$. It can be noted that the vorticity $\boldsymbol{\omega}_N$ and pressure p_N are nearly the same in Figures 3 and 4.

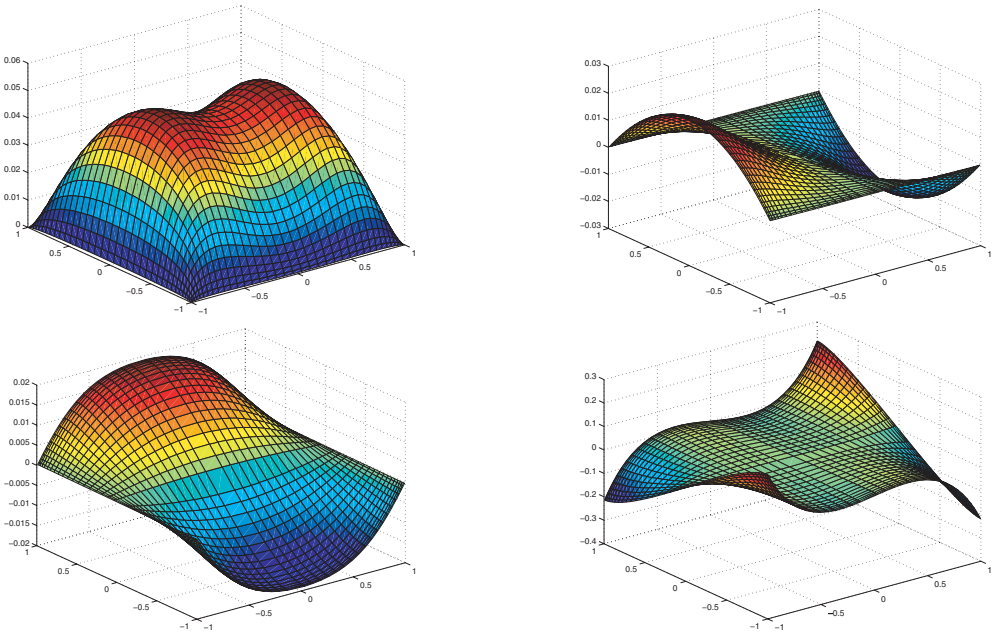


FIG. 3. *The solution* $(\boldsymbol{\omega}, u_x, u_y, p)$ *for the data* $\boldsymbol{f}$ *defined in* (6.6) *and* $g = 0$.

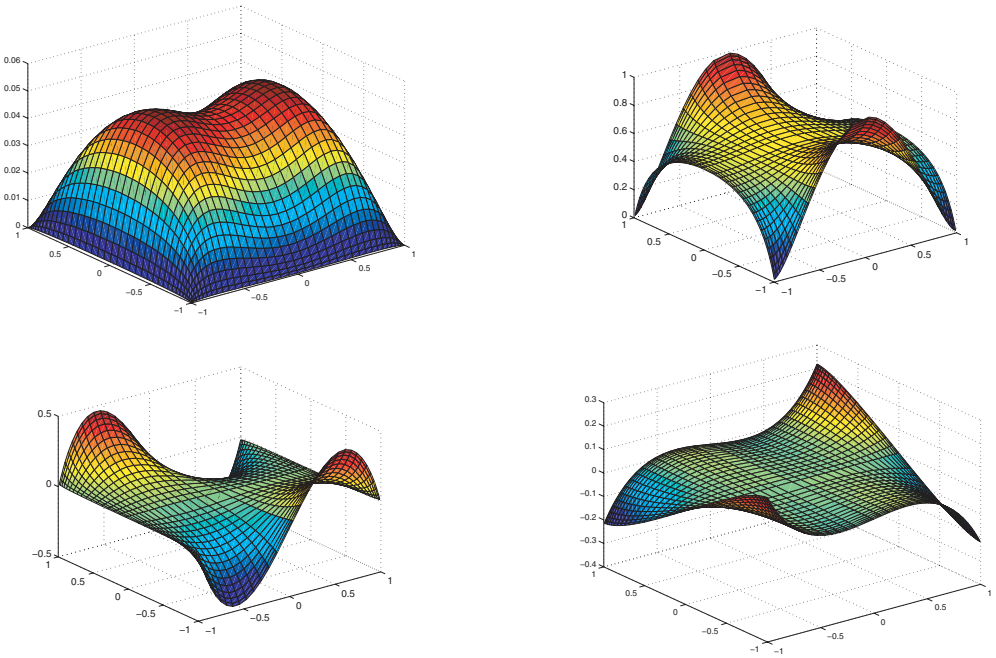


FIG. 4. *The solution* $(\boldsymbol{\omega}, u_x, u_y, p)$ *for the data* $(\boldsymbol{f}, g)$ *defined in* (6.6)–(6.7).

Three-dimensional experiments. We now work in the cube $\Omega =]-1, 1[^3$, with $\nu = 1$ and always in the case $g = 0$ of homogeneous boundary conditions. We consider a given solution constructed thanks to the formulas $\boldsymbol{u} = \mathbf{curl}\,\boldsymbol{\psi}$ and $\boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u}$, with

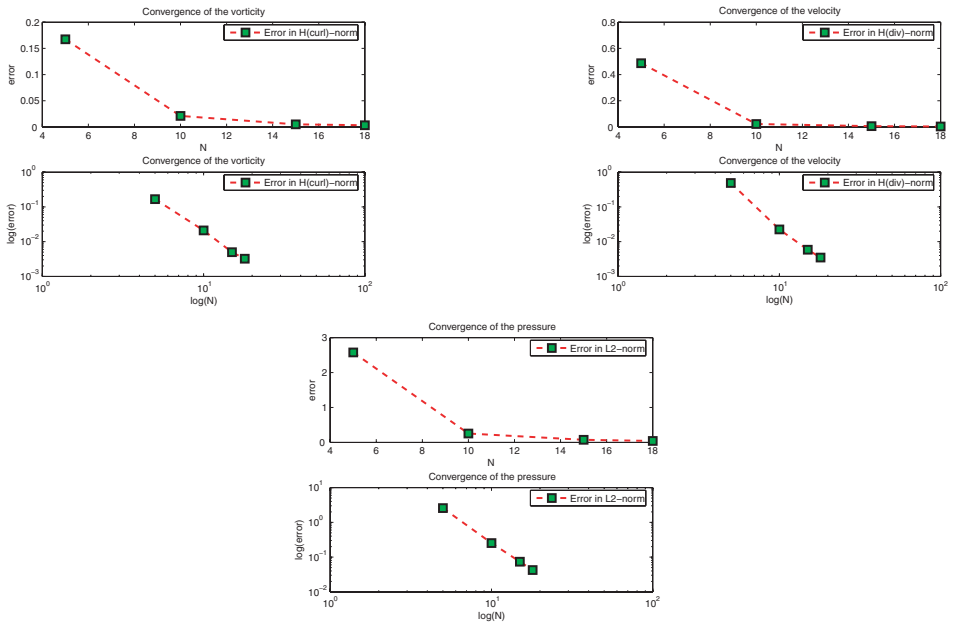


FIG. 5. *Error curves for the solution defined from* (6.8).

$\boldsymbol{\psi} = (\psi_x, \psi_y, \psi_z)$ and p defined by

$$
\begin{aligned}
&\psi_x(x,y,z) = (1-y^2)^3(1-z^2)^{\frac{7}{2}}, \quad \psi_y(x,y,z) = (1-x^2)^{\frac{7}{2}}(1-z^2)^3,\\
&\psi_z(x,y,z) = (1-x^2)^3(1-y^2)^{\frac{7}{2}}, \quad p(x,y,z) = \frac{x(1-x^2)^{\frac{3}{2}}}{(1+y^2)^{\frac{1}{2}}(1+z^2)^{\frac{1}{2}}}.
\end{aligned} \tag{6.8}
$$

Figure 5 presents the convergence curves of the relative errors on $\boldsymbol{\omega}$, $\boldsymbol{u}$, and p, both in standard and logarithmic scales, for N varying from 5 to 18. It can be noted that the regularity of the solution is the same as for the two-dimensional solution defined from (6.5) and that the slopes of the error are very similar to those in Figure 2. We do not present the convergence curves for a solution of class $\mathcal{C}^\infty$ since they are exactly the same as in Figure 1.

Figure 6 presents, from left to right and top to bottom, the curves of isovalues in the plane $x = 0$ of the three components of the vorticity and the velocity and of the pressure corresponding to the data $\boldsymbol{f} = (f_x, f_y, f_z)$, with

$$
f_x = x, \qquad f_y = 0, \qquad f_z = yz^2, \tag{6.9}
$$

obtained with $N = 18$.

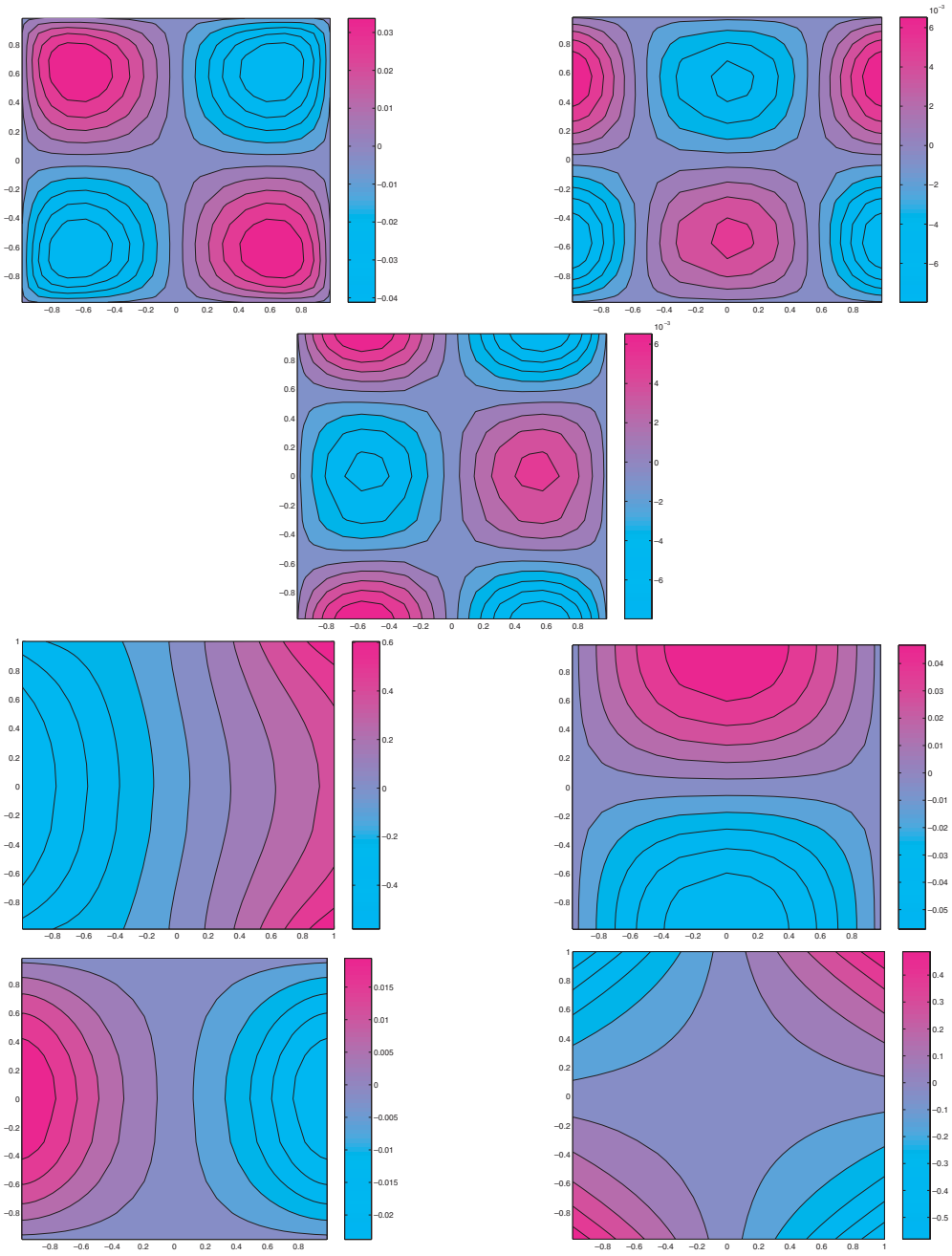


FIG. 6. *The solution* $(\omega_x, \omega_y, \omega_z, u_x, u_y, u_z, p)$ *for the data* $\boldsymbol{f}$ *defined in* (6.9).

As a conclusion, both two- and three-dimensional experiments are in perfect agreement with the error estimates derived in sections 4 and 5 and bring to light the efficiency of the spectral discretization for this type of problem.

REFERENCES

[1] M. Amara, D. Capatina-Papaghiuc, E. Chacon-Vera, and D. Trujillo, *Vorticity–velocity–pressure formulation for Navier–Stokes equations*, Comput. Vis. Sci., 6 (2004), pp. 47–52.

[2] M. Amara, D. Capatina, and D. Trujillo, *Stabilized finite element method for the Navier–Stokes equations, with non standard boundary conditions*, J. Sci. Comput., submitted.

[3] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[4] M. Azaïez, F. Ben Belgacem, M. Grundmann, and H. Khallouf, *Staggered grids hybrid-dual spectral element method for second order elliptic problems. Application to high-order time splitting for Navier-Stokes equations*, Comput. Meth. Appl. Mech. Engrg., 166 (1998), pp. 183–199.

[5] M. Azaïez, C. Bernardi, and M. Grundmann, *Méthodes spectrales pour les équations du milieu poreux*, East-West J. Numer. Math., 2 (1994), pp. 91–105.

[6] C. Bègue, C. Conca, F. Murat, and O. Pironneau, *Les équations de Stokes et de Navier–Stokes avec des conditions aux limites sur la pression*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. IX, H. Brezis and J.-L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1988, pp. 179–264.

[7] F. Ben Belgacem and C. Bernardi, *Spectral element discretization of the Maxwell equations*, Math. Comp., 68 (1999), pp. 1497–1520.

[8] J.-M. Bernard, *Spectral discretizations of the Stokes equations with non standard boundary conditions*, J. Sci. Comput., 20 (2004), pp. 355–377.

[9] C. Bernardi, M. Dauge, and Y. Maday, *Polynomials in the Sobolev World*, Internal Report, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, France, 2003.

[10] C. Bernardi and V. Girault, *Espaces duaux des domaines des opérateurs divergence et rotationnel avec trace nulle*, Internal Report, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, France, 2003.

[11] C. Bernardi, F. Hecht, and O. Pironneau, *Coupling Darcy and Stokes equations for porous media with cracks*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 7–35.

[12] C. Bernardi and Y. Maday, *Polynomial approximation of some singular functions*, Appl. Anal., 42 (1991), pp. 1–32.

[13] C. Bernardi and Y. Maday, *Spectral Methods*, in Handbook of Numerical Analysis V, P.G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 1997, pp. 209–485.

[14] C. Bernardi and Y. Maday, *Uniform inf-sup conditions for the spectral discretization of the Stokes problem*, Math. Models Methods Appl. Sci., 9 (1999), pp. 395–414.

[15] A. Buffa, M. Costabel, and M. Dauge, *Algebraic convergence for anisotropic edge elements in polyhedral domains*, Numer. Math., 101 (2005), pp. 29–65.

[16] M. Costabel and M. Dauge, *Espaces fonctionnels Maxwell: Les gentils, les méchants et les singularités*, http://perso.univ-rennes1.fr/monique.dauge (1998).

[17] M. Costabel and M. Dauge, *Computation of resonance frequencies for Maxwell equations in non-smooth domains*, in Topics in Computational Wave Propagation, M. Ainsworth, P. Davies, D. Duncan, P. Martin, and B. Rynne, eds., Springer, Berlin, 2003, pp. 125–161.

[18] F. Dubois, *Vorticity–velocity–pressure formulation for the Stokes problem*, Math. Methods Appl. Sci., 25 (2002), pp. 1091–1119.

[19] F. Dubois, M. Salaün, and S. Salmon, *Vorticity–velocity–pressure and stream function-vorticity formulations for the Stokes problem*, J. Math. Pures Appl. (9), 82 (2003), pp. 1395–1451.

[20] A. Ern, J.-L. Guermond, and L. Quartapelle, *Vorticity–velocity formulation of the Stokes problem in 3D*, Math. Methods Appl. Sci., 22 (1999), pp. 531–546.

[21] V. Girault, *Incompressible finite element methods for Navier–Stokes equations with nonstandard boundary conditions in* $\mathbb{R}^3$, Math. Comp., 51 (1988), pp. 55–74.

[22] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer, Berlin, 1986.

[23] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.

[24] J.-C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[25] F.-X. Roux, *Utilisation d'éléments finis conformes* $H(\mathrm{rot})$ *pour la résolution de l'équation de Navier-Stokes tridimensionnelle*, Séminaire du Laboratoire d'Analyse Numérique, Paris, France, 1984.

[26] S. Salmon, *Développement numérique de la formulation tourbillon–vitesse–pression pour le problème de Stokes*, Thesis, Université Pierre et Marie Curie, Paris, France, 1999.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 851–864

A FOURTH-ORDER COMMUTATOR-FREE EXPONENTIAL INTEGRATOR FOR NONAUTONOMOUS DIFFERENTIAL EQUATIONS*

MECHTHILD THALHAMMER†

Abstract. In the present work, we study the convergence behavior of commutator-free exponential integrators for abstract nonautonomous evolution equations

$$u'(t) = A(t)u(t), \qquad 0 < t \le T.$$

In particular, we focus on a fourth-order scheme that relies on the composition of two exponentials involving the values of the linear operator family A at the Gaussian nodes

$$u_1 = \mathrm{e}^{h(a_2A_1+a_1A_2)}\,\mathrm{e}^{h(a_1A_1+a_2A_2)}\,u_0, \quad a_i = \tfrac{1}{4} \pm \tfrac{\sqrt{3}}{6}, \quad c_i = \tfrac{1}{2} \mp \tfrac{\sqrt{3}}{6}, \quad A_i = A(c_ih), \quad i = 1, 2.$$

We prove that the numerical scheme is stable and derive an error estimate with respect to the norm of the underlying Banach space. The theoretically expected order reduction is illustrated by a numerical example for a parabolic initial-boundary value problem subject to a homogeneous Dirichlet boundary condition.

Key words. exponential integrators, commutator-free methods, nonautonomous differential equations, parabolic evolution equations, stability, convergence

AMS subject classifications. 65L05, 65M12, 65J10

DOI. 10.1137/05063042

1. Introduction. In the present paper, we consider a nonautonomous differential equation involving a time-dependent linear operator A

$$u'(t) = A(t)u(t), \quad 0 < t \le T, \qquad u(0) \text{ given.} \tag{1.1}$$

Our setting includes parabolic initial-boundary value problems that take the form (1.1) when written as an abstract initial value problem on a Banach space. The objective of this work is to analyze the error behavior of the fourth-order commutator-free exponential integrator

$$\begin{aligned} u_1 &= \mathrm{e}^{h(a_2A_1+a_1A_2)}\,\mathrm{e}^{h(a_1A_1+a_2A_2)}\,u_0, \\ a_i &= \tfrac{1}{4} \pm \tfrac{\sqrt{3}}{6}, \quad c_i = \tfrac{1}{2} \mp \tfrac{\sqrt{3}}{6}, \quad A_i = A(c_ih), \qquad i = 1, 2, \end{aligned} \tag{1.2}$$

to explain the substantial order reduction for problems of parabolic type. For that purpose, we derive a representation for the defect of (1.2) which remains valid within the framework of sectorial operators and analytic semigroups. In situations where $A(t)$ is a bounded linear operator, the Campbell–Baker–Hausdorff formula is a powerful tool for the error analysis of (1.2) and higher-order schemes, respectively. However, it is problematic to justify its validity in the context of parabolic evolution equations.

*Received by the editors May 2, 2005; accepted for publication (in revised form) November 2, 2005; published electronically April 12, 2006. This work was supported by Fonds zur Förderung der wissenschaftlichen Forschung (FWF) under project H210-N13.

http://www.siam.org/journals/sinum/44-2/63042.html

†Institut für Mathematik, Universität Innsbruck, Technikerstrasse 13, 6020 Innsbruck, Austria (Mechthild.Thalhammer@uibk.ac.at).

Therefore, in this paper, we follow a different approach based on the variation-of-constants formula.

Numerical schemes that involve the evaluation of the exponential and related functions were proposed in the middle of the past century; for a historical review, see [24]. At present, a variety of works confirms the renewed interest in such exponential integrators; as a small selection, we mention the recent works [5, 8, 14, 16, 19, 20] and refer to the references given therein. A reason for these research activities are advances in the computation of the product of a matrix exponential with a vector; see, for instance, [10, 15, 25]. As a consequence, numerical integrators based on the Magnus expansion [23] and related method classes [2, 3, 6, 7, 17, 21] are practicable in the numerical solution of nonautonomous stiff differential equations; see also [11, 12, 30] and references cited therein.

The excellent error behavior of interpolatory Magnus integrators for time-dependent Schrödinger equations is explained in Hochbruck and Lubich [14]. There, it is proved that the exponential midpoint rule applied to ordinary differential equations (1.1)

$$u_1 = \mathrm{e}^{hA_1}\, u_0, \qquad A_1 = A\big(\tfrac{h}{2}\big), \tag{1.3}$$

is convergent of order 2 without any restriction on the size of $h\big\|A(t)\big\|$. Moreover, under a mild stepsize restriction, a fourth-order error bound is valid for the Magnus integrator

$$u_1 = \mathrm{e}^{ha_1(A_1+A_2)+h^2a_2[A_2,A_1]}\, u_0,$$
$$a_1 = \tfrac{1}{2}, \quad a_2 = \tfrac{\sqrt{3}}{12}, \quad c_i = \tfrac{1}{2} \mp \tfrac{\sqrt{3}}{6}, \quad A_i = A(c_i h), \qquad i = 1, 2,$$

where $[A_1, A_2] = A_1A_2 - A_2A_1$ denotes the matrix commutator. In [11], we considered the numerical scheme (1.3) in the context of parabolic evolution equations and showed that the full convergence order 2 is obtained when the error is measured in the norm of the underlying Banach space, provided that the data and the exact solution of (1.1) are sufficiently smooth in time.

The purpose of the present work is to investigate the convergence properties of higher-order methods for linear nonautonomous parabolic problems (1.1). Provided that the time-dependent sectorial operator $A(t)$ is Hölder-continuous with respect to t, it is ensured that any linear operator defined through $B = \alpha A(\xi_1) + (1-\alpha)A(\xi_2)$ with $\alpha, \xi_1, \xi_2 \in \mathbb{R}$ generates an analytic semigroup $\big(\mathrm{e}^{tB}\big)_{t\geq 0}$, that is, numerical schemes such as (1.2) remain well defined for abstract evolution equations (1.1). For that reason, we focus on commutator-free exponential integrators that rely on the composition of exponentials involving linear combinations of values of A. We show that the fourth-order scheme (1.2) is stable; however, unless the operator familiy A fulfills unnatural requirements, a substantial order reduction is encountered. For instance, for one-dimensional initial-boundary value problems subject to a homogeneous Dirichlet boundary condition, the order of convergence with respect to a discrete L^p-norm is $2 + \kappa$, where $0 \leq \kappa < (2p)^{-1}$, in general.

The present work is organized as follows. In section 2, we first state the fundamental hypotheses on the nonautonomous evolution equation (1.1). The considered commutator-free exponential integration scheme is then introduced in section 3. The numerical approximation is based on the composition of two exponentials that involve the values of A at certain nodal points. Sections 4 and 5 are concerned with a stability and convergence analysis for parabolic problems. In section 5.1, we derive

an expansion of the numerical solution defect which remains well defined for abstract differential equations (1.1) involving an unbounded linear operator $A(t)$, provided that the data and the exact solution of the problem are sufficiently many times differentiable with respect to time. The main result, a convergence estimate for the fourth-order scheme (1.2), is given in section 5.2. Important tools for its proof are the stability bound and the representation of the defect derived before. Section 6 is devoted to a numerical example that illustrates the expected order reduction.

2. Parabolic problems. Henceforth, we denote by $(X, \|\cdot\|_X)$ the underlying Banach space. Our basic requirements on the unbounded linear operator family A defining the right-hand side of the differential equation in (1.1) are that of [11, 30]. For a detailed treatise of time-dependent evolution equations we refer to [22, 29]. The monographs [13, 27] delve into the theory of sectorial operators and analytic semigroups.

HYPOTHESIS 1. *We assume that the densly defined and closed linear operator $A(t) : D \subset X \to X$ is uniformly sectorial for $0 \le t \le T$. Thus, there exist constants $a \in \mathbb{R}$, $0 < \phi < \frac{\pi}{2}$, and $M > 0$ such that for all $0 \le t \le T$ the resolvent of $A(t)$ satisfies the condition*

$$\left\| \big(\lambda I - A(t)\big)^{-1} \right\|_{X \leftarrow X} \le \frac{M}{|\lambda - a|} \tag{2.1}$$

for any complex number $\lambda \notin S_\phi(a) = \big\{\lambda \in \mathbb{C} : |\arg(a - \lambda)| \le \phi\big\} \cup \{a\}$. The graph norm of $A(t)$ and the norm in D fulfill the following relation with a constant $K > 0$:

$$K^{-1}\|x\|_D \le \|x\|_X + \big\|A(t)x\big\|_X \le K\|x\|_D, \qquad x \in D, \qquad 0 \le t \le T.$$

Moreover, it holds $A \in \mathscr{C}^\vartheta\big([0,T], L(D,X)\big)$ for some $0 < \vartheta \le 1$, i.e., the bound

$$\big\|A(t) - A(s)\big\|_{X \leftarrow D} \le L(t-s)^\vartheta, \qquad 0 \le s \le t \le T, \tag{2.2}$$

is valid with a constant $L > 0$.

For any $0 \le s \le T$ the sectorial operator $\Omega = A(s)$ generates an analytic semigroup $\big(\mathrm{e}^{t\Omega}\big)_{t \ge 0}$ on X which is defined by means of the integral formula of Cauchy

$$\mathrm{e}^{t\Omega} = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^\lambda \big(\lambda I - t\Omega\big)^{-1} \,\mathrm{d}\lambda, \quad t > 0, \qquad \mathrm{e}^{t\Omega} = I, \quad t = 0. \tag{2.3}$$

Here, Γ denotes a path that surrounds the spectrum of Ω.

Henceforth, for $0 < \mu < 1$, we denote by X_μ some intermediate space between the Banach spaces $D = X_1$ and $X = X_0$ such that the norm in X_μ satisfies the bound $\|x\|_{X_\mu} \le K\|x\|_D^\mu \|x\|_X^{1-\mu}$ with a constant $K > 0$ for all elements $x \in D$. Examples for intermediate spaces are real interpolation spaces (see Lunardi [22]) or fractional power spaces (see Henry [13]). Then, for all $0 \le \mu \le \nu \le 1$ and integers $k \ge 0$ the following bound is valid:

$$\big\|t^{k+\nu-\mu}\, \Omega^k\, \mathrm{e}^{t\Omega}\big\|_{X_\nu \leftarrow X_\mu} \le M, \qquad 0 \le t \le T, \tag{2.4}$$

with a constant $M > 0$. As a consequence, the linear operator φ_m which is given by

$$\varphi_m(t\Omega) = \frac{1}{(m-1)!\, t^m} \int_0^t \mathrm{e}^{(t-\tau)\Omega}\, \tau^{m-1} \mathrm{d}\tau, \quad t > 0, \qquad \varphi_m(0) = \frac{1}{(m-1)!}\, I, \tag{2.5a}$$

for integers $m \geq 1$, remains bounded on X_μ for any $0 \leq t \leq T$ and $0 \leq \mu \leq 1$. In the subsequent sections, we make use of the identities

$$\mathrm{e}^{t\Omega} = I + t\Omega\,\varphi_1(t\Omega), \qquad \varphi_{m-1}(t\Omega) = \frac{1}{(m-1)!}\,I + t\Omega\,\varphi_m(t\Omega), \quad m \geq 1. \tag{2.5b}$$

Furthermore, it is substantial that the relation

$$\varphi_m(t\Omega) - \varphi_m(0) = t\Omega\,\chi(t\Omega) \tag{2.5c}$$

holds with a linear operator $\chi(t\Omega)$ that is bounded on X_μ; see [13, 22] and also [11, 30].

3. Commutator-free exponential integrator. In this section, we introduce an integration method for linear nonautonomous parabolic problems (1.1) which relies on the composition of two exponentials. We note that the considered scheme is an example of a *Crouch–Grossman method* [9].

For a constant stepsize $h > 0$ the associated grid points are denoted by $t_j = jh$ for $j \geq 0$. The numerical approximation $u_{n+1} \approx u(t_{n+1})$ to the true solution of (1.1) is given by the recurrence formula

$$u_{n+1} = \mathrm{e}^{\widetilde{\zeta} h C_n}\,\mathrm{e}^{\zeta h B_n}\,u_n, \qquad n \geq 0. \tag{3.1a}$$

Here, we employ the following abbreviations:

$$\begin{aligned} A_{ni} &= A(t_n + c_i h), \qquad i = 1, 2, \\ B_n &= \alpha A_{n1} + \beta A_{n2}, \qquad C_n = \gamma A_{n1} + \delta A_{n2}. \end{aligned} \tag{3.1b}$$

Throughout, we assume that the nodal points $\zeta, \widetilde{\zeta}, c_1, c_2 \in \mathbb{R}$ and the coefficients $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ satisfy

$$0 < \zeta < 1, \qquad \widetilde{\zeta} = 1 - \zeta, \qquad 0 \leq c_1 \leq c_2 \leq 1, \qquad \alpha + \beta = 1, \qquad \gamma + \delta = 1. \tag{3.1c}$$

The following remark shows that relation (3.1a) remains well defined within the analytical framework of section 2.

Remark 1. Under the assumptions of Hypothesis 1, the linear operator

$$\alpha A_{n1} + (1-\alpha) A_{n2} = A_{n2} + \alpha\big(A_{n1} - A_{n2}\big), \qquad \alpha \in \mathbb{R},$$

is sectorial; see also [13, Theorem 1.3.2]. Therefore, the commutator-free exponential integrator (3.1) is well defined for abstract evolution equations (1.1).

4. Stability. The stability properties of the commutator-free exponential integrator (3.1) are determined by the behavior of the evolution operator

$$\prod_{i=m}^{n} \mathrm{e}^{\widetilde{\zeta} h C_i}\,\mathrm{e}^{\zeta h B_i} = \mathrm{e}^{\widetilde{\zeta} h C_n}\,\mathrm{e}^{\zeta h B_n}\,\mathrm{e}^{\widetilde{\zeta} h C_{n-1}}\,\mathrm{e}^{\zeta h B_{n-1}}\,\cdots\,\mathrm{e}^{\widetilde{\zeta} h C_m}\,\mathrm{e}^{\zeta h B_m}, \tag{4.1}$$

where $n \geq m \geq 0$. The following result implies that the numerical solution u_n remains bounded for arbitrarily chosen stepsizes $h > 0$ as long as $nh \leq T$.

THEOREM 1 (stability). *Under the requirements of Hypothesis* 1 *on* A*, the discrete evolution operator* (4.1) *fulfills the bound*

$$\left\| \prod_{i=m}^{n} \mathrm{e}^{\widetilde{\zeta} h C_i}\,\mathrm{e}^{\zeta h B_i} \right\|_{X \leftarrow X} \leq M, \qquad 0 \leq mh \leq nh \leq T,$$

with a constant $M > 0$ that does not depend on n and h.

Proof. As in our preceeding works [11, 30], the proof of the above stability result relies on the telescopic identity and the integral formula of Cauchy. In the present situation, it is useful to compare the discrete evolution operator (4.1) with the linear operator

$$\prod_{i=m}^{n} \mathrm{e}^{\widetilde{\zeta} h A_{m2}} \, \mathrm{e}^{\zeta h A_{m2}} = \prod_{i=m}^{n} \mathrm{e}^{h A_{m2}} = \mathrm{e}^{(t_{n+1}-t_m) A_{m2}},$$

which satisfies the well-known bound

$$\left\| \mathrm{e}^{(t_{n+1}-t_m) A_{m2}} \right\|_{X \leftarrow X} + \left\| (t_{n+1} - t_m) \, A_{m2} \, \mathrm{e}^{(t_{n+1}-t_m) A_{m2}} \right\|_{X \leftarrow X} \leq C$$

for $0 \leq t_m \leq t_n \leq T$. Therefore, it suffices to estimate the difference

$$\begin{aligned}
\Delta_m^n &= \prod_{i=m}^{n} \mathrm{e}^{\widetilde{\zeta} h C_i} \, \mathrm{e}^{\zeta h B_i} - \mathrm{e}^{(t_{n+1}-t_m) A_{m2}} \\
&= \sum_{j=m}^{n-1} \Delta_{j+1}^n \Big(\mathrm{e}^{\widetilde{\zeta} h C_j} \, \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} \Big) \mathrm{e}^{(t_j - t_m) A_{m2}} \\
&\qquad + \sum_{j=m}^{n} \mathrm{e}^{(t_{n+1}-t_{j+1}) A_{m2}} \Big(\mathrm{e}^{\widetilde{\zeta} h C_j} \, \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} \Big) \mathrm{e}^{(t_j - t_m) A_{m2}}.
\end{aligned}$$

For this purpose, it is notable that the following relation holds true:

$$\mathrm{e}^{\widetilde{\zeta} h C_j} \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} = \Big(\mathrm{e}^{\widetilde{\zeta} h C_j} - \mathrm{e}^{\widetilde{\zeta} h A_{m2}} \Big) \, \mathrm{e}^{\zeta h B_j} + \mathrm{e}^{\widetilde{\zeta} h A_{m2}} \Big(\mathrm{e}^{\zeta h B_j} - \mathrm{e}^{\zeta h A_{m2}} \Big).$$

By means of the integral formula of Cauchy, the resolvent identity

$$(\lambda I - \Omega_1)^{-1} - (\lambda I - \Omega_2)^{-1} = (\lambda I - \Omega_1)^{-1} (\Omega_1 - \Omega_2) (\lambda I - \Omega_2)^{-1},$$

and the relations given in (3.1), we receive

$$\begin{aligned}
&\Big(\mathrm{e}^{\widetilde{\zeta} h C_j} \, \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} \Big) \, \mathrm{e}^{(t_j - t_m) A_{m2}} \\
&\quad = \frac{\widetilde{\zeta} h}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^{\lambda} \, (\lambda - \widetilde{\zeta} h C_j)^{-1} \big(\gamma (A_{j1} - A_{j2}) + A_{j2} - A_{m2} \big) \\
&\qquad\qquad \times (\lambda - \widetilde{\zeta} h A_{m2})^{-1} \, \mathrm{e}^{\zeta h B_j} \, \mathrm{e}^{(t_j - t_m) A_{m2}} \, \mathrm{d}\lambda \\
&\qquad + \frac{\zeta h}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^{\lambda} \, \mathrm{e}^{\widetilde{\zeta} h A_{m2}} \, (\lambda - \zeta h B_j)^{-1} \big(\alpha (A_{j1} - A_{j2}) + A_{j2} - A_{m2} \big) \\
&\qquad\qquad \times (\lambda - \zeta h A_{m2})^{-1} \, \mathrm{e}^{(t_j - t_m) A_{m2}} \, \mathrm{d}\lambda.
\end{aligned}$$

With the help of the resolvent bound (2.1), the Hölder estimate (2.2) for A, and (2.4) it thus follows

$$\begin{aligned}
\left\| \Big(\mathrm{e}^{\widetilde{\zeta} h C_j} \, \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} \Big) \, \mathrm{e}^{(t_j - t_m) A_{m2}} \right\|_{X \leftarrow X} &\leq M h^{\vartheta}, \qquad j = m, \\
\left\| \Big(\mathrm{e}^{\widetilde{\zeta} h C_j} \, \mathrm{e}^{\zeta h B_j} - \mathrm{e}^{h A_{m2}} \Big) \, \mathrm{e}^{(t_j - t_m) A_{m2}} \right\|_{X \leftarrow X} &\leq M h (t_j - t_m)^{-1+\vartheta}, \qquad j > m.
\end{aligned}$$

Consequently, a further application of (2.4) together with a Gronwall-type inequality with a weakly singular kernel (see also [4, 26]) yields the desired stability bound. ∎

5. Convergence. In this section, we analyze the convergence behavior of the considered commutator-free exponential integrator for parabolic problems (1.1). As a first step, we next derive a useful relation for the defect of (3.1) by means of a suitable linearization of the differential equation and an application of the variation-of-constants formula. Similar techniques have been used in the study of exponential splitting methods; see [1, 18, 28] and references therein. The following considerations also explain the definition of the numerical method.

5.1. Expansion of the defect. Replacing in (3.1) the numerical by the exact solution values defines the defect of the method

$$u(t_{n+1}) = \mathrm{e}^{\widetilde{\zeta} h C_n}\, \mathrm{e}^{\zeta h B_n}\, u(t_n) + d_{n+1}, \qquad n \geq 0. \tag{5.1}$$

Our basic approach is to consider the initial value problem (1.1) on the subinterval $[t_n, t_{n+1}]$ and to derive an analogous relation to (3.1a) for the exact solution values. For that purpose, we set

$$G_n(t) = \big(A(t) - B_n\big)u(t), \qquad H_n(t) = \big(A(t) - C_n\big)u(t). \tag{5.2}$$

On the one hand, rewriting the right-hand side of the differential equation in (1.1) as $u'(t) = B_n u(t) + G_n(t)$ and applying the variation-of-constants formula (see [22]) yields the following relation for the solution value at time $t_n + \zeta h$:

$$u(t_n + \zeta h) = \mathrm{e}^{\zeta h B_n}\, u(t_n) + \int_0^{\zeta h} \mathrm{e}^{(\zeta h - \tau) B_n}\, G_n(t_n + \tau)\, \mathrm{d}\tau.$$

On the other hand, by linearizing (1.1) around C_n and inserting the above representation for $u(t_n + \zeta h)$, we further obtain

$$\begin{aligned} u(t_{n+1}) = \mathrm{e}^{\widetilde{\zeta} h C_n}\, \mathrm{e}^{\zeta h B_n}\, u(t_n) + \mathrm{e}^{\widetilde{\zeta} h C_n} \int_0^{\zeta h} \mathrm{e}^{(\zeta h - \tau) B_n}\, G_n(t_n + \tau)\, \mathrm{d}\tau \\ + \int_0^{\widetilde{\zeta} h} \mathrm{e}^{(\widetilde{\zeta} h - \tau) C_n}\, H_n(t_n + \zeta h + \tau)\, \mathrm{d}\tau. \end{aligned}$$

Consequently, the defect of the numerical method (3.1) equals

$$d_{n+1} = \mathrm{e}^{\widetilde{\zeta} h C_n} \int_0^{\zeta h} \mathrm{e}^{(\zeta h - \tau) B_n}\, G_n(t_n + \tau)\, \mathrm{d}\tau + \int_0^{\widetilde{\zeta} h} \mathrm{e}^{(\widetilde{\zeta} h - \tau) C_n}\, H_n(t_n + \zeta h + \tau)\, \mathrm{d}\tau. \tag{5.3}$$

In order to derive a suitable expansion of d_{n+1}, it is useful to introduce some additional notation.

The time derivatives of the linear operator A and the exact solution u of (1.1) at time t_n are denoted by

$$A_n^{(i)} = A^{(i)}(t_n), \quad i \geq 0, \qquad \widehat{u}_n^{(j)} = u^{(j)}(t_n), \quad j \geq 0. \tag{5.4a}$$

For the coefficients of the numerical scheme, we define

$$\mu_i = \alpha c_1^i + \beta c_2^i, \qquad \nu_i = \gamma c_1^i + \delta c_2^i, \qquad i = 1, 2, 3; \tag{5.4b}$$

see (3.1). We note that for a sufficiently differentiable function $f : [t_n, t_{n+1}] \to X$ a Taylor series expansion yields

$$
\begin{aligned}
f(t_n+\tau) &= \sum_{i=0}^{m} \frac{\tau^i}{i!}\, f_n^{(i)} + R\big(\tau^{m+1}, f^{(m+1)}\big), \qquad 0 \le \tau \le h,\\
R\big(\tau^{m+1}, f^{(m+1)}\big) &= \frac{1}{m!}\,\tau^{m+1} \int_0^1 (1-\sigma)^m f^{(m+1)}(t_n+\sigma\tau)\,\mathrm{d}\sigma,
\end{aligned}
\tag{5.5}
$$

where $f_n^{(i)} = f^{(i)}(t_n)$. Thus, provided that the quantity

$$
\big\|f^{(m+1)}\big\|_{X,\infty} = \max_{t_n \le t \le t_{n+1}} \big\|f^{(m+1)}(t)\big\|_X
$$

is well defined, the remainder fulfills

$$
\big\|R\big(\tau^{m+1}, f^{(m+1)}\big)\big\|_X \le M h^{m+1} \big\|f^{(m+1)}\big\|_{X,\infty}, \qquad 0 \le \tau \le h,
$$

with some constant $M > 0$. Terms that satisfy an estimate of this form are henceforth denoted by $\mathscr{R}\big(h^{m+1}, f^{(m+1)}\big)$. In particular, the abbrevitation $\mathscr{R}\big(h^k, A^{(i)}\, u^{(j)}\big)$ signifies that the bound

$$
\big\|\mathscr{R}\big(h^k, A^{(i)}\, u^{(j)}\big)\big\|_X \le M h^k \max_{t_n \le s,t \le t_{n+1}} \big\|A^{(i)}(s)\, u^{(j)}(t)\big\|_{X,\infty}
$$

holds true.

Provided that the involved derivatives of A and u are well defined, the following representation is valid for the defect d_{n+1} given by (5.1). We recall formula (2.5a) for the linear operator φ_m.

LEMMA 1. *The numerical solution defect of* (3.1) *fulfills the relation*

$$
\begin{aligned}
d_{n+1} = \sum_{(i,j)\in\mathscr{J}} & h^{i+j+1}\, \Phi_{ij}\, A_n^{(i)} \widehat{u}_n^{(j)} + \mathscr{R}\big(h^5, A^{(4)} u\big)\\
& + \mathscr{R}\big(h^5, A''' u'\big) + \mathscr{R}\big(h^5, A'' u''\big) + \mathscr{R}\big(h^5, A' u'''\big),
\end{aligned}
$$

where $\Phi_{ij} = \Phi_{ij}(hB_n, hC_n)$ *is defined through*

$$
\begin{aligned}
\Phi_{ij} = \frac{1}{i!j!} \Big\{ & \zeta^{j+1} \mathrm{e}^{\widetilde{\zeta} hC_n} \Big((i+j)!\, \zeta^i \varphi_{i+j+1}(\zeta hB_n) - j!\, \mu_i \varphi_{j+1}(\zeta hB_n) \Big)\\
& + \sum_{\ell=j+1}^{i+j} \ell!\, \tbinom{i+j}{\ell} \zeta^{i+j-\ell} \widetilde{\zeta}^{\ell+1} \varphi_{\ell+1}(\widetilde{\zeta} hC_n)\\
& + \sum_{\ell=0}^{j} \ell!\, \zeta^{j-\ell} \widetilde{\zeta}^{\ell+1} \Big(\tbinom{i+j}{\ell} \zeta^i - \nu_i \tbinom{j}{\ell} \Big) \varphi_{\ell+1}(\widetilde{\zeta} hC_n) \Big\}
\end{aligned}
$$

and $\mathscr{J} = \big\{(1,0),(2,0),(1,1),(3,0),(2,1),(1,2)\big\}$.

Proof. We first derive a useful relation for the maps G_n and H_n defined in (5.2). With the help of (5.5), by combining the expansions

$$
\begin{aligned}
A(t_n+\tau) - B_n &= \sum_{i=0}^{3} \frac{1}{i!} \big(\tau^i - \mu_i h^i\big) A_n^{(i)} + \mathscr{R}\big(h^4, A^{(4)}\big),\\
u(t_n+\tau) &= \sum_{j=0}^{2} \frac{1}{j!}\, \tau^j \widehat{u}_n^{(j)} + R\big(\tau^3, u^{(3)}\big),
\end{aligned}
$$

we receive the following representation:

$$G_n(t_n+\tau) = \sum_{(i,j)\in\mathscr{J}} \frac{1}{i!j!}\big(\tau^i - \mu_i h^i\big)\tau^j A_n^{(i)}\widehat{u}_n^{(j)} + \mathscr{R}\big(h^4\big), \tag{5.6a}$$

$$\mathscr{R}\big(h^4\big) = \mathscr{R}\big(h^4, A^{(4)}u\big) + \mathscr{R}\big(h^4, A'''u'\big) + \mathscr{R}\big(h^4, A''u''\big) + \mathscr{R}\big(h^4, A'u'''\big);$$

see also (3.1b)–(3.1c) and (5.4). Similarly, it follows

$$H_n(t_n+\zeta h+\tau) = \sum_{(i,j)\in\mathscr{J}} \frac{1}{i!j!}\big((\zeta h+\tau)^i - \nu_i h^i\big)(\zeta h+\tau)^j A_n^{(i)}\widehat{u}_n^{(j)} + \mathscr{R}\big(h^4\big). \tag{5.6b}$$

We next insert the above expansions (5.6) into (5.3) and express the resulting integrals by means of (2.5a). Altogether, this yields the given result. ∎

In the situation of Section 2, a reasonable smoothness assumption on (1.1) is that the linear operator A and the exact solution u are sufficiently differentiable with respect to the variable t. Precisely, we suppose $A^{(4)}(t)$ and $u^{(4)}(t)$ to be bounded in the underlying Banach space X for all $0 \le t \le T$. The following remark states that then the expansion of Lemma 1 is well-defined. However, unless the exact solution satisfies additional (unnatural) requirements such as $A'(t)u(t) \in D$ for $0 \le t \le T$, in general, it is not possible to further expand the defect.

Remark 2. Provided that $u'(t) \in X$ it follows from the differential equation in (1.1) that $A(t)u(t) \in X$ and therefore $u(t) \in D$ for $0 \le t \le T$. Differentiating (1.1) with respect to the variable t implies $A(t)u'(t) = u''(t) - A'(t)u(t) \in X$, and, as a consequence, $u'(t) \in D$ for any $0 \le t \le T$. Similarly, it follows $u^{(j-1)}(t) \in D$ if $u^{(j)}(t) \in X$ for $0 \le t \le T$ and $j = 3, 4$. Thus, under the regularity requirements $A \in \mathscr{C}^4\big([0,T], L(D,X)\big)$ and $u \in \mathscr{C}^4\big([0,T], X\big)$, the representation of the defect given in Lemma 1 is well defined.

5.2. Error estimate. With the help of the stability estimate and the expansion of the defect given in sections 4 and 5.1, we are able to prove the following convergence result.

THEOREM 2 (convergence). *Assume that the requirements of Hypothesis* 1 *are fulfilled and that further* $A \in \mathscr{C}^4\big([0,T], L(D,X)\big)$ *and* $u \in \mathscr{C}^4\big([0,T], X\big)$. *Then, provided that* $A^{(i)}(t)\,u^{(j)}(t)$ *belongs to the intermediate space* X_κ *with* $0 \le \kappa < 1$ *for* $0 \le t \le T$ *and* $(i,j) \in \big\{(1,0),(2,0),(1,1)\big\}$, *the fourth-order commutator-free exponential integrator* (1.2) *satisfies the error estimate*

$$\|u_n - u(t_n)\|_X \le C\Big(\big\|u_0 - u(0)\big\|_X + h^{2+\kappa}\big(1 + |\log h|\big)\Big), \qquad 0 \le t_n \le T,$$

with some constant $C > 0$ *independent of* n *and* h.

Proof. In order to obtain a suitable relation for the global error $e_n = u_n - u(t_n)$, we first resolve the recurrence formula (3.1a) for the numerical approximation

$$u_n = \prod_{i=0}^{n-1} \mathrm{e}^{\widetilde{\zeta}hC_i}\,\mathrm{e}^{\zeta hB_i}\,u_0, \qquad n \ge 0.$$

Furthermore, by using (5.1), we receive $e_n = e_n^{(1)} + e_n^{(2)}$ with

$$e_n^{(1)} = \prod_{i=0}^{n-1} \mathrm{e}^{\widetilde{\zeta}hC_i}\,\mathrm{e}^{\zeta hB_i}\big(u_0 - u(0)\big), \qquad e_n^{(2)} = -\sum_{j=0}^{n-1}\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta}hC_i}\,\mathrm{e}^{\zeta hB_i}\,d_{j+1}. \tag{5.7}$$

We next estimate the terms in (5.7) with respect to the norm of the underlying Banach space X. An application of Theorem 1 shows that the first term is bounded by a constant times the error of the initial value

$$\big\|e_n^{(1)}\big\|_X \le \left\|\prod_{i=0}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\right\|_{X\leftarrow X} \|u_0 - u(0)\|_X \le C\|u_0 - u(0)\|_X.$$

For estimating the second term $e_n^{(2)}$, we employ the representation of the defect given in Lemma 1. Making use of the fact that the sums involving the remainder and the terms where $i + j \ge 3$ are bounded by constant times h^3, we receive

$$\begin{aligned}\big\|e_n^{(2)}\big\|_X \le\ & h^2 \sum_{j=0}^{n-1} \left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{10}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} \big\|A_j' \widehat{u}_j\big\|_{X_\kappa} \\ &+ h^3 \sum_{j=0}^{n-1} \left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{20}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} \big\|A_j'' \widehat{u}_j\big\|_{X_\kappa} \\ &+ h^3 \sum_{j=0}^{n-1} \left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{11}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} \big\|A_j' \widehat{u}_j'\big\|_{X_\kappa} \\ &+ Ch^3.\end{aligned} \tag{5.8}$$

We note that the coefficients of the fourth-order scheme (1.2) satisfy the conditions

$$\begin{aligned}\Phi_{20}(0,0) &= \tfrac{1}{2}\Big\{\zeta\big(\tfrac{1}{3}\,\zeta^2 - \mu_2\big) + \widetilde{\zeta}\Big(\tfrac{1}{3}\,\widetilde{\zeta}^2 + \zeta\widetilde{\zeta} + \zeta^2 - \nu_2\Big)\Big\} = 0, \\ \Phi_{11}(0,0) &= \zeta^2\big(\tfrac{1}{3}\,\zeta - \tfrac{1}{2}\,\mu_1\big) + \widetilde{\zeta}\Big(\tfrac{1}{3}\,\widetilde{\zeta}^2 + \tfrac{1}{2}\,\widetilde{\zeta}(2\zeta - \nu_1) + \zeta(\zeta - \nu_1)\Big) = 0.\end{aligned} \tag{5.9a}$$

Therefore, similar arguments as in the proof of Theorem 1 show the refined bounds

$$\begin{aligned}\left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{20}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} &\le Mh(t_n - t_j)^{-1+\kappa}, \\ \left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{11}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} &\le Mh(t_n - t_j)^{-1+\kappa};\end{aligned}$$

see also (2.4) and (2.5c). In relation (5.8), it remains to estimate the sum involving Φ_{10}. For that purpose, we apply (2.5b) together with suitable Taylor expansions of B_j and C_j. Moreover, the coefficients of (1.2) fulfill

$$\begin{aligned}\Phi_{10}(0,0) &= \zeta\big(\tfrac{1}{2}\,\zeta - \mu_1\big) + \tfrac{1}{2}\,\widetilde{\zeta}^2 + \widetilde{\zeta}(\zeta - \nu_1) = 0, \\ \Psi_{10}(0,0) &= \zeta^2\big(\tfrac{1}{6}\,\zeta - \tfrac{1}{2}\,\mu_1\big) + \widetilde{\zeta}\Big(\tfrac{1}{6}\,\widetilde{\zeta}^2 + \tfrac{1}{2}\,\widetilde{\zeta}(\zeta - \nu_1) + \zeta\big(\tfrac{1}{2}\,\zeta - \mu_1\big)\Big) = 0.\end{aligned} \tag{5.9b}$$

As a consequence, we finally obtain the refined estimate

$$\left\|\prod_{i=j+1}^{n-1} \mathrm{e}^{\widetilde{\zeta} hC_i}\, \mathrm{e}^{\zeta hB_i}\, \Phi_{10}(hB_j, hC_j)\right\|_{X\leftarrow X_\kappa} \le Mh^{1+\kappa}\big(1 + |\log h| + (t_{n+1} - t_m)^{-1}\big).$$

Altogether, this implies

$$\begin{aligned}\big\|e_n^{(2)}\big\|_X &\le Ch^{3+\kappa}\sum_{j=0}^{n-1}\big(1+|\log h|+(t_n-t_j)^{-1}\big)\\ &\quad + Ch^4\sum_{j=0}^{n-1}(t_n-t_j)^{-1+\kappa}+Ch^3 \le Ch^{2+\kappa}\big(1+|\log h|\big),\end{aligned}$$

which proves the given error estimate. □

Remark 3. Going over the proof of Theorem 2 shows that the essential conditions for a fractional convergence order of $2+\kappa$ are (5.9). We note that the conditions for a classical convergence order 3 are equivalent to the relations in (5.9). However, it is not possible to construct a commutator-free exponential integrator of classical order 3 that is based on the evaluation of one exponential only, that is, the validity of relation (5.9) implies $0<\zeta<1$ in (3.1).

6. Numerical example. In this section, we illustrate the error estimate of Theorem 2 by a numerical example for a parabolic initial boundary value problem subject to a homogeneous Dirichlet boundary condition. We start with a brief discussion of the considered time integration schemes. For notational simplicity, we give only the first step and denote $A_i = A(c_i h)$.

Method 1. For parabolic problems (1.1), it follows from the error estimate given in our previous work [11] that the exponential midpoint rule

$$u_1 = \mathrm{e}^{hA_1}\,u_0, \qquad c_1 = \tfrac{1}{2},$$

is convergent of order 2 with respect to the norm of the underlying Banach space.

Method 2. The commutator-free exponential integration scheme

$$\begin{gathered}u_1 = \mathrm{e}^{(1-\zeta)h(a_1A_1+(1-a_1)A_2)}\,\mathrm{e}^{\zeta hA_1}\,u_0,\\ \zeta = \tfrac{\sqrt{3}}{3}, \qquad a_1 = \tfrac{1}{4}-\tfrac{\sqrt{3}}{4}, \qquad c_i = \tfrac{1}{2}\mp\tfrac{\sqrt{3}}{6}, \qquad i=1,2,\end{gathered}$$

has a classical convergence order 3.

Method 3. The numerical method

$$u_1 = \mathrm{e}^{h(a_2A_1+a_1A_2)}\,\mathrm{e}^{h(a_1A_1+a_2A_2)}\,u_0, \qquad a_i = \tfrac{1}{4}\pm\tfrac{\sqrt{3}}{6}, \qquad c_i = \tfrac{1}{2}\mp\tfrac{\sqrt{3}}{6}, \qquad i=1,2,$$

is the unique scheme of the form (3.1) that satisfies the conditions for a classical convergence order 4; see also (1.2).

In the numerical example, as an illustration, the fourth-order commutator-free exponential integrator given before is compared with a fourth-order interpolatory Magnus integrator. To explain the stability and error behavior of this method for parabolic problems is beyond the purpose of the present work.

Method 4. The fourth-order interpolatory Magnus integrator

$$u_1 = \mathrm{e}^{ha_1(A_1+A_2)+h^2a_2[A_2,A_1]}\,u_0, \qquad a_1 = \tfrac{1}{2}, \qquad a_2 = \tfrac{\sqrt{3}}{12},$$

requires the evaluation of the linear operator $\big[A_2,A_1\big] = A_2A_1 - A_1A_2$.

TABLE 6.1
Numerical temporal convergence orders in a discrete L^1-norm for spatial discretizations of grid length $\Delta x = (M+1)^{-1}$.

Stepsize h	1/2	1/4	1/8	1/16	1/32
Method 1 (M = 50)	2.0076	1.9632	1.9597	1.9699	1.9822
Method 1 (M = 100)	2.0075	1.9631	1.9595	1.9696	1.9818
Method 2 (M = 50)	1.0924	1.9634	2.2295	2.3162	2.4248
Method 2 (M = 100)	1.0949	1.9604	2.2267	2.3153	2.4181
Method 3 (M = 50)	2.2597	2.1983	2.3386	2.4337	2.4999
Method 3 (M = 100)	2.2591	2.1960	2.3348	2.4227	2.4782
Method 4 (M = 50)	3.3250	3.5115	3.3419	3.0490	2.8486
Method 4 (M = 100)	3.0426	3.4011	3.4838	3.2384	2.9488

TABLE 6.2
Numerical temporal convergence orders in a discrete L^2-norm for spatial discretizations of grid length $\Delta x = (M+1)^{-1}$.

Stepsize h	1/2	1/4	1/8	1/16	1/32
Method 1 (M = 50)	2.0120	1.9740	1.9723	1.9786	1.9879
Method 1 (M = 100)	2.0120	1.9739	1.9722	1.9785	1.9878
Method 2 (M = 50)	1.1979	1.9223	2.0992	2.1336	2.1732
Method 2 (M = 100)	1.1985	1.9208	2.0977	2.1303	2.1666
Method 3 (M = 50)	2.0197	2.0409	2.1271	2.1917	2.2331
Method 3 (M = 100)	2.0194	2.0397	2.1244	2.1859	2.2210
Method 4 (M = 50)	3.3204	3.5217	2.9654	2.4024	2.3609
Method 4 (M = 100)	3.0341	3.4204	3.3656	2.6010	2.3197

We consider a one-dimensional initial boundary value problem for a real-valued function $U : [0,1] \times [0,T] \to \mathbb{R} : (x,t) \mapsto U(x,t)$ comprising the partial differential equation

$$\partial_t U(x,t) = \mathscr{A}(x,t)\, U(x,t), \qquad 0 < x < 1, \quad 0 < t \le T, \tag{6.1a}$$

subject to a homogeneous Dirichlet boundary condition and an initial condition

$$U(0,t) = 0 = U(1,t), \quad 0 \le t \le T, \qquad U(x,0) = U_0(x), \quad 0 \le x \le 1. \tag{6.1b}$$

The differential equation involves a second-order differential operator

$$\mathscr{A}(x,t) = \alpha(x,t)\, \partial_x^2 + \beta(x,t)\, \partial_x + \gamma(x,t) \tag{6.1c}$$

which we assume to satisfy the condition of strong ellipticity. We further suppose that the space and time-dependent coefficients α, β, and γ fulfill suitable regularity and boundedness requirements. For $v \in \mathscr{C}_0^\infty(0,1)$ we define $u(t)$ and $A(t)$ through $\big(u(t)\big)(x) = U(x,t)$ and $\big(A(t)v\big)(x) = \mathscr{A}(x,t)v(x)$. Then, problem (6.1) can be cast into the abstract framework of section 2 for

$$X = L^p(0,1), \qquad D = W^{p,2}(0,1) \cap W_0^{p,1}(0,1), \qquad 1 < p < \infty;$$

see [11] and references therein. In view of the numerical experiment, we choose

$$\alpha(x,t) = \mathrm{e}^{x-t}, \qquad \beta(x,t) = xt, \qquad \gamma(x,t) = x^2\big(1 + \mathrm{e}^t\big).$$

The admissible values of κ in Theorem 2 are $0 \le \kappa < (2p)^{-1}$. Thus, the expected fractional convergence order in $X = L^p(0,1)$ is $2 + \kappa$, where $\kappa < (2p)^{-1}$.

TABLE 6.3
Numerical temporal convergence orders in a discrete L^∞-norm for spatial discretizations of grid length $\Delta x = (M+1)^{-1}$.

Stepsize h	1/2	1/4	1/8	1/16	1/32
Method 1 (M = 50)	2.0250	2.0065	2.0208	2.0226	2.0149
Method 1 (M = 100)	2.0250	2.0063	2.0207	2.0222	2.0129
Method 2 (M = 50)	1.2328	1.7318	1.8169	1.8604	1.9092
Method 2 (M = 100)	1.2341	1.7313	1.8135	1.8559	1.9072
Method 3 (M = 50)	1.7384	1.8369	1.9113	1.9649	1.9851
Method 3 (M = 100)	1.7391	1.8347	1.9103	1.9604	1.9736
Method 4 (M = 50)	3.3042	3.0169	1.9200	2.0864	2.1880
Method 4 (M = 100)	3.0257	3.4434	2.0132	1.9839	2.0752

In the numerical experiment, we discretize the problem in space by symmetric finite differences of grid length $\Delta x = (M+1)^{-1}$. In time, we apply the exponential integrators given above with stepsize $h = 2^{-i}$ for $1 \le i \le 5$ and integrate the problem up to time $T = 1$. A reference solution is determined for a temporal stepsize $h = 2^{-10}$. The numerical temporal order of convergence with respect to a discrete L^p-norm is determined in a standard way from the numerical solution values. The obtained numbers for $p = 2$ and the limiting cases $p = 1$ and $p = \infty$ are displayed in Tables 6.1, 6.2, and 6.3. The convergence order 2 for the exponential midpoint rule (Method 1) is explained by a convergence result proved in [11]. For the commutator-free exponential integrators of classical order 3 (Method 2) and classical order 4 (Method 3), respectively, the values of approximately $2 + (2p)^{-1}$ are in accordance with the convergence orders predicted by Theorem 2.

7. Conclusions. In the present work, we studied the convergence properties of a commutator-free exponential integrator that relies on the composition of two exponentials for parabolic initial value problems of the form (1.1). In particular, we focused on the fourth-order scheme (1.2), which is based on the Gaussian nodes. We showed that the exponential integration scheme remains stable for arbitrarily large stepsizes. But, it is seen from the theoretical investigations and as well in a numerical experiment that a substantial order reduction occurs, in general. For instance, for one-dimensional parabolic initial-boundary value problems under a homogeneous Dirichlet boundary condition a fractional convergence order of at most $2 + (2p)^{-1}$ can be expected in the norm of the function space L^p. The order reduction is explained by the fact that even if the exact solution of the initial boundary value problem belongs to the domain of the differential operator and further is temporally smooth, it in general does not fulfill additional boundary conditions, that is, combinations of the form $A(s)A(t)u(t)$ are not well defined for all $0 \le s, t \le T$.

For that reason, concerning the derivation of high-order exponential integrators for nonautonomous parabolic problems, it seems more promising to employ a suitable linearization and to base the numerical schemes on explicit exponential methods of Runge–Kutta or multistep type. Also, the error analysis for nonautonomous parabolic equations is of theoretical value as it gives insight into how to construct and study numerical methods for quasi-linear equations which are of particular interest in view of practical applications. For example, quasi-linear parabolic problems are used in the modeling of diffusion processes with state-dependent diffusitivity and arise in the study of fluids in porous media, see [12]. We intend to investigate this approach in a future work.

Acknowledgments. I am grateful to Jitse Niesen. His talk and interest at MaGIC 2005 motivated me to finish this work. I thank Alexander Ostermann for inspiring discussions on exponential integrators.

REFERENCES

[1] Ch. Besse, B. Bidegaray, and S. Descombes, *Order estimates in time of splitting methods for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 40 (2002), pp. 26–40.
[2] S. Blanes, F. Casas, J.A. Oteo, and J. Ros, *Magnus and Fer expansions for matrix differential equations: The convergence problem*, J. Phys. A, 31 (1998), pp. 259–268.
[3] S. Blanes and P.C. Moan, *Splitting methods for the time-dependent Schrödinger equation*, Phys. Lett. A, 265 (2000), pp. 35–42.
[4] H. Brunnner and P.J. van der Houwen, *The Numerical Solution of Volterra Equations*, CWI Monogr. 3, North–Holland, Amsterdam, 1986.
[5] M.P. Calvo and C. Palencia, *A class of explicit multistep exponential integrators for semilinear problems*, Numer. Math., 102 (2006), pp. 367–381.
[6] E. Celledoni, *Eulerian and Semi-Lagrangian Schemes Based on Commutator Free Exponential Integrators*, CRM Proc. Lecture Notes, ISSN 1065-8580, 39 (2005).
[7] E. Celledoni, A. Martinsen, and B. Owren, *Commutator-free Lie group methods*, Future Generation Computer Systems (FGCS), 19 (2003), pp. 341–352.
[8] S.M. Cox and P.C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.
[9] P.E. Crouch and R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci., 3 (1993), pp. 1–33.
[10] J. Van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 27 (2006), pp. 1438–1457.
[11] C. González, A. Ostermann, and M. Thalhammer, *A second-order Magnus integrator for non-autonomous parabolic problems*, J. Comput. Appl. Math., 189 (2006), pp. 142–156.
[12] C. González and M. Thalhammer, *A second-order Magnus type integrator for quasilinear parabolic problems*, Math. Comp., in press.
[13] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin, 1981.
[14] M. Hochbruck and Ch. Lubich, *On Magnus integrators for time-dependent Schrödinger equations*, SIAM J. Numer. Anal., 41 (2003), pp. 945–963.
[15] M. Hochbruck and Ch. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
[16] M. Hochbruck and A. Ostermann, *Explicit exponential Runge-Kutta methods for semilinear parabolic problems*, SIAM J. Numer. Anal., 43 (2005), pp. 1069–1090.
[17] A. Iserles, *On the global error of discretization methods for highly-oscillatory ordinary differential equations*, BIT, 42 (2002), pp. 561–599.
[18] T. Jahnke and Ch. Lubich, *Error bounds for exponential operator splittings*, BIT, 40 (2000), pp. 735–744.
[19] A.K. Kassam and L.N. Trefethen, *Fourth-order time stepping for stiff PDEs*, SIAM J. Sci. Comput., 26 (2005), pp. 1214–1233.
[20] S. Krogstad, *Generalized integrating factor methods for stiff PDEs*, J. Comput. Phys., 203 (2005), pp. 72–88.
[21] Y.Y. Lu, *A fourth order Magnus scheme for Helmholtz equation*, J. Comput. Appl. Math., 173 (2005), pp. 247–258.
[22] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, Switzerland, 1995.
[23] W. Magnus, *On the exponential solution of a differential equation for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.
[24] B.V. Minchev and W.M. Wright, *A Review of Exponential Integrators for First Order Semilinear Problems*, Tech. report 2/05, Department of Mathematics, Norwegian University of Science and Technolgy (NTNU), April 2005.
[25] C. Moler and Ch. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.
[26] A. Ostermann and M. Thalhammer, *Non-smooth data error estimates for linearly implicit Runge-Kutta methods*, IMA J. Numer. Anal., 20 (2000), pp. 167–184.
[27] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[28] Q. Sheng, *Global error estimates for exponential splitting*, IMA J. Numer. Anal., 14 (1993), pp. 27–56.
[29] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.
[30] M. Thalhammer, *A second-order Magnus type integrator for non-autonomous semilinear parabolic problems*, Universität Innsbruck, 2005, preprint.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 865–878

ROBUST EIGENVALUE COMPUTATION FOR SMOOTHING OPERATORS*

RADU ALEXANDRU TODOR†

Abstract. Robust quasi-relative Galerkin discretization error estimates are derived for the eigenvalue problem associated to a nonnegative compact operator $\mathcal{K}$ acting in a Hilbert space. Trace discretization error estimates for arbitrarily small positive powers of $\mathcal{K}$ are obtained as a consequence. Coupled with bounds on eigenfunction oscillations, the results are then applied to the case of an integral operator with (piecewise) smooth kernel K on a bounded domain and in the context of the h finite element method.

Key words. eigenvalue computation, Galerkin FEM, integral operator

AMS subject classifications. 45C05, 65N30, 47G10

DOI. 10.1137/040616449

1. Introduction.

We consider a bounded domain $D \subset \mathbb{R}^d$ and a symmetric kernel $K \in L^2(D \times D)$ defining a symmetric, nonnegative compact integral operator

$$\mathcal{K} : L^2(D) \to L^2(D), \qquad (\mathcal{K}u)(x) = \int_D K(x,x')u(x')\,dx' \quad \lambda\text{-a.e. } x \in D, \tag{1.1}$$

and for all $u \in L^2(D)$, where λ denotes here the Lebesgue measure. Such operators arise frequently in statistics and the theory of random fields as covariance operators (a typical example is given by the Gaussian kernel $K(x,x') = \exp(-|x-x'|^2)$; see [11] for further examples) and the computation of their spectral decomposition (eigenelements) is relevant for many applications, of which we mention here only the random field representation via the Karhunen–Loève expansion (see, e.g., [8]). This in turn has an important impact on the accuracy of all practical algorithms based on the Karhunen–Loève expansion of a random field, like, e.g., solving PDEs with stochastic data via polynomial chaos and stochastic Galerkin method. See [4] and the references therein for details and further examples.

1.1. Problem formulation.

The eigenvalue computation for the integral operator $\mathcal{K}$ given by (1.1) using the Galerkin method applied with a finite element space family $\mathcal{S} := (S_\hbar)_{\hbar\in\mathfrak{H}} \subset L^2(D)$ ($\hbar \in \mathfrak{H}$ stands here and in the following for the discretization parameter) consists in solving the discrete variational problem of finding $(\lambda_{\hbar,m}, \phi_{\hbar,m})_{m\in\mathbb{N}_+} \subset \mathbb{R} \times S_\hbar$ such that $\|\phi_{\hbar,m}\|_{L^2(D)} = 1\ \forall m \in \mathbb{N}_+$ and

$$\int_D \int_D K(x,x')\phi_{\hbar,m}(x')\psi(x)\,dx'dx = \lambda_{\hbar,m} \int_D \phi_{\hbar,m}(x)\psi(x)\,dx \quad \forall \psi \in S_\hbar. \tag{1.2}$$

Equation (1.2) shows that $(\lambda_{\hbar,m}, \phi_{\hbar,m})_{m\in\mathbb{N}_+}$ is nothing but the eigenvalue sequence of the nonnegative compact operator $\mathcal{K}_\hbar := P_\hbar \mathcal{K} P_\hbar$ in $L^2(D)$, where $P_\hbar$ denotes the $L^2(D)$ orthogonal projection onto $S_\hbar$.

*Received by the editors October 6, 2004; accepted for publication (in revised form) October 31, 2005; published electronically April 21, 2006. This work was supported in part under the IHP network Breaking Complexity of the EC, contract HPRN-CT-2002-00286, with support by the Swiss Federal Office for Science and Education under grant BBW 02.0418.
http://www.siam.org/journals/sinum/44-2/61644.html
†Seminar for Applied Mathematics, ETH Zürich, Switzerland (todor@math.ethz.ch).

The discretization error analysis of the eigenvalue problem (1.2) in the presence of a Galerkin scheme follows in general from abstract results on compact operators in Hilbert spaces. (See, e.g., [7], [3], [2], [5], or [9] for similar results in Banach spaces.) Denoting by $(\lambda_m, \phi_m)_{m\in\mathbb{N}_+}/(\lambda_{\hbar,m}, \phi_{\hbar,m})_{m\in\mathbb{N}_+}$ the exact/discrete eigenelements of $\mathcal{K}$, the uniform error estimates obtained in an abstract setting in [12], [7] apply

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq \|(I-P_\hbar)\mathcal{K}(I-P_\hbar)\|_{\mathcal{B}(L^2(D))} \quad \forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}. \tag{1.3}$$

On the other hand, due to the positivity of $\mathcal{K}$, it trivially holds

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq \lambda_m \quad \forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}. \tag{1.4}$$

While (1.3) describes the exact error asymptotics w.r.t. $\hbar$, (1.4) retains the correct scaling with λ_m. In fact, (1.3) and (1.4) are to some extent (but not fully) generalized by the following exact error representation formula (see [12], [3])

$$0 \leq \lambda_m - \lambda_{\hbar,m} = c_{\hbar,m}\|(I-P_\hbar)E^{\mathcal{K}}_{\{\lambda_m\}}\phi_{\hbar,m}\|^2_{L^2(D)}\lambda_m \quad \forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}, \tag{1.5}$$

where $E^{\mathcal{K}}_A$ denotes the spectral projector of $\mathcal{K}$ onto the Borel set $A \subset \mathbb{R}$ and

$$\forall m \in \mathbb{N}_+ \quad c_{\hbar,m} \to 1 \quad \text{ as } S_\hbar \nearrow L^2(D),$$

that is, as the finite element space $S_\hbar$ gets refined.

However, (1.5) has a rather strong *asymptotic* character: for a fixed $m \in \mathbb{N}_+$, (1.5) becomes sharper than (1.4) and (1.3) only for fine enough finite element spaces depending on m ($\hbar \in \mathfrak{H}_m \subset \mathfrak{H}$) and in fact on the size of the spectrum gap around λ_m. The *estimated* constant $c_{\hbar,m}$ scales unfavorably with m (for rough finite element spaces), as the inverse of the spectrum gap size around λ_m. Consequently, the smaller the gap, the larger the estimated preasymptotic domain $\mathfrak{H}_m$ (that is, the set of those $\hbar \in \mathfrak{H}$ where (1.3), (1.4) are sharper than (1.5)).

1.2. Main results. The purpose of this work is to provide *robust* eigenvalue convergence rates of the type (1.5) for the particular case under consideration, that of an integral operator with smooth kernel, if standard finite element spaces corresponding to the h version of FEM are employed. The main result reads as follows.

THEOREM 1.1. *Let $\mathcal{K}$ be the integral operator* (1.1) *with (piecewise, in the sense of Definition* 3.1*) smooth kernel K and let $\mathcal{T} = (T_h)_{h>0}$ be a regular triangulation of D with meshwidth h. Setting $\hbar := h$ and defining, for $p \in \mathbb{N}_+$, $S_\hbar$ to be the space of discontinuous piecewise polynomials of degree $p-1$ on T_h, we have that for any $s > 0$ there exists $c_{K,\mathcal{T},p,s} > 0$ such that*

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq c_{K,\mathcal{T},p,s}(h^{2p}\lambda_m^{1-s} + h^{4p}\lambda_m^{-2s}) \quad \forall m \in \mathbb{N}_+, \forall h > 0 \tag{1.6}$$

and

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq c_{K,\mathcal{T},p,s}h^{2p}\lambda_m^{1/2-s} \quad \forall m \in \mathbb{N}_+, \forall h > 0. \tag{1.7}$$

Note that (1.7) follows by interpolation with logarithmic weight $1/2$ between (1.4) and the second term in the upper bound (1.6). Comparing (1.7) to (1.5), we see that (1.7) retains the correct asymptotic behaviour w.r.t. $\hbar$ and trades a factor $\lambda_m^{1/2}$ for robustness in $m \in \mathbb{N}_+$.

The proof of Theorem 1.1 is based on the following abstract error estimate we prove for a symmetric nonnegative compact operator $\mathcal{K}$ acting in a separable Hilbert

space $(H, \langle\cdot,\cdot\rangle_H)$. Concerning notations, $\mathcal{B}(H)$ denotes throughout this work the space of bounded linear operators acting in the separable Hilbert space $(H, \langle,\rangle_H)$, and we occasionally use traditional subscripts (e.g., $\mathcal{B}_\infty$, $\mathcal{B}_p$, $\mathcal{B}_{\mathrm{sym}}$) for spaces of operators with additional properties (compactness, compactness and ℓ_p summability of the eigenvalue sequence for $p > 0$, symmetry). Further, generic constants are denoted by c and all the variables on which they depend are included as subscripts.

PROPOSITION 1.2. *Let $\mathcal{K} \in \mathcal{B}(H)$ be a nonnegative compact operator with eigenelements $(\lambda_m, \phi_m)_{m\in\mathbb{N}_+}$ and let $\mathcal{S} := (S_\hbar)_{\hbar\in\mathfrak{H}} \subset H$ be a finite element space family. Then the following estimate holds:*

$$0 \le \lambda_m - \lambda_{\hbar,m} \le c_m \Psi(m,\hbar)^2 \lambda_m + c_{\mathcal{K},m} \Psi(m,\hbar)^4 \quad \forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}, \tag{1.8}$$

where

$$\begin{aligned} \Phi(m,\hbar) &:= \|(I - P_\hbar)\phi_m\|_H, \\ \Psi(m,\hbar) &:= \max_{1\le j\le m} \Phi(j,\hbar), \end{aligned} \tag{1.9}$$

and

$$c_m := m^{3/2} + m, \quad c_{\mathcal{K},m} := m \sum_{j=1}^{m} \lambda_j.$$

Here $(\lambda_{\hbar,m})_{m\in\mathbb{N}_+}$ are the eigenvalues of $\mathcal{K}_\hbar := P_\hbar \mathcal{K} P_\hbar$.

Remark 1.3. In general (i.e., for standard finite element spaces), one expects the maximum in (1.9) to be attained for $j = m$, that is, $\Psi(m,\hbar) = \Phi(m,\hbar)$ for all $m \in \mathbb{N}_+$. This is essentially due to the increasing difficulty of approximating high frequency eigenfunctions, as $m \to \infty$ (or, equivalently, to the larger size of the estimated preasymptotic range $\mathfrak{H}_m$ as compared to $\mathfrak{H}_j$, for $j < m$).

Remark 1.4. Note that the first term on the right-hand side of (1.8) corresponds for any fixed $m \in \mathbb{N}_+$ to the correct asymptotics of (1.5) w.r.t. $\hbar$, whereas the second (corrector) gives a size estimate of the preasymptotic range,

$$\mathfrak{H}_m \subseteq \{\hbar \in \mathfrak{H} \mid \Psi(m,\hbar)^2 \ge \lambda_m / c_{\mathcal{K},m}\}.$$

The application of Proposition 1.2 to the case of the integral operator (1.1) and in the context of standard h FEM is then discussed in the second part of this work. Using the Gagliardo–Nirenberg inequalities (see, e.g., [1]) to bound eigenfunction oscillations and the standard h FEM error estimate, we prove the following upper bound on the functional Ψ.

THEOREM 1.5. *Let $\mathcal{K}$ be the integral operator* (1.1) *with (piecewise, in the sense of Definition* 3.1*) smooth kernel K and let $\mathcal{T} = (T_h)_{h>0}$ be a regular triangulation of D with meshwidth h. Setting $\hbar := h$ and defining $S_\hbar$ to be the space of discontinuous piecewise polynomials of degree $p - 1 \in \mathbb{N}$ on T_h, we have that for any $s > 0$ there exists $c_{K,\mathcal{T},p,s} > 0$ such that*

$$\Phi(m,\hbar) \le \Psi(m,\hbar) \le c_{K,\mathcal{T},p,s} \lambda_m^{-s} h^p \quad \forall m \in \mathbb{N}_+, \forall h > 0. \tag{1.10}$$

Further applications of Proposition 1.2 to the case of an operator $\mathcal{K}$ with a weaker/stronger smoothing effect (corresponding to finite Sobolev/analytic regularity of K), in the context of the h/p FEM, respectively, as well as similar estimates for eigenspaces (known to be more sensitive to perturbations than the eigenvalues) will be addressed in a forthcoming paper.

The article proceeds as follows. We devote section 2 to a proof of Proposition 1.2 and to its consequences (e.g., robust eigenvalue convergence rates for arbitrarily small positive powers of $\mathcal{K}$). We then discuss, as an application, the case of an integral operator in section 3. Proposition 1.2 takes a particularly simple form in this case, due to explicit control of eigenvalue decay and eigenfunction oscillations. Eigenvalue decay rates (in terms of the kernel regularity) are standard and briefly reviewed in section 3.1. Eigenfunction oscillations for smooth kernels are investigated in section 3.2 and shown (see Theorem 1.5) to be significantly milder than the eigenvalue decay rate.

2. Robust eigenvalue error estimates. Here we prove the main abstract discretization error estimate (1.8) and then discuss its optimality plus some direct consequences.

2.1. Abstract estimate. The main tool for the proof of Proposition 1.2, as formulated in the introduction, will be the minimax principle.

Proof of Proposition 1.2. First we note that the lower bound 0 for $\lambda_m - \lambda_{\hbar,m}$ follows trivially from the minimax principle. To check the upper bound, we fix $m \in \mathbb{N}_+$ and $\hbar \in \mathfrak{H}$ and write, using again the minimax principle,

$$\begin{aligned} \lambda_{\hbar,m} &= \max_{\substack{U\subset H\\ \dim U\geq m}} \min_{\substack{\phi\in U\\ \|\phi\|_H=1}} \langle \mathcal{K}_\hbar\phi,\phi\rangle_H \\ &= \max_{\substack{U\subset H\\ \dim U\geq m}} \min_{\substack{\phi\in U\\ \|\phi\|_H=1}} \left\{\langle \mathcal{K}\phi,\phi\rangle_H + \langle(\mathcal{K}_\hbar-\mathcal{K})\phi,\phi\rangle_H\right\}. \end{aligned} \tag{2.1}$$

The identity

$$P_\hbar\mathcal{K}P_\hbar - \mathcal{K} = -(I-P_\hbar)\mathcal{K} - \mathcal{K}(I-P_\hbar) + (I-P_\hbar)\mathcal{K}(I-P_\hbar)$$

and the fact that $\mathcal{K}$ is nonnegative ensure then

$$\langle(\mathcal{K}_\hbar-\mathcal{K})\phi,\phi\rangle_H \geq -2|\langle\mathcal{K}\phi,(I-P_\hbar)\phi\rangle_H| \quad \forall\phi\in H. \tag{2.2}$$

Using (2.2) in (2.1) we obtain

$$\begin{aligned} \lambda_{\hbar,m} &\geq \max_{\substack{U\subset H\\ \dim U\geq m}} \min_{\substack{\phi\in U\\ \|\phi\|_H=1}} \left\{\langle \mathcal{K}\phi,\phi\rangle_H - 2|\langle\mathcal{K}\phi,(I-P_\hbar)\phi\rangle_H|\right\} \\ &\geq \max_{\substack{U\subset H\\ \dim U\geq m}} \min_{\substack{\phi\in U\\ \|\phi\|_H=1}} \left\{\langle \mathcal{K}\phi,\phi\rangle_H - 2\|(I-P_\hbar)\mathcal{K}\phi\|_H\|(I-P_\hbar)\phi\|_H\right\}. \end{aligned} \tag{2.3}$$

At this stage we choose U to be the subspace of H spanned by the first m eigenfunctions $\phi_1,\phi_2,\ldots,\phi_m$ of $\mathcal{K}$. Expanding $\phi = \sum_{j=1}^m \alpha_j\phi_j$ with $\alpha_j\in\mathbb{C}$ and using definition (1.9) we obtain

$$\|(I-P_\hbar)\mathcal{K}\phi\|_H \leq \sum_{j=1}^m |\alpha_j|\lambda_j\Phi(j,\hbar) \leq \Psi(m,\hbar)\sum_{j=1}^m|\alpha_j|\lambda_j. \tag{2.4}$$

On the other hand,

$$\|(I-P_\hbar)\phi\|_H \leq \sum_{j=1}^m|\alpha_j|\Phi(j,\hbar) \leq \left(\sum_{j=1}^m\Phi(j,\hbar)^2\right)^{1/2} \leq \sqrt{m}\Psi(m,\hbar). \tag{2.5}$$

From (2.3), (2.4), (2.5) we obtain

$$\lambda_{\hbar,m} \geq \min_{\sum_{j=1}^m |\alpha_j|^2=1} \left\{ \sum_{j=1}^m \lambda_j |\alpha_j|^2 - 2\sqrt{m}\Psi(m,\hbar)^2 \sum_{j=1}^m \lambda_j |\alpha_j| \right\}. \tag{2.6}$$

Setting $\varepsilon := \sqrt{m}\Psi(m,\hbar)^2$, we distinguish two cases, as follows.

If $\varepsilon \geq 1/\sqrt{m}$, then it holds

$$\lambda_{\hbar,m} \geq 0 \geq \lambda_m - m\Psi(m,\hbar)^2\lambda_m,$$

which is in fact a stronger estimate than (1.8).

Otherwise $\varepsilon < 1/\sqrt{m}$ and we apply Lemma 2.1 below to obtain from (2.6)

$$\lambda_{\hbar,m} \geq \lambda_m - (m^{3/2}+m)\Psi(m,\hbar)^2\lambda_m - c_{\mathcal{K},m}\Psi(m,\hbar)^4$$

with $c_{\mathcal{K},m} = m\sum_{j=1}^m \lambda_j$, which concludes the proof. ∎

LEMMA 2.1. *If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ is a nonincreasing sequence of nonnegative real numbers and $\varepsilon \in [0, 1/\sqrt{m}[$, then*

$$\min_{\sum_{j=1}^m t_j^2=1} \sum_{j=1}^m \lambda_j(t_j^2 - 2\varepsilon t_j) \geq (1-\varepsilon(m+\sqrt{m}))\lambda_m - \varepsilon^2 \sum_{j=1}^m \lambda_j. \tag{2.7}$$

Proof. Obviously, the minimum on the left-hand side (l.h.s.) of (2.7) is attained at a location $\tilde{t} := (\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m)$ on the unit m dimensional sphere, with nonnegative coordinates. Using Lagrange multipliers for $f, g : \mathbb{R}^m \to \mathbb{R}$ given by

$$f(t) := \sum_{j=1}^m \lambda_j(t_j^2 - 2\varepsilon t_j) \quad \text{and} \quad g(t) := \sum_{j=1}^m t_j^2 - 1 \quad \forall t = (t_1, t_2, \ldots, t_m) \in \mathbb{R}^m,$$

we obtain the existence of a real $\lambda \neq 0$ such that for the location $\tilde{t}$ of the minimum of f restricted to $g^{-1}(\{0\})$ it holds $\nabla f - \lambda \nabla g = 0$, that is,

$$\tilde{t}_j = \frac{\varepsilon\lambda_j}{\lambda_j - \lambda} \quad \forall 1 \leq j \leq m. \tag{2.8}$$

Imposing $g(\tilde{t}) = 0$ we obtain that λ solves the equation

$$\sum_{j=1}^m \frac{\lambda_j^2}{(\lambda_j - \lambda)^2} = \frac{1}{\varepsilon^2}. \tag{2.9}$$

In order to estimate λ, we first remark that $\tilde{t}_j \geq 0$ $\forall 1 \leq j \leq m$ implies that λ is the unique solution of (2.9) situated in the interval $]-\infty, \lambda_m[$. Further, the condition $0 \leq \varepsilon < 1/\sqrt{m}$ ensures the positivity of λ, so that $\lambda \in]0, \lambda_m[$. As a consequence, the mth term of the sum in (2.9) is the largest one, which then implies

$$\frac{\lambda_m^2}{(\lambda_m - \lambda)^2} \leq \frac{1}{\varepsilon^2} \leq \frac{m\lambda_m^2}{(\lambda_m - \lambda)^2}$$

or, equivalently,

$$\lambda_m - \sqrt{m}\varepsilon\lambda_m \leq \lambda \leq \lambda_m - \varepsilon\lambda_m. \tag{2.10}$$

Computing f at $\tilde{t}$ given by (2.8) then yields

$$\begin{aligned}\min_{\substack{t\in\mathbb{R}^m\\ g(t)=0}} f(t) = f(\tilde{t}) &= \lambda - \varepsilon^2 \sum_{1\le j\le m} \frac{\lambda_j^2}{\lambda_j-\lambda}\\ &= \lambda - \varepsilon^2\left(m\lambda + \sum_{j=1}^m \lambda_j + \lambda^2 \sum_{j=1}^m \frac{1}{\lambda_j-\lambda}\right)\\ &\overset{(2.10)}{\ge} (1-\varepsilon^2 m)\lambda - \varepsilon^2\sum_{j=1}^m \lambda_j - \varepsilon m\lambda^2/\lambda_m\\ &\overset{(2.10)}{\ge} (1-\varepsilon m)\lambda - \varepsilon^2\sum_{j=1}^m \lambda_j \overset{(2.10)}{\ge} (1-\varepsilon(m+\sqrt{m}))\lambda_m - \varepsilon^2\sum_{j=1}^m \lambda_j,\end{aligned}$$

which concludes the proof. □

Some comments on the optimality of the estimate (1.8) in Proposition 1.2 are now in order.

Remark 2.2. If $\lambda_1 = \lambda_2 = \cdots = \lambda_m$, then the minimum in (2.7) can be explicitly computed and equals $(1-2\sqrt{m}\varepsilon)\lambda_m$. The lower bound in (2.7) is therefore in this case suboptimal, due to the $O(\varepsilon^2)$ term. However, this is a rather special case, since in general the eigenvalue sequence is not constant, unless $\mathcal{K} = 0$.

We argue in the following that although sharper estimates for $c_m, c_{\mathcal{K},m}$ in (1.8) can be obtained, the upper bound (1.8) obtained through (2.3) is qualitatively optimal, (and serves our purpose of proving Theorem 1.1), in the sense that the second term on the right-hand side (r.h.s.) of (1.8) does not scale with (any positive power of) λ_m. A careful analysis of the proof presented in this section reveals that its main weakness lies in the use of the Cauchy–Schwarz inequality twice, to obtain (2.3) and (2.5).

Remark 2.3. Improving on (2.3) seems to be the key point in obtaining a qualitatively better eigenvalue error estimate, but this requires further favorable properties (e.g., diagonal dominance) of the matrix

$$(\langle\phi_j, (I-P_\hbar)\phi_{j'}\rangle_H)_{1\le j,j'\le m}.$$

Remark 2.4. Improving on (2.5) by using, e.g., $\|(I-P_\hbar)\phi\|_H \le \Psi(m,\hbar)\sum_{j=1}^m |\alpha_j|$ leads to a sharper eigenvalue error estimate via the (more difficult) minimization problem

$$\lambda_{\hbar,m} \ge \lambda := \min_{\sum_{j=1}^m |\alpha_j|^2=1} \left\{\sum_{j=1}^m \lambda_j|\alpha_j|^2 - 2\Psi(m,\hbar)^2 \sum_{j=1}^m \lambda_j|\alpha_j| \sum_{j=1}^m |\alpha_j|\right\}. \tag{2.11}$$

Choosing, for $m \ge 2$, $\alpha_2 = \alpha_3 = \cdots = \alpha_{m-1} = 0$, we have

$$\lambda \le \tilde{\lambda} := \min_{t_1^2+t_m^2=1} \left\{\lambda_1(1-2\varepsilon)t_1^2 + \lambda_m(1-2\varepsilon)t_m^2 - 2\varepsilon(\lambda_1+\lambda_m)t_1t_m\right\} \tag{2.12}$$

with $\varepsilon := \Psi(m,\hbar)^2$.

From Lemma 2.5 below it follows that if $\lambda_m < \lambda_1$ and $0 \le \varepsilon \le \min\{1/2, \lambda_1\} \le \sqrt{2}(\lambda_1-\lambda_m)$ (which both hold for m large enough depending on $\mathcal{K}$), then $\tilde{\lambda}$ satisfies, with some $c \in [1/2, 1]$,

$$\tilde{\lambda} = (1-2\varepsilon)\lambda_m - c\varepsilon^2\frac{(\lambda_1+\lambda_m)^2}{(1-2\varepsilon)(\lambda_1-\lambda_m)}.$$

We conclude that if $(\lambda_j)_{1\le j\le m}$ is not a constant sequence (and this is always the case for the eigenvalue sequence of a compact operator if m is large enough), then the minimization problem (2.11) leads to an estimate qualitatively similar to (1.8). (The term containing $\Psi(m,\hbar)^4$ does not scale with λ_m but with a constant bounded away from 0.)

LEMMA 2.5. *If $0<\beta<\alpha$ and $0\le\gamma\le\sqrt{2}(\alpha-\beta)$, then*

$$\min_{x^2+y^2=1}\{\alpha x^2+\beta y^2-2\gamma xy\}=\beta-\frac{2\gamma^2}{\alpha-\beta+\sqrt{(\alpha-\beta)^2+4\gamma^2}} \tag{2.13}$$

$$\in\left[\beta-\frac{\gamma^2}{\alpha-\beta},\beta-\frac{1}{2}\frac{\gamma^2}{\alpha-\beta}\right]. \tag{2.14}$$

Proof. Note first that (2.14) follows from (2.13) using the assumption on γ. It remains to check (2.13). To this end, we note first that for the location $(\tilde{x},\tilde{y})$ of the minimum in (2.13) and the Lagrange multiplier $\lambda\in\mathbb{R}$ it holds

$$\begin{cases}(\alpha-\lambda)\tilde{x} &= \gamma\tilde{y},\\ (\beta-\lambda)\tilde{y} &= \gamma\tilde{x},\end{cases} \tag{2.15}$$

from which we obtain (assuming also without loss of generality $\tilde{x},\tilde{y}>0$) that $\lambda\in\,]-\infty,\beta]$ solves $(\alpha-\lambda)(\beta-\lambda)=\gamma^2$, that is, λ equals the r.h.s. of (2.13). The system (2.15) has then the solution

$$\tilde{x}=\left(\frac{\beta-\lambda}{\alpha+\beta-2\lambda}\right)^{1/2},\quad \tilde{y}=\left(\frac{\alpha-\lambda}{\alpha+\beta-2\lambda}\right)^{1/2}.$$

Upon inserting these values into the quadratic form (2.13) we obtain that the minimum we look for equals λ, which concludes the proof. ∎

Proposition 1.2 takes a particularly simple form if the operator $\mathcal{K}$ satisfies the following conditions.

ASSUMPTION 2.6. *$\mathcal{K}\in\cap_{p>0}\mathcal{B}_p(H)$ is nonnegative and for any $s>0$ there exists $c_{\mathcal{K},\mathcal{S},s}>0$ such that*

$$\Phi(m,\hbar)=\|(I-P_\hbar)\phi_m\|_H\le c_{\mathcal{K},\mathcal{S},s}\lambda_m^{-s}\Upsilon(\hbar)\quad\forall m\in\mathbb{N}_+,\forall\hbar\in\mathfrak{H}, \tag{2.16}$$

where the functional $\Upsilon:\mathfrak{H}\to\mathbb{R}$ describes the approximation property *of the finite element space family $\mathcal{S}$.*

Keeping in mind the standard h FEM error estimate, we note that (2.16) can be viewed as a combination of *good $\mathcal{K}$-eigenfunction approximability* through the finite elements in $\mathcal{S}$ and a *mild eigenfunction oscillation* condition. We have the next theorem.

THEOREM 2.7. *If Assumption* 2.6 *is satisfied, then for any $s>0$ there exists a constant $c_{\mathcal{K},\mathcal{S},s}>0$ such that*

$$0\le\lambda_m-\lambda_{\hbar,m}\le c_{\mathcal{K},\mathcal{S},s}(\Upsilon(\hbar)^2\lambda_m^{1-s}+\Upsilon(\hbar)^4\lambda_m^{-2s})\quad\forall m\in\mathbb{N}_+,\forall\hbar\in\mathfrak{H}, \tag{2.17}$$

where $(\lambda_m)_{m\in\mathbb{N}_+}$ and $(\lambda_{\hbar,m})_{m\in\mathbb{N}_+}$ are the eigenvalue sequences of $\mathcal{K}$ and $P_\hbar\mathcal{K}P_\hbar$, respectively.

Proof. Condition (2.16) ensures that

$$\Psi(m,\hbar)\le c_{\mathcal{K},\mathcal{S},s}\Upsilon(\hbar)\lambda_m^{-s}\quad\forall m\in\mathbb{N}_+,\forall\hbar\in\mathfrak{H},\forall s>0,$$

whereas $\mathcal{K} \in \cap_{p>0}\mathcal{B}_p(H)$ gives

$$c_m = m^{3/2} + m \leq c_{\mathcal{K},s}\lambda_m^{-s}, \quad c_{\mathcal{K},m} = m\sum_{j=1}^{m}\lambda_j \leq c_{\mathcal{K},s}\lambda_m^{-s} \quad \forall m \in \mathbb{N}_+, \forall s > 0,$$

which concludes the proof via (1.8). ∎

COROLLARY 2.8. *If Assumption* 2.6 *is satisfied, then for any* $s > 0$ *there exists a constant* $c_{\mathcal{K},\mathcal{S},s} > 0$ *such that*

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq c_{\mathcal{K},\mathcal{S},s}(\Upsilon(\hbar)^2\lambda_m^{1-s} + \Upsilon(\hbar)^{4\alpha}\lambda_m^{1-\alpha-2s\alpha}) \tag{2.18}$$

$\forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}, \forall \alpha \in [0,1]$, *where* $(\lambda_m)_{m\in\mathbb{N}_+}$ *and* $(\lambda_{\hbar,m})_{m\in\mathbb{N}_+}$ *are the eigenvalue sequences of* $\mathcal{K}$ *and* $P_\hbar\mathcal{K}P_\hbar$ *respectively. In particular* ($\alpha = 1/2$)

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq c_{\mathcal{K},\mathcal{S},s}\Upsilon(\hbar)^2\lambda_m^{1/2-s} \quad \forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}. \tag{2.19}$$

Proof. The desired estimate (2.18) follows immediately by interpolation with logarithmic weight α between the second term in the upper bound (2.17) and the trivial estimate $0 \leq \lambda_m - \lambda_{\hbar,m} \leq \lambda_m$, via the inequality $\min\{x, y+z\} \leq y + x^\alpha z^{1-\alpha}$ $\forall x, y, z \geq 0$. ∎

We note next that in contrast to (2.19), the trace discretization error retains the correct scaling factor (λ_m and not $\lambda_m^{1/2}$), in the following sense.

THEOREM 2.9. *If Assumption* 2.6 *is satisfied, then for any* $0 < s < 1$ *there exists a constant* $c_{\mathcal{K},\mathcal{S},s} > 0$ *such that*

$$0 \leq \text{Tr } \mathcal{K} - \text{Tr } P_\hbar\mathcal{K}P_\hbar \leq c_{\mathcal{K},\mathcal{S},s}\Upsilon(\hbar)^2\sum_{m\in\mathbb{N}_+}\lambda_m^{1-s} \quad \forall \hbar \in \mathfrak{H}. \tag{2.20}$$

Proof. The lower bound follows trivially from $\lambda_{\hbar,m} \leq \lambda_m$ $\forall m \in \mathbb{N}_+$.

Further, the identity

$$\mathcal{K} - P_\hbar\mathcal{K}P_\hbar = (I - P_\hbar)\mathcal{K} + \mathcal{K}(I - P_\hbar) - (I - P_\hbar)\mathcal{K}(I - P_\hbar)$$

and the fact that $\mathcal{K}$ is nonnegative ensure

$$\langle(\mathcal{K} - P_\hbar\mathcal{K}P_\hbar)\phi, \phi\rangle_H \leq 2|\langle\mathcal{K}\phi, (I - P_\hbar)\phi\rangle_H| \quad \forall \phi \in H. \tag{2.21}$$

Using (2.21) and Assumption 2.6 it follows

$$\begin{aligned}\text{Tr } \mathcal{K} - \text{Tr } P_\hbar\mathcal{K}P_\hbar &= \sum_{m=1}^{\infty}\langle(\mathcal{K} - P_\hbar\mathcal{K}P_\hbar)\phi_m, \phi_m\rangle_H \leq 2\sum_{m=1}^{\infty}|\langle\mathcal{K}\phi_m, (I - P_\hbar)\phi_m\rangle_H| \\ &\leq 2\sum_{m=1}^{\infty}\lambda_m\|(I - P_\hbar)\phi_m\|_H^2 \leq c_{\mathcal{K},\mathcal{S},s}\Upsilon(\hbar)^2\sum_{m=1}^{\infty}\lambda_m^{1-s},\end{aligned}$$

and the proof is concluded. ∎

2.2. $\mathcal{K}^\delta$ spectrum approximation. For a given $\delta > 0$, a simple argument based on the Lipschitz continuity of

$$]0,\infty[\ni x \to x^\delta \in]0,\infty[$$

shows that the computed eigenvalues of $\mathcal{K}^\delta$, namely, $(\lambda^\delta_{\hbar,m})_{m\in\mathbb{N}_+}$, are good approximations of the exact eigenvalues $(\lambda^\delta_m)_{m\in\mathbb{N}_+}$ of $\mathcal{K}^\delta$.

THEOREM 2.10. *If Assumption* 2.6 *is satisfied, then for any* $0 \le \alpha < \delta \le 1$,

$$0 \le \sum_{m=1}^{\infty} (\lambda^\delta_m - \lambda^\delta_{\hbar,m}) \le c_{K,\mathcal{S},\alpha,\delta} \max\{\Upsilon(\hbar)^2, \Upsilon(\hbar)^{4\alpha}\} \quad \forall \hbar \in \mathfrak{H}. \tag{2.22}$$

Proof. From the Lipschitz condition

$$\lambda^\delta_m - \lambda^\delta_{\hbar,m} \le \lambda_m^{\delta-1}(\lambda_m - \lambda_{\hbar,m})$$

and (2.17) we obtain that for any $s > 0$ there exists a constant $c_{K,\mathcal{S},s} > 0$ such that

$$0 \le \lambda^\delta_m - \lambda^\delta_{\hbar,m} \le c_{K,\mathcal{S},s}(\Upsilon(\hbar)^2 \lambda_m^{\delta-s} + \Upsilon(\hbar)^4 \lambda_m^{\delta-1-2s}) \tag{2.23}$$

$\forall m \in \mathbb{N}_+, \forall \hbar \in \mathfrak{H}$. Interpolation with logarithmic weight $\alpha \in [0, \delta[$ between the second term in the r.h.s. of (2.23) and the trivial estimate $\lambda^\delta_m - \lambda^\delta_{\hbar,m} \le \lambda^\delta_m$ yields

$$0 \le \lambda^\delta_m - \lambda^\delta_{\hbar,m} \le c_{K,\mathcal{S},s}(\Upsilon(\hbar)^2 \lambda_m^{\delta-s} + \Upsilon(\hbar)^{4\alpha} \lambda_m^{\delta-(1+2s)\alpha}). \tag{2.24}$$

Choosing $s > 0$ small enough to ensure $(1 + 2s)\alpha < \delta$ and summing (2.24) over $m \in \mathbb{N}_+$ we obtain (2.22). ∎

3. Application to integral operators. We check the validity of Assumption 2.6 in the case of an integral operator $\mathcal{K}$ with piecewise smooth kernel K, if standard h FEM is used to construct the finite element space family $\mathcal{S}$. We begin with a review of eigenvalue decay rates in terms of the kernel regularity, which are immediately seen to ensure $\mathcal{K} \in \cap_{p>0}\mathcal{B}_p(H)$. We then show that (2.16) is a consequence of standard h FEM error estimates and Gagliardo–Nirenberg inequalities ensuring mild eigenfunction oscillations for $\mathcal{K}$.

3.1. Eigenvalue decay. The results we present in this section are standard (see, e.g., [6], [10]), following from the abstract theory of Weyl/approximation/entropy numbers via approximation of K by discrete, finite rank (separable w.r.t. (x, x')) kernels. Roughly speaking, the smoother the kernel the faster the eigenvalue decay, with finite Sobolev regularity implying algebraic decay and analyticity giving rise to quasi-exponential decay.

Remarkably, all these results hold for piecewise regular kernels on product subdomains of D, in the sense of Definition 3.1. Note that general piecewise regularity allowing singularities on the diagonal set of $D \times D$ ensure in general only a slower eigenvalue decay. (See, e.g., [6] and [4] for examples with known exact eigenelements.)

DEFINITION 3.1. *If D is a bounded domain in $\mathbb{R}^d$ and $p, q \in [0, \infty[$, a measurable function $K : D \times D \to \mathbb{R}$ is said to be piecewise $H^{p,q}$ on $D \times D$ if there exists a finite family $\mathcal{D} = (D_j)_{j\in\mathcal{J}}$ of subdomains of D such that*

i. $D_j \cap D_{j'} = \emptyset \quad \forall j, j' \in \mathcal{J}$ *with* $j \ne j'$,
ii. $D \setminus \bigcup_{j\in\mathcal{J}} D_j$ *is a null set in* $\mathbb{R}^d$,
iii. $\overline{D} \subset \bigcup_{j\in\mathcal{J}} \overline{D_j}$,
iv. $K|_{D_j \times D_{j'}} \in H^{p,q}(D_j \times D_{j'}) := H^p(D_j) \otimes H^q(D_{j'}) \quad \forall j, j' \in \mathcal{J}$.

We denote by $H^{p,q}_{\mathcal{D}}(D^2)$ the space of piecewise $H^{p,q}$ functions on $D \times D$ in the sense given above.

Moreover, if there exists also a finite family $\mathcal{G} = (G_j)_{j\in\mathcal{J}}$ *of open sets in* $\mathbb{R}^d$ *such that*

v. $\overline{D_j} \subset G_j \quad \forall j \in \mathcal{J}$,

vi. $K|_{D_j\times D_{j'}}$ *has an* $H^{p,q}$ *continuation to* $G_j \times G_{j'} \quad \forall j, j' \in \mathcal{J}$,

then we say that K *is piecewise* $H^{p,q}$ *on a covering of* $D \times D$ *and we denote by* $H^{p,q}_{\mathcal{D},\mathcal{G}}(D^2)$ *the corresponding space.*

Similarly we introduce spaces of piecewise regular functions defined on D*, which we denote by* $H^p_{\mathcal{D}}(D)$*,* $H^p_{\mathcal{D},\mathcal{G}}(D)$*, etc.*

For kernels with finite Sobolev regularity the next proposition holds.

PROPOSITION 3.2. *Let* $D \subset \mathbb{R}^d$ *be a bounded domain and* $K \in L^2(D \times D)$ *be a symmetric kernel defining a compact nonnegative integral operator* $\mathcal{K}$ *via* (1.1)*. If* $K \in H^{p,0}_{\mathcal{D},\mathcal{G}}(D^2)$ *for some* $p \in [0, \infty[$*, then there exists a constant* $c_K > 0$ *such that*

$$0 \le \lambda_m \le c_K m^{-p/d} \quad \forall m \in \mathbb{N}_+. \tag{3.1}$$

COROLLARY 3.3. *Let* $D \subset \mathbb{R}^d$ *be a bounded domain and* $K \in L^2(D \times D)$ *be a symmetric kernel defining a compact nonnegative integral operator* $\mathcal{K}$ *via* (1.1)*. If* K *is piecewise smooth (i.e., piecewise* $H^{p,q}_{\mathcal{D},\mathcal{G}}(D^2)$ $\forall p, q \in [0, \infty[$*) on a covering of* $D \times D$ *and* $(\lambda_m)_{m\in\mathbb{N}_+}$ *denotes the eigenvalue sequence of* $\mathcal{K}$*, then for any* $s > 0$ *there exists a constant* $c_{K,s} > 0$ *such that*

$$0 \le \lambda_m \le c_{K,s} m^{-s} \quad \forall m \in \mathbb{N}_+ \tag{3.2}$$

so that $\mathcal{K} \in \cap_{p>0}\mathcal{B}_p(L^2(D))$.

Example 3.4. One is often interested in Gaussian kernels of the form

$$K(x, x') := \sigma^2 \exp(-|x - x'|^2/(\gamma^2\Lambda^2)) \quad \forall (x, x') \in D \times D, \tag{3.3}$$

where $\sigma, \gamma > 0$ are real parameters (standard deviation, correlation length) and Λ is the diameter of the domain D. K given by (3.3) has an entire continuation to $\mathbb{C}^d$ and defines a nonnegative compact operator via (1.1).

Note that for piecewise analytic kernels the next proposition holds.

PROPOSITION 3.5. *Let* $K \in L^2(D\times D)$ *be a symmetric kernel defining a compact nonnegative integral operator via* (1.1)*. If* $K \in \mathcal{A}_{\mathcal{D},\mathcal{G}}(D^2)$ *(defined analogously to* $H^{p,q}_{\mathcal{D},\mathcal{G}}$*) and* $(\lambda_m)_{m\in\mathbb{N}_+}$ *denotes the eigenvalue sequence of* $\mathcal{K}$*, then there exist constants* $c_{1,K}, c_{2,K} > 0$ *such that*

$$0 \le \lambda_m \le c_{1,K} e^{-c_{2,K} m^{1/d}} \quad \forall m \in \mathbb{N}_+. \tag{3.4}$$

Since K given by (3.3) admits an analytic continuation to the whole complex space $\mathbb{C}^d \times \mathbb{C}^d$, the eigenvalue decay is in this case even faster than in (3.4).

PROPOSITION 3.6. *If* $K \in L^2(D \times D)$ *is given by* (3.3)*, then for the eigenvalue sequence* $(\lambda_m)_{m\in\mathbb{N}_+}$ *of* $\mathcal{K}$ *defined by* (1.1) *it holds*

$$0 \le \lambda_m \le c_{\sigma,\gamma,D} \frac{(1/\gamma\Lambda)^{m^{1/d}}}{\Gamma(m^{1/d}/2)} \quad \forall m \in \mathbb{N}_+. \tag{3.5}$$

3.2. Eigenfunction oscillations. We show next that the piecewise smoothness assumption on the kernel allows also a good control of the eigenfunctions and their derivatives in the $L^\infty(D)$ norm. Roughly speaking, the eigenfunctions are shown to be bounded from above, asymptotically as $m \to \infty$, by any negative power of the

corresponding eigenvalue. In other words, the eigenfunction oscillations are much weaker than the eigenvalue decay rate.

First we note that piecewise regularity of eigenfunctions follows from that of the kernel K.

PROPOSITION 3.7. *If $K \in H^{p,q}_{\mathcal{D},\mathcal{G}}(D^2)$, then the eigenfunctions of $\mathcal{K}$ given by* (1.1) *corresponding to nontrivial eigenvalues belong to $H^p_{\mathcal{D},\mathcal{G}}(D)$.*

Proof. The conclusion follows at once from the eigenvalue equation

$$\phi_m(x) = \frac{1}{\lambda_m} \sum_{j' \in \mathcal{J}} \int_{D_{j'}} K(x,x')\phi_m(x')\,dx' \quad \forall x \in D_j, \tag{3.6}$$

which can be naturally extended to G_j by replacing K by its $H^{p,q}$ continuation on $G_j \times G_{j'}$. ∎

Remark 3.8. *Similarly, if $K \in H^{p,q}_{\mathcal{D}}(D^2)$, then the eigenfunctions of $\mathcal{K}$ corresponding to nontrivial eigenvalues belong to $H^p_{\mathcal{D}}(D)$.*

The following result due to Ehrling–Nirenberg–Gagliardo (see [1, Theorem 4.14]) is essential for our analysis.

THEOREM 3.9. *Let $D \subset \mathbb{R}^d$ be a bounded domain having the uniform cone property and $\varepsilon_0 \in (0,\infty)$, $n \in \mathbb{N}$, $p \in [1,\infty)$. Then there exists $c_{\varepsilon_0,n,p,D} > 0$ such that for any $\varepsilon \in (0,\varepsilon_0]$, $l \in \{0,1,\dots,n-1\}$ and $u \in W^{n,p}(D)$,*

$$|u|_{l,p} \le c_{\varepsilon_0,n,p,D} \left\{ \varepsilon |u|_{n,p} + \varepsilon^{-l/(n-l)} |u|_{0,p} \right\}, \tag{3.7}$$

where

$$|u|^p_{l,p} := \int_D \sum_{|\alpha|=l} |\partial^\alpha u(x)|^p \,dx.$$

PROPOSITION 3.10. *For $D \subset \mathbb{R}^d$ a bounded domain and K piecewise smooth on $D \times D$ such that the domains D_j in Definition* 3.1 *all have the uniform cone property, we denote by $(\lambda_m,\phi_m)_{m\in\mathbb{N}_+}$ the eigenelements of the associated integral operator $\mathcal{K}$ via* (1.1)*, such that $\|\phi_m\|_{L^2(D)} = 1$ $\forall m \in \mathbb{N}_+$. Then for any $s > 0$ and any multiindex $\alpha \in \mathbb{N}^d$ there exists $c_{K,s,\alpha} > 0$ such that*

$$\|\partial^\alpha \phi_m\|_{L^\infty(D_j)} \le c_{K,s,\alpha} \, |\lambda_m|^{-s} \quad \forall m \in \mathbb{N}_+, \forall j \in \mathcal{J}. \tag{3.8}$$

Proof. We first note that the eigenvalue equation (3.6) implies (by differentiating and applying the Cauchy–Schwarz inequality to estimate the resulting integrals) for any $\alpha \in \mathbb{N}^d$ the existence of a constant $c_{K,\alpha} > 0$ such that

$$\|\partial^\alpha \phi_m\|_{L^\infty(D_j)} \le \; c_{K,\alpha} |\lambda_m|^{-1} \quad \forall m \in \mathbb{N}_+, \forall j \in \mathcal{J}. \tag{3.9}$$

We apply now Theorem 3.9 on D_j with $p = 2$, $\varepsilon_0 := \max_{m\in\mathbb{N}_+} |\lambda_m|$ and choose in (3.7) $\varepsilon = \lambda_m$, $u = \phi_m$ for an arbitrary $m \in \mathbb{N}_+$ (we assume w.l.o.g. $\lambda_m \neq 0$). It follows that for any $n \in \mathbb{N}$ there exists $c_{\varepsilon_0,n,D_j} > 0$ such that for all $l \in \{0,1,\dots,n-1\}$

$$\begin{aligned} |\phi_m|_{D_j,l,2} &\le c_{\varepsilon_0,n,D_j} \left\{ \lambda_m |\phi_m|_{D_j,n,2} + \lambda_m^{-l/(n-l)} |\phi_m|_{D_j,0,2} \right\} \\ &\le c_{\varepsilon_0,n,D_j,K} \left\{ 1 + \lambda_m^{-l/(n-l)} \right\} \le c_{\varepsilon_0,n,D_j,K} \lambda_m^{-l/(n-l)}, \end{aligned} \tag{3.10}$$

due to (3.9).

Now, for any $s > 0$ and $\alpha \in \mathbb{N}^d$ we choose $l = \lceil d/2 \rceil + |\alpha|$ and $n > l$ such that $l/(n-l) < s$. From (3.10) and the Sobolev embedding theorems we deduce then

$$\begin{aligned}\|\partial^\alpha \phi_m\|_{L^\infty(D_j)} &\le c_{\alpha,D_j}\|\phi_m\|_{H^l(D_j)} \le c_{\alpha,D_j}\sum_{k=0}^{l}|\phi_m|_{D_j,k,2}\\ &\le c_{\varepsilon_0,n,D_j,K,\alpha}\sum_{k=0}^{l}\lambda_m^{-k/(n-k)}\\ &\le c_{\varepsilon_0,n,D_j,K,\alpha}\lambda_m^{-l/(n-l)} \le c_{\varepsilon_0,n,D_j,K,\alpha}\lambda_m^{-s}\end{aligned}$$

$\forall m \in \mathbb{N}_+$, and the proof is concluded. ∎

Remark 3.11. Under the regularity assumptions of Proposition 3.10 the estimate (3.8) is optimal in the sense that for any α it fails to hold with $s = 0$. This can be seen, e.g., on $D :=]0,1[$ by taking $K := \sum_{m\ge1}\lambda_m\,\phi_m \otimes \phi_m$ with

$$\lambda_m := e^{-m}, \quad \phi_m(x) := m\phi(m^2x - m) \quad \forall x \in]0,1[, \forall m \in \mathbb{N}_+,$$

where $\phi \in C_0^\infty(]0,1[)$ satisfies $\|\phi\|_{L^2(]0,1[)} = 1$.

Remark 3.12. It can be shown that further assumptions, like stationarity of the kernel, i.e., $K(x,x') = k(x-x')$ for some $k : \mathbb{R}^d \to \mathbb{R}$, lead to the uniform L^∞ boundedness of the eigenfunctions (but not of their derivatives).

Remark 3.13. If K is not piecewise smooth in the sense of Definition 3.1 (for instance, if the singularities of K lie on the diagonal set of $D \times D$, as it is the case for some usual stationary kernels like, e.g., $K(x,x') = k(x-x') = \exp(-|x-x'|^{1+\delta})$ with $0 \le \delta < 1$), then estimates of type (3.8) hold true only for $s > c_k > 0$, where the constant $c_k \in]0,1]$ depends on the Sobolev regularity of k in $\mathbb{R}^d$. (See also [4] for such examples with known exact eigenelements.)

3.3. *h* FEM. For the integral operator (1.1) with smooth kernel K we can check now the validity of Assumption 2.6, if the standard h FEM is used to construct the finite element space family $\mathcal{S}$. For a fixed $p \in \mathbb{N}_+$ and with $\hbar := h \in]0,\infty]$ we define $S_\hbar = S_h := S_h^p$ to be the space of discontinuous piecewise polynomials of total degree at most $p-1$ on a regular mesh T_h in $\overline{D}$ of width h and subordinate to $\mathcal{D}$, i.e., to the covering $(\overline{D_j})_{j\in\mathcal{J}}$. With this choice the next proposition holds.

PROPOSITION 3.14. *If $K \in L^2(D\times D)$ is piecewise smooth on a covering of $D \times D$, defining a compact nonnegative integral operator $\mathcal{K}$ via* (1.1) *and $p \in \mathbb{N}_+$, $S_\hbar = S_h^p$ $\forall \hbar = h \in]0,\infty]$, then Assumption* 2.6 *holds with $\Upsilon(\hbar) := h^p$.*

Proof. The fast eigenvalue decay has been established in Corollary 3.3, so that we only have to check (2.16). To this end, we note that the standard h FEM approximation property holds for S_h,

$$\|\phi - P_\hbar\phi\|_{L^2(D)} \le c_{p,\mathcal{D}}h^p|\phi|_p \quad \forall h > 0,\, \forall \phi \in H^p_{\mathcal{D}}(D), \tag{3.11}$$

where

$$|\phi|_p^2 := \sum_{j\in\mathcal{J}}\sum_{|\alpha|=p}\|\partial^\alpha\phi\|^2_{L^2(D_j)} \quad \forall \phi \in H^p_{\mathcal{D}}(D).$$

The conclusion follows then by applying (3.11) to $\phi := \phi_m \in H^p_{\mathcal{D}}(D)$ (in view of Remark 3.8), and using Proposition 3.10 to estimate $|\phi_m|_p$ in terms of a given $s > 0$ and λ_m. ∎

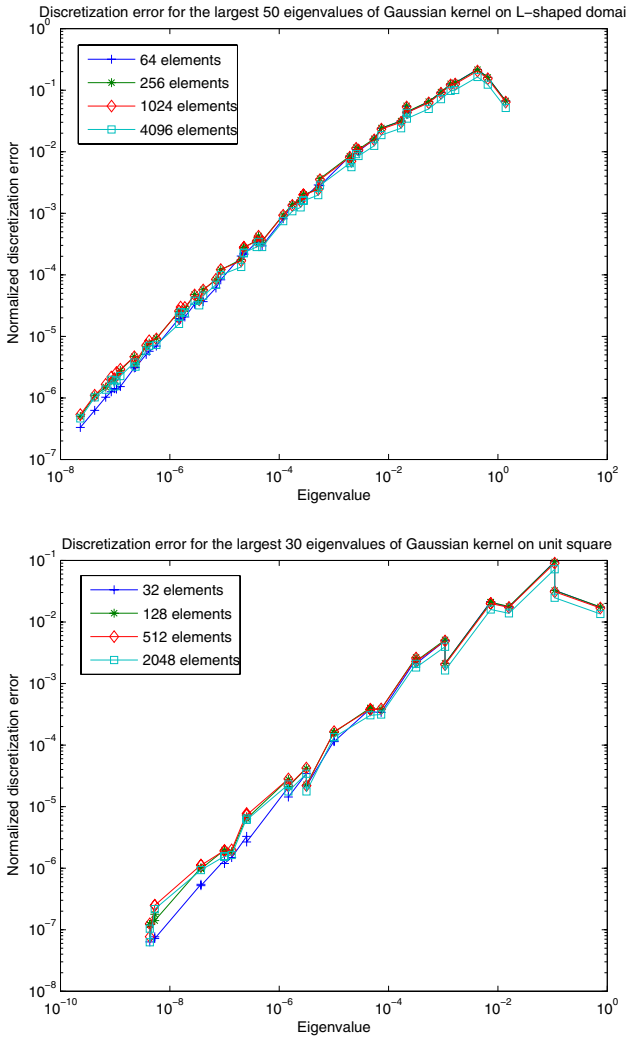


FIG. 1. *Eigenvalue discretization error for Gaussian kernel* $K(x,x') = \exp(-|x-x'|^2)$ *on two-dimensional L-shaped domain (top) and unit square (bottom).*

Theorem 1.5 is now just a reformulation of Proposition 3.14, whereas Theorem 1.1 follows directly from the main abstract result, Theorem 2.7.

Remark 3.15. For the h FEM applied to the integral operator (1.1) with piecewise smooth kernel, (2.22) becomes

$$0 \leq \sum_{m=1}^{\infty} (\lambda_m^{\delta} - \lambda_{h,m}^{\delta}) \leq c_{K,\mathcal{S},\alpha,\delta} h^{p \min\{2,4\alpha\}} \quad \forall h \in]0,1], \forall \alpha < \delta.$$

3.4. Numerical tests. The results of the eigenvalues computation for the case of a Gaussian kernel $K(x,x') = \exp(-|x-x'|^2)$ on an L-shaped domain and on the unit square, respectively, are presented in Figure 1. In both cases we employ piecewise constant elements on a regular mesh, at discretization levels 2, 3, 4, and

5 (corresponding to 64, 256, 1024, and 4096 elements for the L-shaped and to 32, 128, 512, and 2048 elements for the unit square, respectively). We use the results of an overkill computation on level 6 (with 16384 and 8192 elements, respectively) as exact eigenvalues. In Figure 1 we plot the normalized (w.r.t. meshwidth h) discretization error $(\lambda_m - \lambda_{m,\hbar})/h^2$ versus the corresponding eigenvalue λ_m for the first 50 eigenvalues ($m = 1, 2, \ldots, 50$) in the case of the L-shaped domain and the first 30 eigenvalues for the unit square. Note that the separability of the Gaussian kernel on the unit square ensures the existence of multiple eigenvalues. In both cases the slope of the resulting curve appears to be close to 1, validating thus the theoretical result (Theorem 1.1), which predicted a slope of at least 1/2. Moreover, Figure 1 and the exact error representation formula (1.5) suggest the following question.

OPEN QUESTION 3.16. *Is it true that in the case of an analytic (or just smooth) kernel K, the curve slopes in Figure* 1 *come arbitrarily close to* 1*? More precisely, does the estimate stronger than* (1.7)*,*

$$0 \leq \lambda_m - \lambda_{\hbar,m} \leq c_{K,\mathcal{T},p,s} h^{2p} \lambda_m^{1-s} \quad \forall m \in \mathbb{N}_+, \forall h > 0, \forall s > 0, \tag{3.12}$$

hold under the assumptions of Theorem 1.1*?*

Acknowledgments. The author would like to thank Prof. Christoph Schwab of ETH Zürich for helpful discussions during the preparation of this article as well as Dr. Gregor Schmidlin for providing the software and assistance with the eigenvalue computation.

REFERENCES

[1] W. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1978.
[2] I. BABUŠKA AND J. E. OSBORN, *Eigenvalue problems*, in Handb. Numer. Anal., 2, North–Holland, Amsterdam, 1991, pp. 641–787.
[3] F. CHATELIN, *Spectral Approximations of Linear Operators*, Academic Press, New York, 1983.
[4] R. GHANEM AND P. D. SPANOS, *Stochastic Finite Elements: A Spectral Approach*, Springer, New York, 1991.
[5] T. KATO, *Perturbation Theory for Linear Operators*, Grundlehren Text Ed. 132, Springer, New York, 1984.
[6] H. KÖNIG, *Eigenvalue Distribution of Compact Operators*, Oper. Theory Adv. Appl., 16, Birkhäuser, Basel, 1986.
[7] M. A. KRASNOSELSKI, G. M. VAINIKKO, P. P. ZABREIKO, Y. B. RUTITSKII, AND V. Y. STETSENKO, *Approximate Solution of Operator Equations*, Wolters, Groeningen, 1972.
[8] M. LOÈVE, *Probability Theory*, vols. I and II, Springer, New York, 1978.
[9] J. E. OSBORN, *Spectral approximation for compact operators*, Math. Comput., 29 (1975), pp. 712–725.
[10] A. PIETSCH, *Eigenvalues and s-Numbers*, Math. Anwendungen 43, Geest & Portig, Leipzig, W. Germany, 1987.
[11] M. SCHLATHER, *Introduction to Positive Definite Functions and to Unconditional Simulation of Random Fields*, Technical Report ST-99-10, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK, 1999.
[12] G. M. VAINIKKO, *Error estimates of the Bubnov-Galerkin method in an eigenvalue problem*, USSR Comput. Math. Math. Phys., 5 (1965), pp. 587-607.

SIAM J. NUMER. ANAL.
Vol. 44, No. 2, pp. 879–895

CONVERGENT DIFFERENCE SCHEMES FOR DEGENERATE ELLIPTIC AND PARABOLIC EQUATIONS: HAMILTON–JACOBI EQUATIONS AND FREE BOUNDARY PROBLEMS*

ADAM M. OBERMAN†

Abstract. Convergent numerical schemes for degenerate elliptic partial differential equations are constructed and implemented. Simple conditions are identified which ensure that nonlinear finite difference schemes are monotone and nonexpansive in the maximum norm. Explicit schemes endowed with an explicit CFL condition are built for time-dependent equations and are used to solve stationary equations iteratively. Explicit and implicit formulations of monotonicity for first- and second-order equations are unified. Bounds on orders of accuracy are established. An example of a scheme which is stable, but nonmonotone and nonconvergent, is presented. Schemes for Hamilton–Jacobi equations, obstacle problems, one-phase free boundary problems, and stochastic games are built and computational results are presented.

Key words. finite difference schemes, partial differential equations, monotone schemes, viscosity solution, Hamilton–Jabobi equation, free boundary problems

AMS subject classifications. 65N06, 65N12, 65M06, 65M12, 35B50, 35J60, 35R35, 35K65, 49L25

DOI. 10.1137/S0036142903435235

1. Introduction.

We devise practical techniques for building convergent numerical schemes for a class of nonlinear partial differential equations. This is the class of *degenerate elliptic* (in the sense of Crandall, Ishii, and Lions [11]) partial differential equations, for which unique viscosity solutions exist. The class includes Hamilton–Jacobi equations, which are nonlinear first-order equations; elliptic equations which may be degenerate; and fully nonlinear second-order equations. It also includes free boundary problems and the equation for the value function from control and game theory.

The approximation theory developed by Barles and Souganidis [5] provides the following criteria for the convergence of approximation schemes: monotone, consistent, and stable schemes converge to the unique viscosity solution of a degenerate elliptic equation. Despite the clear requirements of the theory, building monotone schemes remains a challenge for many important equations. The finite difference method is the natural method for building monotone schemes, but conditions which ensure monotonicity are different for first- and second-order equations, and for explicit and implicit schemes.

For linear elliptic equations, Motzkin and Wasow [28] introduced the notion that a scheme is of "positive type." These linear schemes respect the discrete maximum principle. This formulation of monotonicity was further studied in [7] and later generalized to nonlinear elliptic equations by Kuo and Trudinger [23, 24, 25, 26]. Related notions for linear parabolic equations have been studied [1, 21, 35].

*Received by the editors September 20, 2003; accepted for publication (in revised form) September 12, 2004; published electronically April 21, 2006. An earlier version of this paper was selected for the Leslie Fox prize in numerical analysis.
http://www.siam.org/journals/sinum/44-2/43523.html
†Department of Mathematics, Simon Fraser University, Burnaby, BC V5A IS6, Canada (aoberman@math.sfu.ca).

For conservation laws, monotonicity[1] is associated with entropy solutions. In this setting, monotone schemes are contractions in ℓ^1 [13] and are at most first-order accurate [18]. Higher-order accuracy is achieved by essentially nonoscillatory (ENO) or weighted ENO (WENO) schemes [19], which are not monotone: they selectively use high-order (nonmonotone) interpolation in smooth regions of the solution and monotone schemes in nonsmooth regions.

For Hamilton–Jacobi equations, monotonicity is necessary for convergence. Early numerical papers studied explicit schemes for time-dependent equations on uniform grids [10, 33]. A number of methods have since been developed, which include fast marching [32], fast sweeping [34], semi-Lagrangian [16], central [27], and ENO [31].

For certain second-order equations, which include fully nonlinear equations and degenerate linear elliptic equations, monotonicity is necessary for convergence. An early result of Motzkin and Wasow illustrated difficulties associated with monotone schemes: even for linear elliptic equations, in general it is not possible to build monotone schemes using a narrow stencil [28]. Very large stencil schemes for quasi-linear equations were studied by Crandall and Lions [12]. Wide stencil schemes have been used to solve certain degenerate second-order equations [29, 30].

We identify a class of nonlinear finite difference schemes which we call *degenerate elliptic*. Degenerate elliptic schemes are monotone. They also enjoy a strong form of stability: they are nonexpansive in the maximum norm. The class includes implicit or explicit schemes for first- or second-order equations on structured or unstructured grids.

Degenerate elliptic schemes are built in simple ways from building blocks consisting of schemes for basic equations. They begin with an implicit scheme for the spatial part of the equation. This scheme may then be extended to an explicit scheme for the time-dependent equation, or equivalently, to an iterative method for the stationary equation. The explicit scheme is endowed with a nonlinear CFL condition which is easily calculated.

A guiding principle of this work is that, in order to build effective numerical schemes, it is essential to have a thorough understanding of the underlying equations. In this manner, schemes can be built that inherit desirable properties from the equations.

THEOREM 1. *The solution operator of a degenerate elliptic partial differential equation is monotone and nonexpansive in the maximum norm, provided mild analytic conditions hold so that it is well defined.*

THEOREM 2. *The solution operator of a degenerate elliptic finite difference scheme is monotone and nonexpansive in the maximum norm, provided mild analytic conditions hold so that it is well defined.*

THEOREM 3. *A scheme is monotone and nonexpansive in the ℓ^∞ norm if and only if it is degenerate elliptic.*

Remark 1. Monotonicity by itself does not ensure stability. For example, $u_j^{n+1} = 2u_j^n$ is unstable; examples with worse growth rates are easily constructed.

While degenerate ellipticity is stronger than monotonicity for abstract schemes, the condition occurs naturally for schemes built using the finite difference method; see section 2.3. For these schemes, the two conditions are equivalent.

Remark 2. Theorem 3 recalls a theorem from [14], which states that monotonicity is equivalent to nonexpansivity in ℓ^∞, for mappings which are invariant under

[1]Not to be confused with "monotonicity preserving," which means that increasing functions on the line remain increasing.

translation by a constant.

THEOREM 4. *The accuracy of a monotone finite difference scheme is at most first order for first-order equations and at most second order for second-order equations.*

Remark 3. Useful numerical solutions can be obtained with first- or second-order schemes. Despite the fact that singularities occur, the accuracy requirements are not as high in this setting are they are for conservation laws.

Contents. Section 1.1 demonstrates the equivalence of the explicit and implicit formulations of monotonicity. Section 1.2 contains basic examples to illustrate the definitions. Section 1.3 contains an example which shows monotonicity is necessary for convergence.

Section 2 is the bulk of the theory. Section 2.1 summarizes relevant aspects of degenerate elliptic equations. Section 2.2 defines the class of degenerate elliptic *equations*, followed by section 2.3, which defines the class of degenerate elliptic *schemes*. Section 2.4 provides the nonlinear CFL-type condition for explicit schemes. Section 2.5 establishes properties of the solution operator. Section 2.6 contains proofs of Theorems 1, 2, 3, and 4.

A technique is developed in section 3 to build schemes for complicated equations using building blocks consisting of schemes for simpler equations. This technique is used to build schemes for various equations, including Hamilton–Jacobi equations, obstacle problems, one-phase free boundary problems, and stochastic games. In section 4 computational results are presented.

1.1. Equivalent formulations of monotonicity. In its most general formulation, monotonicity means that the comparison principle holds. This global property was used in [5] to prove convergence of nondiscrete approximation schemes.

For the purpose of building schemes, it is useful to have an easily verified local condition which guarantees monotonicity. The condition comes in two forms. The explicit formulation, usually seen for time-dependent equations, is

$$u_i = H^i\left(u|_{j=N(i)}\right), \tag{1}$$

where $N(i)$ is the list of neighbors of u_i. For example, it appears as $u_i^{n+1} = H(u_{j-1}^n, u_j^n, u_{j+1}^n)$ in the case of three-point explicit schemes [10]. The explicit formulation (1) is monotone if H^i is a nondecreasing function of each variable. The implicit formulation, usually seen for stationary elliptic equations, is

$$F^i\left(u_i, u|_{j=N(i)}\right) = 0. \tag{2}$$

For example, linear schemes $\sum_{i=0}^n a_i u(x+idx)$ are monotone (of "positive type") if $a_0 \geq 0$ and $a_i \leq 0$ for $i \neq 0$ [28]. For nonlinear equations, schemes are monotone if F^i is nondecreasing in the first variable and nonincreasing in the remaining variables [24].

Remark 4. The two formulations are formally equivalent. To put the explicit form into an implicit form is trivial. To go from the implicit form to the explicit form, differentiate implicitly to obtain $D_{u_i}F^i du_i + \sum_{j=N(i)} D_{u_j}F^i du_j = 0$. Fixing all but neighbor u_k, we obtain $du_i/du_k = -D_{u_k}F^i/D_{u_i}F^i \geq 0$. Use the implicit function theorem to solve for u_i as a nondecreasing function of the neighbors.

1.2. Illustration of the definitions. Consider the standard centered difference scheme for $-u_{xx}$, $(u_i - u_{i-1} + u_i - u_{i+1})/dx^2$. The scheme is degenerate elliptic, since, as in Definition 2, it is a nondecreasing function of the differences between the reference variable and its neighbors, $u_i - u_{i-1}$ and $u_i - u_{i+1}$. Solving for u_i gives

$u_i = \frac{1}{2}(u_{i+1} + u_{i-1})$. This puts the scheme in the explicit form of monotonicity, since the righthand side is a nondecreasing function of its arguments.

The implicit Euler scheme for the heat equation, $u_t - u_{xx} = 0$, is also degenerate elliptic,

$$\frac{1}{dt}(u_i^n - u_i^{n-1}) + \frac{1}{dx^2}\left(u_i^n - u_{i-1}^n + u_i^n - u_{i+1}^n\right) = 0.$$

On the other hand, the explicit Euler scheme,

$$u_i^n = (1 - 2dt/dx^2)u_i^{n-1} + (dt/dx^2)(u_{i-1}^{n-1} + u_{i+1}^{n-1}),$$

is monotone if and only if $0 \le dt \le dx^2/2$. In this case, the scheme is also degenerate elliptic,

$$\left(1 - 2dt/dx^2\right)\left(u_i^n - u_i^{n-1}\right) + dt/dx^2(u_i^n - u_{i-1}^{n-1} + u_i^n - u_{i+1}^{n-1}) = 0.$$

Remark 5. The restriction on the time step coincides with the usual CFL condition [9], which is a condition for stability in ℓ^2. In general, these conditions do not coincide.

Remark 6. The example generalizes naturally to nonlinear schemes. As an exercise for the reader, repeat the example with $|u_x|$ instead of $-u_{xx}$. Use the discretization $\max\{u_i - u_{i-1}, u_i - u_{i+1}, 0\}/dx$. The resulting nonlinear CFL condition is $0 \le dt \le dx$.

1.3. A stable but nonconvergent scheme. In this section, we give an example of a difference scheme which is stable but nonmonotone and nonconvergent. The example involves the linear, but degenerate, second-order elliptic equation,

$$-(u_{xx} + 2u_{xy} + u_{yy}) = -\frac{d^2u}{dv^2} = 0, \qquad v = (1,1),$$

along with Dirichlet boundary conditions on the unit square. Continuous functions of the form $f(x - y)$, whose level sets are straight lines in the direction of v, are viscosity solutions. The equation is degenerate: it has a zero eigenvalue in the direction perpendicular to v.

Two discretizations. We present two consistent, second-order accurate difference schemes on a uniform grid with spacing h. For the first scheme, simply use the centered second difference in the diagonal direction,

$$\frac{1}{h^2}\left(u(x+h, y+h) - 2u(x,y) + u(x-h, y-h)\right).$$

For the second scheme, use centered differences for u_{xx} and u_{yy}, and a symmetric centered difference for u_{xy}, to obtain

$$\begin{aligned}\frac{1}{h^2}\big(&2u(x+h,y) + 2u(x,y+h) + 2u(x-h,y) + 2u(x,y-h)\\ &- 6u(x,y) - u(x+h,y-h) - u(x-h,y+h)\big).\end{aligned}$$

The first scheme is degenerate elliptic. The second scheme is not, since the coefficients of the values at grid points $(x+h, y-h)$ and $(x-h, y+h)$ are negative.

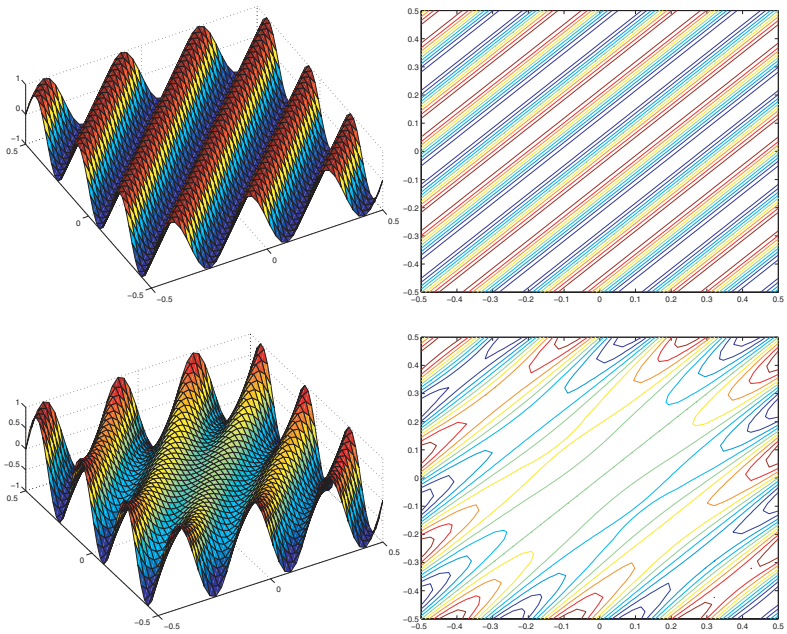


FIG. 1. *Solution and level sets, computed using the first (top) and second (bottom) scheme.*

Numerical experiments. The first scheme converges. The second scheme gives errors of a size comparable to the data, independent of the grid spacing. The solutions were found by an explicit, stable iteration scheme. Computational results using Dirichlet boundary values $\sin(6\pi(x-y))$ are presented in Figure 1. Note that the level sets fail to be straight lines for the second scheme.

Stability analysis. We verify stability in ℓ^2 directly. Consider for the sake of analysis a periodic, 2×2 grid. The grid functions

$$v_1 = \begin{pmatrix} + & + \\ + & + \end{pmatrix}, \qquad v_2 = \begin{pmatrix} + & - \\ + & - \end{pmatrix}, \qquad v_3 = \begin{pmatrix} + & + \\ - & - \end{pmatrix}, \qquad v_4 = \begin{pmatrix} + & - \\ - & + \end{pmatrix},$$

which consist of horizontal, vertical, and diagonal stripes, form a simultaneous set of eigenvectors for the schemes, with eigenvalues $\{0, -4, -4, 0\}$ and $\{0, -4, -4, -8\}$, respectively. Thus the operators are stable. The explicit scheme with time step dt corresponds to adding the identity to dt times the linear map, so it has the same eigenvectors, with eigenvalues $\lambda \mapsto 1 + dt\lambda$. Taking $dt \leq 1/2, 1/4$, respectively, gives a scheme with eigenvalues in the unit circle, and that is thus stable in ℓ^2.

Conclusion. Despite the stability of the second scheme, it is nonconvergent. For this equation, monotonicity is necessary for convergence. We offer a heuristic explanation: while the equation is sensitive to data only in the diagonal direction, the second scheme uses data from grid points in other directions.

2. Theory.

2.1. Viscosity solutions.

We have endeavored to make this article accessible to readers who are not familiar with the theory of viscosity solutions. The standard reference is [11]. An introduction to the first-order case, with applications to

control theory, is [15]. A readable introductory article with valuable exercises is the contribution by Crandall [3]. The complete first-order theory can be found in [2, 4].

Viscosity solutions are weak solutions defined for the class of degenerate elliptic and parabolic equations. In this class, under mild analytic assumptions [20], there exist unique viscosity solutions. These solutions are stable in the maximum norm under perturbations: a perturbation to the data of size ϵ results in an error in the solution of size at most ϵ. Solutions are also stable under perturbations of the *equation*, as long as the resulting equation is still in the class. For example, adding ϵ times the Laplacian regularizes the equation (this is where the term *viscosity solutions* comes from). It is also common to add ϵ time u to the equation. In addition, replacing the equation with a finite difference approximation is a valid perturbation, as long as the approximation is monotone [5]. In this case, ϵ may represent the grid spacing.

Remark 7 (nonsmooth functions). Although solutions need not be smooth (or even differentiable), the definition of viscosity solutions requires only verifying inequalities for smooth test functions. In particular, when verifying consistency for numerical schemes, we may work freely with smooth functions.

2.2. Degenerate elliptic equations. Let Ω be a domain in $\mathbb{R}^n$, Du and D^2u denote the gradient and Hessian of u, respectively, and $F(x, r, p, X)$ be a continuous real valued function defined on $\Omega \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{S}^n$, with $\mathbb{S}^n$ being the space of symmetric $n \times n$ matrices. Write $F[u](x) \equiv F(x, u(x), Du(x), D^2u(x))$. Consider the nonlinear, degenerate elliptic partial differential equation with Dirichlet boundary conditions,

$$\begin{cases} F[u](x) = 0 & \text{for } x \text{ in } \Omega, \\ u(x) = g(x) & \text{for } x \text{ on } \partial\Omega, \end{cases}$$

or the initial-boundary value problem for the degenerate parabolic partial differential equation,

$$\begin{cases} u_t(t,x) = -F[u](t,x) & \text{for } (t,x) \text{ in } \Omega \equiv [0,t) \times \Omega, \\ u(t,x) = g(t,x) & \text{for } (t,x) \text{ on } \partial\Omega \equiv \{t=0\} \times \Omega \cup [0,t) \times \partial\Omega. \end{cases}$$

In both cases, $\partial\Omega$ is the correct set on which boundary conditions are set for the equation, not the topological boundary.

Definition 1. *The equation F is* degenerate elliptic *if*

$$F(x, r, p, X) \le F(x, s, p, Y) \text{ whenever } r \le s \text{ and } Y \le X,$$

where $Y \le X$ means that $Y - X$ is a nonnegative definite symmetric matrix.

Example. The obstacle problem, $\min(-u_{xx}, u - g(x)) = 0$, is degenerate elliptic. The Hamilton–Jacobi equation, $u_t - |u_x| = 0$, is degenerate parabolic.

Given a degenerate elliptic equation, F, consider the solution mapping, S, which takes continuous boundary data, g, to the continuous solution, u, assuming it is well-defined. We say that S is *monotone* if for all continuous functions g, h on $\partial\Omega$,

$$g(x) \le h(x) \text{ for all } x \in \partial\Omega \text{ implies } S(g)(x) \le S(h)(x) \text{ for all } x \in \Omega. \tag{3}$$

Likewise, the solution mapping is nonexpansive in the maximum norm if

$$\max_{x\in\Omega} |S(g)(x) - S(h)(x)| \le \max_{x\in\partial\Omega} |g(x) - h(x)|. \tag{4}$$

These conditions generalize the maximum principle, with equivalence when constants (or zero) are solutions.

2.3. Degenerate elliptic schemes. We begin with the definition of a finite difference scheme on an unstructured grid. We regard a scheme as an *equation* which holds at each grid point, and thereby study monotonicity and stability properties of the solution operator.

For the purpose of convergence, we implicitly assume the existence of an interpolation operator, which takes grid functions to functions on the domain. We also require a sequence of grids indexed by a small parameter. Typically, the small parameter is dx, the maximum distance between neighboring grid points, but we might want to allow for $d\theta$, the directional resolution [29, 30]. The interpolation operator and the sequence of approximations puts us in the framework of the convergence theory in [5].

Define an unstructured grid on the domain Ω as a directed graph consisting of a set of points, $x_i \in \Omega, i = 1, \ldots, N$, each endowed with a list of neighbors, $N(i)$. A *grid function* is a real valued function defined on the grid, with values $u_i = u(x_i)$. The scheme is represented at each grid point by an equation of the form

$$F^i[u] \equiv F^i\left(u_i, \left.\frac{u_i - u_j}{|x_i - x_j|}\right|_{j=N(i)}\right), \quad i = 1, \ldots, N. \tag{5}$$

A finite difference scheme is local: it depends only on the value at the reference points, and on the first-order approximations to the derivatives in the direction of the neighbors. Higher-order approximations are obtained by taking linear combinations of the first-order derivatives.

From now on, we suppress the explicit dependence on $|x_i - x_j|$ and write

$$F^i[u] \equiv F^i\left(u_i, u_i - u_j|_{j=N(i)}\right) \equiv F^i(u_i, u_i - u_j),$$

where u_j is shorthand for the list of neighbors $u_j|_{j=N(i)}$.

A *boundary* point is a grid point with no neighbors. Dirichlet boundary conditions are imposed at boundary points by setting $F^i[u] = u_i - g(x_i)$. A *solution* is a grid function which satisfies $F[u] = 0$. If, for arbitrary boundary data g, there exists a unique solution u, we write $u = S(g)$ for the *solution operator*. We regard the solution operator as a mapping from the Dirichlet data on the boundary points to grid functions.

We now define degenerate elliptic schemes.

DEFINITION 2. *The scheme F is* degenerate elliptic *if each component F^i is nondecreasing in each variable.*

Remark 8. We emphasize that the scheme is a nondecreasing function of u_i and the differences $u_i - u_j$.

2.4. The nonlinear CFL condition. Write $\|x\|_\infty$ for the maximum norm, $\max_i |x_i|$.

While schemes may be nonlinear and nondifferentiable, we assume that they are globally Lipschitz continuous, with constant K. The resulting restriction on the time step is simply that $dt \leq K^{-1}$. We can also allow for schemes which are only locally Lipschitz continuous by allowing for the time step to depend on the data, $dt = K(u)^{-1}$. Higher-order time-stepping methods which still maintain monotonicity and nonexpansivity may also be used [17].

DEFINITION 3 (Lipschitz continuity). *The finite difference scheme, F, is Lipschitz continuous if there is a constant K such that for all $i = 1, \ldots, N$, $x, y \in \mathbb{R}^{|N(i)|+1}$,*

$$|F^i(x) - F^i(y)| \leq K\|x - y\|_\infty. \tag{6}$$

DEFINITION 4 (the explicit Euler map). *For $\rho > 0$, define $S_\rho : \mathbb{R}^N \to \mathbb{R}^N$ by*

$$S_\rho(u) = u - \rho F[u]. \tag{7}$$

This map is the explicit Euler discretization, with time step ρ, of the ordinary differential equation $du/dt + F[u] = 0$.

DEFINITION 5 (nonlinear CFL condition). *Let F be a Lipschitz continuous, degenerate elliptic scheme, with Lipschitz constant K. The nonlinear CFL condition for the Euler map S_ρ is*

$$\rho \leq \frac{1}{K}. \tag{CFL}$$

2.5. Existence, uniqueness, and comparison for schemes. Given $u, v \in \mathbb{R}^N$, define $u \vee v = \max(u, v), u^+ = \max(u, 0)$, $u^- = \min(u, 0)$, componentwise and define $u \leq v$ to mean $u_i \leq v_i$ for $i = 1, \ldots, N$

DEFINITION 6 (proper schemes). *The finite difference scheme is* proper *if there exists $\delta > 0$ such that for $i = 1, \ldots, N$ and for all $x \in \mathbb{R}^{|N(i)|}$ and $x_0, y_0 \in \mathbb{R}$,*

$$x_0 \leq y_0 \text{ implies that } F^i(x_0, x) - F^i(y_0, x) \leq \delta(x_0 - y_0). \tag{8}$$

Remark 9. If a scheme is not proper, we can consider instead $F[u] + \epsilon u$. By taking ϵ to be small enough (for example, smaller than the discretization error), we can assume the scheme is proper without any loss of generality.

Remark 10. This property is introduced to simplify the existence proof. It can be relaxed for the proof of comparison. An alternative approach would be to generalize the "marching to the boundary" argument of [28].

THEOREM 5 (comparison of sub- and supersolutions). *Let F be a proper, degenerate elliptic finite difference scheme. If $F[u] \leq F[v]$, then $u \leq v$. In particular, solutions are unique.*

Proof. Suppose $u \not\leq v$ and let i be an index for which

$$u_i - v_i = \max_{j=1,\ldots,N} \{u_j - v_j\} > 0, \tag{i}$$

so that

$$u_i - u_j \geq v_i - v_j, \quad j = 1, \ldots, N. \tag{ii}$$

(See Figure 2.) Then we obtain a contradiction as follows:

$$\begin{aligned} F[u]^i = F^i(u_i, u_i - u_j) &\geq F^i(u_i, v_i - v_j) && \text{by (ii) and Definition 2,} \\ &> F^i(v_i, v_i - v_j) = F[v]^i && \text{by (i) and (8).} \end{aligned}$$

Uniqueness follows, since if u, v are solutions, then $F[u] = F[v] = 0$, so $u \geq v$ and $u \leq v$, and thus $u = v$. ∎

The next result combines the Lipschitz continuity property with the degenerate elliptic property of the scheme to give an ordered Lipschitz continuity property.

LEMMA 1 (ordered Lipschitz continuity property). *Let F be a Lipschitz continuous, degenerate elliptic scheme, with Lipschitz constant K. Then for all $i = 1, \ldots, N$ and $x, y \in \mathbb{R}^{|N(i)|+1}$,*

$$-K\|(x - y)^-\|_\infty \leq F^i(x) - F^i(y) \leq K\|(x - y)^+\|_\infty. \tag{9}$$

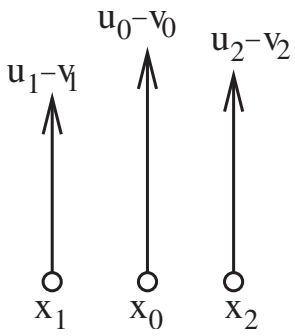


FIG. 2. *Discrete local maximum of* $u_i - v_i$ *at* $i = 0$.

Proof. Given x, y, use Definition 2 and (6) to compute

$$F(x) - F(y) \le F(x \vee y) - F(y) \le K\|x \vee y - y\|_\infty = K\|(x-y)^+\|_\infty.$$

The other inequality is similar. ∎

THEOREM 6 (the Euler map is monotone). *Let F be a Lipschitz continuous, degenerate elliptic scheme. Then the Euler map* (7) *is monotone provided (CFL) holds.*

Proof. Suppose $u \le v$. Compute for an arbitrary index i,

$$\begin{aligned} S^i_\rho(u) - S^i_\rho(v) &= u_i - v_i + \rho\left(F^i(v_i, v_i - v_j) - F^i(u_i, u_i - u_j)\right) \\ &\le u_i - v_i + \rho K\|(v_i - u_i, v_i - u_i + u_j - v_j)^+\|_\infty \quad \text{by (9)} \\ &\le (1 - \rho K)(u_i - v_i) \quad \text{since } u \le v \\ &\le 0 \quad \text{by (CFL).} \end{aligned}$$ ∎

THEOREM 7 (the Euler map is a contraction). *Let F be a Lipschitz continuous, degenerate elliptic scheme. Then the Euler map* (7) *is a contraction in $\mathbb{R}^N$ equipped with the maximum norm, provided* (CFL) *holds. If, in addition, F is proper, and strict inequality holds in* (CFL)*, then the Euler map is a strict contraction.*

Proof. We will show that

$$\|S_\rho(u) - S_\rho(v)\|_\infty \le r\|u - v\|_\infty \tag{i}$$

for $r = \max(1 - \rho\delta, \rho K)$. We assume without loss of generality that $\rho\delta, \rho K < 1/2$.

We proceed to find upper and lower bounds on $S^k_\rho(u) - S^k_\rho(v)$ for an arbitrary index k. The lower bound will follow easily, while the upper bound will rely on careful application of the ordered Lipschitz continuity property.

1. Assume $u_k \ge v_k$. The alternative will follow by a similar argument.
2. For the lower bound, use (9) in the definition of the Euler map (7) to obtain

$$\begin{aligned} S^k_\rho(u) - S^k_\rho(v) &\ge u_k - v_k - \rho K\| \left(u_k - v_k, u_k - v_k - (u_j - v_j)\right)^- \|_\infty \\ &\ge -\rho K\|u - v\|_\infty, \end{aligned} \tag{ii}$$

since $u_k \ge v_k$.

3. For the upper bound, add and subtract $\rho F^k(v_k, u_k - u_j)$ to $S^k_\rho(u) - S^k_\rho(v)$,

$$\begin{aligned} S^k_\rho(u) - S^k_\rho(v) = u_k - v_k - \rho\left(F^k(u_k, u_k - u_j) - F^k(v_k, u_k - u_j)\right) \\ + \rho\left(F^k(v_k, v_k - v_j) - F^k(v_k, u_k - u_j)\right). \end{aligned} \tag{iii}$$

Use (8) to estimate the second to last term in (iii),

$$-\left(F^k\left(u_k, u_k - u_j\right) - F^k(v_k, u_k - u_j)\right) \leq -\delta(u_k - v_k). \tag{iv}$$

Next use (9) to estimate the last term in (iii),

$$\begin{aligned} F^k\left(v_k, v_k - v_j\right) - F^k(v_k, u_k - u_j) &\leq K\|((u_j - v_j) - (u_k - v_k))^+\|_\infty, \\ &\leq K\left(\|u - v\|_\infty - (u_k - v_k)\right), \end{aligned} \tag{v}$$

since $u_k \geq v_k$. Combining (iv) and (v) gives

$$\begin{aligned} S^k_\rho(u) - S^k_\rho(v) &\leq (1 - \rho\delta - \rho K)(u_k - v_k) + \rho K\|u - v\|_\infty \\ &\leq (1 - \rho\delta - \rho K)\|u - v\|_\infty + \rho K\|u - v\|_\infty \\ &\leq (1 - \rho\delta)\|u - v\|_\infty. \end{aligned} \tag{vi}$$

4. Combining (ii) and (vi) gives (i) as desired. ∎

THEOREM 8. *A proper, Lipschitz continuous degenerate elliptic scheme has a unique solution. The iterates of the Euler map converge to the solution for arbitrary initial data, provided strict inequality holds in* (CFL).

Proof. By Theorem 7, S_ρ is a strict contraction on $\mathbb{R}^N$, equipped with the maximum norm. Thus by Banach's fixed point theorem, iterates of S_ρ converges to a unique fixed point from arbitrary initial data. Such a fixed point is a solution. ∎

Remark 11. Since the error tolerance is on the order of the spatial discretization error, the number of iterations need not be prohibitive. Experimentally, the number of iterations is on the order of the diameter of the graph, when the time step is optimal.

2.6. Proofs. We begin by establishing a link between the degenerate ellipticity condition and the comparison principle.

LEMMA 2 (exercise in [3]). *The function $F(x, r, p, X)$ is degenerate elliptic if and only if whenever x is a nonnegative local maximum of $u - v$, for $u, v \in C^2$, $F[u](x) \geq F[v](x)$.*

Proof. If x is a local maximum, $u \geq v$, $Dv = Du$, and $D^2u \leq D^2v$, at x. Then $F(x, u, Du, D^2u) = F(x, u, Dv, D^2u) \geq F(x, v, Dv, D^2u) \geq F(x, v, Dv, D^2v)$. ∎

LEMMA 3. *The scheme F is degenerate elliptic if and only if whenever x_i is a nonnegative maximum of $u - v$, for u, v grid functions, $F^i[u] \geq F^i[v]$.*

Proof. Let i be an index for which $u_i - v_i = \max_{j=1,\ldots,N}\{u_j - v_j\} \geq 0$, so that $u_i - u_j \geq v_i - v_j$, $j = 1, \ldots, N$. Then $F^i[u] = F^i\left(u_i, u_i - u_j\right) \geq F^i\left(v_i, u_i - u_j\right) \geq F^i\left(v_i, v_i - v_j\right) = F^i[v]$. ∎

Proof of Theorem 1. The proof is formal, but can be made rigorous. Let $\epsilon > 0$, set $F^\epsilon[u] = F[u] + \epsilon u$, and let S^ϵ be the corresponding solution operator. Let $u^\epsilon = S^\epsilon(g), v^\epsilon = S^\epsilon(h)$. If x is a strict local max of $u^\epsilon - v^\epsilon$, then as in Lemma 2, $F^\epsilon[u](x) > F^\epsilon[v](x)$, which contradicts $F^\epsilon[u] = F^\epsilon[v] = 0$. So the maximum of $u^\epsilon - v^\epsilon$ occurs on the boundary. Likewise, the minimum of $u^\epsilon - v^\epsilon$ occurs on the boundary. Stability of viscosity solutions implies that $u^\epsilon \to u, v^\epsilon \to v$, and thus sending $\epsilon \to 0$ allows the same conclusions to hold for u, v. Thus (4) follows; assuming $g \leq h$ gives (3). ∎

Proof of Theorem 2. We can assume without loss of generality that F is proper. Then as in Lemma 3, we can show that the max and min of $S(g), S(h)$ occur on the boundary. The conclusion follows as in the proof of Theorem 1. □

Proof of Theorem 3. We have already shown that degenerate elliptic schemes are monotone and nonexpansive. Now, suppose a scheme given in explicit form (1) is monotone and nonexpansive. Then H^i is a nondecreasing function for each i. Nonexpansivity means

$$|H^i(x) - H^i(y)| \le \|x - y\|_\infty$$

for all x, y. Estimate $|H^i(x) - H^i(y)| \le \|DH^i\|_1 \|x - y\|_\infty$. Since equality may hold for some x, y, we require $\|DH^i\|_1 \le 1$. Locally define $F^i = (1 - \sum_{j=N(i)} D_j H^i)u_i + \sum_{j=N(i)} D_j H^i(u_i - u_j)$. Applying the implicit function theorem gives F^i in the required form. □

Remark 12. The linear scheme $u^{n+1} = Mu^n$ is monotone if and only if $m_{ij} \ge 0$ and nonexpansive in the ℓ^∞ norm if and only if $\sum_j |m_{ij}| \le 1$ for each i. The differentiable scheme, $u^{n+1} = F(u^n)$, is monotone (respectively, nonexpansive) if the gradient $DF(u)$ is monotone (nonexpansive) for every u. In the linear case, nonexpansivity in ℓ^∞ does not imply nonexpansivity in ℓ^2, or in ℓ^1, as simple examples illustrate.

Accuracy. Given the equation $F(x, u(x), Du(x), D^2u(x))$ and the scheme $F^i(u_i, u_i - u_j|_{j=N(i)})$, fix $x = x_i$ and set $h_j = |x_i - x_j|$, $j = N(i)$. Assume the h_j are of the same order so that the expression $O(h)$ is meaningful. The order of accuracy of the scheme is the best possible number r in the expression

$$F(x_i, u(x_i), Du(x_i), D^2u(x_i)) - F^i(u(x_i), u(x_i) - u(x_j)|_{j=N(i)}) = O(h^r),$$

over all functions u which have all derivatives defined in a neighborhood of x_i.

Proof of Theorem 4. It is sufficient to show that higher-order accuracy is impossible for functions of a particular form. By considering functions of the form $u(x) = g(n \cdot x)$, where n is a direction vector, we reduce to an equation in one space dimension. Considering functions with constant values $u(x_i)$ and constant first or second derivatives further reduces the equation to the form $H(u_x)$ or $H(u_{xx})$. While some reductions yield trivial equations, any nontrivial equation will give a nontrivial reduction for some choice of the direction n, $u(x_i)$, and the derivatives.

Redefine $h_j = x_j - x_i$, and expand u in Taylor series, $u(x_j) = \sum_{k=0}^\infty \frac{h_j^k}{k!} \frac{\partial^k u}{\partial x^k}(x_i)$. Apply the series to each u_j in the expression for $F^i[u]$, dropping the u_i dependence to give

$$\text{(i)} \qquad F^i[u] = F^i\left(- \sum_{k=1}^\infty \frac{h_j^k}{k!} \frac{\partial^k u}{\partial x^k}(x_i) \Big|_{j=N(i)} \right).$$

Consider first the case when both the scheme and the equation are linear. Write $F^i[u] = \sum_{j=1}^{|N(i)|} a_j(u_i - u_j)$, $a_j > 0$, and insert the Taylor expansion into (i) to obtain $-\sum_{j=1}^{|N(i)|} \sum_{k=1}^\infty \frac{a_j h_j^k}{k!} \frac{\partial^k u}{\partial x^k}(x_i)$. Observe that the coefficients of $\partial^k u/\partial x^k$ have the same sign for even values of k, and thus no cancellation is possible. In addition, the coefficients are homogeneous of order k in h. After dividing by the leading coefficient, a scheme for u_x will have a nonzero coefficient of u_{xx} of $O(h)$, and a scheme for u_{xx} will have a nonzero coefficient of u_{xxxx} of $O(h^2)$.

We now treat the general case. Since F is nondecreasing, there is still no cancellation among the terms with even values of k. An expression containing u_x to $O(1)$ appears with a u_{xx} term of $O(h)$, and similarly u_{xx} appears with a u_{xxxx} term of $O(h^2)$. Since the higher derivatives do not appear at all in the expression for H, the error is $O(h)$ and $O(h^2)$ for the first-order and second-order equations, respectively. □

3. Building elliptic schemes. In this section we construct examples of degenerate elliptic schemes. Using simple schemes as building blocks, we build schemes for nontrivial equations. We begin with parallel observations for equations and schemes which, when taken together, give a technique for building schemes.

Many types of operations (addition, min, max, nondecreasing transformations) may be used to combine schemes. On the other hand, selection criteria ("if" statements) generally do not preserve ordering properties and must be used with care.

Observation 1 (see [11, p. 8]). Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a nondecreasing function. If F_1 and F_2 are degenerate elliptic functions, then so is $F = g(F_1, F_2)$.

Observation 2. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a nondecreasing function. If F_1 and F_2 are degenerate elliptic finite difference schemes, then so is $F = g(F_1, F_2)$.

Example (order preserving operations). The constant scheme $F[u] = u - g$ is degenerate elliptic. If F, F_1, F_2 are degenerate elliptic, then so are $F^+ = \max(F, 0)$, and $F^- = \min(F, 0)$ as well as $\min(F_1, F_2), \max(F_1, F_2)$, and $aF_1 + bF_2$ for $a, b \in \mathbb{R}^N, a, b \geq 0$.

Example ("if" statements). If $F_1[u], F_2[u]$ are degenerate elliptic, and $G[u]$ is an equation, then

$$F[u] = \begin{cases} F_2[u] = 0 & \text{if } G[u] > 0, \\ F_1[u] = 0 & \text{otherwise} \end{cases}$$

is not usually degenerate elliptic. If, however, G is degenerate elliptic, and $F_2[u] \geq F_1[u]$ for all u, then F is degenerate elliptic. This follows by using Lemma 3. If $u - v$ has a nonnegative local max at i, then $G^i[u] \geq G^i[v]$, and thus $F^i[u] \geq F^i[v]$.

Example (distance function and eikonal equation). Starting from the upwind schemes $u_x = (u_j - u_{j-1})/dx$ and $-u_x = (u_j - u_{j+1})/dx$, which are degenerate elliptic, write $|u_x| = \max(u_x, -u_x)$, $-|u_x| = \min(u_x, -u_x)$ and apply the observations to build the schemes

$$|u_x| = \frac{1}{dx} \max\left(u_j - u_{j-1}, u_j - u_{j+1}, 0\right), \quad -|u_x| = \frac{1}{dx} \min\left(u_j - u_{j-1}, u_j - u_{j+1}, 0\right),$$

accurate to $O(dx)$. Next write $u_x^2 = |u_x|^2$ to obtain the scheme

$$u_x^2 = \frac{1}{dx^2} \max\left(u_j - u_{j-1}, u_j - u_{j+1}, 0\right)^2,$$

which is accurate to $O(dx)$. Schemes for $|Du|$ and $|Du|^p$ in higher dimensions are easily built.

Example (obstacle problems). Let F_1 be a degenerate elliptic scheme for $F[u]$. The obstacle problem

$$\min(F[u], u - g(x)) = 0$$

is degenerate elliptic, and the scheme $\min(F_1, u - g)$, is consistent and degenerate elliptic. This example can be generalized to double obstacle problems.

Example (finite maxima and minima of schemes [5]). A degenerate elliptic scheme for $\min_{\gamma\in\Gamma}\max_{\beta\in B}\{F_{\gamma\beta}\}$, where the index sets are finite, can be built from schemes for each $F_{\gamma\beta}$ by taking the corresponding finite minima and maxima over the schemes.

Example (nonlinear one-dimensional equations). The degenerate elliptic equation $F(x,u_{xx})$ is nonincreasing in u_{xx}, and thus the scheme $F^i[u]=F(x_i,(2u_i-u_{i+1}-u_{i-1})/dx^2)$ is degenerate elliptic. If $H(x,u_x)$ is increasing in u_x, then the scheme $F^i[u]=H(x_i,(u_i-u_{i-1})/dx)$ is also degenerate elliptic. Likewise, if $H(x,u_x)$ is decreasing in u_x, then $F^i[u]=H(x_i,(u_i-u_{i+1})/dx)$ is degenerate elliptic. Simply combining the previous two schemes with an if statement will not yield a degenerate elliptic scheme for general H.

Example (one-phase free boundary problems). Consider

$$\begin{cases} F[u]=0 & \text{in } \{u>0\},\\ H(x,Du)=0 & \text{on } \partial\{u=0\}, \end{cases}$$

where $F[u]=F(x,Du,D^2u)$ is degenerate elliptic. Time-dependent versions may also be considered. This one-phase free boundary problem (see [8, 22] for examples) is degenerate elliptic when the boundary condition is interpreted in the viscosity sense,

$$\min(F,H)\le 0 \text{ and } \max(F,H)\ge 0 \quad \text{on } \partial\{u=0\}.$$

Eliminate the free boundary by extending to a computational domain large enough to contain $\{u=0\}$, and consider instead

$$\begin{cases} F[u]=0 & \text{in } \{u>0\},\\ \min(F[u],H(x,Du))=0 & \text{on } \{u\le 0\}. \end{cases}$$

Given F_1,F_2, degenerate elliptic schemes for F and H, respectively, the following scheme is consistent and degenerate elliptic:

$$F^i[u]=\begin{cases} F_1^i[u] & \text{if } u_i>0,\\ \min(F_1^i[u],F_2^i[u]) & \text{if } u_i\le 0. \end{cases}$$

4. Computations.

Example (front propagation). For $u_t=|u_x|$, (CFL) gives $\rho\le dx$. Setting $\rho=dx$ gives the exact solution for piecewise linear initial data. For $|u_x|^2$, the function $F(x,y,0)=\max(x,y,0)^2$ is locally, but not globally, Lipschitz, with constant $K=\max(x,y,0)$. (Simply differentiating overestimates the constant by a factor of 2.) This leads to $dt\le dx^2/\max_j\{|u_j-u_{j-1}|\}$. The solution was computed using sinusoidal initial data and periodic boundary conditions, with 500 grid points. The solutions are displayed in Figure 3.

Homogenization of Hamilton–Jacobi equations. Next we consider one- and two-dimensional nonconvex Hamilton–Jacobi homogenization problems. The problem involves solve $H(x,Du)=\overline{H}$, for the function u and the constant $\overline{H}$, using periodic boundary conditions. The value $\overline{H}$ is unique, although the function u is not. The solution can be obtained by solving the time-dependent problem $u_t=H(x,Du)$ for a long time, because (see [6]) $u_t\to\overline{H}, u(t,\cdot)\to u+\overline{H}t$ as $t\to\infty$.

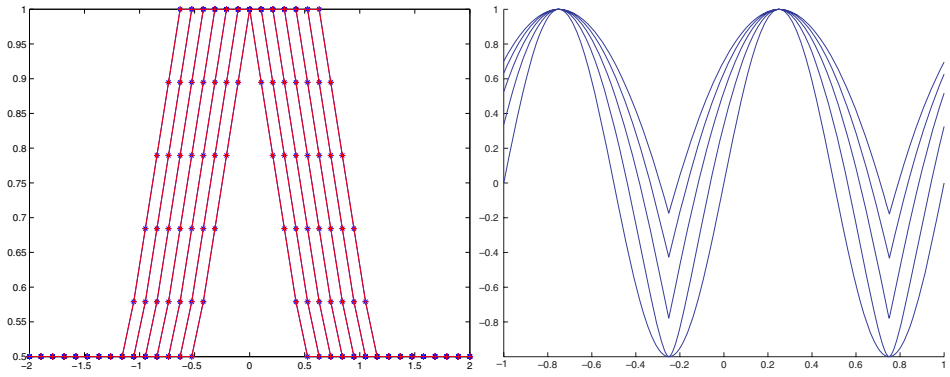


FIG. 3. *Snapshots of the solution of* $u_t = |u_x|$, *and* $u_t = |u_x|^2$.

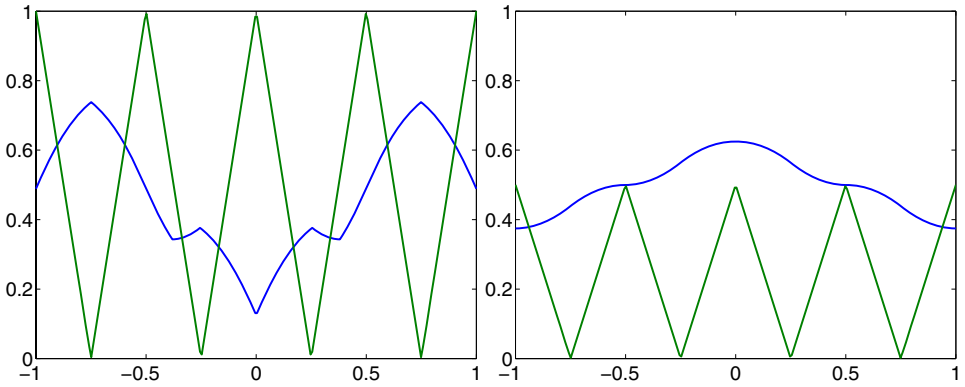


FIG. 4. *Solutions* u *plotted with the zig-zag functions.*

Example (one-dimensional nonconvex Hamilton–Jacobi equation). Set $H(x, u_x) = \max(|u_x|, 1 - |u_x|) + f(x)$, where $f(x)$ is periodic on $[-1, 1]$. With $f(x)$ the "zigzag" function $f(x) = 2|2x \pmod 1 - .5|$, we found $\overline{H} \approx .48$. The solutions were also computed with $f(x) = |2x \pmod 1 - .5|$. The solutions, along with the functions $f(x)$, are displayed in Figure 4, plotted so that the average of the solution is $\overline{H}$. In the second case, with the hindsight afforded by the numerical solution, we found an exact piecewise quadratic solution, with $\overline{H}(u) = .5$. Modifying the solution by reflecting the portion between $-.5$ and $.5$ (where $u_x = 0$) in the line $y = .5$ gives another solution with $\overline{H} = .5$.

Example (two-dimensional nonconvex Hamilton–Jacobi equation). Set $H(u_x, u_y) = u_x^2 - u_y^2$ in a periodic domain, with periodic boundary conditions. With initial data $\sin(x)\sin(y)$, a nontrivial steady solution was computed, shown in Figure 5. Examination of the numerical solution reveals an exact solution, which is piecewise quadratic, with u_x, u_y as piecewise linear functions with slopes ± 1.

Free boundary problems.

Example (two-dimensional double obstacle problem). We have the equation

$$-\max\left(u - h, \min\left(u_{xx} + u_{yy}, u - g\right)\right) = 0,$$

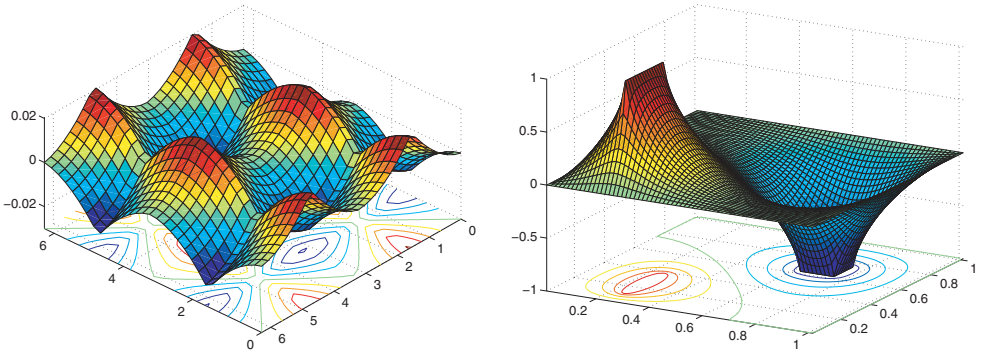


FIG. 5. *Piecewise quadratic solution of* $u_x^2 - u_y^2 = 0$*, solution of the double obstacle problem.*

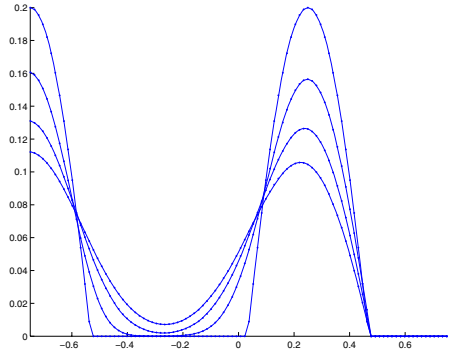


FIG. 6. *Snapshots in time of the solution to the Stefan problem. Two bumps move together and merge.*

where the obstacle functions h, g are characteristic functions of a square and a line on different parts of the domain. The solution is displayed in Figure 5.

Example. The one-phase Stefan problem

$$\begin{cases} u_t - \Delta u = 0 & \text{in } \{u > 0\}, \\ u_t - |Du|^2 = 0 & \text{on } \partial\{u = 0\} \end{cases}$$

is solved in one dimension with sinusoidal initial data. Snapshots of the solution are shown in Figure 6.

Example (a nonconvex, fully nonlinear second-order equation). The fully nonlinear, uniformly elliptic second-order equation $-\max(\min(L^1u, L^2u)L^3u) + 1 = 0$, where $L^1u = u_{xx} + u_{yy}$, $L^2u = \frac{1}{2}u_{xx} + 2u_{yy}$, $L^3u = \frac{1}{2}u_{xx} + u_{yy}$, is solved in the unit square with Dirichlet boundary values $\frac{1}{2}\max(\min(x^2+y^2, \frac{1}{2}x^2+2y^2), \frac{1}{2}x^2+y^2)$. The solution is displayed in Figure 7. The boundary shown is composed of two parts: the dotted lines correspond to the boundary of the set $\{L^1u \le L^2u\}$. The heavy lines correspond to the boundary of the set $L^3u \ge \min(L^1u, L^2u)$.

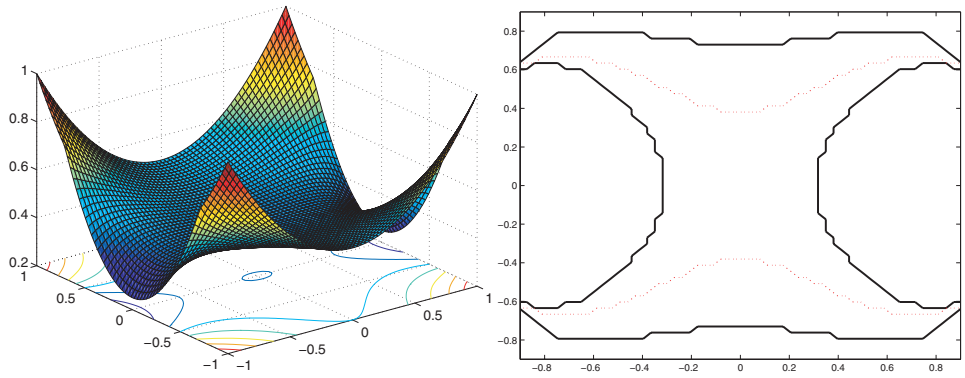


FIG. 7. *Solution and free boundary for the fully nonlinear second-order equation* $-\max(\min(L^1u, L^2u)L^3u) + 1 = 0$.

Acknowledgments. It is a pleasure to acknowledge many valuable discussions with P. E. Souganidis. I am grateful to Ward E. Cheney for writing advice, and to the University of Texas at Austin for its hospitality during the course of this work.

REFERENCES

[1] D. G. ARONSON, *The stability of finite difference approximations to second order linear parabolic differential equations*, Duke Math. J., 30 (1963), pp. 117–127.

[2] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[3] M. BARDI, M. G. CRANDALL, L. C. EVANS, H. M. SONER, AND E. P. SOUGANIDIS, *Viscosity Solutions and Applications*, Lecture Notes in Math. 1660, Springer-Verlag, Berlin, 1997.

[4] G. BARLES, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[5] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[6] G. BARLES AND P. E. SOUGANIDIS, *On the large time behavior of solutions of Hamilton-Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.

[7] J. H. BRAMBLE, B. E. HUBBARD, AND V. THOMÉE, *Convergence estimates for essentially positive type discrete Dirichlet problems*, Math. Comp., 23 (1969), pp. 695–709.

[8] L. A. CAFFARELLI AND J. L. VAZQUEZ, *Viscosity solutions for the porous medium equation*, in Differential Equations: La Pietra 1996 (Florence), Proc. Sympos. Pure Math. 65, AMS, Providence, RI, 1999, pp. 13–26.

[9] R. COURANT, K. O. FRIEDRICHS, AND H. LEWY, *On the partial difference equations of mathematical physics*, IBM J. Res. Develop., 11 (1967), pp. 215–234.

[10] M. G. CRANDALL AND P.-L. LIONS, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[11] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[12] M. G. CRANDALL AND P.-L. LIONS, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–41.

[13] M. G. CRANDALL AND A. MAJDA, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[14] M. G. CRANDALL AND L. TARTAR, *Some relations between nonexpansive and order preserving mappings*, Proc. Amer. Math. Soc., 78 (1980), pp. 385–390.

[15] L. C. EVANS, *Partial Differential Equations*, Graduate Studies in Math. 19, AMS, Providence, RI, 1998.

[16] M. FALCONE AND R. FERRETTI, *Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods*, J. Comput. Phys., 175 (2002), pp. 559–575.

[17] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
[18] A. Harten, J. M. Hyman, and P. D. Lax, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297–322.
[19] A. Harten and S. Osher, *Uniformly high-order accurate nonoscillatory schemes*, I, SIAM J. Numer. Anal., 24 (1987), pp. 279–309.
[20] R. Jensen, P.-L. Lions, and P. E. Souganidis, *A uniqueness result for viscosity solutions of second order fully nonlinear partial differential equations*, Proc. Amer. Math. Soc., 102 (1988), pp. 975–978.
[21] F. John, *On integration of parabolic equations by difference methods*, I. *Linear and quasi-linear equations for the infinite interval*, Comm. Pure Appl. Math., 5 (1952), pp. 155–211.
[22] I. C. Kim, *Uniqueness and existence results on the Hele-Shaw and the Stefan problems*, Arch. Ration. Mech. Anal., 168 (2003), pp. 299–328.
[23] H. J. Kuo and N. S. Trudinger, *On the discrete maximum principle for parabolic difference operators*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 719–737.
[24] H. J. Kuo and N. S. Trudinger, *Discrete methods for fully nonlinear elliptic equations*, SIAM J. Numer. Anal., 29 (1992), pp. 123–135.
[25] H.-J. Kuo and N. S. Trudinger, *Maximum principles for difference operators*, in Partial Differential Equations and Applications, Lecture Notes in Pure and Appl. Math. 177, Dekker, New York, 1996, pp. 209–219.
[26] H.-J. Kuo and N. S. Trudinger, *Positive difference operators on general meshes*, Duke Math. J., 83 (1996), pp. 415–433.
[27] C.-T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.
[28] T. S. Motzkin and W. Wasow, *On the approximation of linear elliptic differential equations by difference equations with positive coefficients*, J. Math. Phys., 31 (1953), pp. 253–259.
[29] A. M. Oberman, *A convergent monotone difference scheme for motion of level sets by mean curvature*, Numer. Math., 99 (2004), pp. 365–379.
[30] A. M. Oberman, *A convergent difference scheme for the infinity Laplacian: Construction of absolutely minimizing Lipschitz extensions*, Math. Comp., 74 (2005), pp. 1217–1230.
[31] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
[32] J. A. Sethian, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.
[33] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.
[34] Y.-H. R. Tsai, L.-T. Cheng, S. Osher, and H.-K. Zhao, *Fast sweeping algorithms for a class of Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.
[35] O. B. Widlund, *Stability of parabolic difference schemes in the maximum norm*, Numer. Math., 8 (1966), pp. 186–202.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 897–907

NUMERICAL CUBATURE USING ERROR-CORRECTING CODES*

GREG KUPERBERG†

Dedicated to Włodzimierz and Krystyna Kuperberg on the occasion of their 40th anniversary

Abstract. We present a construction for improving numerical cubature formulas with equal weights and a convolution structure, in particular equal-weight product formulas, using linear error-correcting codes. The construction is most effective in low degree with extended BCH codes. Using it, we obtain several sequences of explicit, positive, interior cubature formulas with good asymptotics for each fixed degree t as the dimension $n \to \infty$. Using a special quadrature formula for the interval [G. Kuperberg, *Adv. in Appl. Math.*, 34 (2005), pp. 853–870, arXiv:math.PR/0408360], we obtain an equal-weight t-cubature formula on the n-cube with $O(n^{\lfloor t/2 \rfloor})$ points, which is within a constant of the Stroud lower bound. We also obtain t-cubature formulas on the n-sphere, n-ball, and Gaussian $\mathbb{R}^n$ with $O(n^{t-2})$ points when t is odd. When μ is spherically symmetric and $t = 5$, we obtain $O(n^2)$ points. For each $t \geq 4$, we also obtain explicit, positive, interior formulas for the n-simplex with $O(n^{t-1})$ points; for $t = 3$, we obtain $O(n)$ points. These constructions asymptotically improve the nonconstructive Tchakaloff bound.

Some related results were recently found independently by Victoir [*SIAM J. Numer. Anal.*, 42 (2004), pp. 209–227], who also noted that the basic construction more directly uses orthogonal arrays.

Key words. cubature formulas, orthogonal arrays, error-correcting codes

AMS subject classifications. 65D32, 05B15, 94B05

DOI. 10.1137/040615572

1. General results.

Let μ be a normalized measure on $\mathbb{R}^n$ with finite moments. A *cubature formula of degree* t, or t-cubature formula, for μ is a set of points $F = \{\vec{p}_a\} \subset \mathbb{R}^n$ and a weight function $\vec{p}_a \mapsto w_a \in \mathbb{R}$ such that

$$\int P(\vec{x}) d\mu = P(F) = \sum_{a=1}^{N} w_a P(\vec{p}_a)$$

for polynomials P of degree at most t. (If $n = 1$, then F is also called a *quadrature formula*.) The formula F is *equal-weight* if the w_a are all equal; it is *positive* if $w_a > 0$ for all a; and otherwise it is *negative*. Let X be the support of μ. The formula F is *interior* if every point $\vec{p}_a$ is in the interior of X; it is *boundary* if every $\vec{p}_a$ is in X and some $\vec{p}_a \in \partial X$; and otherwise it is *exterior*. These properties of cubature formulas are often abbreviated. For example, PI means positive and interior and EB means equal-weight and boundary. (Exterior formulas are denoted "O," for outside.) An equal-weight formula is abbreviated "E" and is also called a (geometric) *t-design* or a *Chebyshev-type formula*.

The main use of a cubature formula is to numerically integrate a function f which is approximately a polynomial. In this application, formulas with many points or nonexplicit points are impractical; exterior formulas are ill-founded if f is defined only on X; and formulas with large negative weights are ill-conditioned on the class

*Received by the editors September 23, 2004; accepted for publication (in revised form) November 22, 2005; published electronically May 5, 2006. This material is based upon work supported by the National Science Foundation under grant 0306681.

http://www.siam.org/journals/sinum/44-3/61557.html

†Department of Mathematics, University of California, Davis, CA 95616 (greg@math.ucdavis.edu).

of continuous functions [20, Chap. 1]. Thus PI formulas with few points are the best kind.

By Tchakaloff's theorem [20, p. 61], every measure μ on $\mathbb{R}^n$ has a PI t-cubature formula with at most $\binom{n+t}{t}$ points, the same as the dimension of the vector space of relevant polynomials, $\mathbb{R}[\vec{x}]_{\le t}$. (If ∂X has nonzero measure, it may be only a PB formula.) Tchakaloff's theorem has a short proof, but it is computationally nonconstructive. Many known formulas with n small, or with n large and $t \le 2$, are better than the Tchakaloff bound [20, 4]. But if n is large, $t \ge 3$, and μ is reasonably natural, most explicit formulas in the existing literature either are negative, are exterior, or have exponentially many points.

In this article we present a new method to thin equal-weight cubature formulas with a convolution structure, in particular product formulas for product measures. (By *thinning* a formula, we mean removing some of its points without reducing its cubature degree.) The thinned formulas are efficient in high dimensions and low degree. The method also applies to some nonproduct measures that are related to product measures, in particular spheres and simplices with uniform measure. Victoir [21] independently obtained the basic construction when $q = 2$ (where q is the prime power parameter in Theorem 1.1), together with some other generalizations not considered by this author. However, many of our asymptotic bounds and derived constructions are new.

If F and G are two cubature formulas, we define their *convolution* $F * G$ to be their sum as sets, $F + G$. The weight w_a of $\vec{p}_a$ in $F * G$ is given by a product rule:

$$w_a = \sum_{\substack{\vec{p}_a = \vec{p}_b + \vec{p}_c \\ \vec{p}_b \in F, \vec{p}_c \in G}} w_b w_c.$$

Convolution of cubature formulas is related to convolution of measures in two ways: First, it is convolution of measures if cubature formulas are interpreted as atomic measures. Second, if F is a t-cubature formula for μ and G is a t-cubature formula for ν, then $F * G$ is a t-cubature formula for $\mu * \nu$. In particular, product formulas and product measures are convolutions in independent directions.

We also recall some basic facts from coding theory. For each prime power q, there is a unique finite field $\mathbb{F}_q$ with q elements. A *linear error-correcting code* of length ℓ, dimension k, and distance t over $\mathbb{F}_q$ is a k-dimensional vector subspace of $\mathbb{F}_q^\ell$ such that each nonzero vector has at least t nonzero coordinates. It is also called an $[\ell, k, t]_q$ code. A code C is a *zero-sum code* if the coordinates of every $\vec{a} \in C$ sum to 0.

THEOREM 1.1. *Let t, n, and ℓ be positive integers, let q be a prime power, and let μ be a measure on $\mathbb{R}^n$. For each $1 \le i \le \ell$, let F_i be an equal-weight formula with q elements such that the convolution*

$$F = F_1 * F_2 * \cdots * F_\ell$$

is a t-cubature formula for μ. Then an $[\ell, k, t+1]_q$ code C yields a thinning $G \subset F$ with $q^{\ell-k}$ points. In addition, if each F_i is centrally symmetric, t is odd, and either q is odd or C is a zero-sum code, then C need only be an $[\ell, k, t]_q$ code.

Theorem 1.1 can be strengthened further using the notion of an orthogonal array [9]. Linear error-correcting codes are dual to linear orthogonal arrays, and the proof actually uses orthogonal arrays rather than codes. In some cases nonlinear orthogonal arrays are slightly better than linear ones. See sections 2 and 4.

The most effective case of Theorem 1.1 is in the asymptotic limit $n \to \infty$ with t and q fixed. Recall that a function $f(n)$ is *quasi-linear* if $f(n) = O((\log n)^\alpha n)$ for some

α. Quasi linearity is also written $f(n) = \widetilde{O}(n)$. Say that a family $\{F\}$ of cubature formulas is quasi-linear (abbreviated "QL") if the points and weights of each F can be generated in quasi-linear time in the length of the output.

THEOREM 1.2. *Assume all variables as in Theorem* 1.1. *Then G can have $O(\ell^\alpha)$ points (with the constant depending only on q), where*

$$\alpha = t - 1 - \left\lfloor \frac{t-1}{q} \right\rfloor.$$

If each G is centrally symmetric and t is odd, then

$$\alpha = t - 2 - \left\lfloor \frac{t-2}{q} \right\rfloor.$$

Moreover, G is quasi-linear as $\ell \to \infty$, assuming precomputation of each F_i.

We can make some comparisons between Theorem 1.2 and other asymptotic bounds. To properly state these other bounds, we recall that a function $f(n)$ is of class $\Omega(g(n))$ if $f(n) \geq Cg(n)$ for some constant C (the reverse of $O(g(n))$) and that $f(n)$ is $\Theta(g(n))$ if it is both $O(g(n))$ and $\Omega(g(n))$.

If μ is an m-fold product with $m \propto n$, then we can take $\ell \propto n$ in Theorem 1.2, so that $O(\ell^\alpha) = O(n^\alpha)$. In comparison, the Tchakaloff upper bound is $O(n^t)$ points, or it is $O(n^{t-1})$ points when t is odd and μ is centrally symmetric (section 4). Thus Theorem 1.2 is asymptotically better than Tchakaloff's theorem for all such product measures. Tchakaloff's theorem also does not guarantee equal weights.

Another comparison is with the cardinality of exact determination. A t-cubature formula F is *overdetermined*, *underdetermined*, or *exactly determined* if the parameters of its points provide fewer, more, or the same number of degrees of freedom, respectively, as the constraints imposed by integrating all polynomials of degree t. The cardinality of exact determination is $\Theta(n^{t-1})$ for general μ and $\Theta(n^{t-2})$ when t is odd and μ is centrally symmetric. Thus for product measures, the formulas in Theorem 1.2 are asymptotically exactly determined (up to a constant factor that depends on t) when q is large. But when $q < t-1$, or $q < t-2$ in the odd and centrally symmetric case, they are asymptotically overdetermined.

A third comparison is with the Stroud lower bound: Any t-cubature formula in n dimensions, not necessarily interior or positive, requires $\Omega(n^{\lfloor t/2 \rfloor})$ points. Theorem 1.2 achieves the Stroud bound (up to a constant factor) when $q = 2$.

A final comparison is with an interesting thinning construction of Novak and Ritter for products of quadrature formulas [16]. (It is similar to an earlier construction due to Grundmann and Möller for the n-simplex [6].) They produce t-cubature formulas with $O(n^{\lfloor t/2 \rfloor})$ points, which is within a constant factor of the Stroud bound and better than Theorem 1.2 when $q > 2$. Crucially, their formulas are not positive, although they can be made interior. They also require that the factors of μ be 1-dimensional. The Novak–Ritter construction does generalize to convolutions, as long as each factor formula has collinear points.

Theorem 1.2 can be used to construct interesting cubature formulas for several infinite sequences of regions and measures considered by Stroud [20, Chaps. 7 and 8].

THEOREM 1.3. *For any t:*

1. *the n-cube C_n with uniform measure has a QLEI t-cubature formula with $O(n^{\lfloor t/2 \rfloor})$ points;*
2. *the cubical shell $C_n - rC_n$ has a QLEI t-cubature formula with $O(n^{\lfloor t/2 \rfloor + 1})$ points.*

For any odd $t \geq 3$,

1. *$\mathbb{R}^n$ with Gaussian weight function has a QLEI t-cubature formula with $O(n^{t-2})$ points;*
2. *any spherically symmetric measure on $\mathbb{R}^n$ has a QLPI t-cubature formula with $O(n^{t-2})$ points, which includes the n-ball B_n, the spherical shell $B_n - rB_n$, and the $(n-1)$-sphere S^{n-1} with uniform measure and $\mathbb{R}^n$ with radial exponential weight function $\exp(-||\vec{x}||_2)$.*

For any $t \geq 2$,

1. *the n-simplex Δ_n has a QLPI t-cubature formula with $O(n^{t-1})$ points;*
2. *the n-cross-polytope C_n^* with uniform measure has a QLPI t-cubature formula with $O(n^{\lfloor 3t/2 \rfloor - 1})$ points.*

All cases of Theorem 1.3 other than the cross polytope C_n^* improve the Tchakaloff bound. On the other hand, the construction for the cube C_n matches the Stroud bound up to a constant factor. We admit that this t-dependent factor is very generous when t is large: For each $t = 2s+1$, it approaches $2 \cdot s^s \cdot s!$ as $n \to \infty$ in the favorable case $n = 2^m$. By contrast, the Novak–Ritter formulas use only 2^s more points than the Stroud bound as $n \to \infty$.

Theorem 1.3 partially solves a problem of Stroud [20, p. 18]: Are there PI 5-cubature formulas for C_n, B_n, or Δ_n with $O(n^2)$ or $O(n^3)$ points? Theorem 1.3 provides QLPI 5-cubature formulas with $O(n^2)$ points for C_n, $O(n^3)$ points for B_n, and $O(n^4)$ points for Δ_n. In section 3, we will establish a special QLPI 5-cubature formula for B_n with $O(n^2)$ points and QLPB and QLPI 3-cubature formulas for Δ_n with $O(n)$ points. Thus the only remaining case of Stroud's question is the n-simplex in degree 4 or 5.

REMARK 1. *The formula in Theorem* 1.3 *for S^{n-1} is technically a QLPB formula if we take the definition of boundary in general topology. However, we take boundary in the sense of geometric topology, so that Theorem* 1.3 *is correct as stated.*

2. Proofs. *Proof of Theorem* 1.1. First, identify an affinely independent set of q points in $\mathbb{R}^{q-1}$ with the finite field $\mathbb{F}_q$. For each $1 \leq i \leq \ell$, choose a linear map $\pi_i : \mathbb{R}^q \to \mathbb{R}^n$ that sends $\mathbb{F}_q$ to F_i and define $\pi : \mathbb{R}^{(q-1)\ell} \to \mathbb{R}^n$ to be their direct sum:

$$\pi = \pi_1 \oplus \pi_2 \oplus \cdots \oplus \pi_\ell.$$

Because $F_1 * F_2 * \cdots * F_\ell$ is a t-design for the measure μ on $\mathbb{R}^n$, the identity

$$\int P(\vec{x}) d\mu = \frac{1}{q^\ell} \sum_{\vec{p} \in \mathbb{F}_q^\ell} P(\pi(\vec{p}))$$

holds for any polynomial P of degree at most t on $\mathbb{R}^n$. Now suppose that we thin the set $F = \pi(\mathbb{F}_q^\ell)$ to a set $G = \pi(A)$ for some set $A \subset \mathbb{F}_q^\ell$. Since π is linear, if we want G to be a t-cubature formula for μ as F is, it suffices that

$$\frac{1}{q^\ell} \sum_{\vec{p} \in \mathbb{F}_q^\ell} P(\vec{p}) = \frac{1}{|A|} \sum_{\vec{p} \in A} P(\vec{p}) \tag{2.1}$$

for any polynomial P on $\mathbb{R}^{(q-1)\ell}$ of degree at most t. If P is a monomial, then as a function on $\mathbb{F}_q^\ell$, it depends on at most t coordinates. Conversely, any function on $\mathbb{F}_q^\ell$ that depends on at most t coordinates is realized by a polynomial of degree at most t. It follows that (2.1) is equivalent to the statistical property that the projection of

A onto any t of the I coordinates of $\mathbb{F}_q^\ell$ is constant-to-1. Such a set A is called an *orthogonal array* of strength t.

If C is an $[\ell, k, t+1]_q$ code, then the dual space C^* (in the sense of linear algebra over $\mathbb{F}_q$) is a linear orthogonal array of strength t. Since C has dimension k, C^* has dimension $\ell - k$ and therefore has $q^{\ell-k}$ points. Thus we can let $G = \pi(C^*)$.

The refinement when t is odd and each F_i is centrally symmetric is as follows. If q is odd, we replace $\mathbb{R}^{q-1}$ by $\mathbb{R}^{(q-1)/2}$, and we position $\mathbb{F}_q$ as a centrally symmetric set that does not lie in a hyperplane. (In other words, the points of $\mathbb{F}_q$ are the vertices of an affinely regular cross polytope, plus the origin.) We further demand that negation in $\mathbb{F}_q$ coincides with negation in $\mathbb{R}^{(q-1)/2}$. Then any centrally symmetric subset $A \subset \mathbb{F}_q^\ell$ is centrally symmetric in $\mathbb{R}^{(q-1)\ell/2}$. In this case both sides of (2.1) vanish when P is an odd polynomial. Thus A need only be an orthogonal array of strength $t-1$. In particular, this is so if $A = C^*$, because C^* is a vector space over $\mathbb{F}_q$ and vector spaces are centrally symmetric sets.

Finally, if t is odd, q is even, and C is a zero-sum code, then C^* contains the vector $(1, 1, \ldots, 1)$ and therefore is invariant under addition by this vector. In this case we replace $\mathbb{R}^{q-1}$ in the general construction by $\mathbb{R}^{q/2}$ and realize $\mathbb{F}_q$ as a centrally symmetric set (the vertices of a regular cross polytope). We further demand that adding 1 in $\mathbb{F}_q$ coincides with negation in $\mathbb{R}^{q/2}$. Then once again C^* is centrally symmetric and need only be an orthogonal array of strength $t-1$. □

The following lemma establishes Theorem 1.2 as a corollary of Theorem 1.1.

LEMMA 2.1. *Let q be a prime power, let $m, t \in \mathbb{Z}_{\geq 0}$, and let*

$$\alpha = t - 1 - \left\lfloor \frac{t-1}{q} \right\rfloor.$$

Then there is a $[q^m, k, u]_q$ zero-sum code C with

$$u \geq t+1, \qquad k \geq q^m - m\alpha - 1.$$

The code in Lemma 2.1 is called an (extended, narrow-sense) *BCH code* [14, 3, 1, 10]. We will use the duals of BCH codes to thin cubature formulas. As it happens, the dual of a BCH code of this type is another BCH of the same type.

Proof. It is easier to define the dual code C^* and show that it is an orthogonal array. Since it is a linear space, it suffices to show that every coordinate projection $\pi_I : C^* \to \mathbb{F}_q^I$ with $|I| \leq t$ is onto. There is an important $\mathbb{F}_q$-linear function

$$\mathrm{Tr}_q : \mathbb{F}_{q^m} \to \mathbb{F}_q$$

called the *trace*. (It is analogous to the taking the real part of a complex number.) First, we interpret $\mathbb{F}_q^{q^m}$ as the space of all functions from $\mathbb{F}_{q^m}$ to $\mathbb{F}_q$. We define C^* as the set of all functions

$$f : \mathbb{F}_{q^m} \to \mathbb{F}_q, \qquad f(x) = \mathrm{Tr}_q(P(x)),$$

where P is a polynomial of degree at most $t-1$. If $I \subseteq \mathbb{F}_{q^m}$ and $|I| \leq t$, the polynomial P can achieve any desired values on I by Lagrange interpolation. Thus the distance of C is at least $t+1$.

The space of polynomials of degree $t-1$ on $\mathbb{F}_{q^m}$ has $\mathbb{F}_{q^m}$-dimension t, and therefore $\mathbb{F}_q$-dimension mt. But taking the trace reduces the dimension in two ways. To give

an explicit example, suppose that $q = 2$, $t = 3$, and m is arbitrary. Then C^* is the set of all

$$f(x) = \mathrm{Tr}_2(ax^2 + bx + c).$$

The apparent dimension of C^* is $3m$. But f depends only on the trace of c, so c contributes 1 rather than m to the dimension of C^*. Moreover, $\mathrm{Tr}_2(bx) = \mathrm{Tr}_2(b^2x^2)$, so the linear term can be removed from f, with the conclusion that

$$\dim C^* \leq m + 1.$$

In general, the constant term of P contributes 1 to the dimension and the other $t-1$ terms contribute m each, except that $\lfloor \frac{t-1}{q} \rfloor$ terms are superfluous by the Frobenius automorphism $x \mapsto x^q$. Thus

$$\dim C^* \leq m\alpha + 1,$$

as desired.

Since constants are polynomials of degree 0, C^* contains constant vectors. Therefore C is a zero-sum code. □

On the face of it, Lemma 2.1 establishes Theorem 1.2 only when $\ell = q^m$. If $q^{m-1} < \ell < q^m$, we can project a BCH code from $\mathbb{F}_q^{q^m}$ to $\mathbb{F}_q^n$. This preserves the $O(t^\alpha)$ bound at the expense of worsening the constant factor. If ℓ is not much more than q^{m-1}, we can slightly improve the projected code with a projection that annihilates up to $\alpha - 1$ independent vectors in C^*. (See Theorem 3.1 for an example.)

REMARK 2. *The inequalities for u and k in Lemma* 2.1 *become sharp as $m \to \infty$.*

Proof of Theorem 1.3. The simplest case to consider is with uniform measure and $\mathbb{R}^n$ with Gaussian weight function. This fits Theorem 1.2 with $\ell = n$, provided that for each t we find an EI, centrally symmetric t-quadrature formula with Gaussian weight and with q points for some prime power q. Since we assume no upper bound on q, the very general Seymour–Zaslavsky theorem [18] establishes the existence of such formulas. (The proof of the theorem, but not the statement, shows that there is a Q such that we can take any $q \geq Q$. In particular we can take $q = 2^k$. We can then symmetrize the formula by taking the multiset union of it and its reflection.) On the other hand, since q is large, $\lfloor (t-2)/q \rfloor = 0$. Thus Theorem 1.2 produces formulas with $O(n^{t-2})$ points.

We will need the same construction for the orthant $\mathbb{R}^n_{\geq 0}$ with exponential weight function $\exp(-||\vec{x}||_1)$. This measure does not have central symmetry, and the end result is formulas with $O(n^{t-1})$ points, again with $\ell = n$.

The next simplest case is the n-simplex Δ_n. Recall that Δ_n has barycentric coordinates

$$x_0 + x_1 + \cdots + x_n = 1$$

which realize it as a subset of the orthant $\mathbb{R}^{n+1}_{\geq 0}$. If $P(\vec{x})$ is a polynomial of degree t on Δ_n, then it can be *homogenized*: it can be expressed as a homogeneous polynomial of degree t by attaching a factor of $\left(\sum_i x_i\right)^{t-s}$ to each term of degree s. In this case

$$\int_{\Delta_n} P(\vec{x}) d\vec{x} = \frac{1}{(n+t)!} \int_{\mathbb{R}^{n+1}_{\geq 0}} P(\vec{x}) \exp(-||\vec{x}||_1) d\vec{x}.$$

Therefore we can project any nonexterior cubature formula for $\mathbb{R}^{n+1}_{\geq 0}$ radially onto Δ_n without loss of degree, although the weights change. (If the origin happens to

be a cubature point, discard it.) In particular, we can project the cubature formulas provided by Theorem 1.2 as explained previously. The formulas still have $O(n^{t-1})$ points, although the weights are no longer equal.

The same argument works for the sphere $S^{n-1} \subset \mathbb{R}^n$ for centrally symmetric formulas. Every polynomial P on S^{n-1} can be expressed as $P_S + P_A$, where P_S is centrally symmetric and P_A is centrally antisymmetric. The integral of P_A vanishes, as does its sum with respect to any centrally symmetric formula. Meanwhile every term of P_S has even degree, so it can be expressed as a homogeneous polynomial on $\mathbb{R}^n$ using the equation

$$x_1^2 + \cdots + x_n^2 = 1$$

for the unit sphere. Then

$$\int_{S^{n-1}} P(\vec{x}) d\Omega = \frac{2}{(\frac{n+t}{2} - 1)!} \int_{\mathbb{R}^n} P(\vec{x}) \exp(-||\vec{x}||_2^2) d\vec{x},$$

where Ω is usual surface volume on S^{n-1}. Again, any centrally symmetric cubature formula can be radially projected and the weights can be adjusted.

Formulas for the ball B_n and the spherical shell $B_n - rB_n$ can be derived from formulas for the sphere S^{n-1} using radial separation of variables [20, Thm. 2.8]. The result is a product formula where the radial factor can be Gaussian quadrature. The number of points in this factor does not increase with dimension.

The cross-polytope C_n^* is the union of 2^n simplices. Thus we can obtain formulas for C_n^* by repeating formulas for Δ_n. In degree t, we do not need all 2^n copies; instead we can repeat it in the pattern of the BCH code over $\mathbb{F}_2$ defined by polynomials of degree $t-1$ over $\mathbb{F}_{2^m}$. Such a code has $O(n^{\lfloor t/2 \rfloor})$ vectors and the formula for Δ_n has $O(n^{t-1})$ points, so the total is $O(n^{\lfloor 3t/2 \rfloor - 1})$ points.

The n-cube $C_n = [-1, 1]^n$ is in some ways the most interesting case. As in the Gaussian case, it is a straight application of Theorem 1.2 using an equal-weight quadrature formula. But in this case we will carefully choose the quadrature formula on $[-1, 1]$ to itself be a convolution of $s = \lfloor t/2 \rfloor$ formulas with two points. For example, the Chebyshev 5-quadrature formula has points at

$$\pm\sqrt{\frac{5+\sqrt{5}}{30}} \pm \sqrt{\frac{5-\sqrt{5}}{30}}.$$

This is evidently a convolution, as is any centrally symmetric, equal-weight formula with four points. Elsewhere [13] we show that the 2^s points

$$\pm z_1 \pm z_2 \pm \cdots \pm z_s$$

form a Chebyshev-type $(2s+1)$-quadrature formula for $[-1, 1]$ with constant weight if and only if the z_i's are the roots of the polynomial

$$Q(x) = x^s - \frac{x^{s-1}}{3} + \frac{x^{s-2}}{45} - \cdots + \frac{(-1)^s}{1 \cdot 3 \cdot 15 \cdots (4^s - 1)}.$$

We also show that all roots of Q are real and that the resulting quadrature formula is interior. The n-fold product power of this formula is thus a convolution of sn pairs of points, so we can apply Theorem 1.2 with $\ell = sn$ and $q = 2$.

Finally, the $O(n^{\lfloor t/2 \rfloor})$ formula for the n-cube C_n yields a $O(n^{\lfloor t/2 \rfloor + 1})$ formula for the cube surface ∂C_n just by repeating the formula for C_{n-1} on each facet of C_n. Then radial separation of variables produces a product formula for the cubical shell $C_n - rC_n$ which also has $O(n^{\lfloor t/2 \rfloor + 1})$ points. □

3. Special constructions and examples. In this section we will consider some examples and special constructions with concern for constant factors. For this purpose, we spell out more precisely the notion of an orthogonal array. Let A be a finite set. If a subset $T \subset A^n$ has the property that its projection $T \to A^I$ is a constant-to-1 map for every $|I| \le t$, then T is an *orthogonal array* of strength t, or an $\mathrm{OA}(|T|, n, |A|, t)$ [9]. If $A = \mathbb{F}_q$ and $T = C^*$ is the dual of an $[n, k, t]_q$ code, then T is an $\mathrm{OA}(q^{n-k}, n, q, t-1)$. We will also say that T is an $[n, n-k, t^*]_q$ to refer to its linear structure and indicate its dual distance.

If $|A| = q$ is a prime power and t is fixed, then BCH codes are the best presently known $\mathbb{F}_q$-linear orthogonal arrays in the limit $n \to \infty$. But a few nonlinear arrays are slightly better.

A *Hadamard matrix* of order n is an $n \times n$ matrix with entries ± 1 and with orthogonal rows (and therefore orthogonal columns). It is easy to show that a Hadamard matrix is equivalent to an $\mathrm{OA}(2n, n, 2, 3)$. A $[2^m, m+1, 4^*]_2$ BCH code, which is also called a first-order Reed–Muller code, yields a Hadamard matrix of order 2^m. But there are also Hadamard matrices for other values of n, for example when $n-1$ is prime and 4 divides n. The Hadamard conjecture asserts that there is a Hadamard matrix of every order n divisible by 4.

For any even $m \ge 4$, there is a Kerdock code which is a nonlinear $\mathrm{OA}(2^{2m}, 2^m, 2, 5)$. It has half as many points as the corresponding $[2^m, 2m+1, 6^*]_2$ BCH code [9, 8, 11]. For any even $m \ge 6$, there is a Delsarte–Goethals code which is a nonlinear $\mathrm{OA}(2^{3m-2}, 2^m, 2, 7)$ [5]. It has one-fourth as many points as the corresponding $[2^m, 3m+1, 8^*]_2$ BCH code.

The following result comes from thinning some cubature formulas of Stroud, some of whose points have a product structure.

THEOREM 3.1. *Let $n \ge 6$, and let*

$$k = \begin{cases} 4m, & 2^{2m-1} < n \le 2^{2m}, \\ 4m+2, & 2^{2m} < n \le 2^{2m} + 2^m, \\ 4m+3, & 2^{2m} + 2^m < n \le 2^{2m+1}. \end{cases}$$

Then the sphere S^{n-1}, $\mathbb{R}^n$ with Gaussian measure, and the ball B_n admit QLPI 5-cubature formulas with $2^k + 2n$ points.

Proof. The formulas S_n:5-3, U_n:5-2, and $E_n^{r^2}$:5-3 listed in Stroud [20, pp. 270, 294, and 317] have $2^n + 2n$ points with 2^n of them lying on the vertices of a cube. These 2^n points can be thinned to either the $[2^{2m+1}, 4m+3, 6^*]_2$ BCH code or the Kerdock $\mathrm{OA}(2^{4m}, 2^{2m}, 2, 5)$ and then projected down to n dimensions.

If $2^{2m} < n \le 2^{2m} + 2^m$, then the $[2^{2m+1}, 4m+3, 6^*]_2$ BCH code can be reduced in half by carefully choosing the projection. The code has a vector of weight $2^{2m} - 2^m$, so when n is only slightly larger than 2^{2m}, we can choose a projection that annihilates this vector.

In each of the three cases, the result is a formula with $2^k + 2n$ points. ∎

Actually, Theorem 3.1 is not quite optimal, because it uses a convenient set of good distance-6 linear codes and nonlinear strength-5 orthogonal arrays rather than the best ones presently known. A complicated map of the best presently known linear codes over $\mathbb{F}_2$ of length $n \le 256$ is provided by the "best codes" functions in Magma [22]. Undoubtedly this map could be augmented by nonlinear orthogonal arrays, but we know of no effort to do so. When n is a power of 2, Kerdock and BCH codes are the best presently known choices.

Victoir [21] also established Theorem 3.1 (with BCH codes). If $n = 2^m$ and Stroud's formulas for S^{n-1} is thinned using a BCH code, it then has equal weights and is therefore a 5-design. Interestingly, in this case it has a transitive symmetry group and was previously found by Calderbank et al. [2]. Similar constructions were found by König [12], by Sidelnikov [19], and by Schechtman, interpreting the work of Hajela [7].

We can obtain a good 3-cubature formula for the cube C_n by a straightforward application of Theorem 1.2 using the 2-point Gaussian quadrature formula for the interval $[-1, 1]$. Thinning the product formula using a BCH code yields a 2^{j+1}-point formula when $2^{j-1} < n \le 2^j$. When $n = 2^j$, or more generally whenever there is a Hadamard matrix of order n, the product formula can be thinned to the $2n$ vertices of a certain regular cross-polytope inside C_n. A formula due to Stroud, C_n:3-1 (see [20, p. 230]), also uses the vertices of a regular cross-polytope but not the same one.

We can obtain a 3-cubature formula with $O(n)$ points for Δ_{n-1} with a similar construction. Using known Hadamard matrices, the formula has $3n + o(n)$ points; if the Hadamard conjecture holds, it has between $3n - 1$ and $3n + 5$ points. First, the positive ray $\mathbb{R}_{\ge 0}$ with exponential weight has an equal-weight 2-quadrature formula with points at 0 and 2. If we apply Theorem 1.1 to this formula and a Hadamard matrix of order n, the result is a $2n$-point formula F on $\mathbb{R}^n_{\ge 0}$ which also has degree 2. However, if our interest is integration on Δ_{n-1}, we need only consider homogeneous polynomials on $\mathbb{R}^n_{\ge 0}$. The formula F correctly integrates every degree 3 monomial other than x_i^3. We can fix F for these monomials, without changing its sum for $x_i^2 x_j$ or $x_i x_j x_k$, by adding a point at $(1, 0, 0, \dots)$ (and permutations) with weight 2.

The projected formula on Δ_{n-1} consists of these points and weights in barycentric coordinates:

$$\begin{array}{cc} (1, 0, 0, \dots, 0)_S, & \frac{2}{n(n+1)(n+2)}, \\ (\frac{2}{n}, \frac{2}{n}, \dots, \frac{2}{n}, 0, 0, \dots, 0)_H, & \frac{n}{2(n+1)(n+2)}, \\ (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}), & \frac{4n}{(n+1)(n+2)}. \end{array}$$

The subscript S denotes full symmetrization, as in Stroud's notation. The subscript H denotes symmetrization in the pattern of a Hadamard design. (See section 4.) This produces a formula with $3n-1$ points, provided that there exists a Hadamard matrix of order n. When there is none, we can use a Hadamard matrix of order $\ell > n$. The formula on $\Delta_{\ell-1}$ with $3\ell - 1$ points can be projected onto Δ_{n-1}, as in the proof of Theorem 1.3. We can take $\ell = n + o(n)$ by letting $\ell - 1$ be the first prime after n which is 3 mod 4. If the Hadamard conjecture holds, we can take $\ell = 4\lceil n/4 \rceil$.

Stroud asked for a practical, PI 5-cubature formula for C_{100}. Following Theorem 1.3, we can find one by thinning the product formula coming from the 4-point Chebyshev quadrature on $[-1, 1]$. This product formula is the convolution of 200 pairs of points, so we can thin it using the Kerdock $\mathrm{OA}(2^{16}, 2^8, 2, 5)$, projected to 200 dimensions. The cubature formula therefore has $2^{16} = 65536$ points, which would have been fairly practical even in 1971 when Stroud asked the question. (The Kerdock code used here was discovered shortly afterward [11], but the BCH codes were known in 1959 [1, 10].)

Victoir [21] found another thinning of the same Chebyshev product formula with $4^{12} = 16777216$ points, which the author tied in the first version of this paper.

Note that the Chebyshev–Kerdock 5-cubature formula for C_{100} is overdetermined. The threshold of exact determination for centrally symmetric 5-cubature formulas on

C_{100} is 87651 points. Meanwhile the centrally symmetric Tchakaloff bound is 8852652 points, while the Stroud lower bound is 5050 points.

Finally, Schürer [17] compared the numerical accuracy of various cubature and quasi-Monte Carlo methods for the integration of various test functions defined on C_n with $2 \leq n \leq 100$. He assumed a more modern limit of 2^{25} evaluations of the integrand. For much of this test regime we can suggest the following cubature formulas: Start with the power of the convolutional 7-quadrature formula [13] for $[-1, 1]$, whose points are approximately at

$$\pm.500128 \pm .243941 \pm .153942.$$

Then thin the n-fold product power of this formula using a Delsarte–Goethals code. The result is an EI 7-cubature formula with at most 2^{23} points up to dimension $\lfloor 256/3 \rfloor = 85$.

4. Other comments. Victoir [21] proposes thinning symmetric cubature formulas rather than product or convolution formulas. The enabling result of symmetric cubature formulas is Sobolev's theorem: If a linear action of a finite group G preserves μ, then a cubature formula consisting of orbits of G need only be checked for G-invariant polynomials. Victoir extends Sobolev's theorem with a G-invariant generalization of Tchakaloff's theorem: A PI cubature formula needs only as many orbits as the dimension of $\mathbb{R}[\vec{x}]^G_{\leq t}$, the space of G-invariant polynomials of degree at most t. One important special case is when G is the 2-element central symmetry group. If μ is a measure on $\mathbb{R}^n$ with central symmetry and t is odd, the bound from this version of Tchakaloff's theorem is $O(n^{t-1})$ points.

Even if a cubature formula F uses very few orbits of G, some of the orbits might be very large. Victoir proposes thinning each large orbit separately. He notes that this can be done using linear programming, among other methods; linear programming on a set of G-orbits should be much easier than general numerical methods to find positive cubature formulas for μ. If $G = (\mathbb{Z}/2)^n$ is the group of independent sign changes of all n coordinates, then an orbit of Gq can be identified with $\mathbb{F}_2^k$ for some $k \leq n$. In this case Victoir found the constructions of Theorems 1.1 and 1.2. (In the case of 5-cubature on C_n, he found a special construction with $O(n^3)$ points with elements of both Theorem 1.1 and the n-cube case of Theorem 1.3.)

If G is the group of coordinate permutations, then an orbit whose points have two distinct coordinates can be identified with the set of k-subsets of an n-set. A geometric t-design T within this orbit is also a traditional combinatorial t-design, or an $(n, k, t) - \lambda$ design. Namely, T is a collection of blocks of size k in a set of n such that each t-subset is contained in exactly λ blocks. In particular, an $(n, \frac{n}{2}, 3) - \frac{n}{4}$ design is called a *Hadamard design*, because it comes from the rows of a Hadamard matrix.

These constructions motivate the notion of a weighted orthogonal array. We define it as a finite set A and a measure μ on A^n that projects to uniform measure on each A^I with $|I| \leq t$. More generally, μ might project to σ^n for some reference measure σ on S. Such arrays could improve of Theorem 1.1; the factor formulas would not need to have equal weights.

Finally, cubature formulas coming from Theorem 1.1 could be viewed as quasi-Monte Carlo methods. They are similar to some constructions of (t, m, s)-nets, which are quasi-Monte Carlo methods first defined and largely developed by Niederreiter [15]. Nonetheless, PI cubature formulas and discrepancy-based quasi-Monte Carlo methods are thought to have complementary advantages [17]. We believe that the

improved asymptotics presented here could change the standing of cubature among numerical methods for integration.

Acknowledgments. The author would like to thank Hermann König, Eric Rains, and Hong Xiao for useful discussions. The author is also indebted to the late Arthur Stroud for his excellent introduction to the cubature problem.

REFERENCES

[1] R. C. Bose and D. K. Ray-Chaudhuri, *On a class of error correcting binary group codes*, Information and Control, 3 (1960), pp. 68–79.
[2] A. R. Calderbank, R. H. Hardin, E. M. Rains, P. W. Shor, and N. J. A. Sloane, *A group-theoretic framework for the construction of packings in Grassmannian spaces*, J. Algebraic Combin., 9 (1999), pp. 129–140, arXiv:math.CO/0208002.
[3] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., Grundlehren Math. Wiss. 290, Springer-Verlag, New York, 1993.
[4] R. Cools, *An encyclopaedia of cubature formulas*, J. Complexity, 19 (2003), pp. 445–453.
[5] P. Delsarte and J.-M. Goethals, *Alternating bilinear forms over $GF(q)$*, J. Combin. Theory Ser. A, 19 (1975), pp. 26–50.
[6] A. Grundmann and H. M. Möller, *Invariant integration formulas for the n-simplex by combinatorial methods*, SIAM J. Numer. Anal., 15 (1978), pp. 282–290.
[7] D. Hajela, *Construction techniques for some thin sets in duals of compact abelian groups*, Ann. Inst. Fourier (Grenoble), 36 (1986), pp. 137–166.
[8] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, *The $\mathbf{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 301–319, arXiv:math.CO/0207208.
[9] A. S. Hedayat, N. J. A. Sloane, and J. Stufken, *Orthogonal Arrays: Theory and Applications*, Springer Series in Statistics, Springer-Verlag, New York, 1999.
[10] A. Hocquenghem, *Codes correcteurs d'erreurs*, Chiffres, 2 (1959), pp. 147–156.
[11] A. M. Kerdock, *A class of low-rate nonlinear binary codes*, Information and Control, 20 (1972), pp. 182–187.
[12] H. König, *Cubature formulas on spheres*, in Advances in Multivariate Approximation (Witten-Bommerholz, 1998), Math. Res. 107, Wiley-VCH, Berlin, 1999, pp. 201–211.
[13] G. Kuperberg, *Special moments*, Adv. in Appl. Math., 34 (2005), pp. 853–870, arXiv:math.PR/0408360.
[14] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland Math. Library 16, North–Holland, Amsterdam, 1977.
[15] H. Niederreiter, *Point sets and sequences with small discrepancy*, Monatsh. Math., 104 (1987), pp. 273–337.
[16] E. Novak and K. Ritter, *Simple cubature formulas with high polynomial exactness*, Constr. Approx., 15 (1999), pp. 499–522.
[17] R. Schürer, *A comparison between (quasi-)Monte Carlo and cubature rule based methods for solving high-dimensional integration problems*, Math. Comput. Simulation, 62 (2003), pp. 509–517.
[18] P. D. Seymour and T. Zaslavsky, *Averaging sets: A generalization of mean values and spherical designs*, Adv. in Math., 52 (1984), pp. 213–240.
[19] V. M. Sidelnikov, *Spherical 7-designs in 2^n-dimensional Euclidean space*, J. Algebraic Combin., 10 (1999), pp. 279–288.
[20] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall Series in Automatic Computation, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[21] N. Victoir, *Asymmetric cubature formulae with few points in high dimension for symmetric measures*, SIAM J. Numer. Anal., 42 (2004), pp. 209–227.
[22] Magma, http://magma.maths.usyd.edu.au/.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 908–935

NUMERICAL CUBATURE FROM ARCHIMEDES' HAT-BOX THEOREM*

GREG KUPERBERG†

Dedicated to Krystyna Kuperberg on the occasion of her 60th birthday

Abstract. Archimedes' hat-box theorem states that uniform measure on a sphere projects to uniform measure on an interval. This fact can be used to derive Simpson's rule. We present various constructions of, and lower bounds for, numerical cubature formulas using moment maps as a generalization of Archimedes' theorem. We realize some well-known cubature formulas on simplices as projections of spherical designs. We combine cubature formulas on simplices and tori to make new formulas on spheres. In particular, S^n admits a 7-cubature formula (and sometimes a 7-design) with $O(n^4)$ points. We establish a local lower bound on the density of a positive interior cubature formula on a simplex using the moment map.

Along the way we establish other quadrature and cubature results of independent interest. For each t, we construct a lattice trigonometric $(2t+1)$-cubature formula in n dimensions with $O(n^t)$ points. We derive a variant of the Möller lower bound using vector bundles. And we show that Gaussian quadrature is very sharply locally optimal among positive quadrature formulas.

Key words. cubature formulas, moment maps, projective space, lattices

AMS subject classification. 65D32

DOI. 10.1137/040615584

1. Introduction.

Let μ be a measure on $\mathbb{R}^n$ with finite moments. A *cubature formula of degree t* for μ is a set of points $F = \{\vec{p}_a\} \subset \mathbb{R}^n$ and a weight function $\vec{p}_a \mapsto w_a \in \mathbb{R}$ such that

$$\int P(\vec{x}) d\mu = P(F) \stackrel{\text{def}}{=} \sum_{a=1}^{N} w_a P(\vec{p}_a)$$

for polynomials P of degree at most t. (If $n = 1$, then F is also called a *quadrature formula*.) The formula F is *equal-weight* if all w_a are equal, *positive* if all w_a are positive, and *negative* if at least one w_a is negative. Let X be the support of μ. The formula $\mathbb{F}$ is *interior* if every point $\vec{p}_a$ is in the interior of X; it is *boundary* if every $\vec{p}_a$ is in X and some $\vec{p}_a$ is in ∂X; and otherwise it is *exterior*. We will mainly consider positive interior (PI) and positive boundary (PB) cubature formulas, and we will also assume that μ is normalized so that total measure is 1. PI formulas are the most useful in numerical analysis [30, Chap. 1]. This application also motivates the main question of cubature formulas, which is to determine how many points are needed for a given formula and a given degree t. Equal-weight formulas that are either interior or boundary (EI or EB) are important for other applications, in which context they are also called *t-designs*.

Our starting point is to use Archimedes' hat-box theorem [2] to relate quadrature on the interval $[-1, 1]$ and cubature on the unit sphere S^2, both with uniform measure. (Lest the reader be misled, this starting point is much simpler and easier than the generalizations that will eventually follow.) Archimedes' theorem says that the orthogonal projection π from S^2 to the z coordinate preserves normalized uniform

*Received by the editors September 23, 2004; accepted for publication (in revised form) November 22, 2005; published electronically May 5, 2006. This work was supported by NSF grant DMS 0306681.
http://www.siam.org/journals/sinum/44-3/61558.html

†Department of Mathematics, University of California, Davis, CA 95616 (greg@math.ucdavis.edu).

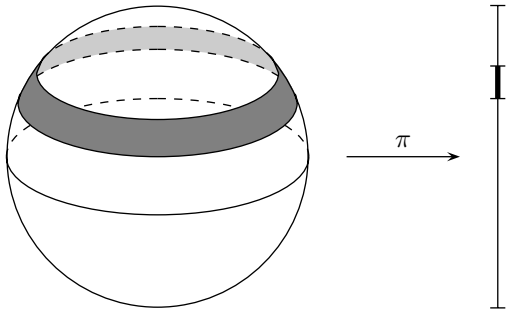


FIG. 1. *Archimedes' hat-box theorem.*

measure. In plainer terms, for any interval $I \subset [a, b]$ or other measurable set, the area of $\pi^{-1}(I)$ is proportional to the length of I; see Figure 1. (It is called the hat-box theorem because the surface area of a hemispherical hat equals the area of the side of a cylindrical box containing it.) Therefore if F is a t-cubature formula on S^2, its projection $\pi(F)$ is a t-cubature formula on $[-1, 1]$.

The 2-sphere S^2 has five especially nice cubature formulas given by the vertices of the Platonic solids. Their cubature properties follow purely from a symmetry argument of Sobolev [27]. Suppose that G is the group of common symmetries of a putative cubature formula F and its measure μ. If $P(\vec{x})$ is a polynomial and $P_G(\vec{x})$ is the average of its G-orbit, then

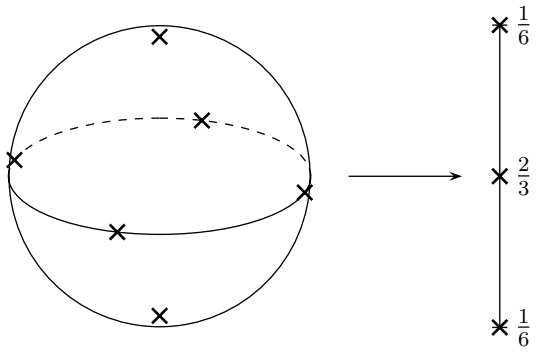
$$\int P_G(\vec{x}) d\mu = \int P(\vec{x}) d\mu, \qquad P_G(F) = P(F).$$

Therefore it suffices to check F for G-invariant polynomials. In particular, if every G-invariant polynomial of degree $\leq t$ is constant, then any G-orbit is a t-design.

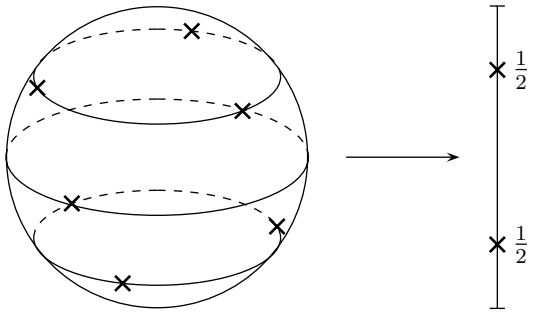
By Sobolev's theorem, the vertices of a regular octahedron form a 3-design on S^2. If we project this formula using Archimedes' theorem, the result is Simpson's rule. Another projection of the same six points yields 2-point Gauss–Legendre quadrature. Figure 2 shows both projections. The eight vertices of a cube are also a 3-design. One projection is again 2-point Gauss–Legendre quadrature; another is Simpson's $\frac{3}{8}$ rule. Finally the twelve vertices of a regular icosahedron form a 5-design by symmetry. One projection of these twelve points is 4-point Gauss–Lobatto quadrature.

The rest of this article applies toric moment maps, which generalize Archimedes' theorem to higher dimensions, to the cubature problem. Section 2 shows that several well-known quadrature formulas on the interval and cubature formulas on simplices are projections of higher-dimensional symmetric formulas. Section 4 combines formulas on tori with formulas on simplices and moment maps to make formulas on spheres and projective spaces. In particular, it constructs a PI 7-cubature formula on the sphere S^n with $O(n^4)$ points. Finally, section 6 uses moment maps to establish a local lower bound for the density of points in any PI cubature formula on a simplex. A similar lower bound holds for an arbitrary simple convex polytope.

Along the way we establish some other quadrature and cubature results that are not derived from moment maps but are of independent interest. Section 3 establishes new lattice cubature formulas on tori that are similar to cubature formulas based on error-correcting codes [18]. In particular, it constructs for each t a trigonometric



(a) Simpson's rule



(b) 2-point Gauss-Legendre rule

FIG. 2. *Two projections of the octahedron rule.*

$(2t+1)$-cubature formula on $[0, 2\pi)^n$ of lattice type with $O(n^t)$ points. This improves a construction of Cools, Novak, and Ritter with $O(n^{2t})$ points and negative weights [5], and agrees up to a t-dependent constant factor with the Stroud-type lower bound [20, 22, 31]. Section 5 presents a refinement of this well-known lower bound in odd degree. It is similar to the Möller bound [19] but applies to some new cases. Section 6 also establishes that Gaussian quadrature is very sharply locally optimal among all positive quadrature formulas (Theorem 6.3). This bound might be previously known, since Gaussian quadrature has been widely studied, but the author could not find it in the literature.

2. Projection constructions. The immediate higher-dimensional generalization of Archimedes' theorem replaces the sphere S^2 by the complex manifold $\mathbb{C}P^n$. This manifold has a natural metric and a natural real algebraic structure. Concretely, assume that the projective coordinates $(z_0 : z_1 : \cdots : z_n)$ of $\mathbb{C}P^n$ are normalized so that

$$|z_0|^2 + |z_1|^2 + \cdots + |z_n|^2 = 1.$$

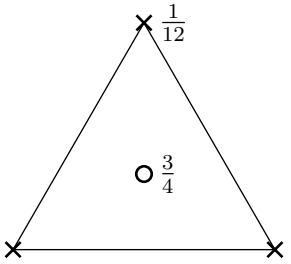


FIG. 3. *A 2-dimensional generalization of Simpson's rule.*

Then the coordinates $z_k\overline{z}_j$ together embed $\mathbb{C}P^n$ into $\mathbb{C}^{(n+1)^2}$ as a real algebraic variety (with $\mathbb{C}^{(n+1)^2}$ interpreted as a $2(n+1)^2$-dimensional real vector space) and a Riemannian manifold. This embedding is familiar in quantum mechanics as the density matrix (or density operator) formalism [21, sect. 2.4]. The induced metric is called the Fubini–Study metric. Since the metric yields a measure on $\mathbb{C}P^n$, and since it is a real algebraic variety, we can consider cubature formulas on it.

There is a projection $\pi : \mathbb{C}P^n \to \Delta_n$ to the n-simplex given by

$$\pi(z_0 : z_1 : \cdots : z_n) = (|z_0|^2, |z_1|^2, \ldots, |z_n|^2),$$

using normalized coordinates for $\mathbb{C}P^n$ and barycentric coordinates for Δ_n. It is linear and preserves normalized measure. In more abstract terms, π has these properties because $\mathbb{C}P^n$ is a projective toric variety and π is its moment map. Archimedes' theorem is a description of the moment map of $\mathbb{C}P^1 \cong S^2$. Thus, if F is an interior t-cubature formula on $\mathbb{C}P^n$, then $\pi(F)$ is a t-cubature formula on Δ_n.

Ivanović [13] and Wootters and Fields [34] defined one interesting family of 2-designs on $\mathbb{C}P^{q-1}$ for $q = p^k$ a prime power. If p is odd, then the 2-design is the orbit of a standard basis vector e_k in the group generated by cyclic permutation and linear operators of the form

$$L(e_k) = \omega^{\mathrm{Tr}_p(ak^2+bk+c)} e_k,$$

where ω is a pth root of unity and Tr_p is the $\mathbb{F}_p$ trace function on $\mathbb{F}_q$. The construction is more complicated when $p = 2$. In either case, the standard basis projects to the vertices of Δ_{q-1}, and the other q^2 vectors project to the center. The result is a standard degree 2 generalization of Simpson's rule for Δ_{q-1}, shown in Figure 3 when $q = 3$.

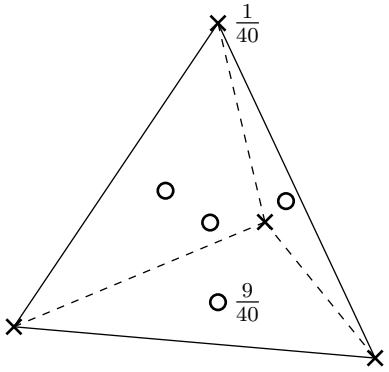
Other interesting designs and cubature formulas on $\mathbb{C}P^{n-1}$ come from designs and formulas on S^{2n-1}. The generalized Hopf fibration

$$h : S^{2n-1} \to \mathbb{C}P^{n-1}$$

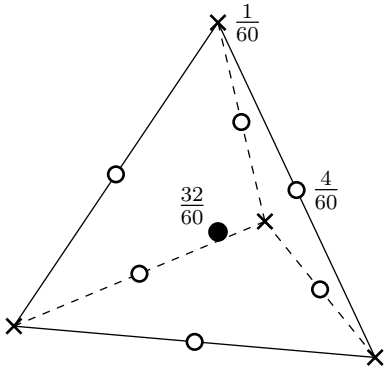
is a quadratic, volume-preserving map from S^{2n-1} to $\mathbb{C}P^{n-1}$. Namely, if we place S^{2n-1} in $\mathbb{C}^n$, h takes each point to the complex line containing it. The map h projects a $2t$- or $2t+1$-cubature formula on S^{2n-1} to a t-cubature formula on $\mathbb{C}P^{n-1}$.

One interesting example is the 240 roots of the E_8 root system, which are a 7-design as well as the solution to the sphere kissing problem in $\mathbb{R}^8$ [4, sect. 14.2]. The root system has two natural positions in $\mathbb{C}^4$. In the first position, it is generated from the two points

$$(1, 1, 1, 0), \qquad (1 - \omega, 0, 0, 0)$$



(a) 8 points



(b) 11 points

FIG. 4. 3-*cubature formulas on* Δ_3 *from the* E_8 *root system.*

by freely permuting the first three coordinates, applying the map $(a, b, c, d) \mapsto (d, a, -b, c)$, and multiplying any one coordinate by ω, a cube root of unity. In the second position, it is generated from the three points

$$(1, 1, 1, 1), \qquad (2, 0, 0, 0), \qquad (1 + i, 1 + i, 0, 0)$$

by freely permuting the four coordinates and multiplying any two coordinates by i. These two positions respectively exhibit the Eisenstein and Gaussian lattice structures of the E_8 lattice. The Hopf fibration sends the Eisenstein position of the root system to a 40-point 3-design in $\mathbb{C}P^3$ and the Gaussian position to a 60-point 3-design. Then the moment map projects these two 3-designs to 3-cubature formulas for the tetrahedron Δ_3 that appear in Abramowitz and Stegun [1, p. 895]. They have 8 and 11 points, respectively, and are shown in Figure 4.

The composition $\pi \circ h$ of the moment map and the Hopf fibration is a torus fibration $\tau_2 : \mathbb{R}^{2n} \to \Delta_{n-1}$ that does not fully depend on the complex structure $\mathbb{R}^{2n} = \mathbb{C}^n$ but only on the decomposition of $\mathbb{R}^{2n}$ into n orthogonal planes. Explicitly, the map is

$$\tau_2(x_1, \ldots, x_{2n}) = (x_1^2 + x_2^2, x_3^2 + x_4^2, \ldots, x_{2n-1}^2 + x_{2n}^2).$$

This projection is analogous to a map $\tau_1 : S^{n-1} \to \Delta_{n-1}$ defined by Xu [35]:

$$\tau_1(x_1, \ldots, x_n) = (x_1^2, \ldots, x_n^2).$$

The Xu map does not preserve uniform measure. Rather, it takes uniform measure on the sphere to the measure with weight function

$$w_1(\vec{y}) = \frac{2^n \pi^{n/2}}{\frac{n}{2}! n \sqrt{y_0 y_1 y_2 \cdots y_{n-1}}}$$

in barycentric coordinates.

In the case of the E_8 root system, one interesting set of orthogonal planes are the four eigenplanes of the abelian subgroup of $\mathrm{Aut}(E_8)$ of the form $C_5 \times C_5$. Rains [25] has computed the corresponding 3-cubature formula on Δ_3 using Magma [37]. In barycentric coordinates on Δ_3, its points and weights are the orbits of the two weighted points

$$\begin{aligned} \vec{p}_1 &= \frac{1}{10}(0, 0, 5 - \sqrt{5}, 5 + \sqrt{5}), \qquad & w_1 &= \frac{1}{24}, \\ \vec{p}_2 &= \frac{1}{10}(2, 2, 3 + \sqrt{5}, 3 - \sqrt{5}), \qquad & w_2 &= \frac{5}{24}, \end{aligned}$$

under the action of the coordinate permutations (34) and (13), (24). In particular, it has eight points. In conclusion, at least three interesting 3-cubature formulas for Δ_3 arise as projections of E_8 root system. The root system model explains the simple rational values of the weights.

The E_8 lattice is one of four widely studied and highly symmetric lattices in low dimensions; the other three are the Coxeter–Todd lattice K_{12} in $\mathbb{R}^{12}$, the Barnes–Wall lattice Λ_{16} in $\mathbb{R}^{16}$, and the Leech lattice Λ_{24} in $\mathbb{R}^{24}$ [4, Chap. 4]. In each case, the set of short vectors has transitive symmetry, and in each case, Sobolev's theorem establishes its degree as a spherical design.

The 756 short vectors of K_{12} form a 7-design on S^{11}. In one of its several presentations as an Eisenstein lattice in $\mathbb{C}^6$ (the "3-base" presentation [4, sect. 7.8]), the short vectors are generated from the two points

$$(1, 1, 1, 1, 1, 1), \qquad (1 - \omega, \omega - 1, 0, 0, 0, 0)$$

by freely permuting coordinates, multiplying the coordinates by powers of ω whose exponents sum to 0, and negating all coordinates. The projection τ_2 sends these points to a 16-point 3-cubature formula on Δ_5 generated from the points

$$\begin{aligned} \vec{p}_1 &= \frac{1}{2}(1, 1, 0, 0, 0, 0), \qquad & w_1 &= \frac{1}{42}, \\ \vec{p}_2 &= \frac{1}{6}(1, 1, 1, 1, 1, 1), \qquad & w_2 &= \frac{27}{42}, \end{aligned}$$

by freely permuting coordinates. This formula was found by Stroud [29, 30].

The 4320 short vectors of Λ_{16} form a 7-design on S^{15}. In its simplest position (which exhibits its Gaussian lattice structure), the short vectors are generated from the two vectors

$$\begin{gathered} (1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0), \\ (2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \end{gathered}$$

by permuting coordinates under the group $\mathrm{GL}(4,2) \ltimes (\mathbb{Z}/2)^4$ of affine automorphisms of $(\mathbb{Z}/2)^4$, together with sign changes that keep the coordinate sums divisible by 4. The projection τ_2 sends these points to a 51-point 3-cubature formula on Δ_7 generated from the points

$$\begin{aligned}
\vec{p}_1 &= (1,0,0,0,0,0,0,0), & w_1 &= \frac{1}{1080},\\
\vec{p}_2 &= \frac{1}{2}(1,1,0,0,0,0,0,0), & w_2 &= \frac{1}{270},\\
\vec{p}_3 &= \frac{1}{4}(1,1,1,1,0,0,0,0), & w_3 &= \frac{4}{135},\\
\vec{p}_4 &= \frac{1}{8}(1,1,1,1,1,1,1,1), & w_4 &= \frac{64}{135},
\end{aligned}$$

under the action of the affine group $\mathrm{GL}(3,2) \ltimes (\mathbb{Z}/2)^3$. This is not an optimal PI 3-cubature formula, because the orbit of $\vec{p}_2$ can be eliminated, leaving only 23 points. But it does have a novel property: Instead of full symmetrization, the orbit of $\vec{p}_3$ is in the pattern of the (8, 4, 3) Steiner system. This is as good as full symmetrization for 3-cubature, because any monomial of degree 3 involves at most three coordinates. The structure of this Barnes–Wall projection led the author to relate cubature to combinatorial t-designs and orthogonal arrays [18].

The above position of Λ_{16} is compatible with its Gaussian lattice structure. Rains found another interesting position which is compatible with an Eisenstein lattice structure. The corresponding 3-cubature formula on Δ_7 has 50 points. They are generated from

$$\begin{aligned}
\vec{p}_1 &= (1,0,0,0,0,0,0,0), & w_1 &= \frac{1}{720},\\
\vec{p}_2 &= \frac{1}{4}(1,1,1,1,0,0,0,0), & w_2 &= \frac{1}{90},\\
\vec{p}_3 &= \frac{1}{3}(1,1,0,0,1,0,0,0), & w_3 &= \frac{1}{80},\\
\vec{p}_4 &= \frac{1}{12}(4,0,4,0,1,1,3,3), & w_4 &= \frac{1}{60},\\
\vec{p}_5 &= \frac{1}{12}(4,0,4,0,1,1,1,1), & w_5 &= \frac{1}{40},\\
\vec{p}_6 &= \frac{1}{12}(3,1,3,1,1,1,1,1), & w_6 &= \frac{1}{30},\\
\vec{p}_7 &= \frac{1}{12}(3,1,1,1,1,1,3,1), & w_7 &= \frac{1}{30},
\end{aligned}$$

by the coordinate permutations (12), (13), (24), (57), (68) and (15), (26), (37), (48).

The 196560 short vectors of the Leech lattice form an 11-design on S^{23}. The lattice has a space Eisenstein lattice structure, which Conway and Sloane call the *complex Leech lattice* [4, sect. 7.8]. The complex basis that they give leads to a 5-cubature

formula on Δ_{11} generated by the points

$$\begin{aligned}
\vec{p}_1 &= \frac{1}{2}(1,1,0,0,0,0,0,0,0,0,0,0), & w_1 &= \frac{1}{10920},\\
\vec{p}_2 &= \frac{1}{6}(1,1,1,1,1,1,0,0,0,0,0,0), & w_2 &= \frac{9}{3640},\\
\vec{p}_3 &= \frac{1}{18}(7,1,1,1,1,1,1,1,1,1,1,1), & w_3 &= \frac{27}{1820},\\
\vec{p}_4 &= \frac{1}{18}(4,4,1,1,1,1,1,1,1,1,1,1), & w_4 &= \frac{27}{3640},
\end{aligned}$$

by the action of the Mathieu group M_{12}. In other words, the coordinates of $\vec{p}_2$ are permuted in the pattern of the (12, 6, 5) Steiner system, and the points of the other coordinate are permuted freely. The total is 276 points. Another interesting basis of the plane consists of the mutual eigenplanes of the $(\mathbb{Z}/5)^3$ subgroup of the isometry group of the Leech lattice. Rains has computed that the corresponding 5-cubature formula on Δ_{11} has 498 points, consisting of 22 orbits of the surviving coordinate permutations. However, since none of the Barnes–Wall formulas on Δ_7 is optimal, it is not clear that the smaller of these formulas is either.

3. Torus constructions. The constructions in the next section depend on an auxiliary case that generally works out better than cubature on S^n, $\mathbb{C}P^n$, and Δ_n, namely cubature on algebraic tori. There is a developed theory for a special case of this problem known as trigonometric cubature [6, 7]. We will describe a more general class of problems, with one new result for the classic trigonometric cubature problem (Theorem 3.3).

Consider a torus group $T \cong (S^1)^n$ together with a faithful linear action on some real vector space $V \cong \mathbb{R}^N$. Then we can identify T with any faithful orbit O to give it a real algebraic structure. Since T is a compact group, it also comes with Haar measure (i.e., uniform measure). Given both structures, we can then consider cubature formulas for T. If a cubature formula F is a t-design and forms a subgroup of T, then it is called a *lattice formula* or an *additive t-design*.

PROPOSITION 3.1. *The lattice cubature problem on T is equivalent to a lattice packing problem as follows:*

1. *The real algebraic structure on T does not depend on the orbit O or the base point chosen on O. The ring of polynomials on T is the same as the character ring $R(T)$.*
2. *Every character $\chi : T \to \mathbb{C}$ is homogeneous as a polynomial on T. Its degree defines a norm on $\widehat{T}$, the character group of T. The norm is generated by unit steps corresponding to the characters that appear in $V \otimes \mathbb{C}$.*
3. *The characters that are constant on a subgroup $F \subset T$ form a sublattice $\widehat{F} \subseteq \widehat{T}$. This correspondence is a bijection between finite subgroups and sublattices such that $|F| = [\widehat{T} : \widehat{F}]$.*
4. *The subgroup F is a t-design if and only if $\widehat{F}$ has minimum distance $d = t+1$.*

The proof of Proposition 3.1 is lengthy but routine and can be left as an exercise for the reader. It is essentially established in the literature when $T = T(\mathrm{SO}(2n))$ acts on $\mathbb{R}^{2n}$ by separate rotations in n orthogonal planes. This case is equivalent to the (cubic) trigonometric cubature problem, defined as cubature formulas on the n-cube $[0, 2\pi)^n$, which are exact for trigonometric polynomials of degree t [7]. All of the arguments generalize without change.

When $T = T(\mathrm{SO}(2n))$, $\widehat{T}$ is naturally identified with $\mathbb{Z}^n$, and its norm is the ℓ_1

or taxicab norm. Another torus of interest to us is $T = T(\mathrm{PSU}(n+1))$, the group of diagonal unitary matrices with determinant 1 modulo its center. It acts on $\mathbb{C}^{(n+1)^2}$, interpreted as the space of $(n+1) \times (n+1)$ complex matrices, by conjugation. In this case $\widehat{T} = A_n$, the root lattice of $\mathrm{PSU}(n+1)$, and its norm is defined by taking the roots of A_n as unit steps.

THEOREM 3.2. *Given a real algebraic torus T of dimension n, let $K \subset \widehat{T} \otimes \mathbb{R}$ be the real convex hull of the unit steps in $\widehat{T}$. Let $\delta_L(K)$ be the lattice packing density of K, and let $\operatorname{Vol} K$ be the volume of K normalized by $\widehat{T}$. Let $t \geq 0$ and let $d = t+1$. Then the best additive t-design F on T has at least*

$$\frac{d^n(\operatorname{Vol} K)}{2^n \delta_L(K)} \leq |F| \leq \frac{d^n(\operatorname{Vol} K)}{2^n \delta_L(K)}(1 + O(t^{-1}))$$

points.

Theorem 3.2 has been noted independently by several people for trigonometric cubature, but may originally be due to Frolov [11]. In outline, a lattice $\widehat{F} \subset \widehat{T}$ with minimum distance d produces a packing of the dilated body $\frac{d}{2}K$. The packing density $\delta_L(K)$ then yields a lower bound on the index of $\widehat{F}$. On the other hand, if Λ is the center lattice of the best packing of K, then when t is large, $\frac{d}{2}\Lambda$ can be approximated by a sublattice of $\widehat{T}$. This establishes the upper bound.

Note also that the best Λ has rational coordinates relative to $\widehat{T}$ (or they can be made rational if Λ is not unique), because K is a rational polytope. Thus there exist special distances d such that the best F has exactly

$$\frac{d^n(\operatorname{Vol} K)}{2^n \delta_L(K)}$$

points. Also if some d achieves exactitude, then so does kd for every $k > 1$.

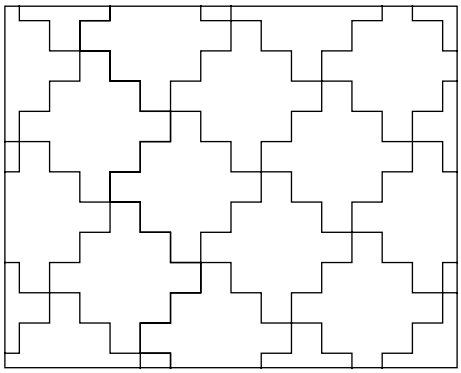
If $T = T(\mathrm{SO}(2n))$ is the standard cubic n-torus, then K is the n-cross polytope C_n^*. For example, Minkowski established that the lattice packing density of the regular octahedron C_3^* is $\frac{18}{19}$. Thus there exists an additive 5-design on $T(\mathrm{SO}(6))$ with 38 points [11, 23].

Since C_2^* is a square, its packing density is 1. Noskov [23] found the best discrete approximation to this packing for every distance d to obtain lattice rules for $T(\mathrm{SO}(4))$. If $d = 2s$, then the best approximation is exact and there is a $(2s-1)$-design with $2s^2$ points. If $d = 2s+1$, then the best approximation corresponds to the tiling of $\mathbb{Z}^2$ by the discrete ℓ_1 ball of radius s, or the tiling of the plane $\mathbb{R}^2$ by certain Aztec diamonds, as shown in Figure 5(a). The ball and the corresponding $2s$-design have $s^2 + (s+1)^2$ points.

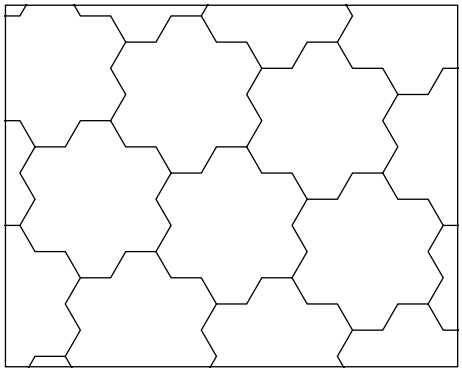
Noskov's designs have a counterpart for $T(\mathrm{PSU}(3))$, where $T(\widehat{\mathrm{PSU}(3)}) = A_2$ is the triangular lattice. If we identify A_2 with the Eisenstein integers $\mathbb{Z}[\omega]$, then the highest-density lattice with minimum distance d is the ideal generated by

$$\left\lfloor \frac{d}{2} \right\rfloor - \omega \left\lfloor \frac{d+1}{2} \right\rfloor.$$

When $d = 2s$, the dual $(d-1)$-design has $3s^2$ points and exactly matches the tiling of the plane by regular hexagons. When $d = 2s+1$, it has $3s^2 + 3s + 1$ points and corresponds to a tiling of the plane by the hexagonal polyhex of order s (an "Afghan hexagon"), as shown in Figure 5(b).



(a) Aztec diamonds for $T(\mathrm{SO}(4))$.



(b) Afghan hexagons for $T(\mathrm{PSU}(3))$.

FIG. 5. *Polyomino and polyhex tilings that lead to lattice rules.*

THEOREM 3.3. *Let $t \geq 0$. The torus $T(\mathrm{SO}(2n))$ has a $(2t+1)$-design with $O(n^t)$ points. More precisely it has a $2t$-design with $(2n)^t(1+o(1))$ points as $n \to \infty$ and a $(2t+1)$-design with twice as many points. The torus $T(\mathrm{PSU}(n+1))$ has a t-design with $n^t(1+o(1))$ points as $n \to \infty$.*

Remark 3.1. Theorem 3.3 can be compared with a prior result by Cools, Novak, and Ritter [5], who obtained negative interior (NI) t-cubature formulas for $T(\mathrm{SO}(2n))$ with $O(n^t)$ points. Another comparison is with the lower bound due to Stroud [31], Noskov [22], and Mysovskikh [20] for trigonometric $2t$-cubature:

$$|F| \geq \frac{(2n)^t(1-o(1))}{t!}.$$

The Möller bound applies to trigonometric $(2t+1)$-cubature in its interpretation as cubature on $T(\mathrm{SO}(2n))$ because it is a centrally symmetric algebraic variety. It yields the inequality

$$|F| \geq \frac{2(2n)^t(1-o(1))}{t!}.$$

Section 5 establishes an analogous lower bound for t-cubature on $T(\mathrm{PSU}(n))$ (Corollary 5.4):

$$|F| \geq \frac{n^t(1-o(1))}{\lceil t/2 \rceil! \lfloor t/2 \rfloor!}.$$

Thus for each t, Theorem 3.3 is asymptotically optimal to within a constant factor, even though the lower bounds do not require that F be positive or interior.

Proof. By Proposition 3.1, our task is to find suitable lattices in $\mathbb{Z}^n = T(\widehat{\mathrm{SO}(2n)})$ and $A_n = T(\widehat{\mathrm{PSU}(n+1)})$. Our task is fulfilled by Craig lattices [4, sect. 8.6] in A_n and skew analogues of Craig lattices in $\mathbb{Z}^n$. We describe the A_n case first.

We can model A_n as the set of points in $\mathbb{Z}^{n+1}$ with zero coordinate sum. Let $p \geq n+1$ be prime, and index the standard basis $\{\vec{e}_a\}$ of $\mathbb{Z}^{n+1}$ by some subset $N \subset \mathbb{Z}/p$. Let $p > t > 0$, and define a linear map $\phi : \mathbb{Z}^{n+1} \to (\mathbb{Z}/p)^t$ by

$$\phi(\vec{e}_a) = (a, a^2, \dots, a^t).$$

We define the lattice

$$\Lambda^{(t)}(A_n) = \ker \phi \cap A_n.$$

Plainly the index of $\Lambda^{(t)}(A_n)$ is at most $p^t = n^t(1+o(1))$. (If n is large and $p \approx n$, it is p^t, because any lower power of p would violate the Stroud-type bound.)

We claim that the distance of $\Lambda^{(t)}(A_n)$ is $t+1$. To show this, we will show that ϕ is injective on the simplex $\Delta_n^{(t)} \subset \mathbb{Z}_{\geq 0}^{n+1}$ of nonnegative vectors with coordinate sum t. A vector $\vec{x} \in A_n$ with root-step length at most t can be expressed as the difference of two vectors in $\Delta_n^{(t)}$; therefore injectivity shows that none of these vectors lie in $\Lambda^{(t)}(A_n)$.

We can interpret a vector $\vec{x} \in \Delta_n^{(t)}$ as a multiset of S over the set $\{0, \dots, n\}$ with $|S| = t$: if

$$\vec{x} = \sum_a m_a \vec{e}_a,$$

then m_a is the multiplicity of $a \in S$. In this interpretation, $\phi(\vec{x})$ is the list of power sums

$$\sum_{a \in S} a^k$$

for $1 \leq k \leq t$. By standard inversion formulas [28], these power sums determine the elementary symmetric functions of the elements of S when $p > t$, which are the coefficients of the polynomial

$$\prod_{a \in S} (x - a).$$

Thus $\phi(\vec{x})$ determines S as a multiset and the vector $\vec{x}$, and it is injective on $\Delta_n^{(t)}$.

For $\mathbb{Z}^n$ (with the ℓ_1 norm), let $p > 2n$ be prime. Index the standard basis $\{\vec{e}_a\}$ of $\mathbb{Z}^n$ by some subset $N \subset \mathbb{Z}/p$ such that N is disjoint from $-N$. Define $\phi : \mathbb{Z}^n \to (\mathbb{Z}/p)^t$ by

$$\widehat{\phi}(\vec{e}_a) = (a, a^3, a^5, \dots, a^{2t-1}),$$

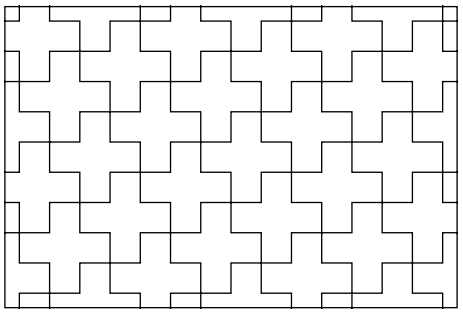


FIG. 6. *A Hamming-like "plus" tiling.*

and define

$$\Lambda^{(t)}(\mathbb{Z}^n) = \ker \widehat{\phi}.$$

Then the index of $\Lambda^{(t)}(\mathbb{Z}_n)$ is again at most (and usually exactly) $p^t = n^1(1 + o(1))$. Its distance property can be explained by embedding $\mathbb{Z}^n$ isometrically into A_{2n} using the map

$$\alpha : \vec{e}_a \mapsto \vec{e}_a - \vec{e}_{-a}.$$

Then

$$\Lambda^{(t)}(\mathbb{Z}^n) = \alpha^{-1}(\Lambda^{(2t)}(A_n)).$$

Since $\Lambda^{(2t)}(A_n)$ has distance at least $2t + 1$, so does $\Lambda^{(t)}(\mathbb{Z}^n)$. We can boost the distance to $2t + 2$ by passing to its even-sum sublattice. □

Remark 3.2. When $t = 1$, the number p in the proof of Theorem 3.3 need not be prime, and the lattices $\Lambda^{(1)}(\mathbb{Z}_n)$ and $\Lambda^{(1)}(A_n)$ produce lattice tilings of the ball of ℓ_1-radius 1 in $\mathbb{Z}^n$ and the combinatorial simplex $\Lambda^{(1)}$ in A_n. For example, when $n = 2$, these tilings are equivalent to familiar tilings of the plus pentomino (Figure 6) and the triangle trihex. The plus tiling resembles combinatorial tilings coming from Hamming codes [4, sect. 3.2]. More generally, Craig lattices resemble low-distance BCH codes. This resemblance is what led the author to Theorem 3.3.

4. Fibration constructions. The projection construction in section 2, while instructive, is backwards in a sense: It is harder to make t-cubature formulas for $\mathbb{C}P^{n-1}$ and S^{2n-1} than for Δ_{n-1} for most values of n and t. In this section we will use the same projections to lift cubature formulas to spheres and projective spaces from simplices. The construction also requires the definition and construction of cubature formulas on tori from section 3.

THEOREM 4.1. *Let $\alpha : X \to Y$ be one of the three projections h, π, or τ_2, and let T be a generic fiber. Let $s = 2t+1$ when $X = S^{2n-1}$ and $s = t$ when $X = \mathbb{C}P^{n-1}$. Given an interior (or boundary) t-cubature formula F for Y and an interior s-cubature formula F_T for T, there is a twisted product s-cubature formula $F_X = F_T \ltimes F_Y$ for X. It satisfies $|F_X| = |F_T|\ |F_Y|$ and it inherits positivity from its factors. In the boundary case, $|F_X| \le |F_T|\ |F_Y|$.*

Note that in the three cases, T is isomorphic to S^1, $T(\mathrm{SO}(2n))$, and $T(\mathrm{PSU}(n))$, respectively.

Proof. Let σ_Y be the discrete measure on Y corresponding to the cubature formula F_Y, and let $\sigma_X = \alpha^*(\sigma_Y)$ be the pull-back of σ_Y to X. In other words, for each point p of weight w in F_Y, σ_X has a term consisting of uniform measure on the torus fiber $\alpha^{-1}(p)$. Also let μ_X and μ_Y be normalized uniform measure on X and Y.

We claim that

$$\int_X P(\vec{x})d\mu_X = \int_X P(\vec{x})d\sigma_X$$

for any polynomial of P of degree s; in other words μ_X and σ_X are *s-cubature equivalent* [18]. If we assume the natural group structure on T, then T acts on X in each of the three cases with Y as the set of orbits. Then

$$\int_X P(\vec{x})d\sigma_X = \int_X P_T(\vec{x})d\sigma_X, \qquad \int_X P(\vec{x})d\mu_X = \int_X P_T(\vec{x})d\mu_X,$$

where P_T is the average of P with respect to the action of T. The polynomial P_T then descends to a polynomial P_Y on Y of degree t, and

$$\int_X P_T(\vec{x})d\sigma_X = P(F_Y), \qquad \int_X P_T(\vec{x})d\mu_X = \int_Y P_Y(\vec{y})d\mu_Y$$

because α preserves measure.

The measure σ_X evidently has a twisted product s-cubature formula $F_X = F_T \ltimes F_Y$ given by replacing each fiber by a copy of F_T. (A singular fiber corresponding to a boundary point of T can be replaced by a projection of F_T.) Since μ_X and σ_X are s-cubature equivalent, F_X is a cubature formula for μ_X as well. □

Remark 4.1. The proof of Theorem 4.1 is analogous to that of Sobolev's theorem with the finite group G replaced by the torus T. Indeed, the argument works for any compact group.

The simplest case of Theorem 4.1 is the Hopf map h. In this case the theorem says that a t-cubature formula F on $\mathbb{C}P^{n-1}$ lifts to a $(2t+1)$-cubature formula F' on S^{2n-1} with $(2t+2)|F|$ points. This relation was also observed by König [16].

Corollary 4.2. *The n-sphere S^n has a 7-cubature formula with $O(n^4)$ points for all n, more precisely $4n^4(1+o(1))$ points. The 3-sphere S^3 has a $(2s+1)$-cubature formula with*

$$|F| = \begin{cases} (s+1)(s^2+3), & s \text{ odd}, \\ (s+1)(s^2+s+2), & s \text{ even}, \end{cases}$$

points.

Proof. The simplex Δ_n has a 3-cubature formula with $O(n)$ points [18] constructed using Hadamard designs. This can be combined with the 7-design on $T(\mathrm{SO}(2n))$ with $O(n^3)$ points provided by Theorem 3.3, for a total of $O(n^4)$ points. More precisely, the formula on Δ_n has points at the corners, each of which lifts to $O(n)$ points; a point in the center, which lifts to $O(n^3)$ points; and $2n+o(n)$ points on $\lfloor n/2 \rfloor$-dimensional faces, each of which lift to $2n^3(1+o(1))$ points. Only the last family of points is significant, and it comprises $4n^4(1+o(1))$ points.

Noskov's formulas from section 3 include a $(2s+1)$-design on the square torus $T(\mathrm{SO}(4))$ with $2(s+1)^2$ points. When s is odd, this can be combined with the Gauss–Lobatto s-quadrature formula on the interval Δ_2 with $\frac{s+3}{2}$ points. Two of the fibers are circles and can be replaced by $2(s+1)$ points instead of $2(s+1)^2$ points.

The total is then $(s+1)(s^2+3)$ points. When s is even, it can be combined with the Gauss–Radau s-quadrature formula with $\frac{s+2}{2}$ points. In this case one fiber is a circle. □

Remark 4.2. The first part of Corollary 4.2 actually yields a 7-design on S^{n-1} with $O(n^6)$ points whenever there is a Hadamard matrix of order n. In this case the weights of the 3-cubature formulas on Δ_{n-1} are $\frac{2}{n(n+1)(n+2)}$ at the corners, $\frac{n}{2(n+1)(n+2)}$ at the faces, and $\frac{4n}{(n+1)(n+2)}$ at the center. Thus the weights are all commensurable up to a factor of $2n^2$ (note that n is even), and the cubature formula can be interpreted as a 7-design with this multiplicity factor. Moreover, copies of the lattice formulas on the torus fibers can be shifted to make the design multiplicity-free. Better yet, the design need have only $O(n^4)$ points if, for example, $n = 4 \cdot 7^k$. In this case the prime p used in the proof of Theorem 3.3 can be replaced by the prime power 7^{k+1}. The number of points on each fiber then compensates for all but a bounded part of the factor of $2n^2$ in the weights.

The previous best construction of 7-designs on S^{n-1} is due to Sidelnikov [26] and requires $O(2^{k(k+1)/2})$ points when $n = 2^k$.

A useful variant of Theorem 4.1 involves the moment map $\tau_2 : \mathbb{R}^{2n} \to \mathbb{R}^n$ defined by the same formula as $\tau_2 : S^{2n-1} \to \Delta_{n-1}$, namely,

$$\tau_2(x_1, \ldots, x_{2n}) = (x_1^2 + x_2^2, x_3^2 + x_4^2, \ldots, x_{2n-1}^2 + x_{2n}^2).$$

This τ_2 takes uniform measure on the ball B_n to uniform measure on the simplex

$$\Delta_n' = \Big\{ \vec{x} | x_k \geq 0, \sum x_k \leq 1 \Big\}.$$

When $n = 2$, Noskov's formulas together with some ad hoc cubature formulas for the triangle yield some economical formulas for the 4-ball B_4. For example, there is a PB 3-cubature formula on the triangle $x, y \geq 0,\ x + y \leq 1$ with points and weights generated from

$$\vec{p}_1 = \left(\frac{2}{5}, \frac{2}{5}\right), \qquad w_1 = \frac{25}{48},$$
$$\vec{p}_2 = \left(\frac{161 + 17\sqrt{14}}{1344}, 0\right), \qquad w_2 = \frac{16 - 2\sqrt{14}}{25}$$

by switching the coordinates and negating $\sqrt{14}$. This formula lifts to 1 generic fiber in B_4, which can be replaced with 32 points and 4 singular fibers, which are circles and can be replaced with 8 points each. The result is a PI 7-cubature formula on B_4 with 64 points.

Wandzura and Xiao [33] found competitive PI s-cubature formulas for s up to 30; Figure 7 shows one example. Most of these yield competitive PI $(2s+1)$-cubature formulas on B_4 and S^5. The formulas could probably be improved further with a search on the triangle that favors nodes on the edges.

The map $\tau_2 : \mathbb{R}^{2n} \to \mathbb{R}^n$ also takes Gaussian measure on $\mathbb{R}^{2n}$ to exponential measure on $\mathbb{R}^n_+$. For example, there is a PB exponential 4-cubature formula on $\mathbb{R}^2_+$ with points and weights generated from

$$\vec{p}_1 \approx (1.50766353, 1.50766353), \qquad w_1 \approx 0.354104443,$$
$$\vec{p}_2 \approx (6.29508677, 1.76717584), \qquad w_2 \approx 0.00876905581,$$
$$\vec{p}_3 \approx (0.285606152, 0), \qquad w_3 \approx 0.556110610,$$
$$\vec{p}_4 \approx (3.27491992, 0), \qquad w_4 \approx 0.0722468398$$

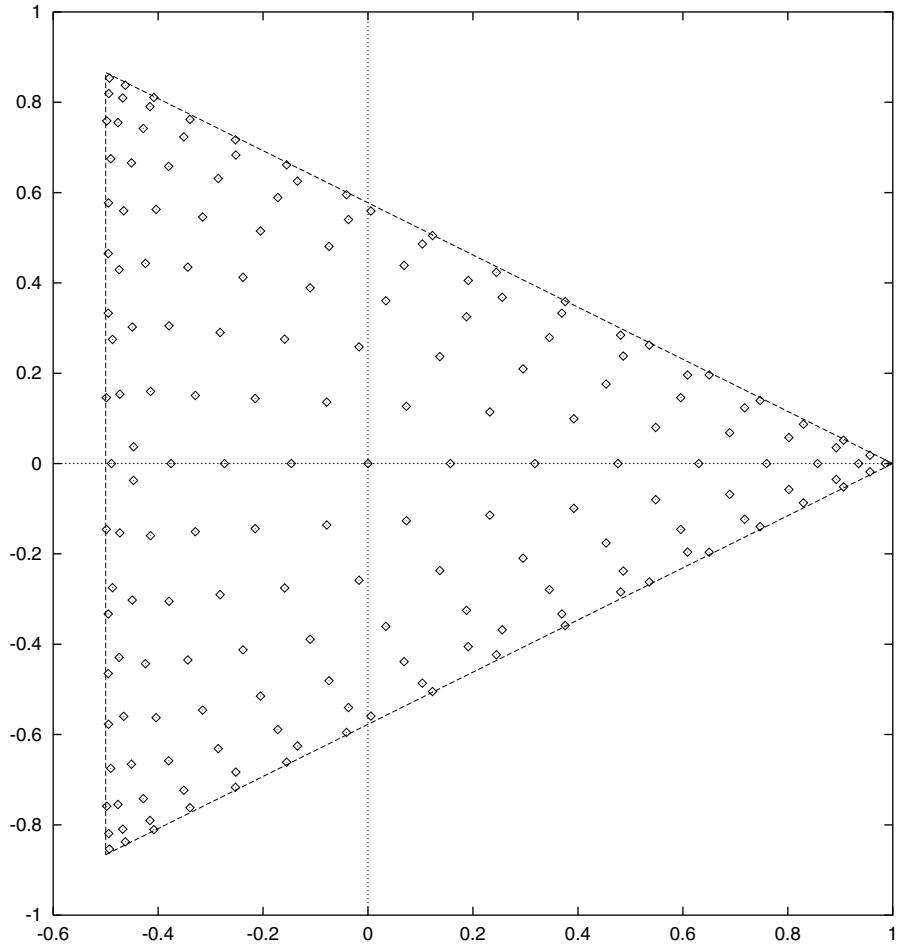


FIG. 7. *A* 175*-point* 30*-cubature formula found by Wandzura and Xiao* [33].

by switching the coordinates. It lifts to 3 generic fibers with 50 points each and 4 singular fibers with 10 points each. The result is a positive Gaussian 7-cubature formula on $\mathbb{R}^4$ with 190 points.

5. An algebraic lower bound. Let X be the Zariski closure of the support of a measure μ on $\mathbb{R}^n$, and let A be the ring of polynomial functions on X. (Recall that the *Zariski closure* of a set S, or closure in the Zariski topology, is the smallest algebraic variety containing S.) In other words, A is the quotient of $\mathbb{R}[\vec{x}\,]$ by the ideal I_X of polynomials that vanish on X. The ring A has a degree filtration coming from the degree filtration of $\mathbb{R}[\vec{x}\,]$. Stroud [31, 30] established an important lower bound on an arbitrary $2t$-cubature formula F for μ (not necessarily positive or interior), as follows.

THEOREM 5.1 (after Stroud). *If F is a $2t$-cubature formula for μ, then*

$$|F| \geq \dim A_{\leq t}.$$

Noskov [22] and Mysovskikh [20] observed that this applies to trigonometric cu-

bature by taking $X = T(\mathrm{SO}(2n))$. (According to Möller [19], the bound was also noted independently in special cases by other authors, e.g., Radon.)

Proof. Define a bilinear form

$$b : A_{\leq t} \times A_{\leq t} \to \mathbb{R}$$

by

$$b(P, Q) = \int_X P(\vec{x})Q(\vec{x})d\mu.$$

The form b is positive-definite because the integrand of $b(P, P)$ is nonnegative; moreover, if the integrand vanishes on X, then $P = 0$ as an element of A. Therefore b is nondegenerate, and its rank is $\dim A_{\leq t}$. On the other hand, the integrand lies in $A_{\leq 2t}$, so a $2t$-cubature formula F leads to the formula

$$b(P, Q) = \sum w_k P(\vec{p}_k)Q(\vec{p}_k).$$

This formula realizes b as a sum of $|F|$ rank 1 forms. Therefore $|F|$ is at least the rank of b, as desired. ∎

An interesting scholium of the proof of Theorem 5.1 is that if F is a $2t$-cubature formula, then its points suffice to interpolate polynomials on X of degree t.

It is curiously difficult to improve the Stroud bound for odd-degree cubature. However, the inference that lower bounds improve mainly in even degrees is not consistent with the Hopf fibrations

$$h : S^{2n+1} \to \mathbb{C}P^n, \qquad h : T(\mathrm{SO}(2n + 2)) \to T(\mathrm{PSU}(n + 1)).$$

On the one hand, these maps are quadratic and double the degree of cubature in passing from the target to the domain; in particular, they do not preserve odd and even. On the other hand, sections 2 and 4 together show that cubature in the domain and target are comparably difficult when $n \gg t$.

The Hopf fibration example suggests a generalization of Stroud's theorem involving group actions and degree doubling.

THEOREM 5.2. *Let μ be a measure on $\mathbb{R}^n$, and let X be the Zariski closure of its support. Let $Y \subset \mathbb{R}^k$ be another affine real algebraic variety on which a compact group G acts. Let A and B be the rings of complex-valued polynomials on X and Y, and suppose that there is a ring isomorphism*

$$A \xrightarrow{\cong} \mathrm{Inv}_G(B)$$

that doubles the filtration degree of A. Let V be a unitary representation of G, and define the filtered vector space

$$M = \mathrm{Inv}_G(B \otimes V).$$

If F is a t-cubature formula for μ, then

$$|F| \geq \frac{\dim M_{\leq t}}{\dim V}.$$

We will always take Y to be the coordinate ring of another algebraic variety Y, which is a principal G-bundle over X, such that the bundle projection $\alpha : Y \to X$

is quadratic. The A-module M can then be understood as the space of polynomial sections of a vector bundle E over X with fiber V. The sections in $M_{\le t}$ then behave like polynomials elements of A, except that their degrees are half-integers. If $Y = X$ and G is trivial, then E is the trivial line bundle and Theorem 5.2 reduces to Theorem 5.1. The hypotheses of Theorem 5.2 have been chosen so that the proof of Theorem 5.1 generalizes to the case when E is not trivial.

Proof. The vector space M (which is naturally an A-module) has an A-valued Hermitian inner product a induced by the Hermitian inner product on V. More precisely, let $\overline{V}$ be the representation conjugate to V and let

$$\overline{M} = \mathrm{Inv}_G(\overline{V} \otimes B)$$

be the corresponding conjugate of M. (Note that A and B are both self-conjugate by hypothesis.) Let

$$\varepsilon : V \otimes \overline{V} \longrightarrow \mathbb{C}$$

be the linearization of the standard Hermitian inner product on V, and let

$$m : B \otimes B \longrightarrow B$$

be the linearization of multiplication on B. Let a' be the composition

$$B \otimes V \otimes \overline{V} \otimes B \xrightarrow{I \otimes \varepsilon \otimes I} B \otimes B \xrightarrow{m} B.$$

We can restrict the domain to

$$M \otimes \overline{M} = \mathrm{Inv}_G(B \otimes V) \otimes \mathrm{Inv}_G(\overline{V} \otimes B).$$

Since the restricted domain is G-invariant, we can then restrict the target to A. Let a be this restriction of a'. Although given as a linear map on $M \otimes \overline{M}$, it can be reinterpreted as a Hermitian inner product on M. In more geometric terms, if M comes from a bundle E over X with fiber V, then $a(f, g)$ is the pointwise inner product of two sections f and g of E.

Note that a is positive-definite in the sense that

$$a(f, f)(\vec{x}) \ge 0$$

for all $\vec{x} \in X$, and if $a(f, f) = 0$, then $f = 0 \in M$. The rest of the proof follows that of Theorem 5.1: Define a complex-valued Hermitian inner product b on $M_{\le t}$ by

$$b(f, g) = \int_X a(f(\vec{x}), g(\vec{x}))d\mu.$$

Then b is also positive-definite, because a is positive-definite and μ is Zariski-dense in X. Thus, b has rank $\dim M_{\le t}$. A cubature formula F realizes b as a sum of $|F|$ terms of rank at most $\dim V$. □

We state the following three special cases of Theorem 5.2 as corollaries.

Corollary 5.3. *If $|F|$ is a $(2t+1)$-cubature formula on $\mathbb{C}P^n$, then*

$$|F| \ge \binom{n+t}{n}\binom{n+t+1}{n}.$$

Proof. Let Y be S^{2n+1}, let G be $S^1 \subset \mathbb{C}$, and let G act by complex rotation on $\mathbb{C}^{n+1} \supset S^{2n+1}$. The bundle projection $\alpha : Y \to X$ is the Hopf map h. Let $V = L_1$ be the tautological representation of S^1, so that E is the tautological line bundle on $\mathbb{C}P^n$. The space $M_{\leq 2t+1}$ is explicitly realized as the space of homogeneous polynomials in $\vec{z}$ and $\overline{\vec{z}}$ of bidegree $(t+1, t)$. The result follows by noting that

$$\dim M_{\leq 2t+1} = \binom{n+t}{n}\binom{n+t+1}{n}$$

and that $\dim L_1 = 1$. ∎

COROLLARY 5.4. *If $|F|$ is a t-cubature formula on $T(\mathrm{PSU}(n+1))$, then*

$$|F| \geq \frac{n^t(1-o(1))}{\lceil t/2 \rceil! \lfloor t/2 \rfloor!}$$

uniformly as $n \to \infty$.

Proof. Let Y be the torus $T(\mathrm{SO}(2n+2))$, and let $G = S^1$ again act by complex rotation in $\mathbb{C}^{n+1}$. In this case $M_{\leq 2t+1}$ is spanned by the space of monomials in $\vec{z}$ and $\overline{\vec{z}}$ of bidegree $(s+1, s)$ with $s \leq t$ and with the relation

$$z_k \overline{z}_k = 1$$

for all k. Its dimension is the number of points in the Minkowski difference

$$\Delta_n^{(t+1)} - \Delta_n^{(t)},$$

where $\Delta_n^{(t)}$ is the discrete simplex defined in the proof of Theorem 3.3. This is very similar to Theorem 5.1 for $2t$-cubature, because

$$\dim A_{\leq t} = |\Delta_n^{(t)} - \Delta_n^{(t)}|.$$

There is no concise formula for either number, but there is a concise estimate for fixed t in the limit $n \to \infty$. If E is either the trivial bundle when t is even or the bundle L_1 (restricted from $\mathbb{C}P^n$) when t is odd, then

$$\dim M_{\leq t} \approx \binom{n+1}{\lceil t/2 \rceil, \lfloor t/2 \rfloor, n+1-t} \approx \frac{n^t}{\lceil t/2 \rceil! \lfloor t/2 \rfloor!}$$

as $n \to \infty$, as desired. ∎

Remark 5.1. When F is a lattice formula, Corollary 5.4 is equivalent to Minkowski's classic upper bound on the density $\widehat{F}$ as a lattice packing of the discrete simplex $\Delta_n^{(t)}$. This and the fact that the Hopf fibration is quadratic led the author to Theorem 5.2.

COROLLARY 5.5. *If μ is a Zariski-dense measure on S^n and F is a $(2t+1)$-cubature formula for μ, then*

$$|F| \geq 2\binom{n+t}{t}.$$

Remark 5.2. Corollary 5.5 matches the Möller bound [19] for cubature on S^n, but it is more general because the measure μ need not be centrally symmetric.

Proof. The idea is to let $Y = \mathrm{Spin}(n+1)$ and $G = \mathrm{Spin}(n)$, where $\mathrm{Spin}(n)$ is the connected double cover of the Lie group $\mathrm{SO}(n)$. Then

$$X = \mathrm{Spin}(n+1)/\mathrm{Spin}(n) = \mathrm{SO}(n+1)/\mathrm{SO}(n) = S^n.$$

Our choice for the representation V is the spinor representation of $\mathrm{Spin}(n)$ when n is odd and a semispinor representation when n is even. The rest of the argument is a review of standard but nontrivial representation theory. We will borrow some specific calculations from Koike [15]. For an introduction to representation theory of Lie groups, see Varadarajan [32] or Fulton and Harris [12].

We divide the argument into steps. It is convenient to postpone the cases with $n \leq 3$ because of notational discrepancies.

1. If G is a compact, connected, and simply connected Lie group of rank n, then it has an irreducible unitary representation V (an irrep) for every list of nonnegative integers $(a_1, a_2, \ldots, a_n)$. The list is usually expressed in the vector form
$$\vec{\lambda} = a_1\vec{\lambda}_1 + a_2\vec{\lambda}_2 + \cdots + a_n\vec{\lambda}_n,$$
where $\vec{\lambda}$ is called the highest weight of V. The standard notation for this V is $V(\vec{\lambda})$.

   Both $\mathrm{Spin}(2n)$ and $\mathrm{Spin}(2n+1)$ are compact, connected, simply connected and have rank n when $n \geq 1$, except that $\mathrm{Spin}(2)$ is not simply connected. We will denote the irrep $V(\vec{\lambda})$ of $\mathrm{Spin}(n)$ by $V(n, \vec{\lambda})$.
2. When $n \geq 4$, the representation $V(n+1, t\vec{\lambda}_1)$ is realized as the space A_t of harmonic polynomials of degree t on S^n. The representation $V(2n+1, \vec{\lambda}_n)$ is the spinor representation of $\mathrm{Spin}(2n+1)$, and its dimension is 2^n. The representations $V(2n, \vec{\lambda}_{n-1})$ and $V(2n, \vec{\lambda}_n)$ are the semispinor representations of $\mathrm{Spin}(2n)$, and the dimension of each is 2^{n-1}.
3. The matrix entries of $V(n, \vec{\lambda})$ for all $\vec{\lambda}$ can be viewed as polynomials on $\mathrm{Spin}(n)$. (Indeed, this works for any compact Lie group.) Their span forms a ring which is generated by the entries of $V(n, \vec{\lambda}_k)$ for all k. We can define a degree filtration by giving every matrix entry of every $V(n, \vec{\lambda}_k)$ degree 2, except for spinor or semispinor representations, which are given degree 1.
4. For any irrep $V(n, \vec{\lambda})$, we can define a representation
$$M(n, \vec{\lambda}) = \mathrm{Ind}_{\mathrm{Spin}(n)}^{\mathrm{Spin}(n+1)} V(n, \vec{\lambda}).$$
There is more than one way to define representation induction in this context. One way is in terms of the algebraic structure in the previous step. In this case, $M(n, \vec{0})$ is the subring of polynomials on $\mathrm{Spin}(n+1)$ which are constant on left cosets of $\mathrm{Spin}(n)$, or in other words the ring of polynomials on
$$S^n = \mathrm{Spin}(n+1)/\mathrm{Spin}(n).$$
Each $M(n, \vec{\lambda})$ is not only a representation of $\mathrm{Spin}(n+1)$, but also a filtered module over $M(n, \vec{0})$. To apply Theorem 5.2, we let
$$M = M(n, \vec{\lambda}), \qquad A = M(n, \vec{0}), \qquad V = V(n, \vec{\lambda}).$$
5. The structure of $M(n, \vec{\lambda})$ can be computed by branching formulas for the restriction of an irrep of $\mathrm{Spin}(n+1)$ to $\mathrm{Spin}(n)$, together with induction-restriction duality. Induction-restriction duality says in general that if H is a subgroup of G, V is a unitary irrep of G, and W is a unitary irrep of H, then the number of times that W appears in the restriction $\mathrm{Res}_G^H V$ equals the number of times that V appears in the induced representation $\mathrm{Ind}_H^G W$.

The restriction formulas were computed by Koike [15, Thms. 11.2 and 11.3]. When $\vec{\lambda} = \vec{\lambda}_n$, Koike's formulas together with the degree filtrations yield the decompositions

$$M(2n, \vec{\lambda}_n)_{\le 2t+1} \cong \bigoplus_{s \le t} V(2n+1, s\vec{\lambda}_1 + \vec{\lambda}_n)$$

and

$$M(2n-1, \vec{\lambda}_n)_{\le 2t+1} \cong \bigoplus_{0 \le s \le t} \big(V(2n, s\vec{\lambda}_1 + \vec{\lambda}_{n-1}) \oplus V(2n, s\vec{\lambda}_1 + \vec{\lambda}_n)\big).$$

6. By the Weyl dimension formula,

$$\begin{aligned}
\dim V(2n+1, s\vec{\lambda}_1 + \vec{\lambda}_n) &= 2^n \binom{2n-1+s}{s}, \\
\dim V(2n, s\vec{\lambda}_1 + \vec{\lambda}_n) &= 2^{n-1} \binom{2n-2+s}{s}, \\
\dim V(2n, s\vec{\lambda}_1 + \vec{\lambda}_{n-1}) &= 2^{n-1} \binom{2n-2+s}{s}.
\end{aligned}$$

Combining the dimension formulas with the decomposition of $M(n, \vec{\lambda})_{\le 2t+1}$ yields

$$\begin{aligned}
|F| &\ge \frac{\dim M(n, \vec{\lambda})_{\le 2t+1}}{\dim V} \\
&= \sum_{0 \le s \le t} 2\binom{n-1+s}{s} = 2\binom{n+t}{t}.
\end{aligned}$$

This completes the proof of the result when $n \ge 4$.

The proof for $n \ge 3$ is not really different, but the notation for the representations changes because the groups involved have a simplified structure. Nonetheless, both Spin(1) and Spin(3) have a spinorial representation V, while Spin(2) has two semispinorial representations, and either one can be called V. In each case, A is the ring of polynomials on S^n with doubled degrees, the elements of V have degree 1, and

$$M = \operatorname{Ind}_{\mathrm{Spin}(n)}^{\mathrm{Spin}(n+1)} V.$$

The notation for the calculations changes, but the final answer is the same. ∎

Remark 5.3. The $n = 2$ case of Corollary 5.5 coincides with the $n = 1$ case of Corollary 5.3, and so does the proof. This was the original inspiration for Corollary 5.5.

6. A local lower bound. Our final application of moment maps is to help establish a local lower bound on the density of points of a PI or PB cubature formula on the simplex Δ_n. The bound was originally inspired by PI cubature formulas due to Wandzura and Xiao [33], which were found by simulated annealing. As in the example shown in Figure 7, the points in these formulas accumulate transversely at the edges of the triangle. Another related result is that the limiting density of the points of Gauss–Legendre quadrature (i.e., the zeros of Legendre polynomials) is $\frac{1}{\pi\sqrt{1-x^2}}$ [8].

This density can be interpreted as the linear projection of uniform measure on a circle, which is related to Archimedes' moment map (see Figure 9 below).

Theorem 6.1 and Corollary 6.2 establish a lower bound on the limiting density of any sequence of PI and PB formulas on Δ_n that generalizes the limiting density of Legendre zeros. Moreover, if the local density is high in certain regions, in particular near the vertices of Δ_n, then the weights there must be low. By this reasoning, Theorem 6.3 and Corollary 6.4 establish that a t-design on Δ_n requires many more points than an efficient t-cubature formula does as $t \to \infty$ (namely, $O(t^{2n})$ points versus $O(t^n)$ points). Along the way, Theorem 6.3 establishes that Gaussian quadrature for an arbitrary weight function is very sharply locally optimal among all positive quadrature formulas. Finally Scholium 6.5 generalizes the results for uniform measure on Δ_n to uniform measure on an arbitrary simple convex polytope.

THEOREM 6.1. *A PI or PB $2t-1$-cubature formula F on the simplex Δ_n is an ε-net with respect to the metric*

$$ds^2 = \frac{dx_0^2}{2x_0} + \frac{dx_1^2}{2x_1} + \cdots + \frac{dx_n^2}{2x_n}$$

in barycentric coordinates, where $\cos 2\varepsilon$ is the highest zero of the Jacobi polynomial $P_t^{(n-1,0)}(x)$.

In the proof and later, we will abbreviate $(n-1,0)$ as "#" in superscripts.

Proof. The idea of the proof is to find, for each $\vec{p} \in \Delta_n$ and each $\varepsilon' > \varepsilon$, a P of degree $2t-1$ on Δ_n such that

$$\int_{\Delta_n} P(\vec{x}) d\vec{x} > 0$$

but $P(\vec{x}) > 0$ only when $\vec{x} \in \Delta_n$ is in the ε'-ball $B_{\varepsilon'}(\vec{p})$ around $\vec{p}$. We can call this ball the *positive island* of $P(\vec{x})$; see Figure 8. The existence of such a polynomial P forces F to have an evaluation point in $B_\varepsilon(\vec{p})$, for otherwise $P(F) \le 0$.

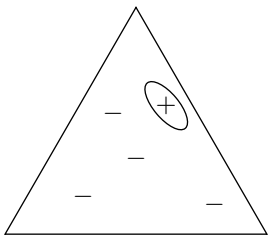


FIG. 8. *A polynomial on Δ_n with a small positive island.*

We first claim that the stated metric is the distance between fibers of the moment map π with respect to the Fubini–Study metric on $\mathbb{C}P^n$. To see this it suffices to check the following: The real locus $\mathbb{R}P^n \subset \mathbb{C}P^n$ is perpendicular to the fibers of π and meets each generic fiber 2^n times. Indeed, π is a bijection on the orthant $\mathbb{R}P^n_{\ge 0}$ with non-negative projective coordinates. Moreover, $\mathbb{R}P^n_{\ge 0}$ is isometric to the orthant $S^n_{\ge 0}$ of a unit n-sphere, and the restriction of π agrees with the restriction of the Xu map τ_1. The metric ds^2 on Δ_n is exactly the push-forward of the standard metric on $S^n_{\ge 0}$ under τ_1. See Figure 9 for an example.

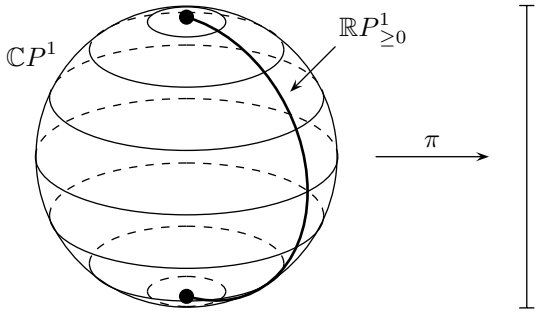


FIG. 9. *The moment map π restricts to the Xu map τ_1.*

Consider the linear projection $\alpha : \Delta_n \to [-1, 1]$ given by

$$\alpha(\vec{x}) = 2x_0 - 1. \tag{1}$$

The map α sends uniform measure on $\mathbb{C}P^n$ to the measure

$$\mu(x) = n2^{1-n}(1-x)^{n-1}.$$

The tth orthogonal polynomial with respect to this measure μ is the Jacobi polynomial $P_t^{\#} = P_t^{(n-1,0)}$. Let $p_t^{\#}$ be its highest zero.

Define a polynomial $Q_\delta : \mathbb{C}P^n \to \mathbb{R}$ by

$$Q_\delta(\vec{z}) = Q_\delta(x) = \frac{P_t^{\#}(x)^2(x - p_t^{\#} + \delta)}{(x - p_t^{\#})^2}, \tag{2}$$

where $x = \alpha(\pi(\vec{z}))$ and $\delta > 0$. It has degree $2t - 1$ as a polynomial in x, as well as a polynomial on $\mathbb{C}P^n$. Moreover, Q_δ vanishes at the zeros of $P_t^{\#}$, except at the highest zero, at which its value is positive. Therefore by Gaussian quadrature (!) with respect to the measure μ,

$$\int_{\mathbb{C}P^n} Q_\delta(\vec{z})d\vec{z} = \int_{-1}^{1} Q_\delta(x)d\mu > 0.$$

At the same time, Q is nonpositive outside of the region

$$x > p_t^{\#} - \delta.$$

This region corresponds to the ball of radius ε' around $(1 : 0 : 0 : \cdots : 0)$, with

$$2(\cos\varepsilon')^2 - 1 = \cos 2\varepsilon' = p_t^{\#} - \delta.$$

This can be confirmed by comparing with the orthant $\mathbb{R}P^n_{\geq 0}$ mentioned previously. Note that $\varepsilon' \to \varepsilon$ as $\delta \to 0$.

Given $\vec{q} \in \mathbb{C}P^n$, define $Q_{\delta,\vec{q}}$ by rotating Q_δ by some isometry of $\mathbb{C}P^n$ that takes $(1 : 0 : 0 : \cdots : 0)$ to $\vec{q}$. Define $Q^T_{\delta,\vec{p}} : \Delta_n \to \mathbb{R}$, where $\vec{p} = \pi(\vec{q})$, by averaging Q over torus fibers:

$$Q^T_{\delta,\vec{p}}(\vec{x}) = \frac{1}{|\pi^{-1}(\vec{x})|} \int_{\pi^{-1}(\vec{x})} Q_{\delta,\vec{p}}(\vec{z})d\vec{z}.$$

Then

$$\int_{\Delta_n} Q^T_{\delta,\vec{p}}(\vec{x})d\vec{x} = \int_{\mathbb{C}P^n} Q_{\delta,\vec{p}}(\vec{z})d\vec{z} > 0,$$

and $Q^T_{\delta,\vec{p}}$ is nonpositive outside of the ball of radius ε' around $\vec{p} = \pi(\vec{q})$ in the induced metric on Δ_n. Thus, $Q^T_{\delta,\vec{p}}$ has the desired properties. □

Remark 6.1. A somewhat weaker version of Theorem 6.1 holds when F is positive and exterior, but with real evaluation points. Polynomials similar to Q^T_δ can be constructed directly as products of factors that vanish on quadratic surfaces in $\mathbb{R}^n \supset \Delta_n$, with only one unsquared factor that vanishes on the boundary of $B_{\varepsilon'}(\vec{p})$. As it happens, the boundary of $B_{\varepsilon'}(\vec{p})$ is a quadratic surface. We did not refine this sketched argument into a proof with explicit estimates.

COROLLARY 6.2. *Any sequence of PI or PB t-cubature formulas on Δ_n has limiting point density $\Omega(t^n \prod_k x_k^{-1/2})$, where $\vec{x} \in \Delta_n$ is fixed and given in barycentric coordinates, and $t \to \infty$.*

Proof. The corollary follows from computing the volume form corresponding to the metric ds^2 in the statement of Theorem 6.1 and estimating the covering radius ε. The asymptotic behavior of zeros of Jacobi polynomials is given in Abramowitz and Stegun [1, p. 787]. The key step in the estimate is the limit

$$\lim_{t\to\infty} \frac{P_t^{(a,b)}(\cos\frac{\theta}{t})}{P_t^{(a,b)}(1)} = 2^a \theta^{-a} a! J_a(\theta), \tag{3}$$

where $J_a(z)$ is the ordinary Bessel function of the first kind. Convergence to the limit is analytic in θ. Thus

$$\lim_{t\to\infty} t\theta^{(a,b)}_{t,t+1-k} = j_{a,k}$$

for every fixed k, where $\cos\theta^{(a,b)}_{t,k}/t$ is the kth zero of $P_t^{(a,b)}(x)$ and $j_{a,k}$ is the kth zero of $J_a(x)$. The estimate can be established directly in our geometry by noting that $P_t^{\#}(2|z_0|^2 - 1)$ is a harmonic function on $\mathbb{C}P^n$. The harmonic equation on $\mathbb{C}P^n$ is then locally approximated by the harmonic (or Helmholtz) equation on $\mathbb{R}^{2n}$, whose radial solutions are derived from Bessel functions.

In our case,

$$\cos 2\varepsilon = \cos\frac{\theta^{\#}_{t,t}}{t}$$

for $(2t-1)$ cubature. Thus

$$\varepsilon = \frac{j_{n-1,1}}{2t}(1 + o(1)) = \Theta(t^{-1}),$$

which is also $\Theta(t^{-1})$ for t-cubature. □

THEOREM 6.3. *Let μ be an arbitrary normalized measure on $\mathbb{R}$ whose support has at least $2t$ points. Let $p_1, \ldots, p_t$ and $w_1, \ldots, w_t$ be the points and weights of Gaussian t-quadrature for the measure μ. Let F be a positive t-quadrature formula for μ. Then for each $1 \le k \le t$, F has at least one point in the half-open interval $(p_{k-1}, p_k]$, where $p_0 = -\infty$. Moreover, the total weight of all points in $(-\infty, p_1]$ is at most w_1, with equality if and only if F is the Gaussian quadrature formula.*

Note that Theorem 6.3 is further sharpened by symmetry: F must also have at least one point in each half-open interval $[p_k, p_{k+1})$, with $p_{t+1} = \infty$, and its total weight in $[p_t, \infty)$ is at most w_t.

Proof. Let $\phi_t(x)$ be the tth orthonormal polynomial with respect to μ (with either sign), and let A_t be the leading coefficient of $\phi_t(x)$. If $k = 1$, let

$$P(x) = \frac{\phi_t(x)^2(p_1 + \delta_1 - x)}{(x - p_1)^2}$$

with $\delta_1 > 0$. If $k > 1$, let

$$P(x) = \frac{\phi_t(x)^2(x - p_{k-1} - \delta_0)(p_k + \delta_1 - x)}{(x - p_{k-1})^2(x - p_k)^2}$$

with $\delta_1 \gg \delta_0 > 0$. In both cases,

$$\int_{\mathbb{R}} P(x)d\mu > 0$$

by Gaussian quadrature, while P is positive only on the interval $(p_{k-1} + \delta_0, p_k + \delta_1)$. Therefore F has at least one point in this interval. Since F has only finitely many points, the limit $\delta_1 \to 0$ establishes that F has a point in $(p_{k-1}, p_k]$.

For the second claim, let

$$P(x) = \frac{\phi_t(x)^2}{(x - p_1)^2}.$$

Then by Gaussian quadrature,

$$\int_{\mathbb{R}} P(x)d\mu = w_1 P(p_1).$$

Let $q_1, \ldots, q_k$ be the points of F which are at most p_1, and let $v_1, \ldots, v_k$ be their weights. Then

$$\int_{\mathbb{R}} P(x)d\mu = P(F) \geq \sum_{j=1}^{k} w_j P(q_j) \geq P(p_1) \sum_{j=1}^{k} w_j.$$

The first inequality holds because P is nonnegative, the second because P decreases on $(-\infty, p_1]$. ∎

COROLLARY 6.4. *The least weight of any positive t-cubature formula on Δ_n (with uniform measure) is $O(t^{-2n})$, uniformly in t. Any t-design on Δ_n has $\Omega(t^{2n})$ points.*

Proof. If F is a t-cubature formula on Δ_n, the map α (see (1)) sends it to a t-quadrature formula $\alpha(F)$ on $[-1, 1]$ with Jacobi-polynomial measure. If F is positive, then the least weight of $\alpha(F)$ is at least that of F. On the other hand, Theorem 6.3 establishes that the least weight $\alpha(F)$ is at least the last Christoffel weight w_t.

The first claim follows by estimating this weight. One of the standard formulas for the general Christoffel weight w_k is

$$w_k = -\frac{A_{t+1}||\phi_t(x)||_\mu^2}{A_t \phi_t'(p_k)\phi_{t+1}(p_k)},$$

where $\phi_t(x)$ is the tth orthogonal polynomial, A_t is its leading coefficient, and p_k is its kth root. In our case, $\phi_t = P_t^\#$, $p_k = p_{t,k}^\#$, and $k = t$. We compute

$$\|P_t^\#\|_\mu^2 = \frac{n}{2t+n} = \Theta(t^{-1}),$$
$$A_t = 2^{-t}\binom{n-1+2t}{t} = \Theta(2^t).$$

To estimate $(P_t^\#)'(p_{t,t}^\#)$ and $P_{t+1}^\#(p_{t,t}^\#)$, we again appeal to the limit in (3). Differentiating both sides by θ, we obtain

$$\lim_{t\to\infty} -\frac{(P_t^\#)'(\cos\theta/t)(\sin\theta/t)}{tP_t^\#(1)} = -a2^a\theta^{-a}a!J_a'(\theta).$$

Note that $P_t^\#(1) = \binom{t+n-1}{t} = \Theta(t^{n-1})$. For a fixed value of θ, the various parts of the limit yield

$$(P_t^\#)'\left(\cos\frac{\theta}{t}\right) = \Theta(t^{n+1}).$$

By the same token

$$(P_t^\#)'\left(\cos\frac{\theta_t}{t}\right) = \Theta(t^{n+1})$$

when θ_t approaches a fixed value of θ, as is the case when $\theta_t = \theta_{t,t}^\#$ is given by

$$p_{t,t}^\# = \cos\frac{\theta_{t,t}^\#}{t}.$$

By a similar calculation,

$$(P_{t+1}^\#)(p_{t,t}^\#) = \Theta(t^{n-2}).$$

The conclusion is that

$$w_t = \Theta(t^{2n}),$$

as desired. □

SCHOLIUM 6.5. *Let $K \subset \mathbb{R}^n$ be a convex n-polytope with N facets. Let F be a t-cubature formula on K with uniform measure. If F is PI or PB and if K is simple, then F is an ε-net with respect to the metric*

$$ds^2 = \frac{dx_0^2}{x_1} + \frac{dx_2^2}{x_2} + \cdots + \frac{dx_N^2}{x_N},$$

where $\varepsilon = O(1/t)$. If F is positive, then its least weight is $O(t^{-2n})$. If it is a t-design, then it has at least $O(t^{2n})$ points.

Proof sketch. The proof of Theorem 6.1 retains its strength if $K \subseteq \Delta_n$ and we pass from Δ_n to K, provided that the positive island of the polynomial $Q_{\delta,\vec{p}}^T$ lies within K. In this case

$$\int_K Q_{\delta,\vec{p}}^T(\vec{x})d\vec{x} \geq \int_{\Delta_n} Q_{\delta,\vec{p}}^T(\vec{x})d\vec{x} > 0.$$

In order to properly position $Q^T_{\delta,\vec{p}}$ for all $\vec{p} \in K$, we need several embeddings of K into Δ_n. For each vertex $\vec{x} \in K$, choose a linear embedding L that sends $\vec{x}$ to some vertex of Δ_n, and that sends the facets incident to $\vec{x}$ to facets of Δ_n. (Equivalently, for each vertex $\vec{x} \in K$, choose a simplex $L^{-1}(\Delta_n) \supseteq K$ whose facets include all facets of K that meet at $\vec{x}$. See Figure 10.) Then there exists a finite set of L such that the positive islands of polynomials of the form $Q^T_{\delta,vq} \circ L$ together cover K. The formula F must have a point in each island, which establishes that F is an ε-net.

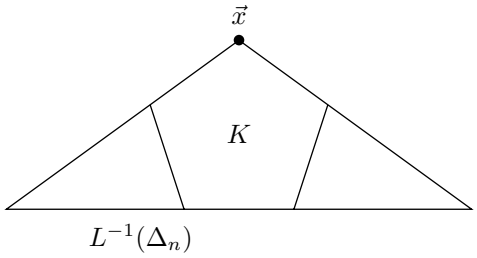


FIG. 10. *A simplex $L^{-1}(\Delta_n)$ whose facets contain facets of K that meet at $\vec{x}$.*

Similarly, the proofs of Theorem 6.3 and Corollary 6.4 retain their strength if a uniform measure on K projects by a map α to a measure ν on $[-1, 1]$ which is dominated by

$$\mu(x) = 2^{-n} n (1 - x)^{n-1}$$

and which agrees with μ in a neighborhood of 1. (Of course, μ cannot dominate ν if ν is normalized, so this condition on ν must be dropped.) In this case

$$\int_{\mathbb{R}} P(x) d\nu \geq \int_{\mathbb{R}} P(x) d\mu > 0$$

for the first half of Theorem 6.3 for the interval $(-\infty,\ p_1)$, while

$$\int_{\mathbb{R}} P(x) d\nu \leq \int_{\mathbb{R}} P(x) d\mu$$

for the second half of Theorem 6.3. A suitable projection α can be realized by positioning K in Δ_n so that it touches the vertex $x_0 = 1$, and then restricting the usual map α to K. □

7. Other comments. In this article we have studied the toric moment map on $\mathbb{C}P^n$ and on $\mathbb{C}^n$ restricted to S^{2n-1} (which can be interpreted as the level surface of an invariant Hamiltonian on $\mathbb{C}^n$) as it applies to the cubature problem. Many of the constructions apply equally well to arbitrary toric varieties. To begin with, every complex projective variety X inherits both a metric and an affine real structure from $\mathbb{C}P^n$. If X is toric, it also has a volume-preserving moment map whose image is a centrally symmetric polytope. However, the variety X rarely has much more symmetry than its moment map image.

The duality between toric cubature (particularly trigonometric cubature) and lattice packings explored in section 3 suggests a different limit of the cubature problem. Let $K \subset \mathbb{R}^n$ be a centrally symmetric convex body. For simplicity let $F = \{\vec{p}_a\}$ be a periodic discrete subset of $\mathbb{R}^n$ with a periodic weight function $\vec{p}_a \mapsto w_a$. Since

it is periodic, it has a well-defined Fourier transform $\widehat{F}$. In this setting, F is a Fourier K-cubature formula if and only if $\widehat{F}$ is disjoint from the interior of K. The (continuous) Fourier cubature problem is to minimize the density of F among all K-cubature formulas or all positive K-cubature formulas. If F is a lattice with equal weights, then $\widehat{F}$ has the same property, and the Fourier K-cubature problem reduces to finding the best lattice packing of K. It would be interesting to find examples of nonlattice formulas that are better than the best lattice formula.

We conjecture that a version of Corollary 5.5 holds, using Theorem 5.2 and the same spinor bundles, for any centrally symmetric subvariety $X \subset S^n$. That is, we conjecture Möller's bound for these varieties, even when the measure μ is not centrally symmetric.

Theorem 6.1 shows why some tempting approaches to constructing efficient PI or PB formulas on the simplex Δ_n, even the triangle Δ_2, are bound to fail. For example, if the points of a putative cubature formula F are fixed in advance, the question of whether it admits nonnegative weights for t-cubature reduces to linear programming. But if the points are arranged in some lattice with spacing $1/k$, Theorem 6.1 shows that the weights can only be nonnegative if $k = \Omega(t^2)$, so that $|F| = \Omega(t^{2n})$.

We believe that the requirement that K be simple in Scholium 6.5 is not essential. More generally we conjecture that similar results hold if K is not convex. We also conjecture that the bounds in Theorem 6.1 and Scholium 6.5 are sharp to within a constant factor. The cubature formulas found by Wandzura and Xiao support this conjecture, at least when $K = \Delta_n$.

The proof of Theorem 6.1 was partly inspired by the linear programming method for bounding kissing numbers, t-designs, and sphere packings [3, 9, 10, 14, 24]. Xu observed that the method for t-designs also yields bounds on PI t-cubature [36]. In fact, it yields an upper bound on the ℓ_2 norm of the weights of a PI t-cubature formula, which implies a lower bound on the number of points. We conjecture that linear programming methods could be used to improve the constants in Theorem 6.1.

Krylov [17] established that if $\{F_t\}$ is a sequence of interior t-cubature formulas for a measure μ, then $\{f(F_t)\}$ converges to $\int_X f(\vec{x})d\mu$ for every continuous f if and only if the ℓ_1 norm of the weights of F_t is bounded as $t \to \infty$. We conjecture then that Theorem 6.1 still holds, assuming a bound on the ℓ_1 norm of the coefficients of F instead of assuming that F is positive.

Acknowledgments. The author would like to thank Yael Karshon, Włodzimierz Kuperberg, Thomas Strohmer, and Hong Xiao for useful discussions. The author would especially like to thank Noam Elkies and Eric Rains, who obtained some of the examples and provided many other useful comments.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Appl. Math. Ser. 55, U.S. Government Printing Office, Washington, DC, 1964.

[2] Archimedes of Syracuse, *On the Sphere and Cylinder*, ca. 225 BC.

[3] H. Cohn and N. Elkies, *New upper bounds on sphere packings.* I, Ann. of Math. (2), 157 (2003), pp. 689–714, arXiv:math.MG/0110009.

[4] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., Grundlehren Math. Wiss. 290, Springer-Verlag, New York, 1993.

[5] R. Cools, E. Novak, and K. Ritter, *Smolyak's construction of cubature formulas of arbitrary trigonometric degree*, Computing, 62 (1999), pp. 147–162.

[6] R. Cools and J. N. Lyness, *Three- and four-dimensional K-optimal lattice rules of moderate trigonometric degree*, Math. Comp., 70 (2001), pp. 1549–1567.

[7] R. Cools and I. H. Sloan, *Minimal cubature formulae of trigonometric degree*, Math. Comp., 65 (1996), pp. 1583–1600.
[8] J. S. Dehesa, *Orthogonal polynomials in transport theories*, J. Phys. A, 14 (1981), pp. 297–302.
[9] P. Delsarte, *Bounds for unrestricted codes, by linear programming*, Philips Res. Rep., 27 (1972), pp. 272–289.
[10] P. Delsarte, J. M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geom. Dedicata, 6 (1977), pp. 363–388.
[11] K. K. Frolov, *The connection of quadrature formulas and sublattices of the lattice of integer vectors*, Dokl. Akad. Nauk SSSR, 232 (1977), pp. 40–43.
[12] W. Fulton and J. Harris, *Representation Theory*, 3rd ed., Grad. Texts Math. 129, Springer-Verlag, New York, 1998.
[13] I. D. Ivanović, *Formal state determination*, J. Math. Phys., 24 (1983), pp. 1199–1205.
[14] G. A. Kabatjanskiĭ and V. I. Levenšteĭn, *Bounds for packings on the sphere and in space*, Problemy Peredači Informacii, 14 (1978), pp. 3–25.
[15] K. Koike, *Representations of spinor groups and the difference characters of* $\mathrm{SO}(2n)$, Adv. Math., 128 (1997), pp. 40–81.
[16] H. König, *Cubature formulas on spheres*, in Advances in Multivariate Approximation (Witten-Bommerholz, 1998), Math. Res. 107, Wiley-VCH, Berlin, 1999, pp. 201–211.
[17] V. I. Krylov, *Approximate Calculation of Integrals*, Translated by A. H. Stroud, Macmillan, New York, 1962.
[18] G. Kuperberg, *Numerical cubature using error-correcting codes*, SIAM J. Numer. Anal., 44 (2006), pp. 897–907; also arXiv:math.NA/0402047.
[19] H. M. Möller, *Lower bounds for the number of nodes in cubature formulae*, in Numerische Integration (Tagung, Math. Forschungsinst., Oberwolfach, 1978), Internat. Ser. Numer. Math., 45, Birkhäuser, Basel, 1979, pp. 221–230.
[20] I. P. Mysovskikh, *Cubature formulas that are exact for trigonometric polynomials*, Dokl. Akad. Nauk SSSR, 296 (1987), pp. 28–31.
[21] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, UK, 2000.
[22] M. V. Noskov, *Cubature formulas for the approximate integration of periodic functions*, Metody Vychisl., 185 (1985), pp. 15–23.
[23] M. V. Noskov, *Formulas for the approximate integration of periodic functions*, Metody Vychisl., 178 (1988), pp. 19–22.
[24] A. M. Odlyzko and N. J. A. Sloane, *New bounds on the number of unit spheres that can touch a unit sphere in n dimensions*, J. Combin. Theory Ser. A, 26 (1979), pp. 210–214.
[25] E. Rains, *personal communication*, University of California, Davis, 2003.
[26] V. M. Sidelnikov, *Spherical* 7*-designs in* 2^n*-dimensional Euclidean space*, J. Algebraic Combin., 10 (1999), pp. 279–288.
[27] S. L. Sobolev, *Cubature formulas on the sphere which are invariant under transformations of finite rotation groups*, Dokl. Akad. Nauk SSSR, 146 (1962), pp. 310–313.
[28] R. P. Stanley, *Enumerative Combinatorics*, vol. 2, Cambridge University Press, Cambridge, UK, 1999.
[29] A. H. Stroud, *Some approximate integration formulas of degree* 3 *for an n-dimensional simplex*, Numer. Math., 9 (1966), pp. 38–45.
[30] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[31] A. H. Stroud, *Quadrature methods for functions of more than one variable*, Ann. New York Acad. Sci., 86 (1960), pp. 776–791.
[32] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Grad. Texts in Math. 102, Springer-Verlag, New York, 1984.
[33] S. Wandzura and H. Xiao, *Symmetric quadrature rules on a triangle*, Comput. Math. Appl., 45 (2003), pp. 1829–1840.
[34] W. K. Wootters and B. D. Fields, *Optimal state-determination by mutually unbiased measurements*, Ann. Phys., 191 (1989), pp. 363–381.
[35] Y. Xu, *Orthogonal polynomials and cubature formulae on spheres and on simplices*, Methods Appl. Anal., 5 (1998), pp. 169–184.
[36] Y. Xu, *Lower bound for the number of nodes of cubature formulae on the unit ball*, J. Complexity, 19 (2003), pp. 392–402.
[37] *Magma*, http://magma.maths.usyd.edu.au/.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 936–948

HOW DATA DEPENDENT IS A NONLINEAR SUBDIVISION SCHEME? A CASE STUDY BASED ON CONVEXITY PRESERVING SUBDIVISION*

THOMAS POK-YIN YU†

This paper is dedicated to the memory of Wong Suk Ling (1935–2004)

Abstract. The regularity of the limit function of a linear subdivision scheme is essentially irrelevant to the initial data. How data dependent, then, is the regularity of the limit of a nonlinear subdivision scheme? The answer is the most obvious—*it depends*. In this paper, we prove that the nonlinear convexity preserving subdivision scheme developed independently by Floater and Micchelli [M. S. Floater and C. A. Micchelli, *Approximation Theory*, Marcel Dekker, New York, 1998, pp. 209–224] and Kuijt and van Damme [F. Kuijt and R. van Damme, *Constr. Approx.*, 14 (1998), pp. 609–630] exhibits a rather strong nonlinear, data-dependent, behavior: For any $\nu \in (1, 2)$, there exists initial convex data such that the critical Hölder regularity of the limit curve is exactly ν. We also show that the limit function of any initial data always has Hölder regularity less than 2, unless if restricted to a subset of the domain at which the initial data is sampled from the convex branch of a rational polynomial of degree 2 over degree 1.

This result stands in contrast to what is reported in several recent publications on nonlinear subdivision schemes [I. Ur Rahman, I. Drori, V. C. Stodden, D. L. Donoho, and P. Schröder, *SIAM J. Multiscale Modeling and Simulation*, submitted, 2005; G. Xie and T. P.-Y. Yu, *Constr. Approx.*, 22 (2005), pp. 219–254; G. Xie and T. P.-Y. Yu, *Advances in Constructive Approximation*, 2004, pp. 519–533; I. Daubechies, O. Runborg, and W. Sweldens, *Constr. Approx.*, 3 (2004), pp. 399–463; T. P.-Y. Yu, *Cutting corners on the sphere*, preprint available at http://www.rpi.edu/∼yut/Papers/CuttingCorners.pdf, 2005], in which various families of nonlinear subdivision schemes are either proved or empirically observed to have rather weak nonlinearity in the sense that they produce limit curves with smoothness insensitive to initial data.

Key words. subdivision/refinement scheme, nonlinear subdivision scheme, convexity preserving subdivision, Hölder regularity, homogeneous map, real projective plane

AMS subject classifications. 26A15, 26A16, 26A18, 41A05, 42C40

DOI. 10.1137/050628751

1. Introduction.

Subdivision is a method for taking coarsely described data and recursively generating (typically smooth) data at finer and finer resolution. It can be used to rapidly generate curves and surfaces with built-in multiple level-of-details. Linear subdivision in the regular grid setting is also well known to be connected to wavelet construction via the MRA framework due to Mallat and Meyer; and this connection had been explored and exploited in various nonlinear settings [8, 17, 2, 18, 5] to construct various nonlinear "wavelet" transforms. Despite the interest in both wavelets and computer-aided geometric design, there has been little theory on nonlinear subdivision schemes.

In recent years, it was either observed empirically or proved that certain nonlinear subdivision schemes exhibit the following *weakly nonlinear* property: for "most" initial data, the limit curve produced by the nonlinear subdivision scheme has a critical

*Received by the editors April 7, 2005; accepted for publication (in revised form) December 2, 2005; published electronically May 5, 2006. This research was partially supported by a NSF CAREER Award (CCR9984501). This research was partially finished at the Institute for Mathematical Sciences, National University of Singapore, in August, 2004, during the author's visit, which was supported by the Institute.

http://www.siam.org/journals/sinum/44-3/62875.html

†Department of Mathematics, Drexel University, 3141 Chestnut Street, 206 Korman Center, Philadelphia, PA 19104 (yut@drexel.edu).

Hölder regularity exactly the same as that produced by a related linear subdivision scheme. Notable examples include (i) median- and p-mean-interpolating subdivision schemes [21, 20, 8, 17, 16], (ii) refinement schemes of manifold-valued data [18, 19, 22], and (iii) refinement schemes arising from normal multiresolution analysis [5]. A conceptually related discovery is reported in [3], where it is shown that an irregular grid variant of Dubuc's 4-point interpolatory subdivision scheme [9] has the exact same critical Hölder regularity as that in the regular grid setting, so long as the irregularity of the successively refined grids is somehow controlled; and it is conjectured [3, 23] that a similar phenomenon holds for a wilder class of irregular grid subdivision schemes.

Since the regularity of the limit function of a linear subdivision scheme is essentially irrelevant to the initial data, the aforementioned weakly nonlinear subdivision schemes share the same property of data-independence.

In this paper, we show that the nonlinear convexity preserving subdivision scheme (2.1) by [14, 12] produces limit curves with critical Hölder exponents quite heavily dependent on the initial data, unlike the behavior of a linear or weakly nonlinear subdivision scheme. Note that the convexity preserving subdivision scheme when applied to strictly convex data is simply based on the harmonic mean and, similar to the aforementioned weakly nonlinear schemes, may *not* occur to be data dependent at first glance. There exist nonlinear refinement schemes that are more data adaptive in appearance, e.g., the edge adapted or ENO refinement schemes in [2, 7].

While this paper is intended to be self-contained, the proof of our main result (Theorem 2.1) uses a key idea from [21], which is described by the commutative diagram in Figure 1.

We reiterate the lame statement that a lot is known about linear subdivision, but little is known about their nonlinear counterparts; there are currently many unsolved open questions in the nonlinear subdivision literature; see the nonexhaustive list: [18, 19, 5, 21, 20, 8, 17, 13, 16, 15, 14, 12, 2, 22]. Subdivision schemes in various geometric and nonlinear settings are of recent practical interest because of their natural connection with multiscale representations of different data types.

2. Convexity preserving subdivision. In [14, 12], the following nonlinear subdivision scheme is introduced: $f_{j+1} = Sf_j$, where $S : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is defined by

$$f_{j+1,2k} = f_{j,k}, \quad f_{j+1,2k+1} = \frac{f_{j,k} + f_{j,k+1}}{2} - \frac{1}{8} H\big((\triangle^2 f_j)_{k-1}, (\triangle^2 f_j)_k\big). \tag{2.1}$$

Here Δ is a forward difference operator, and $H(\cdot,\cdot)$ denotes harmonic mean, i.e., $H(a,b) = 2ab/(a+b)$ if $ab > 0$, and we define $H(a,b) = 0$ if $ab \leqslant 0$.

It is helpful to bear in mind the linear counterpart of (2.1) based on replacing the nonlinear harmonic mean by the linear arithmetic mean:

$$\begin{aligned} f_{j+1,2k} = f_{j,k}, \quad f_{j+1,2k+1} &= \frac{f_{j,k} + f_{j,k+1}}{2} - \frac{1}{8}\text{Average}\big((\triangle^2 f_j)_{k-1}, (\triangle^2 f_j)_k\big) \\ &= \frac{9}{16}(f_{j,k} + f_{j,k+1}) - \frac{1}{16}(f_{j,k-1} + f_{j,k+2}). \end{aligned} \tag{2.2}$$

This is the well-known subdivision scheme by Dubuc [9]; we denote its subdivision operator by $\overline{S} : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$.

For $r = 1, 2$, there exist (nonlinear) subdivision operators $S^{[r]}$ such that

$$S^{[r]} \circ \Delta^r = \Delta^r \circ S. \tag{2.3}$$

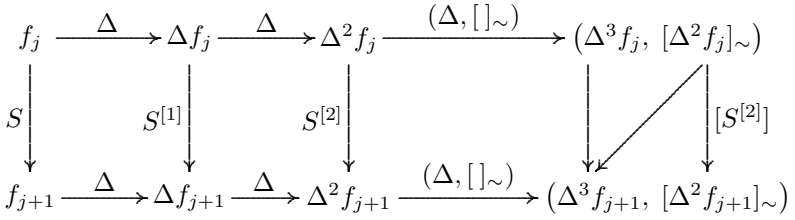


FIG. 1. *Commutation relations for S.*

In particular, $S^{[2]}\Delta^2 v = \Delta^2 S v$; if we write $w = \Delta^2 v$, we have

$$\begin{aligned}\big(S^{[2]}w\big)_{2k} = (\Delta^2 Sv)_{2k} &= (Sv)_{2k} - 2(Sv)_{2k+1} + (Sv)_{2k+2}\\ &= v_k - 2[\frac{v_k + v_{k+1}}{2} - \frac{1}{8}H\big((\triangle^2 v)_{k-1}, (\triangle^2 v)_k\big)] + v_{k+1}\\ &= \frac{1}{4}H(w_{k-1}, w_k),\end{aligned} \tag{2.4}$$

$$\begin{aligned}\big(S^{[2]}w\big)_{2k+1} = (\Delta^2 Sv)_{2k+1} &= (Sv)_{2k+1} - 2(Sv)_{2k+2} + (Sv)_{2k+3}\\ &= \frac{v_k + v_{k+1}}{2} - \frac{1}{8}H\big((\triangle^2 v)_{k-1}, (\triangle^2 v)_k\big) - 2v_{k+1}\\ &\quad + \frac{v_{k+1} + v_{k+2}}{2} - \frac{1}{8}H\big((\triangle^2 v)_k, (\triangle^2 v)_{k+1}\big)\\ &= \frac{w_k}{2} - \frac{1}{8}[H(w_{k-1}, w_k) + H(w_k, w_{k+1})].\end{aligned} \tag{2.5}$$

As a comparison, there exist linear subdivision operators $\overline{S}^{[r]}$, $r = 1, 2, 3, 4$, such that $\overline{S}^{[r]} \circ \Delta^r = \Delta^r \circ \overline{S}$.

It is easy to check, using (2.4)–(2.5), that $S^{[2]}$ is positivity preserving; consequently $S^{[1]}$ is monotonicity preserving and S is convexity preserving. These properties of S are *not* shared by its linear counterpart $\overline{S}$. More in-depth discussions of the relationships among convexity preservation, nonlinear means, and rational interpolation can be found in [14, 12, 11]. We shall later use the fact that S reproduces rational polynomials of degree 2 over degree 1.

We denote by $\mathbb{R}_+$ the set of positive real numbers, and let $\ell_+(\mathbb{Z}) := \{v \mid v : \mathbb{Z} \to \mathbb{R}_+\}$ and $\ell_+^\infty(\mathbb{Z}) := \ell_+(\mathbb{Z}) \cap \ell^\infty(\mathbb{Z})$. By the positivity preserving and locality properties of $S^{[2]}$, we can view it as operators on either $\ell_+(\mathbb{Z})$ or $\ell_+^\infty(\mathbb{Z})$.

By (2.4)–(2.5), we have

$$(S^{[2]}w)_{2k,2k+1,2k+2} = D(w_{k-1}, w_k, w_{k+1}), \tag{2.6}$$

where the nonlinear map $D : \mathbb{R}^3 \to \mathbb{R}^3$ is given by

$$D(w_1, w_2, w_3) = \left(\frac{H(w_1, w_2)}{4},\ \frac{w_2}{2} - \frac{1}{8}\big(H(w_1, w_2) + H(w_2, w_3)\big),\ \frac{H(w_2, w_3)}{4}\right). \tag{2.7}$$

Since D is homogeneous (i.e., $D(\lambda w) = \lambda D(w)$ for all $\lambda \in \mathbb{R}$ and $w \in \mathbb{R}^3$), it induces a quotient map $\pi : \mathbb{P}(\mathbb{R}^3) \to \mathbb{P}(\mathbb{R}^3)$ via the formula $\pi([x]_\sim) = [Dx]_\sim$. Here if V is a real vector space, $\mathbb{P}(V) := V/\sim := \{[v]_\sim : v \in V\}$, where $\sim$ is the equivalence relation defined by $v \sim v' \iff \exists\, c \neq 0$ such that $v = cv'$.

Since D leaves $\mathbb{R}^3_+$ invariant, so does π to $\{[v]_\sim : v \in \mathbb{R}^3_+\}$. Clearly $\{[v]_\sim : v \in \mathbb{R}^3_+\}$ can be identified with $\mathbb{R}^2_+$ by pairing $[x, 1, y]^T_\sim$ with (x, y) (here $x, y > 0$.) Viewing π as a map on $\mathbb{R}^2_+$ under this identification, we have, by (2.7),

$$\pi(x, y) = \left(\frac{2x(1+y)}{2+x+y} \, , \, \frac{2y(1+x)}{2+x+y} \right). \tag{2.8}$$

We have the following facts pertaining to this map:

[P1] (x, x) is a fixed point of π for any $x > 0$.

[P2] For any $x, y > 0$, there is a unique $\overline{x} > 0$ such that $\lim_{n\to\infty} \pi^n(x, y) = (\overline{x}, \overline{x})$.
Proof. [P1] takes care of the case of $x = y$. Assume $x > y$; the other case is symmetrical. The observation that

$$1 < \frac{xy + x}{xy + y} = \frac{\pi(x, y)_1}{\pi(x, y)_2} = \frac{x}{y} \frac{1+y}{1+x} < \frac{x}{y} \tag{2.9}$$

implies that as n increases the first component of $\pi^n(x, y)$ decreases monotonically to a limit value $\overline{x}$, whereas the second component of $\pi^n(x, y)$ increases monotonically to the same value. □

[P3] For any $0 < a < b$, π leaves the square $R := [a, b] \times [a, b]$ invariant, i.e., $\pi(R) \subseteq R$.
Proof. Let $(x, y) \in R$. Without loss of generality, assume $x \geqslant y$. We seek to show that $(x', y') := \pi(x, y)$ continues to belong to R. By (2.9), we have (i) $x' \leqslant x$, (ii) $y' \geqslant y$, and (iii) $x' \geqslant y'$. Assume the contrary that $(x', y') \notin R$; then either $x' < a$ or $y' > b$. If $x' < a$, then $y' < a$ by (iii), so $y < a$ by (ii), contradicting the assumption that $(x, y) \in R$. A similar contradiction can be generated if $y' > b$. □

The following is a more quantitative version of [P2], and is due to Güntürk.

[P4] $L(x, y) := xy/(2 + x + y)$ is invariant under π, i.e., $L(\pi(x, y)) = L(x, y)$. Therefore, for any $(x', y') \in \mathbb{R}^2_+$, the orbit $\{\pi^n(x', y'), : n = 0, 1, 2, \ldots\}$ lies on the level curve $C = \{(x, y) : L(x, y) = x'y'/(2 + x' + y')\}$ and converges (in a monotonic fashion as described in [P2]) to the limit point $(\overline{x}, \overline{x})$, where $\overline{x}$ is the positive root of the quadratic equation $L(\overline{x}, \overline{x}) = x'y'/(2 + x' + y')$. In Lemma 2.5 we show that the curve $xy/(2 + x + y) = 1/4$ corresponds exactly to data sampled from rational polynomials of degree 2 over degree 1.

Let $\Delta : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ or $\mathbb{R}^N \to \mathbb{R}^{N-1}$ $(N \geqslant 2)$ be the forward differencing operator defined by $(\Delta m)_i = m_{i+1} - m_i$. Let

$$\Delta^1 := \Delta, \qquad \Delta^{r+1} := \Delta \circ \Delta^r.$$

(Clearly we are abusing notation when Δ acts on finite vectors. In this case Δ^r maps $\mathbb{R}^N$ to $\mathbb{R}^{N-r}$ for $N \geqslant r + 1$.)

Next we consider the *shrinking factor*: for $(x, y) \neq (1, 1)$, define

$$\begin{aligned} s(x, y) &:= \frac{\|\Delta D([x, 1, y])\|_\infty}{\|\Delta[x, 1, y]\|_\infty} \\ &= \frac{\max\left(\left|\tfrac{1}{2} - \tfrac{3}{8}H(1, x) - \tfrac{1}{8}H(1, y)\right|, \left|\tfrac{1}{2} - \tfrac{1}{8}H(1, x) - \tfrac{3}{8}H(1, y)\right| \right)}{\max(|1 - x|, |1 - y|)}. \end{aligned} \tag{2.10}$$

We define $s(1, 1) = 1/4$ in order to make $s(x, x) = [2(1 + x)]^{-1}$ continuous on $x > 0$. It is elementary to verify the following:

[S1] $s(x,y) \in (0,1/2)$ for $(x,y) \in \mathbb{R}^2_+$.

[S2] $s(x,y)$ is discontinuous at (and only at) $(x,y) = (1,1)$. In a small neighborhood of $(1,1)$, $s(x,y)$ ranges from a little below $1/8$ to a little above $1/4$; precisely one can check that

$$\lim_{\varepsilon\to 0} \inf_{\|(x,y)-(1,1)\|<\varepsilon} s(x,y) = \frac{1}{8} = \lim_{\substack{(x,y)\to(1,1)\\ xy/(2+x+y)=1/4}} s(x,y), \tag{2.11}$$

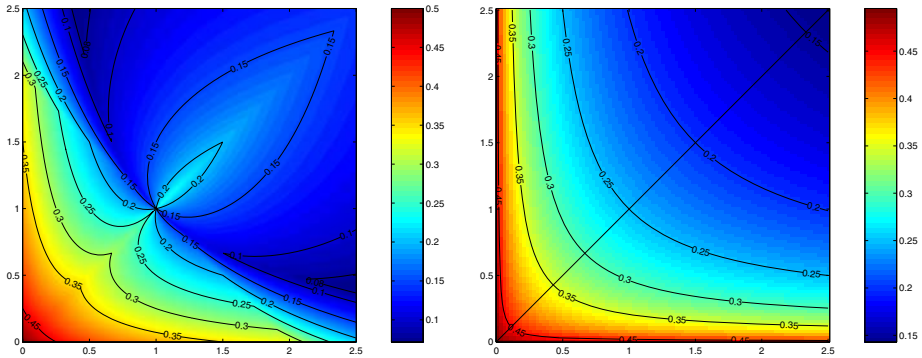
$$\lim_{\varepsilon\to 0} \sup_{\|(x,y)-(1,1)\|<\varepsilon} s(x,y) = \frac{1}{4} = \lim_{\substack{(x,y)\to(1,1)\\ x=y}} s(x,y).$$

See Figure 2. Warning: The 0.25-contour in Figure 2(b), which is exactly the curve $xy/(2+x+y) = 1/4$, can be misleading due to the discontinuity at (1,1): if (x_0,y_0) is any point on this curve not equal to $(1,1)$, then $\lim_{n\to\infty} s(\pi^n(x_0,y_0)) = 1/8 \neq s(\lim_{n\to\infty}\pi^n(x_0,y_0)) = 1/4$.

[S3] $s(x,y) \leqslant s(\pi(x,y))$ with equality holds iff $x = y$.

See also Figure 2. Combining [S3] with [P3] and that $s(x,x)$ is decreasing in x, we have

$$\sup_{\theta\in[a,b]\times[a,b]} s(\theta) = s(a,a) = [2(1+a)]^{-1}. \tag{2.12}$$



(a) (b)

FIG. 2. (a) *Shrinking factor* $s(x,y)$. (b) *Asymptotic shrinking factor* $s(\lim_{n\to\infty}\pi^n(x,y))$.

Let M, N be integers, $M < N$. It is well known in approximation theory (see, e.g., [1, 6]) that for $f \in C([M,N])$, for $\alpha > 0$, we have

$$\begin{aligned} f \in \operatorname{Lip}\alpha \quad &\Longleftrightarrow \quad \exists\, r \in \mathbb{Z}_+,\ r > \alpha, \text{ s.t. } \max_{2^j M\leqslant k\leqslant 2^j N-r} |(\Delta^r f_j)_k| = O(2^{-j\alpha}) \\ &\Longleftrightarrow \quad \forall\, r \in \mathbb{Z}_+,\ r > \alpha, \ \max_{2^j M\leqslant k\leqslant 2^j N-r} |(\Delta^r f_j)_k| = O(2^{-j\alpha}), \end{aligned} \tag{2.13}$$

where f_j is the (length $2^j(N-M)+1$) sequence $(f_j)_k = f(2^{-j}k)$. This equivalence implies that the critical Hölder regularity exponent of f can be determined from the exact asymptotic decay rate of $\max_{2^j M\leqslant k\leqslant 2^j N-r} |(\Delta^r f_j)_k|$ for a large enough differencing order r, i.e.,

$$\sup\{\alpha : f \in \operatorname{Lip}\alpha\} = \sup\Big\{\alpha : \max_{2^j M\leqslant k\leqslant 2^j N-r} |(\Delta^r f_j)_k| = O(2^{-j\alpha})\Big\}. \tag{2.14}$$

For continuous functions defined on the whole real line instead of a compact interval, the equivalence in (2.13) with $M = -\infty$ and $N = +\infty$ holds if we assume that f is bounded. For a (possibly unbounded) continuous function $f : \mathbb{R} \to \mathbb{R}$, we say $f \in \text{Lip } \alpha$ if $f|_{[M,N]} \in \text{Lip } \alpha$ for any $M < N$. Then, under the assumption that there exists $r > \alpha$, $r \in \mathbb{Z}_+$ such that $\Delta^r f_j \in \ell^\infty$ for all j, we have

$$f \in \text{Lip } \alpha \quad \Longleftrightarrow \quad \max_{k\in\mathbb{Z}} |(\Delta^r f_j)_k| = O(2^{-j\alpha}).$$

Remark. Since S is (point-)interpolatory, the subdivision data $S^j f_0$ is exactly the limit function f sampled on the grid $2^{-j}\mathbb{Z}$, so the above result is directly applicable to analyzing the smoothness of f. For other subdivision schemes, linear or nonlinear, more subtle arguments related to stability are needed; see [21, section 3] and the references therein.

As suggested by Figure 1, we shall use $r = 3$ (i.e., third order differences) to analyze the limit functions generated by the nonlinear convexity preserving scheme S.

An essential fact based on the locality property of S is that if we specify the initial data $f_{0,k}$ at the integers $k = M - 2, \ldots, N + 2$, then the limit function restricted to the interval $[M, N]$ is uniquely determined. For $v \in \ell(\mathbb{Z})$, we denote by $S^\infty v$ or $f_v : \mathbb{R} \to \mathbb{R}$ the limit function; it is shown in [14] that f_v is C^1 smooth for arbitrary *strictly convex* initial data v, i.e., $\Delta^2 v \in \ell_+(\mathbb{Z})$.

Remark. Property [S1] already says that $\|\Delta^3 S^j v\|_\infty = O((1/2)^j)$ if $\Delta^2 v \in \ell_+^\infty(\mathbb{Z})$. With a refined argument, one can prove that

$$\|\Delta^3 S^j v\|_\infty = O((1/2 + \epsilon)^j)$$

for an $\epsilon > 0$ dependent on v and can be arbitrarily small. This, in turn, implies the above-mentioned C^1 result in [14].

Our main result is the following theorem.

THEOREM 2.1. *Let $v \in \ell(\mathbb{Z})$ be such that $(\Delta^2 v)_{i-1} = (\Delta^2 v)_{i+1}$ and $(\Delta^2 v)_i > 0$ for all i. Assume*

$$\mu := \frac{(\Delta^2 v)_{2i}}{(\Delta^2 v)_{2i-1}} \in (0, 1). \tag{2.15}$$

(In particular, $\Delta^2 v \in \ell_+^\infty(\mathbb{Z})$ and is a 2-periodic sequence.) Then

$$\sup\{\alpha : f_v \in \text{Lip } \alpha\} = \log_2 2(1 + \mu). \tag{2.16}$$

Therefore, by adjusting the value $\mu \in (0, 1)$, one can construct strictly convex initial data v such that the limit function f_v has a critical Hölder exponent equal to **any** *value in $(1, 2)$.*

Proof. Write $f_j := S^j v \in \ell(\mathbb{Z})$. As $r := 3 > 2 > \log_2 2(1 + \mu)$, it suffices to use $r = 3$ in (2.14) and prove

$$\|\Delta^3 f_j\|_\infty \asymp [2(1 + \mu)]^{-j}, \quad j \to \infty. \tag{2.17}$$

Note that $s(x, x) = [2(1+x)]^{-1}$ according to (2.10). The key point here is to relate the decay rate of $\|\Delta^3 f_j\|_\infty$ to the maps $\pi : \mathbb{R}_+^2 \to \mathbb{R}_+^2$ and $s : \mathbb{R}_+^2 \to (0, 1/2)$ introduced in (2.8) and (2.10).

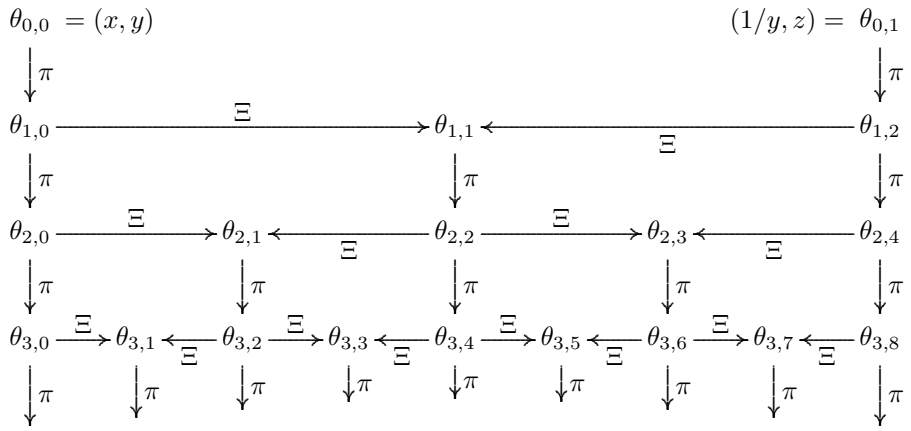


FIG. 3. *Recursive definition of* $\theta_{j,k}$.

(1°) Recall that $\Delta^2 f_j = (S^{[2]})^j w$, where $w := \Delta^2 v$. Recall also that the map $D : \mathbb{R}^3 \to \mathbb{R}^3$ in (2.7) describes the operator $S^{[2]}$ via (2.6). Now, if we write

$$\begin{aligned} \mathbb{R}^3_+ \ni w_{j,k} &:= \big((S^{[2]})^j w\big)_{k-2,\,k-1,\,k} \\ &= \big((\Delta^2 f_j)_{k-2}, (\Delta^2 f_j)_{k-1}, (\Delta^2 f_j)_k\big), \quad k = 0, \dots, 2^j, \end{aligned} \tag{2.18}$$

then, by virtue of (2.6) we have

$$\begin{aligned} w_{j+1,2k} &= D(w_{j,k}), \quad k = 0, \dots, 2^j, \\ w_{j+1,2k+1} &= \big[(w_{j+1,2k})_2, (w_{j+1,2k})_3 = (w_{j+1,2k+2})_1, (w_{j+1,2k+2})_2\big]^T, \\ &\qquad k = 0, \dots, 2^j - 1. \end{aligned} \tag{2.19}$$

Define $\theta_{j,k} = ((w_{j,k})_1, (w_{j,k})_3)/(w_{j,k})_2 \in \mathbb{R}^2_+$. Then (2.19) gives

$$\begin{aligned} \theta_{j+1,2k} &= \pi(\theta_{j,k}), \quad k = 0, \dots, 2^j, \\ \theta_{j+1,2k+1} &= \Xi(\theta_{j+1,2k}, \theta_{j+1,2k+2}), \quad k = 0, \dots, 2^j - 1, \end{aligned} \tag{2.20}$$

where $\Xi : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}^2_+$, $\Xi\big((x, y), (x', y')\big) = (1/y, 1/x')$. (See Figure 3.)

(2°) Under these notations, the assumption (2.15) of the theorem is equivalent to saying

$$\theta_{0,0} = (\mu, \mu) \text{ and } \theta_{0,1} = (1/\mu, 1/\mu), \tag{2.21}$$

and (2.17) is, by symmetry of the data and the subdivision scheme, equivalent to

$$\max_{k=0,\dots,2^j} ||\Delta w_{j,k}||_\infty \asymp [2(1+\mu)]^{-j}, \quad j \to \infty. \tag{2.22}$$

By the overlapping properties of $w_{j,k}$'s (recall (2.18)) we do not need to use all the spatial indices k for a given scale j, and (2.22) is equivalent to

$$\max_{k=0,\dots,2^j} ||\Delta w_{j+1,2k}||_\infty \asymp [2(1+\mu)]^{-(j+1)}, \quad j \to \infty. \tag{2.23}$$

Let $R_\mu := [\mu, 1/\mu] \times [\mu, 1/\mu]$. It is clear that $\Xi(\theta, \theta') \in R_\mu$ if $\theta, \theta' \in R_\mu$, together with property [P3] ($\pi(R_\mu) \subset R_\mu$). We conclude that

$$\theta_{j,k} \in R_\mu \qquad \forall\, j, k. \tag{2.24}$$

So by (2.12), we upper bound all the shrinking factors as

$$s(\theta_{j,k}) \leqslant s(\mu,\mu) = [2(1+\mu)]^{-1} \qquad \forall\, j,k. \tag{2.25}$$

Notice also that

$$\theta_{j,0} = (\mu,\mu) \qquad \forall\, j \geqslant 0. \tag{2.26}$$

Since

$$||\Delta w_{j+1,2k}||_\infty = s(\theta_{j,k})||\Delta w_{j,k}||_\infty,$$

together with (2.25) we have

$$\max_{k=0,\dots,2^j} ||\Delta w_{j+1,2k}||_\infty = O([2(1+\mu)]^{-(j+1)}). \tag{2.27}$$

On the other hand $\max_{k=0,\dots,2^j} ||\Delta w_{j+1,2k}||_\infty = \Omega([2(1+\mu)]^{-(j+1)})$ since

$$\max_k ||\Delta w_{j,k}||_\infty \geqslant ||\Delta w_{j,0}||_\infty = ||\Delta w_{0,0}||_\infty \prod_{l=1}^{j} s(\theta_{l,0}) \overset{(2.26)}{=} ||\Delta w_{0,0}||_\infty\, s(\mu,\mu)^j.$$

So we have proved (2.17). □

In contrast to the nonlinear S, it is well known that for Dubuc's scheme $\overline{S}$,

$$\sup\{\nu : \overline{S}^\infty v \in \operatorname{Lip}\nu\} = \begin{cases} 2 & \text{if } v \in \ell(\mathbb{Z})\backslash\Pi_3|_{\mathbb{Z}}, \\ \infty & \text{if } v \in \Pi_3|_{\mathbb{Z}}. \end{cases} \tag{2.28}$$

In other words, except for initial data sampled from a polynomial of degree 3 or lower, the critical Hölder regularity of the limit curve is 2.[1] This is characteristic of linear subdivision schemes and of the weakly nonlinear schemes mentioned in section 1.

Note. Both (2.16) and (2.17) are not true if $\mu = 1$. When $\mu = 1$, $\Delta^2 v$ is a constant sequence. By (2.4)–(2.5), $\Delta^2 f_j = (1/4)^j \Delta^2 v$, so $\Delta^3 f_j$ is the zero sequence for all j. In this case, the limit function f_v is a quadratic polynomial, which is infinitely smooth.

For general initial strictly convex data, we have the following theorem.

THEOREM 2.2. *Let $w = \Delta^2 v \in \ell_+(\mathbb{Z})$ and $k \in \mathbb{Z}$. Write*

$$x = \frac{w_{k-2}}{w_{k-1}}, \quad y = \frac{w_k}{w_{k-1}}, \quad z = \frac{w_{k+1}}{w_k}.$$

1. *If $\frac{x\,y}{2+x+y} = \frac{1}{4} = \frac{(1/y)\,z}{2+(1/y)+z}$, then $S^\infty v|_{[k,k+1]}$ is a rational polynomial and is C^∞.*
2. *Otherwise, for any $\varepsilon > 0$,*

$$\sup\{\alpha : S^\infty v|_{[k-\varepsilon,k+1+\varepsilon]} \in \operatorname{Lip}\alpha\} < 2.$$

To prove this result we need the following facts pertaining to rational functions.

Remark 2.3. If $R(t) = (at^2+bt+c)/(t-d)$, then $R''(t) = 2(ad^2+bd+c)/(t-d)^3$. This means R is convex on one side of the pole $t = d$ if and only if it is concave on the other side. We refer to the open subset of $\mathbb{R}$ ($\{t : t > d\}$ or $\{t : t < d\}$) at which

[1] More precisely: f_v satisfies $|f_v'(x+t) - f_v'(x)| = O(t\log(1/t))$ and this bound cannot be improved unless for data sampled from a cubic polynomial.

$R''(t) \geqslant 0$ is the *convex branch* of R. If $R(t) = at^2 + bt + c$, its convex branch is defined to be the whole of $\mathbb{R}$ when $a \geqslant 0$, and $\emptyset$ otherwise.

Remark 2.4. Let $t_0 < t_1 < \cdots < t_{n+1}$, and $f_0, f_1, \ldots, f_{n+1} \in \mathbb{R}$. Let $[t_i, \ldots, t_k]f$ be the divided difference of $f_i, \ldots, f_k$ at $t_i, \ldots, t_k$, defined, recursively, by

$$[t_\ell]f := f_\ell, \quad [t_\ell, \ldots, t_{\ell+n}]f := \frac{[t_{\ell+1}, \ldots, t_{\ell+n}]f - [t_\ell, \ldots, t_{\ell+n-1}]f}{t_{\ell+n} - t_\ell} \quad \text{if } n \geqslant 1.$$

If $D_i := [t_i, t_{i+1}, \ldots, t_{i+n}]f$, $i = 0, 1$, are such that $D_0 D_1 > 0$, then there is a unique rational polynomial $R(t)$ of degree n over degree 1 that interpolates f_i at t_i; moreover,

$$R(t) = \frac{D_1 \cdot (t_{n+1} - t)p_0(t) + D_0 \cdot (t - t_0)p_1(t)}{D_1 \cdot (t_{n+1} - t) + D_0 \cdot (t - t_0)}, \tag{2.29}$$

where p_i is the unique Lagrange interpolant of $f_i, \ldots, f_{i+n}$ at $t_i, \ldots, t_{i+n}$, $i = 0, 1$. Recall also Newton's formula

$$p_i(t) = \sum_{k=i}^{i+n} [t_k, \ldots, t_{i+n}]f \prod_{\ell=k+1}^{i+n} (t - t_\ell). \tag{2.30}$$

Note that the numerator of $R(t)$ is a polynomial of degree n or lower: by (2.30) the coefficient of t^n in p_i is D_i, so the coefficient of t^{n+1} in the numerator of $R(t)$ is $D_1 D_0 - D_0 D_1 = 0$. The denominator of $R(t)$ is a constant when $D_0 = D_1$.

Note. When $D_0 \neq D_1$, the rational function (2.29) has a pole at $d = (t_0 D_0 - t_{n+1} D_1)/(D_0 - D_1)$. The condition $D_0 D_1 > 0$ guarantees that $d \neq t_i$, $i = 0, \ldots, n+1$, as $D_1(t_{n+1} - t_i) + D_0(t_i - t_0)$ is positive (resp., negative) if both D_i are positive (resp., negative). However, an example would show that in general there is *no* guarantee that d must stay outside the interval $[t_0, t_{n+1}]$. This last comment adds to the subtlety of the next lemma.

LEMMA 2.5. *Let $f_0, f_1, \ldots, f_4$ be such that $w_i := f_{i+2} - 2f_{i+1} + f_i > 0$ for $i = 0, 1, 2$. Set $x = w_0/w_1$ and $y = w_2/w_1$. Let $\overline{t} \in \mathbb{R}$ and $h > 0$.*

Then there exists a (unique) rational function $R(t) = (at^2 + bt + c)/(t - d)$ such that $R(\overline{t} + ih) = f_i$ for $i = 0, \ldots, 4$ and $[\overline{t}, \overline{t} + 4h] \subset$ (convex branch of R) if and only if

$$\frac{x\, y}{2 + x + y} = \frac{1}{4}. \tag{2.31}$$

Proof. Without loss of generality, we can assume $\overline{t} = 0$ and $h = 1$.

If $f_i = (a(i)^2 + b(i) + c)/(i - d)$, $i = 0, 1, 2, 3, 4$, then $x = (f_2 - 2f_1 + f_0)/(f_3 - 2f_2 + f_1) = (-3 + d)/d$, $y = (f_4 - 2f_3 + f_2)/(f_3 - 2f_2 + f_1) = (-1 + d)/(-4 + d)$. One can then verify (2.31) immediately. Notice that the assumption $[\overline{t}, \overline{t} + 4h] \subset$ (convex branch of R) is the same as saying $d < 0$ or $d > 4$, so there is no worry of division by zero in the definitions of f_i's.

To prove the converse, let

$$D_i := [i, i+1, i+2, i+3]f = (w_{i+1} - w_i)/6, \quad i = 0, 1. \tag{2.32}$$

(We use the divided difference notation as mentioned in Remark 2.4.) Since $D_0 D_1 = (w_1 - w_0)(w_2 - w_1)/36$, $D_0 D_1 > 0$ follows from (2.31) according to the following calculation: $(w_1 - w_0)(w_2 - w_1) = w_1(1 - x)(y - 1) = w_1(1 - x)((2 + x)/(4x - 1) - 1) = 3w_1(1 - x)^2/(4x - 1) > 0$. (Notice that (2.31) implies that $x, y > 1/4$.)

Knowing $D_0D_1 > 0$, Remark 2.4 then applies: The unique rational polynomial of degree 3 over degree 1 that interpolates $f_0, \dots, f_4$ at $0, \dots, 4$ is given by

$$R(x) = \frac{D_1(4-t)p_0(t) + D_0 t p_1(t)}{D_1(4-t) + D_0 t}. \tag{2.33}$$

The goal here is to the show that the numerator of the right-hand side of (2.33) is in fact a polynomial of degree 2 or lower, and also to prove that the pole of R is outside of the interval $[0, 4]$.

The coefficient of t^3 in the numerator of $R(x)$ is

$$\begin{aligned} & D_1(4\,(\text{coeff. of } t^3 \text{ in } p_0) - (\text{coeff. of } t^2 \text{ in } p_0)) + D_0(\text{coeff. of } t^2 \text{ in } p_1) \\ = \;& D_1(4D_0 - (w_1/2 - 6D_0)) + D_0(w_2/2 - 9D_1) \\ \overset{(2.32)}{=} \;& 36w_1(2 + x + y - 4xy) \overset{(2.31)}{=} 0. \end{aligned}$$

The denominator is constant if $D_0 = D_1$; in this case $x = y = 1$ under the condition (2.31). Otherwise, R has its pole at

$$d = \frac{4D_1}{D_1 - D_0} = \frac{4(y-1)}{x+y-2} \overset{(2.31)}{=} \frac{3}{1-x},$$

which stays out of the interval $[0, 4]$ since (2.31) also implies $x > 1/4$. ∎

Proof of Theorem 2.2. Without loss of generality, assume $k = 0$.

(1°) If $\frac{x\,y}{2+x+y} = \frac{1}{4} = \frac{(1/y)\,z}{2+(1/y)+z}$, then, by Lemma 2.5, there exists a unique rational polynomial $R(t)$ of degree 2 over degree 1 such that $f_i = R(i)$, $i = -2, \dots, 3$. It is known from [12] that the subdivision scheme reproduces R on $[0, 1]$, i.e., $S^\infty v|_{[0,1]} = R$. The pole of R is outside the interval $[-2, 3]$. So R is certainly C^∞ on $[0, 1]$.

(2°) To prove the second part of the theorem, we need to first recall the basic setup in part (1°) of the proof of Theorem 2.1. If we define $\theta_{j,k}$ $k = 0, \dots, 2^j$ according to (2.20) with

$$\theta_{0,0} = (x, y), \quad \theta_{0,1} = (1/y, z).$$

(Recall also Figure 3.) We write $C := \{(x, y) \in \mathbb{R}^2_+ : xy/(2 + x + y) = 1/4\}$, $C^+ := \{(x, y) \in \mathbb{R}^2_+ : xy/(2+x+y) > 1/4\}$, and $C^- := \{(x, y) \in \mathbb{R}^2_+ : xy/(2+x+y) < 1/4\}$. Note that $C^+ \supset (1, \infty)^2$ and $C^- \supset (0, 1)^2$.

To prove the theorem, we claim that it suffices to show that

$$\exists\, \overline{j}, \overline{k} \;\text{ s.t. }\; \theta_{\overline{j},\overline{k}} \in C^-. \tag{2.34}$$

By properties [P2] and [P4] of π and the identity $s(x, x) = [2(1 + x)]^{-1}$, (2.34) is equivalent to saying

$$\lim_{n\to\infty} s(\pi^n(\theta_{\overline{j},\overline{k}})) = s^* > 1/4. \tag{2.35}$$

Now, recall from part (1°) of the proof of Theorem 2.1 that

$$s(\theta_{j,k}) = \frac{\max\left(|(\Delta^3 f_{j+1})_{2k-2}|, |(\Delta^3 f_{j+1})_{2k-1}|\right)}{\max\left(|(\Delta^3 f_j)_{k-2}|, |(\Delta^3 f_j)_{k-1}|\right)} \tag{2.36}$$

and

$$\theta_{j,k} = \left(\frac{(\Delta^2 f_j)_{k-2}}{(\Delta^2 f_j)_{k-1}}, \frac{(\Delta^2 f_j)_k}{(\Delta^2 f_j)_{k-1}}\right).$$

If $\theta_{j,k} \in C^-$, then at least one of its two components is strictly less than one, which in turn implies that the denominator of (2.36) is nonzero. So altogether (2.35) implies that

$$\max_{-2 \leqslant k < 2^j} |(\Delta^3 f_j)_k| = \Omega(s_-^j)$$

for any $s_- \in (1/4, s^*)$. Consequently, for any $\varepsilon > 0$, $S^\infty v|_{[-\varepsilon, 1+\varepsilon]}$ has Hölder regularity strictly less than 2.

(3°) So it remains to prove (2.34) under the assumption that either $(x, y) \notin C$ or $(1/y, z) \notin C$.

There is nothing to prove if $(x, y) \in C^-$ or $(1/y, z) \in C^-$. So we can assume without loss of generality that

$$\theta_{0,0} = (x, y) \in C^+ \quad \text{and} \quad \theta_{0,1} = (1/y, z) \in C \cup C^+.$$

As a warmup, let us first consider the case $y = 1$. If $y = 1$, then $1/y = 1$, $z \geqslant 1$ and $x > 1$. But then $\theta_{1,0} = \pi((> 1, 1)) = (> 1, > 1)$, and $\theta_{1,2} = \pi((1, \geqslant 1)) = (\geqslant 1, \geqslant 1)$, so

$$\theta_{1,1} = \Xi(\theta_{1,0}, \theta_{1,2}) = \left(\frac{1}{> 1}, \frac{1}{\geqslant 1} \right) = \big(< 1, \leqslant 1 \big) \in C^-,$$

and we are done with the special case $y = 1$.

We now consider the case $y < 1$; the case of $y > 1$ is similar.

Assume the contrary that no $\theta_{j,k} \in C^-$.

If $y < 1$, then $x > 1$, and $1/y > 1$, so we have

$$\theta_{0,0} = (> 1, < 1), \quad \theta_{0,1} = (> 1, *). \tag{2.37}$$

We know that $\theta_{1,0} = (> 1, *)$, $\theta_{1,2} = (> 1, *)$. Observe that $\theta_{1,0} = (> 1, < 1)$, otherwise $\theta_{1,0} = (> 1, \geqslant 1)$ and

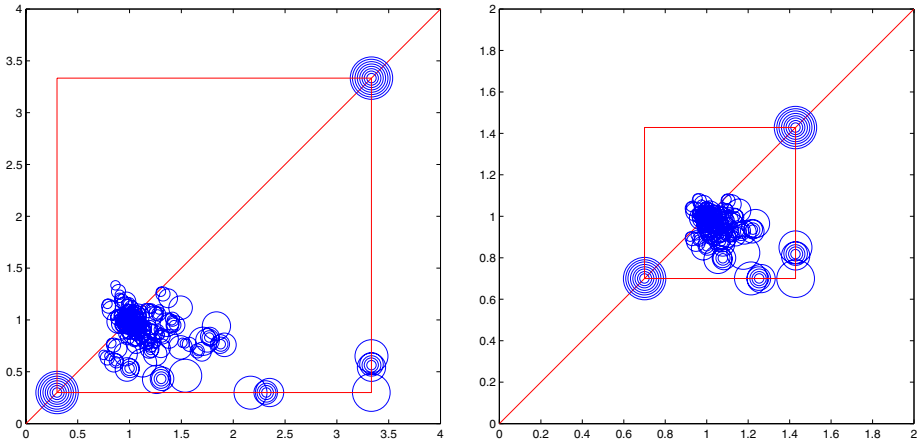
$$\theta_{1,1} = \Xi(\theta_{1,0}, \theta_{1,2}) = \big(1/(\geqslant 1), 1/(> 1) \big) = (\leqslant 1, < 1) \in C^-,$$

violating our assumption. So we now have $\theta_{1,0} = (> 1, < 1)$, $\theta_{1,1} = (> 1, *)$. This brings us back to the same situation as in scale $j = 0$ (recall (2.37)); this means, by induction,

$$\theta_{n,0} = (> 1, < 1), \quad \theta_{n,1} = (> 1, *) \quad \forall\, n.$$

But this contradicts the fact that $\lim_{n\to\infty} \theta_{n,0} = \lim_{n\to\infty} \pi^n(\theta_{0,0}) = (c, c)$, $c > 1$. This completes the proof of the theorem. □

Remark. If $\theta_{0,0}, \theta_{0,1} \in C$, then all $\theta_{j,k}$'s lie on C as well, and are near $(1, 1)$ for j large enough. So, by (2.11), the shrinking factors $s(\theta_{j,k})$ are close to $1/8$ for large j, so $\|\Delta^3 f_j\|_\infty$ decays essentially like $(1/8)^j = 2^{-3j}$. In this very case, we can only say that the critical Hölder regularity is no less than 3, but third order differences may be too "underpower" to determine the optimal Hölder regularity. In other words, the critical smoothness can be higher than 3, and this is exactly the case: Part 1 of Theorem 2.2 says that the smoothness is ∞ ($\gg 3$!).



(a) $x = 0.3$ (b) $x = 0.7$

FIG. 4. *For each $\theta_{j,k} \in \mathbb{R}^2_+$ defined by* (2.20)*, a circle with center $\theta_{j,k}$ and radius which decreases linearly with j is drawn. Notice that $\theta_{j,k}$ tends to cluster at* $(1, 1)$.

3. Observations. For linear subdivision schemes, it is typical that the local Hölder exponents at dyadic irrationals are higher than the global critical Hölder regularity; see [4, section 4]. (Here we assume that the subdivision scheme is binary.) A similar property holds for the nonlinear convexity preserving scheme in this paper, except that it can be more easily explained than those linear examples in [4, section 4]: While Theorem 2.2 says that the global regularity of the limit function f_v for any initial data v is less than 2, a classical result of Alexsandrov (see, e.g., [10]) asserts that f_v, being a convex function on $\mathbb{R}$, must be twice differentiable almost everywhere. (This also partly explains the observed clustering of $\theta_{j,k}$ about $(1, 1)$ for arbitrary initial configuration $x, y, z > 0$ (see, e.g., Figure 4)—a fact that seems difficult to explain by elementary means.)

Acknowledgment. The author thanks Sinan Güntürk for discussions on dynamical systems.

REFERENCES

[1] Z. Ciesielski, *Approximation of splines and its application to Lipschitz classes and to stochastic processes*, in Approximation Theory of Functions, Proceedings of the Conference in Kaluga, 1975, Nauka, Moscow, 1977, pp. 397–400.

[2] A. Cohen, N. Dyn, and B. Matei, *Quasilinear subdivision schemes with applications to ENO interpolation*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 89–116.

[3] I. Daubechies, I. Guskov, and W. Sweldens, *Regularity of irregular subdivision*, Constr. Approx., 15 (1999), pp. 381–426.

[4] I. Daubechies and J. C. Lagarias, *Two-scale difference equations.* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[5] I. Daubechies, O. Runborg, and W. Sweldens, *Normal multiresolution approximation of curves*, Constr. Approx., 3 (2004), pp. 399–463.

[6] Z. Ditzian, *Moduli of smoothness using discrete data*, J. Approx. Theory, 49 (1987), pp. 115–129.

[7] D. L. Donoho, *Minimum entropy segmentation*, in Wavelets Theory, Algorithms and Applications, C.K. Chui et al, eds., Academic Press, San Diego, 1994, pp. 233–269.

[8] D. L. Donoho and T. P.-Y. Yu, *Nonlinear pyramid transforms based on median-interpolation*, SIAM J. Math. Anal., 31 (2000), pp. 1030–1061.

[9] S. Dubuc, *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114 (1986), pp. 185–204.
[10] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[11] M. S. Floater and C. A. Micchelli, *Nonlinear means in geometric modeling*, in Approximation Theory: in Memory of A. K. Varma, N. K. Govil, R. N. Mohapatra, Z. Nashed, A. Sharma, and J. Szabados, eds., Marcel Dekker, New York, 1998, pp. 197–207.
[12] M. S. Floater and C. A. Micchelli, *Nonlinear stationary subdivision*, in Approximation Theory: in Memory of A. K. Varma, N. K. Govil, R. N. Mohapatra, Z. Nashed, A. Sharma, and J. Szabados, eds., Marcel Dekker, New York, 1998, pp. 209–224.
[13] T. N. T. Goodman and T. P.-Y. Yu, *Interpolation of medians*, Adv. Comput. Math., 11 (1999), pp. 1–10.
[14] F. Kuijt and R. van Damme, *Convexity preserving interpolatory subdivision schemes*, Constr. Approx., 14 (1998), pp. 609–630.
[15] P. Oswald, *Smoothness of a nonlinear subdivision scheme*, in Curves and Surface Fitting: Saint-Malo 2002, A. Cohen, J.-L. Merrien, and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2003, pp. 323–332.
[16] P. Oswald, *Smoothness of nonlinear median-interpolation subdivision*, Adv. Comput. Math., 20 (2004), pp. 401–423.
[17] J. S. Pang and T. P.-Y. Yu, *Continuous M-estimators and their interpolation by polynomials*, SIAM J. Numer. Anal., 42 (2004), pp. 997–1017.
[18] I. Ur Rahman, I. Drori, V. C. Stodden, D. L. Donoho, and P. Schröder, *Multiscale representations for manifold-valued data*, Multiscale Modeling and Simulation, 4 (2005), pp. 1201–1232.
[19] J. Wallner and N. Dyn, *Convergence and C^1 analysis of subdivision schemes on manifolds by proximity*, Comput. Aided Geom. Design, 22 (2005), pp. 593–622.
[20] G. Xie and T. P.-Y. Yu, *On a linearization principle for nonlinear p-mean subdivision schemes*, in Advances in Constructive Approximation, M. Neamtu and E. B. Saff, eds., Nashboro Press, Brentwood, TN, 2004, pp. 519–533.
[21] G. Xie and T. P.-Y. Yu, *Smoothness analysis of nonlinear subdivision schemes of homogeneous and affine invariant type*, Constr. Approx., 22 (2005), pp. 219–254.
[22] T. P.-Y. Yu, *Cutting corners on the sphere*, in Wavelets and Splines (Athens, 2005), G. Chen and M.-J. Lai, eds., Nashboro Press, Brentwood, TN, 2006, pp. 496–506.
[23] T. P.-Y. Yu, *On the regularity analysis of interpolatory Hermite subdivision schemes*, J. Math. Anal. Appl., 302 (2005), pp. 201–216.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 949–978

A POSTERIORI ERROR ESTIMATIONS OF SOME CELL CENTERED FINITE VOLUME METHODS FOR DIFFUSION-CONVECTION-REACTION PROBLEMS*

SERGE NICAISE†

Abstract. This paper presents an a posteriori residual error estimator for diffusion-convection-reaction problems approximated by some cell centered finite volume methods on isotropic or anisotropic meshes in $\mathbb{R}^d$, $d = 2$ or 3. For that purpose we built a reconstructed approximation, which is an appropriate interpolant of the finite volume solution. The error is then the difference between the exact solution and this interpolant. The residual error estimator is based on the jump of the normal derivative of the interpolant. We then prove the equivalence between the energy norm of the error and the residual error estimator. Some numerical tests confirm our theoretical results.

Key words. finite volume method, cell centered method, a posteriori error estimates

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/040611483

1. Introduction. The finite volume method is a well-adapted method for the discretization of various partial differential equations and is largely used by engineers [31]. The mathematical analysis of the method has started only recently. Existence and uniqueness results as well as a priori error estimates are now available for quite a large class of problems; see [10] and the references cited there. For finite element methods, a posteriori error estimates are now well understood for a large class of equations; see, for instance, [34]. On the other hand, for finite volume methods, such estimates are not well developed and up to now only a few such results had been obtained. Let us quote [14, 29, 1, 12, 13] for cell centered finite volume methods, [23, 25, 33, 30] for vertex centered methods, and [4, 5, 21, 22, 20] for finite volume element methods.

Recently we obtained a posteriori error estimates of residual type for some cell centered finite volume methods for the Laplace equation in a bounded domain of $\mathbb{R}^d, d = 2$ or 3 [27, 26]. This estimator is based on the use of a reconstructed approximation, namely, an interpolant of Morley type of the finite volume solution. The first goal of the present paper is to extend the previous analysis to diffusion-convection-reaction problems that eventually develop boundary or interior layers. As for the Laplace equation, this requires the introduction of an interpolant, also of Morley type, of the finite volume solution possessing the desired conservation properties. For that purpose, we introduce new finite elements with appropriate degrees of freedom. In contrast with [27, 26] our interpolant is in $H^1(\Omega)$, and therefore the residual error estimator is naturally based on the jump of normal derivatives of the interpolant of the solution. We finally show the equivalence between the energy norm of the error and the residual error estimator. The proof of the upper error bound uses a quasi-orthogonality relation based on the conservation properties of the interpolant. The

*Received by the editors July 12, 2004; accepted for publication (in revised form) December 14, 2005; published electronically May 5, 2006.
http://www.siam.org/journals/sinum/44-3/61148.html
†MACS, ISTV, Université de Valenciennes et du Hainaut-Cambrésis, F-59313 Valenciennes Cedex 9, France (snicaise@univ-valenciennes.fr).

proof of the lower error bound is more standard and simply uses some Green formulas and inverse inequalities as for finite element methods [34, 36, 15, 19].

In certain situations the solution of the diffusion-convection-reaction problem exhibits strong directional features. For instance, if the Peclet number (see below) is large, then boundary or interior layers may occur. Other examples are edge singularities appearing along concave edges in three-dimensional domains. In these cases the use of anisotropic meshes is recommended. Such meshes consist of elements in which the aspect ratio can be very large. Our second goal is to present a residual error estimate valid for anisotropic meshes satisfying minimal assumptions (contrary to the case of a standard residual error estimate for finite element methods; see [15, 16, 18]). This estimate is such that the size of the constant appearing in the upper bound is independent of the coefficients of the operator, while the size of the constant appearing in the lower bound is explicitly given with respect to the coefficients of the operator. These facts are further confirmed numerically.

In contrast with standard practice [35, 19], we do not assume a strong coerciveness assumption (see below), and therefore our residual a posteriori analysis differs from [35, 19].

The outline of the paper is as follows: In section 2 we describe the so-called "cell centered" method for the diffusion-convection-reaction model problem on a triangulation of the domain consisting of triangles, rectangles, or tetrahedra. Some inverse inequalities are recalled in section 3, where we give some further useful interpolation error estimates. Section 4 is devoted to the introduction of some new finite elements (of Morley type) used later on. In section 5 we introduce the Morley interpolant of the approximate solution and prove its main properties. The upper and lower error bounds are then deduced in section 6. The upper error bound is based on the properties of the Morley interpolant, while the lower error bound is proved in a quite standard way. Finally, section 7 is devoted to numerical tests that confirm our theoretical considerations.

2. Discretization of the diffusion-convection-reaction equation. Let Ω be a bounded open subset of $\mathbb{R}^d$, $d = 2$ or 3, with a polygonal ($d = 2$) or polyhedral ($d = 3$) boundary Γ.

We consider the following standard elliptic problem: For $f \in L^2(\Omega)$ let u be the solution of

$$\begin{cases} Au := \operatorname{div}(-\varepsilon\nabla u + \mathbf{v}u) + bu & = f \quad \text{in} \quad \Omega, \\ u & = 0 \quad \text{on} \quad \Gamma, \end{cases} \tag{2.1}$$

where ε is a fixed positive constant, $\mathbf{v}$ a fixed vector function assumed to be sufficiently regular, namely $\mathbf{v} \in C^1(\bar{\Omega}, \mathbb{R}^d)$, and $b \in L^\infty(\Omega, \mathbb{R})$ is a fixed function. This problem is a linearized diffusion-convection-reaction problem appearing in many physical models. In the case of a large Peclet number $Pe \equiv \varepsilon^{-1}\|\mathbf{v}\|_\infty$ and/or large number $\Gamma \equiv \varepsilon^{-1}\|\operatorname{div}\mathbf{v} + b\|_\infty$, the problem is singularly perturbed and the solution may generate sharp boundary or interior layers, where the solution of the limit problem (corresponding to $\varepsilon = 0$) is not smooth or does not satisfy the Dirichlet boundary condition.

In this paper we further assume that

$$\frac{1}{2}\operatorname{div}\mathbf{v} + b \geq 0.$$

This assumption guarantees that there exists a unique weak solution $u \in H_0^1(\Omega)$ of problem (2.1), i.e., satisfying

$$\int_\Omega (\varepsilon \nabla u \cdot \nabla v + \operatorname{div}(\mathbf{v}u)v + buv)\,dx = \int_\Omega fv\,dx \quad \forall v \in H_0^1(\Omega). \tag{2.2}$$

Moreover this solution satisfies

$$|||u||| \lesssim \|f\|, \tag{2.3}$$

where $|||\cdot|||$ is the natural energy norm defined by

$$|||u|||^2 := \int_\Omega \left(\varepsilon|\nabla u|^2 + \left(\frac{1}{2}\operatorname{div}\mathbf{v} + b\right)|u|^2\right).$$

This norm slightly differs from the one used in [35, 19] since in those papers the authors assume that the strong coercivity assumption

$$\frac{1}{2}\operatorname{div}\mathbf{v}(x) + b(x) \geq c_0 > 0 \quad \forall x \in \Omega,$$

holds. This condition excludes the study of convection-diffusion equations (2.1) with $b = 0$ and $\operatorname{div}\mathbf{v} = 0$ that we want to consider here. Therefore their residual a posteriori analysis slightly differs from the one presented here.

As usual, we denote by $L^2(.)$ the Lebesgue spaces and by $H^s(.)$, $s \geq 0$, the standard Sobolev spaces. The usual norm and seminorm of $H^s(D)$ are denoted by $\|\cdot\|_{s,D}$ and $|\cdot|_{s,D}$. For the sake of brevity the $L^2(D)$-norm will be denoted by $\|\cdot\|_D$, and in the case when $D = \Omega$, we will drop the index Ω. The space $H_0^1(\Omega)$ is defined, as usual, by $H_0^1(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\}$. In what follows, the symbol $|\cdot|$ will denote either the Euclidean norm in $\mathbb{R}^d$, $d = 2$ or 3, or the length of a line segment, or the measure of a domain of $\mathbb{R}^d$. Finally, the notation $a \lesssim b$ means here and below that there exists a positive constant C independent of a and b (of the meshsize of the triangulation, as well as the diffusion constant ε) such that $a \leq C\ b$. The notation $a \sim b$ means that $a \lesssim b$ and $b \lesssim a$ hold simultaneously.

2.1. Discretization of the domain. To approximate problem (2.1) by a finite volume scheme we fix a mesh T_h of Ω that satisfies the usual conformity conditions; cf. [6, Chap. 2]. In two dimensions we assume that all elements of T_h are triangles or rectangles, while in three dimensions the mesh is made up of tetrahedra only. For $K \in T_h$ we denote by h_K the diameter of K, and $h = \max_{K\in T_h} h_K$.

We define E_h as the set of edges ($d = 2$) or faces ($d = 3$) of the triangulation, $E_h^{int} = \{E \in E_h / E \subset \Omega\}$ as the set of interior edges/faces of T_h, and $E_h^{ext} = E_h \setminus E_h^{int}$ as the set of exterior edges/faces of T_h.

For an edge E of a two-dimensional element K we introduce $n_{K,E} = (n_x, n_y)$, which is the unit outward normal vector to K along E. Similarly, for a face E of a tetrahedron K let $n_{K,E} = (n_x, n_y, n_z)$ be the unit outward normal vector to K on E. Furthermore, for each edge/face E we fix one of the two normal vectors and denote it by n_E.

The jump of some function v across an edge/face E at a point $y \in E$ is defined as

$$[\![v(y)]\!]_E := \begin{cases} \lim_{\alpha\to+0} v(y + \alpha n_E) - v(y - \alpha n_E) & \forall E \in E_h^{int}, \\ v(y) & \forall E \in E_h^{ext}. \end{cases}$$

Finally, we will need local subdomains (also called patches). As usual, let ω_K be the union of all elements having a common edge/face with K. Similarly, let ω_E be the union of both elements having E as an edge/face.

2.2. The finite volume scheme. Integrating (2.1) on a control volume K and using the divergence formula, we obtain

$$\sum_{E\in E_K}\int_E(-\varepsilon\nabla u+\mathbf{v}u)\cdot n_{K,E}\,ds+\int_K bu\,dx=\int_K f(x)\,dx\quad\forall K\in T_h, \tag{2.4}$$

where E_K is the set of edges/faces of K. The continuous diffusion flux $-\varepsilon\nabla u\cdot n_{K,E}$ is approximated using finite differences and the principle of conservation of flux, the expression $\mathbf{v}u\cdot n_{K,E}$ is approximated by a first order upwind scheme, and the reaction term $\int_K bu$ by a simple quadrature formula (see [10]). These approximations lead to the following system: Find $u_h:=(u_K)_{K\in T_h}$ (u_K being the approximation of $u(x_K)$ for $K\in T_h$ with x_K being the "center" of the box K) that is a solution of

$$\sum_{E\in E_K}\left(-\varepsilon F^D_{K,E}(u_h)+v_{K,E}F^C_E(u_h)\right)+\beta_K F^R_K(u_h)=\int_K f(x)\,dx\quad\forall K\in T_h, \tag{2.5}$$

where $v_{K,E}=\mathcal{M}_E(\mathbf{v}\cdot n_{K,E})$, $\beta_K=\mathcal{M}_K b$; $\mathcal{M}_K g$ and $\mathcal{M}_E g$ denote the mean of g on K and E, respectively, i.e.,

$$\mathcal{M}_K g=\frac{1}{|K|}\int_K g(x)\,dx\quad\forall K\in T_h,\quad \mathcal{M}_E g=\frac{1}{|E|}\int_E g(x)\,d\sigma(x)\quad\forall E\in E_h,$$

while the quantities $F^C_E(u_h)$ and $F^R_K(u_h)$ are defined as

$$F^C_E(u_h):=|E|u_{E,+}, \tag{2.6}$$

where for $E\in E_h^{int}$, $u_{E,+}=u_{K_{E,+}}$, with $K_{E,+}$ being the upstream control volume, i.e., $v_{K_{E,+},E}\geq 0$, while for $E\in\bar K\cap\Gamma$, $u_{E,+}=u_K$ if $v_{K,E}\geq 0$, and $u_{E,+}=0$ otherwise. Similarly,

$$F^R_K(u_h)=|K|u_K.$$

For our purposes, we do not need the exact form of $F^D_{K,E}(u_h)$, but the principle of conservation of flux is required:

$$F^D_{K,E}(u_h)=-F^D_{L,E}(u_h)\text{ for }E=\bar K\cap\bar L.$$

If the mesh T_h is admissible in the sense of [10, Def. 9.1], i.e., satisfies standard orthogonality conditions (see Figure 2.1), then the numerical diffusion flux is defined by

$$F^D_{K,E}(u_h):=\begin{cases}\frac{|E|}{d_E}(u_L-u_K) & \text{if } E=\overline{K}\cap\overline{L},\\ -\frac{|E|}{d_E}u_K & \text{if } E\subset\overline{K}\cap\partial\Omega,\end{cases} \tag{2.7}$$

when $d_E=d(x_K,x_L)$ if $E=\overline{K}\cap\overline{L}$, with $K,L\in T_h$ and $d_E=d(x_K,\Gamma)$ if $E=\partial K\cap\Gamma$. For general meshes, a possible choice for $F^D_{K,E}(u_h)$ is proposed in [7, 8] using the diamond cell method.

For a restricted admissible mesh in the sense of [10, Def. 9.4], if $F^D_{K,E}(u_h)$ is given by (2.7), then system (2.5) is well defined as proved in [9]; see also [10] in the particular case when $\operatorname{div}\mathbf{v}\geq 0$ and b a positive constant. For a general mesh, as is the case here, we simply assume that system (2.5) has a unique solution.

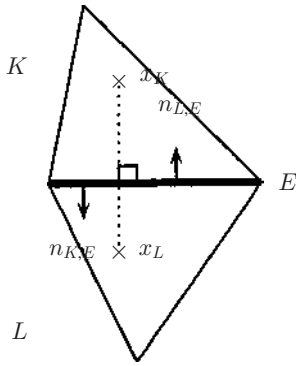


FIG. 2.1. *The standard orthogonality condition.*

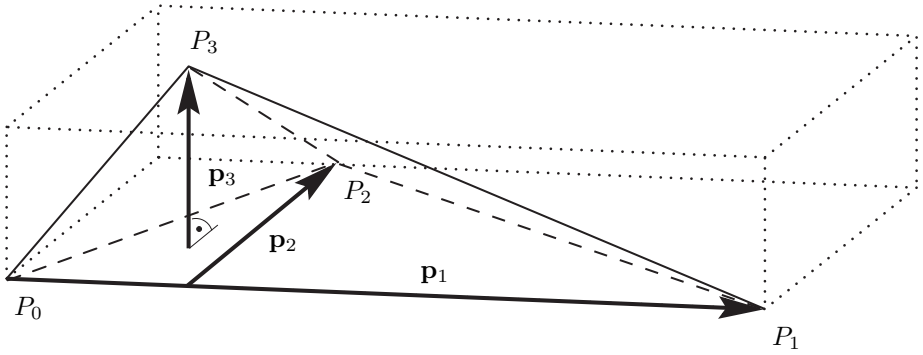


FIG. 2.2. *Notation of tetrahedron K.*

2.3. Some anisotropic quantities. As explained in the introduction, anisotropic discretizations can be very advantageous or, in certain situations, even mandatory. More information and arguments concerning anisotropy can be found in [2, 15].

In this subsection we introduce and describe anisotropic quantities and present their basic properties.

We start with an arbitrary (anisotropic) tetrahedron $K \in T_h$. We enumerate its vertices so that P_0P_1 is the longest edge, $\text{meas}_2(\triangle P_0P_1P_2) \geq \text{meas}_2(\triangle P_0P_1P_3)$, and $\text{meas}_1(P_1P_2) \geq \text{meas}_1(P_0P_2)$. Further, we introduce three orthogonal vectors $\mathbf{p}_{i,K}$ of length $h_{i,K} := |\mathbf{p}_{i,K}|$, as described in Figure 2.2.

The minimal element size is particularly important; thus define

$$h_{min,K} := h_{3,K}.$$

The three main anisotropic directions $\mathbf{p}_{i,K}$ play an important role in several proofs. They span the orthogonal matrix

$$C_K := (\mathbf{p}_{1,K}, \mathbf{p}_{2,K}, \mathbf{p}_{3,K}) \in \mathbb{R}^{3\times 3}.$$

This matrix may be considered as a transformation matrix which defines implicitly the *reference element* $\hat{K}$ via $\hat{K} := C_K^{-1}(K - \vec{P}_0)$; cf. Figure 2.3. In order to facilitate the understanding of this mapping, the circumscribing box of K has been drawn in Figure 2.2. This box is mapped onto the unit cube given in Figure 2.3. Note in particular that the reference element $\hat{K}$ is of size $\mathcal{O}(1)$.

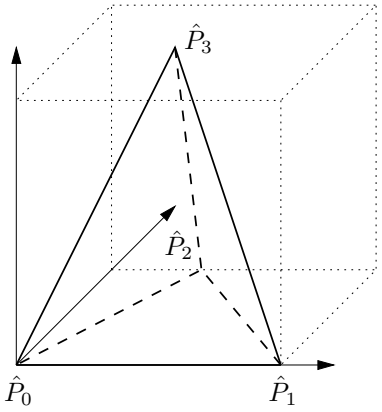


FIG. 2.3. *Reference tetrahedron* $\hat{K}$.

In two dimensions the notation is similar. For a triangle K the enumeration is as in the bottom triangle $P_0P_1P_2$ of Figure 2.2. For a rectangle K we simply take P_0P_1 as the longest edge of K, P_0P_2 as the edge perpendicular to P_0P_1, and simply take $\mathbf{p}_{1,K} = P_0P_1$, $\mathbf{p}_{2,K} = P_0P_2$. In both cases, $h_{min,K} := h_{2,K}$, and C_K becomes a 2×2 matrix.

For an edge/face E of an element K, introduce the height $h_{E,K} = \frac{|K|}{|E|}$.

3. Analytical tools.

3.1. Bubble functions, extension operator, and inverse inequalities. For our further analysis we require standard bubble functions and extension operators that satisfy certain properties recalled here for the sake of completeness.

We need two types of bubble functions, namely, b_K and b_E associated with an element K and an edge E, respectively. For a triangle or a tetrahedron K, denoting by $\lambda_{a_i^K}$, $i = 1, \ldots, d+1$, the barycentric coordinates of K and by $a_i^E, i = 1, \ldots, d$, the vertices of the edge/face $E \subset \partial K$ we recall that

$$b_K = \prod_{i=1}^{d+1} \lambda_{a_i^K} \text{ and } b_E = \prod_{i=1}^{d} \lambda_{a_i^E}.$$

For a rectangle K, we enumerate its vertices in a clockwise sense here. Denoting by $\lambda_{a_i^K}$, $i = 1, \ldots, 4$, the barycentric coordinates of K, namely, $\lambda_{a_i^K}$ is the unique element in $\mathbb{Q}_1(K)$ such that $\lambda_{a_i^K}(a_j^K) = \delta_{i,j}$, we recall that

$$b_K = \lambda_{a_1^K}\lambda_{a_3^K} \text{ and } b_E = \lambda_{a_1^K}(\lambda_{a_2^K} + \lambda_{a_3^K})$$

if the endpoints of the edge E are a_1^K and a_2^K.

We note that

$$b_K = 0 \text{ on } \partial K, \qquad b_E = 0 \text{ on } \partial\omega_E, \qquad \|b_K\|_{\infty,K} = \|b_E\|_{\infty,\omega_E} \sim 1.$$

In two dimensions for an edge $E \subset \partial K$, using temporarily the local coordinates system (x, y) such that E is included in the x-axis, the extension $\mathrm{F}_{\mathrm{ext}}(v_E)$ of $v_E \in C(E)$ to K is defined by $\mathrm{F}_{\mathrm{ext}}(v_E)(x, y) = v_E(x)$. We proceed similarly in three dimensions.

Now we may recall the so-called inverse inequalities that are proved using classical scaling techniques (cf. [34] for the isotropic case and [15] for the anisotropic case).

LEMMA 3.1 (inverse inequalities). *Let $v_K \in \mathbb{P}_{k_0}(K)$ and $v_E \in \mathbb{P}_{k_1}(E)$ for some nonnegative integers k_0 and k_1. Then the following inequalities hold, with the constants in the inequalities depending on the polynomial degrees k_0 or k_1 but not on K, E or v_K, v_E:*

$$\|v_K b_K^{1/2}\|_K \sim \|v_K\|_K, \tag{3.1}$$

$$\|\nabla(v_K b_K)\|_K \lesssim h_{min,K}^{-1}\|v_K\|_K, \tag{3.2}$$

$$\|v_E b_E^{1/2}\|_E \sim \|v_E\|_E, \tag{3.3}$$

$$\|\mathrm{F}_{\mathrm{ext}}(v_E) b_E\|_K \lesssim h_{E,K}^{1/2}\|v_E\|_E, \tag{3.4}$$

$$\|\nabla(\mathrm{F}_{\mathrm{ext}}(v_E) b_E)\|_K \lesssim h_{E,K}^{1/2} h_{min,K}^{-1}\|v_E\|_E. \tag{3.5}$$

3.2. Anisotropic interpolation error estimates. In order to obtain an accurate discrete solution u_h, it is obviously helpful to align the elements of the mesh according to the anisotropy of the solution. It turns out that this intuitive alignment is also necessary to prove sharp upper error bounds. In particular the proof employs specific interpolation error estimates. However, these interpolation estimates do not hold for general meshes; instead the mesh has to have the aforementioned anisotropic alignment with the function to be interpolated.

In order to quantify this alignment, we introduce a so-called alignment measure $m_1(v, T_h)$ which was originally introduced in [16].

DEFINITION 3.2 (alignment measure). *Let $v \in H^1(\Omega)$, and $\mathcal{F} = \{T_h\}$ be a family of triangulations of Ω. Define the* alignment measure *$m_1 : H^1(\Omega) \times \mathcal{F} \mapsto \mathbb{R}$ by*

$$m_1(v, T_h) := \left(\sum_{K \in T_h} h_{min,K}^{-2} \|C_K^\top \nabla v\|_K^2 \right)^{1/2} \Big/ \|\nabla v\|. \tag{3.6}$$

By definition one has $m_1(v, T_h) \geq 1$. For arbitrary isotropic meshes one obtains that $m_1(v, T_h) \sim 1$. The same is achieved for anisotropic meshes T_h that are aligned with the anisotropic function v. Therefore the alignment measure is not an obstacle for reliable error estimation.

Since the focus of our work is a posteriori error estimation, we refer to [17, 16] for discussions concerning this alignment measure.

LEMMA 3.3 (local interpolation error bounds). *Let $v \in H_0^1(\Omega)$; then,*

$$\|v - \mathcal{M}_K v\|_K \lesssim \|C_K^\top \nabla v\|_K \qquad \forall K \in T_h, \tag{3.7}$$

$$h_{E,K}\|v - \mathcal{M}_E v\|_E^2 \lesssim \|C_K^\top \nabla v\|_K^2 \qquad \forall E \in E_K, K \in T_h. \tag{3.8}$$

Proof. The first inequality (3.7) has been proved in [16, Lemma 4]. The same scaling argument and the compact embedding of $H^1(\hat{K})$ into $L^2(\hat{E})$ yield the second estimate. ∎

LEMMA 3.4 (global interpolation error bounds). *Let $v \in H^1(\Omega)$; then,*

$$\sum_{K \in T_h} h_{min,K}^{-2}\|v - \mathcal{M}_K v\|_K^2 \lesssim m_1(v, T_h)^2 \|\nabla v\|^2, \tag{3.9}$$

$$\sum_{K \in T_h} \sum_{E \in E_K \cap E_h^{int}} h_{E,K} h_{min,K}^{-2}\|v - \mathcal{M}_E v\|_E^2 \lesssim m_1(v, T_h)^2 \|\nabla v\|^2. \tag{3.10}$$

Proof. These are direct consequences of the previous lemma and the definition of the alignment measure. ∎

For further analysis of our estimator, we need specific interpolation estimates related to the diffusion-convection-reaction problem [36, 18, 19]. Namely, the error estimates have to be related to the energy norm $|||\cdot|||$ and a local quantity relying on the local meshsize and the local behavior of the functions involved in the operator. This quantity is defined by

$$\alpha_K := \min\{c_K^{-1/2}, \varepsilon^{-1/2} h_{min,K}\},$$

where we set

$$c_K = \min_{x\in K}\left(\frac{1}{2}\operatorname{div}\mathbf{v}(x) + b(x)\right).$$

Here we use the convention that if $c_K = 0$, then the minimum is the second term, namely, $\alpha_K := \varepsilon^{-1/2} h_{min,K}$.

We are now able to prove the following error estimate (compare with Lemma 3.9 of [18]).

LEMMA 3.5 (global interpolation error bounds). *Let $v \in H_0^1(\Omega)$. Then we have*

$$\sum_{K\in T_h} \alpha_K^{-2}\|v - \mathcal{M}_K v\|_K^2 \lesssim m_1(v,T_h)^2 |||v|||^2, \tag{3.11}$$

$$\varepsilon^{1/2}\sum_{K\in T_h}\alpha_K^{-1}\sum_{E\in E_K\cap E_h^{int}}\|v - \mathcal{M}_E v\|_E^2 \lesssim m_1(v,T_h)^2 |||v|||^2. \tag{3.12}$$

Proof. For the first estimate, we split up its left-hand side as follows:

$$\begin{aligned}\sum_{K\in T_h}\alpha_K^{-2}\|v - \mathcal{M}_K v\|_K^2 = &\sum_{K\in T_h:\varepsilon h_{min,K}^{-2}\le c_K} c_K\|v - \mathcal{M}_K v\|_K^2\\ &+ \sum_{K\in T_h:\varepsilon h_{min,K}^{-2} > c_K}\varepsilon h_{min,K}^{-2}\|v - \mathcal{M}_K v\|_K^2.\end{aligned}$$

By the definition of c_K and estimate (3.9), we conclude that

$$\begin{aligned}\sum_{K\in T_h}\alpha_K^{-2}\|v - \mathcal{M}_K v\|_K^2 &\lesssim \|\left(\frac{1}{2}\operatorname{div}\mathbf{v} + b\right)^{1/2} v\|^2 + \varepsilon m_1(v,T_h)^2\|\nabla v\|^2\\ &\lesssim m_1(v,T_h)^2|||v|||^2.\end{aligned}$$

For the second estimate, we first use the same splitting

$$\begin{aligned}\sum_{K\in T_h}\sum_{E\in E_K\cap E_h^{int}}\alpha_K^{-1}\|v - \mathcal{M}_E v\|_E^2 = &\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}}\sum_{E\in E_K\cap E_h^{int}}\alpha_K^{-1}\|v - \mathcal{M}_E v\|_E^2\\ &+ \sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2} > c_K}}\sum_{E\in E_K\cap E_h^{int}}\alpha_K^{-1}\|v - \mathcal{M}_E v\|_E^2.\end{aligned}$$

As before, this leads to

$$\varepsilon^{1/2}\sum_{K\in T_h}\sum_{E\in E_K\cap E_h^{int}}\alpha_K^{-1}\|v-\mathcal{M}_E v\|_E^2 \lesssim \varepsilon^{1/2}\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}} c_K^{1/2}\sum_{E\in E_K\cap E_h^{int}}\|v\|_E^2$$
$$+\varepsilon\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}> c_K}} h_{min,K}^{-1}\sum_{E\in E_K\cap E_h^{int}}\|v-\mathcal{M}_E v\|_E^2.$$

Using the estimate $h_{min,K}\lesssim h_{E,K}$ proved in Lemma 3.1 of [18] and the estimate (3.10), we obtain

$$\varepsilon^{1/2}\sum_{K\in T_h}\sum_{E\in E_K\cap E_h^{int}}\alpha_K^{-1}\|v-\mathcal{M}_E v\|_E^2 \lesssim \sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}}\varepsilon^{1/2}c_K^{1/2}\sum_{E\in E_K\cap E_h^{int}}\|v\|_E^2 + \varepsilon m_1(v,T_h)^2\|\nabla v\|^2. \tag{3.13}$$

For the first term of this right-hand side, using the trace inequality (see, for instance, Lemma 2.4 of [15] or Lemma 3.5 of [18])

$$\|v\|_E^2\lesssim h_{E,K}^{-1}\|v\|_K(\|v\|_K+\|C_K^\top\nabla v\|_K),$$

we may write

$$\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}}\varepsilon^{1/2}c_K^{1/2}\sum_{E\in E_K\cap E_h^{int}}\|v\|_E^2\lesssim\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}}\varepsilon^{1/2}c_K^{1/2}h_{E,K}^{-1}(\|v\|_K^2+\|v\|_K\|C_K^\top\nabla v\|_K). \tag{3.14}$$

Fix for a moment an element K such that $\varepsilon h_{min,K}^{-2}\le c_K$. Using the property $h_{min,K}\lesssim h_{E,K}$ and Young's inequality with a parameter $\eta_K>0$, we may write

$$\varepsilon^{1/2}c_K^{1/2}h_{E,K}^{-1}(\|v\|_K^2+\|v\|_K\|C_K^\top\nabla v\|_K)\lesssim\varepsilon^{1/2}c_K^{1/2}h_{min,K}^{-1}\|v\|_K^2 + \frac{\varepsilon^{1/2}c_K^{1/2}}{2\eta_K}\|v\|_K^2+\frac{\varepsilon^{1/2}c_K^{1/2}h_{min,K}^{-2}\eta_K}{2}\|C_K^\top\nabla v\|_K^2.$$

Since $\varepsilon^{1/2}h_{min,K}^{-1}\le c_K^{1/2}$, the first term has the correct factor, namely,

$$\varepsilon^{1/2}c_K^{1/2}h_{E,K}^{-1}(\|v\|_K^2+\|v\|_K\|C_K^\top\nabla v\|_K) \lesssim c_K\|v\|_K^2+\frac{\varepsilon^{1/2}c_K^{1/2}}{2\eta_K}\|v\|_K^2+\frac{\varepsilon^{1/2}c_K^{1/2}h_{min,K}^{-2}\eta_K}{2}\|C_K^\top\nabla v\|_K^2.$$

For the last two terms we choose $\eta_K=\frac{\varepsilon^{1/2}}{c_K^{1/2}}$ which yields

$$\varepsilon^{1/2}c_K^{1/2}h_{E,K}^{-1}(\|v\|_K^2+\|v\|_K\|C_K^\top\nabla v\|_K)\lesssim c_K\|v\|_K^2+\varepsilon h_{min,K}^{-2}\|C_K^\top\nabla v\|_K^2.$$

This estimate in (3.14) leads to

$$\sum_{\substack{K\in T_h:\\ \varepsilon h_{min,K}^{-2}\le c_K}}\varepsilon^{1/2}c_K^{1/2}\sum_{E\in E_K\cap E_h^{int}}\|v\|_E^2\lesssim\sum_{K\in T_h}(c_K\|v\|_K^2+\varepsilon h_{min,K}^{-2}\|C_K^\top\nabla v\|_K^2) \lesssim m_1(v,T_h)^2|||v|||^2.$$

Inserting this estimate in (3.13) we arrive at (3.12). ∎

4. Some finite elements of Morley type. For our further analysis, we need a continuous interpolant p satisfying

$$\int_E \frac{\partial p}{\partial n_{K,E}}\, ds = F^D_{K,E}(u_h) \qquad \forall E \in E_K, \tag{4.1}$$

$$\int_E \mathbf{v}\cdot n_{K,E} p\, ds = v_{K,E} F^C_{K,E}(u_h) \qquad \forall E \in E_K, \tag{4.2}$$

$$\int_K bp = \beta_K F^R_K(u_h) \tag{4.3}$$

for any $K \in T_h$. This means that we need to use C^0-finite elements having as degrees of freedom the three left-hand sides above. Since these left-hand sides clearly depend on the restriction of $\mathbf{v}$ and b on K, the finite elements depend on these restrictions. The interpolant's construction is quite complicated and could be expansive from a computational point of view. Such elements are even not unique. We therefore build them in a generic way and also give simpler elements in the case of constant coefficients.

4.1. Convection-diffusion elements for constant coefficients. For a convection-diffusion problem with constant coefficients (i.e., for $b = 0$ and $\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0$) we need an interpolant p satisfying only (4.1) and (4.2). In this case we need finite elements having as degrees of freedom the mean of p and the normal derivative of p on each edge. Therefore we need a kind of Morley element [24, 6]. In two dimensions we modify the element of Nilssen, Tai, and Winther [28] and extend this new element to rectangles and to tetrahedra.

4.1.1. Triangles. Here K is a (nondegenerate) triangle with vertices a_i^K, $i = 1, 2, N_f := 3$.

Now, inspired by [28, sect. 4], we take

$$\begin{aligned} &P_K = \mathbb{P}_2(K) \oplus \mathbb{P}_1(K) b_K = \{q + p b_K : q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K)\}, \\ &\Sigma_K = \{p(a_i^K)\}_{i=1,\dots,N_f} \cup \left\{\int_E p\, ds\right\}_{E\in E_K} \cup \left\{\int_E \frac{\partial p}{\partial n_{K,E}}\, ds\right\}_{E\in E_K}. \end{aligned} \tag{4.4}$$

LEMMA 4.1. *The above triple (K, P_K, Σ_K) is a C^0-finite element.*

Proof. It suffices to show that $w \in P_K$ satisfying

$$l(w) = 0 \quad \forall l \in \Sigma_K \tag{4.5}$$

is equal to zero. But w is of the form $w = q + p b_K$ with $q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K)$, and since b_K is identically equal to zero on the edges $E \in E_K$, the conditions $w(a_i^K) = \int_E w\, ds = 0$ are equivalent to

$$q(a_i^K) = \int_E q\, ds = 0.$$

Since the restriction of q to E is of degree 2, these conditions imply that $q_{|E} \equiv 0$ for all $E \in E_K$ and consequently $q \equiv 0$.

The conclusion follows from Lemma 4.1 of [28] since it is proved there that $w = p b_K$ with $p \in \mathbb{P}_1(K)$ satisfying $\int_E \frac{\partial w}{\partial n_{K,E}}\, ds = 0$ for all $E \in E_K$ is identically equal to zero.

The continuity of the element follows from the first part of the proof. ∎

4.1.2. Rectangles. Here K is a (nondegenerate) rectangle with vertices a_i^K, $i = 1, \dots, N_f := 4$ with edges parallel to the axes. Now we take

$$P_K = \{q + pb_K : q \in \mathbb{Q}_2(K), p(x,y) = \alpha x + \beta y + \gamma y^2, \alpha, \beta, \gamma \in \mathbb{R}\}.$$

As before, the degrees of freedom are defined by (4.4).

LEMMA 4.2. *The above triple (K, P_K, Σ_K) is a C^0-finite element.*

Proof. By an affine transformation, it suffices to show the result on the reference element $\hat{K} = (0,1)^2$. For the sake of brevity we will omit the sign $\hat{}$. As $\dim P_K = \operatorname{card} \Sigma_K$, it suffices to show that $w \in P_K$ satisfying (4.5) is equal to zero. But w is of the form $w = q + pb_K$ with $q \in \mathbb{Q}_2(K)$ and p of the form $p(x,y) = \alpha x + \beta y + \gamma y^2, \alpha, \beta, \gamma \in \mathbb{R}$. Since b_K is identically equal to zero on the edges $E \in E_K$, the conditions $w(a_i^K) = \int_E w\, ds = 0$ are equivalent to

$$q(a_i^K) = \int_E q\, ds = 0.$$

Since the restriction of q to E is of degree 2, these conditions imply, as before, that $q_{|E} \equiv 0$ for all $E \in E_K$. Fix the basis $\{\lambda_i(x)\lambda_j(y)\}_{0 \le i,j \le 2}$ of $\mathbb{Q}_2(K)$, where $\lambda_i \in \mathbb{P}_2(0,1)$ satisfy $\lambda_i(x_j) = \delta_{ij}$, where $x_i = \frac{i}{2}, i = 0, 1, 2$. Writing q in this basis, we observe that the property $q_{|E} \equiv 0$ for all $E \in E_K$ implies that $q(x,y) = \delta\lambda_1(x)\lambda_1(y) = \delta b_K(x,y)$ for some $\delta \in \mathbb{R}$.

Returning to w, we conclude that

$$w(x,y) = (\delta + \alpha x + \beta y + \gamma y^2) b_K(x,y).$$

Now, one readily sees that the remainding condition $\int_E \frac{\partial w}{\partial n_{K,E}}\, ds = 0$ for all $E \in E_K$ is equivalent to a homogeneous 4×4 linear system in $\alpha, \beta, \gamma, \delta$ whose unique solution is $\alpha = \beta = \gamma = \delta = 0$. Therefore w is identically equal to zero. ∎

4.1.3. Tetrahedra. Here K is a (nondegenerate) tetrahedron with vertices a_i^K, $i = 1, 2, 3, N_f := 4$.

Inspired by the above triangular example, we choose Σ_K defined by (4.4) and

$$P_K = \left\{ q + pb_K + \sum_{E \in E_K} \alpha_E b_E : p, q \in \mathbb{P}_1(K), \alpha_E \in \mathbb{R} \right\}.$$

As before, we can prove that the above triple (K, P_K, Σ_K) is a C^0-finite element.

4.2. Diffusion-convection-reaction elements for constant coefficients. We build our finite elements by slightly modifying the elements from the previous subsection.

4.2.1. Triangles. We take

$$\begin{aligned} P_K &= \{q + (p + \alpha b_K) b_K : q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K), \alpha \in \mathbb{R}\}, \\ \Sigma_K &= \{p(a_i^K)\}_{i=1,\dots,N_f} \cup \left\{ \int_K p \right\} \cup \left\{ \int_E p\, ds \right\}_{E \in E_K} \cup \left\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \right\}_{E \in E_K}. \end{aligned} \tag{4.6}$$

LEMMA 4.3. *The above triple (K, P_K, Σ_K) is a C^0-finite element.*

Proof. It suffices to show that $w \in P_K$ satisfying (4.5) is equal to zero. Here w is of the form $w = q + (p + \alpha b_K) b_K$ with $q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K)$, and $\alpha \in \mathbb{R}$. As in Lemma 4.1, the conditions $w(a_i^K) = \int_E w\, ds = 0$ imply that $q \equiv 0$.

Next, we consider the condition

$$\int_E \frac{\partial w}{\partial n_{K,E}}\, ds = 0 \qquad \forall E \in E_K,$$

which is reduced to

$$\int_E \frac{\partial (pb_K)}{\partial n_{K,E}}\, ds = 0 \qquad \forall E \in E_K.$$

Thanks to Lemma 4.1 of [28], p is identically equal to zero.

Finally, the condition $\int_K w$ becomes $\alpha \int_K b_K^2 = 0$, which leads to the conclusion. ∎

4.2.2. Rectangles. In the setting of subsection 4.1.2, we take Σ_K defined by (4.6) and

$$P_K = \{q + pb_K : q \in \mathbb{Q}_2(K), p(x,y) = \alpha x + \beta y + \gamma x^2 + \delta y^2, \alpha, \beta, \gamma, \delta \in \mathbb{R}\}.$$

Arguments similar to those used in the proof of Lemma 4.2 allow one to show that the above triple (K, P_K, Σ_K) is a C^0-finite element.

4.2.3. Tetrahedra. Inspired by the above triangular example and the tetrahedral example from subsection 4.1, we choose Σ_K defined by (4.6) and

$$P_K = \left\{ q + pb_K + \alpha b_K^2 + \sum_{E \in E_K} \alpha_E b_E : p, q \in \mathbb{P}_1(K), \alpha_E, \alpha \in \mathbb{R} \right\}.$$

As before, the above triple (K, P_K, Σ_K) is a C^0-finite element.

4.3. Reaction-diffusion elements for constant coefficients. If we consider only reaction-diffusion equations (i.e., $\mathbf{v} = 0$), we may restrict ourselves to an interpolant p satisfying (4.1) and (4.3). Therefore we need to use finite elements having as degrees of freedom the left-hand side of (4.1) and the mean on K. Again we need to slightly modify the previous elements.

4.3.1. Triangles or tetrahedra. We take

$$\begin{aligned} P_K &= \{q + (p + \alpha b_K) b_K : q \in \mathbb{P}_1(K), p \in \mathbb{P}_1(K), \alpha \in \mathbb{R}\}, \\ \Sigma_K &= \{p(a_i^K)\}_{i=1,\dots,N_f} \cup \left\{ \int_K p \right\} \cup \left\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \right\}_{E \in E_K}. \end{aligned} \tag{4.7}$$

LEMMA 4.4. *The above triple (K, P_K, Σ_K) is a C^0-finite element.*

Proof. As usual, it suffices to show that $w \in P_K$ satisfying (4.5) is equal to zero. Here w is of the form $w = q + (p + \alpha b_K) b_K$ with $q \in \mathbb{P}_1(K), p \in \mathbb{P}_1(K)$, and $\alpha \in \mathbb{R}$. As in Lemma 4.1, the conditions $w(a_i^K) = 0$ imply that $q \equiv 0$ since q is of degree at most 1. The remainder of the proof is the same as the one of Lemma 4.3. ∎

4.3.2. Rectangles. In the setting of subsection 4.1.2, we take Σ_K defined by (4.7) and

$$P_K = \{q + pb_K : q \in \mathbb{Q}_1(K), p(x,y) = \eta + \alpha x + \beta y + \gamma x^2 + \delta y^2, \alpha, \beta, \gamma, \delta, \eta \in \mathbb{R}\}.$$

Arguments similar to those used in the proof of Lemma 4.2 allow one to show that the triple (K, P_K, Σ_K) is a C^0-finite element.

4.4. General case.

4.4.1. Triangles or tetrahedra. Inspired by the previous examples, we take

$$P_K^{\mathbf{v},b} = \Big\{ q_0 + \sum_{E\in E_K} \alpha_E \tilde{v}_{K,E} b_E b_K + \Big(\sum_{E\in E_K} \beta_E b_E + \gamma b_K \Big) b_K : q_0 \in \mathbb{P}_1(K), \alpha_E, \beta_E, \gamma \in \mathbb{R} \Big\}, \tag{4.8}$$

$$\Sigma_K^{\mathbf{v},b} = \{p(a_i^K)\}_{i=1,\dots,N_f} \cup \Big\{ \int_E \mathbf{v}\cdot n_{K,E} p\, ds \Big\}_{E\in E_K} \cup \Big\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \Big\}_{E\in E_K} \cup \Big\{ \int_K bp \Big\}, \tag{4.9}$$

where $\tilde{v}_{K,E}$ is any extension from E to K of the function $\mathbf{v}\cdot n_{K,E}$ such that $\tilde{v}_{K,E} \equiv 0$ if $\mathbf{v}\cdot n_{K,E} \equiv 0$ on E. Here there is a slight abuse of notation in the sense that the degree of freedom $\int_E \mathbf{v}\cdot n_{K,E} p\, ds$ (or $\int_K bp$) disappears if $\mathbf{v}\cdot n_{K,E} \equiv 0$ on E (or $b \equiv 0$).

As before, we can prove that the triple $(K, P_K^{\mathbf{v},b}, \Sigma_K^{\mathbf{v},b})$ is a C^0-finite element.

4.4.2. Rectangles. Now we take $P_K^{\mathbf{v},b}$ defined by (4.8), but with q_0 in $\mathbb{Q}_1(K)$ and $\Sigma_K^{\mathbf{v},b}$ defined by (4.9). Again this triple $(K, P_K^{\mathbf{v},b}, \Sigma_K^{\mathbf{v},b})$ is a C^0-finite element.

5. The Morley interpolant.

5.1. Definition. For any $K \in T_h$, we fix a C^0-finite element $(K, P_K^{\mathbf{v},b}, \Sigma_K^{\mathbf{v},b})$ such that

$$\Big\{ \int_E \mathbf{v}\cdot n_{K,E} p\, ds \Big\}_{E\in E_K} \cup \Big\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \Big\}_{E\in E_K} \cup \Big\{ \int_K bp \Big\} \subset \Sigma_K^{\mathbf{v},b}.$$

We refer to the previous section for its existence.

We now introduce the finite element space:

$$V_h := \Big\{ v_h \in H_0^1(\Omega) : v_{h|K} \in P_K^{\mathbf{v},b} \quad \forall K \in T_h, \int_E \frac{\partial v_{h|K}}{\partial n_E}\, ds = \int_E \frac{\partial v_{h|K}}{\partial n_E}\, ds \qquad \forall E \in E_h, K, L \in T_h : E = K \cap L \Big\}.$$

DEFINITION 5.1. *For $u_h = (u_K)_{K\in T_h}$ we define its interpolant (of Morley type) $I_M u_h$ as the unique element v_h in V_h satisfying*

$$\int_E \frac{\partial v_{h|K}}{\partial n_{K,E}}\, ds = F_{K,E}^D(u_h) \quad \forall E \in E_K, K \in T_h, \tag{5.1}$$

$$\int_E \mathbf{v}\cdot n_{K,E} v_h\, ds = v_{K,E} F_E^C(u_h) \quad \forall E \in E_h^{int}, \tag{5.2}$$

$$\int_K b v_h\, dx = \beta_K F_K^R(u_h) \quad \forall K \in T_h. \tag{5.3}$$

5.2. Some useful properties. The key point of our a posteriori analysis is the following basic property of the Morley interpolant.

LEMMA 5.2. *If u_h is solution of* (2.5), *then $I_M u_h$ satisfies*

$$\int_K (A(I_M u_h) - f)\, dx = 0 \quad \forall K \in T_h. \tag{5.4}$$

Proof. By Green's formula we have

$$\begin{aligned}\int_K A(I_M u_h)\, dx &= \int_K (\operatorname{div}(-\varepsilon \nabla I_M u_h + \mathbf{v} I_M u_h) + b I_M u_h)\, dx \\ &= \sum_{E \in E_K} \int_E \left(-\varepsilon \frac{\partial (I_M u_h)}{\partial n_{K,E}} + \mathbf{v} \cdot n_{K,E} I_M u_h\right) ds + \int_K b I_M u_h\, dx.\end{aligned}$$

Using the properties (5.1), (5.2), and (5.3) satisfied by $I_M u_h$ we obtain

$$\int_K A(I_M u_h)\, dx = \sum_{E \in E_K} \left(-\varepsilon F^D_{K,E}(u_h) + v_{K,E} F^C_E(u_h)\right) + \beta_K F^R_K(u_h)$$

and we conclude by using (2.5). ∎

Now we prove a quasi-orthogonality relation that will be used for the upper error bound. We first define the gradient jump of $I_M u_h$ in the normal direction by

$$J_{E,n}(u_h) = \varepsilon \left[\left[\left(\frac{\partial}{\partial n_E}\right)(I_M u_h)\right]\right]_E \quad \forall E \in E_h^{int}.$$

LEMMA 5.3. *If u is a solution of* (2.2) *and u_h is a solution of* (2.5), *then for any $\chi \in H^1_0(\Omega)$, setting $e = u - I_M u_h$, we have that*

$$\begin{aligned}\int_\Omega (\varepsilon \nabla e \cdot \nabla \chi + (\operatorname{div}(\mathbf{v} e) + be)\chi)\, dx &= \sum_{K \in T_h} \int_K (f - A(I_M u_h))(\chi - \mathcal{M}_K \chi)\, dx \\ &\quad + \sum_{E \in E_h^{int}} \int_E J_{E,n}(u_h)(\chi - \mathcal{M}_E \chi)\, ds.\end{aligned} \tag{5.5}$$

Proof. For the sake of brevity denote the left-hand side of (5.5) by $I_1(\chi)$. By (2.2) and using Green's formula on each element K we obtain

$$I_1(\chi) = \int_\Omega f\chi\, dx - \sum_{K \in T_h} \int_K A(I_M u_h)\chi\, dx - \sum_{K \in T_h} \int_{\partial K} \varepsilon \frac{\partial (I_M u_h)}{\partial n_K} \chi\, ds.$$

The continuity of $I_M u_h$ and of χ across the edges/faces (in the sense of trace) and the fact that $\chi = 0$ on Γ lead to

$$I_1(\chi) = \sum_{K \in T_h} \int_K (f - A(I_M u_h))\chi\, dx + \sum_{E \in E_h^{int}} \int_E J_{E,n}(u_h)\chi\, ds.$$

Using the identity (5.4) we arrive at

$$I_1(\chi) = \sum_{K \in T_h} \int_K (f - A(I_M u_h))(\chi - \mathcal{M}_K \chi)\, dx + \sum_{E \in E_h^{int}} \int_E J_{E,n}(u_h)\chi\, ds.$$

The conclusion now follows from the fact that

$$\int_E J_{E,n}(u_h)\,ds = 0 \quad \forall E \in E_h^{int},$$

which is due to (5.1) and the principle of conservation of flux: $F_{K,E}^D(u_h) = -F_{L,E}^D(u_h)$ if $E = K \cap L$, $K, L \in T_h$. ∎

COROLLARY 5.4. *Let the assumptions of Lemma* 5.3 *be satisfied. Then, the following estimate holds:*

$$|I_1(\chi)| \lesssim \Big\{ \sum_{K\in T_h} \Big(\alpha_K^2 \|f - A(I_M u_h)\|_K^2 + \varepsilon^{-1/2}\alpha_K \sum_{E\in E_K\cap E_h^{int}} \|J_{E,n}(u_h)\|_E^2 \Big) \Big\}^{1/2} m_1(\chi, T_h)|||\chi|||. \tag{5.6}$$

Proof. The identity (5.5) and the Cauchy–Schwarz inequality yield

$$|I_1(\chi)| \lesssim \sum_{K\in T_h} \|f - A(I_M u_h)\|_K \|\chi - \mathcal{M}_K\chi\|_K + \sum_{K\in T_h} \sum_{E\in E_K\cap E_h^{int}} \|J_{E,n}(u_h)\|_E \|\chi - \mathcal{M}_E\chi\|_E.$$

The discrete Cauchy–Schwarz inequality then leads to

$$\begin{aligned} |I_1(\chi)| \lesssim{}& \Big(\sum_{K\in T_h} \alpha_K^2 \|f - A(I_M u_h)\|_K^2 \Big)^{1/2} \Big(\sum_{K\in T_h} \alpha_K^{-2} \|\chi - \mathcal{M}_K\chi\|_K^2 \Big)^{1/2} \\ &+ \Big(\sum_{K\in T_h} \varepsilon^{-1/2}\alpha_K \sum_{E\in E_K\cap E_h^{int}} \|J_{E,n}(u_h)\|_E^2 \Big)^{1/2} \\ &\cdot \Big(\sum_{K\in T_h} \varepsilon^{1/2}\alpha_K^{-1} \sum_{E\in E_K\cap E_h^{int}} \|\chi - \mathcal{M}_E\chi\|_E^2 \Big)^{1/2}. \end{aligned}$$

We conclude by applying Lemma 3.5. ∎

Remark 5.5. The fundamental properties above are based only on the definition of the scheme (2.5), on the continuity of the interpolant, and on the interpolation properties (5.1), (5.2), and (5.3). Therefore our further analysis is valid for any finite element $(K, P_K^{\mathbf{v},b}, \Sigma_K^{\mathbf{v},b})$ such that the associated interpolant satisfies these properties. But the finite element and the definition of the interpolant should be appropriately chosen in order to guarantee the convergence of $I_M u_h$ to the exact solution u. This convergence analysis is yet to be performed but, in any case, it is outside the scope of this paper.

6. Error estimators.

6.1. Residual error estimators. The exact element residual is defined by $R_K := f - A(I_M u_h)$ on K. As usual, this exact residual is replaced by a certain finite-dimensional approximation $r_K \in \mathbb{P}_k(K)$ called an approximate element residual.

DEFINITION 6.1 (residual error estimator). *The local and global residual error estimators are defined by*

$$\eta_K^2 := \alpha_K^2 \|r_K\|_K^2 + \varepsilon^{-1/2}\alpha_K \sum_{E\in E_K\cap E_h^{int}} \|J_{E,n}(u_h)\|_E^2, \qquad \eta^2 := \sum_{K\in T_h} \eta_K^2.$$

The local and global approximation terms are defined by

$$\zeta_K^2 := \alpha_K^2 \sum_{K' \subset \omega_K} \|R_{K'} - r_{K'}\|_{K'}^2, \qquad \zeta^2 := \sum_{K \in T_h} \zeta_K^2.$$

6.2. Upper error bound.

THEOREM 6.2. *Let u be a solution of (2.2) and u_h a solution of (2.5), and denote the error by $e := u - I_M u_h$. Then the error is bounded as follows:*

$$|||e||| \lesssim m_1(e, T_h)(\eta + \zeta). \tag{6.1}$$

Proof. As e belongs to $H_0^1(\Omega)$, Green's formula yields

$$\int_\Omega (\mathrm{div}\,(\mathbf{v}e)e + b|e|^2)\,dx = \int_\Omega \left(\frac{1}{2}\mathrm{div}\,\mathbf{v} + b\right)|e|^2\,dx.$$

Therefore we have

$$|||e|||^2 = \int_\Omega (\varepsilon \nabla e \cdot \nabla e + (\mathrm{div}\,(\mathbf{v}e) + be)e)\,dx,$$

or equivalently, with the notation from Lemma 5.3,

$$|||e|||^2 = I_1(e).$$

The conclusion immediately follows from estimate (5.6). ∎

6.3. Lower error bound.

THEOREM 6.3. *Assume that there exists $\delta_1 > 2$ and $\delta_2 \in [3/2, 2)$ such that*

$$\begin{cases} -\mathrm{div}\,\mathbf{v} \le \delta_1 b & \text{in } \{x \in \Omega : \mathrm{div}\,\mathbf{v}(x) > 0\}, \\ -\mathrm{div}\,\mathbf{v} \le \delta_2 b & \text{in } \{x \in \Omega : \mathrm{div}\,\mathbf{v}(x) < 0\}. \end{cases} \tag{6.2}$$

Then for all elements K, the following local lower error bound holds:

$$\eta_K \lesssim (1 + Pe_{\omega_K} + \Gamma_{\omega_K})|||e|||_{\omega_K} + \zeta_K, \tag{6.3}$$

where $|||e|||_\omega^2 := \int_\omega (\varepsilon|\nabla e|^2 + (\frac{1}{2}\mathrm{div}\,\mathbf{v} + b)|e|^2)$; $Pe_{\omega_K} = \max_{K' \subset \omega_K} Pe_{K'}$ is the local patch Peclet number, with the local mesh Peclet number being defined as usual by

$$Pe_K = \varepsilon^{-1}\|\mathbf{v}\|_{\infty,K} h_{min,K};$$

see [11, 2, 19] *for anisotropic meshes and* [32, 35] *for isotropic meshes. The element-wise quantity Γ_{ω_K} is defined similarly; namely, $\Gamma_{\omega_K} = \max_{K' \subset \omega_K} \Gamma_{K'}$, where (cf.* [3, 2])

$$\Gamma_K = \max\{1, \alpha_K \|b + \mathrm{div}\,\mathbf{v}\|_{\infty,K}^{1/2}\}.$$

Proof. Since $b + \frac{1}{2}\mathrm{div}\,\mathbf{v} \ge 0$, we readily see that (6.2) is equivalent to

$$|b + \mathrm{div}\,\mathbf{v}| \le \gamma\left(b + \frac{1}{2}\mathrm{div}\,\mathbf{v}\right) \text{ in } \Omega, \tag{6.4}$$

with $\gamma = 2\max\{\frac{\delta_1 - 1}{\delta_1 - 2}, \frac{\delta_2 - 1}{2 - \delta_2}\} > 0$.

Element residual. For a fixed element K define $w_K = r_K b_K$, which belongs to $H_0^1(\Omega)$. From the definition of R_K we have

$$\begin{aligned}\int_K r_K w_K &= \int_K (r_K - R_K) w_K + \int_K R_K w_K = \int_K (r_K - R_K) w_K + \int_K A(u - I_M u_h) w_K \\ &= \int_K (r_K - R_K) w_K + \int_K (\operatorname{div}(\mathbf{v}e) + be) w_K - \int_K \operatorname{div}(\varepsilon \nabla e) w_K.\end{aligned}$$

Integrating by parts in that last term we obtain

$$\int_K r_K w_K = \int_K (r_K - R_K) w_K + \int_K \left(\varepsilon \nabla e \cdot \nabla w_K + (\operatorname{div}(\mathbf{v}e) + be) w_K\right).$$

Leibniz's rule and the Cauchy–Schwarz inequality yield

$$\begin{aligned}\int_K r_K w_K = \|r_K - R_K\|_K \|w_K\|_K &+ (\varepsilon \|\nabla w_K\|_K + \|\mathbf{v}\|_{\infty,K} \|w_K\|_K) \|\nabla e\|_K \\ &+ \|(\operatorname{div} \mathbf{v} + b) e\|_K \|w_K\|_K.\end{aligned}$$

Using property (6.4), we obtain

$$\begin{aligned}\int_K r_K w_K \le \|r_K - R_K\|_K \|w_K\|_K &+ (\varepsilon \|\nabla w_K\|_K + \|\mathbf{v}\|_{\infty,K} \|w_K\|_K) \|\nabla e\|_K \\ &+ \gamma^{1/2} \|\operatorname{div} \mathbf{v} + b\|_{\infty,K}^{1/2} \left\| \left(\frac{1}{2} \operatorname{div} \mathbf{v} + b\right)^{1/2} e \right\|_K \|w_K\|_K.\end{aligned}$$

By the inverse inequalities (3.1), (3.2) we get

$$\begin{aligned}\alpha_K \|r_K\|_K \lesssim \xi_K &+ \alpha_K (\varepsilon h_{min,K}^{-1} + \|\mathbf{v}\|_{\infty,K}) \|\nabla e\|_K \\ &+ \alpha_K \|\operatorname{div} \mathbf{v} + b\|_{\infty,K}^{1/2} \left\| \left(\frac{1}{2} \operatorname{div} \mathbf{v} + b\right)^{1/2} e \right\|_K.\end{aligned}$$

Using the property $\alpha_K \le \varepsilon^{-1/2} h_{min,K}$, we conclude that

$$\alpha_K \|r_K\|_K \lesssim (1 + Pe_K + \Gamma_K) |||e|||_K + \zeta_K. \tag{6.5}$$

Normal jump. Fix an arbitrary edge/face $E \in E_h^{int}$. Recall that $J_{E,n}(u_h) \in \mathbb{P}_k(E)$ for some $k \in \mathbb{N}$ and let

$$w_E := \mathrm{F}_{\mathrm{ext}}(J_{E,n}(u_h)) b_{E,\gamma_{E,K}} \text{ on } K_{E,\gamma_{E,K}} \subset K \subset \omega_E,$$

where $K_{E,\gamma_{E,K}}$ is the squeezed element associated with K (defined in [36, 15, 19] for triangles and tetrahedra and easily extended to rectangles; we do not go into detail here, though we note that the main properties of $K_{E,\gamma_{E,K}}$ are that it be included in K to have E as edge/face and be of height $\sim \gamma_{E,K} h_{E,K}$) and the parameter $\gamma_{E,K}$ is fixed and is equal to

$$\gamma_{E,K} = \min\left\{1, \frac{h_{min,K}}{h_{E,K}}, \frac{\varepsilon^{1/2}}{c_K^{1/2} h_{E,K}}\right\}.$$

Here the function $b_{E,\gamma_{E,K}}$ is the edge/face bubble associated with E in $K_{E,\gamma_{E,K}}$.

The difference with the choice made in [19] relies on the factor c_K.

By elementwise partial integration we get

$$\int_E J_{E,n}(u_h)w_E = -\varepsilon \sum_{K\subset\omega_E}\int_{\partial K}\frac{\partial e}{\partial n_K}w_E = -\sum_{K\subset\omega_E}\int_K (\varepsilon\nabla e\nabla w_E + \operatorname{div}(\varepsilon\nabla e)w_E)$$
$$= \sum_{K\subset\omega_E}\int_K (-\varepsilon\nabla e\nabla w_E + Aew_E - (\operatorname{div}(\mathbf{v}e) + be)w_E).$$

By Leibniz's rule, the Cauchy–Schwarz inequality, and condition (6.4) we obtain

$$\int_E J_{E,n}(u_h)w_E \lesssim \sum_{K\subset\omega_E}\Big(\|R_K\|_K\|w_E\|_K + (\varepsilon\|\nabla w_E\|_K + \|\mathbf{v}\|_{\infty,K}\|w_E\|_K)\|\nabla e\|_K$$
$$+ \|\operatorname{div}\mathbf{v} + b\|_{\infty,K}^{1/2}\|\Big(\frac{1}{2}\operatorname{div}\mathbf{v} + b\Big)^{1/2}e\|_K\|w_E\|_K\Big).$$

The inverse inequalities from Lemma 2 of [19] (see our Lemma 3.1 and the above properties of $K_{E,\gamma_{E,K}}$) in the previous estimate lead to

$$\|J_{E,n}(u_h)\|_E \lesssim \sum_{K\subset\omega_E}\gamma_{E,K}^{1/2}h_{E,K}^{1/2}[(\varepsilon\min\{\gamma_{E,K}h_{E,K}, h_{min,K}\}^{-1} + \|\mathbf{v}\|_{\infty,K})\|\nabla e\|_K$$
$$+ \|\operatorname{div}\mathbf{v} + b\|_{\infty,K}^{1/2}\|\Big(\frac{1}{2}\operatorname{div}\mathbf{v} + b\Big)^{1/2}e\|_K + \|R_K\|_K].$$

Multiplying this estimate by $\varepsilon^{-1/4}\alpha_K^{1/2}$ and using the definition of $\gamma_{E,K}$ we arrive at

(6.6)
$$\varepsilon^{-1/4}\alpha_K^{1/2}\|J_{E,n}(u_h)\|_E \lesssim (1 + Pe_{\omega_E})\varepsilon^{1/2}\|\nabla e\|_{\omega_E} + \Gamma_{\omega_E}\|\Big(\frac{1}{2}\operatorname{div}\mathbf{v} + b\Big)^{1/2}e\|_{\omega_E} + \zeta_K.$$

The conclusion follows from estimates (6.5) and (6.6). □

Remark 6.4. Condition (6.2) is not restrictive since it is satisfied in the so-called convection-dominated case $b + \frac{1}{2}\operatorname{div}\mathbf{v} \geq c_0 > 0$ in Ω. It also holds if $\operatorname{div}\mathbf{v} + b \geq 0$ and $b \geq 0$ (a.e.) in Ω.

7. Numerical results. In this section we present two examples that illustrate the efficiency and reliability of our estimator and show that our estimator is appropriate for adaptivity. Additionally, for both examples we provide the order of convergence of the error $|||u - I_M u_h|||$. In both cases the order is approximately h, which confirms that $I_M u_h$ is a good approximation to u. The first and second examples concern problems with solutions which exhibit boundary layers and for which the use of anisotropic meshes is recommended. The third example even treats the case when the meshes are not fully aligned with the solution. Finally, for a convection-diffusion problem we present two sequences of meshes obtained by an adaptive process. As the meshes are refined in the region of large variation of the solutions we can verify the reliability of our estimator.

For both examples we investigate the main theoretical results which are the upper and lower error bounds. In order to present the inequalities (6.1) and (6.3) appropriately, we consider the ratios

$$q_{up} := \frac{|||u - I_M u_h|||}{\eta + \xi}, \qquad q_{low} := \max_{K\in T_h}\frac{\eta_K}{(1 + Pe_{\omega_K} + \Gamma_{\omega_K})|||e|||_{\omega_K} + \zeta_K}$$

as a function of the degrees of freedom. The first ratio q_{up} is frequently referred to as the effectivity index. It measures the reliability of the estimator and is related to the global upper bound on the error. The second ratio is related to the local lower error bound and measures the efficiency of the estimator. From our theoretical considerations, both ratios should be bounded from above, which is confirmed below experimentally. Hence our estimator is reliable and efficient.

7.1. A reaction-diffusion problem (Test 1). As a particular problem, in (2.1) we take $\mathbf{v} = (0,0)^\top$ and $b = 1$ in the unit square; i.e., we consider

$$-\varepsilon\ \Delta u\ +\ u =\ 0 \quad \text{in} \quad \Omega = (0,1)^2.$$

We prescribe the exact solution to be

$$u(x,y) = \exp\left(-x/\sqrt{\varepsilon}\right),$$

with the Dirichlet boundary datum being fixed accordingly. This solution exhibits an exponential boundary layer of width $O(\sqrt{\varepsilon}|\ln\varepsilon|)$ along the y-axis.

To approximate the solution to this problem appropriately we use anisotropic meshes. Each mesh is the tensor product of a Shishkin-type mesh in the x-direction and a uniform mesh in the y-direction, both with n subintervals. More precisely we fix the transition point $\tau := \min\{1/2, \sqrt{\varepsilon}|\ln\varepsilon|\}$ and define the rectangular mesh of nodes (x_i, y_j), $0 \le i,j \le n$, with (n is assumed to be even)

$$\begin{aligned} &x_i = i2\tau h \text{ for } 0 \le i \le n/2,\\ &x_i = \tau + (2i-n)(1-\tau)h \text{ for } n/2 \le i \le n,\\ &y_j = jh \text{ for } 0 \le j \le n, \end{aligned}$$

where the meshsize h is equal to $1/n$. Such a mesh is illustrated in Figure 7.1 (left) for $h = 1/8$ and $\tau \approx 0.25$.

For each triangulation, we compute the finite volume approximation u_h as a solution to (2.5), and then compute its interpolant $I_M u_h$ using the finite element of subsection 4.3.2 (with the values of $I_M u_h$ at any interior node a being fixed by $I_M u_h(a) = \frac{1}{4}\sum_{a\in K} u_K$). Figure 7.2 presents the energy norm $|||u - I_M u_h|||$ with respect to n for different values of ε. The rate of convergence is approximately h, which confirms that $I_M u_h$ converges quite well to u.

Now we investigate the upper and lower error bounds. According to previous results on finite element methods [15, 19], the meshes are appropriately chosen to resolve the boundary layer, and consequently the alignment measure $m_1(e, T_h)$ should be moderated. Figure 7.3 presents the values of q_{up} (top) and of q_{low} (bottom) with respect to n for different values of ε. We observe that q_{up} is bounded from above by 0.15 and that q_{low} is bounded from above by 0.5.

7.2. Convection-diffusion problems (Test 2). Here we consider the convection-diffusion problem

$$-\varepsilon\ \Delta u\ +\ \operatorname{div}(\mathbf{v}u) =\ 0 \text{ in } \ \Omega = (0,1)^2, \tag{7.1}$$

where $\mathbf{v}$ is either $(-1,0)^\top$ or $(-1,-1)^\top$. We prescribe the exact solution to be, respectively,

$$u(x,y) = \exp\left(-x/\varepsilon\right) \text{ or } u(x,y) = \exp\left(-(x+y)/\varepsilon\right),$$

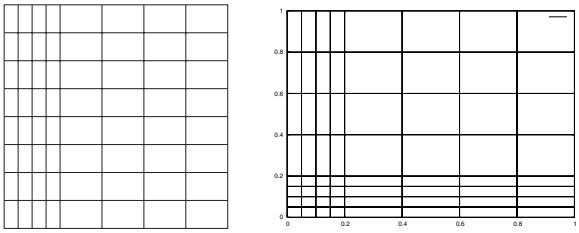


FIG. 7.1. *Shishkin-type meshes on the unit square with* $n = 8$.

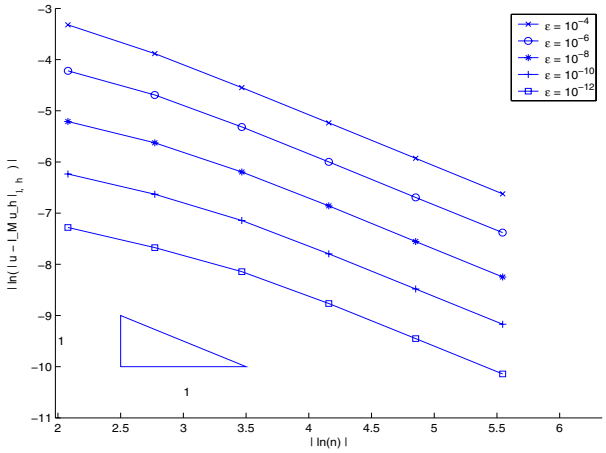


FIG. 7.2. *Illustration of the convergence of* $\|\nabla_h(u - I_M u_h)\|$ *for Test* 1.

with the Dirichlet boundary datum being fixed accordingly. In the first case the solution exhibits an exponential boundary layer of width $O(\varepsilon|\ln\varepsilon|)$ along the y-axis, while in the second case, two exponential boundary layers appear along the x and y axes. Therefore these problems are approximated using similar anisotropic meshes as before with the transition point $\tau := \min\{1/2, \varepsilon|\ln\varepsilon|\}$ (with the tensor product of two Shishkin type meshes in the second case; see Figure 7.1 (right)). Once we have computed the finite volume approximation u_h, solution of (2.5), we compute its interpolant $I_M u_h$ using the finite element of subsection 4.1.2 and the same values at the nodes as before.

Figure 7.4 shows the energy norm $|||u - I_M u_h|||$ with respect to n for different values of ε and confirms a rate of convergence of approximately h.

As before, the meshes are appropriately chosen so the alignment measure $m_1(e, T_h)$ should be moderated. The values of q_{up} and of q_{low} with respect to n for different values of ε are plotted in Figures 7.5 and 7.6. Here we observe similar values for both examples; moreover we see that q_{up} is bounded from above by 0.1 and that q_{low} is approximately 2.2.

7.3. Meshes not aligned with the solutions (Test 3). Here we consider the convection-diffusion problem (7.1) with $\mathbf{v} = (-\cos\theta, \sin\theta)^\top$, with $\theta = 0, 0.1\pi$, and 0.2π, and its prescribed solution given by

$$u(x, y) = \exp\left((-\cos\theta x + \sin\theta y)/\varepsilon\right)$$

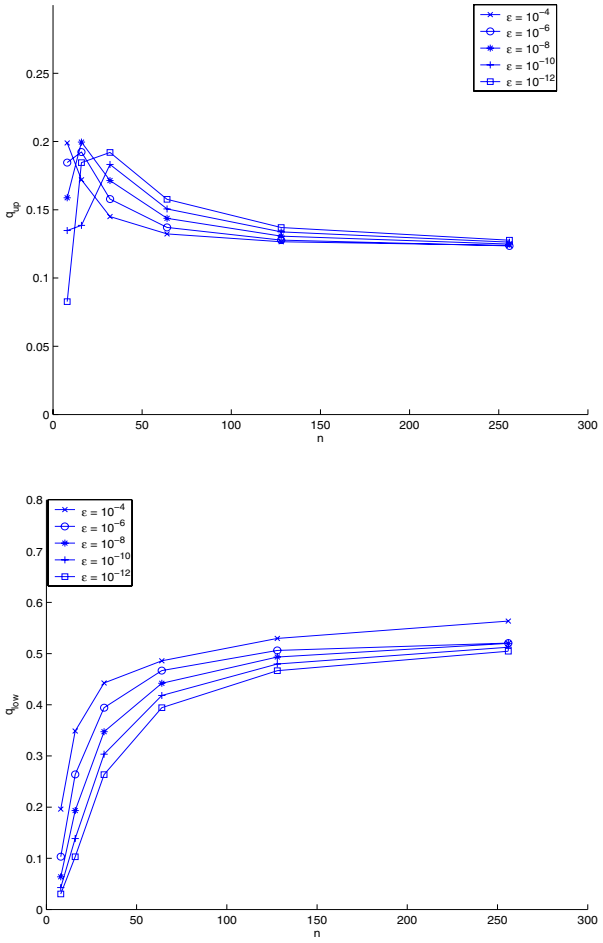


FIG. 7.3. q_{up} *(top) and* q_{low} *(bottom) with respect to* n *for Test* 1.

and $\varepsilon = 0.05$. This solution has a layer along the line $-\cos\theta x + \sin\theta y = 0$ (the case $\theta = 0$ corresponds to the case of the previous subsection but is presented in order to compare the different results). Here we take the sequence of meshes from subsection 7.1 with the transition point $\tau = 4\varepsilon = 0.2$. This means that for $\theta = 0.1\pi$ and 0.2π, the meshes are not fully aligned with the solution and are less and less aligned as θ increases. Nevertheless the convergence rate is h according to Figure 7.7. Moreover the values of q_{up} and q_{low} with respect to n for the different values of θ are plotted in Figures 7.8, where we see that q_{up} varies between 0.1 and 0.14 and that q_{low} is bounded from above by 2.75. From this figure, we further remark that the effectivity index q_{up} grows as θ increases, implying that the matching function $m_1(e, T_h)$ grows as well. This phenomenon was expected since the meshes are less and less aligned as θ increases.

7.4. An adaptive algorithm (Test 4). In view of the upper bound (6.1), we use the following (standard) adaptive process: Starting from an initial mesh T_{h_0}, we

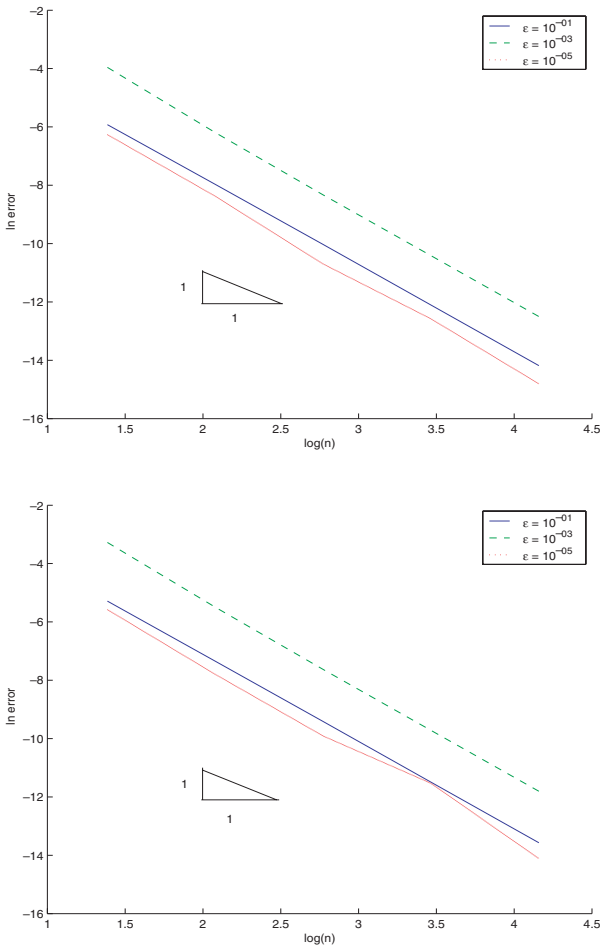


FIG. 7.4. *Illustration of the convergence of* $\|\nabla_h(u - I_M u_h)\|$ *for Test* 2 *for* $\mathbf{v} = (-1, 0)^\top$ (*top*) *and for* $\mathbf{v} = (-1, -1)^\top$ (*bottom*).

mark the elements K for which

$$\eta_K > c \max_{K'} \eta_{K'},$$

for a chosen constant $c \in (0, 1)$. All marked elements are divided into four subelements (standard regular refinement rule), with the other elements being subdivided only to guarantee the conformity of the new mesh. We refine the meshes up to the requested accuracy.

We test this adaptive algorithm for the convection-diffusion problem (7.1) with $\mathbf{v} = (1, 1)^\top$ and the prescribed solution given by

$$u(x, y) = \exp\left(-100((x - 0.5)^2 + (y - 0.5)^2)\right),$$

with the right-hand side and the boundary datum being fixed accordingly. This exact solution is a Gaussian function whose center is the point $(0.5, 0.5)$. For this example we use either meshes consisting of rectangles satisfying the admissibility condition or

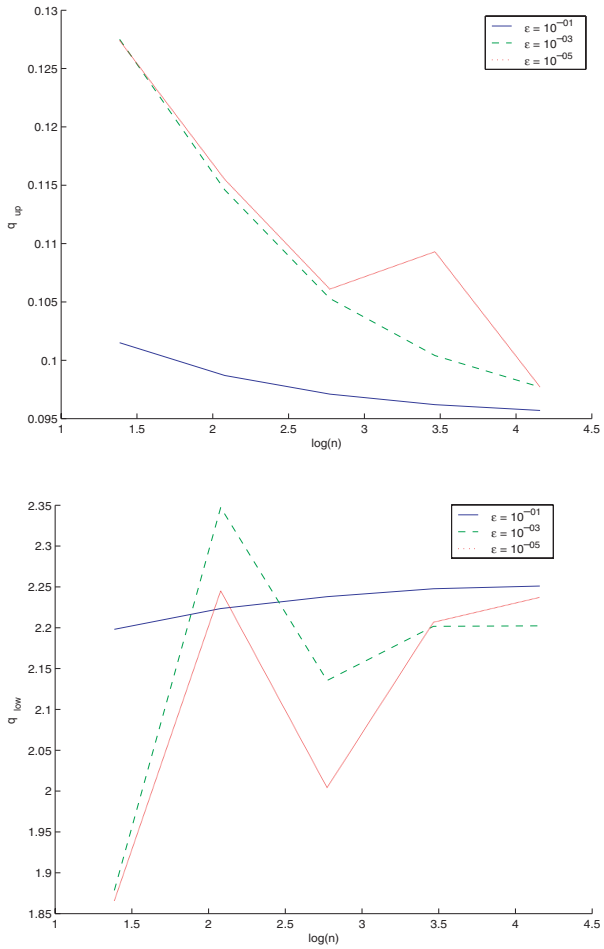


FIG. 7.5. *q_{up} (top) and q_{low} (bottom) with respect to n for Test 2 for* $\mathbf{v} = (-1, 0)^\top$.

meshes consisting of triangles that do not satisfy the admissibility condition (in which case we use the diamond cell technique; see subsection 2.2). The Morley interpolant is computed using the finite elements from subsection 4.1.2 or 4.1.1. The meshes obtained after four iterations are shown in Figure 7.9. In the case of rectangular meshes, anisotropic meshes appear due to the admissibility condition, but this does not affect the convergence of our adaptive process. This also underlines the incompatibility between standard refinement rules and this admissibility condition. From this figure, we can conclude that the meshes are refined in the region of large variation of the solution. Again this confirms the reliability of our estimator.

Figure 7.10 shows the energy norm $|||u - I_M u_h|||$ with respect to the degrees of freedom for the adaptive algorithm in comparison with uniformly refined meshes. In both cases, we see that the adaptive algorithm gives rise to better error bounds and to convergence in h.

The values of q_{up} and q_{low} with respect to the degrees of freedom are plotted in Figures 7.11 and 7.12 and are compared with uniformly refined meshes. There we observe that q_{up} is bounded from above by 0.1 and that q_{low} is bounded from above

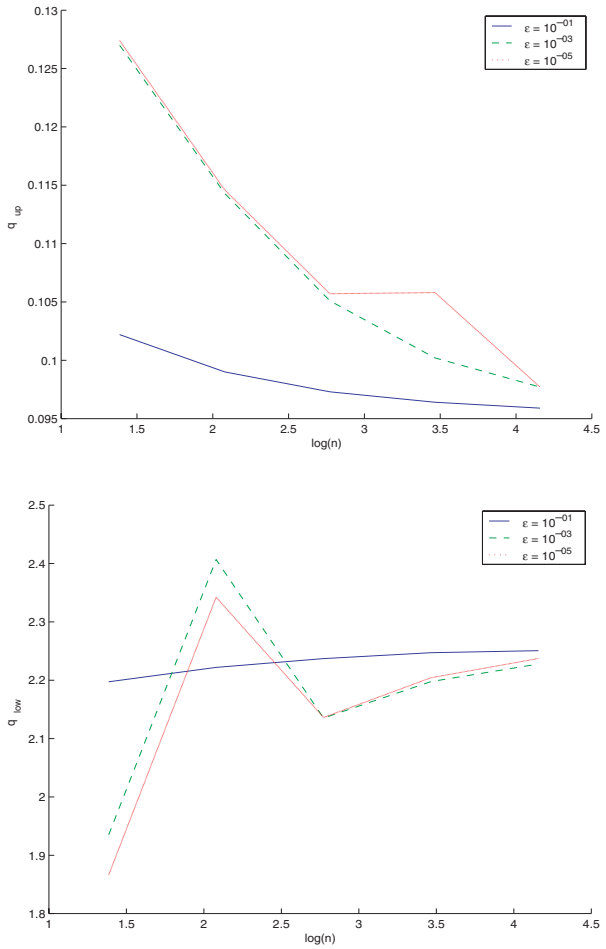


FIG. 7.6. q_{up} *(top) and* q_{low} *(bottom) with respect to* n *for Test* 2 *for* $\mathbf{v} = (-1,-1)^\top$.

by 1.42 (resp., 3.5). From these figures, we can say that the different values of q_{up} remain similar for adaptive meshes and for uniformly refined meshes, while the values of q_{low} have a different behavior but stay in a relatively small range of variations. For rectangular meshes, uniform meshes lead to smallest values of q_{low}, probably due to the overrefinement of adaptive meshes, while the converse holds for triangular meshes because adaptive triangular meshes fit the solution well.

8. Conclusions. We have proposed and rigorously analyzed a new a posteriori error estimate for a cell centered finite volume approximation of diffusion-convection-reaction equations. We have shown that this estimate is reliable and efficient. This estimate is based on the construction of an appropriate interpolant of Morley type. Some numerical experiments confirm our theoretical predictions.

Acknowledgment. I am very grateful to Karim Djadel (Cermics, ENPC, France), who made the numerical experiments presented in section 7.

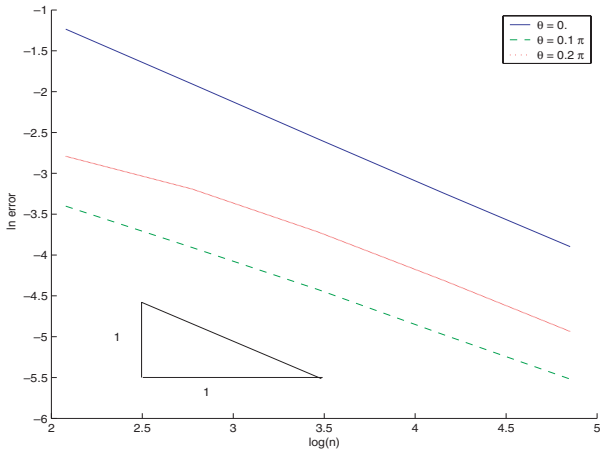


FIG. 7.7. *Illustration of the convergence of* $\|\nabla_h(u - I_M u_h)\|$ *for Test* 3.

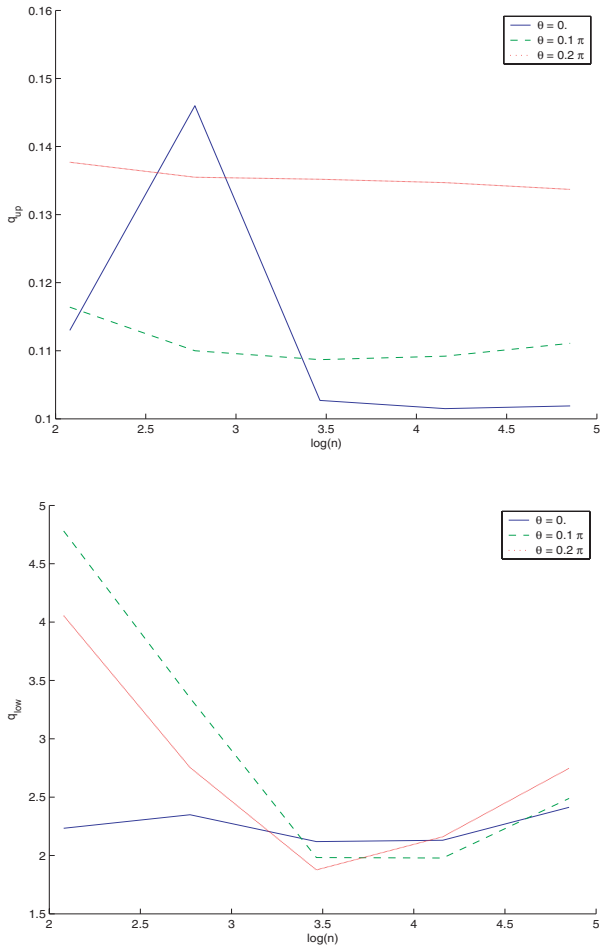


FIG. 7.8. q_{up} *(top) and* q_{low} *(bottom) with respect to* n *for Test* 3.

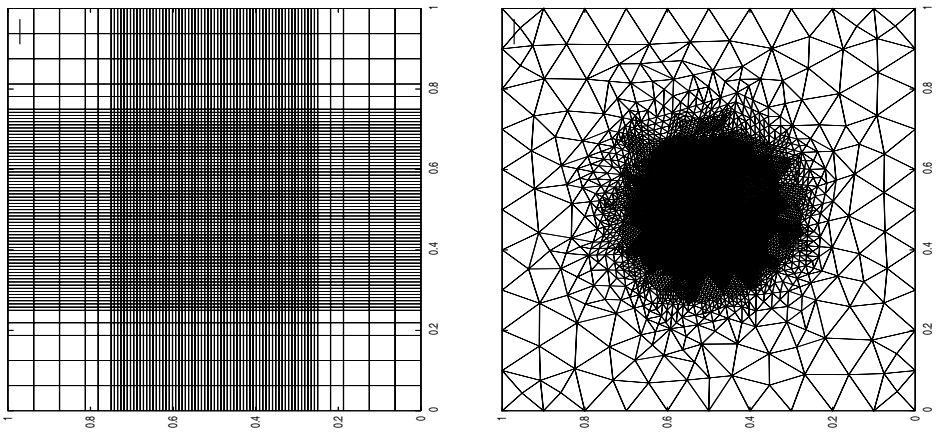


FIG. 7.9. *The last meshes in the refinement sequences for Test* 4.

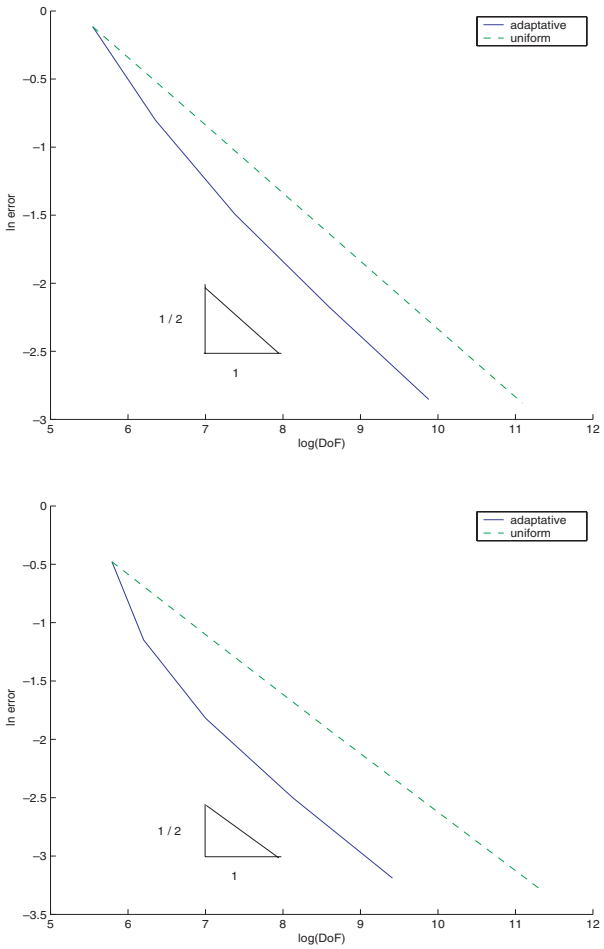


FIG. 7.10. *Illustration of the convergence of* $\|\nabla_h(u - I_M u_h)\|$ *for Test* 4 *for rectangular meshes* (*top*) *and for triangular meshes* (*bottom*).

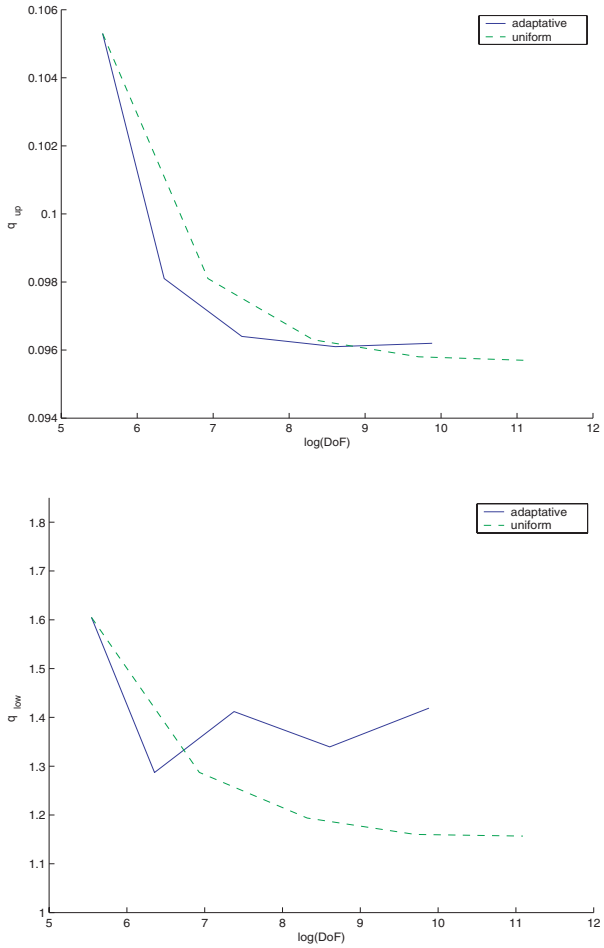


FIG. 7.11. q_{up} *(top) and* q_{low} *(bottom) with respect to* n *for Test* 4 *for rectangular meshes.*

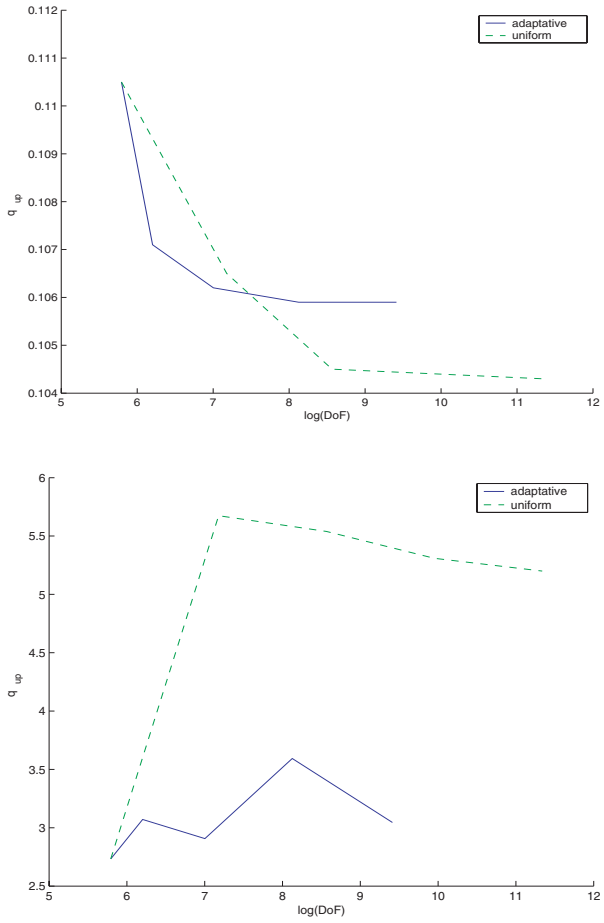


FIG. 7.12. q_{up} (*top*) *and* q_{low} (*bottom*) *with respect to* n *for Test* 4 *for triangular meshes.*

REFERENCES

[1] A. Agouzal and F. Oudin, *A posteriori error estimator for finite volume methods*, Appl. Math. Comput., 110 (2000), pp. 239–250.

[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[3] T. Apel and G. Lube, *Anisotropic mesh refinement in stabilized Galerkin methods*, Numer. Math., 74 (1996), pp. 261–282.

[4] A. Bergam and Z. Mghazli, *Estimateurs a posteriori d'un schéma de volumes finis pour un problème non linéaire*, C. R. Acad. Sci. Paris Sér. I Math. 331 (2000), pp. 475–478.

[5] A. Bergam, Z. Mghazli, and R. Verfürth, *A posteriori estimators for the finite volume discretization of an elliptic problem*, Numer. Math., 95 (2003), pp. 599–624.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[7] Y. Coudière, J.-P. Villa, and P. Villedieu, *Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[8] Y. Coudière and P. Villedieu, *Convergence rate of a finite volume scheme for the linear convection-diffusion equation on locally refined meshes*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1123–1149.

[9] J. Droniou, *Error estimates for the convergence of a finite volume discretization of convection-diffusion equations*, J. Numer. Math., 11 (2003), pp. 1–32.

[10] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, vol. 7, P. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 723–1020.

[11] R. Hangleiter and G. Lube, *Stabilized Galerkin methods and layer-adapted grids for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 165–182.

[12] R. Herbin and M. Ohlberger, *A posteriori error estimate for finite volume approximations of convection diffusion problems*, in Finite Volumes for Complex Applications III, R. Herbin and D. Kröner, eds., Lab. Anal. Topol. Probab. CNRS, Paris, 2002, pp. 739–746.

[13] N. Jullian, *An error indicator for cell-centered finite volumes for linear convection-diffusion problems*, in Finite Volumes for Complex Applications III, R. Herbin and D. Kröner, eds., Lab. Anal. Topol. Probab. CNRS, Paris, 2002, pp. 763–770.

[14] D. Kröner and M. Ohlberger, *A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions*, Math. Comp., 69 (1999), pp. 25–39.

[15] G. Kunert, *A Posteriori Error Estimation for Anisotropic Tetrahedral and Triangular Finite Element Meshes*, Logos Verlag, Berlin, 1999; Ph.D. thesis version available online at http://archiv.tu-chemnitz.de/pub/1999/0012/index.html.

[16] G. Kunert, *An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes*, Numer. Math., 86 (2000), pp. 471–490.

[17] G. Kunert, *A local problem error estimator for anisotropic tetrahedral finite element meshes*, SIAM J. Numer. Anal., 39 (2001), pp. 668–689.

[18] G. Kunert, *Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes*, Adv. Comp. Math., 15 (2001), pp. 237–259.

[19] G. Kunert, *A posteriori error estimation for convection dominated problems on anisotropic meshes*, Math. Methods Appl. Sci., 26 (2003), pp. 589–617.

[20] G. Kunert, Z. Mghazli, and S. Nicaise, *A Posteriori Error Estimation for a Finite Volume Discretization on Anisotropic Meshes*, preprint SFB393/03–016, TU Chemnitz, Chemnitz, Germany, 2003. Available online at http://archiv.tu-chemnitz.de/pub/2003/0005/index.html.

[21] R. Lazarov and S. Tomov, *Adaptive finite volume element method for convection-diffusion-reaction problems in 3-D*, in Scientific Computing and Application, Y. W. P. Minev and Y. Lin, eds., Nova Sci. Publ., Huntington, NY, 2001, pp. 91–106.

[22] R. Lazarov and S. Tomov, *A posteriori error estimates for finite volume approximations of convection-diffusion-reaction equations*, Comput. Geosci., 6 (2002), pp. 483–503.

[23] J. Mackenzie, T. Sonar, and G. Warnecke, *A posteriori error estimates for the cell-vertex finite volume method*, in Adaptive Methods: Algorithms, Theory and Applications, W. Hackbusch and G. Wittum, eds., Vieweg, Braunschweig, 1994, pp. 221–235.

[24] L. Morley, *The triangular equilibrium element in the solution of plate bending problems*, Aero. Quarterly, 19 (1968), pp. 149–169.

[25] K. W. Morton and E. Süli, *A posteriori and a priori error analysis of finite volume methods*, in The Mathematics of Finite Elements and Applications, J. R. Whiteman, ed., John Wiley and Sons, New York, 1994, pp. 267–288.
[26] S. Nicaise, *A posteriori error estimations of some cell-centered finite volume methods*, SIAM J. Numer. Anal., 43 (2005), pp. 1481–1503.
[27] S. Nicaise, *A posteriori residual error estimation of a cell-centered finite volume method*, C. R. Acad. Sci. Paris, Sér., 338 (2004), pp. 419–424.
[28] T. K. Nilssen, X.-C. Tai, and R. Winther, *A robust nonconforming H^2-element*, Math. Comp., 70 (2000), pp. 489–505.
[29] M. Ohlberger, *A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection-diffusion equations*, Numer. Math., 87 (2001), pp. 737–761.
[30] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Anal. Numer., 35 (2001), pp. 355–387.
[31] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Ser. Comput. Methods Mech. Thermal Sci., McGraw–Hill, New York, 1980.
[32] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*, Springer, Berlin, 1996.
[33] T. Sonar and E. Süli, *A dual graph-norm refinement indicator for finite volume approximations of the Euler equations*, Numer. Math., 78 (1998), pp. 619–658.
[34] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley–Teubner, Chichester, Stuttgart, 1996.
[35] R. Verfürth, *A posteriori error estimators for convection-diffusion equations*, Numer. Math., 80 (1998), pp. 641–663.
[36] R. Verfürth, *Robust a posteriori error estimators for the singularly perturbed Helmholtz equation*, Numer. Math., 78 (1998), pp. 479–493.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 979–1004

DISCRETE COMPACTNESS FOR THE *hp* VERSION OF RECTANGULAR EDGE FINITE ELEMENTS*

DANIELE BOFFI†, MARTIN COSTABEL‡, MONIQUE DAUGE‡, AND LESZEK DEMKOWICZ§

Abstract. Discretization of Maxwell eigenvalue problems with edge finite elements involves a simultaneous use of two discrete subspaces of H^1 and $\mathbf{H}(\mathrm{curl})$, reproducing the exact sequence condition. Kikuchi's discrete compactness property, along with appropriate approximability conditions, implies the convergence of discrete eigenpairs to the exact ones. In this paper we prove the discrete compactness property for the edge element approximation of Maxwell's eigenpairs on general *hp* adaptive rectangular meshes. Hanging nodes, yielding 1-irregular meshes, are covered, and the order of the used elements can vary from one rectangle to another, thus allowing for a real *hp* adaptivity. As a particular case, our analysis covers the convergence result for the *p*-method.

Key words. Maxwell equations, *hp* finite elements, discrete compactness, edge elements, eigenvalue approximation

AMS subject classifications. 65N30, 65N25, 78M10

DOI. 10.1137/04061550X

1. Introduction.

The importance of the exact sequence

$$H^1 \xrightarrow{\mathbf{grad}} \mathbf{H}(\mathrm{curl}) \xrightarrow{\mathbf{curl}} \mathbf{H}(\mathrm{div}) \xrightarrow{\mathrm{div}} L^2$$

has been recognized in the analysis of Maxwell equations [2, 12, 13]. In two space dimensions the sequence reduces to

$$H^1 \xrightarrow{\mathbf{grad}} \mathbf{H}(\mathrm{curl}) \xrightarrow{\mathrm{curl}} L^2.$$

In this paper we shall deal with the two-dimensional (2D) case.

The fundamental idea behind the construction of *edge elements* is based on the reproduction of the sequence at the discrete level. This idea had been also successfully exploited in the framework of mixed finite elements for elliptic problems, where it is also known as *commuting diagram property* [34].

Thus, we shall consider discrete subspaces of H^1 and $\mathbf{H}(\mathrm{curl})$ forming part of the discrete exact sequence. It is in this context that Kikuchi [35] introduced the fundamental notion of the *discrete compactness property* which, along with appropriate approximability properties, guarantees the convergence of discrete Maxwell eigenvalues to the exact ones. We also refer the reader to the book [17] for a definition corresponding to discrete compactness in an abstract setting. In another important contribution Caorsi, Fernandes, and Raffetto [16] have demonstrated that the discrete compactness

*Received by the editors September 21, 2004; accepted for publication (in revised form) November 2, 2005; published electronically May 19, 2006.
http://www.siam.org/journals/sinum/44-3/61550.html
†Dipartimento di Matematica "F. Casorati," Università di Pavia, I-27100 Pavia, Italy (daniele.boffi@unipv.it, http://www-dimat.unipv.it/boffi/).
‡IRMAR, Institut Mathématique, Université de Rennes 1, 35042 Rennes, France (costabel@univ-rennes1.fr, http://perso.univ-rennes1.fr/martin.costabel/, dauge@univ-rennes1.fr, http://perso.univ-rennes1.fr/monique.dauge/).
§The Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712 (leszek@ices.utexas.edu, http://www.ices.utexas.edu/∼leszek/).

property is not only a sufficient but also a necessary condition for the convergence of Maxwell eigenvalues without the appearance of any spurious mode.

Let us recall that the difficulty related with the Maxwell curl curl operator is the lack of ellipticity manifested by the presence of the infinite dimensional kernel formed by the gradients. Nevertheless, the Maxwell problem is well posed as soon as the divergence-free constraint is imposed. With this constraint, the Maxwell problem recovers ellipticity properties due to the compact embedding of $\mathbf{H}_0(\mathrm{curl}) \cap \mathbf{H}(\mathrm{div})$ into L^2. The discrete compactness is the correct discrete analogue of the above compact embedding.

For the sake of clarity let us give the definition of discrete compactness property. As with the usual approximability properties, it is related to a sequence of discrete spaces. Let $(\mathbf{X}_n)_n$ be a sequence of finite dimensional subspaces of $\mathbf{H}_0(\mathrm{curl})$ and $(\mathbb{Q}_n)_n$ a related sequence of subspaces of H^1. We say that the sequence $(\mathbb{Q}_n, \mathbf{X}_n)_n$ has the discrete compactness property if the following holds:

Any sequence $\mathbf{u}_n \in \mathbf{X}_n$ of *discrete divergence-free* fields, i.e., satisfying

$$(\mathbf{u}_n, \mathbf{grad}\, \phi_n) = 0 \quad \forall \phi_n \in \mathbb{Q}_n$$

and uniformly bounded in $\mathbf{H}(\mathrm{curl})$, has a subsequence converging in L^2.

The discrete compactness property has been extensively studied in the framework of the h version of edge finite elements, where it is well known to hold true for a variety of edge finite elements on quite general 2D and three-dimensional (3D) meshes (see the review papers [19, 33], the book [38], and the references therein, among which we recall in particular [6, 7, 9, 14, 16, 35, 39]), but it has not been widely investigated for the p and hp version yet. On the other hand, electromagnetic devices very often involve complicated geometries, which in particular may be neither smooth nor convex. The analysis of the singularities arising from reentrant corners or edges and from material discontinuities (see [18, 20]) shows that such situations are to be handled with care. When using edge elements, one might want to locally adapt the meshsize h and the approximation order p, which can possibly vary from one element to another within the same mesh. Such an *hp strategy* is an excellent way to get accurate results (an *exponential convergence* is expected and observed) when even severe singularities are present (see [23, 42] for examples of hp finite element (FE) implementations).

In [8], the analysis of the discrete compactness property for triangular hp finite elements has been tackled, but the proof of the main result relied on a conjectured L^2 estimate which had only been demonstrated numerically. Even for the pure p method, there is no result in this direction available in the literature. In [37] the p version of edge elements has been considered, but the proved results do not apply to eigenvalue approximations.

In this paper, we consider the 2D case of rectangular elements. A rigorous proof of the discrete compactness property is provided for edge elements of the first Nédélec family. Our hypotheses allow for a complete hp refinement, including the presence of hanging nodes. The pure p version of edge elements, being a subset of our setting, is naturally covered by our analysis. The same proof applies to meshes of quadrilaterals obtained by affine transformation from the reference square (i.e., parallelograms) and, more generally, to meshes obtained using the so-called *algebraic mesh generators*.

The case of unstructured quadrilateral meshes presents some issues: It is known that the h version of standard edge elements does not provide optimal results in this case, and even in the lowest order case there is no convergence at all; see [3].

Nevertheless, the results of [31] show that the discrete compactness property holds in the framework of the h version for the modified family of edge elements proposed and analyzed in [3]; another, simpler modification has been proposed and analyzed in [11]. But the validity of the discrete compactness property for the p and hp versions of edge elements remains an open problem in the situation of unstructured meshes.

Our presentation starts with the pure p method on a single square element, which is analyzed in section 3 after the introduction of some preliminary notation in section 2. We consider, in particular, full tensor polynomials (the second Nédélec family of edge elements; see [41]) and standard edge element of the first Nédélec family (see [40]). We show that standard edge elements provide convergent approximation, while the second Nédélec family presents several spurious eigenpairs (more precisely, some discrete eigenvalues come with wrong multiplicity). We then present in section 4 the general hp theory which relies on an L^2 estimate which is proved thanks to the evaluation of an inf-sup constant on the reference element (section 4.3). We make use of the hp edge element spaces presented in [42], which generalize the first family of Nédélec finite elements.

We conclude our paper in section 5, where we recall the consequences of the discrete compactness property on the eigenvalue approximation by a Galerkin method: As a result, the kth nonzero eigenvalue of the Galerkin discretization converges to the kth nonzero Maxwell eigenvalue. We discuss the possibility of proving an exponential convergence rate, such as the one obtained in [22] for the discretization of the Maxwell problem by the weighted regularization method. We then finally comment on the extension of our proofs to the situation of general curvilinear polygons, with meshes obtained using algebraic mesh generators.

2. Preliminary notions and notation.

2.1. Polynomial spaces on the reference square. The square is defined as $\Sigma := I^2$, where I is the interval $(-1, 1)$. We denote the coordinates by $\mathbf{x} = (x, y)$. The outward unit normal vector on the boundary $\partial\Sigma$ is $\mathbf{n}$.

Everywhere p denotes an integer $p \geq 1$. The space of polynomials of degree $\leq p$ on I is denoted by $\mathbb{P}^p(I)$, and its subspace of polynomials φ with zero traces, $\varphi(\pm 1) = 0$, is denoted by $\mathbb{P}^p_0(I)$.

On the square, let $\mathbb{Q}^{p,q}(\Sigma)$ be the space of polynomials of separate degrees p and q in x and y, respectively. This can be expressed as

$$\mathbb{Q}^{p,q}(\Sigma) = \mathbb{P}^p(I) \otimes \mathbb{P}^q(I).$$

Symbol $\mathbb{Q}^p$ will be used for isotropic spaces, $\mathbb{Q}^p = \mathbb{Q}^{p,p}$, and $\mathbb{Q}^p_0$ will denote the polynomials with zero traces, $\mathbb{Q}^p_0 = \mathbb{P}^p_0 \otimes \mathbb{P}^p_0$.

We will study in the following two families of polynomial spaces for the electric field $\mathbf{u} = (u_1, u_2)$ on the square Σ.

2.1.1. Full tensor product spaces. $\mathbf{Q}^p(\Sigma)$ denotes the full space $\mathbb{Q}^p(\Sigma) \times \mathbb{Q}^p(\Sigma)$. This is the space of Lagrange nodal elements on the square and, with appropriate degrees of freedom, this also forms the *second Nédélec family* of edge elements.

We denote by $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ its subspace of the fields $\mathbf{u}$ satisfying the perfect electric conductor boundary condition $\mathbf{u} \times \mathbf{n} = 0$ on $\partial\Sigma$. We have

$$\mathbf{Q}^p_{\mathsf{N}}(\Sigma) = \big[\mathbb{P}^p(I) \otimes \mathbb{P}^p_0(I)\big] \times \big[\mathbb{P}^p_0(I) \otimes \mathbb{P}^p(I)\big]. \tag{1}$$

2.1.2. Classical edge elements. $\mathbf{N}^p(\Sigma)$ denotes the space $\mathbb{Q}^{p-1,p}(\Sigma) \times \mathbb{Q}^{p,p-1}(\Sigma)$ which allows the commuting diagram property with the operator **grad** from the space of scalar polynomials $\mathbb{Q}^p(\Sigma)$. These edge element spaces are also known as the *first Nédélec family* of edge elements. For simplicity, we shall refer to the spaces of this section as standard (Nédélec) edge elements. We denote by $\mathbf{N}^p_{\mathsf{N}}(\Sigma)$ the subspace of fields satisfying the electric boundary condition:

$$\mathbf{N}^p_{\mathsf{N}}(\Sigma) = \left[\mathbb{P}^{p-1}(I) \otimes \mathbb{P}^p_0(I)\right] \times \left[\mathbb{P}^p_0(I) \otimes \mathbb{P}^{p-1}(I)\right]. \tag{2}$$

2.2. Maxwell spectrum in the square. In section 3 we shall describe in terms of one-dimensional (1D) problems the Maxwell spectrum computed with the spaces $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ and $\mathbf{N}^p_{\mathsf{N}}(\Sigma)$.

We first recall the definition of the standard continuous spaces associated with Maxwell equations on a domain Ω: $\mathbf{H}(\mathrm{curl}, \Omega)$ is the space of $L^2(\Omega)$ fields with curl in $L^2(\Omega)$, while $\mathbf{H}_0(\mathrm{curl}, \Omega)$ is the subspace of $\mathbf{H}(\mathrm{curl}, \Omega)$ with perfect electric boundary conditions; $\mathbf{H}(\mathrm{div}, \Omega)$ is the space of $L^2(\Omega)$ fields with divergence in $L^2(\Omega)$.

Let us describe the Maxwell spectrum in the continuous space

$$\mathbf{X}_{\mathsf{N}}(\Sigma) := \mathbf{H}_0(\mathrm{curl}, \Sigma) \cap \mathbf{H}(\mathrm{div}, \Sigma),$$

i.e., the eigenpairs $(\lambda, \mathbf{u})$ with $\mathbf{u} \neq 0$ such that

$$\mathbf{u} \in \mathbf{X}_{\mathsf{N}}(\Sigma) : \int_\Sigma \mathrm{curl}\,\mathbf{u}\ \mathrm{curl}\,\mathbf{v}\, d\mathbf{x} = \lambda \int_\Sigma \mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} \quad \forall \mathbf{v} \in \mathbf{X}_{\mathsf{N}}(\Sigma). \tag{3}$$

2.2.1. The kernel. For $\lambda = 0$, we have the whole space $\mathbf{grad}\, H^1_0(\Sigma)$ of kernel elements.

2.2.2. The genuine Maxwell spectrum. The whole nonzero spectrum corresponds to eigenvectors of the form $\mathbf{u} = \mathbf{curl}\,\varphi$ with φ the nonconstant eigenvector of the Neumann problem on Σ. Let $(\psi_j)_{j\geq 0}$ be the basis of the Neumann eigenvectors on the interval I; they are associated with the eigenvalues $j^2\pi^2/4$. Then the Maxwell spectrum on Σ is

$$\lambda_{j,k} = (j^2 + k^2)\pi^2/4, \quad \mathbf{u}_{j,k}(x, y) = (\psi_j(x)\psi'_k(y), -\psi'_j(x)\psi_k(y)), \quad j + k \geq 1. \tag{4}$$

For comparison purposes, it is convenient to split the whole spectrum into the three following parts:

(a) the kernel;

(b) the nonzero Neumann eigenvalues $j^2\pi^2/4$ associated with the two eigenvectors

$$(0, -\psi'_j(x)) \quad \text{and} \quad (\psi'_j(y), 0);$$

(c) the sum of two nonzero Neumann eigenvalues $(j^2 + k^2)\pi^2/4$ with the eigenvectors $\mathbf{u}_{j,k}$ (and $\mathbf{u}_{k,j}$ if $j \neq k$).

Remark 1. For all nonconstant Neumann eigenvectors ψ_j (i.e., for $j \geq 1$), ψ'_j is a Dirichlet eigenvector associated with the eigenvalue $j^2\pi^2/4$. Let us denote $\varphi_j := -\psi'_j$. Then $(j^2\pi^2/4)\psi_j = -\psi''_j = \varphi'_j$, and we can see that eigenvectors associated with part (c) of the spectrum can be written as

$$\left(\frac{1}{j^2}\,\varphi'_j(x)\varphi_k(y), -\frac{1}{k^2}\,\varphi_j(x)\varphi'_k(y)\right). \tag{5}$$

3. Approximation of Maxwell's spectrum in a square by the *p* method. In this section we characterize explicitly the Maxwell spectrum on the square computed with the full polynomial space $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ and with the Nédélec edge element space $\mathbf{N}^p_{\mathsf{N}}(\Sigma)$.

In contrast with (4), where the 1D generators are the *Neumann* eigenvectors on the interval, at the discrete level we will show that the 1D generators are the *Dirichlet* discrete eigenvectors: for $p \geq 2$, we consider the eigenpairs (λ, β) with $\beta \neq 0$ such that

$$\beta \in \mathbb{P}^p_0(I) : \int_I \beta' w' \, dx = \lambda \int_I \beta w \, dx \quad \forall w \in \mathbb{P}^p_0(I). \tag{6}$$

The dimension of $\mathbb{P}^p_0(I)$ is $p-1$; for $j = 1, \ldots, p-1$, let $(\lambda_j^{[p]}, \beta_j^{[p]})$ be an eigenpair basis of (6) satisfying

$$\lambda_1^{[p]} < \lambda_2^{[p]} < \cdots < \lambda_{p-1}^{[p]}.$$

For each j, $\lambda_j^{[p]}$ tends exponentially to $j^2\pi^2/4$ as $p \to \infty$. This follows from the standard convergence analysis for elliptic eigenvalue problems [5] and best approximation properties of the p-version of the finite element method (FEM); see [43, Chapter 3].

We are going to describe the Maxwell spectrum in $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$, i.e., the eigenpairs $(\lambda, \mathbf{u})$ such that

$$\mathbf{u} \in \mathbf{Q}^p_{\mathsf{N}}(\Sigma) : \quad \int_\Sigma \operatorname{curl} \mathbf{u} \operatorname{curl} \mathbf{v} \, d\mathbf{x} = \lambda \int_\Sigma \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x} \quad \forall \mathbf{v} \in \mathbf{Q}^p_{\mathsf{N}}(\Sigma). \tag{7}$$

THEOREM 1. *The whole Maxwell spectrum* (7) *in* $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ *can be split into four parts:*

(a) *the kernel:* $\lambda = 0$ *and* $\mathbf{u} \in \mathbf{grad}(\mathbb{P}^p_0 \otimes \mathbb{P}^p_0)$*;*

(b) *the Dirichlet discrete eigenvalues* $\lambda_j^{[p]}$ *associated with the two eigenvectors*

$$(0, -\beta_j^{[p]}(x)) \quad \textit{and} \quad (\beta_j^{[p]}(y), 0), \quad j = 1, \ldots, p-1; \tag{8}$$

(c) *the sum of two Dirichlet discrete eigenvalues* $\lambda_j^{[p]} + \lambda_k^{[p]}$ *with the eigenvectors*

$$(\lambda_j^{[p]} \beta_k^{[p]\prime}(x) \beta_j^{[p]}(y), -\lambda_k^{[p]} \beta_k^{[p]}(x) \beta_j^{[p]\prime}(y)), \quad 1 \leq j, k \leq p-1; \tag{9}$$

(d) *the Dirichlet discrete eigenvalues* $\lambda_j^{[p]}$ *associated with two* spurious *eigenvectors*

$$(0, -\beta_j^{[p]}(x) L_p(y)) \quad \textit{and} \quad (L_p(x) \beta_j^{[p]}(y), 0), \tag{10}$$

where L_p *denotes the Legendre polynomial of degree* p*.*

Remark 2. Note that formula (5) transforms into $(j^2 \varphi_k'(x) \varphi_j(y), -k^2 \varphi_k(x) \varphi_j'(y))$ by swapping j and k and multiplying by $j^2 k^2$. The similarity with formula (9) is now obvious.

Remark 3. The previous theorem shows that the space $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ is not suited for the computation of Maxwell's eigenvalues. Indeed, the discrete eigenvalues described in part (d) are redundant, providing a wrong multiplicity to the correct eigenvalues described in part (b). Moreover, the discrete eigenvectors of part (d) do not approximate any physical eigenfunction. The fact that the second Nédélec family produces

spurious modes in the h-version of the FEM has been documented in the literature; see, e.g., [29, 30].

Proof. Let us first check the dimensions of the spaces described in the four above cases:

(a) $(p-1)^2$.
(b) $2(p-1)$.
(c) $(p-1)^2$.
(d) $2(p-1)$.

The sum is $2(p-1)(p+1)$, which is the dimension of $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ (see (1)).

It remains to check that the proposed pairs are eigenpairs of (7).

Case (a). The scalar polynomial space $\mathbb{P}^p_0 \otimes \mathbb{P}^p_0$ is contained in $H^1_0(\Sigma)$; therefore all elements of $\mathbf{grad}(\mathbb{P}^p_0 \otimes \mathbb{P}^p_0)$ belong to the kernel.

For the remaining part of the proof, since p is fixed, let us drop the exponent $[p]$ in the notation of the discrete 1D Dirichlet eigenpairs. By integration by parts we note that the discrete eigenpairs (λ_j, β_j) satisfy

$$\int_I (\beta_j'' + \lambda_j \beta_j)\, w\, dx = 0 \quad \forall w \in \mathbb{P}^p_0(I). \tag{11}$$

On the other hand, again by integration by parts, we obtain that $(\lambda, \mathbf{u})$ is an eigenpair in $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$ if and only if

$$\mathbf{u} \in \mathbf{Q}^p_{\mathsf{N}}(\Sigma) : \int_\Sigma (\mathbf{curl}\,\mathrm{curl}\,\mathbf{u} - \lambda \mathbf{u}) \cdot \mathbf{v}\, d\mathbf{x} = 0 \quad \forall \mathbf{v} \in \mathbf{Q}^p_{\mathsf{N}}(\Sigma). \tag{12}$$

It is clear that all proposed eigenvectors in (b), (c), and (d) belong to $\mathbf{Q}^p_{\mathsf{N}}(\Sigma)$. It remains to compute $\mathbf{curl}\,\mathrm{curl}\,\mathbf{u} - \lambda\mathbf{u}$ in each case and to check (12).

Case (b). For $\lambda = \lambda_j$ and $\mathbf{u} = (0, -\beta_j(x))$, the two components of $\mathbf{curl}\,\mathrm{curl}\,\mathbf{u} - \lambda\mathbf{u}$ are

$$0 \quad \text{and} \quad \beta_j''(x) + \lambda_j \beta_j(x).$$

Then relation (11) yields (12), and the same argument applies to the other vector $(\beta_j(y), 0)$.

Case (c). For $\lambda = \lambda_j + \lambda_k$ and $\mathbf{u}$ given by (9), we have

$$\mathrm{curl}\,\mathbf{u} = (\lambda_j + \lambda_k)\beta_k'(x)\beta_j'(y)$$

and the two components of $\mathbf{curl}\,\mathrm{curl}\,\mathbf{u} - \lambda\mathbf{u}$ are

$$\begin{aligned} &-(\lambda_j + \lambda_k)\beta_k'(x)\beta_j''(y) - (\lambda_j + \lambda_k)\lambda_j \beta_k'(x)\beta_j(y) \\ &+ (\lambda_j + \lambda_k)\beta_k''(x)\beta_j'(y) + (\lambda_j + \lambda_k)\lambda_k \beta_k(x)\beta_j'(y) \end{aligned}$$

which can be written as

$$\begin{aligned} &-(\lambda_j + \lambda_k)\beta_k'(x)\big\{\beta_j''(y) + \lambda_j\beta_j(y)\big\} \\ &+ (\lambda_j + \lambda_k)\beta_j'(y)\big\{\beta_k''(x) + \lambda_k\beta_k(x)\big\}. \end{aligned}$$

Then relation (11) yields (12).

Case (d). For $\lambda = \lambda_j$ and $\mathbf{u} = (0, -\beta_j(x)L_p(y))$, the two components of $\mathbf{curl}\,\mathrm{curl}\,\mathbf{u} - \lambda\mathbf{u}$ are

$$-\beta_j'(x)L_p'(y) \quad \text{and} \quad \beta_j''(x)L_p(y) + \lambda_j\beta_j(x)L_p(y).$$

The second component is orthogonal to any element of $\mathbb{P}_0^p(I) \otimes \mathbb{P}^p(I)$ (see (1) and (11)). It remains to check that the first component is orthogonal to $\mathbb{P}^p(I) \otimes \mathbb{P}_0^p(I)$, i.e.,

$$\int_\Sigma v_j'(x) L_p'(y) w(x) v(y)\, dxdy = 0 \quad \forall w \in \mathbb{P}^p(I), \quad \forall v \in \mathbb{P}_0^p(I). \tag{13}$$

It is sufficient to prove that $\int_I L_p'(y) v(y)\, dy = 0$ for all $v \in \mathbb{P}_0^p(I)$: such a v is given by $(1-y^2)\varphi(y)$ with $\varphi \in \mathbb{P}^{p-2}(I)$. Since the polynomials L_k' are orthogonal on I with respect to the measure $(1-y^2)\, dy$ and since the degree of L_j' is $j-1$, we obtain that

$$\int_I L_p'(y)\varphi(y)(1-y^2)\, dy = 0 \quad \forall \varphi \in \mathbb{P}^{p-2}(I),$$

hence (13). ∎

The next theorem characterizes the Maxwell spectrum in $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$, i.e., the eigenpairs $(\lambda, \mathbf{u})$ such that

$$\mathbf{u} \in \mathbf{N}_{\mathsf{N}}^p(\Sigma) : \quad \int_\Sigma \operatorname{curl} \mathbf{u} \operatorname{curl} \mathbf{v}\, d\mathbf{x} = \lambda \int_\Sigma \mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} \quad \forall \mathbf{v} \in \mathbf{N}_{\mathsf{N}}^p(\Sigma). \tag{14}$$

THEOREM 2. *The three first parts* (a), (b), *and* (c) *of the discrete spectrum described in Theorem* 1 *are the whole discrete Maxwell spectrum computed by the edge element space* $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$.

Proof. We can see that the eigenvectors of parts (a), (b), and (c) all belong to the smaller space $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$. Therefore they are also eigenvectors in this space. We see that the sum of the dimensions of the corresponding eigenspaces is $(p-1)^2 + 2(p-1) + (p-1)^2$, which is equal to $2(p-1)p$, the dimension of $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$ (see (2)). ∎

The conclusion arising from Theorems 1 and 2 is that the space $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$ is to be preferred with respect to $\mathbf{Q}_{\mathsf{N}}^p(\Sigma)$ for the computation of Maxwell's eigenpairs. Indeed, the latter space does not provide correct approximation of the spectrum (see Remark 3).

4. Approximation of Maxwell's spectrum by *hp* rectangular finite elements. In this section we extend the results about the space $\mathbf{N}_{\mathsf{N}}^p(\Sigma)$ to the more involved situation of *hp* refinements, which provides realistic applications to a class of polygonal domains. The structure of this section is as follows. First, we define the FE spaces we are dealing with and make precise the assumptions on the mesh. Then, after establishing an L^2-stability result (see section 4.3), we prove the discrete compactness property which implies the convergence of eigenvalues/eigenvectors. Our proof clearly implies that the discrete compactness property also holds true for the pure p method on a conforming rectangular mesh (i.e., without hanging node) with the standard edge elements of the first Nédélec family.

4.1. De Rham diagram for a variable order quad element. Let $\Sigma = I \times I$ be the master square element. A feature of edge elements is their embedding in a *commuting de Rham diagram* of type (15) relating two exact sequences of spaces on both continuous and discrete levels. We refer the reader to [14, 33] for a systematic description of the standard discrete de Rham diagram of any degree, where the interpolation operators are based on nodal values, edge moments, and volume moments. In view of the construction *hp* finite elements, another class of commuting de Rham diagram has been introduced [26], relying on the so-called *projection-based interpolants*, which

allow variable orders on distinct elements of the same mesh, while preserving the commuting property.

We start by introducing this latter version of the de Rham diagram, involving discrete spaces and interpolation operators on Σ, according to

$$
\begin{array}{ccccccccc}
\mathbb{R} & \longrightarrow & H^{1+\varepsilon}(\Sigma) & \overset{\mathbf{grad}}{\longrightarrow} & \mathbf{H}^{\varepsilon}(\Sigma)\cap\mathbf{H}(\mathrm{curl},\Sigma) & \overset{\mathrm{curl}}{\longrightarrow} & L^2(\Sigma) & \longrightarrow & \mathbf{0} \\
\Big\downarrow id & & \Big\downarrow \Pi & & \Big\downarrow \Pi^{\mathrm{curl}} & & \Big\downarrow \mathrm{P} & & \\
\mathbb{R} & \longrightarrow & \mathbb{Q}^{p|\,p_e}(\Sigma) & \overset{\mathbf{grad}}{\longrightarrow} & \mathbf{N}^{p|\,p_e-1}(\Sigma) & \overset{\mathrm{curl}}{\longrightarrow} & \mathbb{Q}^{p-1}(\Sigma) & \longrightarrow & \mathbf{0}.
\end{array}
\tag{15}
$$

Index p specifies the order in both variables which, for the sake of simplicity of this presentation, we assume to be identical, $\mathbb{Q}^p(\Sigma) = \mathbb{Q}^{p,p}(\Sigma) = \mathbb{P}^p(I)\otimes\mathbb{P}^p(I)$, and with every edge e of the master element, we associate the corresponding order p_e, $e = 1,\dots,4$ (with standard, counterclockwise enumeration of edges), that satisfies the condition

$$
p_e \le p, \quad e = 1,\dots,4. \tag{16}
$$

The polynomial spaces present in the diagram are defined as follows.

- $\mathbb{Q}^{p|\,p_e}(\Sigma)$—the subspace of $\mathbb{Q}^p(\Sigma)$, consisting of polynomials whose traces on edges e reduce to (possibly lower) order p_e, $e = 1,\dots,4$.
- $\mathbf{N}^{p|\,p_e}(\Sigma)$—the subspace of $\mathbf{N}^p(\Sigma)$ (cf. section 2.1.2) of vector-valued polynomials with traces of their tangential components on edges e of (possibly lower) order p_e:

$$
\mathbf{N}^{p|\,p_e}(\Sigma) = \{\mathbf{u}\in\mathbf{N}^p(\Sigma) : u_t|_e := (\mathbf{n}\times\mathbf{u})|_e \in \mathbb{P}^{p_e}(e)\ \forall e\}, \tag{17}
$$

  where $\mathbf{n}$ is the outward unit normal vector.

In particular, $\mathbb{Q}^{p|\,-1}$ provides an alternative notation for the subspace $\mathbb{Q}_0^p$ of polynomials vanishing on the boundary of the element, and $\mathbf{N}^{p|\,-1} = \mathbf{N}_{\mathsf{N}}^p$ stands for the subspace of vector-valued polynomials from the first Nédélec family whose tangential component traces on the boundary are equal to zero. The assumption that edge orders p_e should not exceed corresponding components of order p is realized in practice by implementing the *minimum rule* that sets an edge order p_e to the minimum of orders p corresponding to the adjacent elements.

4.1.1. H^1-conforming projection-based interpolation. Let $\mathbb{P}_0^p(I)$ denote the space of polynomials of degree $\le p$, defined on the interval $I = (-1,1)$ with zero traces at the endpoints. Let $\phi_1(x) = (1-x)/2$, $\phi_2(x) = (x+1)/2$ be the standard 1D linear shape functions. Space $\mathbb{Q}^{p|\,p_e}(\Sigma)$ admits a natural decomposition into vertex bilinear shape functions, edge bubbles, and element bubbles:

$$
\begin{aligned}
\mathbb{Q}^{p|\,p_e}(\Sigma) = {} & \big\{\mathbb{P}^1(I)\otimes\mathbb{P}^1(I)\big\} \\
& \oplus \big\{(\mathbb{P}_0^{p_1}(I)\otimes\mathbb{R}\phi_1)\oplus(\mathbb{R}\phi_2\otimes\mathbb{P}_0^{p_2}(I))\oplus(\mathbb{P}_0^{p_3}(I)\otimes\mathbb{R}\phi_2)\oplus(\mathbb{R}\phi_1\otimes\mathbb{P}_0^{p_4}(I))\big\} \\
& \oplus \big\{\mathbb{P}_0^p(I)\otimes\mathbb{P}_0^p(I)\big\}.
\end{aligned}
\tag{18}
$$

We will alternatively speak of edge bubbles for functions defined on a particular edge (and zero at its ends) or for their extensions to the whole element (and zero on the

other edges). The linear extensions are natural but not essential in the forthcoming discussion. For a particular edge, the corresponding edge bubbles must vanish on the remaining edges and must "live" in the FE space. Similarly, the shape function for a vertex node must vanish at the remaining vertices, and it must be in the FE space; the fact that it is constructed using bilinear functions is secondary.

Given a function $u \in H^{1+\varepsilon}(\Sigma)$, we define its interpolant, $u_p = \Pi u$, as a sum of three contributions,

$$u_p = u_1 + \underbrace{\sum_e u_{2,e,p}}_{u_{2,p}} + u_{3,p}. \tag{19}$$

Interpolation at vertices. Vertex interpolant u_1 interpolates function u at vertices,

$$u_1(a) = u(a) \quad \text{for each vertex } a.$$

The simplest choice of an extension of the vertex values is provided by the bilinear function, but the ultimate value of the interpolant is independent of the choice of the extension as long as the extension "lives" in the FE space.

Projection on edges. We subtract the vertex interpolant u_1 from u and project the difference $u - u_1$, over each edge e, onto the space of edge bubbles,

$$|u - u_1 - u_{2,e,p}|_{1/2,e} \to \min.$$

The projection is done in an $H^{1/2}(e)$ seminorm, and it is equivalent to the solution of a small linear system,

$$\begin{aligned} &\textit{Find edge bubble } u_{2,e,p} \in \mathbb{P}_0^{p_e}(e) \textit{ such that} \\ &(u - u_1 - u_{2,e,p}\,, \phi)_{1/2,e} = 0, \quad \textit{for each edge bubble } \phi \in \mathbb{P}_0^{p_e}(e), \end{aligned}$$

where $(\cdot,\cdot)_{1/2,e}$ denotes the inner product corresponding to edge seminorm $|\cdot|_{1/2,e}$.

Projection on the element. We extend each edge bubble $u_{2,e,p}$ to the whole element. Again, the most natural extension is provided by the element shape functions and corresponds to decomposition (19). We subtract then the total edge interpolant $u_{2,p} = \sum_e u_{2,e,p}$ from the difference $u - u_1$ and project the resulting difference on the element bubbles,

$$|u - u_1 - u_{2,p} - u_{3,p}|_{1,\Sigma} \to \min.$$

Again, the projection is equivalent to a local Dirichlet problem on the element,

$$\begin{aligned} &\textit{Find element bubble } u_{3,p} \in \mathbb{Q}_0^p(\Sigma) \textit{ such that} \\ &(u - u_1 - u_{2,p} - u_{3,p}, \phi)_{1,\Sigma} = 0, \quad \textit{for each element bubble } \phi \in \mathbb{Q}_0^p(\Sigma), \end{aligned}$$

where $(\cdot,\cdot)_{1,\Sigma}$ denotes the H_0^1-inner product.

The interpolation is thus equivalent to the solution of a sequence of local (approximate) Dirichlet problems. We first interpolate at the vertices and then, with the vertex values providing Dirichlet conditions, solve the edge Dirichlet problems. Finally, we use the vertex and edge interpolants to set up the Dirichlet boundary conditions and solve the final Dirichlet problem on the whole element. Remember that it does not matter in which way we construct lifts of the approximate Dirichlet data; the ultimate interpolant is unique. In each of the three steps, we determine a part of the interpolant corresponding to the decomposition (18).

4.1.2. H(curl)-conforming projection-based interpolation. A similar decomposition into edge functions and element bubbles can be constructed for the space $\mathbf{N}^{p|\,p_e-1}(\Sigma)$,

$$\begin{aligned}\mathbf{N}^{p|\,p_e-1}(\Sigma) = \big\{ \big[(\mathbb{P}^{p_1-1}(I)\otimes\mathbb{R}\phi_1)\times\{0\}\big] \oplus \big[\{0\}\times(\mathbb{R}\phi_2\otimes\mathbb{P}^{p_2-1}(I))\big] \\ \oplus \big[(\mathbb{P}^{p_3-1}(I)\otimes\mathbb{R}\phi_2)\times\{0\}\big] \oplus \big[\{0\}\times(\mathbb{R}\phi_1\otimes\mathbb{P}^{p_4-1}(I))\big]\big\} \\ \oplus \big\{ \big[(\mathbb{P}^{p-1}\otimes\mathbb{P}_0^p)(\Sigma)\times\{0\}\big] \oplus \big[\{0\}\times(\mathbb{P}_0^p\otimes\mathbb{P}^{p-1})(\Sigma)\big]\big\}.\end{aligned} \tag{20}$$

Given a vector-valued function $\mathbf{u}\in\mathbf{H}^{\varepsilon}(\mathrm{curl},\Sigma)$, we define its interpolant $\mathbf{u}_p=\Pi^{\mathrm{curl}}\mathbf{u}$ as a sum of two contributions,

$$\mathbf{u}_p = \underbrace{\sum_e \mathbf{u}_{2,e,p}}_{\mathbf{u}_{2,p}} + \mathbf{u}_{3,p}.$$

Edge projections. For each edge e, let $v_t=\mathbf{n}\times\mathbf{v}$ denote the (scalar-valued) tangential component[1] of a field $\mathbf{v}$ on e. We project the tangential component u_t of function $\mathbf{u}$ onto the scalar edge functions,

$$\|u_t - u_{2,e,p,t}\|_{-1/2,e}\to\min.$$

Here the norm $\|\cdot\|_{-1/2,e}$ denotes the norm in the dual space

$$H^{-1/2}(e) = (H_{00}^{1/2}(e))';$$

see, e.g., [36]. We then define the vector edge function $\mathbf{u}_{2,e,p}$ as the tangent vector field on e such that $u_{2,e,p,t}=\mathbf{n}\times\mathbf{u}_{2,e,p}$. The projection problem is equivalent to the variational problem

Find the tangential component $u_{2,e,p,t}\in\mathbb{P}^{p_e-1}(e)$ *of the edge function* $\mathbf{u}_{2,e,p}$ *s. t.*
$(u_t-u_{2,e,p,t},\phi)_{-1/2,e}=0$, *for each edge function* $\phi\in\mathbb{P}^{p_e-1}(e)$,

with $(\cdot,\cdot)_{-1/2,e}$ denoting the inner product corresponding to norm $\|\cdot\|_{-1/2,e}$. Notice that for a constant function ϕ, the inner product reduces to L^2-product, and the equation above incorporates in particular the edge average condition

$$\int_e (u_t - u_{2,e,p,t})\,ds = 0.$$

Element projection. We extend each individual edge function $\mathbf{u}_{2,e,p}$ to the whole element using the edge shape functions according to the splitting (20), sum it up, $\mathbf{u}_{2,p}=\sum_e \mathbf{u}_{2,e,p}$, and subtract the difference from function $\mathbf{u}$. We then solve a local projection problem,

$$\|\,\mathrm{curl}(\mathbf{u}-\mathbf{u}_{2,p}-\mathbf{u}_{3,p})\|_{0,\Sigma}\to\min,$$

subjected to the additional constraint,

$$(\mathbf{u}-\mathbf{u}_{2,p}-\mathbf{u}_{3,p},\mathbf{grad}\,\phi)=0,\quad \text{for each element scalar bubble } \phi.$$

[1]More precisely, using the 3D notation, we have $v_t=(\mathbf{n}\times\mathbf{v})\cdot\mathbf{e}_z$, where $\mathbf{e}_z$ is the unit vector orthogonal to the 2D plane.

The constrained projection problem is equivalent to a Dirichlet mixed problem:

Find element bubble $\mathbf{u}_{3,p}$ and Lagrange multiplier ψ such that

(21)
$$\begin{cases} (\operatorname{curl}(\mathbf{u}-\mathbf{u}_{2,p}-\mathbf{u}_{3,p}),\operatorname{curl}\mathbf{v})+(\mathbf{grad}\,\psi,\mathbf{v})=0 & \text{for every element vector bubble } \mathbf{v},\\ (\mathbf{u}-\mathbf{u}_{2,p}-\mathbf{u}_{3,p},\mathbf{grad}\,\phi)=0 & \text{for every element scalar bubble } \phi. \end{cases}$$

Here the Lagrange multiplier ψ lives in the space of scalar bubbles. Since $\mathbf{grad}\,\psi$ is a vector bubble, the multiplier is identically equal to zero and, for this reason, is sometimes called the *hidden variable*.

Remark 4. In [24], the edge contributions $\mathbf{u}_{2,e,p}$ were split into the Whitney interpolant with constant tangential component and a higher order edge bubble. Also, the choice of "edge" norms in the presentation above is consistent with the latest 3D results (see [25]), and it is slightly different from those used in [24].

Finally, P is the L^2-projection from $L^2(\Sigma)$ onto $\mathbb{Q}^{p-1}(\Sigma)$. With this, we have the following result.

THEOREM 3. *If the edge seminorm $|u|_{1/2,e}$ is selected in such a way that*

$$|u|_{1/2,e}=\left\|\frac{\partial u}{\partial s}\right\|_{-1/2,e},$$

then the de Rham diagram (15) *commutes.*

Proof. The tangential derivative ∂_s is an isomorphism from $H^{1/2}(e)/\mathbb{R}$ onto $H^{-1/2}(e)$ (see [32, p. 31]). By the Bramble–Hilbert lemma, the norm in the quotient space $H^{1/2}(e)/\mathbb{R}$ is equivalent to the standard $|u|_{1/2,e}$-seminorm. Consequently, $\|\partial_s u\|_{-1/2,e}$ defines a seminorm on $H^{1/2}(e)$, equivalent to the standard seminorm.

The commuting projection-based interpolation operators considered in [26] use different projections on edges based on the H^1-seminorm for operator Π and the L^2-norm for operator Π^{curl}. The proof from [26] carries over to the case being presented without any modification, provided the H^1-seminorm is traded for the $H^{1/2}$-seminorm, and the L^2-norm is traded for the $H^{-1/2}$-norm. ∎

4.2. De Rham diagram for *hp* meshes. In this section we consider a polygonal domain Ω with sides parallel to the axes, covered by rectangular meshes aligned along the same axes. Of course, by a global affine transformation, our result generalizes to the situation of nonperpendicular axes.

If we fix a conforming mesh (i.e., such that the intersection of any two distinct elements $\overline{K}$ is either empty or a vertex or a full edge) and consider on each K the mapped spaces $\mathbb{Q}^{p|\,p}(K)$, $\mathbf{N}^{p|\,p-1}(K)$, and $\mathbb{Q}^{p-1}(K)$ with the same p, we can define on the whole domain Ω the corresponding H^1-, $\mathbf{H}(\mathrm{curl})$-, and L^2-conforming discrete spaces Q_p, $\mathbf{X}_p$, and S_p, and the projection-based interpolation is done element by element. The elements are said to be *unconstrained* in this case. Then it is clear that the commutativity properties (15) of the projection-based interpolation operators are still valid on the whole domain Ω. Besides, we note that in this case the discrete spaces coincide with those of the standard p-extension of the edge elements [37, 40].

The adaptation to *hp* meshes containing local refinements, therefore hanging nodes, and variable degrees are by no means obvious.

For the sake of simplicity of the presentation, we shall restrict ourselves to 1-irregular *hp* meshes corresponding to isotropic refinements only and consisting of square elements. Beginning with a standard regular mesh consisting of square elements of the same size, we allow for breaking each element into four elements with the

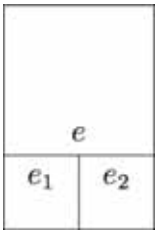


FIG. 1. *Constrained approximation.*

restriction that an element cannot share an edge with more than two small neighbors—the classical "two to one" rule. In other words, the *generation level* for two neighboring elements cannot differ by more than one. The order of elements can be modified locally, element by element, with the *minimum rule* being enforced—the order for an edge is set to the minimum of orders for all adjacent elements. Finally, since for meshes with hanging nodes the projections cannot be done on an element level—the resulting interpolants will no longer be conforming—the global conformity is maintained by means of the *constrained approximation*; see [23, 42].

The situation is illustrated in Figure 1. For a function u defined on "big" edge e, the corresponding "big edge" interpolant is a polynomial defined on the whole edge, whereas the interpolants determined on the small edges e_1, e_2 result in a piecewise interpolant that, in general, is different.

A natural idea is to utilize the constrained approximation concepts. First, do the projections on the big edges and then define the corresponding small edges interpolants by enforcing the global conformity requirements. The resulting interpolants will indeed be globally conforming, but we then lose the commutativity properties. This can be seen by considering the lowest order elements. The last space in the diagram then reduces to piecewise constants and the commutativity property requires that

$$\int_{\partial K} (u_{p,t} - u_t)\, ds = \int_K \operatorname{curl}(\mathbf{u}_p - \mathbf{u})\, d\mathbf{x} = 0$$

for each element K in the mesh. For regular meshes, the condition follows from the edge averaging. In the presence of hanging nodes, however, the condition may not be satisfied. Going back to the situation illustrated in Figure 1, enforcing the averaging condition on "big" edge e *does not imply* the same condition for the restrictions of the original function and its projection on the small edge e_1. Consequently, the condition above is violated for the small element, and the commutativity fails.

A remedy to the problem is to perform the interpolation on groups of elements. The whole mesh is split into *polygonal patches* consisting of single elements or *element clusters* (of minimum size) in such a way that all vertices of the polygonal patches are unconstrained. The decomposition is illustrated with the classical example of the L-shaped domain and h-refinements aimed at resolving the corner singularity shown in Figure 2.

All clusters in this example coincide with either a single element (the white elements) or three elements forming an L-shaped patch (such clusters are indicated in the picture with a grey or black shading). In general, the 1-irregularity rule limits the number of possible cluster shapes to four cases only: clusters of a single, two, three, or four small elements. Our convention is to call patch, denoted by P, the

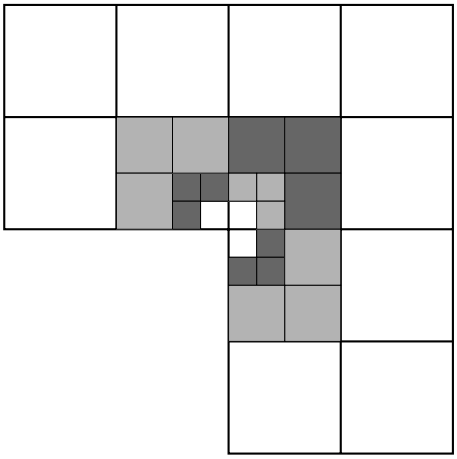


FIG. 2. *Decomposition of a* 1*-irregular mesh into clusters.*

FIG. 3. *L-shaped cluster and patch.*

union of the cluster elements K together with the interior edges. In our example, the L-shaped patches are the union of three squares and two interior edges; see Figure 3. Such patches have six distinct (exterior) edges. Edges of a patch always coincide with either a single element edge or two "small" edges adjacent to a big one.

The space $\mathbb{Q}^{p|\,p_e}(P)$ on the patch is the subspace of continuous functions on P, which are $\mathbb{Q}^p$ on each element K of the cluster, and whose restriction on e belongs to $\mathbb{Q}^{p_e}(e)$ for each edge of the patch. The space $\mathbf{N}^{p|\,p_e-1}(P)$ is defined correspondingly as the subspace of $\mathbf{H}(\mathrm{curl})$ fields on P, which are in $\mathbf{N}^p$ on each element K, and whose tangential restriction on e belongs to $\mathbb{Q}^{p_e}(e)$.

The definition of projection-based interpolation extends now naturally to the patch P: We list only the main steps. The vertices are the corners of P, and we check that there exists a continuous piecewise bilinear vertex interpolant $u_1 \in \mathbb{Q}^{1|\,1}(P)$. The edge bubbles $u_{2,e,p}$ are related with patch edges (and no longer with element edges) and are polynomial on the whole patch edge. These bubbles can be extended inside P as elements of $\mathbb{Q}^{p|\,p_e}(P)$. The patch bubbles $u_{3,p}$ are the functions in the FE space $\mathbb{Q}^{p|\,p_e}(P)$ with zero traces on the patch boundary ∂P.

Similarly we define the edge functions $\mathbf{u}_{2,e,p}$ as vector polynomials on the whole patch edge, tangential to the edge. They can be extended in $\mathbf{N}^{p|\,p_e-1}(P)$. The patch bubbles $\mathbf{u}_{3,p}$ are the elements of $\mathbb{Q}^{p|\,p_e}(P)$ with zero tangential traces on the patch boundary ∂P. The patch bubbles are, therefore, no longer polynomials but *piecewise polynomials* only.

Thus, by the same procedure as before we define the projection operators

$$\Pi = \Pi_P, \quad \Pi^{\mathrm{curl}} = \Pi_P^{\mathrm{curl}}, \quad \text{and} \quad \mathrm{P} = \mathrm{P}_P \tag{22}$$

and obtain a commutative scheme like (15).

Once the interpolation is done on the patches, we regain both the global conformity and commutativity of the interpolation operators on the whole *hp* mesh:

$$\begin{array}{ccccccccc}
\mathbb{R} & \longrightarrow & H^{1+\varepsilon}(\Omega) & \stackrel{\mathbf{grad}}{\longrightarrow} & \mathbf{H}^{\varepsilon}(\Omega)\cap\mathbf{H}(\mathrm{curl},\Omega) & \stackrel{\mathrm{curl}}{\longrightarrow} & L^2(\Omega) & \longrightarrow & \mathbf{0} \\
\downarrow id & & \downarrow \Pi & & \downarrow \Pi^{\mathrm{curl}} & & \downarrow \mathrm{P} & & \\
\mathbb{R} & \longrightarrow & Q_{hp} & \stackrel{\mathbf{grad}}{\longrightarrow} & \mathbf{X}_{hp} & \stackrel{\mathrm{curl}}{\longrightarrow} & S_{hp} & \longrightarrow & \mathbf{0}.
\end{array}$$

Here Q_{hp}, $\mathbf{X}_{hp}$, and S_{hp} denote FE spaces defined on the common domain Ω, corresponding to the H^1-, $\mathbf{H}(\mathrm{curl})$-, and L^2-conforming discretizations, done patch by patch.

4.3. A stability result in L^2. We begin by recalling the inclusion of polynomial spaces,

$$\mathbb{Q}_0^p(\Sigma) \stackrel{\mathbf{grad}}{\longrightarrow} \mathbf{N}_{\mathsf{N}}^p(\Sigma).$$

Here

$$\begin{aligned}
&\mathbb{Q}_0^p(\Sigma) = \mathbb{P}_0^p(I)\otimes\mathbb{P}_0^p(I),\\
&\mathbf{N}_{\mathsf{N}}^p(\Sigma) = [\mathbb{P}^{p-1}(I)\otimes\mathbb{P}_0^p(I)]\times[\mathbb{P}_0^p(I)\otimes\mathbb{P}^{p-1}(I)].
\end{aligned}$$

In this section, we will omit the mention of Σ and I for the spaces $\mathbb{Q}^p$, $\mathbf{N}^p$, and $\mathbb{P}^p$, respectively. We shall denote the L^2-norm on I or Σ by $\|\cdot\|$, with the corresponding L^2-product denoted by $(\cdot,\cdot)$. We hope that the similarity of the latter with the notation for vector components will not lead to confusion.

THEOREM 4. *The following stability condition holds:*

$$\inf_{\mathbf{q}\in\mathbf{N}_{\mathsf{N}}^p}\ \sup_{\mathbf{s}\in\mathbf{grad}\,\mathbb{Q}_0^p\oplus\mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p}\frac{(\mathbf{q},\mathbf{s})}{\|\mathbf{q}\|\,\|\mathbf{s}\|} = C_p, \tag{23}$$

where

$$C_p = \left(\frac{2(2p+1)}{(p+1)(p+2)}\right)^{1/2} = O(p^{-1/2}). \tag{24}$$

The proof of Theorem 4 relies on two lemmas.

LEMMA 5. *Let $a_i>0$, $b_i>0$, $i=1,\dots,n$. Then for any real $v_1,\dots,v_n$*

$$\sup_{u_1,\dots,u_n}\frac{|\sum_{i=1}^n a_iu_iv_i|}{(\sum_{i=1}^n b_iu_i^2)^{1/2}} = \left(\sum_{i=1}^n\frac{a_i^2}{b_i}v_i^2\right)^{1/2}.$$

Proof. Use the Cauchy–Schwarz inequality for the discrete l^2-product and the representation,

$$\sum_{i=1}^n a_iu_iv_i = \sum_{i=1}^n\frac{a_i}{b_i^{1/2}}v_i\,b_i^{1/2}u_i. \qquad \square$$

We recall that (λ_i, β_i), $i = 1, \dots, p-1$, and denote the discrete eigenpairs of the 1D Laplace operator defined in (6) (we omit the exponent $[p]$ for simplicity). The eigenvectors are normalized to satisfy $(\beta_i, \beta_j) = \delta_{ij}$.

LEMMA 6. *The following inequality holds:*

$$\left(\sum_{i=1}^{p-1} \lambda_i^2 v_i^2\right)^{1/2} \geq C_p \left\| \sum_{i=1}^{p-1} v_i \beta_i'' \right\| \quad \forall \mathbf{v} = (v_1, \dots, v_{p-1}) \in \mathbb{R}^{p-1},$$

where C_p is defined in (24).

Proof. It was proved in [8] that the constant

$$C_p = \inf_{u \in \mathbb{P}_0^p} \sup_{f \in \mathbb{P}^{p-2}} \frac{(u,f)}{\|u\|\|f\|} = \inf_{u \in \mathbb{P}_0^p} \sup_{v \in \mathbb{P}_0^p} \frac{(u,v'')}{\|u\|\|v''\|}$$
$$= \inf_{u \in \mathbb{P}_0^p} \sup_{v \in \mathbb{P}_0^p} \frac{(u',v')}{\|u\|\|v''\|} = \inf_{v \in \mathbb{P}_0^p} \sup_{u \in \mathbb{P}_0^p} \frac{(u',v')}{\|u\|\|v''\|}$$

is given by formula (24). Consequently,

$$\sup_{u \in \mathbb{P}_0^p} \frac{(u',v')}{\|u\|} \geq C_p \|v''\| \quad \forall v \in \mathbb{P}_0^p.$$

If we now define

$$u = \sum_{i=1}^{p-1} u_i \beta_i, \quad v = \sum_{i=1}^{p-1} v_i \beta_i,$$

then

$$(u', v') = \sum_{i=1}^{p-1} \lambda_i u_i v_i \qquad \text{and} \qquad \|u\| = \left(\sum_{i=1}^{p-1} u_i^2\right)^{1/2}.$$

Apply Lemma 5 to finish the proof. ∎

Proof of Theorem 4. *Step* 1. Let α_i, $i = 0, \dots, p-1$ be a basis for $\mathbb{P}^{p-1}$ defined as follows:

$$\alpha_i = \begin{cases} 1/\sqrt{2}, & i = 0, \\ \beta_i', & i = 1, \dots, p-1. \end{cases}$$

Polynomials α_i are orthogonal and satisfy

$$\|\alpha_0\|^2 = 1, \quad \|\alpha_i\|^2 = \lambda_i, \ i = 1, \dots, p-1.$$

Any element $\mathbf{q} \in \mathbf{N}_{\mathsf{N}}^p$ can be represented in the form

$$\mathbf{q} = \left(\sum_{i=0}^{p-1} \sum_{j=1}^{p-1} q_{1,ij} \alpha_i \beta_j, \sum_{i=1}^{p-1} \sum_{j=0}^{p-1} q_{2,ij} \beta_i \alpha_j \right). \tag{25}$$

Here and in what follows, we assume that in a tensor product $\alpha\beta$, the first function is always a function of x, and the second is a function of y, i.e., $\alpha\beta = \alpha(x)\beta(y)$.

A direct calculation shows that

$$
\begin{aligned}
&\mathbf{curl}\,\mathrm{curl}\,\mathbf{q}\\
&= \left(\sum_{i=1}^{p-1}\sum_{j=0}^{p-1} q_{2,ij}\beta_i'\alpha_j' - \sum_{i=0}^{p-1}\sum_{j=1}^{p-1} q_{1,ij}\alpha_i\beta_j'', \quad -\sum_{i=1}^{p-1}\sum_{j=0}^{p-1} q_{2,ij}\beta_i''\alpha_j + \sum_{i=0}^{p-1}\sum_{j=1}^{p-1} q_{1,ij}\alpha_i'\beta_j' \right)\\
&= \left(\sum_{i=1}^{p-1}\sum_{j=1}^{p-1} q_{2,ij}\beta_i'\beta_j'' - \sum_{i=0}^{p-1}\sum_{j=1}^{p-1} q_{1,ij}\alpha_i\beta_j'', \quad -\sum_{i=1}^{p-1}\sum_{j=0}^{p-1} q_{2,ij}\beta_i''\alpha_j + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} q_{1,ij}\beta_i''\beta_j' \right)\\
&= \Bigg(-\sum_{j=1}^{p-1} q_{1,0j}\alpha_0\beta_j'' + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1}(q_{2,ij}-q_{1,ij})\beta_i'\beta_j'',\\
&\qquad -\sum_{i=1}^{p-1} q_{2,i0}\beta_i''\alpha_0 - \sum_{i=1}^{p-1}\sum_{j=1}^{p-1}(q_{2,ij}-q_{1,ij})\beta_i''\beta_j' \Bigg).
\end{aligned}
$$

Hence, any element $\mathbf{s} \in \mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p$ can be represented in the form

$$
\mathbf{s} = \left(\sum_{j=1}^{p-1} s_{0j}\alpha_0\beta_j'' + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} s_{ij}\beta_i'\beta_j'', \quad \sum_{i=1}^{p-1} s_{i0}\beta_i''\alpha_0 - \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} s_{ij}\beta_i''\beta_j' \right).
$$

Let $\mathbf{q} \in \mathbf{N}_{\mathsf{N}}^p$ be discrete divergence-free, i.e.,

$$
(\mathbf{q}, \mathbf{grad}\, w) = 0 \quad \forall w \in \mathbb{Q}_0^p.
$$

Selecting $w = \beta_k\beta_l$, $k,l = 1,\dots,p-1$, we conclude that coefficients $q_{1,ij}, q_{2,ij}$ in representation (25) must satisfy the identity

$$
q_{1,kl}\lambda_k + q_{2,kl}\lambda_l = 0.
$$

This leads to the following formulas for the norm of a discrete divergence-free vector and the L^2-product of such a vector with $\mathbf{s} \in \mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p$:

$$
\begin{aligned}
\|\mathbf{q}\|^2 &= \sum_{j=1}^{p-1} q_{1,0j}^2 + \sum_{i=1}^{p-1} q_{2.i0}^2 + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1}\left(\frac{\lambda_i^2}{\lambda_j} + \lambda_i\right) q_{1,ij}^2,\\
(\mathbf{q},\mathbf{s}) &= -\sum_{j=1}^{p-1} q_{1,0j}s_{0j}\lambda_j - \sum_{i=1}^{p-1} q_{2,i0}s_{i0}\lambda_i - \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} q_{1,ij}s_{ij}(\lambda_i\lambda_j + \lambda_i^2).
\end{aligned}
$$

Applying Lemma 5 we get

$$
\sup_{\substack{\mathbf{q}\in\mathbf{N}_{\mathsf{N}}^p\\ (\mathbf{q},\mathbf{grad}\,w)=0\ \forall w\in\mathbb{Q}_0^p}} \frac{(\mathbf{q},\mathbf{s})}{\|\mathbf{q}\|} = \sum_{j=1}^{p-1}\lambda_j^2 s_{0j}^2 + \sum_{i=1}^{p-1}\lambda_i^2 s_{i0}^2 + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1}(\lambda_i^2\lambda_j + \lambda_i\lambda_j^2)s_{ij}^2.
$$

Finally, the norm of $\mathbf{s} \in \mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p$ can be represented in the form

$$
\|\mathbf{s}\|^2 = \left\|\sum_{j=1}^{p-1} s_{0j}\beta_j''\right\|^2 + \left\|\sum_{i=1}^{p-1} s_{i0}\beta_i''\right\|^2 + \sum_{i=1}^{p-1}\lambda_i\left\|\sum_{j=1}^{p-1} s_{ij}\beta_j''\right\|^2 + \sum_{j=1}^{p-1}\lambda_j\left\|\sum_{i=1}^{p-1} s_{ij}\beta_i''\right\|^2.
$$

By Lemma 6 we have

$$\sum_{j=1}^{p-1}\lambda_j^2 s_{0j}^2 \geq C_p^2 \left\| \sum_{j=1}^{p-1} s_{0j}\beta_j'' \right\|^2,$$
$$\sum_{i=1}^{p-1}\lambda_i^2 s_{i0}^2 \geq C_p^2 \left\| \sum_{i=1}^{p-1} s_{i0}\beta_i'' \right\|^2,$$
$$\lambda_i \sum_{j=1}^{p-1}\lambda_j^2 s_{ij}^2 \geq \lambda_i C_p^2 \left\| \sum_{j=1}^{p-1} s_{ij}\beta_j'' \right\|^2,$$
$$\lambda_j \sum_{i=1}^{p-1}\lambda_i^2 s_{ij}^2 \geq \lambda_j C_p^2 \left\| \sum_{i=1}^{p-1} s_{ij}\beta_i'' \right\|^2.$$

Summing up all the inequalities, we get

$$\sup_{\substack{\mathbf{q}\in \mathbf{N}_{\mathsf{N}}^p \\ (\mathbf{q},\,\mathbf{grad}\, w)=0 \ \forall w\in\mathbb{Q}_0^p}} \frac{(\mathbf{q},\mathbf{s})}{\|\mathbf{q}\|} \ \geq\ C_p\|\mathbf{s}\|,$$

or, equivalently, using the equality of inf-sup constants for a bilinear form and its adjoint, we get

$$\inf_{\substack{\mathbf{q}\in \mathbf{N}_{\mathsf{N}}^p \\ (\mathbf{q},\,\mathbf{grad}\, w)=0 \ \forall w\in\mathbb{Q}_0^p}} \sup_{\mathbf{s}\in \mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p} \frac{(\mathbf{q},\mathbf{s})}{\|\mathbf{q}\|\|\mathbf{s}\|} \ \geq\ C_p. \tag{26}$$

Step 2. Use discrete Helmholtz decomposition,

$$\mathbf{q} = \mathbf{q}_0 + \mathbf{grad}\,\phi, \quad (\mathbf{q}_0, \mathbf{grad}\, w) = 0 \quad \forall w\in\mathbb{Q}_0^p, \quad \phi\in\mathbb{Q}_0^p,$$

to extend inequality (26) to arbitrary $\mathbf{q}\in\mathbf{N}_{\mathsf{N}}^p$ and $\mathbf{s}\in\mathbf{grad}\,\mathbb{Q}_0^p \oplus \mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p$.

Step 3. The *equality* in (23) follows, e.g., from the fact that for $\mathbf{q}$ coinciding with eigenvectors from part (b) of the spectrum (see (8)), the 2D inf-sup condition reduces to its 1D counterpart. ∎

The consequence of Theorem 4 is the following L^2-stability result in p-version.

THEOREM 7. *Let* $\mathbf{u}_3\in\mathbf{H}^{\varepsilon}(\Sigma)\cap\mathbf{H}_0(\mathrm{curl},\Sigma)$ *be a divergence-free bubble function on* Σ. *Let* $\mathbf{u}_{3,p}$ *be the projection* $\Pi^{\mathrm{curl}}\mathbf{u}_3$. *Then* $\mathbf{u}_{3,p}$ *is discrete divergence-free and there holds*

$$\|\mathbf{u}_3-\mathbf{u}_{3,p}\|_{0,\Sigma} \leq Cp^{1/2} \inf_{\mathbf{q}_p\in\mathbf{N}_{\mathsf{N}}^p} \|\mathbf{u}_3-\mathbf{q}_p\|_{0,\Sigma}. \tag{27}$$

Proof. Let $\mathbf{q}_p$ be any element of $\mathbf{N}_{\mathsf{N}}^p$. Since

$$(\Pi^{\mathrm{curl}}\mathbf{u}_3, \mathbf{grad}\,\mathbf{q}_p) = (\mathbf{u}_3, \Pi^{\mathrm{curl}}\,\mathbf{grad}\,\mathbf{q}_p) = (\mathbf{u}_3, \mathbf{grad}\,\Pi\mathbf{q}_p),$$

we obtain that $\mathbf{u}_{3,p}$ is discrete divergence-free.

By Theorem 4, there exists $\mathbf{s}\in\mathbf{grad}\,\mathbb{Q}_0^p\oplus\mathbf{curl}\,\mathrm{curl}\,\mathbf{N}_{\mathsf{N}}^p$ so that

$$C_p\|\mathbf{u}_{3,p}-\mathbf{q}_p\|\,\|\mathbf{s}\| \leq (\mathbf{u}_{3,p}-\mathbf{q}_p,\mathbf{s}).$$

Any $\mathbf{s} \in \mathbf{grad}\, \mathbb{Q}_0^p \oplus \mathbf{curl}\, \mathrm{curl}\, \mathbf{N}_{\mathsf{N}}^p$ being orthogonal to $\mathbf{u}_3 - \mathbf{u}_{3,p}$ we get

$$C_p \|\mathbf{u}_{3,p} - \mathbf{q}_p\| \, \|\mathbf{s}\| \leq (\mathbf{u}_3 - \mathbf{q}_p, \mathbf{s}) \leq \|\mathbf{u}_3 - \mathbf{q}_p\| \, \|\mathbf{s}\|.$$

By the triangle inequality we deduce (27). ∎

The best approximation error in the L^2-norm by polynomials in $\mathbf{N}_{\mathsf{N}}^p$ behaves as p^{-1} for fields in H^1 satisfying the boundary conditions of $\mathbf{H}_0(\mathrm{curl})$.

LEMMA 8. *Let $\mathbf{u}_3 \in \mathbf{H}^1(\Sigma) \cap \mathbf{H}_0(\mathrm{curl}, \Sigma)$ be a general bubble function on Σ. There exists $\mathbf{q}_p \in \mathbf{N}_{\mathsf{N}}^p(\Sigma)$ such that*

$$\|\mathbf{u}_3 - \mathbf{q}_p\|_{0,\Sigma} \leq C p^{-1} \|\mathbf{u}_3\|_{1,\Sigma}. \tag{28}$$

Proof. Let u_x and u_y be the two components of $\mathbf{u}_3$. We note that u_x belongs to $L^2(I, H_0^1(I)) \cap H^1(I, L^2(I))$. We take as interpolant for u_x the function $\pi_x^{p-1,0} \otimes \pi_y^{p,1}(u_x)$, where $\pi^{p,0}$ and $\pi^{p,1}$ are the 1D standard projection operators used in spectral and p methods: $\pi^{p,0}$ is the L^2 orthogonal projection on $\mathbb{P}^p(I)$ and $\pi^{p,1}$ is defined as

$$\pi^{p,1}(u)(t) = \int_{-1}^{t} \pi^{p-1,0}(u')(s)\, ds.$$

Both $\pi^{p,0}$ and $\pi^{p,1}$ satisfy the L^2-H^1 error estimate with a factor p^{-1}; see, for instance, [43, Chapter 3]. Moreover, $\pi^{p,0}$ is stable in L^2 and $\pi^{p,1}$ in H^1. The proof of the estimate for $\|u_x - \pi_x^{p-1,0} \otimes \pi_y^{p,1}(u_x)\|_{0,\Sigma}$ then follows. The situation for the second component is similar. ∎

4.4. Discrete compactness. In this section we prove the discrete compactness property for edge finite elements on 1-irregular *hp* square meshes. The discrete compactness property, stated in Theorem 11, is known to be sufficient and in a sense necessary for the good approximation of eigenvalues/eigenvectors (see, for instance, [7, 16, 33, 39]).

For our proof we need L^2 estimates for $\mathbf{u} - \Pi^{\mathrm{curl}}\mathbf{u}$ for divergence-free fields $\mathbf{u}$ on any unconstrained element K (Lemma 9) or any patch P (Lemma 10).

LEMMA 9. *Let K be an unconstrained square element of size $h = h_K$, and let p be the minimum among p_K and $\{p_e,\ e = 1, \ldots, 4\}$. Let $\mathbf{u} \in \mathbf{H}^r(K)$, $0 < r < 1/2$, $\mathrm{curl}\, \mathbf{u} \in L^2(K)$, $\mathrm{div}\, \mathbf{u} = 0$. For every $\varepsilon > 0$, there exists a constant $C > 0$, dependent upon ε but independent of the element and function $\mathbf{u}$, such that*

$$\|\mathbf{u} - \Pi_K^{\mathrm{curl}}\mathbf{u}\| \leq C \left(\frac{h}{p}\right)^{(r-\varepsilon)} (\|\mathbf{u}\|_{r,K} + \|\mathrm{curl}\, \mathbf{u}\|_{0,K}).$$

Here Π_K^{curl} is the projection-based interpolation on K transported from Π^{curl} in (15).

Proof. Step 1. *p-estimate on the master element.* Assume first that $K = \Sigma$ is the master square element. It follows from the integration by parts formula

$$\int_K (\mathrm{curl}\, \mathbf{u})\phi \, d\mathbf{x} = \int_K \mathbf{u} \cdot \mathbf{curl}\, \phi \, d\mathbf{x} + \int_{\partial K} u_t \phi \, ds$$

that the tangential component u_t lives in $H^{-1/2+r}(\partial K)$ and

$$\|u_t\|_{-1/2+r,\partial K} \leq C(\|\mathbf{u}\|_{r,K} + \|\mathrm{curl}\, \mathbf{u}\|_{0,K}),$$

with C denoting a generic constant depending upon the master element only. We decompose function $\mathbf{u}$ into three contributions

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{grad}\, q + \mathbf{u}_3. \tag{29}$$

The terms are constructed as follows.

- $\mathbf{u}_1$ is the lowest degree Whitney interpolant, which means that $\mathbf{u}_1 \in \mathbf{N}^{1|\,0}(K)$, $\operatorname{div} \mathbf{u}_1 = 0$, and the tangential traces of $\mathbf{u}_1$ are the mean values of those of $\mathbf{u}$ on each edge of K.
- Potential q is obtained by integrating tangential component $u_t - u_{1t}$ along the element boundary, starting from any of its vertex nodes. Potential q vanishes at all vertex nodes and

  $$\|q\|_{1/2+r,\partial K} \leq C\|u_t - u_{1t}\|_{-1/2+r,\partial K}.$$

  As the Whitney interpolant depends continuously upon the tangential component u_t itself and it lives in a finite dimensional space, by the standard finite dimensionality argument we conclude that the norm of potential q is controlled by the norm of $\mathbf{u}$ alone:

  $$\begin{aligned}\|q\|_{1/2+r,\partial K} &\leq C\|u_t\|_{-1/2+r,\partial K}\\ &\leq C(\|\mathbf{u}\|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).\end{aligned}$$

  We then extend potential q to the rest of the element using a harmonic (minimum energy) extension. Consequently,

  $$\|q\|_{1+r,K} \leq (\|\mathbf{u}\|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).$$

- $\mathbf{u}_3$ is the residual bubble function: $\mathbf{n}\times\mathbf{u}_3 = 0$ on the boundary ∂K, and

  $$\operatorname{curl}\mathbf{u}_3 = \operatorname{curl}(\mathbf{u}-\mathbf{u}_1), \quad \operatorname{div}\mathbf{u}_3 = \operatorname{div}(\mathbf{u}-\mathbf{u}_1) = 0.$$

  It follows that $\mathbf{u}_3 \in \mathbf{H}^1(K)$ and

  $$\|\mathbf{u}_3\|_{1,K} \leq C\|\operatorname{curl}(\mathbf{u}-\mathbf{u}_1)\|_{0,K} \leq C\|\operatorname{curl}\mathbf{u}\|_{0,K}.$$

We use a similar decomposition for the projection-based interpolant $\Pi_K^{\mathrm{curl}}\mathbf{u}$ of $\mathbf{u}$,

$$\Pi_K^{\mathrm{curl}}\mathbf{u} = \mathbf{u}_1 + \mathbf{grad}\, q_p + \mathbf{u}_{p,3},$$

with the same Whitney interpolant $\mathbf{u}_1$ and $q_p = \Pi_K q$. Thus q_p is only a discrete harmonic function, and $\mathbf{u}_{p,3}$ is only discrete divergence-free. Obviously,

$$\mathbf{u} - \Pi_K^{\mathrm{curl}}\mathbf{u} = \mathbf{grad}(q - q_p) + \mathbf{u}_3 - \mathbf{u}_{p,3}.$$

The first term then admits the estimate (see [24])

$$\begin{aligned}\|\mathbf{grad}(q-q_p)\|_{0,K} &\leq Cp^{-(r-\varepsilon)}\|q\|_{1+r,K}\\ &\leq Cp^{-(r-\varepsilon)}(\|\mathbf{u}\|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).\end{aligned} \tag{30}$$

The estimate of the second term is made possible by Theorem 7: there holds

$$\|\mathbf{u}_3 - \mathbf{u}_{3,p}\|_{0,K} \leq Cp^{1/2} \inf_{\mathbf{F}_{3,p}\in\mathbf{N}_{\mathsf{N}}^p} \|\mathbf{u}_3 - \mathbf{F}_{3,p}\|_{0,K}.$$

The approximation result (28) then gives

$$\begin{aligned}\|\mathbf{u}_3 - \mathbf{u}_{3,p}\|_{0,K} &\le Cp^{-1/2}\|\mathbf{u}_3\|_{1,K} \\ &\le Cp^{-1/2}(\|\mathbf{u}\|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).\end{aligned} \tag{31}$$

Combining (30) and (31), we get the final estimate for the master element,

$$\|\mathbf{u} - \Pi_K^{\mathrm{curl}}\mathbf{u}\|_{0,K} \le Cp^{-(r-\varepsilon)}(\|\mathbf{u}\|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).$$

Step 2. *Scaling argument.* Let K be an arbitrary (unconstrained) square element, and let

$$\Sigma = \hat{K} \ni \boldsymbol{\xi} \to \mathbf{x} \in K$$

be the homothetic transformation from the master element Σ onto K. Recalling the transformation for $\mathbf{H}(\mathrm{curl})$-conforming elements,

$$\hat{\mathbf{u}}(\boldsymbol{\xi}) = \mathbf{u}(\mathbf{x})h,$$

where $h = h_K$ is the element size, we follow the standard scaling argument and Step 1 result to obtain

$$\begin{aligned}\|\mathbf{u} - \Pi_K^{\mathrm{curl}}\mathbf{u}\|_{0,K} &= \|\hat{\mathbf{u}} - \Pi^{\mathrm{curl}}\hat{\mathbf{u}}\|_{0,\Sigma} \\ &\le Cp^{-(r-\varepsilon)}(\|\hat{\mathbf{u}}\|_{r,\Sigma} + \|\operatorname{curl}\hat{\mathbf{u}}\|_{0,\Sigma}).\end{aligned}$$

However, the (projection-based) interpolation reproduces polynomials and, by the Bramble–Hilbert argument and standard interpolation arguments, we get

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}_p\|_{0,K} &\le Cp^{-(r-\varepsilon)}(|\hat{\mathbf{u}}|_{r,\Sigma} + \|\operatorname{curl}\hat{\mathbf{u}}\|_{0,\Sigma}) \\ &\le C\left(\frac{h}{p}\right)^{(r-\varepsilon)}(|\mathbf{u}|_{r,K} + \|\operatorname{curl}\mathbf{u}\|_{0,K}).\end{aligned}$$

This finishes the proof. ∎

We have an analogous but slightly different result for element patches.

LEMMA 10. *Let P be a patch of two, three, or four square elements of the same size, forming a rectangle, an L-shaped domain, and a square, respectively; cf. section* 4.2. *Let p denote the minimum order of all elements and edges constituting the patch. Let h denote the size of the elements forming the patch. Let $\mathbf{u} \in \mathbf{H}^r(P)$, $0 < r < 1/2$, $\operatorname{curl}\mathbf{u} \in L^2(P)$, $\operatorname{div}\mathbf{u} = 0$. There exist constant $C > 0$, independent of the element and function $\mathbf{u}$, and constant r_P, $0 < r_P < r$, such that*

$$\|\mathbf{u} - \Pi_P^{\mathrm{curl}}\mathbf{u}\|_{0,P} \le C\left(\frac{h}{p}\right)^{r_P}(\|\mathbf{u}\|_{r,P} + \|\operatorname{curl}\mathbf{u}\|_{0,P}).$$

By $\Pi_P^{\mathrm{curl}}\mathbf{u}$ we understand the projection-based interpolation (22) *done on the patch.*

Proof. The reasoning follows the same lines as for the preceding lemma. We revisit the main steps and point out the differences.

- $\mathbf{u}_1$ plays on the patch P a similar (but weaker) role as the lowest degree Whitney interpolant on K: $\mathbf{u}_1 \in \mathbf{N}^{1|\,0}(P)$, $\operatorname{div}\mathbf{u}_1 = 0$ in P, and $\mathbf{u}_1$ compensates for the mean value of the tangential trace of $\mathbf{u}$ on the whole boundary ∂P:

$$\int_{\partial P} \mathbf{n} \times (\mathbf{u} - \mathbf{u}_1)\, ds = 0.$$

  For $\mathbf{u}_1$ we may take a field of the form $\gamma\mathbf{e}|_P$, where $\mathbf{e}$ is any nonzero element of $\mathbf{N}^{1|\,0}(\hat{P})$ on the convex hull $\hat{P}$ of P, and γ is a suitable constant.

- The potential q is still obtained by first integrating $u_t - u_{1t}$ along ∂P. Now it does not vanish at the corners, but we still have $q \in H^{1/2+r}(\partial P)$, so we can take its harmonic extension in P to find $q \in H^{1+r}(P)$.
- We still have a decomposition like (29), $\mathbf{u} = \mathbf{u}_1 + \mathbf{grad}\, q + \mathbf{u}_3$, with a divergence-free patch bubble function $\mathbf{u}_3$. But for the L-shaped patches, $\mathbf{u}_3$ is no longer an $\mathbf{H}^1$-function; however, $\mathbf{u}_3$ belongs to $\mathbf{H}^{1/2+r_P}(P)$, with $r_P > 0$ (here r_P is any constant $< \frac{1}{6}$) [18].
- At the discrete level, we have $\Pi_P^{\mathrm{curl}}\mathbf{u} = \mathbf{u}_1 + \mathbf{grad}(\Pi_P q) + \mathbf{u}_{3,p}$. The estimate corresponding to (30), $\|\mathbf{grad}(q - \Pi_P q)\|_{0,P}$, does not follow directly from element estimates, but it can be obtained extending arguments from [24]. Alternatively, the H^1 patch interpolant $\Pi_P q$ can be seen as the Galerkin approximation to the solution of the Laplace equation on the patch, with Dirichlet boundary conditions and right approximation of Dirichlet data (in the $H^{1/2}$-norm). The corresponding estimates can be found in [43].
- The bound on $\|\mathbf{u}_3 - \mathbf{u}_{3,p}\|_{0,P}$ corresponding to (31) does not follow directly from the L^2-stability result for a single element. Instead, we proceed by comparing the patch interpolant $\mathbf{u}_{3,p} = \Pi_P^{\mathrm{curl}}\mathbf{u}_3$ with the union of interpolants $\Pi_K^{\mathrm{curl}}\mathbf{u}_3$ corresponding to elements K contributing to the patch, denoted by $\mathbf{v}_{3,p}$:

$$\mathbf{v}_{3,p}\big|_K = \Pi_K^{\mathrm{curl}}\mathbf{u}_3 \quad \forall K \subset P.$$

  Both operators Π_P^{curl} and $(\Pi_K^{\mathrm{curl}})_{K\subset P}$, acting from $\mathbf{H}_0(\mathrm{curl}, P)\cap\mathbf{H}^{\varepsilon}(P)$, satisfy the commutativity property for the de Rham diagram. The L^2-projections of $\mathrm{curl}\,\mathbf{u}$ done on the whole patch or elementwise are identical. Consequently,

$$\mathrm{curl}\,\mathbf{u}_{3,p} = \mathrm{curl}\,\mathbf{v}_{3,p},$$

  and the two functions may differ only by a gradient of potential ϕ that is zero on the patch boundary ∂P and lives in the patch FE space. It follows from the fact that $\mathbf{u}_{3,p}$ is discrete divergence-free that

$$\begin{aligned}\|\mathbf{u}_3 - \mathbf{u}_{3,p}\|_{0,P} &\le \inf_{\phi} \|\mathbf{u}_3 - \mathbf{u}_{3,p} - \mathbf{grad}\,\phi\|_{0,P}\\ &\le \|\mathbf{u}_3 - \mathbf{v}_{3,p}\|_{0,P}.\end{aligned}$$

  Coming back to the definition of $\mathbf{v}_{3,p}$, we finally obtain

$$\|\mathbf{u}_3 - \mathbf{u}_{3,p}\|_{0,P} \le \sum_{K\subset P} \|\mathbf{u}_3 - \Pi_K^{\mathrm{curl}}\mathbf{u}_3\|_{0,P}.$$

  The estimation can now be done elementwise on each unconstrained element $K \subset P$ utilizing Lemma 9 for $\mathbf{u} := \mathbf{u}_3|_K$, noting that $\mathrm{div}\,\mathbf{u}_3|_K = 0$. ∎

We are ready now to formulate and prove our final result.

THEOREM 11. *Starting with a regular mesh on Ω we perform consecutive hp-refinements, enforcing the 1-irregularity and minimum rules, constructing meshes $\mathfrak{M}_{hp}$. We assume that*

$$\max_{K\in\mathfrak{M}_{hp}} \frac{h_K}{p_K} \to 0. \tag{32}$$

Let $\mathbf{u}_{hp} \in \mathbf{X}_{hp}$ be an arbitrary sequence of FE functions on $\mathfrak{M}_{hp}$, such that $\mathbf{u}_{hp}\times\mathbf{n} = 0$ on $\partial\Omega$. We assume that the functions $\mathbf{u}_{hp}$ are discrete divergence-free, *i.e.,*

$$(\mathbf{u}_{hp}, \mathbf{grad}\,\phi_{hp}) = 0 \quad \forall \phi_{hp} \in \mathbb{Q}_{hp}.$$

We also assume that the $\mathbf{u}_{hp}$ *are uniformly bounded in the space* $\mathbf{H}(\mathrm{curl},\Omega)$*:*

$$\|\operatorname{curl}\mathbf{u}_{hp}\| \le 1.$$

Then there exists a subsequence $\mathbf{u}_{hp}$*, (denoted with the same symbol) converging strongly in* $L^2(\Omega)$ *to a limit*[2] $\mathbf{u}$*:*

$$\|\mathbf{u}_{hp}-\mathbf{u}\| \to 0.$$

Proof. Step 1. Following Kikuchi's reasoning (see [35]), we introduce a sequence of divergence-free functions $\mathbf{u}^{hp}$, satisfying the same essential boundary conditions, such that

$$\operatorname{curl}\mathbf{u}^{hp} = \operatorname{curl}\mathbf{u}_{hp}, \quad (\mathbf{u}^{hp},\, \mathbf{grad}\,\phi) = 0 \quad \forall \phi \in H^1_0(\Omega).$$

We have

$$\mathbf{u}_{hp} = \mathbf{u}^{hp} + \mathbf{grad}\, q^{hp}, \tag{33}$$

where q^{hp} is the solution to

$$\begin{aligned} &q^{hp} \in H^1_0(\Omega)\\ &(\mathbf{grad}\, q^{hp},\, \mathbf{grad}\,\phi) = (\mathbf{u}_{hp},\, \mathbf{grad}\,\phi) \quad \forall \phi \in H^1_0(\Omega). \end{aligned}$$

It follows from the regularity results of [18] that

$$\mathbf{u}^{hp} \in \mathbf{H}^r(\Omega), \quad r>0,$$

with a uniform bound on the $\mathbf{H}^r$ norm,

$$\|\mathbf{u}^{hp}\|_{\mathbf{H}^r(\Omega)} \le C.$$

By a standard compactness argument, there exists a subsequence $\mathbf{u}^{hp}$ converging strongly in $L^2(\Omega)$ to a limit $\mathbf{u}$. We are going to prove that $\mathbf{grad}\, q^{hp} \to 0$, and thus obtain that $\mathbf{u}_{hp}$ converges to the same limit $\mathbf{u}$.

Step 2. Applying the interpolation operator to both sides of the equation, and using the commutativity of interpolation and the fact that the interpolation preserves FE spaces, we get

$$\mathbf{u}_{hp} = \Pi^{\mathrm{curl}}\mathbf{u}^{hp} + \mathbf{grad}\,\Pi q^{hp}. \tag{34}$$

Subtracting (34) from (33) we get

$$-\mathbf{grad}(q^{hp} - \Pi q^{hp}) = \mathbf{u}^{hp} - \Pi^{\mathrm{curl}}\mathbf{u}^{hp}\,.$$

It follows from (33) that $\mathbf{grad}\, q^{hp}$ is orthogonal to all discrete gradients. Consequently,

$$\|\mathbf{grad}\, q^{hp}\| = \inf_{q_{hp}\in Q_{hp}} \|\mathbf{grad}(q^{hp}-q_{hp})\| \le \|\mathbf{grad}(q^{hp}-\Pi q^{hp})\| = \|\mathbf{u}^{hp} - \Pi^{\mathrm{curl}}\mathbf{u}^{hp}\|.$$

It is sufficient, therefore, to prove that the interpolation error of functions $\mathbf{u}^{hp}$ converges uniformly to zero.

[2] Notice that the limit satisfies $\|\operatorname{curl}\mathbf{u}\| \le 1$ and that $\mathbf{u}$ is divergence-free.

Step 3. Applying Lemmas 9 and 10, we obtain

$$\|\mathbf{u}^{hp} - \Pi^{\mathrm{curl}}\mathbf{u}^{hp}\|^2_{0,\Omega} = \sum_P \|\mathbf{u}^{hp} - \Pi^{\mathrm{curl}}\mathbf{u}^{hp}\|^2_{0,P}$$
$$\leq C \sum_P \left(\frac{h_P}{p_P}\right)^{2r_P} (\|\mathbf{u}^{hp}\|_{r,P} + \|\operatorname{curl}\mathbf{u}^{hp}\|_{0,P})^2.$$

Here r is the global regularity constant and $r_P < r$ denote the patch constants discussed in Lemma 10 (we also consider unconstrained elements K as one-element patches and, applying Lemma 9, take r_P as any number between 0 and r in this case). As r_P depends only upon the shape of the patch and the number of different patches is finite, the L^2-interpolation error must converge to zero if the maximum ratio of patch size and (minimum) order converges to zero,

$$\max_P \frac{h_P}{p_P} \to 0.$$

Notice, finally, that the 1-irregularity and max rules imply that the last condition follows from assumption (32). ∎

Remark 5. Examining our proof, we see that we have proved the following property: There exists a sequence δ_{hp} converging to 0 such that

(35) $\forall \mathbf{u}_{hp} \in \mathbf{X}_{hp}$, discrete divergence-free,
$\exists \mathbf{u}^{hp} \in \mathbf{H}_0(\mathrm{curl},\Omega)$ with $\operatorname{div}\mathbf{u}^{hp} = 0$: $\|\mathbf{u}^{hp} - \mathbf{u}_{hp}\| \leq C\delta_{hp}(\|\mathbf{u}_{hp}\| + \|\operatorname{curl}\mathbf{u}_{hp}\|)$.

Here $C > 0$ does not depend on $\mathbf{u}_{hp}$. Condition (35) implies the discrete compactness property; cf. [6, 7]. It also implies the quasi optimality of the discrete electric Maxwell problems for any fixed frequency which is not an eigenfrequency of the continuous problem; see [10, 15, 27].

5. Conclusions. Relying on our main result, Theorem 11, and on [16], we can conclude the convergence of eigenvalue approximation along the following lines.

Let Ω be a simply connected polygonal domain with sides parallel to the coordinate axes. We consider the Maxwell eigenvalue problem on Ω:

(36) Find $\mathbf{u} \in \mathbf{H}(\mathrm{curl},\Omega), \mathbf{u} \neq 0$, and $\lambda \neq 0$:
$$\int_\Omega \operatorname{curl}\mathbf{u}\, \operatorname{curl}\mathbf{v}\, d\mathbf{x} = \lambda \int_\Omega \mathbf{u}\cdot\mathbf{v}\, d\mathbf{x} \quad \forall \mathbf{v} \in \mathbf{H}(\mathrm{curl},\Omega).$$

The condition $\lambda \neq 0$ implies that eigenvectors $\mathbf{u}$ are divergence-free. The converse is also true since Ω is simply connected. The eigenvalues λ have a finite multiplicity and form an increasing sequence of positive numbers without accumulation point. We denote by $\lambda^1 \leq \lambda^2 \leq \cdots \leq \lambda^k \leq \cdots$ the sequence of eigenvalues with repetition according to their multiplicities.

We choose a sequence of *hp* FE spaces $\mathbf{X}_{hp}$ satisfying the assumptions of Theorem 11 and define the approximated problems by the following problem:

(37) Find $\mathbf{u}_{hp} \in \mathbf{X}_{hp}, \mathbf{u}_{hp} \neq 0$, and $\lambda_{hp} \neq 0$:
$$\int_\Omega \operatorname{curl}\mathbf{u}_{hp}\, \operatorname{curl}\mathbf{v}_{hp}\, d\mathbf{x} = \lambda_{hp} \int_\Omega \mathbf{u}_{hp}\cdot\mathbf{v}_{hp}\, d\mathbf{x} \quad \forall \mathbf{v}_{hp} \in \mathbf{X}_{hp}.$$

The discrete eigenvectors $\mathbf{u}_{hp}$ are discrete divergence-free. We denote by $\lambda_{hp}^1 \le \lambda_{hp}^2 \le \cdots \le \lambda_{hp}^k \le \cdots$ the sequence of eigenvalues with repetition according to their multiplicities.

Assumption (32) guarantees that the conditions (CAS) (approximation in $\mathbf{H}(\mathrm{curl}, \Omega)$) and (CDK) (approximation in the kernel of the curl operator) of [16] are satisfied. Theorem 11 yields condition (DCP) of discrete compactness. Thus [16, Theorem 6.9] yields that the sequence of problems (37) is a *spurious-free spectrally correct* approximation of problem (36). As a consequence, we have

$$\forall k \ge 1, \quad \lambda_{hp}^k \longrightarrow \lambda^k \quad \text{as} \quad \max_{K \in \mathfrak{M}_{hp}} \frac{h_K}{p_K} \to 0. \tag{38}$$

Let us recall that the spectral correctness alone would provide a weaker statement, according to which the correct numbering of discrete eigenvalues which ensures (38) should be done by discarding *small* eigenvalues (and not only *zero* eigenvalues). "Small" means that the maximal size ε_{hp} of the discarded ones tends to zero as h_K/p_K tends to zero. The consequence of the spurious-free property is that ε_{hp} is equal to zero.

According to [16, Theorem 6.11] the three conditions (CAS), (CDK), and (DCP) imply condition (CHN) too. Condition (CHN) is the one which allows the application of the theory of [28]; see also [16, Remark 4.11]. This implies, for example, that if λ^k is a simple eigenvalue, there holds the estimate

$$|\lambda^k - \lambda_{hp}^k| \le C_k \left(\min_{\mathbf{v}_{hp} \in \mathbf{X}_{hp}} \|\mathbf{u}^k - \mathbf{v}_{hp}\|_{\mathbf{H}(\mathrm{curl},\Omega)} \right)^2. \tag{39}$$

Here $\mathbf{u}^k$ is a normalized eigenvector associated with λ^k. Less sharp estimates can be deduced by this argument in the case of multiple eigenvalues.

In [21, 22], it is proved that the regularity of the eigenvectors $\mathbf{u}^k$ can be described in terms of weighted analytic spaces (close to the countably normed spaced of [4]), via a decomposition $\nabla\varphi^k + \mathbf{w}^k$ where the potential φ^k concentrates the strongest singularities. Combining this with the approximation result proved in [1] for Raviart–Thomas elements, it is possible to deduce an exponential estimate in our case:

$$|\lambda^k - \lambda_{hp}^k| \le C_k \, e^{-b_k N^{1/3}}, \quad b_k > 0. \tag{40}$$

Here N is the dimension of $\mathbf{X}_{hp}$. Note that $N^{1/3}$ is a $\mathcal{O}(p)$.

Our final comment concerns the validity of our result for meshes obtained using the so-called algebraic mesh generators. This is the case when the actual physical domain is partitioned into a finite number of (possibly curvilinear) quadrilaterals, each of them being the image of a reference unit square through a smooth map. The maps are compatible in the sense that parametrizations for two quadrilaterals adjacent to a common edge provide an identical parametrization for the edge. The original Maxwell problem can then be restated on a collection of reference square domains with appropriate interface conditions and modified material properties resulting from the change of coordinates. According to the result of Caorsi, Fernandes, and Raffetto [16], the discrete compactness property for constant material data implies the corresponding discrete compactness property for the case of general, possibly anisotropic, material data. As the discretization in the original domain with parametric, exact geometry elements is equivalent to the discretization of the modified problem on the reference

squares using square elements discussed in this paper, our analysis applies to such a case as well. We emphasize that the situation is essentially different when unstructured mesh generators are used, and the geometry of individual quadrilateral elements is no longer controlled by global (and sufficiently smooth) maps; cf. [3].

REFERENCES

[1] M. Ainsworth and K. Pinchedez, *hp-approximation theory for* **BDFM** *and* **RT** *finite elements on quadrilaterals*, SIAM J. Numer. Anal., 40 (2002), pp. 2047–2068.

[2] D. N. Arnold, *Differential Complexes and Numerical Stability*, in Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), Higher Education Press, Beijing, China, 2002, pp. 137–157.

[3] D. N. Arnold, D. Boffi, and R. S. Falk, *Quadrilateral H(div) finite elements*, SIAM J. Numer. Anal. 2004, 42 (2005), pp. 2429–245.

[4] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data*. Part I. *Boundary value problems for linear elliptic equation of second order*, SIAM J. Math. Anal., 19 (1988), pp. 172–203.

[5] I. Babuška and J. Osborn, *Eigenvalue Problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal. II, North–Holland, Amsterdam, 1991, pp. 641–787.

[6] D. Boffi, *Fortin operator and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[7] D. Boffi, *A note on the de Rham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[8] D. Boffi, L. Demkowicz, and M. Costabel, *Discrete compactness for p and hp 2D edge finite elements*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1673–1687.

[9] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[10] D. Boffi and L. Gastaldi, *Edge finite elements for the approximation of Maxwell resolvent operator*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 293–305.

[11] D. Boffi, F. Kikuchi, and J. Schöberl, *Edge element computation of Maxwell's eigenvalues on general quadrilateral meshes*, Math. Models Methods Appl. Sci., 16 (2006), pp. 265–273.

[12] A. Bossavit, *Mixed finite elements and the complex of Whitney forms*, in The Mathematics of Finite Elements and Applications, VI (Uxbridge, 1987), Academic Press, London, 1988, pp. 137–144.

[13] A. Bossavit, *Un nouveau point de vue sur les éléments finis mixtes*, Matapli (Bulletin de la Société de Mathématiques Appliquëes et Industrielles), 20 (1989), pp. 23–35.

[14] A. Buffa, M. Costabel, and M. Dauge, *Algebraic convergence for anisotropic edge elements in polyhedral domains*, Numer. Math., 101 (2005), pp. 29–65.

[15] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell transmission problems in Lipschitz domains*, Numer. Math., 95 (2003), pp. 459–485.

[16] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.

[17] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.

[18] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[19] M. Costabel and M. Dauge, *Computation of resonance frequencies for Maxwell equations in non-smooth domains*, in Topics in Computational Wave Propagation, Lect. Notes Comput. Sci. Eng. 31, Springer, Berlin, 2003, pp. 125–161.

[20] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.

[21] M. Costabel, M. Dauge, and C. Schwab, *Exponential convergence of hp-FEM for Maxwell's equations with weighted regularization in polygonal domains*, Math. Models Methods Appl. Sci., 15 (2005), pp. 575–622.

[22] M. Costabel, M. Dauge, and C. Schwab, *Exponential convergence of the weighted regularization method for Maxwell eigenvalue problems*, in Proceedings of the International Conference on Electromagnetics in Advanced Applications (ICEAA), Torino, Italy, 2005.

[23] L. Demkowicz, *Fully automatic hp-adaptivity for Maxwell's equations*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 605–624.
[24] L. Demkowicz and I. Babuška, *p interpolation error estimates for edge finite elements of variable order in two dimensions*, SIAM J. Numer. Anal., 41 (2003), pp. 1195–1208.
[25] L. Demkowicz and A. Buffa, *H^1, $H(curl)$ and $H(div)$-conforming projection-based interpolation in three dimensions*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 267–296.
[26] L. Demkowicz, P. Monk, L. Vardapetyan, and W. Rachowicz, *de Rham diagram for hp finite element spaces*, Comput. Math. Appl., 39 (2000), pp. 29–38.
[27] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.
[28] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation. I. The problem of convergence*, RAIRO Anal. Numér., 12 (1978), pp. 97–112, iii.
[29] B. M. Dillon, P. T. S. Liu, and J. P. Webb, *Spurious modes in quadrilateral and triangular edge elements*, COMPEL, 13, Suppl. A (1994), pp. 311–316.
[30] P. Fernandes and M. Raffetto, *Counterexamples to the currently accepted explanation for spurious modes and necessary and sufficient conditions to avoid them*, IEEE Trans. on Magnetics, 38 (2002), pp. 653–656.
[31] F. Gardini, *Discrete compactness property for quadrilateral finite element spaces*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 41–56.
[32] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, 1992.
[33] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.
[34] J. Douglas, Jr., and J.E. Roberts, *Mixed finite element methods for second order elliptic problems*, Mat. Apl. Comput., 1 (1982), pp. 91–103.
[35] F. Kikuchi, *On a discrete compactness property for the Nédélec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 479–490.
[36] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[37] P. Monk, *On the p- and hp-extension of Nédélec's curl-conforming elements*, J. Comput. Appl. Math., 53 (1994), pp. 117–137.
[38] P. Monk, *Finite element methods for Maxwell's equations*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2003.
[39] P. Monk and L. Demkowicz, *Discrete compactness and the approximation of Maxwell's equations in $\mathbb{R}^3$*, Math. Comp., 70 (2000), pp. 507–523.
[40] J.-C. Nédélec, *Mixed finite elements in $\mathbf{R}^3$*, Numer. Math., 35 (1980), pp. 315–341.
[41] J.-C. Nédélec, *A new family of mixed finite elements in $\mathbb{R}^3$*, Numer. Math., 50 (1986), pp. 57–81.
[42] W. Rachowicz and L. Demkowicz, *An hp-adaptive finite element method for electromagnetics. Part* I. *Data structure and constrained approximation*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 307–337.
[43] Ch. Schwab, *p- and hp-finite element methods. Theory and applications in solid and fluid mechanics*, Numerical Mathematics and Scientific Computation, The Clarendon Press, Oxford University Press, New York, 1998.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1005–1025

OPTIMALITY OF MULTILEVEL PRECONDITIONERS FOR LOCAL MESH REFINEMENT IN THREE DIMENSIONS*

BURAK AKSOYLU†‡ AND MICHAEL HOLST§

Abstract. In this article, we establish optimality of the Bramble–Pasciak–Xu (BPX) norm equivalence and optimality of the wavelet modified (or *stabilized*) hierarchical basis (WHB) preconditioner in the setting of local 3D mesh refinement. In the analysis of WHB methods, a critical first step is to establish the optimality of BPX norm equivalence for the refinement procedures under consideration. While the available optimality results for the BPX norm have been constructed primarily in the setting of uniformly refined meshes, a notable exception is the local 2D red-green result due to Dahmen and Kunoth. The purpose of this article is to extend this original 2D optimality result to the local 3D red-green refinement procedure introduced by Bornemann, Erdmann, and Kornhuber, and then to use this result to extend the WHB optimality results from the quasi-uniform setting to local 2D and 3D red-green refinement scenarios. The BPX extension is reduced to establishing that locally enriched finite element subspaces allow for the construction of a scaled basis which is formally Riesz stable. This construction turns out to rest not only on the shape regularity of the refined elements, but also critically on a number of geometrical properties we establish between neighboring simplices produced by the Bornemann–Erdmann–Kornhuber (BEK) refinement procedure. It is possible to show that the number of degrees of freedom used for smoothing is bounded by a constant times the number of degrees of freedom introduced at that level of refinement, indicating that a practical, implementable version of the resulting BPX preconditioner for the BEK refinement setting has provably optimal (linear) computational complexity per iteration. An interesting implication of the optimality of the WHB preconditioner is the a priori H^1-stability of the L_2-projection. The existing a posteriori approaches in the literature dictate a reconstruction of the mesh if such conditions cannot be satisfied. The theoretical framework employed supports arbitrary spatial dimension $d \geq 1$ and requires no coefficient smoothness assumptions beyond those required for well-posedness in H^1.

Key words. finite element approximation theory, multilevel preconditioning, BPX, hierarchical bases, wavelets, three dimensions, local mesh refinement, red-green refinement

AMS subject classifications. 65M55, 65N55, 65N22, 65F10

DOI. 10.1137/S0036142902406119

1. Introduction.

In this article, we analyze the impact of local mesh refinement on the stability of multilevel finite element spaces and on optimality (linear space and time complexity) of multilevel preconditioners. Adaptive refinement techniques have become a crucial tool for many applications, and access to optimal or near-optimal multilevel preconditioners for locally refined mesh situations is of primary concern to computational scientists. The preconditioners which can be expected to have somewhat favorable space and time complexity in such local refinement scenarios are the hierarchical basis (HB) method [9], the Bramble–Pasciak–Xu (BPX)

*Received by the editors April 19, 2002; accepted for publication (in revised form) December 21, 2005; published electronically May 19, 2006.

http://www.siam.org/journals/sinum/44-3/40611.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803.

‡Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803 (burak@cct.lsu.edu). This author was supported in part by the Burroughs Wellcome Fund, in part by NSF grants ACI-9721349 and DMS-9872890, and in part by DOE grant W-7405-ENG-48/B341492. Additional support was provided by Intel, Microsoft, Alias|Wavefront, Pixar, and the Packard Foundation.

§Department of Mathematics, University of California at San Diego, La Jolla, CA 92093 (mholst@math.ucsd.edu). This author was supported in part by NSF CAREER Award DMS-9875856 and standard grants DMS-0208449, DMS-9973276, and DMS-0112413, in part by DOE grant SCI-DAC-21-6993, and in part by a Hellman Fellowship.

preconditioner [16], and the wavelet modified (or stabilized) hierarchical basis (WHB) method [35]. While there are optimality results for both the BPX and WHB preconditioners in the literature, these are primarily for quasi-uniform meshes and/or two space dimensions (with some exceptions noted below). In particular, there are few hard results in the literature on the optimality of these methods for various realistic local mesh refinement hierarchies, especially in three space dimensions. In this article, the first in a series of two articles [2] on local refinement and multilevel preconditioners, we first assemble optimality results for the BPX norm equivalence in local refinement scenarios in three spatial dimensions. Building on the extended BPX results, we then develop optimality results for the WHB method in local refinement settings. The material forming this series is based on the first author's Ph.D. dissertation [1]; a comprehensive presentation of this article can be found in [3, 4, 5, 6].

Through some topological or geometrical abstraction, if local refinement is extended to d spatial dimensions, then the main results are valid for any dimension $d \geq 1$ and for nonsmooth PDE coefficients $p \in L_\infty(\Omega)$. Throughout this article, we consider primarily the $d = 3$ case. But, when the abstraction to generic d is clear, we simply state the argument by using this generic d.

The problem class we focus on here is linear second order PDEs of the form

$$-\nabla \cdot (p\ \nabla u) + q\ u = f, \quad u = 0 \ \text{ on } \partial\Omega. \tag{1.1}$$

Here, $f \in L_2(\Omega)$, $p, q \in L_\infty(\Omega)$, $p : \Omega \to L(\mathbb{R}^d, \mathbb{R}^d)$, $q : \Omega \to \mathbb{R}$, where p is a symmetric positive definite matrix function and q is a nonnegative function. Let $\mathcal{T}_0$ be a shape regular and quasi-uniform initial partition of Ω into a finite number of d simplices, and generate $\mathcal{T}_1, \mathcal{T}_2, \ldots$ by refining the initial partition using red-green local refinement strategies in $d = 3$ spatial dimensions. Denote by $\mathcal{S}_j$ the simplicial linear C^0 finite element space corresponding to $\mathcal{T}_j$ equipped with zero boundary values. The set of nodal basis functions for $\mathcal{S}_j$ is denoted by $\Phi^{(j)} = \{\phi_i^{(j)}\}_{i=1}^{N_j}$, where $N_j = \dim\ \mathcal{S}_j$ is equal to the number of interior nodes in $\mathcal{T}_j$, representing the number of degrees of freedom (DOF) in the discrete space. Successively refined finite element spaces will form the following nested sequence:

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots \subset \mathcal{S}_j \subset \cdots \subset H_0^1(\Omega).$$

Let the bilinear form and the functional associated with the weak formulation of (1.1) be denoted as

$$a(u, v) = \int_\Omega p\ \nabla u \cdot \nabla v + q\ u\ v\ dx, \quad b(v) = \int_\Omega f\ v\ dx, \quad u, v \in H_0^1(\Omega).$$

We consider primarily the following Galerkin formulation: Find $u \in \mathcal{S}_j$, such that

$$a(u, v) = b(v) \quad \forall v \in \mathcal{S}_j. \tag{1.2}$$

The finite element approximation in $\mathcal{S}_j$ has the form $u^{(j)} = \sum_{i=1}^{N_j} u_i \phi_i^{(j)}$, where $u = (u_1, \ldots, u_{N_j})^T$ denotes the coefficients of $u^{(j)}$ with respect to $\Phi^{(j)}$. The resulting *discretization operator* $A^{(j)} = \{a(\phi_k^{(j)}, \phi_l^{(j)})\}_{k,l=1}^{N_j}$ must be inverted numerically to determine the coefficients u from the linear system

$$A^{(j)} u = F^{(j)}, \tag{1.3}$$

where $F^{(j)} = \{b(\phi_l^{(j)})\}_{l=1}^{N_j}$. Our task is to solve (1.3) with optimal (linear) complexity in both storage and computation, where the finite element spaces $\mathcal{S}_j$ are built on locally refined meshes.

Optimality of the BPX norm equivalence with generic local refinement was shown by Bramble and Pasciak in [14], where the impact of the local smoother and the local projection operator on the estimates was carefully analyzed. The two primary results on optimality of the BPX norm equivalence in the local refinement settings are due to Dahmen and Kunoth [19] and Bornemann and Yserentant [12]. Both works consider only two space dimensions, and in particular, the refinement strategies analyzed are restricted to 2D red-green refinement and 2D red refinement, respectively. In this paper, we extend the framework developed in [19] to a practical, implementable 3D local red-green refinement procedure introduced by Bornemann, Erdmann, and Kornhuber [11]. We will refer to this as the BEK refinement procedure.

HB methods [9, 7, 37] are particularly attractive in the local refinement setting because (by construction) each iteration has linear (optimal) computational and storage complexity. Unfortunately, the resulting preconditioner is not optimal due to condition number growth: in two dimensions the growth is slow, and the method is quite effective (nearly optimal), but in three dimensions the condition number grows much more rapidly with the number of unknowns [26]. To address this instability, one can employ L_2-orthonormal wavelets in place of the HB, giving rise to an optimal preconditioner [23]. However, the complicated nature of traditional wavelet bases, in particular the nonlocal support of the basis functions and problematic treatment of boundary conditions, severely limits computational feasibility. WHB methods have been developed [34, 35] as an alternative, and they can be interpreted as a wavelet modification (or *stabilization*) of the HB. These methods have been shown to optimally stabilize the condition number of the systems arising from HB methods on quasi-uniform meshes, in both two and three space dimensions, and retain a comparable cost per iteration.

There are two main results and one secondary result in this article. The main results establish the optimality of the BPX norm equivalence and also optimality of the WHB preconditioner—as well as optimal computational complexity per iteration—for the resulting locally refined 3D finite element hierarchy. Both the BPX and WHB preconditioners under consideration are additive Schwarz preconditioners. The BPX analysis here relies heavily on the techniques of the Dahmen–Kunoth [19] framework and can be seen as an extension to three spatial dimensions, with the realistic BEK refinement procedure [11] being the application of interest. The WHB framework relies on the optimality of the BPX norm equivalence. Hence, the WHB results are established after the BPX results.

The secondary result is the H^1-stability of L_2-projection onto finite element spaces built through the BEK local refinement procedure. This is currently under intensive study in the finite element community due to its relationship to multilevel preconditioning. The existing theoretical results, due primarily to Carstensen [18] and Bramble, Pasciak, and Steinbach [15] involve a posteriori verification of somewhat complicated mesh conditions after local refinement has taken place. If such mesh conditions are not satisfied, one has to redefine the mesh. However, an interesting consequence of the BPX optimality results for locally refined 2D and 3D meshes established here is H^1-stability of L_2-projection restricted to the same locally enriched finite element spaces. This result appears to be the first a priori H^1-stability result for L_2-projection on finite element spaces produced by practical and easily implementable 2D and 3D local refinement procedures.

Outline of the paper. In section 2, we introduce some basic approximation theory tools used in the analysis such as Besov spaces and Bernstein inequalities. The framework for the main norm equivalence is also established here. In section 3,

we list the BEK refinement conditions. We give several theorems about the generation and size relations of the neighboring simplices, thereby establishing local (patchwise) quasi uniformity. This gives rise to an L_2-stable Riesz basis in section 3.1; one can then establish the Bernstein inequality.

In section 4, we explicitly give an upper bound for the nodes introduced in the refinement region. This implies that one application of the BPX preconditioner to a function has linear (optimal) computational complexity. In section 5, we use the geometrical results from section 3 to extend the 2D Dahmen–Kunoth results to the 3D BEK refinement procedure by establishing the desired norm equivalence. While it is not possible to establish a Jackson inequality due to the nature of local adaptivity, in section 6 the remaining inequality in the norm equivalence is handled directly using approximation theory tools, as in the original work [19]. In section 7, we introduce the WHB preconditioner as well as the operator used in its definition. In section 8, we state the fundamental assumption for establishing basis stability and set up the main theoretical results for the WHB framework, namely, optimality of the WHB preconditioner in the 2D and 3D local red-green refinements. The results in section 8 rest completely on the BPX results in section 5 and on the Bernstein inequalities, the latter of which rest on the geometrical results established in section 3. The first a priori H^1-stability result for L_2-projection on the finite element spaces produced is established in section 9. We conclude in section 10.

2. Preliminaries and the main norm equivalence. The basic restriction on the refinement procedure is that it remains *nested*. In other words, tetrahedra of level j which are not candidates for further refinement will never be touched in the future. Let Ω_j denote the refinement region, namely, the union of the supports of basis functions which are introduced at level j. Due to nested refinement, $\Omega_j \subset \Omega_{j-1}$, the following hierarchy holds:

$$\Omega_J \subset \Omega_{J-1} \subset \cdots \subset \Omega_0 = \Omega. \tag{2.1}$$

In the local refinement setting, in order to maintain optimal computational complexity, the smoother is restricted to a local space $\tilde{\mathcal{S}}_j$, typically

$$\mathcal{S}_j^f \subseteq \tilde{\mathcal{S}}_j \subset \mathcal{S}_j, \tag{2.2}$$

where $\mathcal{S}_j^f := (I_j - I_{j-1})\ \mathcal{S}_j$ and $I_j : L_2(\Omega) \to \mathcal{S}_j$ denotes the finite element interpolation operator. DOF corresponding to $\mathcal{S}_j^f$ and $\tilde{\mathcal{S}}_j$ will be denoted by $\mathcal{N}_j^f$ and $\tilde{\mathcal{N}}_j$, respectively, where f stands for *fine*. Equation (2.2) indicates that $\mathcal{N}_j^f \subseteq \tilde{\mathcal{N}}_j$; typically, $\tilde{\mathcal{N}}_j$ consists of fine DOF and their corresponding coarse fathers.

The BPX preconditioner (also known as the parallelized or additive multigrid) is defined as follows:

$$Xu := \sum_{j=0}^{J} 2^{j(d-2)} \sum_{i \in \tilde{\mathcal{N}}_j} (u, \phi_i^{(j)}) \phi_i^{(j)}. \tag{2.3}$$

Success of the BPX preconditioner in locally refined regimes relies on the fact that the BPX smoother acts on a local space as in (2.2). As mentioned above, it acts on a slightly bigger set than fine DOF (examples of these are given in [13]). The choice of such a set is crucial because computational cost per iteration will eventually determine the overall computational complexity of the method. Hence, in section 4

we show that the overall computational cost of the smoother is $O(N)$, meaning that the BPX preconditioner is optimal per iteration. We would like to emphasize that one of the main goals of this paper, as in the earlier works of Dahmen and Kunoth [19] and Bornemann and Yserentant [12] in the purely 2D case, is to establish the optimality of the BPX norm equivalence,

$$c_1 \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2 \leq \|u\|_{H^1}^2 \leq c_2 \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2, \tag{2.4}$$

where Q_j is the L_2-projection. We note that in the uniform refinement setting, it is straightforward to link the BPX norm equivalence to the optimality of the BPX preconditioner,

$$c_1(Xu, u) \leq \|u\|_{H^1}^2 \leq c_2(Xu, u),$$

due to the projector relationships between the Q_j operators. However, in the local refinement scenario the precise link between the norm equivalence and the preconditioner is more subtle and essentially remains open.

The rest of this section is dedicated to setting up the framework to establish the main norm equivalence (2.4), which will be formalized in Theorem 2.1 at the end of this section. We borrow several tools from approximation theory, including the modulus of smoothness, $\omega_k(f, t, \Omega)_p$, which is a finer scale of smoothness than differentiability. It is a central tool in the analysis here and it naturally gives rise to the notion of *Besov spaces*. For further details and definitions, see [19, 29]. Besov spaces are defined to be the collection of functions $f \in L_p(\Omega)$ with a finite Besov norm defined as

$$\|f\|_{B^s_{p,q}(\Omega)}^q := \|f\|_{L_p(\Omega)}^q + |f|_{B^s_{p,q}(\Omega)}^q,$$

where the seminorm is given by

$$|f|_{B^s_{p,q}(\Omega)} := \|\{2^{sj}\omega_k(f, 2^{-j}, \Omega)_p\}_{j\in\mathbb{N}_0}\|_{l_q},$$

with k any fixed integer larger than s.

Besov spaces become the primary function space setting in the analysis by realizing Sobolev spaces as Besov spaces:

$$H^s(\Omega) \cong B^s_{2,2}(\Omega), \quad s > 0.$$

The primary motivation for employing the Besov space stems from the fact that the characterization of functions which have a given upper bound for the error of approximation sometimes calls for a finer scale of smoothness than that provided by Sobolev classes functions.

The Bernstein inequality is defined as

$$\omega_{k+1}(u, t)_p \leq c\ (\min\{1, t2^J\})^\beta \|u\|_{L_p}, \quad u \in \mathcal{S}_j, \quad j = 0, \ldots, J, \tag{2.5}$$

where c is independent of u and j. Usually $k =$ degree of the element, and in the case of linear finite elements, $k = 1$. Here β is determined by the global smoothness of the approximation space as well as p. For C^r finite elements, $\beta = \min\{1 + r + \frac{1}{p}, k + 1\}$.

Let θ_J be defined as

$$\theta_{j,J} := \sup_{u \in \mathcal{S}_J} \frac{\|u - Q_j u\|_{L_2}}{\omega_2(u, 2^{-j})_2}, \quad \theta_J := \max\{1, \theta_{j,J} : j = 0, \ldots, J\}. \tag{2.6}$$

Following [19] we have the following theorem.

THEOREM 2.1. *Suppose the Bernstein inequality* (2.5) *holds for some real number* $\beta > 1$. *Then, for each* $0 < s < \min\{\beta, 2\}$, *there exist constants* $0 < c_1,\ c_2 < \infty$ *independent of* $u \in \mathcal{S}_J,\ J = 0, 1, \ldots,$ *such that the following norm equivalence holds:*

$$\frac{c_1}{\theta_J^2} \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2 \le \|u\|_{H^1}^2 \le c_2 \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2, \quad u \in \mathcal{S}_J. \tag{2.7}$$

Proof. See [19, Thm. 4.1]. ∎

We would like to elaborate on the difficulties one faces within the local refinement framework. In order for the Bernstein inequality to hold, one needs to establish that the underlying basis is an L_2-stable Riesz basis as in (3.8). This crucial property heavily depends on local quasi uniformity of the mesh. Hence, the Bernstein inequality is established in section 5 through local quasi uniformity and L_2-stability of the basis in the Riesz sense.

A Jackson-type inequality cannot hold in a local refinement setting. This poses a major difficulty in the analysis because one has to calculate θ_J directly. The missing crucial piece of the optimal norm equivalence in (2.7), namely, $\theta_J = O(1)$ as $J \to \infty$, will be shown in (6.12) and as a result (2.4) will hold. This required the operator $\tilde{Q}_j$ to be bounded locally and to fix polynomials of degree 1 as will be shown in section 6.

3. The BEK refinement procedure. Our interest is in showing optimality of the BPX norm equivalence for the local 3D red-green refinement introduced by Bornemann, Erdmann, and Kornhuber [11]. This 3D red-green refinement is practical and easy to implement; numerical experiments were presented in [11]. A similar refinement procedure was analyzed by Bey [10]; in particular, the same green closure strategy was used in both papers. While these refinement procedures are known to be asymptotically nondegenerate (and thus produce shape regular simplices at every level of refinement), shape regularity is insufficient for constructing a stable Riesz basis for finite element spaces on locally adapted meshes. To construct a stable Riesz basis we will need to establish patchwise quasi uniformity as in [19]; as a result, d-vertex adjacency relationships that are independent of the shape regularity of the elements must be established between neighboring tetrahedra as done in [19] for triangles.

We first list a number of geometric assumptions concerning the underlying mesh. Let $\Omega \subset \mathbb{R}^3$ be a polyhedral domain. We assume that the triangulation $\mathcal{T}_j$ of Ω at level j is a collection of tetrahedra with mutually disjoint interiors which cover $\Omega = \bigcup_{\tau \in \mathcal{T}_j} \tau$. We want to generate successive refinements $\mathcal{T}_0, \mathcal{T}_1, \ldots$ which satisfy the following conditions.

Assumption 3.1 (*nestedness*). Each tetrahedron (son) $\tau \in \mathcal{T}_j$ is covered by exactly one tetrahedron (father) $\tau' \in \mathcal{T}_{j-1}$, and any corner of τ is either a corner or an edge midpoint of τ'.

Assumption 3.2 (*conformity*). The intersection of any two tetrahedra $\tau, \tau' \in \mathcal{T}_j$ is either empty, a common vertex, a common edge, or a common face.

Assumption 3.3 (*nondegeneracy*). The interior angles of all tetrahedra in the refinement sequence $\mathcal{T}_0, \mathcal{T}_1, \ldots$ are bounded away from zero.

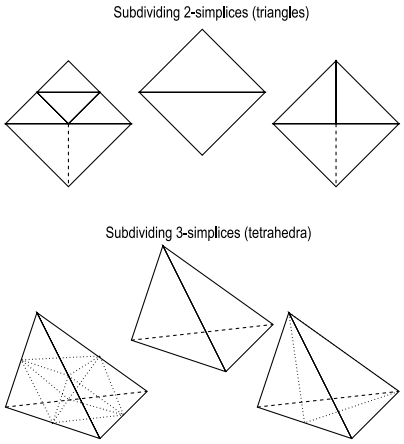


FIG. 3.1.

A regular (red) refinement subdivides a tetrahedron τ into eight equal volume subtetrahedra. We connect the edges of each face as in 2D regular refinement. We then cut off four subtetrahedra at the corners which are congruent to τ. An octahedron with three parallelograms remains in the interior. Cutting the octahedron along the two faces of these parallelograms, we obtain four more subtetrahedra which are not necessarily congruent to τ. We choose the diagonal of the parallelogram so that the successive refinements always preserve nondegeneracy [1, 10, 27, 38]. A sketch of regular refinement (octasection and quadrasection in three and two dimensions, respectively) as well as bisection is given in Figure 3.1.

If a tetrahedron is marked for regular refinement, the resulting triangulation violates conformity Assumption 3.2. Nonconformity is then remedied by irregular (green) refinement. In three dimensions, there are altogether $2^6 = 64$ possible edge refinements, of which 62 are irregular. One must pay extra attention to irregular refinement in the implementation due to the large number of possible nonconforming configurations. Bey [10] gives a methodical way of handling irregular cases. Using symmetry arguments, the 62 irregular cases can be divided into nine different types. To ensure that the interior angles remain bounded away from zero, we enforce the following additional conditions. (Identical assumptions were made in [19] for their 2D refinement analogue.)

Assumption 3.4. Irregular tetrahedra are not refined further.

Assumption 3.5. Only tetrahedra $\tau \in \mathcal{T}_j$ with $L(\tau) = j$ are refined for the construction of $\mathcal{T}_{j+1}$, where $L(\tau) = \min\{j : \tau \in \mathcal{T}_j\}$ denotes the level of τ.

One should note that the restrictive character of Assumptions 3.4 and 3.5 can be eliminated by a modification on the sequence of the tetrahedralizations [10]. On the other hand, it is straightforward to enforce both assumptions in a typical local refinement algorithm by minor modifications of the supporting datastructures for tetrahedral elements (cf. [22]). In any event, the proof technique (see (6.8) and (6.9)) requires both assumptions hold. The last refinement condition enforced for the possible 62 irregularly refined tetrahedra is stated as the following.

Assumption 3.6. If three or more edges are refined and do not belong to a common face, then the tetrahedron is refined regularly.

We note that the d-vertex adjacency generation bound for simplices in $\mathbb{R}^d$ which are adjacent at d vertices *is the primary result required in the support of a basis func-*

tion so that Assumption 3.6 *holds, and depends critically on the particular details of the local refinement procedure rather than on shape regularity of the elements.* The generation bound for simplices which are adjacent at $d-1, d-2, \ldots$ vertices follows by using the shape regularity and the generation bound established for d-vertex adjacency. We provide rigorous generation bounds for all the adjacency types mentioned in the lemmas to follow when $d = 3$. The 2D version appeared in [19]; the 3D extension is described below.

LEMMA 3.1. *Let τ and τ' be two tetrahedra in $\mathcal{T}_j$ sharing a common face f. Then*

$$|L(\tau) - L(\tau')| \leq 1. \tag{3.1}$$

Proof. If $L(\tau) = L(\tau')$, then $0 \leq 1$, and there is nothing to show. Without loss of generality, assume that $L(\tau) < L(\tau')$. The proof requires a detailed and systematic analysis. To show the line of reasoning, we first list as follows the facts used in the proof:

1. $L(\tau') \leq j$ because by assumption $\tau' \in \mathcal{T}_j$. Then $L(\tau) < j$.
2. By assumption $\tau \in \mathcal{T}_j$, meaning that τ was never refined from level $L(\tau)$, in which it was born, to level j.
3. Let τ'' be the father of τ'. Then $L(\tau'') = L(\tau') - 1 < j$.
4. $L(\tau) < L(\tau')$ by assumption, implying $L(\tau) \leq L(\tau'')$.
5. By (2), τ belongs to all the triangulations from $L(\tau)$ to j, in particular $\tau \in \mathcal{T}_{L(\tau'')}$, where by fact 3 $L(\tau'') < j$.

f is the common face of τ and τ' on level j. By fact 5 both $\tau,\ \tau'' \in \mathcal{T}_{L(\tau'')}$. Then, Assumption 3.2 implies that f must still be the common face of τ and τ''. Hence, τ' must have been irregular.

On the other hand, $L(\tau) \leq L(\tau') - 1 = L(\tau'')$. Next, we proceed by eliminating the possibility that $L(\tau) < L(\tau'')$. If so, we repeat the above reasoning, and τ'' becomes irregular. τ'' is already the father of the irregular τ', contradicting Assumption 3.4 for level $L(\tau'')$. Hence, $L(\tau) = L(\tau'') = L(\tau') - 1$ concludes the proof. ∎

By Assumptions 3.4 and 3.5, every tetrahedron at any $\mathcal{T}_j$ is geometrically similar to some tetrahedron in $\mathcal{T}_0$ or to a tetrahedron arising from an irregular refinement of some tetrahedron in $\mathcal{T}_0$. Then, there exist absolute constants $c_1,\ c_2$ such that

$$c_1 \operatorname{diam}(\bar{\tau})\ 2^{-L(\tau)} \leq \operatorname{diam}(\tau) \leq c_2 \operatorname{diam}(\bar{\tau})\ 2^{-L(\tau)}, \tag{3.2}$$

where $\bar{\tau}$ is the father of τ in the initial mesh. The lemma below follows by shape regularity and (3.1).

LEMMA 3.2. *Let τ, τ' and ζ, ζ' be the tetrahedra in $\mathcal{T}_j$ sharing a common edge (two vertices) and a common vertex, respectively. Then there exist finite numbers V and E depending on the shape regularity such that*

$$|L(\tau) - L(\tau')| \leq V, \tag{3.3}$$

$$|L(\zeta) - L(\zeta')| \leq E. \tag{3.4}$$

Consequently, simplices in the support of a basis function are comparable in size as indicated in (3.5). This is usually called *patchwise quasi uniformity*. Furthermore, it was shown in [1] that patchwise quasi uniformity (3.5) holds for 3D marked tetrahedron bisection introduced by Liu and Joe [24] and for 2D newest vertex bisection introduced by Sewell [30] and Mitchell [25]. Due to the restrictive nature of the proof technique (see (6.8) and (6.9)), we focus on refinement procedures which obey Assumptions 3.4 and 3.5. However, due to the strong geometrical results available for

purely bisection-based local refinement procedures, it should be possible to establish the main results of this paper for purely bisection-based strategies.

LEMMA 3.3. *There is a constant depending on the shape regularity of $\mathcal{T}_j$ and the quasi uniformity of $\mathcal{T}_0$, such that*

$$\frac{\operatorname{diam}(\tau)}{\operatorname{diam}(\tau')} \le c \quad \forall \tau, \tau' \in \mathcal{T}_j, \quad \tau \cap \tau' \neq \emptyset. \tag{3.5}$$

Proof. τ and τ' are either face-adjacent (d vertices), edge-adjacent ($d-1$ vertices), or vertex-adjacent, and are handled by (3.1), (3.4), (3.3), respectively. Therefore we have

$$\begin{aligned}\frac{\operatorname{diam}(\tau)}{\operatorname{diam}(\tau')} &\le c\ 2^{|L(\tau)-L(\tau')|}\ \frac{\operatorname{diam}(\bar{\tau})}{\operatorname{diam}(\bar{\tau}')} \quad \text{(by (3.2))} \\ &\le c\ 2^{\max\{1,E,V\}}\ \gamma^{(0)} \quad \text{(by (3.1), (3.4), (3.3) and quasi uniformity of } \mathcal{T}_0). \ \square\end{aligned}$$

3.1. L_2-stable Riesz basis. Since patchwise quasi uniformity is established by (3.5), we can now take the first step in establishing the norm equivalence in section 5. In other words, our motivation is to form a stable basis in the following sense [29]:

$$\left\| \sum_{x_i \in \mathcal{N}_j} u_i \phi_i^{(j)} \right\|_{L_2(\Omega)} \eqsim \|\{\text{volume}^{1/2}(\text{supp } \phi_i^{(j)})\ u_i\}_{x_i \in \mathcal{N}_j}\|_{l_2}. \tag{3.6}$$

The basis stability (3.6) will then guarantee that the Bernstein inequality (2.5) holds. For a stable basis, functions with small supports have to be augmented by an appropriate scaling so that $\|\phi_i^{(j)}\|_{L_2(\Omega)}$ remains roughly the same for all basis functions. This is reflected in volume(supp $\phi_i^{(j)}$) by defining

$$L_{j,i} = \min\{L(\tau) : \tau \in \mathcal{T}_j,\ x_i \in \tau\}. \tag{3.7}$$

Then

$$\text{volume}(\text{supp } \phi_i^{(j)}) \eqsim 2^{-dL_{j,i}}.$$

We prefer to use an equivalent notion of basis stability; a basis is called an *L_2-stable Riesz basis* if

$$\left\| \sum_{x_i \in \mathcal{N}_j} \hat{u}_i \hat{\phi}_i^{(j)} \right\|_{L_2(\Omega)} \eqsim \|\{\hat{u}_i\}_{x_i \in \mathcal{N}_j}\|_{l_2}, \tag{3.8}$$

where $\hat{\phi}_i^{(j)}$ denotes the scaled basis, and the relationship between (3.6) and (3.8) is given as follows:

$$\hat{\phi}_i^{(j)} = 2^{d/2L_{j,i}}\ \phi_i^{(j)}, \quad \hat{u}_i = 2^{-d/2L_{j,i}}\ u_i, \quad x_i \in \mathcal{N}_j. \tag{3.9}$$

Then (3.8) forms the sufficient condition to establish the Bernstein inequality (2.5). This crucial property helps us to prove Theorem 8.2.

Remark 3.1. The analysis is done purely with basis functions, completely independent of the underlying mesh geometry. Furthermore, our construction works for any d-dimensional setting with the scaling (3.9). However, it is not clear how to define face-adjacency relations for $d > 3$. If such relations can be defined through some topological or geometrical abstraction, then our framework naturally extends to d-dimensional local refinement strategies, and hence the optimality of the BPX and WHB preconditioners can be guaranteed in $\mathbb{R}^d$, $d \ge 1$. One such generalization was given by Brandts, Korotov, and Krizek in [17]; also see the references therein.

4. Local smoothing computational complexity. In [11], the smoother is chosen to act on the local space

$$\tilde{\mathcal{S}}_j = \text{span}\left[\bigcup\{\phi_i^{(j)}\}_{i=N_{j-1}+1}^{N_j}\bigcup\{\phi_i^{(j)} \neq \phi_i^{(j-1)}\}_{i=1}^{N_{j-1}}\right].$$

Other choices for $\tilde{\mathcal{N}}_j$ are also possible, e.g., DOF which intersect the refinement region Ω_j [2, 14]. The only restriction is that $\tilde{\mathcal{N}}_j \subset \Omega_j$. For this particular choice, $\tilde{\mathcal{N}}_j = \{i = N_{j-1}+1,\ldots,N_j\}\bigcup\{i : \phi_i^{(j)} \neq \phi_i^{(j-1)}, \quad i = 1,\ldots,N_{j-1}\}$, the following result from [11] establishes a bound for the number of nodes used for smoothing (those created in Ω_j by the BEK procedure) so that the BPX preconditioner has provably optimal (linear) computational complexity per iteration.

LEMMA 4.1. *The total number of nodes used for smoothing satisfies the bound*

$$\sum_{j=0}^{J} \tilde{N}_j \leq \frac{5}{3}N_J - \frac{2}{3}N_0. \tag{4.1}$$

Proof. See [11, Lem. 1]. ∎

A similar result for 2D red-green refinement was given by Oswald [29, p. 95]. In the general case of local smoothing operators, which involve smoothing over newly created basis functions plus some additional sets of local neighboring basis functions, one can extend the arguments from [11] and [29] using shape regularity.

5. Establishing optimality of the BPX norm equivalence. In this section, we extend the Dahmen–Kunoth framework to three spatial dimensions; the extension closely follows the original work [19]. However, the general case for $d \geq 1$ spatial dimensions is not in the literature, and therefore we present it below.

For linear g, the element mass matrix gives rise to the following useful formula:

$$\|g\|_{L_2(\tau)}^2 = \frac{\text{volume}(\tau)}{(d+1)(d+2)}\left(\sum_{i=1}^{d+1} g(x_i)^2 + \left[\sum_{i=1}^{d+1} g(x_i)\right]^2\right), \tag{5.1}$$

where $i = 1,\ldots,d+1$ and x_i is a vertex of τ, $d = 2, 3$. In view of (5.1), we have that

$$\|\hat{\phi}_i^{(j)}\|_{L_2(\Omega)}^2 = 2^{dL_{j,i}}\ \frac{\text{volume}(\text{supp}\ \hat{\phi}_i^{(j)})}{(d+1)(d+2)}.$$

Since the min in (3.7) is attained, there exists at least one $\tau \in \text{supp}\ \hat{\phi}_i^{(j)}$ such that $L(\tau) = L_{j,i}$. By (3.2) we have

$$2^{L_{j,i}} \eqsim \frac{\text{diam}(\tau)}{\text{diam}(\bar{\tau})}. \tag{5.2}$$

Also,

$$\text{volume}(\text{supp}\ \hat{\phi}_i^{(j)}) \eqsim \sum_{i=1}^{E} \text{diam}^d(\tau_i), \quad \tau_i \in \text{supp}\ \hat{\phi}_i^{(j)}. \tag{5.3}$$

By (3.5), we have

$$\text{diam}(\tau_i) \eqsim \text{diam}(\tau). \tag{5.4}$$

Combining (5.3) and (5.4), we conclude that

$$\text{volume}(\text{supp } \hat{\phi}_i^{(j)}) \eqsim E \text{ diam}^d(\tau). \tag{5.5}$$

Finally, (5.2) and (5.5) yield

$$2^{dL_{j,i}} \text{volume}(\text{supp } \hat{\phi}_i^{(j)}) \eqsim E \frac{1}{\text{diam}^d(\bar{\tau})}.$$

E is a uniformly bounded constant by shape regularity. One can view the size of any tetrahedron in $\mathcal{T}_0$, in particular of size $\bar{\tau}$, as a constant. The reason is the following: Assumptions 3.4 and 3.5 force every tetrahedron at any $\mathcal{T}_j$ to be geometrically similar to some tetrahedron in $\mathcal{T}_0$ or to a tetrahedron arising from an irregular refinement of some tetrahedron in $\mathcal{T}_0$, and hence to some tetrahedron of a fixed finite collection. Combining the two arguments above, we have established that

$$\|\hat{\phi}_i^{(j)}\|_{L_2(\Omega)} \eqsim 1, \quad x_i \in \mathcal{N}_j. \tag{5.6}$$

Let $g = \sum_{x_i \in \mathcal{N}_j} \hat{u}_i \hat{\phi}_i^{(j)} \in \mathcal{S}_j$. For any $\tau \in \mathcal{T}_j$ we have that

$$\|g\|^2_{L_2(\tau)} \leq c \sum_{x_i \in \mathcal{N}_{j,\tau}} |\hat{u}_i|^2 \|\hat{\phi}_i^{(j)}\|^2_{L_2(\Omega)}, \tag{5.7}$$

where $\mathcal{N}_{j,\tau} = \{x_i \in \mathcal{N}_j : x_i \in \tau\}$, which is uniformly bounded in $\tau \in \mathcal{T}_j$ and $j \in \mathbb{N}_0$. By the scaling (3.9), we get equality in the estimate below. The inequality is a standard inverse inequality, where one bounds $g(x_i)$ using formula (5.1) and by handling the volume in the formula by (3.2):

$$|\hat{u}_i|^2 = 2^{-dL_{j,i}} |g(x_i)|^2 \leq c\, 2^{-dL_{j,i}} 2^{dL_{j,i}} \|g\|^2_{L_2(\tau)}. \tag{5.8}$$

Now, we are ready to establish that our basis is an L_2-stable Riesz basis as in (3.8). This is achieved by simply summing up over $\tau \in \mathcal{T}_j$ in (5.7) and (5.8) and using (5.6). L_2-stability in the Riesz sense allows us to establish the Bernstein inequality (2.5).

LEMMA 5.1. *For the scaled basis* (3.9), *the Bernstein inequality* (2.5) *holds for* $\beta = 3/2$.

Proof. Equation (5.6) with (5.7) and (5.8) asserts that the scaled basis (3.9) is stable in the sense of (3.8). Hence, (2.5) holds by [29, Thm. 4]. Note that the proof actually works independently of the spatial dimension. ∎

6. Lower bound in the norm equivalence. The Jackson inequality for Besov spaces is defined as follows:

$$\inf_{g \in \mathcal{S}_J} \|f - g\|_{L_p} \leq c\, \omega_\alpha(f, 2^{-J})_p, \quad f \in L_p(\Omega), \tag{6.1}$$

where c is a constant independent of f and J, and α is an integer. In the uniform refinement setting, (6.1) is used to obtain the lower bound in (2.7). However, in the local refinement setting, (6.1) holds only for functions whose singularities are somehow well captured by the mesh geometry. For instance, if a mesh is designed to pick up the singularity at $x = 0$ of $y = 1/x$, then on the same mesh we will not be able to recover a singularity at $x = 1$ of $y = 1/(x-1)$. Hence, the Jackson inequality (6.1) cannot hold in a general setting, i.e., for $f \in W_p^k$. In order to get the lower bound in (2.7), we focus on estimating θ_J directly, as in [19] for the 2D setting.

To begin, we borrow the quasi-interpolant construction from [19], extending it to the 3D setting. Let $\tau \in \mathcal{T}_j$ be a tetrahedron with vertices x_1, x_2, x_3, x_4. Clearly the restrictions of $\hat{\phi}_i^{(j)}$ to τ are linearly independent over τ, where $x_i \in \{x_1, x_2, x_3, x_4\}$. Then, there exists a unique set of linear polynomials $\psi_1^\tau, \psi_2^\tau, \psi_3^\tau, \psi_4^\tau$ such that

$$\int_\tau \hat{\phi}_k^{(j)}(x,y,z)\psi_l^\tau(x,y,z)dxdydz = \delta_{kl}, \quad x_k, x_l \in \{x_1, x_2, x_3, x_4\}. \tag{6.2}$$

For $x_i \in \mathcal{N}_j$ and $\tau \in \mathcal{T}_j$, define a function for $x_i \in \tau$,

$$M_i^{(j)}(x,y,z) = \begin{cases} \frac{1}{E_i}\psi_i^\tau(x,y,z), & (x,y,z) \in \tau, \\ 0, & (x,y,z) \notin \text{supp } \hat{\phi}_i^{(j)}, \end{cases} \tag{6.3}$$

where E_i is the number of tetrahedra in $\mathcal{T}_j$ in supp $\hat{\phi}_i^{(j)}$. By (6.2) and (6.3), we obtain

$$(M_k^{(j)}, \hat{\phi}_l^{(j)}) = \int_\Omega M_k^{(j)}(x,y,z)\hat{\phi}_l(x,y,z)\ dxdydz = \delta_{kl}, \quad x_k, x_l \in \mathcal{N}_j. \tag{6.4}$$

We can now define a quasi interpolant, and in fact a *projection* onto $\mathcal{S}_j$, such that

$$(\tilde{Q}_j f)(x,y,z) = \sum_{x_i \in \mathcal{N}_j} (f, M_i^{(j)})\hat{\phi}_i^{(j)}(x,y,z). \tag{6.5}$$

As remarked earlier, due to (6.3) the slice operator term $\tilde{Q}_j - \tilde{Q}_{j-1}$ will vanish outside the refined set Ω_j defined in (2.1). One can easily observe by (5.6) and (6.4) that

$$\|M_i^{(j)}\|_{L_2(\Omega)} \eqsim 1, \quad x_i \in \mathcal{N}_j,\ j \in \mathbb{N}_0. \tag{6.6}$$

Letting $\Omega_{j,\tau} = \bigcup\{\tau' \in \mathcal{T}_j :\ \tau \cap \tau' \neq \emptyset\}$, we can conclude from (5.6) and (6.6) that

$$\|\tilde{Q}_j f\|_{L_2(\tau)} = \| \sum_{x_k \in \mathcal{N}_{j,\tau}} (f, M_l^{(j)})\hat{\phi}_k^{(j)}\|_{L_2(\tau)} \leq c\|f\|_{L_2(\Omega_{j,\tau})}. \tag{6.7}$$

We define now a subset of the triangulation where the refinement activity stops, meaning that all tetrahedra in $\mathcal{T}_j^*,\ j \leq m$ also belong to $\mathcal{T}_m$:

$$\mathcal{T}_j^* = \{\tau \in \mathcal{T}_j :\ L(\tau) < j,\ \Omega_{j,\tau} \cap \tau' = \emptyset\ \forall \tau' \in \mathcal{T}_j \text{ with } L(\tau') = j\}. \tag{6.8}$$

Due to the local support of the dual basis functions $M_i^{(j)}$ and the fact that $\tilde{Q}_j$ is a projection, one gets for $g \in \mathcal{S}_J$,

$$\|g - \tilde{Q}_j g\|_{L_2(\tau)} = 0, \quad \tau \in \mathcal{T}_j^*. \tag{6.9}$$

Since $\tilde{Q}_j$ is a projection onto linear finite element space, it fixes polynomials of degree at most 1 (i.e., $\Pi_1(\mathbb{R}^3)$). Using this fact and (6.7), we arrive at

$$\begin{aligned} \|g - \tilde{Q}_j g\|_{L_2(\tau)} &\leq \|g - P\|_{L_2(\tau)} + \|\tilde{Q}_j(P - g)\|_{L_2(\tau)} \\ &\leq c\ \|g - P\|_{L_2(\Omega_{j,\tau})}, \quad \tau \in \mathcal{T}_j \setminus \mathcal{T}_j^*. \end{aligned} \tag{6.10}$$

We would like to bound the right-hand side of (6.10) in terms of a modulus of smoothness in order to reach a Jackson-type inequality. Following [19], we utilize a modified modulus of smoothness for $f \in L_p(\Omega)$,

$$\tilde{\omega}_k(f,t,\Omega)_p^p = t^{-s} \int_{[-t,t]^s} \|\Delta_h^k f\|_{L_p(\Omega_{k,h})}^p\ dh.$$

The two moduli of smoothness can be shown to be equivalent:

$$\tilde{\omega}_{k+1}(f,t,\Omega)_p \eqsim \omega_{k+1}(f,t,\Omega)_p.$$

The equivalence in the one-dimensional setting can be found in [20, Lem. 5.1].

For τ a simplex in $\mathbb{R}^d$ and $t = \operatorname{diam}(\tau)$, a Whitney estimate shows that [21, 28, 33]

$$\inf_{P\in\Pi_k(\mathbb{R}^d)} \|f-P\|_{L_p(\tau)} \le c\tilde{\omega}_{k+1}(f,t,\tau)_p, \tag{6.11}$$

where c depends only on the smallest angle of τ but not on f and t. The reason why $\tilde{Q}_j$ works well for tetrahedralization in three dimensions is the fact that the Whitney estimate (6.11) remains valid for any spatial dimension. $\mathcal{T}_j \setminus \mathcal{T}_j^*$ is the part of the tetrahedralization $\mathcal{T}_j$, where refinement is active at every level. Then, in view of (3.5),

$$\operatorname{diam}(\Omega_{j,\tau}) \eqsim 2^{-j}, \quad \tau \in \mathcal{T}_j \setminus \mathcal{T}_j^*.$$

Taking the inf over $P \in \Pi_1(\mathbb{R}^3)$ in (6.10) and using the Whitney estimate (6.11), we conclude that

$$\|g - \tilde{Q}_j g\|_{L_2(\tau)} \le c\tilde{\omega}_2(g, 2^{-j}, \Omega_{j,\tau})_2.$$

Recalling (6.9) and summing over $\tau \in \mathcal{T}_j \setminus \mathcal{T}_j^*$ gives rise to

$$\|g - \tilde{Q}_j g\|_{L_2(\Omega)} \le c\tilde{\omega}_2(g, 2^{-j}, \Omega)_2 \le \tilde{c}\ \omega_2(g, 2^{-j}, \Omega)_2,$$

where we have switched from the modified modulus of smoothness to the standard one. Since Q_j is an orthogonal projection, we have the following:

$$\|g - Q_j g\| \le \|g - \tilde{Q}_j g\|.$$

Using the above inequality with (2.6), one then has

$$v_J = O(1), \quad J \to \infty. \tag{6.12}$$

7. The WHB preconditioner. In local refinement, HB methods enjoy an optimal complexity of $O(N_j - N_{j-1})$ per iteration per level (resulting in $O(N_J)$ overall complexity per iteration) by using only DOF corresponding to $\mathcal{S}_j^f$. However, HB methods suffer from suboptimal iteration counts or, equivalently, suboptimal condition numbers. The BPX decomposition $\mathcal{S}_j = \mathcal{S}_{j-1} \oplus (Q_j - Q_{j-1})\mathcal{S}_j$ gives rise to basis functions which are not locally supported, but they decay rapidly outside a local support region. This allows for locally supported approximations; in addition, the WHB methods [34, 35, 36] can be viewed as an approximation of the wavelet basis stemming from the BPX decomposition [23]. A similar wavelet-like multilevel decomposition approach was taken in [32], where the orthogonal decomposition is formed by a discrete L_2-equivalent inner product. This approach utilizes the same BPX two-level decomposition [31, 32]. The WHB preconditioner is defined as follows:

$$Hu := \sum_{j=0}^{J} 2^{j(d-2)} \sum_{i\in\mathcal{N}_j^f} (u, \psi_i^{(j)})\psi_i^{(j)}, \tag{7.1}$$

where $\psi_i^{(j)} = (\tilde{Q}_j - \tilde{Q}_{j-1})\phi_i^{(j)}$. The WHB preconditioner uses the modified basis (whereas the BPX preconditioner uses the standard nodal basis) where the projection

operator used is defined as in (7.5). In the WHB setting, these operators are chosen to satisfy the following three properties [5]:

$$\tilde{Q}_j \mid_{\mathcal{S}_j} = I, \tag{7.2}$$

$$\tilde{Q}_j \tilde{Q}_k = \tilde{Q}_{\min\{j,k\}}, \tag{7.3}$$

$$\|(\tilde{Q}_j - \tilde{Q}_{j-1})u^{(j)}\|_{L_2} \eqsim \|u^{(j)}\|_{L_2}, \quad u^{(j)} \in (I_j - I_{j-1})\mathcal{S}_j. \tag{7.4}$$

As indicated in (2.2), the WHB smoother acts on only the fine DOF, i.e., $\mathcal{N}_j^f$, and hence is an approximation to a fine-fine discretization operator $A_{ff}^{(j)} : \mathcal{S}_j^f \to \mathcal{S}_j^f$, where $\mathcal{S}_j^f := (\tilde{Q}_j - \tilde{Q}_{j-1})\mathcal{S}_j$ and f stands for fine. On the other hand, the BPX smoother acts on a slightly bigger set than fine DOF, typically $\mathcal{N}_j^f \subseteq \tilde{\mathcal{N}}_j$, the union of fine DOF and their corresponding coarse fathers.

The WHB preconditioner introduced in [34, 35] is, in some sense, the best of both worlds. While the condition number of the HB preconditioner is stabilized by inserting Q_j into the definition of $\tilde{Q}_j$, somehow employing the operators $I_j - I_{j-1}$ at the same time guarantees optimal computational and storage cost per iteration. The operators which will be seen to meet both goals at the same time are

$$\tilde{Q}_k = \prod_{j=k}^{J-1} I_j + Q_j^a(I_{j+1} - I_j), \tag{7.5}$$

with $\tilde{Q}_J = I$. The exact L_2-projection Q_j is replaced with a computationally feasible approximation $Q_j^a : L_2 \to \mathcal{S}_j$. To control the approximation quality of Q_j^a, a small fixed tolerance γ is introduced:

$$\|(Q_j^a - Q_j)u\|_{L_2} \leq \gamma \|Q_j u\|_{L_2} \quad \forall u \in L_2(\Omega). \tag{7.6}$$

In the limiting case $\gamma = 0$, $\tilde{Q}_k$ reduces to the exact L_2-projection on $\mathcal{S}_J$ by (7.2):

$$\tilde{Q}_k = Q_k I_{k+1} Q_{k+1} \cdots I_{J-1} Q_{J-1} I_J = Q_k Q_{k+1} \cdots Q_{J-1} = Q_k.$$

Following [34, 35], the properties (7.2), (7.3), and (7.4) can be verified for $\tilde{Q}_k$ as follows:

• Property (7.2): Let $u^{(k)} \in \mathcal{S}_k$. Since $(I_{j+1} - I_j)u^{(k)} = 0$ and $I_j u^{(k)} = u^{(k)}$ for $k \leq j$, then $[I_j + Q_j^a(I_{j+1} - I_j)](u^{(k)}) = u^{(k)}$, verifying (7.2) for $\tilde{Q}_k$. It also implies

$$\tilde{Q}_k^2 = \tilde{Q}_k. \tag{7.7}$$

• Property (7.3): Let $k \leq l$; then by (7.7)

$$\tilde{Q}_k \tilde{Q}_l = [(I_k + Q_k^a(I_{k+1} - I_k)) \cdots (I_{l-1} + Q_{l-1}^a(I_l - I_{l-1}))\ \tilde{Q}_l]\tilde{Q}_l = \tilde{Q}_k. \tag{7.8}$$

Since $\tilde{Q}_k u \in \mathcal{S}_k$ and $\mathcal{S}_k \subset \mathcal{S}_l$, then by (7.2) we have

$$\tilde{Q}_l(\tilde{Q}_k u) = \tilde{Q}_k u. \tag{7.9}$$

Finally, (7.3) then follows from (7.8) and (7.9).

• Property (7.4): This is an implication of Lemma 7.1.

For an overview, we list the corresponding slice spaces for the preconditioners of interest:

HB:	$\mathcal{S}_j^f$	$=$	$(I_j - I_{j-1})\mathcal{S}_j$,
BPX:	$\mathcal{S}_j^f$	$=$	$(Q_j - Q_{j-1})\mathcal{S}_j$,
WHB:	$\mathcal{S}_j^f$	$=$	$(\tilde{Q}_j - \tilde{Q}_{j-1})\mathcal{S}_j = (I - Q_{j-1}^a)(I_j - I_{j-1})\mathcal{S}_j$, where $\tilde{Q}_j$ is given in (7.5).

The WHB smoother acts only on the fine DOF. Then, in the generic multilevel preconditioner notation, the WHB preconditioner can be written in the following form:

$$Bu := \sum_{j=0}^{J} B_{ff}^{(j)^{-1}} (\tilde{Q}_j - \tilde{Q}_{j-1})u. \tag{7.10}$$

B_{ff} is chosen to be a spectrally equivalent operator to fine-fine discretization operator $A_{ff}^{(j)}$. Since the smoother and property (7.4) both rely on a well-conditioned $A_{ff}^{(j)}$, we discuss this next.

7.1. Well-conditioned $A_{ff}^{(j)}$. The lemma below is essential to extend the existing results for quasi-uniform meshes (see [34, Lem. 6.1] or [35, Lem. 2]) to the locally refined ones. $\mathcal{S}_j^{(f)} = (I_j - I_{j-1})\mathcal{S}_j$ denotes the HB slice space.

LEMMA 7.1. *Let $\mathcal{T}_j$ be constructed by the local refinements under consideration. Let $\mathcal{S}_j^f = (I - \tilde{Q}_{j-1})\mathcal{S}_j^{(f)}$ be the modified hierarchical subspace, where $\tilde{Q}_{j-1}$ is any L_2-bounded operator. Then, there are constants c_1 and c_2 independent of j such that*

$$c_1 \|\phi^f\|_X^2 \le \|\psi^f\|_X^2 \le c_2 \|\phi^f\|_X^2, \quad X = H^1, L_2, \tag{7.11}$$

holds for any $\psi^f = (I - \tilde{Q}_{j-1})\phi^f \in \mathcal{S}_j^f$ with $\phi^f \in \mathcal{S}_j^{(f)}$.

Proof. The Cauchy–Schwarz like inequality [8] is central to the proof: There exists $\delta \in (0,1)$ independent of the mesh size or level j such that

$$(1-\delta^2)(\nabla\phi^f, \nabla\phi^f) \le (\nabla(\phi^c + \phi^f), \nabla(\phi^c + \phi^f)) \quad \forall \phi^c \in \mathcal{S}_{j-1}, \phi^f \in \mathcal{S}_j^{(f)}, \tag{7.12}$$

$$(1-\delta^2)\|\phi^f\|_{L_2}^2 \le c|\phi^c + \phi^f|_{H^1}^2 \quad \text{(by the Poincaré inequality and (7.12))}. \tag{7.13}$$

Combining (7.12) and (7.13), we get $(1-\delta^2)\|\phi^f\|_{H^1}^2 \le \|\phi^c + \phi^f\|_{H^1}^2$. Choosing $\phi^c = -\tilde{Q}_{j-1}\phi^f$, we get the lower bound $(1-\delta^2)\|\phi^f\|_{H^1}^2 \le \|\psi^f\|_{H^1}^2$.

Let Ω_j^f denote the support of basis functions corresponding to $\mathcal{N}_j^f$. Due to nested refinement, triangulation on Ω_j^f is quasi uniform. One can analogously introduce a triangulation hierarchy, where all the simplices are exposed to uniform refinement: $\mathcal{T}_j^f := \{\tau \in \mathcal{T}_j : L(\tau) = j\} = \mathcal{T}_j|_{\Omega_j^f}$. Hence, $\mathcal{T}_j^f$ becomes a quasi-uniform tetrahedralization and the inverse inequality holds for $\mathcal{S}_j^f$. The upper bound is derived by using a father-son size relation, the inverse inequalities, and L_2-boundedness of $\tilde{Q}_{j-1}$. Hence, one gets

$$\|\psi^f\|_{H^1}^2 \le c_0 2^{2j} \|\psi^f\|_{L_2}^2 \le c_0 2^{2j} \left(1 + \|\tilde{Q}_{j-1}\|_{L_2}\right)^2 \|\phi^f\|_{L_2}^2 \le c 2^{2j} \|\phi^f\|_{L_2}^2.$$

The slice space $\mathcal{S}_j^{(f)}$ is oscillatory. Then there exists c such that $\|\phi^f\|_{L_2}^2 \le c 2^{-2j} \|\phi^f\|_{H^1}^2$. Hence, $\|\psi^f\|_{H^1}^2 \le c\|\phi^f\|_{H^1}^2$. The case for $X = L_2$ can be established similarly. ∎

Using the above tools, one can establish that $A_{ff}^{(j)}$ is well conditioned. Namely,

$$c_1 2^{2j} \leq \lambda_{j,\min}^f \leq \lambda_{j,\max}^f \leq c_2 2^{2j}, \tag{7.14}$$

where $\lambda_{j,\min}^f$ and $\lambda_{j,\max}^f$ are the smallest and largest eigenvalues of $A_{ff}^{(j)}$, and both c_1 and c_2 are independent of j. For details see [34, Lem. 4.3] or [35, Lem. 3].

8. The fundamental assumption and WHB optimality. As in the BPX splitting, the main ingredient in the WHB splitting is the L_2-projection. Hence, the stability of the BPX splitting is still important in the WHB splitting. The lower bound in the BPX norm equivalence is the *fundamental assumption* for the WHB preconditioner. Utilizing a local projection $\tilde{Q}_j$, the BPX lower bound was verified earlier for 3D a local red-green (BEK) refinement procedure. The same result easily holds for the projection Q_j. Dahmen and Kunoth [19] verified a BPX lower bound for the 2D red-green refinement procedures.

Before getting to the stability result we remark that the existing perturbation analysis of WHB is one of the primary insights in [34, 35]. Although not observed in [34, 35], the result does not require substantial modification for locally refined meshes. Let $e_j := (\tilde{Q}_j - Q_j)u$ be the error; then the following holds.

LEMMA 8.1. *Let γ be as in* (7.6). *There exists an absolute c satisfying*

$$\sum_{j=0}^{J} 2^{2j} \|e_j\|_{L_2}^2 \leq c\gamma^2 \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2 \quad \forall u \in \mathcal{S}_J. \tag{8.1}$$

Proof. See [34, Lem. 5.1] or [35, Lem. 1]. ∎

We arrive now at the primary result, which indicates that the WHB slice norm is optimal in the class of locally refined meshes under consideration.

THEOREM 8.2. *If there exists sufficiently small γ_0 such that* (7.6) *is satisfied for $\gamma \in [0, \gamma_0)$, then*

$$\|u\|_{\text{WHB}}^2 = \sum_{j=0}^{J} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|_{L_2}^2 \eqsim \|u\|_{H^1}^2, \quad u \in \mathcal{S}_J. \tag{8.2}$$

Proof. Observe that

$$\begin{aligned} (\tilde{Q}_j - \tilde{Q}_{j-1})u &= (\tilde{Q}_j - Q_j)u - (\tilde{Q}_{j-1} - Q_{j-1})u + (Q_j - Q_{j-1})u \\ &= e_j - e_{j-1} + (Q_j - Q_{j-1})u. \end{aligned} \tag{8.3}$$

This gives

$$\begin{aligned} \sum_{j=0}^{J} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|_{L_2}^2 &\leq c \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2 + c \sum_{j=0}^{J} 2^{2j} \|e_j\|_{L_2}^2 \\ &\leq c(1+\gamma^2) \sum_{j=0}^{J} 2^{2j} \|(Q_j - Q_{j-1})u\|_{L_2}^2 \quad \text{(using (8.1))} \\ &\leq c\|u\|_{H^1}^2. \end{aligned}$$

Let us now proceed with the upper bound. The Bernstein inequality (2.5) holds for $\mathcal{S}_j$ [1, 19] for the local refinement procedures. Hence, we are going to utilize an inequality

involving the Besov norm $\|\cdot\|_{B^1_{2,2}}$ which naturally fits into our framework when the moduli of smoothness is considered in (2.5). The following important inequality holds, provided that (2.5) holds [29, p. 39]:

$$\|u\|^2_{B^1_{2,2}} \leq c \sum_{j=0}^{J} 2^{2j} \|u^{(j)}\|^2_{L_2} \tag{8.4}$$

for any decomposition such that $u = \sum_{j=0}^{J} u^{(j)}$, $u^{(j)} \in \mathcal{S}_j$, in particular for $u^{(j)} = (\tilde{Q}_j - \tilde{Q}_{j-1})u$. Then the upper bound holds due to $H^1(\Omega) \cong B^1_{2,2}(\Omega)$. ∎

Remark 8.1. The following equivalence is used for the upper bound in the proof of Theorem 8.2 on uniformly refined meshes [35, Lem. 4]:

$$c_1 \|u\|^2_{H^1} \leq \inf_{\substack{u=\sum_{j=0}^{J} u^{(j)}, \\ u^{(j)} \in \mathcal{S}_j}} \sum_{j=0}^{J} 2^{2j} \|u^{(j)}\|^2_{L_2} \leq c_2 \|u\|^2_{H^1}.$$

Let us emphasize that the left-hand side holds in the presence of the Bernstein inequality (2.5), and the right-hand side holds in the simultaneous presence of Bernstein and Jackson inequalities. However, the Jackson inequality cannot hold under local refinement procedures (cf. the counterexample in section 6). That is why we can utilize only the left-hand side of the above equivalence as in (8.4).

Now, we have all the required estimates at our disposal for establishing the optimality of WHB preconditioner for 2D and 3D red-green refinement procedures for $p \in L_\infty(\Omega)$. We would like to emphasize that our framework supports any spatial dimension $d \geq 1$, provided that the necessary geometrical abstractions are in place.

THEOREM 8.3. *If a BPX lower bound holds and if there exists sufficiently small* γ_0 *such that* (7.6) *is satisfied for* $\gamma \in (0, \gamma_0)$, *then for* B *in* (7.10),

$$(Bu, u) \eqsim \|u\|^2_{H^1}.$$

Proof. $B^{(j)}_{ff}$ is spectrally equivalent to $A^{(j)}_{ff}$. Since $A^{(j)}_{ff}$ is a well-conditioned matrix, by using (7.14) it is spectrally equivalent to $2^{2j} I$. The result follows from Theorem 8.2. ∎

An extension to a multiplicative WHB preconditioner is also possible under additional assumptions. These results will not be reported here.

9. H^1-stable L_2-projection. The involvement of $\tilde{Q}_j$ in the multilevel decomposition makes it the most crucial element in the stabilization. We then come to the central question, Which choice of $\tilde{Q}_j$ can provide an optimal preconditioner? The following theorem sets a guideline for picking $\tilde{Q}_j$. It shows that H^1-stability of $\tilde{Q}_j$ is actually a *necessary condition* for obtaining an optimal preconditioner.

THEOREM 9.1 (see [34, 35]). *If* $\tilde{Q}_j$ *induces an optimal preconditioner, namely for* $u \in \mathcal{S}_J$, $\sum_{j=0}^{J} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|^2_{L_2} \eqsim \|u\|^2_{H^1}$, *then there exists an absolute constant* c *such that*

$$\|\tilde{Q}_k u\|_{H^1} \leq c \, \|u\|_{H^1} \quad \forall k \leq J.$$

Proof. Using the multilevel decomposition and (7.3), we get $\tilde{Q}_k u = \sum_{j=0}^{k} (\tilde{Q}_j - \tilde{Q}_{j-1})u$. Since $\tilde{Q}_j$ induces an optimal preconditioner, there exist two absolute

constants σ_1 and σ_2:

$$\sigma_1 \|u\|_{H^1}^2 \leq \sum_{j=0}^{J} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|_{L_2}^2 \leq \sigma_2 \|u\|_{H^1}^2 \quad \forall u \in \mathcal{S}_J. \tag{9.1}$$

Using (9.1) for $\tilde{Q}_k u$, we get

$$\|\tilde{Q}_k u\|_{H^1}^2 \leq \frac{1}{\sigma_1} \sum_{j=0}^{k} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|_{L_2}^2 \leq \frac{1}{\sigma_1} \sum_{j=0}^{J} 2^{2j} \|(\tilde{Q}_j - \tilde{Q}_{j-1})u\|_{L_2}^2 \leq \frac{\sigma_2}{\sigma_1} \|u\|_{H^1}^2. \qquad \square$$

As a consequence of Theorem 9.1 we have the following corollary.

COROLLARY 9.2. *L_2-projection restricted to $\mathcal{S}_j$, $Q_j|_{\mathcal{S}_j} : L_2 \to \mathcal{S}_j$ is H^1-stable on $2D$ and $3D$ locally refined meshes by red-green refinement procedures.*

Proof. Optimality of the BPX norm equivalence on the above locally refined meshes was already established. Application of Theorem 9.1 with Q_j proves the result. Alternatively, the same result can be obtained through Theorem 9.1 applied to the WHB framework. Theorem 8.2 will establish the optimality of the WHB preconditioner for the local refinement procedures. Hence, the operator $\tilde{Q}_j$ restricted to $\mathcal{S}_j$ is H^1-stable. Since $\tilde{Q}_j$ is none other than Q_j in the limiting case, we can also conclude the H^1-stability of the L_2-projection. $\square$

Our stability result appears to be the first a priori H^1-stability for the L_2-projection on these classes of locally refined meshes. H^1-stability of L_2-projection is guaranteed for the subset $\mathcal{S}_j$ of $L_2(\Omega)$, but not for all of $L_2(\Omega)$. This problem is currently undergoing intensive study in the finite element and approximation theory communities. The existing theoretical results, mainly those in [15, 18], involve a posteriori verification of somewhat complicated mesh conditions after refinement has taken place. If such mesh conditions are not satisfied, one has to redefine the mesh. The mesh conditions mentioned require that the simplex sizes do not change drastically between regions of refinement. In this context, quasi uniformity in the support of a basis function becomes crucial. This type of local quasi uniformity is usually called *patchwise quasi uniformity*. Local quasi uniformity requires neighbor generation relations as in (3.1), neighbor size relations, and shape regularity of the mesh. It was shown in [1] that patchwise quasi uniformity holds also for 3D marked tetrahedron bisection [24] and for 2D newest vertex bisection [25, 30]. These are promising refinement procedures for which H^1-stability of the L_2-projection can be established.

10. Conclusion. In this article, we examined the Bramble–Pasciak–Xu (BPX) norm equivalence in the setting of local 3D mesh refinement. In particular, we extended the 2D optimality result for BPX due to Dahmen and Kunoth to the local 3D red-green refinement procedure introduced by Bornemann, Erdmann, and Kornhuber (BEK procedure). The extension involved establishing that the locally enriched finite element subspaces produced by the BEK procedure allow for the construction of a scaled basis which is formally Riesz stable. This in turn rested entirely on establishing a number of geometrical relationships between neighboring simplices produced by the local refinement algorithms. We remark again that shape regularity of the elements produced by the refinement procedure is insufficient to construct a stable Riesz basis for finite element spaces on locally adapted meshes. The d-vertex adjacency generation bound for simplices in $\mathbb{R}^d$ is the primary result required to establish patchwise quasi uniformity for stable Riesz basis construction, and this result depends critically

on the particular details of the local refinement procedure rather than on shape regularity of the elements. We also noted in section 3 that these geometrical properties have been established in [1] for purely bisection-based refinement procedures that have been shown to be asymptotically nondegenerate, and therefore also allow for the construction of a stable Riesz basis.

We also examined the wavelet modified hierarchical basis (WHB) methods of Vassilevski and Wang and extended their original quasi-uniformity-based framework and results to local 2D and 3D red-green refinement scenarios. A critical step in the extension involved establishing the optimality of the BPX norm equivalence for the local refinement procedures under consideration, as established in the first part of this article. With the local refinement extension of the WHB analysis framework presented here, we established the optimality of the WHB preconditioner on locally refined meshes in both two and three dimensions under the minimal regularity assumptions required for well-posedness. An interesting implication of the optimality of WHB preconditioner was the a priori H^1-stability of the L_2-projection. Existing a posteriori approaches in the literature dictate a reconstruction of the mesh if such conditions cannot be satisfied.

The theoretical framework established here supports arbitrary spatial dimension $d \geq 1$, and therefore allows for extension of the optimality results, the H^1-stability of L_2-projection results, and the various supporting results to arbitrary $d \geq 1$. We indicated clearly which geometrical properties must be re-established to show BPX optimality for spatial dimension $d \geq 4$. All of the results here require no smoothness assumptions on the PDE coefficients beyond those required for well-posedness in H^1.

To address the practical computational complexity of implementable versions of the BPX and WHB preconditioners, we indicated how the number of degrees of freedom (DOF) used for the smoothing step can be shown to be bounded by a constant times the number of DOF introduced at that level of refinement. This indicates that practical implementable versions of the BPX and WHB preconditioners for the local 3D refinement setting considered here have provably optimal (linear) computational complexity per iteration. A detailed analysis of both the storage and per-iteration computational complexity questions arising with BPX and WHB implementations can be found in the second article in our series [2].

Acknowledgments. The authors thank R. Bank, P. Vassilevski, and J. Xu for many enlightening discussions.

REFERENCES

[1] B. Aksoylu, *Adaptive Multilevel Numerical Methods with Applications in Diffusive Biomolecular Reactions*, Ph.D. thesis, Department of Mathematics, University of California at San Diego, La Jolla, CA, 2001. Available online at http://www.cct.lsu.edu/~burak/pubs/Aksoylu_thesis.ps.gz

[2] B. Aksoylu, S. Bond, and M. Holst, *An odyssey into local refinement and multilevel preconditioning* III: *Implementation and numerical experiments*, SIAM J. Sci. Comput., 25 (2003), pp. 478–498.

[3] B. Aksoylu, M. Holst, and S. Bond, *Implementation and Theoretical Aspects of the BPX Preconditioner in the Three Dimensional Local Mesh Refinement Setting*, Tech. report, ICES Report 04-50, Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, 2004. Available online at http://www.ices.utexas.edu/research/reports/2004/0450.pdf

[4] B. Aksoylu and M. Holst, *An Odyssey into Local Refinement and Multilevel Preconditioning* I: *Optimality of the BPX Preconditioner*, Tech. report, ICES Report 05-03, Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, 2005. Available online at http://www.ices.utexas.edu/research/reports/2005/0503.pdf

[5] B. Aksoylu and M. Holst, *An Odyssey into Local Refinement and Multilevel Preconditioning* II: *Stabilizing Hierarchical Basis Methods*, Tech. report, ICES Report 05-04, Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, 2005. Available online at http://www.ices.utexas.edu/research/reports/2005/0504.pdf

[6] B. Aksoylu, A. Khodakovsky, and P. Schröder, *Multilevel solvers for unstructured surface meshes*, SIAM J. Sci. Comput., 26 (2005), pp. 1146–1165.

[7] R. E. Bank, *Hierarchical basis and the finite element method*, Acta Numer., 5 (1996), pp. 1–43.

[8] R. E. Bank and T. F. Dupont, *Analysis of a Two-Level Scheme for Solving Finite Element Equations*, Tech. report, CNA–159, Center for Numerical Analysis, University of Texas at Austin, Austin, TX, 1980.

[9] R. E. Bank, T. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.

[10] J. Bey, *Tetrahedral grid refinement*, Computing, 55 (1995), pp. 271–288.

[11] F. Bornemann, B. Erdmann, and R. Kornhuber, *Adaptive multilevel methods in three space dimensions*, Intl. J. Numer. Meth. Eng., 36 (1993), pp. 3187–3203.

[12] F. Bornemann and H. Yserentant, *A basic norm equivalence for the theory of multilevel methods*, Numer. Math., 64 (1993), pp. 455–476.

[13] J. H. Bramble and J. E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.

[14] J. H. Bramble and J. E. Pasciak, *New estimates for multilevel algorithms including the V-cycle*, Math. Comp., 60 (1993), pp. 447–471.

[15] J. H. Bramble, J. E. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 71 (2001), pp. 147–156.

[16] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[17] J. Brandts, S. Korotov, and M. Krizek, *The strengthened Cauchy-Bunyakowski-Schwarz inequality for n-simplicial linear finite elements,* in Proceedings of the Third Conference on Numerical Analysis and Applications (Bulgaria, 2004), Lecture Notes in Comput. Sci. 3401, Z. Li et al., eds., Springer, Berlin, 2005, pp. 203–210.

[18] C. Carstensen, *Merging the Bramble-Pasciak-Steinbach and the Crouzeix-Thome criterion for H^1-stability of the L^2-projection onto finite element spaces*, Math. Comp., 71 (2001), pp. 157–163.

[19] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[20] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin, 1993.

[21] R. A. DeVore and V. A. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[22] M. Holst, *Adaptive numerical treatment of elliptic systems on manifolds*, Adv. Comput. Math., 15 (2002), pp. 139–191.

[23] S. Jaffard, *Wavelet methods for fast resolution of elliptic problems*, SIAM J. Numer. Anal., 29 (1992), pp. 965–986.

[24] A. Liu and B. Joe, *Quality local refinement of tetrahedral meshes based on bisection*, SIAM J. Sci. Comput., 16 (1995), pp. 1269–1291.

[25] W. F. Mitchell, *Unified Multilevel Adaptive Finite Element Methods for Elliptic Problems*, Ph.D. thesis, Computer Science Department, University of Illinois at Urbana-Champaign, Urbana, IL, 1988.

[26] M. E. G. Ong, *Hierarchical Basis Preconditioners for Second Order Elliptic Problems in Three Dimensions*, Ph.D. thesis, University of Washington, Seattle, WA, 1989.

[27] M. E. G. Ong, *Uniform refinement of a tetrehedron*, SIAM J. Sci. Comput., 15 (1994), pp. 1134–1144.

[28] P. Oswald, *On function spaces related to finite element approximation theory*, Z. Anal. Anwendungen, 9 (1990), pp. 43–64.

[29] P. Oswald, *Multilevel Finite Element Approximation. Theory and Applications*, Teubner Skri. Numer., B. G. Teubner, Stuttgart, 1994.

[30] E. G. Sewell, *Automatic Generation of Triangulations for Piecewise Polynomial Approximation*, Ph.D. thesis, Department of Mathematics, Purdue University, West Lafayette, IN, 1972.

[31] R. Stevenson, *Robustness of the additive multiplicative frequency decomposition multi-level method*, Computing, 54 (1995), pp. 331–346.

[32] R. Stevenson, *A robust hierarchical basis preconditioner on general meshes*, Numer. Math., 78 (1997), pp. 269–303.

[33] E. A. Storozhenko and P. Oswald, *Jackson's theorem in the spaces $L_p(\mathbb{R}^k)$, $0 < p < 1$*, Siberian Math., 19 (1978), pp. 630–639.

[34] P. S. Vassilevski and J. Wang, *Stabilizing the hierarchical basis by approximate wavelets*, I: *Theory*, Numer. Linear Algebra Appl., 4 (1997), pp. 103–126.

[35] P. S. Vassilevski and J. Wang, *Wavelet-like methods in the design of efficient multilevel preconditioners for elliptic PDEs*, in Multiscale Wavelet Methods For Partial Differential Equations, W. Dahmen, A Kurdila, and P. Oswald, eds., Academic Press, New York, 1997, pp. 59–105.

[36] P. S. Vassilevski and J. Wang, *Stabilizing the hierarchical basis by approximate wavelets* II: *Implementation and numerical results*, SIAM J. Sci. Comput., 20 (1998), pp. 490–514.

[37] H. Yserentant, *On the multilevel splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

[38] S. Zhang, *Multilevel Iterative Techniques*, Ph.D. thesis, Pennsylvania State University, State College, PA, 1988.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1026–1048

ON THE EVALUATION OF HIGHLY OSCILLATORY INTEGRALS BY ANALYTIC CONTINUATION*

DAAN HUYBRECHS† AND STEFAN VANDEWALLE†

Abstract. We consider the integration of one-dimensional highly oscillatory functions. Based on analytic continuation, rapidly converging quadrature rules are derived for a general class of oscillatory integrals with an analytic integrand. The accuracy of the quadrature increases both for the case of a fixed number of points and increasing frequency, and for the case of an increasing number of points and fixed frequency. These results are then used to obtain quadrature rules for more general oscillatory integrals, i.e., for functions that exhibit some smoothness but that are not analytic. The approach described in this paper is related to the steepest descent method, but it does not employ asymptotic expansions. It can be used for small or moderate frequencies as well as for very high frequencies. The approach is compared with the oscillatory integration techniques recently developed by Iserles and Nørsett.

Key words. oscillatory functions, steepest descent method, numerical quadrature

AMS subject classifications. 65D30, 41A60

DOI. 10.1137/050636814

1. Introduction.

Consider the oscillatory integral

$$I := \int_a^b f(x) e^{i\omega g(x)} \, \mathrm{d}x \tag{1.1}$$

with $\omega > 0$ and with $f(x)$ and $g(x)$ smooth functions. Integrals of this form abound in mathematical models and computational algorithms for oscillatory phenomena in science and engineering. Recently, much progress has been made in numerical quadrature techniques for (1.1). Methods have been devised that compute an accurate approximation to the value of the integral with low computational complexity and with a number of operations that actually decreases as ω increases to infinity [12, 13, 14, 15, 16, 17, 18]. This is in contrast to most classical integration approaches, based on polynomial interpolation, that rapidly deteriorate in the presence of strong oscillations. In order to appreciate the inner workings of these methods, one should understand the asymptotic behavior of the oscillatory integral (1.1) for large values of the parameter ω.

The value of I at large frequencies depends on the behavior of the smooth functions f and g near the endpoints a and b, and near the so-called stationary points. The latter are the solutions to the equation $g'(x) = 0$ on $[a, b]$; they represent points in which the integrand locally does not oscillate. An intuitive justification of this property may be that, away from the endpoints and the stationary points, the oscillations of the integrand increasingly cancel out. Mathematically, the property is reflected in the asymptotic expansion of I. We say that a stationary point ξ has order r if $g^{(j)}(\xi) = 0$, $j = 1, \ldots, r$, but $g^{(r+1)}(\xi) \neq 0$; i.e., the first r derivatives of the

*Received by the editors July 25, 2005; accepted for publication (in revised form) January 5, 2006; published electronically May 19, 2006.

http://www.siam.org/journals/sinum/44-3/63681.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium (daan.huybrechs@cs.kuleuven.be, stefan.vandewalle@cs.kuleuven.be). The first author was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

oscillator vanish. Assuming one stationary point ξ of order r in the interval $[a, b]$, the asymptotic expansion of (1.1) has the form

$$I \sim \sum_{j=0}^{\infty} \frac{a_j}{\omega^{(j+1)/(r+1)}}, \tag{1.2}$$

where the coefficients a_j depend only on a finite number of function values and derivatives of f and g at the critical points a, b, and ξ [20]. The coefficients are in general not easily obtained, although the leading order coefficient a_0 is given by the method of stationary phase. Still, the mere existence of the asymptotic expansion reveals a lot of information about I. For example, an immediate consequence is that $|I| = O(\omega^{-1/(r+1)})$.

A first efficient method is to simply truncate the asymptotic expansion (1.2) after a finite number of terms. By construction, the truncation error decays as a power of $1/\omega$. This *asymptotic method* was described by Iserles and Nørsett in [14]. The problem of the unknown coefficients in the presence of stationary points is solved by constructing a uniform asymptotic expansion, based on factoring out the moment $\mu_0 = \int_a^b e^{i\omega g(x)}\,\mathrm{d}x$, or similar higher order moments. The coefficients in this expansion can be computed explicitly if the moments themselves are known a priori. A disadvantage of such an approach is that the error of an asymptotic expansion is essentially uncontrollable, since asymptotic expansions tend to diverge. This is especially true for smaller frequencies.

A second approach, proposed also in [14], is to extend Filon's method for oscillatory integrals (see [9, 5]) by considering Hermite interpolation of f. The result is a quadrature rule for I with a classical form, involving function values and derivatives of f. The error of this approach is controllable and may be very small. A disadvantage is that the weights of the rule are given by oscillatory integrals themselves, and they cannot always be explicitly computed. We will revisit *Filon-type methods* in section 6.

An entirely different approach was proposed by Levin in [17]. If the indefinite integral is written as $F(x)e^{i\omega g(x)}$, then we immediately have $I = F(b)e^{i\omega g(b)} - F(a)e^{i\omega g(a)}$. It was observed in [17] that $F(x)$ is a smooth function in the absence of stationary points. Moreover, it satisfies the nonoscillatory differential equation

$$F'(x) + i\omega F(x)g'(x) = f(x). \tag{1.3}$$

This system can be solved for $F(x)$ by collocation. The method was generalized in [7, 8] to more general oscillatory functions that satisfy a linear ordinary differential equation; for example, Bessel functions. The accuracy of the methods improves with ω if the boundary points are included in the collocation. Recently, it was shown in [18] that collocating also the derivatives of f in the endpoints can arbitrarily increase the order of accuracy as a function of $1/\omega$. In some cases, the order can also be increased by adding internal points. This *Levin-type method* allows for an accurate evaluation of the integral, without the need for moments. The accuracy is increased simply by solving the differential equation more accurately.

For the particular case of an oscillating factor of the form $\cos(\omega x)$ or $\sin(\omega x)$, specialized quadrature rules using first order derivatives were developed in [16]. An *exponentially fitting quadrature rule* with n points has an error of order ω^{-n}. The weights depend on ω and converge to zero.

The approach taken in this paper achieves a similar high convergence rate as a function of ω. We will show that it solves some of the problems of the other methods

and introduces some peculiarities of its own, thus adding to the spectrum of available approaches that appear to complement each other. For example, we present a case that exhibits a significantly faster convergence rate for increasing ω. The method we describe for approximating (1.1) depends on two simple observations. First, the oscillatory function $e^{i\omega g(x)}$ decays exponentially fast for a complex $g(x)$ along a path with a growing imaginary part. Second, the oscillatory function $e^{i\omega g(x)}$ does *not* oscillate for complex $g(x)$ along a path with fixed real part. These observations are exploited numerically in combination with a corollary to Cauchy's theorem; i.e., the value of a line integral of an analytic function along a path between two points in the complex plane does not depend on the exact path taken (see, e.g., [11]). The same observations also provide the foundation for the steepest descent method [1, 2]. In that method, an asymptotic expansion of the form (1.2) is developed for I. The method was used already by Cauchy and Riemann and developed further by Debye [6]. Methods in the complex plane have since been considered for oscillatory integrals several times in specific applications and for Laplace transforms (see, e.g., [21, 4, 3]). We will present a rather general implementation of the steepest descent method that is also valid for small values of ω. We prove convergence estimates of the numerical scheme as a function of the frequency, and we extend the method to functions f and g that are not analytic. A quadrature rule is proposed that has the same order of accuracy as the Filon-type method. The implementation is entirely numerical; hence we shall refer to the method as the *numerical steepest descent method*.

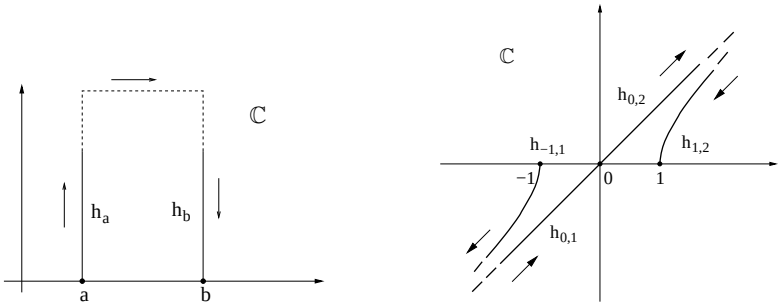
We start this paper in section 2 with some practical and motivating examples that illustrate most of the theory described later. In section 3 we describe and analyze the idealized setting that gives the best possible convergence. It is shown that a suitable n-point quadrature rule in that setting leads to a convergence of $O(\omega^{-2n-1})$. This setting comes with the most restrictions but still covers many important applications. The first requirement is that the functions f and g in (1.1) be analytic in an (infinitely) large region of the complex plane containing the integration interval $[a, b]$. Further, it is assumed that there are no stationary points in $[a, b]$ and that the equation $g(x) = c$ should be "easily solvable." This rather vague description will be made more precise further on. We then proceed by relaxing the requirements one by one until a more generally applicable method is obtained. This increase in generality will, at times, come with a loss in convergence rate. In section 4 we will allow stationary points. We relax the "easy solvability" requirement in section 5. We drop the requirement that f and g should be analytic in sections 6 and 7, respectively. Some final remarks conclude the paper in section 8.

2. Some motivating examples. Consider the following integral, which frequently appears in Fourier analysis applications:

$$\int_a^b f(x)e^{i\omega x}\,\mathrm{d}x. \tag{2.1}$$

This integral has the form of (1.1) with $g(x) = x$. The integrand is highly oscillatory along the real axis if ω is sufficiently large. An important observation is that the function $e^{i\omega x}$ decays rapidly for complex values of x with a positive imaginary part, since $e^{i\omega x} = e^{-\omega \Im x} e^{i\omega \Re x}$. The speed of the decay actually grows as the frequency parameter ω increases. Additionally, the function $e^{i\omega x}$ does not oscillate if the real part of the argument x remains fixed.

Based on these observations, integral (2.1) can be reformulated in such a way that the difficulty—the highly oscillatory nature—is removed. To that end, the integration



(a) $g(x) = x$ (b) $g(x) = x^2$

FIG. 2.1. *Illustration of the integration paths for $g(x) = x$ and $g(x) = x^2$.*

on interval $[a, b]$ is replaced with a path in the complex plane, as illustrated in the left panel of Figure 2.1. The first, vertical part of the path is of the form $z = h_a(p) := a+ip$ for $p \in [0, P]$. The second part is horizontal and connects the points $h_a(P) := a + iP$ to the point $h_b(P) := b + iP$. Finally, the third part connects $h_b(P)$ to b with the vertical path $z = h_b(p)$ for $p \in [0, P]$. Now assume that f is analytic and that f itself does not grow exponentially large in the complex plane. Letting P go to infinity, and using paths parameterized by $h_a(p)$ and $h_b(p)$, for $p \in [0, \infty)$, we can write (2.1) as

$$\int_a^b f(x)e^{i\omega x}\,\mathrm{d}x = e^{i\omega a}\int_0^\infty f(a+ip)e^{-\omega p}\,\mathrm{d}p - e^{i\omega b}\int_0^\infty f(b+ip)e^{-\omega p}\,\mathrm{d}p. \tag{2.2}$$

The integral along the path that connects the endpoints of $h_a(P)$ and $h_b(P)$ vanishes for $P = \infty$ and can therefore be discarded. Both integrals in the right-hand side of (2.2) are well behaved. They can be evaluated efficiently by standard numerical integration techniques, e.g., by Gauss–Laguerre integration [5]. It can be expected from (2.2) that the accuracy of any numerical integration scheme will increase with increasing ω, thanks to the faster decay of the integrand. This expectation will be confirmed both theoretically and numerically in the subsequent sections. One also sees that, asymptotically, the behavior of f around $x = a$ and $x = b$ completely determines the value of (2.1).

Next, we consider the function $g(x) = x^2$ and the corresponding integral

$$\int_{-1}^{1} f(x)e^{i\omega x^2}\,\mathrm{d}x. \tag{2.3}$$

Again, we can remove the integration difficulty by a careful selection of an integration path in the complex plane. The path is drawn in the right panel of Figure 2.1. The following notation is used for the parameterization: $h_{xj}(p) = (-1)^j\sqrt{x^2 + ip}$. Integrating along any such path for $p \in [0, \infty)$ leads to an integrand with the desired decay properties, since $e^{i\omega h_{xj}(p)^2} = e^{i\omega x^2}e^{-\omega p}$. One can see that, for general g, a similar result is obtained if the path satisfies $g(h_x(p)) = g(x) + ip$. This path can be found by using the inverse of g, if it exists, i.e., $h_x(p) = g^{-1}(g(x) + ip)$. Returning to the example function $g(x) = x^2$, however, we note that the inverse of $y = g(x)$ is multivalued: we have $x = -\sqrt{y}$ corresponding to the restriction $g_1 := g|_{[-1,0]}$, and

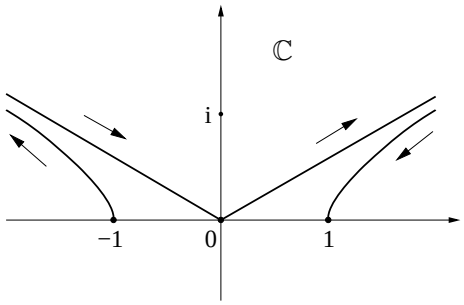


FIG. 2.2. *Illustration of the integration path for* $g(x) = x^3$.

$x = \sqrt{y}$ corresponding to $g_2 := g|_{[0,1]}$. The paths leaving -1 and arriving at 1 are uniquely determined by the requirement that $h_{xj}(0) = x$. Hence,

$$h_{-1,1}(p) = -\sqrt{1+ip} \quad \text{and} \quad h_{1,2}(p) = \sqrt{1+ip}.$$

Contrary to the first example, the integral along the path that connects the limiting endpoints of $h_{-1,1}(p)$ and $h_{1,2}(p)$ cannot be discarded. Since $h_{-1,1}(p)$ and $h_{1,2}(p)$ have opposite signs, any connecting path should cross the real axis. Additionally we require the connecting path to be such that the integrand along the path is nonoscillatory. The solution is to pass explicitly through the point $x = 0$ via two new paths

$$h_{0,1}(p) = -\sqrt{ip} \quad \text{and} \quad h_{0,2}(p) = \sqrt{ip}.$$

The point $x = 0$ is such that the paths corresponding to the two inverses coincide at $x = 0$. We can now rewrite (2.3) as

$$\begin{aligned}\int_{-1}^{1} f(x)e^{i\omega x^2}\mathrm{d}x =& e^{i\omega}\int_0^\infty f(h_{-1,1}(p))e^{-\omega p}h'_{-1,1}(p)\mathrm{d}p - \int_0^\infty f(h_{0,1}(p))e^{-\omega p}h'_{0,1}(p)\mathrm{d}p \\ &+ \int_0^\infty f(h_{0,2}(p))e^{-\omega p}h'_{0,2}(p)\mathrm{d}p - e^{i\omega}\int_0^\infty f(h_{1,2}(p))e^{-\omega p}h'_{1,2}(p)\mathrm{d}p.\end{aligned}$$

These four integrals are well behaved, although the derivatives $h'_{0,1}(p)$ and $h'_{0,2}(p)$ introduce a weak singularity of the form $1/\sqrt{p}$ for $p \to 0$. The integrands do not oscillate, and their decay is exponentially fast.

Note that $\xi = 0$ is a stationary point because $g'(\xi) = 0$. More general stationary points, where also higher order derivatives of g vanish, are handled in a similar way. Consider, e.g., $g(x) = x^3$ and its inverse $g^{-1}(y) = \sqrt[3]{y}$. The cubic root has three branches in the complex plane, and the optimal path $h_x(p) = g^{-1}(g(x) + ip)$ at the point x is found by taking the branch corresponding to the inverse of g that is valid at x, i.e., for which $h_x(0) = x$. At $\xi = 0$, we have that $g'(\xi) = g''(\xi) = 0$ and the three branches coincide. For this example, integral (1.1) can again be decomposed into four contributions, each of which corresponds to a nonoscillating integral. The integration path is drawn in Figure 2.2.

3. The ideal case: Analytic integrand and no stationary points.

3.1. An approximate decomposition of the oscillatory integral. The ideal setting for our approach has three conditions: both f and g are analytic functions, there are no stationary points in the integration interval $[a, b]$ (i.e., $g'(x) \neq 0$),

and the equation $g(x) = z$ is easily solvable, preferably by analytical means. None of these conditions is crucial in order to obtain a convergent quadrature method, as we will relax all conditions later on. But, the ideal case leads to the highest convergence rate among all cases described and is most suited to demonstrate our approach: the problem of evaluating (1.1) can be transformed into the problem of integrating two integrals on $[0, \infty)$ with a smooth integrand that does not oscillate and that decays exponentially fast. This will be proved in this section in Theorem 3.3. First, we give a basic lemma for the approximation of an integral with an integrand that becomes small in some region S of the complex plane.

LEMMA 3.1. *Assume u is analytic in a simply connected complex region $D \subset \mathbb{C}$ with $[a, b] \subset D$, and there exists a bounded and connected region $S \subset D$ such that $|u(z)| \leq \epsilon \ \forall z \in S$. If the shortest distance between any two points p and q of S along a curve that lies in S can be bounded from above by a constant $M > 0$, then there exists a function $F(x)$, $x \in [a, b]$, such that the integral of u can be approximated by*

$$\int_a^x u(z)\,\mathrm{d}z \approx F(a) - F(x) \tag{3.1}$$

with an error e that satisfies $|e| \leq M\epsilon$. The function F is of the form

$$F(x) = \int_{\Gamma_x} u(z)\,\mathrm{d}z \tag{3.2}$$

with Γ_x any path in D that starts at x and ends in S.

Proof. Let Γ_x be a curve in D from x to an arbitrary point in S, denoted by $q(x)$, and Γ_a a curve in D from a to $q(a) \in S$. Choose κ as the shortest path in S that connects $q(a)$ and $q(x)$. Since u is analytic in D, the integration path between a and x may be chosen as the concatenation of Γ_a, κ, and $-\Gamma_x$. The integral is written as

$$\int_a^x u(z)\,\mathrm{d}z = F(a) + \int_\kappa u(z)\,\mathrm{d}z - F(x) \quad \text{with} \quad \left| \int_\kappa u(z)\,\mathrm{d}z \right| \leq M\epsilon.$$

This proves the result. ∎

Note that F is not completely determined by the conditions of this lemma. In particular, the endpoint $q(x)$ of Γ_x may be an arbitrary function of x.

If g is analytic, then the oscillating function $e^{i\omega g(x)}$ in the integrand of (1.1) is also analytic as a function of x. This function is small in absolute value if

$$|e^{i\omega g(x)}| \leq \epsilon \iff e^{-\omega \Im g(x)} \leq \epsilon \iff \Im g(x) \geq \frac{-\log(\epsilon)}{\omega}.$$

Hence, if the inverse of g exists, we can find a suitable region S that is required for Lemma 3.1 with points given by $g^{-1}(c + id)$ for $d \geq d_0 := \frac{-\log(\epsilon)}{\omega}$. Note that, in general, the inverse of an analytic function may be multivalued. Each single-valued branch of the inverse has branch points that are located at the points ξ where $g'(\xi) = 0$, and it is discontinuous across branch cuts that extend from one branch point to another, or from a branch point to infinity. By explicitly excluding the presence of branch points locally, a single-valued branch of the inverse can be found that is analytic in a neighborhood of $[a, b]$. We can then characterize the error of the decomposition given in Lemma 3.1 for the particular case of integral (1.1) as a function of ω.

THEOREM 3.2. *Assume f and g are analytic in a bounded and open complex neighborhood D of $[a,b]$, and assume $g'(z) \neq 0$, $z \in D$. Then there exists an approximation of the form* (3.1) *for* (1.1)*, with an error that has order $O(e^{-\omega d_0})$ as a function of ω for a real constant $d_0 > 0$.*

Proof. Define $S := \{z : \Im g(z) \geq d_0\} \cap D$ with $d_0 > 0$. A positive constant d_0 can always be found such that S is nonempty because g is analytic. In order to prove this, consider a point $x \in [a,b]$. Since g is analytic at x, the equation $g(z) = g(x) + id_0$ always has a solution z for sufficiently small $d_0 > 0$ [11]. Additionally, d_0 can be chosen small enough such that $z \in D$, because D contains an open neighborhood of x. The necessary geometrical conditions on S required by Lemma 3.1 follow from the continuity properties of g. We have

$$\forall x \in S : |f(x)e^{i\omega g(x)}| \leq |f(x)|e^{-\omega d_0}.$$

Since S is finite (because D is bounded), there exists a constant $C > 0$ such that $|f(x)| \leq C$, $x \in S$. The result is established by Lemma 3.1 with $u(x) = f(x)e^{i\omega g(x)}$ and $\epsilon = Ce^{-\omega d_0}$. ∎

Theorem 3.2 shows that the error in the approximation $I \approx F(a) - F(b)$ for (1.1) decays exponentially fast as the frequency parameter ω increases. It requires only that f and g are analytic in a finite neighborhood of $[a,b]$. The function F is given by an integral along a curve that originates in x, and it leads to a point z such that $g(z)$ has a positive imaginary part. The result follows from the exponential decay of the integrand, which is the first of the two observations about the integrand made in the introduction.

3.2. An exact decomposition of the oscillatory integral. Next, we will take the second observation into account: $e^{i\omega g(x)}$ does not oscillate along a path where $g(x)$ has a fixed real part. This will lead to a particularly useful choice for the path Γ_x in definition (3.2) of F.

Let $h_x(p)$ be a parameterization for Γ_x, $p \in [0,P]$; then we find a suitable path as the solution to

$$g(h_x(p)) = g(x) + ip, \quad x \in [a,b].$$

If the inverse of g exists, we have the unique solution $h_x(p) = g^{-1}(g(x) + ip)$. The path $h_x(p)$ is also called the path of steepest descent [1, 2]. This can be understood as follows. Define $k(x,y) := ig(z) = u(x,y) + iv(x,y)$, with $z = x + iy$. Then we have $e^{i\omega g(z)} = e^{\omega k(x,y)}$. It can be shown that the path is such that $v(x,y) = v(x_0,y_0)$ is constant and that the descent of $u(x,y)$ is maximal. In particular, the direction of steepest descent coincides with $-\nabla u$ at each point $z = x + iy$.

Using this path in the definition of F, the decomposition for (1.1) becomes

$$\begin{aligned}\int_a^x f(z)e^{i\omega g(z)}\,\mathrm{d}z &\approx F(a) - F(x)\\ &= e^{i\omega g(a)}\int_0^P f(h_a(p))e^{-\omega p}h_a'(p)\,\mathrm{d}p - e^{i\omega g(x)}\int_0^P f(h_x(p))e^{-\omega p}h_x'(p)\,\mathrm{d}p.\end{aligned}$$

The integrands in the right-hand side do not oscillate, and they decay exponentially fast as the integration parameter p or the frequency parameter ω increases.

In the following theorem, we will consider the limit case $P \to \infty$ in which the error of the approximation vanishes. This will require stronger analyticity conditions for

both f and g. Additionally, the function f can no longer be assumed to be bounded. The result of the theorem will hold if f does not grow faster than polynomially in the complex plane along the suggested integration path.

THEOREM 3.3. *Assume that the functions f and g are analytic in a simply connected and sufficiently (infinitely) large complex region D containing the interval $[a,b]$, and that the inverse of g exists on D. If the following conditions hold in D:*

$$\exists m \in \mathbb{N} : |f(z)| = O(|z|^m), \tag{3.3}$$

$$\exists \omega_0 \in \mathbb{R} : |g^{-1}(z)| = O(e^{\omega_0 |z|}), \quad |z| \to \infty, \tag{3.4}$$

then there exists a function $F(x)$ for $x \in [a,b]$ such that

$$\int_a^x f(z) e^{i\omega g(z)} \, \mathrm{d}z = F(a) - F(x) \qquad \forall \omega > (m+1)\omega_0, \tag{3.5}$$

where $F(x)$ is of the form

$$F(x) := \int_{\Gamma_x} f(z) e^{i\omega g(z)} \, \mathrm{d}z, \tag{3.6}$$

with Γ_x a path that starts at x. A parameterization $h_x(p)$, $p \in [0,\infty)$, for Γ_x exists such that the integrand of (3.6) *is $O(e^{-\omega p})$.*

Proof. We will use $u(z)$ to denote the integrand of (1.1). Using the fact that $|u(z)| = |f(z)e^{i\omega g(z)}| = |f(z)|e^{-\omega \Im g(z)}$, and conditions (3.3) and (3.4), we can state

$$c + id \in D \Rightarrow |u(g^{-1}(c+id))| = O(e^{(m\omega_0 - \omega)d}), \quad d \to \infty. \tag{3.7}$$

If $\omega > m\omega_0$, then (3.7) characterizes the exponential decay of the integrand in the complex plane. We will now choose an integration path from the point a to the region where the integrand becomes small and from that region back to the point $x \in [a,b]$. We will show that the contribution along the line that connects both paths can be discarded. This will establish the existence of Γ_a and Γ_x in (3.6) and the independence of Γ_a and Γ_x.

Assume an integration path for I that consists of three connected parts, parameterized as $h_a(p)$ and $h_x(p)$ with $p \in [0,P]$, and $\kappa(p)$ with $p \in [a,x]$. The parameterizations can be chosen differentiable and satisfy $h_a(0) = a$, $h_x(0) = x$, $h_a(P) = \kappa(a)$, and $h_x(P) = \kappa(x)$. We have

$$\begin{aligned} \int_a^x u(z) \, \mathrm{d}z = &\int_0^P u(h_a(p)) h_a'(p) \, \mathrm{d}p + \int_a^x u(\kappa(p)) \kappa'(p) \, \mathrm{d}p \\ &- \int_0^P u(h_x(p)) h_x'(p) \, \mathrm{d}p. \end{aligned} \tag{3.8}$$

Since the inverse function g^{-1} exists, we can choose the points $h_a(P)$ and $h_x(P)$ as follows: $h_a(P) = g^{-1}(g(a) + iP)$ and $h_x(P) = g^{-1}(g(x) + iP)$. Hence, by (3.7),

$$|u(h_a(P))| = O(e^{(m\omega_0 - \omega)P}) \quad \text{and} \quad |u(h_x(P))| = O(e^{(m\omega_0 - \omega)P}).$$

We will now show that, as $P \to \infty$, the second integral vanishes. Equation (3.8) is then of the form (3.5), with Γ_a and Γ_x parameterized by $h_a(p)$ and $h_x(p)$, respectively, with $p \in [0,\infty)$. The contribution of the integral along $\kappa(p)$ is bounded by

$$\left| \int_a^x u(\kappa(p)) \kappa'(p) \, \mathrm{d}p \right| \leq \max_{p \in [a,x]} |u(\kappa(p))| \max_{p \in [a,x]} |\kappa'(p)| \, |x - a|. \tag{3.9}$$

Defining the path $\kappa(p) := g^{-1}(g(p)+iP)$, we have by (3.7) that $|u(\kappa(p))| = O(e^{(m\omega_0-\omega)P})$, $p \in [a, x]$. We can write the second factor in the bound (3.9) as

$$\kappa'(p) = \frac{\partial g^{-1}}{\partial y}(g(p) + iP)\frac{\mathrm{d}g}{\mathrm{d}p}(p).$$

The derivative of $g(p)$ with respect to p is bounded on $[a, b]$ because g is analytic. The factor $\frac{\partial g^{-1}}{\partial y}(g(p) + iP)$ is bounded by $O(e^{\omega_0 P})$. Combining the asymptotic behavior of the factors in (3.9), the second term in (3.8) vanishes for $P \to \infty$ and $\forall\ x \in [a, b]$ if $\omega > (m + 1)\omega_0$. This proves the result. ∎

Remark 3.4. Note that f and g should be analytic in a simply connected region D that contains the paths h_a, h_b, and $\kappa(p)$ in order to apply Cauchy's theorem. The unique existence of the inverse of g is a necessary condition: if $g'(z) = 0$ with $z \in D$, then the point z is a branch point of the inverse function. The path $\kappa(p)$ may cross the branch cut that originates at z, and Cauchy's theorem cannot be applied.

Remark 3.5. Conditions (3.3) and (3.4) are sufficient but not necessary. For example, the limit case also applies when $f(x) = e^x$ and $g(x) = x$. If, however, $f(x) = e^{-x^2}$ and $g(x) = x$, the integrand always diverges at infinity along the steepest descent path, regardless of the size of ω. In that case, the path should be truncated at a finite distance from the real axis. The accuracy of the decomposition is then described by Theorem 3.2; i.e., the error decays exponentially fast.

3.3. Evaluation of $F(x)$ by Gauss–Laguerre quadrature. Next, we consider the evaluation of $F(x)$ as defined by (3.6). The parameterization of the path $h_x(p)$ is such that it solves the equation

$$g(h_x(p)) = g(x) + ip. \tag{3.10}$$

The integrand of (1.1) along this path is nonoscillatory and exponentially decaying,

$$f(h_x(p))e^{i\omega g(h_x(p))} = f(h_x(p))e^{i\omega g(x)}e^{-\omega p}.$$

In the simplest, yet important, case $g(x) := x$ the suggested path is $h_x(p) = x + ip$.

An efficient approach for infinite integrals with an exponentially decaying integrand is Gauss–Laguerre quadrature [5]. Laguerre polynomials are orthogonal with respect to e^{-x} on $[0, \infty]$. A Gauss–Laguerre rule with n points is exact for polynomials up to degree $2n - 1$. The integral $F(x)$ with the suggested path can be written as

$$\begin{aligned} F(x) &= \int_0^\infty f(h_x(p))e^{i\omega(g(x)+ip)}h_x'(p)\,\mathrm{d}p = e^{i\omega g(x)}\int_0^\infty f(h_x(p))h_x'(p)e^{-\omega p}\,\mathrm{d}p \\ &= \frac{e^{i\omega g(x)}}{\omega}\int_0^\infty f\left(h_x\left(\frac{q}{\omega}\right)\right)h_x'\left(\frac{q}{\omega}\right)e^{-q}\,\mathrm{d}q, \end{aligned}$$

with $q = \omega p$ in the last expression. Applying a Gauss–Laguerre quadrature rule with n points x_i and weights w_i yields a quadrature rule

$$F(x) \approx Q_F[f, g, h_x] := \frac{e^{i\omega g(x)}}{\omega}\sum_{i=1}^{n} w_i f\left(h_x\left(\frac{x_i}{\omega}\right)\right)h_x'\left(\frac{x_i}{\omega}\right). \tag{3.11}$$

The rule requires the evaluation of f in a complex neighborhood of x.

THEOREM 3.6. *Assume functions f and g satisfy the conditions of Theorem* 3.3. *Let I be approximated by the quadrature formula*

$$I \approx Q[f,g] := Q_F[f,g,h_a] - Q_F[f,g,h_b], \tag{3.12}$$

where Q_F is evaluated by an n-point Gauss–Laguerre quadrature rule as specified in (3.11). *Then the quadrature error behaves asymptotically as $O(\omega^{-2n-1})$.*

Proof. A formula for the error of the n-point Gauss–Laguerre quadrature rule applied to the integral $\int_0^\infty f(x)e^{-x}\mathrm{d}x$ is given by [5]

$$E = \frac{(n!)^2}{(2n)!} f^{(2n)}(\zeta), \quad \zeta \in [0,\infty).$$

Using this formula, one can derive an expression for the error $E := F(a) - Q_F[f,g,h_a]$:

$$\begin{aligned} E &= \frac{e^{i\omega g(a)}}{\omega} \frac{(n!)^2}{(2n)!} \left. \frac{\mathrm{d}^{2n}(f(h_a(q/\omega))h_a'(q/\omega))}{\mathrm{d}q^{2n}} \right|_{q=\zeta} \\ &= \frac{e^{i\omega g(a)}}{\omega^{2n+1}} \frac{(n!)^2}{(2n)!} \left. \frac{\mathrm{d}^{2n}(f(h_a(q))h_a'(q))}{\mathrm{d}q^{2n}} \right|_{q=\zeta/\omega} \end{aligned} \tag{3.13}$$

with $\zeta \in \mathbb{C}$. The error behaves asymptotically as $O(\omega^{-2n-1})$. The absolute error for the approximation to (1.1) is composed of two contributions of the form (3.13), and, hence, has the same high order of convergence. ∎

Remark 3.7. The decomposition $I = F(a) - F(b)$ is of a similar type as the decomposition of I in [14] based on asymptotic expansions. There, the terms in the expansions are given by a combination of f, g, and their derivatives, evaluated in the points a and b. Yet, the numerical properties of our approach are different: the convergence rate $O(\omega^{-2n-1})$ when using an n-point quadrature rule for both $Q_F[f,g,h_a]$ and $Q_F[f,g,h_b]$ should be compared to the rate $O(\omega^{-n-1})$ when using an n-term asymptotic expansion of I evaluated in a and b.

Example 3.8. We end this section with a numerical example to illustrate the sharpness of our convergence result. The absolute error for different values of ω and of n is given in Table 3.1 for the functions $g(x) = x$ and $f(x) = 1/(1+x)$ on $[0,1]$. The parameterization for Γ_x is given by $h_x(p) = g(x) + ip$. The behavior as a function of ω follows the theory until machine precision is reached. The relative error scales only slightly worse, since $|I| = O(\omega^{-1})$.

One should note that decomposition (3.5) is exact for all positive values of the parameter $\omega > (m+1)\omega_0 > 0$. The conditions from Theorem 3.3 yield the minimal frequency parameter $(m+1)\omega_0$. The method itself is therefore not asymptotic—only

TABLE 3.1

Absolute error of the approximation of I by $Q_F[f,g,h_a] - Q_F[f,g,h_b]$ with n quadrature points for the functions $f(x) = 1/(1+x)$ and $g(x) = x$ on $[0,1]$. The last row shows the value of $\log_2(e_{40}/e_{80})$: this value should approximate $2n+1$.

$\omega \setminus n$	1	2	3	4	5
10	$1.0E-3$	$3.1E-5$	$1.9E-6$	$1.7E-7$	$2.1E-8$
20	$1.2E-4$	$1.1E-6$	$2.3E-8$	$7.5E-10$	$3.2E-11$
40	$1.7E-5$	$3.9E-8$	$2.1E-10$	$2.0E-12$	$2.8E-14$
80	$2.0E-6$	$1.2E-9$	$1.7E-12$	$4.2E-15$	$1.6E-17$
Rate	3.1	5.0	6.9	8.9	10.8

the convergence estimate is. Table 3.1 shows an absolute error of $2.1E-8$ (relative error $1.4E-7$) for $\omega = 10$ with a number of quadrature points as small as $n = 5$. The corresponding integral is not highly oscillatory at all. In order to achieve the same absolute error with standard Gaussian quadrature on $[0, 1]$, we had to choose a rule with 10 points. Considering the fact that we evaluate both $Q_F[f, g, h_a]$ and $Q_F[f, g, h_b]$ with $n = 5$ points, the amount of work is the same. Thus, even at relatively low frequencies, our approach is competitive with conventional quadrature on the real axis. For higher frequencies, obviously, the new approach may be many orders of magnitude faster.

4. The case of stationary points.

4.1. A new decomposition for the oscillatory integral. At a stationary point ξ, the derivative of g vanishes and the integrand $f(x)e^{i\omega g(x)}$ does not oscillate, at least locally. The contribution of the integrand and its derivatives at ξ can therefore not be neglected. The theorems of section 3 do not apply, because the inverse of g does not exist uniquely due to the branch point at ξ.

In order to illustrate the problem, consider the following situation. Assume that the equation $g'(x) = 0$ has one solution ξ and $\xi \in [a, b]$. Now define the restrictions

$$g_1 := g|_{[a,\xi]} \quad \text{and} \quad g_2 := g|_{[\xi,b]} \tag{4.1}$$

of g on the intervals $[a, \xi]$ and $[\xi, b]$, respectively. Then the unique inverse of g on $[a, b]$ does not exist, but a single-valued branch g_1^{-1} can be found that satisfies $g_1^{-1}(g_1(x)) = x$, $x \in [a, \xi]$. This branch is analytic everywhere except at the point ξ and along a branch cut that can be chosen arbitrarily but that always originates at ξ. Similarly, a single-valued branch g_2^{-1} exists that satisfies $g_2^{-1}(g_2(x)) = x$, $x \in [\xi, b]$. Both branches satisfy $g(g_i^{-1}(z)) = z$, $i = 1, 2$, in their domain of analyticity. The integrand is small in the region S_1 with points of the form $g_1^{-1}(c+id)$, $d \geq d_0$, or in the region S_2 with points of the form $g_2^{-1}(c + id)$, $d \geq d_0$. It is easy to see that S_1 and S_2 are not connected: applying g on both sides of the equality $g_1^{-1}(y) = g_2^{-1}(z)$ leads to $y = z$, which is only possible if $z = \xi \notin S_1, S_2$. The path (3.10) that solves $g(h_x(p)) = g(x) + ip$, as suggested in section 3, leads to a path in S_1 for a and to a path in S_2 for b.

The solution is therefore to split the integration interval into the two subintervals $[a, \xi]$ and $[\xi, b]$. This procedure can be repeated for any number of stationary points. The analogues of Theorems 3.2 and 3.3 can be stated as follows.

THEOREM 4.1. *Assume that the functions f and g are analytic in a bounded and open complex neighborhood D of $[a, b]$. If the equation $g'(x) = 0$ has only one solution ξ in D and $\xi \in (a, b)$, then there exist functions $F_j(x)$, $j = 1, 2$, such that*

$$\int_s^t f(z)e^{i\omega g(z)}\,\mathrm{d}z = F_1(s) - F_1(\xi) + F_2(\xi) - F_2(t) + O(e^{-\omega d_0}), \quad d_0 > 0,$$

for $s \in [a, \xi]$ and $t \in [\xi, b]$, where $F_j(x)$ is of the form

$$F_j(x) := \int_{\Gamma_{x,j}} f(z)e^{i\omega g(z)}\,\mathrm{d}z \tag{4.2}$$

with $\Gamma_{x,j}$ a path that starts at x.

Proof. Define $g_2(x)$ as in (4.1). A decomposition for $\int_\xi^t f(x)e^{i\omega g_2(x)}\,\mathrm{d}x$ can be found using the proof of Theorem 3.2 with two modifications. First, the equation $g(z) = g(x) + id_0$ now has at least two solutions locally around $x = \xi$. We choose

the solution that corresponds to the single-valued branch g_2^{-1} of the inverse of g that satisfies $g_2^{-1}(g(x)) = x$, $x \in [\xi, b]$. The branch cut can always be chosen such that it does not prevent us from applying Cauchy's theorem. Second, the set S in the proof is now defined as $S := \{z : \Im g(z) \geq d_0 \text{ and } g_2^{-1}(g(z)) = z\} \cap D$; i.e., the set is restricted to one connected part of D, where the integrand is small, as opposed to the set of all points, where the integrand is small. The latter set would not be connected in this case. With these modifications, the proof shows the existence of F_2 such that

$$\int_\xi^t f(z)e^{i\omega g_2(z)}\,\mathrm{d}z = F_2(\xi) - F_2(t) + O(e^{-\omega d_0}).$$

The same reasoning can be applied in order to find a decomposition on the interval $[a, \xi]$. This leads to the result. ∎

The next theorem is the limit case of Theorem 4.1, where the error vanishes. The notation g_1^{-1} denotes a branch of the multivalued inverse of g that satisfies $g_1^{-1}(g_1(x)) = x$, $x \in [a, \xi]$. The notation g_2^{-1} is similar.

THEOREM 4.2. *Assume that the functions f and g are analytic in a simply connected and sufficiently (infinitely) large complex region D containing the interval $[a, b]$. Assume further that the equation $g'(x) = 0$ has only one solution ξ in D and $\xi \in (a, b)$. Define g_1 and g_2 as in* (4.1)*. If the following conditions hold:*

$$\begin{aligned}&\exists m \in \mathbb{N} : |f(z)| = O(|z|^m),\\ &\exists \omega_0 \in \mathbb{R} : |g_1^{-1}(z)| = O(e^{\omega_0|z|}) \text{ and } |g_2^{-1}(z)| = O(e^{\omega_0|z|}), \quad |z| \to \infty,\end{aligned}$$

then there exist functions $F_j(x)$, $j = 1, 2$, of the form (4.2) *such that*

$$\int_s^t f(z)e^{i\omega g(z)}\,\mathrm{d}z = F_1(s) - F_1(\xi) + F_2(\xi) - F_2(t) \quad \forall \omega > (m+1)\omega_0 \tag{4.3}$$

for $s \in [a, \xi]$ and $t \in [\xi, b]$. A parameterization $h_{\xi,j}(p)$, $p \in [0, \infty)$, for $\Gamma_{x,j}$ exists such that the integrand of (4.2) *is $O(e^{-\omega p})$.*

Theorems 4.1 and 4.2 are easily extended to the case when $\xi = a$ (or $\xi = b$) by discarding the two terms $F_1(a) - F_1(\xi)$ (or $F_2(\xi) - F_2(b)$).

Example 4.3. We consider the function $g(x) = (x - 1/2)^2$ with a stationary point at $\xi = 1/2$. The inverse of g, i.e., $g^{-1}(y) = 1/2 \pm \sqrt{y}$, is a two-valued function. One branch is valid on the interval $[0, \xi]$, the other on $[\xi, 1]$. The paths suggested by (3.10) on $[0, \xi]$ that originate at the points 0 and ξ, respectively, are given by

$$h_{0,1}(p) = 1/2 - \sqrt{1/4 + ip} \quad \text{and} \quad h_{\xi,1}(p) = 1/2 - \sqrt{ip}.$$

The paths on $[\xi, 1]$ for the points $1/2$ and 1 are parameterized by

$$h_{\xi,2}(p) = 1/2 + \sqrt{ip} \quad \text{and} \quad h_{1,2}(p) = 1/2 + \sqrt{1/4 + ip}.$$

These paths correspond to the two inverse functions. We have found the decomposition $I = F_1(a) - F_1(\xi) + F_2(\xi) - F_2(b)$.

Note that the paths $h_{\xi,1}$ and $h_{\xi,2}$ that originate in the point ξ introduce a numerical problem. Their derivatives, which appear in the integrand of the line integral, behave like $1/\sqrt{p}$, $p \to 0$, at ξ. This weak singularity is integrable but prevents convergence of the Gauss–Laguerre quadrature rules. We will require a new method to evaluate $F_j(\xi)$.

4.2. The evaluation of $F_j(x)$ by generalized Gauss–Laguerre quadrature. The previous example showed a numerical problem for the evaluation of $F_j(x)$ by numerical quadrature: the integrand of $F_j(\xi)$ along the path suggested by (3.10) becomes weakly singular at the stationary point ξ. A similar singularity occurs if higher order derivatives of $g(\xi)$ also vanish. Assume that $g^{(k)}(\xi) = 0$, $k = 1, \ldots, r$. The Taylor expansion of g is then

$$g(x) = g(\xi) + 0 + \cdots + 0 + g^{(r+l)}(\xi)\frac{(x-\xi)^{r+1}}{(r+1)!} + O((x-\xi)^{l+2}).$$

The path $h_{\xi,j}(p)$ solves the equation $g(h_{\xi,j}(p)) = g(\xi) + ip$. Its behavior at $p = 0$ is

$$h_{\xi,j}(p) \sim \xi + \sqrt[r+1]{\frac{(r+1)!\,p}{g^{(r+1)}(\xi)}}\,i. \tag{4.4}$$

The derivative has a singularity of the form $p^{\frac{1}{r+1}-1}$, $p \to 0$.

Fortunately, these types of singularities can be handled efficiently by generalized Gauss–Laguerre quadrature. Generalized Laguerre polynomials are orthogonal with respect to the weight function $x^\alpha e^{-x}$, $\alpha > -1$ [5]. Function $F_j(\xi)$ with optimal path $h_{\xi,j}(p)$ is given by

$$F_j(\xi) = \frac{e^{i\omega g(\xi)}}{\omega}\int_0^\infty f\left(h_{\xi,j}\left(\frac{q}{\omega}\right)\right)h'_{\xi,j}\left(\frac{q}{\omega}\right)e^{-q}\,\mathrm{d}q. \tag{4.5}$$

Generalized Gauss–Laguerre quadrature will be used with n points x_i and weights w_i that depend on the value of $\alpha = 1/(r+1) - 1 = -r/(r+1)$. The function $F_j(x)$ is then approximated by

$$Q_F^\alpha[f, g, h_{\xi,j}] := \frac{e^{i\omega g(\xi)}}{\omega}\sum_{i=1}^{n} w_i\, f\left(h_{\xi,j}\left(\frac{x_i}{\omega}\right)\right)h'_{\xi,j}\left(\frac{x_i}{\omega}\right)x_i^{-\alpha}. \tag{4.6}$$

This expression is similar to (3.11) but includes the factor $x_i^{-\alpha}$ to regularize the singularity.

THEOREM 4.4. *Assume functions f and g satisfy the conditions of Theorem* 4.2. *Assume that $g^{(k)}(\xi) = 0$, $k = 1, \ldots, r$, and $g^{(r+1)}(\xi) \neq 0$. Let the function $F_j(\xi)$ be approximated by the quadrature formula*

$$F_j(\xi) \approx Q_F^\alpha[f, g, h_{\xi,j}]$$

with $\alpha = -r/(r+1)$. Then the quadrature error has order $O(\omega^{-2n-1-\alpha})$.

Proof. The error formula for an n-point generalized Gauss–Laguerre quadrature rule is

$$\frac{n!\Gamma(n+\alpha+1)}{(2n)!}f^{(2n)}(\zeta), \quad 0 < \zeta < \infty. \tag{4.7}$$

We can repeat the arguments of the proof of Theorem 3.6. An expression for the error $e := F_j(\xi) - Q_F^\alpha[f, g, h_{\xi,j}]$ can be derived by using (4.7). This leads to

$$\begin{aligned} e &= \frac{e^{i\omega g(\xi)}}{\omega}\frac{n!\Gamma(n+\alpha+1)}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_{\xi,j}(q/\omega))h'_{\xi,j}(q/\omega)q^{-\alpha})}{\mathrm{d}q^{2n}}\right|_{q=\zeta} \\ &= \frac{e^{i\omega g(\xi)}}{\omega^{2n+1}}\frac{n!\Gamma(n+\alpha+1)}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_{\xi,j}(q))h'_{\xi,j}(q)(\omega q)^{-\alpha})}{\mathrm{d}q^{2n}}\right|_{q=\zeta/\omega} \end{aligned}$$

with $\zeta \in \mathbb{C}$. Hence, the error is asymptotically of the order $O(\omega^{-2n-1-\alpha})$. ∎

Remark 4.5. Generalized Gauss–Laguerre quadrature converges rapidly only if the function $v(x)$ in an integrand of the form $v(x)x^\alpha e^{-x}$ has polynomial behavior. Depending on f, the function $f(h_{\xi,j}(p))$ may not resemble a polynomial very well, due to the root in (4.4) for small p. An alternative to generalized Gauss–Laguerre quadrature with $\alpha = -1/2$ is to remove the singularity by the transformation $u = \sqrt{p}$ or $p = u^2$. The same transformation also removes the square root behavior of $h_{\xi,j}(p)$. The integrand after the transformation decays like e^{-u^2}. In that case, variants of the classical Hermite polynomials that are orthogonal with respect to e^{-u^2} on the half-range interval $[0, \infty)$ can be used with corresponding Gaussian quadrature rules as constructed by Gautschi [10]. A similar convergence analysis yields the order $O(\omega^{-n-1})$ in this case.

We can now characterize the approximation of (1.1) in the presence of several stationary points.

THEOREM 4.6. *Assume that f and g are analytic in a sufficiently large region $D \subset \mathbb{C}$ and that the equation $g'(x) = 0$ has l solutions $\xi_i \in (a, b)$. Define $r_i := (\min_{k>1} g^{(k)}(\xi_i) \neq 0) - 1$ and $r := \max_i r_i$. If the conditions of Theorem* 4.2 *are satisfied on each subinterval $[\xi_i, \xi_{i+1}]$, and on $[a, \xi_1]$ and $[\xi_r, b]$, then* (1.1) *can be approximated by*

$$
\begin{aligned}
I \approx Q[f,g] := {} & Q_F[f,g,h_{a,0}] - Q_F^{\alpha_1}[f,g,h_{\xi_1,0}] \\
& + \sum_{i=1}^{l-1} \left(Q_F^{\alpha_i}[f,g,h_{\xi_i,i}] - Q_F^{\alpha_{i+1}}[f,g,h_{\xi_{i+1},i}] \right) \\
& + Q_F^{\alpha_l}[f,g,h_{\xi_l,l}] - Q_F[f,g,h_{b,l}]
\end{aligned} \tag{4.8}
$$

with $\alpha_i = -r_i/(r_i + 1)$, with a quadrature error of the order $O(\omega^{-2n-1/(r+1)})$.

Proof. This follows from a repeated application of the decomposition given by Theorem 4.2 and from the approximation of each term $F_i(x)$ by $Q_F^{\alpha_i}[f, g, h_{x,i}]$ as in Theorem 4.4. ∎

Theorem 4.6 can easily be extended to the case when $g'(a) = 0$ or $g'(b) = 0$. If, e.g., $g'(a) = 0$, we can set $\xi_1 = a$ and use the general decomposition (4.8) with the first two terms left out.

Example 4.7. We return to Example 4.3 of this section in order to illustrate the convergence results. The approximation of (1.1) for the function $g(x) = (x - 1/2)^2$ is given by

$$
Q[f,g] = Q_F[f,g,h_{0,1}] - Q_F^{-1/2}[f,g,h_{1/2,1}] + Q_F^{-1/2}[f,g,h_{1/2,2}] - Q_F[f,g,h_{1,2}].
$$

Theorem 4.6 predicts an error of the order $O(\omega^{-2n-1/2})$. The sharpness of this estimate can be verified by the results in Table 4.1.

4.3. The case of complex stationary points. So far, we have required the stationary point $\xi \in [a, b]$ to be real. But even for functions g that are real valued on the real axis, the equation $g'(x) = 0$ may have complex solutions. The value of $g'(x)$ on $[a, b]$ can become very small if a complex stationary point ξ lies close to the real axis. We may therefore expect that such a point contributes to the value of the integral (1.1). Here, we will not pursue the extension of the theory to the case of complex stationary points in any detail. Instead, we will restrict ourselves to a number of remarks that address some of the relevant issues.

A first observation is that Theorem 4.1 can still be applied if the region D is chosen small enough such that it does not contain ξ. This means that the contribution of

TABLE 4.1
Absolute error of the approximation of I by $Q[f,g]$ using (generalized) Gauss–Laguerre quadrature with $f(x) = 1/(1+x)$ and $g(x) = (x-1/2)^2$ on $[0,1]$. The last row shows the value of $\log_2(e_{80}/e_{160})$: this value should approximate $2n+1/2$.

$\omega \setminus n$	1	2	3	4	5
10	$4.7E-3$	$7.1E-4$	$1.7E-4$	$4.9E-5$	$1.7E-5$
20	$7.8E-4$	$5.6E-5$	$7.2E-6$	$1.3E-6$	$2.7E-7$
40	$1.2E-4$	$2.8E-6$	$1.5E-7$	$1.2E-8$	$1.3E-9$
80	$1.6E-5$	$1.0E-7$	$1.7E-9$	$5.0E-11$	$2.1E-12$
160	$2.3E-6$	$3.4E-9$	$1.6E-11$	$1.3E-13$	$1.6E-15$
Rate	2.8	4.9	6.8	8.6	10.4

ξ to the value of I, if any, decays exponentially fast as ω increases. Still, for small values of ω, the error may become prohibitively large if ξ lies close to the real axis.

In order to resolve this problem, one must first know which stationary points can contribute to the error of the approximations of section 4. In general, the question can be answered by inspecting the integration paths. A stationary point contributes if it lies in the interior of the domain bounded by the integration interval on the real axis and the complex integration path (including the limiting connecting part at infinity). In order to obtain an exact decomposition, the integration path should be changed to pass through ξ explicitly. Specifically, the decomposition should include two additional terms for ξ of the form (4.5).

As a final remark, we note that the integral of the form (4.5) has a factor $e^{i\omega g(\xi)}$ with $g(\xi) = c + id$ complex. If $d > 0$, then the contribution decays exponentially as $e^{-\omega d}$. We know from Theorem 4.1 that the error introduced by discarding complex stationary points should decay exponentially. Hence, complex stationary points for which $d \leq 0$ cannot contribute to the value of I.

5. The case when the oscillator is not easily invertible. Theorems 3.3 and 4.2 continue to hold for paths different from the one implicitly defined by (3.10). The value of $F(a)$ does not depend on the path taken, and does not even depend on the limiting endpoint of the path, as long as the imaginary part of $g(x)$ grows infinitely large. We have merely suggested (3.10), which yields a nonoscillatory integrand with exponential decay, as being suitable for Gauss–Laguerre quadrature. Other integration techniques may be applied for other paths with different numerical properties. We will not explore these possibilities in depth here.

We restrict the discussion to an approach that is useful when the inverse function of g is known to exist, but when the suggested path is not easily obtained by analytical means. As ω increases, we see from (3.11) that $Q_F[f,g,h_a]$ requires function values in a complex region around a of diminishing size. Therefore, it is reasonable to assume that approximating the path defined by (3.10) locally around $x = a$ is acceptable. Use of the first order Taylor approximation $g(x) \approx g(a) + g'(a)(x-a)$ to replace the left-hand side of (3.10) leads to the path

$$h_a(p) = a + \frac{ip}{g'(a)}.$$

The second order Taylor approximation leads to the path

$$h_a(p) = a - \frac{g'(a) - \sqrt{g'(a)^2 + 2ipg^{(2)}(a)}}{g^{(2)}(a)}.$$

In the case of stationary points the path can be approximated by using (4.4).

The general expression for the integral along the approximate path is given by

$$F(a) = \int_0^\infty f(h_a(p))e^{i\omega g(h_a(p))}h_a'(p)\,\mathrm{d}p.$$

Computing $F(a)$ by Gauss–Laguerre quadrature yields a numerical approximation with an error given by

$$\begin{aligned} E &= \omega^{-1}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q/\omega))e^{i\omega g(h_a(q/\omega))}h_a'(q/\omega)e^q)}{\mathrm{d}q^{2n}}\right|_{q=\zeta} \\ &= \omega^{-2n-1}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q))h_a'(q)e^{i\omega g(h_a(q))}e^{\omega q})}{\mathrm{d}q^{2n}}\right|_{q=\zeta/\omega}. \end{aligned}$$

The order of convergence is not necessarily $O(\omega^{-2n-1})$ in this case because the derivative still depends on ω. However, the function $e^{i\omega g(h_a(q))}$ is a good approximation to $e^{i\omega g(a)}e^{-\omega q}$, and we can expect the quadrature to converge. This will be illustrated further on.

The results can be improved to preserve the original convergence rate of $O(\omega^{-2n-1})$ at the cost of a little extra work to determine the optimal path. The optimal path depends only on $g(x)$, and on the interval $[a,b]$, and can therefore be reused for different functions f. The extra computations have to be done once for each combination of $g(x)$ and $[a,b]$.

The Taylor approximation of the path can be used to generate suitable starting values for a Newton–Raphson iteration, which is applied to find the root x of the equation

$$g(x) - g(a) - ip = 0. \tag{5.1}$$

For the set of n (fixed) values for p that are required by the quadrature rule, the iteration yields the points $x = h_a(p)$ on the path. The values of $h_a'(p)$, i.e., $\frac{\mathrm{d}x}{\mathrm{d}p}$, are found by taking the derivative of (5.1) with respect to p,

$$g'(x)\frac{\mathrm{d}x}{\mathrm{d}p} = i.$$

With the Newton–Raphson method, the points on the optimal path and the derivatives at these points can be found to high precision. Since the Taylor approximation is already a good approximation for large ω, the required number of iterations is small.

Example 5.1. We consider the second order Taylor approximation of the path for $f(x) = 1/(1+x)$ and $g(x) = (x^2+x+1)^{1/3}$. The absolute error is shown in Table 5.1. Use of the Newton–Raphson iteration for the same example yields an error of order $O(\omega^{-2n-1})$. This is shown in Table 5.2. The number of iterations per quadrature point varied between 1 and 4.

6. Generalization to a nonanalytic function $f(x)$. If $f(x)$ is not analytic in a complex region surrounding $[a,b]$, then the method presented thus far will not work. If $f(x)$ is piecewise analytic (e.g., piecewise polynomial), the integration can be split into the integrals corresponding to the analytic parts of f. More generally, however, we need to resort to another approach. For a suitable analytic function $\tilde{f}$ that approximates f, we can expect the integral

$$\tilde{I} := \int_a^b \tilde{f}(x)e^{i\omega g(x)}\,\mathrm{d}x$$

to approximate the value of I.

TABLE 5.1
Absolute error of approximation of $F(a) - F(b)$ by Gauss–Laguerre quadrature with $f(x) = 1/(1+x)$ and $g(x) = (x^2+x+1)^{1/3}$ on $[0,1]$ and second order Taylor approximation of the optimal path. The last row shows the value of $\log_2(e_{160}/e_{320})$.

$\omega \setminus n$	1	2	3	4	5
20	$1.4E-2$	$2.7E-3$	$7.4E-4$	$2.4E-4$	$8.9E-5$
40	$2.5E-3$	$2.6E-4$	$4.6E-5$	$1.0E-5$	$2.5E-6$
80	$3.8E-4$	$1.8E-5$	$1.7E-6$	$2.0E-7$	$2.9E-8$
160	$5.2E-5$	$1.1E-6$	$4.0E-8$	$2.1E-9$	$1.5E-10$
320	$6.7E-6$	$6.8E-8$	$7.7E-10$	$1.6E-11$	$4.4E-13$
Rate	3.0	4.0	5.7	7.0	8.4

TABLE 5.2
The same example as in Table 5.1, but using Newton–Raphson iterations to compute the optimal path. The number of iterations per quadrature point varied between 1 and 4. The last row shows the value of $\log_2(e_{320}/e_{640})$: this value should approximate $2n+1$.

$\omega \setminus n$	1	2	3	4	5
20	$1.1E-2$	$2.4E-3$	$7.4E-4$	$2.5E-4$	$7.5E-5$
40	$2.1E-3$	$2.4E-4$	$4.4E-5$	$1.0E-5$	$2.4E-6$
80	$3.3E-4$	$1.5E-5$	$1.2E-6$	$1.5E-7$	$2.3E-8$
160	$4.5E-5$	$6.1E-7$	$1.8E-8$	$8.7E-10$	$6.2E-11$
320	$5.9E-6$	$2.1E-8$	$1.8E-10$	$2.7E-12$	$6.2E-14$
640	$7.2E-7$	$6.7E-10$	$1.5E-12$	$6.3E-15$	$4.3E-17$
Rate	3.0	5.0	6.9	8.8	10.5

This leads to a *Filon-type method* that was already mentioned in the introduction. Filon's method was extended by Iserles and Nørsett in [14]. Since the value of I depends on the value of f and its derivatives at $x = a$ and $x = b$, they successfully used Hermite interpolation in the points a and b, in the stationary points, and in a few other regular points in the interval $[a, b]$. In [14, Thm. 2.3] it was shown that interpolating $f^{(j)}(x)$ at a and b, $j = 0, \ldots, s-1$, with a polynomial of degree $2s-1$ leads to a quadrature rule with an error of order $O(\omega^{-s-1})$.

Since polynomials are analytic, we can also use the Hermite approximation in our approach. This enables the computation of the weights of the Filon-type quadrature rule for general oscillators. (Note that the complex approach also enables the computation of the moments in the *asymptotic method* of [14].) It does not improve the convergence rate of the method. We can improve on the Filon-type method, however, in a different way. Thanks to the decomposition of (1.1) as $I = F(a) - F(b)$, it is possible to use different approximations around a and b, and, hence, to approximate $F(a)$ and $F(b)$ independently. Since $F(a)$ depends only on the behavior of f around a, the approximating Hermite polynomial can have a much lower degree. In the theorem below, we show that we can obtain a similar accuracy as in [14, Thm. 2.3] with two independently constructed polynomials of degree $s-1$ instead of one polynomial of degree $2s-1$.

THEOREM 6.1. *Assume that f is a smooth function and g is analytic. Let $f_a(x)$ and $f_b(x)$ be the Hermite interpolating polynomials of degree $s-1$ that satisfy*

$$f_a^{(k)}(a) = f^{(k)}(a) \quad \textit{and} \quad f_b^{(k)}(b) = f^{(k)}(b), \quad k = 0, \ldots, s-1.$$

Then the approximation of (1.1) *by*

$$F_{f_a}(a) - F_{f_b}(b) := \int_0^\infty f_a(h_a(p))e^{i\omega g(h_a(p))}h_a'(p)\,\mathrm{d}p - \int_0^\infty f_b(h_b(p))e^{i\omega g(h_b(p))}h_b'(p)\,\mathrm{d}p$$

along the paths $h_a(p)$ *and* $h_b(p)$ *that satisfy* (3.10) *has an error of order* $O(\omega^{-s-1})$.

Proof. First, we consider the approximation with the Hermite interpolating polynomial $\tilde{f}(x)$ of degree $2s-1$ that satisfies $\tilde{f}^{(k)}(a) = f^{(k)}(a)$ and $\tilde{f}^{(k)}(b) = f^{(k)}(b)$, $k = 0, \ldots, s-1$. Since $\tilde{f}$ is analytic, it can be used to approximate (1.1) as $I \approx F_{\tilde{f}}(a) - F_{\tilde{f}}(b)$. This approximation has an error of $O(\omega^{-s-1})$ by [14, Thm. 2.3]. Now consider the approximation of $F_{\tilde{f}}(a)$ by $F_{f_a}(a)$. Since $\tilde{f}(x)$ is a polynomial, we can write $F_{\tilde{f}}(a)$ as

$$F_{\tilde{f}}(a) = \sum_{k=0}^{2s-1} \tilde{f}^{(k)}(a)\frac{\mu_k(a)}{k!}, \tag{6.1}$$

where the $\mu_k(a)$ are the moments of the form

$$\begin{aligned}\mu_k(a) &:= \int_0^\infty (h_a(p)-a)^k e^{i\omega g(h_a(p))}h_a'(p)\,\mathrm{d}p \\ &= \int_0^\infty \frac{e^{i\omega g(a)}}{\omega}(h_a(q/\omega)-a)^k e^{-q}h_a'(q/\omega)\,\mathrm{d}q.\end{aligned} \tag{6.2}$$

Although q goes to infinity, the behavior for small q/ω dominates due to the factor e^{-q} (this follows from Watson's lemma [1]). Since $(h_a(q/\omega)-a) \sim \omega^{-1}$, we see that $\mu_k(a) \sim \omega^{-k-1}$. For $F_{f_a}(a)$, we have

$$F_{f_a}(a) = \sum_{k=0}^{s-1} f_a^{(k)}(a)\frac{\mu_k(a)}{k!}. \tag{6.3}$$

The first discarded moment, $\mu_s(a)$, is of order $O(\omega^{-s-1})$. The approximation of $F_{\tilde{f}}(b)$ by $F_{f_b}(b)$ has an error of the same order. This concludes the proof. ∎

There are two ways to proceed: either $f_a(x)$ can be evaluated explicitly in the quadrature evaluation of $F_{f_a}(a)$, or the moments (6.2) can be precomputed with the previous techniques and used in the summation (6.3). The result is a quadrature rule for integrals of type (1.1) for fixed g, a, and b, and using function values and derivatives of f at a and b. Define $w_{i,1} = \frac{\mu_k(a)}{k!}$ and $w_{i,2} = -\frac{\mu_k(b)}{k!}$. Then

$$I \approx Q_\mu[f] := \sum_{i=0}^{s-1} w_{i,1}f^{(i)}(a) + \sum_{i=0}^{s-1} w_{i,2}f^{(i)}(b) \tag{6.4}$$

is a quadrature rule with an error of order $O(\omega^{-s-1})$. For a fixed frequency, this *localized Filon-type method* is exact for polynomials up to degree $s-1$, while the regular Filon-type method is exact for polynomials up to degree $2s-1$. Hence, the simplified construction comes at a cost; the order of accuracy as a function of ω is the same, but we can expect the coefficient to be much larger.

We can generalize the result to include stationary points. The same reasoning applies, but we need to interpolate more derivatives in order to achieve a similar convergence rate. That rate depends on the smallest value of r for which $g^{(r+1)}(\xi) \neq 0$ with ξ a stationary point.

THEOREM 6.2. *Assume that g is analytic and that $g^{(k)}(\xi) = 0$, $k = 1, \ldots, r$, and $g^{(r+1)}(\xi) \neq 0$. Let f be sufficiently smooth, and let $f_\xi(x)$ be the Hermite interpolating polynomial of degree $s(r+1)-1$ that satisfies*

$$f_\xi^{(k)}(\xi) = f^{(k)}(\xi), \quad j = 0, \ldots, s(r+1)-1.$$

Then the sequence $F_{f_\xi,j}(\xi)$ converges for increasing values of s to a limit with an error of order $O(\omega^{-s-1/(r+1)})$.

Proof. The proof follows essentially the same lines as the proof of Theorem 6.1 and uses the moments $\mu_{k,j}(\xi)$, defined using the path $h_{\xi,j}$,

$$\mu_{k,j}(\xi) := \int_0^\infty \frac{e^{i\omega g(\xi)}}{\omega} \left(h_{\xi,j}\left(\frac{q}{\omega}\right) - \xi \right)^k e^{-q} h'_{\xi,j}\left(\frac{q}{\omega}\right) \mathrm{d}q. \tag{6.5}$$

The derivative of the parameterization $h_{\xi,j}$ in the integrand has an integrable singularity of the form $(q/\omega)^{-r/(r+1)}$ at the stationary point ξ and leads to a factor $\omega^{r/(r+1)}$. By (4.4) we have $(h_{\xi,j}(q/\omega) - \xi) \sim \omega^{-1/(r+1)}$. This makes $\mu_{k,j}(\xi) \sim \omega^{r/(r+1)-k/(r+1)-1} = \omega^{(-k-1)/(r+1)}$. The first discarded moment $\mu_{k,j}(\xi)$ in the sum $F_{f_\xi,j}$ of the form (6.3) has the index $k = s(r+1)$, which leads to the result. ∎

Theorem 6.2 shows only that the value $F_{f_\xi}(\xi)$ converges with a specific rate if more derivatives of f are interpolated. It does not explicitly state that $F_{f_\xi}(\xi)$ can be used in a decomposition to approximate (1.1). The existence of an analytic function $\tilde{f}$ that can be used to approximate the value of (1.1) with an arbitrary accuracy, provided f is smooth enough, was proved in [14, Thm. 3.3] using Hermite interpolation.

Assume there is one stationary point $\xi \in (a, b)$, and $g^{(r+1)}(\xi) \neq 0$. Then we can extend the definition of quadrature rule (6.4) to

$$I \approx Q_\mu[f] := \sum_{i=0}^{s-1} w_{i,1} f^{(i)}(a) + \sum_{i=0}^{s(r+1)-1} w_{i,2} f^{(i)}(\xi) + \sum_{i=0}^{s-1} w_{i,3} f^{(i)}(b), \tag{6.6}$$

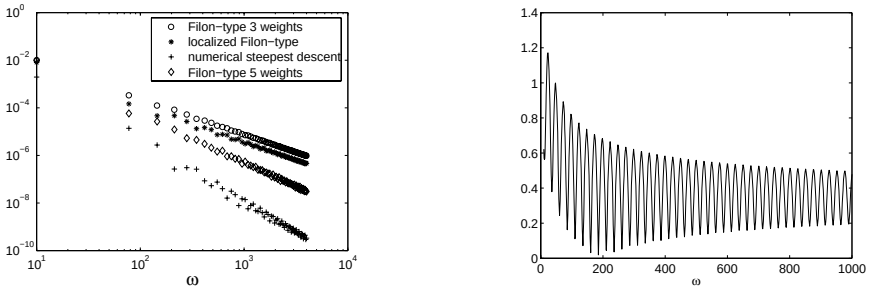
with $w_{i,1} = \frac{\mu_k(a)}{k!}$, $w_{i,3} = -\frac{\mu_k(b)}{k!}$, and $w_{i,2} = \frac{-\mu_{k,1}(\xi)+\mu_{k,2}(\xi)}{k!}$. This rule has an absolute error of order $O(\omega^{-s-1/(r+1)})$ and a relative error of order $O(\omega^{-s})$.

Example 6.3. We consider the functions $f(x) = 1/(1+x)$ and $g(x) = (x-1/3)^2$ on $[0, 1]$. Since f is analytic, we could use the previous techniques. However, here we will use only the values of the first few derivatives of f at 0 and 1 and at the stationary point $\xi = 1/3$. The results are shown in Table 6.1 for varying degrees of interpolation. The convergence rate is limited to the convergence rate at the stationary point. According to Theorem 6.2, in order to obtain an error of order $O(\omega^{-s-1/(r+1)})$, we need to interpolate up to the derivative of order $m = s(r+1)-1$. Hence, solving the latter expression for s, we expect a convergence rate of $(m+2)/(r+1)$. The rate is actually higher in the columns with even m, due to the cancellation of the moments at ξ with odd index. For a more general function g there is no exact cancellation, but the difference of the moments at ξ, i.e., $\mu_{k,1}(\xi) - \mu_{k,2}(\xi)$, can have lower order than predicted by Theorem 6.2. This cancellation does not occur if the stationary point ξ is the endpoint of the integration interval.

TABLE 6.1

Absolute error of the approximation of I for $f(x) = 1/(1+x)$ and $g(x) = (x-1/3)^2$ on $[0,1]$. We approximate f by interpolating m derivatives. The last row shows the value of $\log_2(e_{1280}/e_{2560})$: this value should approximate $(m+2)/2$ for odd m and $(m+3)/2$ for even m.

$\omega \setminus m$	0	1	2	3	4
160	$1.0E-4$	$1.8E-4$	$9.5E-7$	$9.7E-7$	$8.6E-9$
320	$6.5E-5$	$6.5E-5$	$1.7E-7$	$1.7E-7$	$7.6E-10$
640	$2.8E-5$	$2.3E-5$	$3.1E-8$	$3.0E-8$	$6.7E-11$
1280	$8.1E-6$	$8.2E-6$	$5.4E-9$	$5.4E-9$	$5.9E-12$
2560	$3.2E-6$	$2.9E-6$	$9.5E-10$	$9.5E-10$	$5.2E-13$
Rate	1.4	1.5	2.5	2.5	3.5



(a) (b)

FIG. 6.1. *A numerical comparison between the regular and localized Filon-type methods and the numerical steepest descent method (Example 6.4).* (a) *Absolute error for four methods.* (b) *The absolute error for numerical steepest descent, scaled by $\omega^{5/2}$.*

Example 6.4. We make a numerical comparison between the regular Filon-type method, the localized Filon-type method, and the numerical steepest descent method for $f(x) = 1/(1+x^2)$ and $g(x) = (x-1/2)^2$ on $[-1,1]$. Filon-type methods for this integral suffer from Runge's phenomenon: the interpolation error for the function f is large [19]. We choose $s = 1$; i.e., we use only function values of f in $\{-1, 1/2, 1\}$ and no derivatives. The order of the Filon-type methods is then $O(\omega^{-3/2})$. We choose $n = 1$ in Theorem 4.6. The order of the numerical steepest descent method is then $O(\omega^{-5/2})$, using four evaluations of f in the complex plane. We also interpolate two additional derivatives at $1/2$ for the Filon-type method: this yields a quadrature rule with five weights and order $O(\omega^{-2})$. The results are illustrated in Figure 6.1.

7. Generalization to a nonanalytic function $g(x)$. If $g(x)$ is piecewise analytic, the integration interval can be split into subintervals, where the function is analytic. Otherwise we can try to approximate $g(x)$ by an analytic function $\tilde{g}(x)$ on $[a,b]$. We should take care to not introduce new stationary points and to make sure that we accurately approximate all stationary points of $g(x)$. Alternatively, we can approximate $g(x)$ locally around the special points, possibly by using different functions for each point. This will turn out to be easier and will yield the same convergence rate.

When $g(x)$ is smooth, it can be approximated arbitrarily well by an analytic function $\tilde{g}(x)$ on $[a,b]$. Hence, there exist analytic $\tilde{g}$ such that the integral

$$\tilde{I} := \int_a^b f(x) e^{i\omega\tilde{g}(x)}\,\mathrm{d}x = \tilde{F}(a) - \tilde{F}(b) \tag{7.1}$$

is arbitrarily close to the value of I. Such function $\tilde{g}$ may be difficult to find, however, and, if found, intractable for numerical purposes. Hermite interpolation in the points a and b is not a solution in this case, as the resulting polynomial may introduce stationary points that the original function g did not have. However, owing to decomposition (7.1), it becomes possible to do Hermite interpolation in a and b separately by different polynomials.

THEOREM 7.1. *Assume that f and $\tilde{g}$ are analytic. Let $g_a(x)$ be the Hermite interpolating polynomial of degree s that satisfies*

$$g_a^{(k)}(a) = \tilde{g}^{(k)}(a), \quad k = 0, \dots, s.$$

Then the approximation of $\tilde{F}(a)$ by $F_{g_a}(a)$ has an error of order $O(\omega^{-s-1})$.

Proof. The error $e := \tilde{F}(a) - F_{g_a}(a)$ can be written as

$$\begin{aligned} e &= \int_0^\infty f(h_a(p))(e^{i\omega\tilde{g}(h_a(p))} - e^{i\omega g_a(h_a(p))})h_a'(p)\,\mathrm{d}p \\ &= \int_0^\infty f(h_a(p))e^{i\omega g_a(h_a(p))}(e^{i\omega(\tilde{g}(h_a(p))-g_a(h_a(p)))} - 1)h_a'(p)\,\mathrm{d}p \\ &= \frac{e^{i\omega g_a(a)}}{\omega}\int_0^\infty f\left(h_a\left(\frac{q}{\omega}\right)\right)e^{-q}(e^{i\omega(\tilde{g}(h_a(\frac{q}{\omega}))-g_a(h_a(\frac{q}{\omega})))} - 1)h_a'\left(\frac{q}{\omega}\right)\mathrm{d}q. \end{aligned} \tag{7.2}$$

The path $h_a(p)$ was chosen as the solution of (3.10) with respect to the approximation $g_a(x)$. Using a Taylor approximation around a, we have

$$\tilde{g}(x) - g_a(x) = (\tilde{g}^{(s+1)}(a) - g_a^{(s+1)}(a))\frac{(x-a)^{s+1}}{(s+1)!} + O((x-a)^{s+2}).$$

Because $h_a(q/\omega) - a \sim \omega^{-1}$, we have

$$e^{i\omega(\tilde{g}(h_a(q/\omega))-g_a(h_a(q/\omega)))} - 1 \sim i\omega(\tilde{g}(h_a(q/\omega)) - g_a(h_a(q/\omega))) \sim \omega^{-s}.$$

The error e is therefore of order $O(\omega^{-s-1})$. ∎

The value of $\tilde{I}$, defined by (7.1), is completely determined by the derivatives of $\tilde{g}$ at a and b. If $\tilde{I} - I$ is small, it follows from Theorem 7.1 that $\tilde{g}$ should satisfy $\tilde{g}^{(j)}(a) = g^{(j)}(a)$ and $\tilde{g}^{(j)}(b) = g^{(j)}(b)$, $j = 0, \dots, s$, for some maximal order s that depends on the smoothness of g. Hence, $\tilde{g}$ need not be explicitly constructed.

At a stationary point ξ, more derivatives are needed. The convergence rate depends on the minimal value of r for which $\tilde{g}^{(r+1)}(\xi) \neq 0$.

THEOREM 7.2. *Assume that f and $\tilde{g}$ are analytic and that $\tilde{g}^{(k)}(\xi) = 0$, $k = 1, \dots, r$, and $\tilde{g}^{(r+1)}(\xi) \neq 0$. Let $g_\xi(x)$ be the Hermite interpolating polynomial of degree $(s+1)(r+1)-1$ that satisfies*

$$g_\xi^{(k)}(\xi) = \tilde{g}^{(k)}(\xi), \quad k = 0, \dots, (s+1)(r+1)-1.$$

Then the approximation of $\tilde{F}_j(\xi)$ by $F_{g_\xi,j}(\xi)$ has an error of order $O(\omega^{-s-1/(r+1)})$.

Proof. The proof follows the same lines as the proof of Theorem 7.1. The difference is that, similar to the situation in the proof of Theorem 6.2, we have $h_{\xi,j}(q/\omega) - \xi \sim \omega^{-1/(r+1)}$ and $h_{\xi,j}'(q/\omega) \sim \omega^{r/(r+1)}$. This leads to

$$e^{i\omega(\tilde{g}(h_{\xi,j}(q/\omega))-g_\xi(h_{\xi,j}(q/\omega)))} - 1 \sim \omega^{-s}.$$

TABLE 7.1

Absolute error of the approximation of $\tilde{F}(a)$ by $F_a(a)$ for $f(x) = 1/(1+x)$ and $g(x) = (x - 1/2)^2(x-2)e^{x^2}$ at $a = 0$. We approximate g by interpolating m derivatives. The last row shows the value of $\log_2(e_{400}/e_{800})$: this value should approximate $m+1$.

$\omega \setminus m$	1	2	3	4
100	$6.1E-5$	$7.6E-7$	$1.3E-8$	$1.9E-10$
200	$1.5E-5$	$9.5E-8$	$8.4E-10$	$6.1E-12$
400	$3.8E-6$	$1.2E-8$	$5.3E-11$	$1.9E-13$
800	$9.6E-7$	$1.5E-9$	$3.3E-12$	$6.0E-15$
Rate	2.0	3.0	4.0	5.0

TABLE 7.2

Absolute error of the approximation of I by $\tilde{I}$ for $f(x) = 1/(1+x)$ and $g(x) = (x-1/2)^2(x-2)e^{x^2}$ on $[0,1]$. We approximate g by interpolating m derivatives. The last row shows the value of $\log_2(e_{400}/e_{800})$: this value should approximate $m/2$ for odd m and $(m+1)/2$ for even m.

$\omega \setminus m$	2	3	4
100	$1.6E-4$	$2.7E-4$	$1.8E-7$
200	$5.5E-5$	$9.8E-6$	$3.2E-8$
400	$2.0E-5$	$3.5E-6$	$5.6E-9$
800	$6.9E-6$	$1.2E-6$	$9.9E-10$
Rate	1.5	1.5	2.5

The error estimate for this case is analogous to (7.2) in the proof of Theorem 7.1. Adding all contributions, it is of order $O(\omega^{-1-s+r/(r+1)}) = O(\omega^{-s-1/(r+1)})$. □

Example 7.3. We illustrate the convergence with two examples. The function $g(x) = (x-1/2)^2(x-2)e^{x^2}$ is approximated by a polynomial of degree m in the endpoints $a = 0$ and $b = 1$, and in the stationary point $\xi = 1/2$. The resulting errors are displayed in Tables 7.1 and 7.2. Table 7.1 shows the error in approximating only $\tilde{F}(a)$. Table 7.2 shows the error of the approximation of I. The latter error is dominated by the error made at the stationary points but follows the theory. As in the last example for a nonanalytic function f, the convergence rate is actually higher for even m because the difference of the terms at ξ in the decomposition of I can have an order lower than predicted by Theorem 7.2. Note that it is not possible to approximate $g(x)$ by a fixed constant since in that case also $e^{i\omega g_a(x)} = e^{i\omega c}$ reduces to a constant. At a stationary point with r vanishing derivatives, the minimal number of derivatives to interpolate is $r+1$.

8. Concluding remarks. We have presented an approach to compute highly oscillatory integrals of the form (1.1). The method is quite general and leads to high order convergence when the frequency increases. The (generalized) Gauss–Laguerre quadrature rules yields the typical Gauss rule convergence exponent of approximately $2n$, but here as a function of $1/\omega$. This is made possible by transforming the integrand into a numerically well behaved one, i.e., one that is not oscillatory and that has exponential decay which becomes faster with increasing ω.

The approach by Iserles and Nørsett has led us to consider the use of Hermite interpolation for functions $f(x)$ that are not analytic. The resulting polynomial is analytic, and this enables the use of our rapidly converging complex approach. Owing to our decomposition of the integral into a sum of a number of functions that each depend only on one point, this approach could be simplified considerably in our setting. Vice versa, the methods developed in this paper may be used to compute generalized moments of the form $\int_0^1 p(x)e^{i\omega g(x)}\mathrm{d}x$ with $p(x)$ a polynomial of low degree. Such

moments are assumed to be available in the approach of Iserles and Nørsett, but an analytical value may not always be available. The details of the latter method can be found in [14].

Acknowledgments. The authors wish to thank Arieh Iserles and Sheehan Olver for many insightful discussions on the topic of oscillatory integrals (suggesting the term *numerical steepest descent* in the process) and to thank the anonymous referees for a number of helpful suggestions and references.

REFERENCES

[1] M. J. Ablowitz and A. S. Fokas, *Complex Variables: Introduction and Applications*, Cambridge University Press, Cambridge, UK, 1997.
[2] N. Bleistein and R. Handelsman, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.
[3] S. N. Chandler-Wilde and D. C. Hothersall, *Efficient calculation of the Green function for acoustic propagation above a homogeneous impedance plane*, J. Sound Vibration, 180 (1995), pp. 705–724.
[4] K. Davies, M. Strayer, and G. White, *Complex-plane methods for evaluating highly oscillatory integrals in nuclear physics* I, J. Phys. G: Nucl. Phys., 14 (1988), pp. 961–972.
[5] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Comput. Sci. Appl. Math., Academic Press, New York, 1984.
[6] P. Debye, *Näherungsformeln für die Zylinderfunktionen für grosse Werte des Arguments und unbeschränkt veranderliche Werte des Index*, Math. Anal., 67 (1909), pp. 535–558.
[7] G. A. Evans, K. C. Chung, and J. R. Webster, *A method to generate generalised quadrature rules for oscillatory integrals*, App. Numer. Math., 34 (2000), pp. 85–93.
[8] G. A. Evans and K. C. Chung, *Some theoretical aspects of generalised quadrature methods*, J. Complexity, 19 (2003), pp. 272–285.
[9] L. N. G. Filon, *On a quadrature formula for trigonometric integrals*, Proc. Roy. Soc. Edinburgh, 49 (1928), pp. 38–47.
[10] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Clarendon Press, Oxford, UK, 2004.
[11] P. Henrici, *Applied and Computational Complex Analysis,* Vol. I, John Wiley & Sons, New York, 1974.
[12] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* I: *Fourier transforms*, IMA J. Numer. Anal., 24 (2004), pp. 365–391.
[13] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* II: *Irregular oscillators*, IMA J. Numer. Anal., 25 (2005), pp. 25–44.
[14] A. Iserles and S. Nørsett, *Efficient quadrature of highly oscillatory integrals using derivatives*, Proc. Royal Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 1383–1399.
[15] A. Iserles and S. Nørsett, *On quadrature methods for highly oscillatory integrals and their implementation*, BIT, 44 (2004), pp. 755–772.
[16] K. J. Kim, R. Cools, and L. G. Ixaru, *Quadrature rules using first derivatives for oscillatory integrands*, J. Comput. Appl. Math., 140 (2002), pp. 479–497.
[17] D. Levin, *Fast integration of rapidly oscillatory functions*, J. Comput. Appl. Math., 67 (1996), pp. 95–101.
[18] S. Olver, *Moment-free numerical integration of highly oscillatory functions*, IMA J. Numer. Anal., 26 (2006), pp. 213–227.
[19] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, UK, 1981.
[20] E. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.
[21] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1049–1072

FULLY DISCRETE FINITE ELEMENT APPROXIMATIONS OF THE NAVIER–STOKES–CAHN-HILLIARD DIFFUSE INTERFACE MODEL FOR TWO-PHASE FLUID FLOWS*

XIAOBING FENG†

Abstract. This paper develops and analyzes some fully discrete finite element methods for a parabolic system consisting of the Navier–Stokes equations and the Cahn–Hilliard equation, which arises as a diffuse interface model for the flow of two immiscible and incompressible fluids. In the model the two sets of equations are coupled through an extra phase induced stress term in the Navier–Stokes equations and a fluid induced transport term in the Cahn–Hilliard equation. Fully discrete mixed finite element methods are proposed for approximating the coupled system, it is shown that the proposed numerical methods satisfy a mass conservation law, and a discrete energy law which is analogous to the basic energy law for the phase field model. The convergence of the numerical solutions to the solutions of the phase field model and its sharp interface limit is established by utilizing the discrete energy law. As a by-product, the convergence result also provides a constructive proof of the existence of weak solutions to the Navier–Stokes-Cahn–Hilliard phase field model. Numerical experiments are also presented to validate the theory and to show the effectiveness of the combined phase field and finite element approach.

Key words. two-phase fluids, phase field model, Cahn–Hilliard equation, Navier–Stokes equations, finite element methods

AMS subject classifications. 65M60, 35K55, 76D05

DOI. 10.1137/050638333

1. Introduction.

Interfacial dynamics in the mixture of different fluids, solids, or gas has been one of the fundamental issues in hydrodynamics and materials science. It plays an increasingly important role in many current scientific, engineering, and industrial applications (cf. [7, 15] and the references therein). In the classical approaches, the interface is usually considered as a free curve/surface that evolves in time along with fluid. The movement of the interface at each time is determined by a set of interfacial balance conditions. In the case of two immiscible incompressible fluids, the dynamics of the fluid mixture is described by the following two-phase Navier–Stokes equations:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{g} \quad \text{in } \Omega_T \setminus \Gamma_t, \tag{1.1}$$

$$\text{div } \mathbf{u} = 0 \quad \text{in } \Omega_T \setminus \Gamma_t, \tag{1.2}$$

$$[(\nu D(\mathbf{u}) - pI)\mathbf{n}] = \alpha\kappa\mathbf{n} \quad \text{on } \Gamma_t, \tag{1.3}$$

$$[\mathbf{u}] = 0 \quad \text{on } \Gamma_t, \tag{1.4}$$

with a given set of initial and boundary conditions. Here $\Omega \subset \mathbf{R}^d\,(d = 1, 2, 3)$ is a bounded domain, $\Omega_T = \Omega \times (0, T)$, Γ_t denotes the (free) interface at the time t with the normal $\mathbf{n}$ and the mean curvature κ, $\alpha > 0$ is the surface tension constant. $D(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$ denotes the deformation tensor, I is the $d \times d$ identity matrix. $[\mathbf{u}]$ denotes the jump of the $\mathbf{u}$ across the interface Γ_t. Clearly, (1.3) and (1.4) are the

*Received by the editors August 17, 2005; accepted for publication (in revised form) December 5, 2005; published electronically June 2, 2006. This work is partially supported by the NSF grant DMS-0410266.

http://www.siam.org/journals/sinum/44-3/63833.html

†Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu).

interfacial conditions for the fluid mixture, which are the mathematical descriptions of the balance of the normal stress and the balance of the movement.

Computationally, the above free interface problem is difficult to solve directly due to the existence of the surface tension on the interface. In addition, during the evolution the fluid interface may experience topological changes such as self-intersection, pinch-off, splitting, and fattening. When that happens, the classical solution of the free interface problem ceases to exist. In such a situation it is delicate and difficult to develop a proper notion of generalized solutions, it becomes even more challenging to compute the generalized solutions when they are defined.

To overcome the difficulties, an alternative approach for solving interface problems is the *diffuse interface (or mean field) theory*, which was originally developed as methodology for modeling and approximating solid-liquid phase transitions in which the effects of surface tension and nonequilibrium thermodynamic behavior may be important at the surface [28, 12, 24]. In the theory, the interface is represented as a thin layer of finite thickness. The method uses an auxiliary function (called phase field function/variable) to indicate the "phase." The phase-field function assumes distinct values in the bulk phases away from the interfacial regions over which the phase function varies smoothly, and the interface itself can be associated with an intermediate contour or level set of the phase function (cf. [32] and the references therein). The diffuse interface models converge to their corresponding sharp interface models as the width of the interfacial layer ε tends to zero.

This is the second paper in a series (cf. [18]) which devotes to finite element numerical analysis of two-phase fluid flows based on the phase field (diffuse interface) approach. In [18] finite element methods were developed and analyzed for the Navier–Stokes-Allen–Cahn phase field model proposed in. The goal of this paper is to carry out a parallel finite element numerical analysis for a mass-conserved diffuse interface model for two-phase fluids proposed in [27, 30], which consists of the Navier–Stokes equations and the Cahn–Hilliard equation. In the model the two sets of equations are coupled through an extra phase induced stress term in the Navier–Stokes equations and a fluid induced transport term in the Cahn–Hilliard equation. We develop and analyze some fully discrete mixed finite element methods for the Navier–Stokes-Cahn–Hilliard phase field model. It is proved that the proposed numerical methods satisfy a mass conservation law, and a discrete energy law which exactly mimics the basic energy law for the phase field model. The convergence of the numerical solutions to the solutions of the phase field model and its sharp interface limit is then established by utilizing the discrete energy law.

It should be noted that the Navier–Stokes-Cahn–Hilliard phase field model (in different forms, see section 2 for more details) for two phase fluids has been studied numerically by several authors [2, 3, 27, 30, 31], among them [27, 30] are most closely related to the current paper. In [27] Jacqmin derived the mathematical model based on physical arguments, and then proposed some finite difference compact schemes for the model. Impressive numerical experiments also were reported in the paper although no convergence analysis was given. In [30], inspired by their early experiences on the nematic liquid crystal flows, Liu and Shen rederived the Navier–Stokes-Cahn–Hilliard phase field model based on the (heuristic) mathematical arguments, and then proposed some Fourier-spectral element methods under the periodic assumptions. *Local-in-time* stability estimates were also established for the proposed Fourier-spectral element methods.

The remainder of this paper is organized as follows. In section 2, we present the mass conserved Navier–Stokes-Cahn–Hilliard phase field model for two-phase fluids

to be studied in this paper. We then demonstrate that various phase field models appeared in [27, 30] are actually mathematically equivalent up to an additive function to the pressure. However, it is shown later in this paper that numerically the phase field model in the potential form is favored since finite element methods based on this form are shown to fulfill a *global-in-time* energy (or stability) estimate which exactly mimics the basic energy law of the differential model. In section 3, we first re-establish the basic energy law (in a slightly different form) associated with the Navier–Stokes-Cahn–Hilliard phase field model, and then derive some additional a priori energy estimates which show explicit dependence on the physical parameters $\varepsilon, \lambda, \gamma$, and ν. In section 4, we propose and analyze a family of fully discrete mixed finite element methods for the phase field model. It is proved that the proposed numerical methods enjoy a discrete energy law which mimics the basic energy law for the differential problem. It is this discrete energy law which paves the way for us to establish the convergence of the fully discrete methods to the phase field model as the mesh sizes $h, \tau \to 0$, and to the sharp interface model (1.1)–(1.4) as the mesh sizes h, τ, and the capillary width ε all tend to zero, provided that the phase field model converges to the sharp interface model. Our main idea is to rewrite the flow equations in the potential form by introducing a new "pressure" $\overline{p} = p + \frac{\lambda}{2}|\nabla\varphi|^2 + \frac{\lambda}{\varepsilon^2}F(\varphi) + \lambda\varphi(\Delta\varphi - \frac{1}{\varepsilon^2}F'(\varphi))$ in the place of the original pressure p. As a by-product, our convergence result also provides a rigorous proof of the existence of weak solutions to the phase field model, which clearly is of interests in itself. Finally, in section 5 we present some numerical experiment results to validate our theoretical results and to show the effectiveness of the combined phase field and finite element approach.

2. The Navier–Stokes-Cahn–Hilliard phase field model. The phase field model for two immiscible and incompressible fluids with comparable densities (which are taken to be 1) and viscosities $\nu > 0$ to be studied in this paper takes the form [30, 3]

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p + \lambda\,\mathrm{div}\,(\nabla\varphi\otimes\nabla\varphi) = \mathbf{g} \quad \text{in } \Omega_T, \tag{2.1}$$

$$\varphi_t + \mathbf{u}\cdot\nabla\varphi + \gamma\Delta\big(\Delta\varphi - \frac{1}{\varepsilon^2}f(\varphi)\big) = 0 \quad \text{in } \Omega_T, \tag{2.2}$$

$$\mathrm{div}\ \mathbf{u} = 0 \quad \text{in } \Omega_T. \tag{2.3}$$

To close the system, it will be complemented with the following initial and boundary conditions:

$$\mathbf{u}(\cdot,0) = \mathbf{u}_0^\varepsilon(\cdot), \qquad \varphi(\cdot,0) = \varphi_0^\varepsilon(\cdot), \qquad \text{in } \Omega, \tag{2.4}$$

$$\mathbf{u} = 0, \qquad \frac{\partial\varphi}{\partial\mathbf{n}} = \frac{\partial\Delta\varphi}{\partial\mathbf{n}} = 0, \qquad \text{on } \partial\Omega_T := \partial\Omega\times(0,T]. \tag{2.5}$$

Note that we have suppressed the superscript ε in $(\mathbf{u}^\varepsilon, \varphi^\varepsilon, p^\varepsilon)$ for the notation brevity. Here the vector $\mathbf{u}(x,t) \in \mathbf{R}^d$ and the scalar $p(x,t) \in \mathbf{R}$ denote the velocity and the pressure of the fluid mixture at the space time point (x,t), respectively. The scalar function φ is called a *phase function* and is used to indicate the fluid phases. φ assumes distinct values in the bulk phases away from a thin layer (called the interfacial region) over which φ varies smoothly, and the interface itself can be associated with the zero level set $\{\varphi = 0\}$ of φ ($f(\varphi) = F'(\varphi)$ and $F(\varphi) = \frac{1}{4}(\varphi^2-1)^2$). The positive constants λ, γ, and ε are the surface tension, the elastic relaxation time, and the capillary width (width of the interfacial layer), respectively. $\nabla\varphi\otimes\nabla\varphi$ stands for the $d\times d$ rank-one matrix $(\nabla\varphi)^T\nabla\varphi$ with entries $\varphi_{x_i}\varphi_{x_j}$. We especially note that $\varepsilon << 1$.

Equation (2.1) without the stress term $\lambda \text{div}\,(\nabla\varphi \otimes \nabla\varphi)$ is the Navier–Stokes equations [33] and (2.2) without the convection term $u \cdot \nabla\varphi$ is the Cahn–Hilliard equation [23, 32]. In the literature, the phase equation is always given by (2.2). On the other hand, the flow equations often appear differently in different papers due to how the phase induced force is expressed in the equations. In (2.1), since $\nabla\varphi \otimes \nabla\varphi$ is a phase induced stress tensor (its divergence represents the phase induced force), hence, we may regard (2.1) as the flow equations in the *stress form*. Using the differential identity

$$\text{div}\,(\nabla\varphi \otimes \nabla\varphi) = \Delta\varphi\nabla\varphi + \frac{1}{2}\nabla|\nabla\varphi|^2, \tag{2.6}$$

(2.1) can be rewritten as

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla\widehat{p} + \lambda\,\Delta\varphi\nabla\varphi = \mathbf{g}, \tag{2.7}$$

where

$$\widehat{p} := p + \frac{\lambda}{2}|\nabla\varphi|^2. \tag{2.8}$$

By introducing the *chemical potential* (cf. [11, 23, 32]),

$$w := -\Delta\varphi + \frac{1}{\varepsilon^2}f(\varphi), \tag{2.9}$$

and noticing the fact that $F'(\varphi) = f(\varphi)$, (2.7) can be rewritten as

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla\widetilde{p} - \lambda\, w\nabla\varphi = \mathbf{g}, \tag{2.10}$$

where

$$\widetilde{p} := \widehat{p} + \frac{\lambda}{\varepsilon^2}F(\varphi) = p + \frac{\lambda}{2}|\nabla\varphi|^2 + \frac{\lambda}{\varepsilon^2}F(\varphi). \tag{2.11}$$

It was based exactly on $(\mathbf{u}, \widetilde{p}, \varphi, w)$ formulation that convergent finite element methods were successfully developed in [18] for the related Navier–Stokes-Allen–Cahn phase field model of for two-phase fluids.

It is natural to ask if the success of [18] can be extended to the above Navier–Stokes-Cahn–Hilliard phase field model. It turns out (see section 4 for details) that one can show that mixed finite methods based on $(\mathbf{u}, \widetilde{p}, \varphi, w)$ formulation for the Navier–Stokes-Cahn–Hilliard phase field model do satisfy a discrete energy law which mimics the basic energy law associated with the phase field model. However, the numerical solutions may *not* satisfy the mass conservation law

$$\int_\Omega \varphi(x,t)\,dx = \int_\Omega \varphi(x,0)\,dx \qquad \forall t \in [0,T],$$

associated with the Navier–Stokes-Cahn–Hilliard phase field model (2.1)–(2.5).

To overcome the difficulty, we define another new "pressure" $\overline{p}$ as

$$\overline{p} := \widetilde{p} - \lambda\varphi w = p + \frac{\lambda}{2}|\nabla\varphi|^2 + \frac{\lambda}{\varepsilon^2}F(\varphi) + \lambda\varphi\Big(\Delta\varphi - \frac{1}{\varepsilon^2}F'(\varphi)\Big), \tag{2.12}$$

and rewrite (2.10) as

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla\overline{p} + \lambda\,\varphi\nabla w = \mathbf{g}, \tag{2.13}$$

which it turns out is exactly the flow equations proposed by Jacqmin in [27] using physical arguments. Since the phase induced force is expressed as the gradient of the chemical potential w, we may regard (2.13) as the flow equation in the *potential form*. We also note that with help of the chemical potential w, (2.2) can be rewritten as

$$\varphi_t + \mathbf{u} \cdot \nabla \varphi - \gamma \Delta w = 0. \tag{2.14}$$

Equations (2.9) and (2.14) is known as the mixed (or split) formulation for the Cahn–Hilliard equation (cf. [16, 19, 20]).

In section 4, we shall present some fully discrete mixed finite element methods for the Navier–Stokes-Cahn–Hilliard phase field model based on the $(\mathbf{u}, \overline{p}, \varphi, w)$ formulation, which consists of (2.13), (2.14), (2.3), and (2.9). It is shown that such numerical methods not only satisfy the global-in-time discrete energy law but also fulfill the mass conservation law. We like to emphasize that although the flow equations presented above all are mathematically equivalent, numerical methods based on these equations are often different, and the differences could be significant.

3. A priori energy estimates. The standard space notations are used in this paper; we refer to [1, 33] for their exact definitions. In particular, B^* denotes the dual space of a Banach space B, and $\mathbf{B}$ denotes the vector Banach space B^d. $(\cdot, \cdot)$ is used to denote the standard $L^2(\Omega)$ inner product, $\langle \cdot, \cdot \rangle$ stands for the dual product between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, and

$$L_0^2(\Omega) = \{q \in L^2(\Omega);\ (q, 1) = 0\},$$

$$\mathbf{V} = \{\mathbf{v} \in \mathbf{H}_0^1(\Omega);\ \text{div } \mathbf{v} = 0 \text{ in } \Omega\},$$

$$\mathbf{H} = \{\mathbf{v} \in \mathbf{L}^2(\Omega);\ \text{div } \mathbf{v} = 0 \text{ in } \Omega \text{ and } \mathbf{v} \cdot \mathbf{n}|_{\partial\Omega} = 0\}.$$

In addition, we use π to denote the L^2-orthogonal projection from $\mathbf{L}^2(\Omega)$ onto $\mathbf{H}$, and $\tilde{\Delta} = \pi\Delta$ to denote the Stokes operator (cf. [33]).

Throughout the paper, unless stated otherwise, c and C will be used to denote generic positive constants which is independent of $\mathbf{u}, p, \varphi$, and ε.

Existence and uniqueness (for $d = 2$) of weak solutions of system (2.1)–(2.3) were heuristically outlined in [30]. A key ingredient of the proof is to establish the basic energy law for the phase field model (see (3.9) below). In this section, we shall first reestablish the basic energy law in a slightly different form. We shall also derive some additional uniform (in ε) a priori estimates, in particular on $\mathbf{u}_t, \varphi_t$, and $\overline{p}$, for weak solutions to the system (2.1)–(2.3), which will be needed in section 4. Special attention will be given to tracing the explicit dependence of the a priori estimates on the capillary width ε, and the parameters λ and γ.

LEMMA 3.1. *Suppose that* $\mathbf{g} \in L^2((0,T); \mathbf{L}^2(\Omega))$*, and the initial values* $\mathbf{u}_0^\varepsilon$ *and* φ_0^ε *satisfy* $|\varphi_0^\varepsilon| \le 1$ *and* $\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}_0^\varepsilon, \phi_0^\varepsilon) < \infty$*, i.e., the initial energy is bounded, then every regular solution* $(\mathbf{u}, \varphi, p)$ *of system* (2.1)–(2.3) *satisfies the following estimates: for all* $T \in [0, \infty]$

$$\int_\Omega \varphi(x,t)\, dx = \int_\Omega \varphi_0^\varepsilon(x)\, dx \quad \forall t \in (0,T), \tag{3.1}$$

$$\operatorname*{ess\,sup}_{t\in[0,T]} \{\|\mathbf{u}(t)\|_{L^2}^2 + \lambda \|\nabla\varphi(t)\|_{L^2}^2 + \lambda\varepsilon^{-2}(F(\varphi(t)), 1)\} \le C, \tag{3.2}$$

$$\int_0^T \nu \,\|\nabla \mathbf{u}(t)\|_{L^2}^2\, dt + \left\{ \begin{array}{l} \displaystyle\int_0^T \lambda\gamma \,\|\nabla w(t)\|_{L^2}^2 \, dt \\ \displaystyle\int_0^T \lambda\gamma^{-1} \,\|\varphi_t(t) + \mathbf{u}(t)\cdot\nabla\varphi(t)\|_{H^{-1}}^2 \, dt \end{array} \right\} \le C, \tag{3.3}$$

$$\int_0^T \left\{ \|\mathbf{u}_t(t)\|_{V^*}^{\frac{12}{6+d}} + \|\mathbf{u}_t(t)\|_{(V\cap L^\infty)^*}^2 \right\} dt \le C, \tag{3.4}$$

$$\operatorname*{ess\,sup}_{t\in[0,T]} \left\| \int_0^t \overline{p}(s)\, ds \right\|_{L^2} \le C, \tag{3.5}$$

$$\int_0^T \|\varphi_t(t)\|_{H^{-1}}^2 \le C, \tag{3.6}$$

where $\overline{p} = p + \frac{\lambda}{2}|\nabla\varphi|^2 + \frac{\lambda}{\varepsilon^2}F(\varphi)$. *In addition, there holds*

$$\int_0^T \|\Delta\varphi(t)\|_{L^2}^2 \, dt \le C\,\varepsilon^{-2}, \tag{3.7}$$

$$\int_0^T \|\nabla\Delta\varphi\|_{L^2}^2 \, dt \le C\,\varepsilon^{-\frac{6(4+d)}{6-d}}. \tag{3.8}$$

Proof. Testing (2.1) with $\mathbf{u}$, (2.14) with $\lambda\gamma^{-1}\Delta^{-1}(\varphi_t + \mathbf{u}\cdot\nabla\varphi)$ or with w, (2.9) with φ_t, using the differential identity (2.6), and adding the resulting equations yield

$$\frac{d}{dt}\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u},\varphi) + \nu\,\|\nabla\mathbf{u}\|_{L^2}^2 + \left\{ \begin{array}{l} \lambda\gamma\,\|\nabla w\|_{L^2}^2 \\ \nu\gamma^{-1}\,\|\varphi_t + \mathbf{u}\cdot\nabla\varphi\|_{H^{-1}} \end{array} \right\} = \int_\Omega \mathbf{g}\cdot\mathbf{u}\, dx, \tag{3.9}$$

where

$$\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u},\varphi) := \int_\Omega \left[\frac{1}{2}|\mathbf{u}|^2 + \frac{\lambda}{2}|\nabla\varphi|^2 + \frac{\lambda}{\varepsilon^2}F(\varphi) \right] dx. \tag{3.10}$$

The identity (3.9) is known (cf. [27, 30]) as the basic energy law for the system (2.1)–(2.5).

Next, the estimates (3.2) and (3.3) follow easily from integrating (3.9) in t from 0 to T and using the inequality

$$|(\mathbf{g},\mathbf{u})| \le \frac{1}{4}\,\|\mathbf{u}\|_{L^2}^2 + \|\mathbf{g}\|_{L^2}^2.$$

To show (3.4), we test (2.13) with $\mathbf{v}\in\mathbf{V}\cap\mathbf{L}^\infty(\Omega)$ to get

$$\begin{aligned} (\mathbf{u}_t,\mathbf{v}) &= -\nu\,(\nabla\mathbf{u},\nabla\mathbf{v}) - ((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{v}) - \lambda\,(\varphi\nabla w,\mathbf{v}) + (\mathbf{g},\mathbf{v}) \\ &\le \nu\,\|\nabla\mathbf{u}\|_{L^2}\,\|\nabla\mathbf{v}\|_{L^2} + \lambda\,\|\nabla w\|_{L^2}\,\|\varphi\|_{L^3}\,\|\mathbf{v}\|_{L^6} + \|\mathbf{g}\|_{L^2}\,\|\mathbf{v}\|_{L^2} \\ &\quad -((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{v}). \end{aligned} \tag{3.11}$$

For the last term above we have

$$((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{v}) \le C \left\{ \begin{array}{l} \|\nabla\mathbf{u}\|_{L^2}^{\frac{6+d}{6}}\,\|u\|_{L^2}^{\frac{6-d}{6}}\,\|\mathbf{v}\|_{L^6} + \|\nabla\mathbf{u}\|_{L^2}\,\|u\|_{L^2}\,\|\mathbf{v}\|_{L^6}, \\ \|\nabla\mathbf{u}\|_{L^2}\,\|u\|_{L^2}\,\|\mathbf{v}\|_{L^\infty}, \end{array} \right. \tag{3.12}$$

here we have used the interpolation inequality (cf. [1])

$$\|\mathbf{u}\|_{L^3} \le C\,\|\nabla\mathbf{u}\|_{L^2}^{\frac{d}{6}}\,\|\mathbf{u}\|_{L^2}^{\frac{6-d}{d}} + C\,\|\mathbf{u}\|_{L^2}\,. \tag{3.13}$$

Equation (3.4) then follows from combining (3.11), (3.12), (3.13), (3.2), and (3.3).

To verify (3.5), testing (2.13) with $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$ and integrating the resulted equation in t from 0 to $\tau(\le T)$ yield

$$\left(\int_0^\tau p(t)dt, \operatorname{div}\mathbf{v}\right) = (\mathbf{u}(\tau) - \mathbf{u}_0, \mathbf{v}) + \nu\left(\int_0^\tau \nabla\mathbf{u}(t)dt, \nabla\mathbf{v}\right) + \left(\int_0^\tau (\mathbf{u}\cdot\nabla\mathbf{u})(t)dt, \mathbf{v}\right) - \lambda\left(\int_0^\tau (\varphi\nabla w)(t)dt, \mathbf{v}\right).$$

Using the estimates (3.2), (3.3), (3.11)–(3.13), and the Sobolev inequality (cf. [1]) we conclude that

$$\left(\int_0^\tau p(t)dt, \operatorname{div}\mathbf{v}\right) \le C\,\|\mathbf{v}\|_{H^1} \qquad \forall\mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.14}$$

Equation (3.5) then follows from (3.14) and an application of the inf-sup inequality (cf. [33]).

To show (3.6), we first notice that $\nabla w \in L^2(\Omega_T)$ implies that $\Delta w \in L^2((0,T); H^{-1})$. Then (3.6) follows from (2.14), (3.2), (3.3), and the following inequality:

$$\|\mathbf{u}\cdot\nabla\varphi\|_{L^{\frac{6}{5}}} \le \|\mathbf{u}\|_{L^3}\,\|\nabla\varphi\|_{L^2}\,.$$

To show (3.7), testing (2.9) with $\Delta\varphi$ we get

$$\begin{aligned}\|\Delta\varphi\|_{L^2}^2 &= -(w, \Delta\varphi) + \frac{1}{\varepsilon^2}(f(\varphi), \Delta\varphi) \le \|\nabla w\|_{L^2}\,\|\nabla\varphi\|_{L^2} - \frac{1}{\varepsilon^2}(f'(\varphi), |\nabla\varphi|^2)\\ &\le \left(\frac{1}{2} + \frac{1}{\varepsilon^2}\right)\|\nabla\varphi\|_{L^2}^2 + \frac{1}{2}\,\|\nabla w\|_{L^2}^2\,.\end{aligned}$$

Here we have used the fact that $f'(\varphi) = 3\phi^2 - 1$. The assertion then follows from the above inequality, (3.2), and (3.3).

Finally, applying the operator ∇ on both sides of (2.9) yields

$$\nabla\Delta\varphi = -\nabla w + \frac{1}{\varepsilon^2}f'(\varphi)\nabla\varphi = -\nabla w + \frac{3}{\varepsilon^2}\varphi^2\nabla\varphi - \frac{1}{\varepsilon^2}\nabla\varphi,$$

which, (3.2), (3.3), and the interpolation inequality (cf. [1])

$$\|\varphi\|_{L^\infty} \le C\left(\|\Delta\varphi\|_{L^2}^{\frac{d}{2(6-d)}}\,\|\varphi\|_{L^6}^{\frac{3(4-d)}{2(6-d)}} + \|\varphi\|_{L^6}\right)$$

imply that

$$\int_0^T \|\nabla\Delta\varphi\|_{L^2}^2\,dt \le C\big(1 + \varepsilon^{-4-\frac{2d}{6-d}}\big).$$

Hence, (3.8) holds. The proof is complete. ∎

Remark 3.1. (a) Compare with a priori estimates for the Navier–Stokes-Allen–Cahn phase field model obtained in [18], here for the Navier–Stokes-Cahn–Hilliard model we get better uniform estimates for w, $\mathbf{u}_t$, and $\overline{p}$. However, the estimate for φ_t is weaker.

(b) The estimate (3.1) is often known as the mass conservation property of the Cahn–Hilliard equation. It should be noted that this property does not hold for the Navier–Stokes-Allen–Cahn model.

(c) It is well known that no maximum principle holds for the fourth order Cahn–Hilliard equation. Although L^∞-estimate is known [10] for the Cauchy problem of the Cahn–Hilliard equation, to the best of our knowledge, it is not clear if such an estimate still holds for the initial-boundary value problem for the Cahn–Hilliard equation, in particular, with the presence of a flow. Hence, throughout this paper, we do *not* assume any L^∞-estimate for φ.

(d). We emphasize that the constant C in (3.2)–(3.8) is independent of T.

The next lemma derives a priori estimates in higher norms for $(\mathbf{u},\varphi)$ under stronger assumptions on the initial data $(\mathbf{u}_0^\varepsilon,\varphi_0^\varepsilon)$.

LEMMA 3.2. *In addition to the assumptions of Lemma* 3.1*, suppose that* $\mathbf{u}_0^\varepsilon\in\mathbf{V}$*,* $\varphi_0^\varepsilon\in H^2(\Omega)$*, then every regular solution* $(\mathbf{u},\varphi)$ *of problem* (2.1)–(2.3) *satisfies the following estimates: for any* $T\in(0,\infty)$

$$\operatorname*{ess\,sup}_{t\in[0,T]}\|\Delta\varphi(t)\|_{L^2}^2+\gamma\int_0^T\left\|\Delta^2\varphi(s)\right\|_{L^2}^2\,ds\le c_1\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}},\tag{3.15}$$

$$\int_0^T\left[\,\|\varphi_t(s)+\mathbf{u}(s)\cdot\nabla\varphi(s)\|_{L^2}^2+\gamma^2\,\|\Delta w(s)\|_{L^2}^2\,\right]ds\le c_2\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}},\tag{3.16}$$

$$\int_0^T\|\varphi_t(s)\|_{L^2}^2\,ds\le c_3\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}}.\tag{3.17}$$

Moreover, when $d=2$ *there also hold the following additional estimates:*

$$\operatorname*{ess\,sup}_{t\in[0,T]}\|\nabla\mathbf{u}\|_{L^2}^2+\nu\int_0^T\|\Delta\mathbf{u}(s)\|_{L^2}^2\,ds\le c_4\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}},\tag{3.18}$$

$$\int_0^T\left[\,\|\mathbf{u}_t(s)\|_{L^2}^2+\|\nabla\overline{p}(s)\|_{L^2}^2\,\right]ds\le c_5\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}}.\tag{3.19}$$

Here $c_j=c_j(u_0,\varphi_0,\nu,g,\lambda,\gamma,T)$ *for* $j=1,2,3,4,5$ *are some positive constants.*

Proof. Testing (2.2) with $\Delta^2\varphi$ and using Young's inequality yields

$$\frac{d}{dt}\|\Delta\varphi\|_{L^2}^2+\gamma\left\|\Delta^2\varphi\right\|_{L^2}^2\le\frac{\gamma}{\varepsilon^4}\|\Delta f(\varphi)\|_{L^2}^2+\frac{1}{\gamma}\|\mathbf{u}\cdot\Delta\varphi\|_{L^2}^2.\tag{3.20}$$

By (3.2) and the interpolation inequality (cf. [1]):

$$\|\nabla\varphi\|_{L^\infty}\le C\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d}{6}}\|\nabla\varphi\|_{L^2}^{\frac{6-d}{6}}+C\|\nabla\varphi\|_{L^2},$$

we have

$$\begin{aligned}\|\mathbf{u}\cdot\Delta\varphi\|_{L^2}^2&\le\|\mathbf{u}\|_{L^2}^2\|\nabla\phi\|_{L^\infty}^2\\&\le C\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d}{3}}\|\nabla\varphi\|_{L^2}^{\frac{6-d}{3}}+C\|\nabla\varphi\|_{L^2}^2\\&\le\frac{\gamma}{4}\left\|\Delta^2\varphi\right\|_{L^2}^2+C.\end{aligned}\tag{3.21}$$

To estimate the first term on the right-hand side of (3.20), using the differential identity

$$\Delta f(\varphi) = f'(\varphi)\Delta\varphi + f''(\varphi)|\nabla\varphi|^2,$$

we have

$$\|\Delta f(\varphi)\|_{L^2} \le 3\,\|\varphi\|_{L^\infty}^2\,\|\Delta\varphi\|_{L^2} + 6\,\|\varphi\|_{L^\infty}\,\|\nabla\varphi\|_{L^4}^2\,.$$

The above inequality, (3.2), and the interpolation inequalities (cf. [1])

$$\|\varphi\|_{L^\infty} \le C\left(\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d}{2(12-d)}}\,\|\varphi\|_{L^6}^{\frac{3(8-d)}{2(12-d)}} + \|\varphi\|_{L^6}\right) \tag{3.22}$$

$$\|\Delta\varphi\|_{L^2} \le C\left(\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d}{6}}\,\|\nabla\varphi\|_{L^2}^{\frac{6-d}{6}} + \|\nabla\varphi\|_{L^2}\right) \tag{3.23}$$

$$\|\nabla\varphi\|_{L^4} \le C\left(\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d}{12}}\,\|\nabla\varphi\|_{L^2}^{\frac{12-d}{12}} + \|\nabla\varphi\|_{L^2}\right), \tag{3.24}$$

imply that

$$\|\Delta f(\varphi)\|_{L^2} \le C\Big(\left\|\Delta^2\varphi\right\|_{L^2}^{\frac{d(18-d)}{6(12-d)}} + 1\Big) \le \varepsilon^2\sqrt{\frac{\gamma}{8}}\left\|\Delta^2\varphi\right\|_{L^2} + C\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}}. \tag{3.25}$$

Now (3.15) follows immediately from combining (3.20), (3.21), and (3.25).

Applying the operator Δ to both sides of (2.9) we get

$$\Delta w = -\Delta^2\varphi + \frac{1}{\varepsilon^2}\Delta f(\varphi).$$

It then follows from the above equation, (3.15), and (3.25) that

$$\int_0^T \|\Delta w(s)\|_{L^2}^2\,ds \le C\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}},$$

which (2.14) in turn implies that

$$\int_0^T \|\varphi_t(s) + \mathbf{u}(s)\cdot\nabla\varphi(s)\|_{L^2}^2\,ds \le C\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}}.$$

Hence, (3.16) holds.

Clearly, (3.17) is a trivial consequence of (3.16) and (3.21). To show (3.18), we test (2.13) with $-\Delta\mathbf{u}$ to get

$$\frac{1}{2}\frac{d}{dt}\|\nabla\mathbf{u}\|_{L^2}^2 + \nu\,\|\Delta\mathbf{u}\|_{L^2}^2 = ((\mathbf{u}\cdot\nabla)\mathbf{u}, \Delta\mathbf{u})) + \lambda\big(\varphi\nabla w, \Delta\mathbf{u}\big) - (\mathbf{g}, \Delta\mathbf{u}). \tag{3.26}$$

Using (3.2) and interpolation inequalities, three terms on the right-hand side of (3.26) can be bounded as follows:

$$\begin{aligned}\big|((\mathbf{u}\cdot\nabla)\mathbf{u}, \Delta\mathbf{u})\big| &\le \|\Delta\mathbf{u}\|_{L^2}\,\|\mathbf{u}\|_{L^4}\,\|\nabla\mathbf{u}\|_{L^4} \\ &\le C\,\|\Delta\mathbf{u}\|_{L^2}^{\frac{4+d}{4}}\,\|\mathbf{u}\|_{L^2}^{\frac{4-d}{4}}\,\|\nabla\mathbf{u}\|_{L^2}\,, \qquad \text{(see [33])} \\ &\le \frac{\nu}{4}\,\|\Delta\mathbf{u}\|_{L^2}^2 + C\,\|\nabla\mathbf{u}\|_{L^2}^{2+\frac{2d}{4-d}}\,,\end{aligned} \tag{3.27}$$

$$\begin{aligned}\big|\lambda(\varphi\nabla w, \Delta\mathbf{u})\big| &\le \|\Delta\mathbf{u}\|_{L^2}\,\|\nabla w\|_{L^4}\,\|\phi\|_{L^4} \\ &\le C\,\|\Delta\mathbf{u}\|_{L^2}\left\{\|\Delta w\|_{L^2}^{\frac{d}{4}}\,\|\nabla w\|_{L^2}^{\frac{4-d}{4}} + \|\nabla w\|_{L^2}\right\} \\ &\le \frac{\nu}{4}\,\|\Delta\mathbf{u}\|_{L^2}^2 + C\,\|\Delta w\|_{L^2}^2 + C\,\|\nabla w\|_{L^2}^2\,,\end{aligned} \tag{3.28}$$

$$\big|(\mathbf{g}, \Delta\mathbf{u})\big| \le \frac{\nu}{4}\left\|\Delta^2\mathbf{u}\right\|_{L^2}^2 + C\,\|\mathbf{g}\|_{L^2}^2\,. \tag{3.29}$$

For $d = 2$, (3.18) now follows from applying the Gronwall's inequality to (3.26) after substituting (3.27)–(3.29) into it.

Testing (2.13) with $\mathbf{u}_t$ and utilizing (3.2), (3.28), (3.29), and (3.18) we obtain

$$\int_0^T \|\mathbf{u}_t(s)\|_{L^2}^2 \, ds \le C\,\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}},$$

which together with (2.13), (3.18), and (3.27) in turn implies that

$$\int_0^T \|\nabla \overline{p}(s)\|_{L^2}^2 \, ds \le C\,\varepsilon^{-\frac{2d(18-d)}{d^2-24d+72}}.$$

Hence, (3.19) holds. The proof is complete. □

Remark 3.2. (a) When $d = 3$, estimates (3.18) and (3.19) only hold local in time as is the case for the Navier–Stokes equations (cf. [33]). In addition, the difficulty of extending these estimates to all times is caused exactly by the nonlinear term $(\mathbf{u}\cdot\nabla)\mathbf{u}$, not by the nonlinear coupling terms.

(b) Unlike the Navier–Stokes-Allen–Cahn model for two-phase fluids (cf. [18]), the higher order norm estimates for the solution of the Navier–Stokes-Cahn–Hilliard model depend on ε^{-1} only *polynomially*, instead of *exponentially*. This important fact may give the possibility to derive a priori error estimates, which depend on ε^{-1} polynomially, for numerical solutions to the Navier–Stokes-Cahn–Hilliard model; see section 4 for further discussions.

4. Fully discrete finite element approximations and convergence analysis. In this section, we shall first give the weak formulation of the problem (2.1)–(2.5) based on the mixed (or split) setting using variables $(\mathbf{u}, \overline{p}, \varphi, w)$. We then introduce a family of fully discrete finite element methods based on this mixed (or split) weak formulation. The implicit Euler time-stepping will be used as a prototype scheme for time discretization and for presenting the idea of our convergence analysis. Essentially, any stable finite element for the Navier–Stokes equations can be used for the spatial discretizations of $\mathbf{u}$ and $\overline{p}$, and any of Ciarlet–Raviart family of mixed elements for the biharmonic operator can be used for the spatial discretizations of φ and w. The highlight of this section is to establish a discrete energy law, which mimics exactly the basic energy law (3.9) for the Navier–Stokes-Cahn–Hilliard phase field model, for the proposed finite element methods. Utilizing this discrete energy law we then show the convergence of the numerical solutions to the weak solution of (2.13), (2.14), (2.3), and (2.9) as $h, \tau \to 0$, and to the solution of its sharp interface model (1.1)–(1.4) as $h, \tau, \varepsilon \to 0$, provided that the phase field model converges to the sharp interface model (cf. Conjecture 4.1).

4.1. Weak formulation. The mixed weak formulation of (2.13), (2.14), (2.3), and (2.9) used in this paper is defined as follows: Find $(\mathbf{u}, \overline{p}, \varphi, w)$ such that

$$\begin{aligned}
&\mathbf{u} \in L^\infty((0,T);\mathbf{L}^2(\Omega)) \cap L^2((0,T);\mathbf{H}_0^1(\Omega)) \cap L^2((0,T);\mathbf{V}^*),\\
&\int_0^t \overline{p}(s)ds \in L^\infty((0,T);L_0^2(\Omega)),\\
&\varphi \in L^\infty((0,T);H^1(\Omega)) \cap H^1((0,T);H^{-1}(\Omega)),\\
&w \in L^2((0,T);H^1(\Omega)),
\end{aligned}$$

and (2.13) holds in the distribution sense. Moreover, for all $(\mathbf{v}, q, \psi, \chi) \in \mathbf{V} \times L_0^2(\Omega) \times H^1(\Omega) \times H^1(\Omega)$ there hold

$$\langle \mathbf{u}_t, \mathbf{v}\rangle + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + ((\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{v}) + \lambda(\varphi \nabla w, \mathbf{v}) = (\mathbf{g}, \mathbf{v}), \tag{4.1}$$

$$(\text{div } \mathbf{u}, q) = 0, \tag{4.2}$$

$$\langle \varphi_t, \psi\rangle - (\varphi \mathbf{u}, \nabla\psi) + \gamma(\nabla w, \nabla\psi) = 0, \tag{4.3}$$

$$(\nabla\varphi, \nabla\chi) + \frac{1}{\varepsilon^2}(f(\varphi), \chi) = (w, \chi), \tag{4.4}$$

with the initial conditions $\mathbf{u}(0) = \mathbf{u}_0^\varepsilon$ and $\varphi(0) = \varphi_0^\varepsilon$.

Remark 4.1. (a) We note that the second term on the left-hand side of (4.3) is obtained after performing an integration by parts to the coupling term. It turns out this simple step is quite important for the construction of the finite element methods which not only satisfy the discrete energy law but also fulfill the mass conservation law (see Lemma 4.2).

(b) The well-posedness of (4.1)–(4.4) can be proved by the standard techniques such as the Galerkin method using a priori estimates derived in the previous section (cf. [33]). In fact, as a by-product, our convergence result (see section 4.3) also provides an alternative and constructive proof of the existence of weak solutions.

4.2. Formulation of fully discrete finite element method. Let $J_\tau = \{t_m\}_{m=0}^M$ be a quasi-uniform partition of $[0, T]$ of mesh size $\tau := \frac{T}{M}$, and $d_t v^m := (v^m - v^{m-1})/\tau$. Let $\mathcal{T}_h$ be a quasi-uniform "triangulation" of the domain Ω of mesh size $h \in (0, 1)$ and $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$ ($K \in \mathcal{T}_h$ are tetrahedrons in the case $d = 3$). For a nonnegative integer r, let $P_r(K)$ denote the space of polynomials of degree less than or equal to r on K. We introduce the finite element spaces

$$\begin{aligned}
M_h &= \{q_h \in L_0^2(\Omega);\ q_h|_K \in P_0(K)\},\\
\mathbf{X}_h &= \{v_h \in \mathbf{C}^0(\overline{\Omega}) \cap \mathbf{H}_0^1(\Omega);\ \mathbf{v}_h|_K \in \mathbf{P}_2(K)\},\\
\mathbf{V}_h &= \{\mathbf{v}_h \in \mathbf{X}_h;\ (\text{div } \mathbf{v}_h, q_h) = 0 \quad \forall q_h \in M_h\},\\
Y_h &= \{\psi_h \in C^0(\overline{\Omega});\ \psi_h|_K \in P_r(K),\ r \ge 1\}.
\end{aligned}$$

It is well known that [9, 25] the P_2-P_0 mixed finite element space $(\mathbf{X}_h, M_h)$ is a stable pair for the Navier–Stokes equations since it satisfies the inf-sup condition

$$\sup_{\mathbf{v}_h\in\mathbf{X}_h} \frac{(\text{div } \mathbf{v}_h, q_h)}{\|\nabla \mathbf{v}_h\|_{L^2}} \ge c\,\|q_h\|_{L^2} \qquad \forall q_h \in M_h. \tag{4.5}$$

Remark 4.2. The above finite element spaces $(\mathbf{X}_h, M_h)$ are chosen for the convenience of presentation, our convergence analysis in fact holds for *any* stable mixed finite element pairing (X_h, M_h) for the Navier–Stokes equations. In addition, the convergence analysis also holds for stabilized finite element approximations of the Navier–Stokes equations (cf. [9]).

It is also well known [13, 34, 16, 19] that $Y_h \times Y_h$ is a stable pair for the biharmonic operator and there holds the inf-sup condition

$$\sup_{\psi_h\in Y_h} \frac{(\nabla\psi_h, \nabla\chi_h)}{\|\psi_h\|_{H^1}} \ge c\,\|\chi_h\|_{H^1} \qquad \forall \chi_h \in Y_h. \tag{4.6}$$

We now are ready to introduce our fully discrete mixed finite element methods for problem (2.1)–(2.5). Find $\{(\mathbf{u}_h^m, \overline{p}_h^m, \varphi_h^m, w_h)\}_{m=1}^M \in \mathbf{X}_h \times M_h \times Y_h \times Y_h$ such that

for all $(\mathbf{v}_h, q_h, \psi_h, \chi_h) \in \mathbf{X}_h \times M_h \times Y_h \times Y_h$

$$(d_t\mathbf{u}_h^m, \mathbf{v}_h) + \nu\big(\nabla\mathbf{u}_h^m, \nabla\mathbf{v}_h\big) + ((\mathbf{u}_h^m \cdot \nabla)\mathbf{u}_h^m, \mathbf{v}_h) + \frac{1}{2}(\mathbf{u}_h^m \mathrm{div}\ \mathbf{u}_h^m, \mathbf{v}_h) - \big(\overline{p}_h^m, \mathrm{div}\ \mathbf{v}_h\big) + \lambda\big(\varphi_h^m \nabla w_h^m, \mathbf{v}_h\big) = (\mathbf{g}(t_m), \mathbf{v}_h), \tag{4.7}$$

$$\big(\mathrm{div}\ \mathbf{u}_h^m, q_h\big) = 0, \tag{4.8}$$

$$(d_t\varphi_h^m, \psi_h) - \big(\varphi_h^m\mathbf{u}_h^m, \nabla\psi_h\big) + \gamma\big(\nabla w_h^m, \nabla\psi_h\big) = 0, \tag{4.9}$$

$$(\nabla\varphi_h^m, \nabla\chi_h) + \frac{1}{\varepsilon^2}\big(f_h^m, \chi_h\big) = \big(w_h^m, \chi_h\big), \tag{4.10}$$

with the initial conditions $\mathbf{u}_h^0 = \mathbf{u}_{0h}$, and $\varphi_h^0 = \varphi_{0h}$. Here

$$f_h^m = \frac{1}{4}\big\{|\varphi_h^m|^2 + |\varphi_h^{m-1}|^2 - 2\big\}\big\{\varphi_h^m + \varphi_h^{m-1}\big\}, \tag{4.11}$$

Remark 4.3. (a) The f_h^m factor in the above scheme can be replaced by $\widetilde{f}_h^m := (\varphi_h^m)^3 - \varphi_h^{m-1}$. It is not hard to check that the resulted scheme will still satisfy an almost same discrete energy law that (to be given in the next subsection) satisfied by the above scheme, *provided* that a mesh constraint on τ is met (cf. section 3 of [19]).

(b) The solvability of (4.7)–(4.8) can be verified by using a fixed point argument in finite dimensional spaces (cf. [33]) and the discrete energy law to be given in the next subsection.

4.3. Convergence analysis. Since the phase field model couples two sets of well-known equations, the Navier–Stokes equations and the Cahn–Hilliard equation, it should not be hard to derive a priori error estimates for the above fully discrete mixed finite element schemes using the standard techniques as presented in [6, 25, 26, 33] and in [14, 17, 16]. However, since these standard techniques use the Gronwall type arguments at the end, the anticipated error estimates will definitely depend on $\frac{1}{\varepsilon}$ *exponentially*! Such error estimates clearly are not informative and have no practical usefulness for small ε. We refer interested readers to [19, 20] for more discussions in this direction.

One way to overcome this difficulty is to derive better error estimates which only depend on $\frac{1}{\varepsilon}$ *polynomially*, the best situation one can expect. For the Cahn–Hilliard equation, which is (2.2) with $\mathbf{u} = 0$, such error estimates were obtained in [19, 20] using a nonstandard technique. The key idea of this technique is to make use of a spectrum estimate result for the linearized Cahn–Hilliard operator (cf. [4] and the reference therein). In order to adapt this technique for analyzing the scheme (4.1)–(4.4), one needs a similar spectrum estimate result for the linearized operator associated with the coupled system (2.1)–(2.3). Unfortunately, to our knowledge, such a desired spectrum estimate has not been proved in the literature, although it is believed to be true.

In this paper, we shall take a different approach to address the convergence. Instead of proving the convergence by first establishing a rate of convergence (i.e., an error estimate), we shall prove the convergence directly. As expected, the crux of carrying out such a proof is to derive uniform (in ε) a priori estimates for the numerical solutions; in particular, to establish a discrete energy law, which must mimic the basic energy law (3.9). It should be noted that not every numerical method will meet such a criterion. The goal of this subsection is to prove that the fully discrete finite element method proposed in section 4.2 indeed is one exception. We verify our claim in the

next lemma by establishing a discrete counterpart of the basic energy law (3.9) for the numerical scheme (4.7)–(4.8).

LEMMA 4.1. *Let* $(\mathbf{u}_h^m, \overline{p}_h^m, \varphi_h^m, w_h^m)$ *solves* (4.7)–(4.10)*, then there holds that*

$$\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}_h^\ell, \varphi_h^\ell) + \tau \sum_{m=1}^{\ell} \Big[\frac{\tau}{2} \|d_t \mathbf{u}_h^m\|_{L^2}^2 + \frac{\tau\lambda}{2} \|d_t \nabla \varphi_h^m\|_{L^2}^2 + \nu \|\nabla \mathbf{u}_h^m\|_{L^2}^2 + \lambda\gamma \|\nabla w_h^m\|_{L^2}^2 \Big] = \tau \sum_{m=1}^{\ell} (\mathbf{g}(t_m), \mathbf{u}_h^m) + \mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}_h^0, \varphi_h^0) \tag{4.12}$$

for all $0 \le \ell \le M$*. Here* $\mathcal{J}_{\varepsilon,\lambda}(\cdot,\cdot)$ *is defined by* (3.10)*.*

Proof. The desired estimate (4.12) follows from setting $\mathbf{v}_h = \mathbf{u}_h^m$ in (4.7), $q_h = \overline{p}_h^m$ in (4.8), $\psi_h = w_h^m$ in (4.9), $\chi_h = d_t \varphi_h^m$ in (4.10), adding the resulting equations, using the identities

$$\begin{aligned}
&(d_t \mathbf{u}_h^m, \mathbf{u}_h^m) = \frac{1}{2} \big\{ d_t \|\mathbf{u}_h^m\|_{L^2}^2 + \tau \|d_t \mathbf{u}_h^m\|_{L^2}^2 \big\},\\
&(d_t \nabla \mathbf{u}_h^m, \nabla \mathbf{u}_h^m) = \frac{1}{2} \big\{ d_t \|\nabla \mathbf{u}_h^m\|_{L^2}^2 + \tau \|d_t \nabla \mathbf{u}_h^m\|_{L^2}^2 \big\},\\
&((\mathbf{u}_h^m \cdot \nabla)\mathbf{u}_h^m, \mathbf{u}_h^m) + \frac{1}{2} (\mathbf{u}_h^m \mathrm{div}\ \mathbf{u}_h^m, \mathbf{u}_h^m) = 0,\\
&(d_t \varphi_h^m, f_h^m) = \frac{1}{4} d_t \big\| (\varphi_h^m)^2 - 1 \big\|_{L^2}^2 .
\end{aligned}$$

and applying the operator $\tau \sum_{m=1}^{\ell}$ to the combined equation. ∎

The discrete energy law immediately implies the following uniform (in ε) a priori estimates for $(\mathbf{u}_h^m, \overline{p}_h^m, \varphi_h^m, w_h^m)$.

LEMMA 4.2. *Let* $(\mathbf{u}_h^m, \overline{p}_h^m, \varphi_h^m, w_h^m)$ *solves* (4.7)–(4.10)*, and suppose that* $\mathbf{g} \in L^2((0,T); \mathbf{H}^{-1}(\Omega))$ *and there exists a positive constant* C_0 *such that* $\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}_0^\varepsilon, \varphi_0^\varepsilon) \le C_0$*, then there hold the following estimates:*

$$\int_\Omega \varphi_h^m \, dx = \int_\Omega \varphi_h^0, \qquad \text{for } m \ge 1, \tag{4.13}$$

$$\max_{0 \le m \le M} \big\{ \|\mathbf{u}_h^m\|_{L^2}^2 + \lambda \|\nabla \varphi_h^m\|_{L^2}^2 + \lambda \varepsilon^{-2} (F(\varphi_h^m), 1) \big\} \le C, \tag{4.14}$$

$$\sum_{m=1}^{M} \big[\big\|\mathbf{u}_h^m - \mathbf{u}_h^{m-1}\big\|_{L^2}^2 + \lambda \big\|\nabla \varphi_h^m - \nabla \varphi_h^{m-1}\big\|_{L^2}^2 \big] \le C, \tag{4.15}$$

$$\tau \sum_{m=1}^{M} \big[\nu \|\nabla \mathbf{u}_h^m\|_{L^2}^2 + \lambda\gamma \|\nabla w_h^m\|_{L^2}^2 \big] \le C, \tag{4.16}$$

$$\tau \sum_{m=1}^{M} \|d_t \mathbf{u}_h^m\|_{\mathbf{V}^*}^{\frac{12}{6+d}} \le C, \tag{4.17}$$

$$\max_{0 \le \ell \le M} \Big\| \tau \sum_{m=1}^{\ell} \overline{p}_h^m \Big\|_{L^2} \le C, \tag{4.18}$$

$$\tau \sum_{m=1}^{M} \|d_t \varphi_h^m\|_{H^{-1}}^2 \le C, \tag{4.19}$$

for some positive constant $C = C(\mathbf{g}, C_0)$*.*

Proof. Equation (4.13) follows from setting $\psi_h = 1$ in (4.9), and (4.14)–(4.19) are the immediate consequences of the discrete energy law (4.12).

To show (4.17), let P_h denote the L^2-projection operator from $\mathbf{L}^2(\Omega)$ to $\mathbf{V}_h$. For any $\mathbf{v} \in \mathbf{V}$ setting $\mathbf{v}_h = P_h\mathbf{v}$ in (4.7), using the stability property of P_h and an inverse inequality (cf. [13, 25]) we get

$$\begin{aligned}(d_t\mathbf{u}_h^m, \mathbf{v}) &= -\nu\big(\nabla\mathbf{u}_h^m, \nabla P_h\mathbf{v}\big) - \lambda\,(\varphi_h^m\nabla w_h^m, P_h\mathbf{v}) - ((\mathbf{u}_h^m\cdot\nabla)\mathbf{u}_h^m, P_h\mathbf{v})\\ &\qquad -\frac{1}{2}\,(\mathbf{u}_h^m \mathrm{div}\ \mathbf{u}_h^m, P_h\mathbf{v}) + (\mathbf{g}, P_h\mathbf{v})\\ &\le c\nu\,\|\nabla\mathbf{u}_h^m\|_{L^2}\,\|\nabla\mathbf{v}\|_{L^2} + \lambda\,\|\nabla w_h^m\|_{L^2}\,\|\varphi_h^m\|_{L^3}\,\|\mathbf{v}\|_{L^6}\\ &\qquad + \|\nabla\mathbf{u}_h^m\|_{L^2}\,\|\mathbf{u}_h^m\|_{L^3}\,\|\mathbf{v}\|_{L^6} + c\,\|\mathbf{g}\|_{H^{-1}}\,\|\nabla\mathbf{v}\|_{L^2}\\ &\le c\big\{\nu\,\|\nabla\mathbf{u}_h^m\|_{L^2} + \lambda\,\|\nabla w_h^m\|_{L^2} + \|\mathbf{g}\|_{H^{-1}}\big\}\,\|\nabla\mathbf{v}\|_{L^2}\\ &\qquad + \|\nabla\mathbf{u}_h^m\|_{L^2}^{\frac{6+d}{6}}\,\|\mathbf{u}_h^m\|_{L^2}^{\frac{6-d}{6}}\,\|\nabla\mathbf{v}\|_{L^2}\,.\end{aligned}$$

It follows from the above estimate and (4.14)–(4.16) that

$$\tau\sum_{m=1}^{M}\|d_t\mathbf{u}_h^m\|_{V^*}^{\frac{12}{6+d}} \le C.$$

Hence, (4.17) holds.

To show (4.18), we apply the operator $\tau\sum_{m=1}^{\ell}$ to (4.7) to get

$$\begin{aligned}\left(\tau\sum_{m=1}^{\ell}\overline{p}_h^m, \mathrm{div}\mathbf{v}_h\right) &= \big(\mathbf{u}_h^\ell - \mathbf{u}_h^0, \mathbf{v}_h\big) + \nu\left(\tau\sum_{m=1}^{\ell}\nabla\mathbf{u}_h^m, \nabla\mathbf{v}_h\right)\\ &\quad + \left(\tau\sum_{m=1}^{\ell}\left[(\mathbf{u}_h^m\cdot\nabla)\mathbf{u}_h^m + \frac{1}{2}\mathbf{u}_h^m\mathrm{div}\mathbf{u}_h^m\right], \mathbf{v}_h\right)\\ &\quad + \lambda\left(\tau\sum_{m=1}^{\ell}\varphi_h^m\nabla w_h^m, \mathbf{v}_h\right) - \left(\tau\sum_{m=1}^{\ell}\mathbf{g}(t_m), \mathbf{v}_h\right).\end{aligned}$$

It then follows from (4.14), (4.16), (3.12), (3.13), and the Sobolev inequality (cf. [1]) that

$$\left(\tau\sum_{m=1}^{\ell}\overline{p}_h^m, \mathrm{div}\mathbf{v}_h\right) \le C\,\|\mathbf{v}_h\|_{H^1} \qquad \forall\mathbf{v}_h \in \mathbf{X}_h. \tag{4.20}$$

Hence, (4.18) is an immediate consequence of (4.20) and the inf-sup inequality (4.5).

Finally, to show (4.19), for any $\psi \in H_0^1(\Omega)$ setting $\psi_h = Q_h\psi$ in (4.9), where Q_h denotes the L^2-projection from $L^2(\Omega)$ to Y_h, and using the stability property of the L^2-projection (cf. [8]) we get

$$\begin{aligned}(d_t\varphi_h^m, \psi) &= -\gamma\,(\nabla w_h^m, \nabla Q_h\psi) - (\mathbf{u}_h^m\cdot\nabla\varphi_h^m, \psi)\\ &\le c\gamma\,\|\nabla w_h^m\|_{L^2}\,\|\nabla\psi\|_{L^2} + \|\mathbf{u}_h^m\cdot\nabla\varphi_h^m\|_{L^{\frac{6}{5}}}\,\|\psi\|_{L^6}\\ &\le c\big\{\gamma\,\|\nabla w_h^m\|_{L^2} + \|\nabla\mathbf{u}_h^m\|_{L^2}\,\|\nabla\varphi_h^m\|_{L^2}\big\}\,\|\nabla\psi\|_{L^2}\,.\end{aligned}$$

Now, (4.19) follows immediately from the above estimate and (4.14)–(4.16). The proof is complete. ∎

Remark 4.4. The property (4.13) says that our numerical methods preserve the mass conservation law of the phase field model (cf. (3.1)). This property will be further validated numerically in section 5. We remark that such a mass conservation law does not hold for the Navier–Stokes-Cahn–Hilliard phase field model, nor does it for its numerical approximations developed in [18].

Let $(\mathbf{U}_{\varepsilon,h,\tau}(x,t), \Phi_{\varepsilon,h,\tau}(x,t))$ denote the piecewise linear interpolation (in t) of the fully discrete solution $\{(\mathbf{u}_h^m, \varphi_h^m)\}$, that is,

$$\mathbf{U}_{\varepsilon,h,\tau}(\cdot,t) := \frac{t-t_{m-1}}{\tau}\mathbf{u}_h^m(\cdot) + \frac{t_m-t}{\tau}\mathbf{u}_h^{m-1}(\cdot) \quad \forall t \in [t_{m-1}, t_m], \tag{4.21}$$

$$\Phi_{\varepsilon,h,\tau}(\cdot,t) := \frac{t-t_{m-1}}{\tau}\varphi_h^m(\cdot) + \frac{t_m-t}{\tau}\varphi_h^{m-1}(\cdot) \quad \forall t \in [t_{m-1}, t_m], \tag{4.22}$$

for $1 \le m \le M$, and let $\overline{P}_{\varepsilon,h,\tau}(x,t)$, $\overline{\mathbf{U}}_{\varepsilon,h,\tau}(x,t)$, $\overline{\Phi}_{\varepsilon,h,\tau}(x,t)$, and $\overline{W}_{\varepsilon,h,\tau}(x,t)$ denote the piecewise constant extensions of $\{\overline{p}_h^m\}$, $\{\mathbf{u}_h^m\}$, $\{\varphi_h^m\}$, and $\{w_h^m\}$, respectively. That is,

$$\overline{P}_{\varepsilon,h,\tau}(\cdot,t) := \overline{p}_h^m \qquad \forall t \in [t_{m-1}, t_m], \quad 1 \le m \le M, \tag{4.23}$$

$$\overline{\mathbf{U}}_{\varepsilon,h,\tau}(\cdot,t) := \mathbf{u}_h^m \qquad \forall t \in [t_{m-1}, t_m], \quad 1 \le m \le M, \tag{4.24}$$

$$\overline{\Phi}_{\varepsilon,h,\tau}(\cdot,t) := \varphi_h^m \qquad \forall t \in [t_{m-1}, t_m], \quad 1 \le m \le M, \tag{4.25}$$

$$\overline{W}_{\varepsilon,h,\tau}(\cdot,t) := w_h^m \qquad \forall t \in [t_{m-1}, t_m], \quad 1 \le m \le M. \tag{4.26}$$

We remark that $\mathbf{U}_{\varepsilon,h,\tau}(x,t)$ and $\Phi_{\varepsilon,h,\tau}(x,t)$ are continuous piecewise polynomial functions in space and time, $\overline{P}_{\varepsilon,h,\tau}(x,t)$, $\overline{\mathbf{U}}_{\varepsilon,h,\tau}(x,t)$, $\overline{\Phi}_{\varepsilon,h,\tau}(x,t)$, and $\overline{W}_{\varepsilon,h,\tau}(x,t)$ are right continuous at the nodes $\{t_m\}$.

The main result of this section is the following convergence theorem.

THEOREM 4.3. *Suppose the assumptions of Lemma* 4.2 *hold. For each fixed* $\varepsilon > 0$, *let* $(\mathbf{u}_*^\varepsilon, \overline{p}_*^\varepsilon, \varphi_*^\varepsilon, w_*^\varepsilon)$ *denote the unique solution of problem* (4.1)–(4.4), *and* $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$ *be defined as above. Then we have*

$$\lim_{h,\tau\to 0}\Big(\|\mathbf{U}_{\varepsilon,h,\tau} - \mathbf{u}_*^\varepsilon\|_{L^2(L^2)} + \|\Phi_{\varepsilon,h,\tau} - \varphi_*^\varepsilon\|_{L^2(L^2)} + \big\|\overline{W}_{\varepsilon,h,\tau} - w_*^\varepsilon\big\|_{L^2(L^2)} \Big) = 0, \tag{4.27}$$

$$\int_0^t \overline{P}_{\varepsilon,h,\tau}(s) \longrightarrow \int_0^t \overline{p}_*^\varepsilon(s)ds \qquad \text{weakly } \star \text{ in } L^\infty((0,T); L^2(\Omega)). \tag{4.28}$$

Proof. Since the proof is long, we divide it into three steps.

Step 1*: Extracting convergent subsequences.* The estimates of Lemma 4.2 immediately give the following (uniform in h, τ and ε) estimates:

$$\big\|\overline{\mathbf{U}}_{\varepsilon,h,\tau}\big\|_{L^\infty(L^2)} + \sqrt{\lambda}\,\big\|\nabla\overline{\Phi}_{\varepsilon,h,\tau}\big\|_{L^\infty(L^2)} + \varepsilon^{-1}\sqrt{\lambda}\,\Big\|\overline{\Phi}_{\varepsilon,h,\tau}^2 - 1\Big\|_{L^\infty(L^2)} \le C, \tag{4.29}$$

$$\sqrt{\nu}\,\big\|\nabla\overline{\mathbf{U}}_{\varepsilon,h,\tau}\big\|_{L^2(L^2)} + \sqrt{\lambda\gamma}\,\big\|\nabla\overline{W}_{\varepsilon,h,\tau}\big\|_{L^2(L^2)} \le C, \tag{4.30}$$

$$\Big\|\frac{\partial}{\partial t}\mathbf{U}_{\varepsilon,h,\tau}\Big\|_{L^{\frac{12}{6+d}}(V^*)} \le C, \tag{4.31}$$

$$\Big\|\int_0^t \overline{P}_{\varepsilon,h,\tau}(s)ds\Big\|_{L^\infty(L^2)} \le C, \tag{4.32}$$

$$\Big\|\frac{\partial}{\partial t}\Phi_{\varepsilon,h,\tau}\Big\|_{L^2(H^{-1})} \le C. \tag{4.33}$$

Then there exists a convergent subsequence of $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$ (still denote by the same notation) and a quadruple $(\mathbf{u}^\varepsilon, \overline{p}^\varepsilon, \varphi^\varepsilon, w^\varepsilon)$ such that

$$\begin{aligned}
&\mathbf{u}^\varepsilon \in L^\infty((0,T);\mathbf{L}^2(\Omega)) \cap L^2((0,T);\mathbf{H}_0^1(\Omega)) \cap H^1((0,T);\mathbf{V}^*),\\
&\varphi^\varepsilon \in L^\infty((0,T);H^1(\Omega)) \cap H^1((0,T);\mathbf{H}^{-1}(\Omega)),\\
&\int_0^t \overline{p}^\varepsilon(s)ds \in L^\infty((0,T);L_0^2(\Omega)),\\
&w^\varepsilon \in L^2((0,T);H^1(\Omega)),
\end{aligned}$$

and

$$\overline{\mathbf{U}}_{\varepsilon,h,\tau} \overset{h,\tau\searrow 0}{\longrightarrow} \mathbf{u}^\varepsilon \quad \begin{array}{l}\text{weakly}\star \text{ in } L^\infty((0,T);\mathbf{L}^2(\Omega)),\\ \text{strongly in } L^2((0,T);\mathbf{L}^2(\Omega)),\\ \text{weakly in } L^2((0,T);\mathbf{H}^1(\Omega)),\\ \text{weakly in } H^1((0,T);\mathbf{V}^*),\end{array} \tag{4.34}$$

$$\overline{\Phi}_{\varepsilon,h,\tau} \overset{h,\tau\searrow 0}{\longrightarrow} \varphi^\varepsilon \quad \begin{array}{l}\text{weakly}\star \text{ in } L^\infty((0,T);H^1(\Omega)),\\ \text{strongly in } L^2((0,T);L^2(\Omega)),\\ \text{weakly in } H^1((0,T);H^{-1}(\Omega)),\end{array} \tag{4.35}$$

$$\int_0^t \overline{P}_{\varepsilon,h,\tau}(s)ds \overset{h,\tau\searrow 0}{\longrightarrow} \int_0^t \overline{p}^\varepsilon(s)ds \quad \text{weakly } * \text{ in } L^\infty((0,T);L^2(\Omega)), \tag{4.36}$$

$$\overline{W}_{\varepsilon,h,\tau} \overset{h,\tau\searrow 0}{\longrightarrow} w^\varepsilon \quad \begin{array}{l}\text{weakly in } L^2((0,T);H^1(\Omega)),\\ \text{strongly in } L^2((0,T);L^2(\Omega)).\end{array} \tag{4.37}$$

From (4.15) we also have

$$\begin{aligned}
\big\|\mathbf{U}_{\varepsilon,h,\tau} - \overline{\mathbf{U}}_{\varepsilon,h,\tau}\big\|_{L^2(L^2)}^2 &= \sum_{m=1}^M \big\|\mathbf{u}_h^m - \mathbf{u}_h^{m-1}\big\|_{L^2}^2 \int_{t_{m-1}}^{t_m} \left(\frac{t-t_{m-1}}{\tau}\right)^2 dt\\
&= \frac{\tau}{3}\sum_{m=1}^M \big\|\mathbf{u}_h^m - \mathbf{u}_h^{m-1}\big\|_{L^2}^2 \overset{\tau\searrow 0}{\longrightarrow} 0,\\
\big\|\nabla(\Phi_{\varepsilon,h,\tau} - \overline{\Phi}_{\varepsilon,h,\tau})\big\|_{L^2(L^2)}^2 &= \frac{\tau}{3}\sum_{m=1}^M \big\|\nabla(\varphi_h^m - \phi_h^{m-1})\big\|_{L^2}^2 \overset{\tau\searrow 0}{\longrightarrow} 0.
\end{aligned}$$

Hence, the sequences $\{\mathbf{U}_{\varepsilon,h,\tau}\}$ and $\{\overline{\mathbf{U}}_{\varepsilon,h,\tau}\}$ converge to the same limit as $h,\tau \to 0$, so do the sequences $\{\Phi_{\varepsilon,h,\tau}\}$ and $\{\overline{\Phi}_{\varepsilon,h,\tau}\}$.

Step 2*: Passing to the limit.* We now want to pass to the limit in (4.7)–(4.10) and show that $(\mathbf{u}^\varepsilon, \overline{p}^\varepsilon, \varphi^\varepsilon, w^\varepsilon)$ is a weak solution of problem (4.1)–(4.4) with the initial values $\mathbf{u}^\varepsilon(0) = \mathbf{u}_0^\varepsilon$ and $\varphi^\varepsilon(0) = \varphi_0^\varepsilon$. To the end, we rewrite (4.7)–(4.10) as

$$\begin{aligned}
&((\mathbf{U}_{\varepsilon,h,\tau})_t, \mathbf{v}_h) + \nu\big(\nabla\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \nabla\mathbf{v}_h\big) + \big((\overline{\mathbf{U}}_{\varepsilon,h,\tau}\cdot\nabla)\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \mathbf{v}_h\big)\\
&\qquad + \frac{1}{2}\big(\overline{\mathbf{U}}_{\varepsilon,h,\tau}\mathrm{div}\ \overline{\mathbf{U}}_{\varepsilon,h,\tau}, \mathbf{v}_h\big) + \lambda\big(\overline{\Phi}_{\varepsilon,h,\tau}\nabla\overline{W}_{\varepsilon,h,\tau}, \mathbf{v}_h\big) = (\overline{\mathbf{g}}_\tau, \mathbf{v}_h),
\end{aligned} \tag{4.38}$$

$$\big(\mathrm{div}\ \overline{\mathbf{U}}_{\varepsilon,h,\tau}, q_h\big) = 0, \tag{4.39}$$

$$((\Phi_{\varepsilon,h,\tau})_t, \psi_h) - \big(\overline{\Phi}_{\varepsilon,h,\tau}\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \nabla\psi_h\big) + \gamma\big(\nabla\overline{W}_{\varepsilon,h,\tau}, \nabla\psi_h\big) = 0, \tag{4.40}$$

$$\big(\nabla\overline{\Phi}_{\varepsilon,h,\tau}, \nabla\chi_h\big) + \frac{1}{\varepsilon^2}\big(\overline{f}_{\varepsilon,h,\tau}, \chi_h\big) = \big(\overline{W}_{\varepsilon,h,\tau}, \chi_h\big), \tag{4.41}$$

for $(\mathbf{v}_h, q_h, \psi_h, \chi_h) \in \mathbf{V}_h \times M_h \times Y_h \times Y_h$. Here $\overline{f}_{\varepsilon,h,\tau}$ and $\overline{g}_\tau$ denotes the constant extensions of $\{f_h^m\}$ and $\{\mathbf{g}(t_m)\}$, respectively.

For any $\eta \in C^0[0,T]$, we multiply (4.38)–(4.41) by η, respectively, and integrate the resulting equations in t from 0 to T to get

$$\begin{aligned}(4.42)\quad &\int_0^T \big[((\mathbf{U}_{\varepsilon,h,\tau})_t, \mathbf{v}_h) + \nu\big(\nabla\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \nabla\mathbf{v}_h\big) + \big((\overline{\mathbf{U}}_{\varepsilon,h,\tau}\cdot\nabla)\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \mathbf{v}_h\big)\big]\eta(t)dt \\ &+ \int_0^T \left[\frac{1}{2}\left(\overline{\mathbf{U}}_{\varepsilon,h,\tau}\mathrm{div}\ \overline{\mathbf{U}}_{\varepsilon,h,\tau}, \mathbf{v}_h\right) + \lambda\big(\overline{\Phi}_{\varepsilon,h,\tau}\nabla\overline{W}_{\varepsilon,h,\tau}, \mathbf{v}_h\big)\right]\eta(t)dt \\ &\qquad = \int_0^T (\overline{\mathbf{g}}_\tau, \mathbf{v}_h)\eta(t)dt,\end{aligned}$$

$$(4.43)\quad \int_0^T \big(\mathrm{div}\ \overline{\mathbf{U}}_{\varepsilon,h,\tau}, q_h\big)\eta(t)dt = 0,$$

$$(4.44)\quad \int_0^T \big[((\Phi_{\varepsilon,h,\tau})_t, \psi_h) - \big(\overline{\Phi}_{\varepsilon,h,\tau}\overline{\mathbf{U}}_{\varepsilon,h,\tau}, \nabla\psi_h\big) + \gamma\big(\nabla\overline{W}_{\varepsilon,h,\tau}, \nabla\psi_h\big)\big]\eta(t)dt = 0,$$

$$(4.45)\quad \int_0^T \left[\big(\nabla\overline{\Phi}_{\varepsilon,h,\tau}, \nabla\chi_h\big) + \frac{1}{\varepsilon^2}\left(\overline{f}_{\varepsilon,h,\tau}, \chi_h\right)\right]\eta(t)dt = \int_0^T \left(\overline{W}_{\varepsilon,h,\tau}, \chi_h\right)\eta(t)dt.$$

For any $(\mathbf{v}, q, \psi, \chi) \in \mathbf{V} \times L_0^2(\Omega) \times H^1(\Omega) \times H^1(\Omega)$, let $(\mathbf{v}_h, q_h, \varphi_h, \chi_h) \in \mathbf{V}_h \times M_h \times Y_h \times Y_h$ denote the standard finite element (nodal) interpolations of $(\mathbf{v}, q, \psi, \chi)$ in (4.42)–(4.45). Since

$$\begin{aligned}\mathbf{v}_h &\xrightarrow{h\searrow 0} \mathbf{v} \quad \text{strongly in } \mathbf{H}_0^1(\Omega),\\ q_h &\xrightarrow{h\searrow 0} q \quad \text{strongly in } L_0^2(\Omega),\\ \psi_h &\xrightarrow{h\searrow 0} \psi \quad \text{strongly in } H^1(\Omega),\\ \chi_h &\xrightarrow{h\searrow 0} \chi \quad \text{strongly in } H^1(\Omega),\end{aligned}$$

setting $h, \tau \to 0$ in (4.42)–(4.45) and using(4.34)–(4.37) we get $\mathbf{u}^\varepsilon(0) = \mathbf{u}_0^\varepsilon$, $\varphi^\varepsilon(0) = \varphi_0^\varepsilon$, and

$$\begin{aligned}&\int_0^T \left[\langle \mathbf{u}_t^\varepsilon, \mathbf{v}\rangle + \nu\big(\nabla\mathbf{u}^\varepsilon, \nabla\mathbf{v}\big) + ((\mathbf{u}^\varepsilon\cdot\nabla)\mathbf{u}, \mathbf{v}) + \frac{1}{2}\left(\mathbf{u}^\varepsilon\mathrm{div}\ \mathbf{u}^\varepsilon, \mathbf{v}\right)\right]\eta(t)dt \\ &\qquad + \lambda\int_0^T \big(\varphi^\varepsilon\nabla w^\varepsilon, \mathbf{v}\big)\eta(t)dt = \int_0^T (\mathbf{g}, \mathbf{v})\,\eta(t)dt,\\ &\int_0^T (\mathrm{div}\ \mathbf{u}^\varepsilon, q)\,\eta(t)dt = 0,\\ &\int_0^T \big[\langle\varphi_t, \psi\rangle - \big(\varphi^\varepsilon\mathbf{u}^\varepsilon, \nabla\psi\big) + \gamma\big(\nabla w^\varepsilon, \nabla\psi\big)\big]\eta(t)dt = 0,\\ &\int_0^T \left[(\nabla\varphi^\varepsilon, \nabla\chi) + \frac{1}{\varepsilon^2}\big(f(\varphi^\varepsilon), \chi\big)\right]\eta(t)dt = \int_0^T (w^\varepsilon, \chi)\,\eta(t)dt,\end{aligned}$$

which is equivalent to (4.1)–(4.4) since $C^0[0,T]$ is dense in $L^2(0,T)$. In addition, it is easy see that $(\mathbf{u}^\varepsilon, \overline{p}^\varepsilon, \varphi^\varepsilon, w^\varepsilon)$ satisfies (2.13) in the distribution sense. Hence, $(\mathbf{u}^\varepsilon, \overline{p}^\varepsilon, \varphi^\varepsilon, w^\varepsilon)$ is a weak solution of (4.1)–(4.4).

Step 3*: Finishing up.* We have shown above that $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$ has a convergent subsequence and its limit $(\mathbf{u}^\varepsilon, \overline{p}^\varepsilon, \varphi^\varepsilon, w^\varepsilon)$ is a weak solution of (4.1)–(4.4). By the uniqueness, we have $\mathbf{u}^\varepsilon = \mathbf{u}^\varepsilon_*, \overline{p}^\varepsilon = \overline{p}^\varepsilon_*, \varphi^\varepsilon = \varphi^\varepsilon_*$, and $w^\varepsilon = w^\varepsilon_*$. Moreover, the proof also implies that the limit of *every* convergent subsequence of $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$ must be a weak solution of (4.1)–(4.4). Hence, the whole sequence $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$ converge to the unique weak solution $(\mathbf{u}^\varepsilon_*, \overline{p}^\varepsilon_*, \varphi^\varepsilon_*, w^\varepsilon_*)$. The proof is complete. ∎

Remark 4.5. We remark that since the phase field model (2.13), (2.14), (2.3), and (2.9) contains the Navier–Stokes equations, in general, one cannot expect any better results than those for the Navier–Stokes equations. Hence, the uniqueness assumption on the solution $(\mathbf{u}^\varepsilon_*, \overline{p}^\varepsilon_*, \varphi^\varepsilon_*, w^\varepsilon_*)$ can only be justified when $d = 2$. For the case $d = 3$, it can be shown that (2.13), (2.14), (2.3), and (2.9) has a unique local-in-time classical solution, hence, we assume that the uniqueness is understood local-in-time when $d = 3$. Clearly, without the uniqueness assumption *Step* 3 of the above proof does not stand anymore, hence, the convergence stated in Theorem 4.3 only holds for a subsequence, instead of the whole sequence, of $\{(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})\}$.

We now recall that the following convergent result was conjectured in [27, 30], and we also believe it should be true.

CONJECTURE 4.1. *Assume that the sharp interface problem* (1.1)–(1.4) *has a unique regular solution* (u_*, p_*)*. Under the assumptions of Theorem* 4.3 *there hold*

$$\lim_{\varepsilon\to 0} \|\mathbf{u}^\varepsilon_* - u_*\|_{L^2(L^2)} = 0, \tag{4.46}$$

$$\int_0^t \overline{p}^\varepsilon_*(s)\,ds \overset{\varepsilon\searrow 0}{\longrightarrow} \int_0^t p_*(s)\,ds \qquad \textit{weakly} * \textit{ in } L^\infty((0,T); L^2(\Omega)), \tag{4.47}$$

$$\varphi^\varepsilon_* \overset{\varepsilon\searrow 0}{\longrightarrow} \pm 1 \qquad \textit{a.e. in } \Omega_t^\pm \times (0,T). \tag{4.48}$$

Here Ω_t^+ *and* Ω_t^- *denote the outside and inside of* Γ_t *in* Ω *at time* t*, respectively.*

An immediate consequence of Theorem 4.3 and Conjecture 4.1 is the following convergence theorem.

THEOREM 4.4. *Under the assumptions of Conjecture* 4.1 *there hold*

$$\lim_{\varepsilon\to 0} \lim_{h,\tau\to 0} \|U_{\varepsilon,h,\tau} - u_*\|_{L^2(L^2)} = 0, \tag{4.49}$$

$$\int_0^t \overline{P}_{\varepsilon,h,\tau}(s)\,ds \overset{\varepsilon\searrow 0}{\longrightarrow} \int_0^t p_*(s)\,ds \qquad \textit{weakly} * \textit{ in } L^\infty((0,T); L^2(\Omega)), \tag{4.50}$$

$$\Phi_{\varepsilon,h,\tau} \overset{\varepsilon,h,\tau\searrow 0}{\longrightarrow} \pm 1 \qquad \textit{a.e. in } \Omega_t^\pm \times (0,T). \tag{4.51}$$

Remark 4.6. (a) The convergence result of Theorem 4.3 essentially guarantees that the numerical solution $(\mathbf{U}_{\varepsilon,h,\tau}, \overline{P}_{\varepsilon,h,\tau}, \Phi_{\varepsilon,h,\tau}, \overline{W}_{\varepsilon,h,\tau})$ enjoys the same kind convergence to the solution $(\mathbf{u}_*, p_*)$ of the sharp interface problem (1.1)–(1.4) as the phase field solution $(\mathbf{u}^\varepsilon_*, \overline{p}^\varepsilon_*, \varphi^\varepsilon_*, w^\varepsilon_*)$ of (2.13), (2.14), (2.3), and (2.9) does.

(b) We remark that the analogue of convergence result (4.50) for the pressure does not hold for the Navier–Stokes-Allen–Cahn model (cf. [18]). However, this result holds for the Navier–Stokes-Cahn–Hilliard model due to the uniform (in ε) estimate (4.18).

5. Numerical experiments. In this section we provide some 2-D numerical experiments to gauge the fully discrete finite element method developed in the previous sections. In addition, our numerical results reveal some interesting features such

as shrinking, splitting, and merging of fluid interfaces governed by the phase field model (2.13), (2.14), (2.3), and (2.9). In all numerical experiments to be given in the following, we choose $\Omega = [-0.4, 0.4]^2$, $\varepsilon = 10^{-2}$, $\nu = 1$, $\lambda = \gamma = 0.1$, $\mathbf{g} = (1, 0)^t$, $u_0^\varepsilon \equiv 0$, while the initial condition for ϕ is specified in each test. Also, in order to resolve the diffuse interface, we use $\tau = 10^{-5}$ and unstructured spatial meshes with the minimum triangle size $h = 10^{-4}$ in all experiments.

Test 1. In this test, we take the following initial condition for φ:

$$\varphi_0^\varepsilon(x) = \tanh\left(\frac{x_1^2}{0.01} + \frac{x_2^2}{0.0225} - 1\right).$$

Note that the zero level set of φ_0^ε, which gives the initial fluid interface, is the ellipse $\frac{x_1^2}{0.01} + \frac{x_2^2}{0.0225} = 1$. Hence, we have the situation of one elliptical fluid bubble inside another fluid.

Figure 5.1 shows snapshots of color and zero-level set plots of the computed phase function ϕ_h^m at six time steps. In the figure, the red color stands for $\phi_h^m = 1$, the blue color stands for $\phi_h^m = -1$, and the black curve represents the zero-level set of the computed phase function. We notice that the elliptical bubble quickly deforms into a circular bubble while the total mass (the integral of ϕ_h^m over Ω) remains constant in time. In the test, we have

$$\int_\Omega \phi_h^m \, dx \equiv 0.54538 \qquad \text{for } m = 1, 2, \ldots, M.$$

The shape and size of the circular bubble remains unchanged, it should eventually be stabilized (i.e., converge to a stationary solution) due to the dissipative mechanism of the phase field model (2.13), (2.14), (2.3), and (2.9) (cf. Lemma 3.1). We also remark that the interface (zero-level set of φ_h^m) movement is very similar to that of the zero-level set of the solution to the Cahn–Hilliard equation (the equation obtained by setting $u \equiv 0$ in (2.2)) (cf. [21, 22]). As expected, here the zero-level set is pushed to the right by the fluid flow (through the convective term $u \cdot \nabla \phi$) while it is approaching the equilibrium state.

Figure 5.2 displays snapshots of the arrow and streamline plots of the computed velocity field u_h^m at six time steps. The black ellipse in the center of each snapshot stands for the initial fluid interface (i.e., the zero-level set of ϕ_0^ε). We notice that fluid vertices are formed shortly after the initial time step, and the vertices become stronger as the time goes on.

Test 2. In this test, the initial profile of the phase function is taken as

$$\varphi_0^\varepsilon(x) = \tanh\left(\frac{1}{\varepsilon}\left(\frac{x_1^2}{0.0064} + \frac{x_2^2}{0.0225} - 1\right)\left(\frac{x_1^2}{0.0225} + \frac{x_2^2}{0.0064} - 1\right)\right).$$

Note that the zero level set of φ_0^ε, which gives the initial fluid interface, is the union of the following two intersecting ellipses: $\frac{x_1^2}{0.0064} + \frac{x_2^2}{0.0225} = 1$ and $\frac{x_1^2}{0.0225} + \frac{x_2^2}{0.0064} = 1$, which enclose four bullethead-like bubbles inside a fluid.

Figure 5.3 shows snapshots of color and zero-level set plots of the computed phase function φ_h^m at fifteen time steps. Again, the red color stands for $\varphi_h^m = 1$, the blue color stands for $\varphi_h^m = -1$, and the black curve represents the zero-level set of the computed phase function. In this test, we see the fluid bubble first splits into four bubbles, they then deform into four circular bubbles, and finally merge to form a

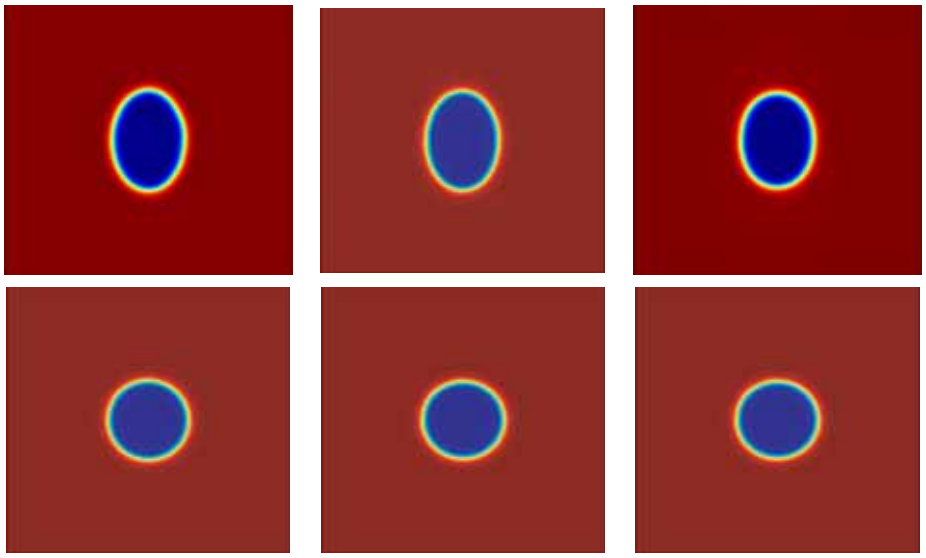


FIG. 5.1. *Color and zero-level set plots of computed phase function* φ_h^m *at* $t_m = 10^{-7}$, $1.1 \times 10^{-6}, 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}$. *The graphs are arranged row-wise.*

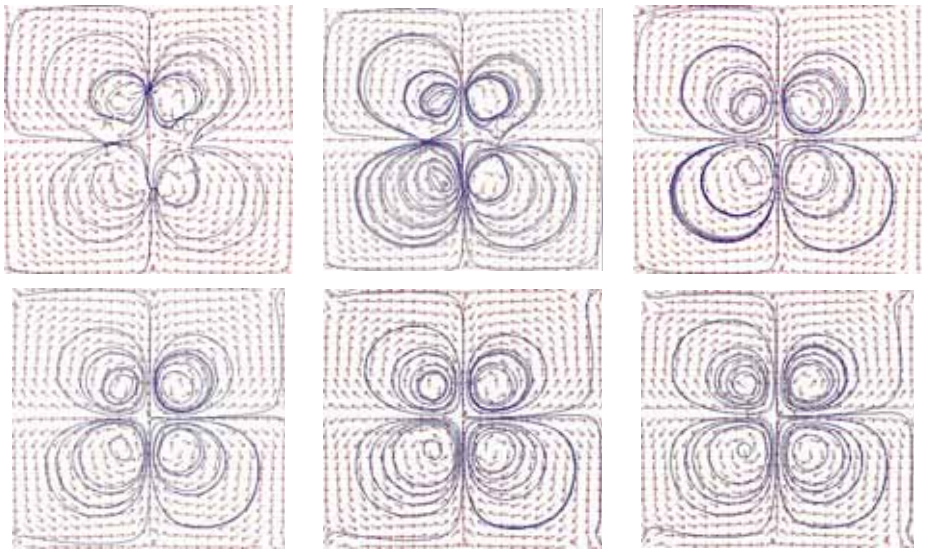


FIG. 5.2. *Arrow and streamline plots of computed velocity field* u_h^m *at* $t_m = 10^{-7}, 1.1 \times 10^{-6}$, $5 \times 10^{-5}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}$. *The graphs are arranged row-wise.*

bigger circular bubble which eventually stabilizes. During the evolution, the total mass (the integral of ϕ_h^m over Ω) remains constant in time. In the test, we have

$$\int_\Omega \phi_h^m \, dx \equiv 0.58313 \qquad \text{for } m = 1, 2, \ldots, M.$$

As expected, the interface (zero-level set of φ_h^m) movement is very similar to that of the zero-level set of the solution to the Cahn–Hilliard equation (the equation obtained

by setting $u \equiv 0$ in (2.2)) (cf. [21, 22]), and it is pushed very slowly off the center to the right by the fluid flow (through the convective term $u \cdot \nabla\varphi$). Another noticeable difference is that, unlike the dynamics of the zero-level set of the solution to the Cahn–Hilliard equation, here the four bullethead-like bubbles seem to evolve at slightly different speed and the bottom bubble disappears a couple of time steps earlier than the top one, which in turn is taken a couple of time steps earlier than the left bubble. We think that this phenomenon is caused by the fluid flow through the convective term $u \cdot \nabla\varphi$.

Figure 5.4 displays snapshots of the arrow and streamline plots of the computed velocity field u_h^m at nine time steps. The black ellipses in the center of each snapshot stand for the initial fluid interface (i.e., the zero-level set of φ_0^ε). We notice that fluid vertices are formed shortly after the initial time step, and more vertices are produced as the time goes on.

Acknowledgment. The author would like to thank the referees for their helpful comments and suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic press, New York, 1975.
[2] D. M. Anderson and G. B. McFadden, *A diffuse-interface description of internal waves in a near-critical fluid*, Phys. Fluids, 9 (1997), pp. 1870–1879.
[3] D. M. Anderson, G. B. McFadden, and A. A. Wheeler, *Diffuse-interface methods in fluid mechanics*, Ann. Rev. Fluid Mech., 30 (1998), pp. 139–165.
[4] N. D. Alikakos, P. W. Bates, and X. Chen, *Convergence of the Cahn–Hilliard equation to the Hele–Shaw model*, Arch. Rational Mech. Anal., 128 (1994), pp. 165–205.
[5] J. W. Barrett, X. Feng, and A. Prohl, *Convergence of a fully discrete finite element approximation of an Ericksen-Leslie model for the flow of liquid crystals*, Numer. Math., submitted.
[6] J. Bercovier and O. Pironneau, *Error estimates for finite element solution of the Stokes problem in the primitive variables*, Numer. Math., 33 (1979), pp. 211–224.
[7] J. Bear, *Dynamics of Fluids in Porous Media*, Dover Publications, Inc., New York, 1972.
[8] J. H. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–576.
[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[10] L. A. Caffarelli and N. E. Muler, *An L^∞ bound for solutions of the Cahn–Hilliard equation*, Arch. Rational Mech. Anal., 133 (1995), pp. 129–144.
[11] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[12] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.
[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[14] Q. Du and R. A. Nicolaides, *Numerical analysis of a continuum model of phase transition*, SIAM J. Numer. Anal., 28 (1991), pp. 1310–1322.
[15] D. A. Edwards, H. Brenner, and D. T. Wasan, *Interfacial Transport Process and Rheology*, Butterworths/Heinemann, London, 1991.
[16] C. M. Elliott, D. A. French, and F. A. Milner, *A second order splitting method for the Cahn–Hilliard equation*, Numer. Math., 54 (1989), pp. 575–590.
[17] C. M. Elliott and Z. Songmu, *On the Cahn–Hilliard equation*, Arch. Rational Mech. Anal., 96 (1986), pp. 339–357.
[18] X. Feng, Y. he, and C. Liu, *Analysis of finite element approximations of a phase field model for two-phase fluids*, Math. Comp., accepted.
[19] X. Feng and A. Prohl, *Error analysis of a mixed finite element method for the Cahn–Hilliard equation*, Numer. Math., 99 (2004), pp. 47–84.
[20] X. Feng and A. Prohl, *Numerical analysis of the Cahn–Hilliard equation and approximation of the Hele–Shaw problem*, Interfaces Free Bound., 7 (2005), pp. 1–28.

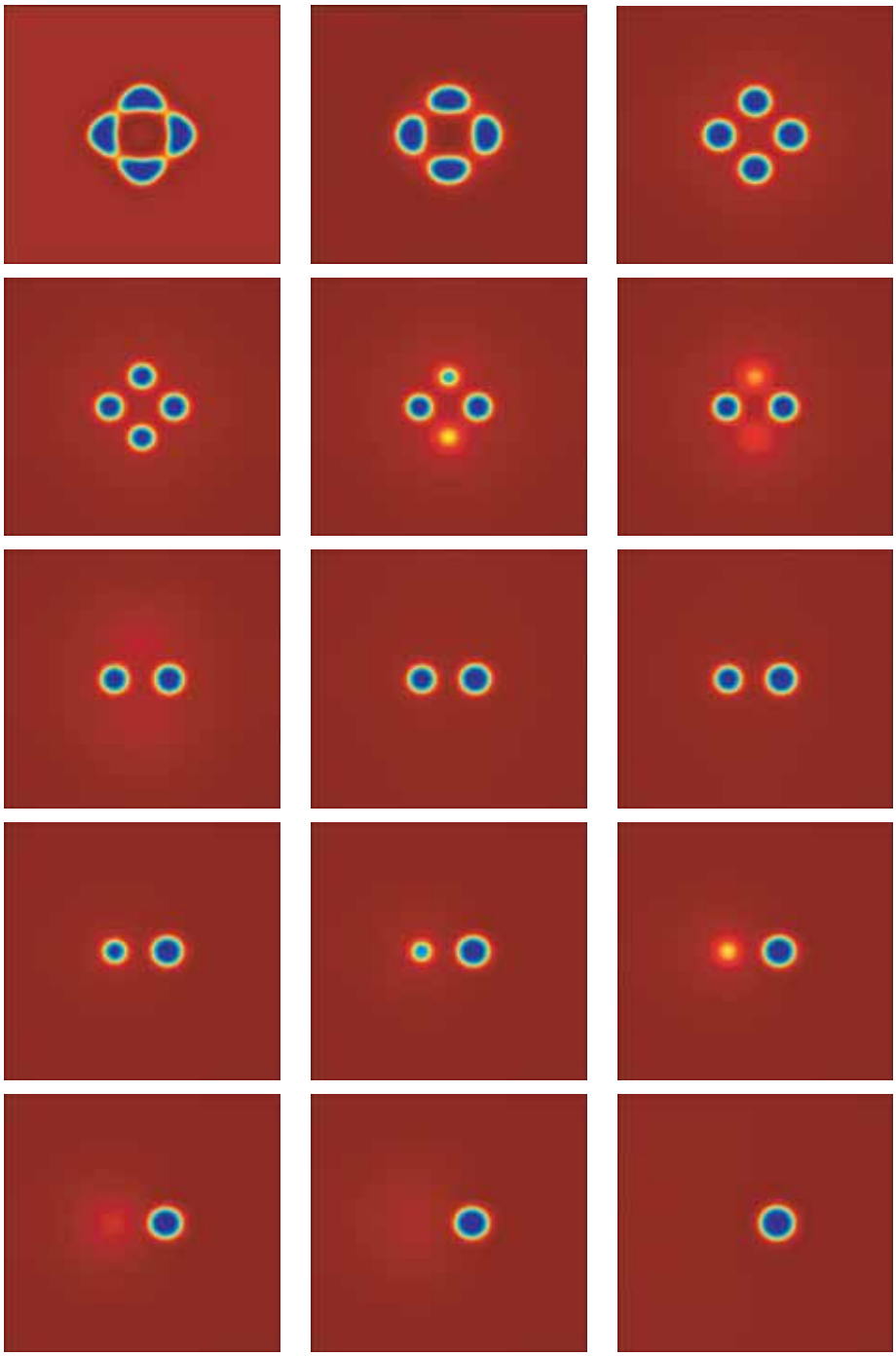


FIG. 5.3. *Color and zero-level set plots of computed phase function* φ_h^m *at* $t_m = 10^{-7}, 10^{-6}, 10^{-5}, 3\times10^{-5}, 4.2\times10^{-5}, 4.3\times10^{-5}, 4.5\times10^{-5}, 5.5\times10^{-5}, 5.8\times10^{-5}, 7\times10^{-5}, 7.5\times10^{-5}, 7.6\times10^{-5}, 7.8\times10^{-5}, 8\times10^{-5}, 9\times10^{-5}$. *The graphs are arranged row-wise.*

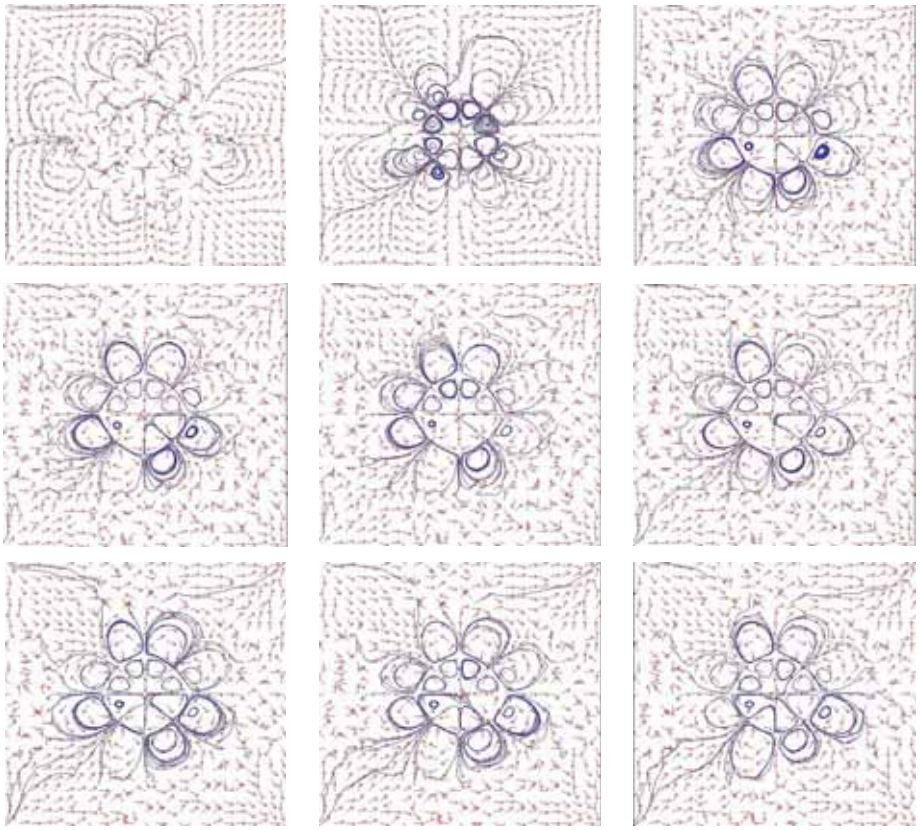


FIG. 5.4. *Arrow and streamline plots of computed velocity field* u_h^m *at* $t_m = 10^{-7}, 10^{-6}, 10^{-5}, 3\times10^{-5}, 4\times10^{-5}, 5\times10^{-5}, 7\times10^{-5}, 8\times10^{-5}, 9\times10^{-5}$. *The graphs are arranged row-wise.*

[21] X. FENG AND A. PROHL, *Analysis of a fully discrete finite element method for the phase field model and approximation of its sharp interface limits*, Math. Comp., 73 (2003), pp. 541–567.

[22] X. FENG AND H. WU, *A posteriori error estimates for finite element approximations of the Cahn–Hilliard equation and the Hele-Shaw flow*, M^2AN, submitted.

[23] P. FIFE, *Dynamics of Internal Layers and Diffusive Interfaces*, SIAM, Philadelphia, 1988.

[24] G. FIX, *Phase field method for free boundary problems*, in Free Boundary Problems, A. Fasano and M. Primicerio, eds., Pitman, London, 1983, pp. 580–589.

[25] V. GIRAULT AND P. A. RAVIART, *Finite Element Method for Navier–Stokes Equations: Theory and algorithms*, Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[26] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[27] D. JACQMIN, *Calculation of two-phase Navier–Stokes flows using phase-field modeling*, J. Comput. Phys., 155 (1999), pp. 96–127.

[28] J. S. LANGER, *Models of patten formation in first-order phase transitions*, in Directions in Condensed Matter Physics, World Science Publishers, 1986, pp. 164–186.

[29] G. M. LIEBERMAN, *Second Order Parabolic Differential Equations*, World Scientific Publishing, Singapore, 1996.

[30] C. LIU AND J. SHEN, *A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method*, Phys. D, 179 (2003), pp. 211–228.

[31] J. LOWENGRUB AND I. TRUSKINOVSKY, *Quasi-incompressible Cahn–Hilliard fluids and topological transitions*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 2617–2654.

[32] G. B. MCFADDEN, *Phase-field models of solidification*, Contemp. Math., 306, 295 (2002), pp. 107–145.

[33] R. Temam, *Navier–Stokes Equations*, Theory and Numerical Analysis, AMS Chelasea Publishing, Providence, RI, 2001.
[34] R. Scholz, *A mixed method for 4th order problems using linear finite elements*, RAIRO Anal. Numér., 12 (1978), pp. 85–90.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1073–1094

FAST RECONSTRUCTION METHODS FOR BANDLIMITED FUNCTIONS FROM PERIODIC NONUNIFORM SAMPLING*

THOMAS STROHMER† AND JARED TANNER‡

Abstract. A well-known generalization of Shannon's sampling theorem states that a bandlimited function can be reconstructed from its periodic nonuniformly spaced samples if the effective sampling rate is at least the Nyquist rate. Analogous to Shannon's sampling theorem this generalization requires that an infinite number of samples be available, which, however, is never the case in practice. Most existing reconstruction methods for periodic nonuniform sampling yield very low order (often not even first order) accuracy when only a finite number of samples is given. In this paper we propose a fast, numerically robust, root-exponential accurate reconstruction method. The efficiency and accuracy of the algorithm is obtained by fully exploiting the sampling structure and utilizing localized Fourier analysis. We discuss applications in analog-to-digital conversion where nonuniform periodic sampling arises in various situations. Finally, we demonstrate the performance of our algorithm by numerical examples.

Key words. Shannon's sampling theorem, oversampling, nonuniform periodic sampling, analog-to-digital conversion, Gevrey regularity, uniform interleaved sampling

AMS subject classifications. 41A05, 41A30, 42C15, 65T50, 94A12, 94A20

DOI. 10.1137/040609586

1. Introduction.

The classical Shannon sampling theorem plays a crucial role in signal processing and communications, indicating how to transfer between analog signals and discrete sequences [26]. Shannon's sampling theorem states that if a function[1] belongs to the space of bandlimited functions B_σ, i.e.,

$$f(t) := \frac{1}{\sqrt{2\pi}} \int_{-\sigma}^{\sigma} e^{2\pi i w t} F(w) dw, \qquad F(w) \in L_0^2[-\sigma, \sigma], \tag{1.1}$$

then it can be recovered exactly from its equidistant samples

$$f(t) \equiv \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2\sigma}\right) \frac{\sin(2\pi\sigma t - \pi k)}{2\pi\sigma t - \pi k} := \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2\sigma}\right) \operatorname{sinc}(2\pi\sigma t - \pi k). \tag{1.2}$$

Shannon's sampling theorem assumes that an infinite number of samples is available, which is of course never true in practice. Truncation of the cardinal series (1.2) results in rather poor approximation of the original bandlimited signal, and the truncation error is of the unacceptable low order of $1/\sqrt{L}$, where L is the number of samples; cf. [26]. In the presence of noise or quantization errors convergence may even break down completely [4]. To avoid these problems in practice, one usually

*Received by the editors June 7, 2004; accepted for publication (in revised form) January 26, 2006; published electronically June 2, 2006. This work was partially supported by NSF DMS grant 0208568.

http://www.siam.org/journals/sinum/44-3/60958.html

†Department of Mathematics, University of California, Davis, CA 95616-8633 (strohmer@math.ucdavis.edu).

‡Department of Mathematics, University of California, Davis, CA 95616-8633. Current address: Department of Mathematics, University of Utah, Salt Lake City, UT (tanner@math.utah.edu). The work of this author was supported by NSF DMS VIGRE grant 0135345.

[1]Throughout this paper we will use lowercase letters to designate functions in the time domain, and uppercase for the Fourier transform of the same function.

resorts to oversampling of the signal, since this gives rise to vastly better convergence rates and as a result greater robustness to noise.

While oversampling is therefore desirable in practice, it is not always easily done in real world applications. For example, consider advanced wireless communication systems, where demand for data rate is steadily increasing, requiring communication systems that use transmission signals with a (baseband) bandwidth in the range of tens of megahertz up to one gigahertz (as in currently developed ultrawideband systems). Such a huge bandwidth necessitates very high sampling rates in the analog-to-digital conversion which puts enormous demands on the analog sampling devices. While it is possible to construct signal acquisition systems that sample a signal even at nanosecond scales with high precision, such devices become increasingly expensive. More specifically, a linear increase in precision of a sampling device often goes hand in hand with a superlinear increase in the costs of constructing such a device.

One possible way to remedy this problem is to combine several analog-to-digital converters (ADCs) with lower sampling rate to obtain one virtual sensor with high sampling rate. We describe this concept in more detail. A standard ADC uniformly (over)samples an analog signal (a continuous-time function) at rate T^{-1}, say, where T is the time between two successive samples. The so obtained discrete-time signal is then subject to quantization, and the quantized signal is further processed by a digital signal processor (DSP). Instead of using one ADC with sampling rate T^{-1} we could run N ADCs in parallel, each operating at the slower rate $(NT)^{-1}$. The sampling instances of the nth ADC are chosen at $\{(kN+n)T\}_{k\in\mathbb{Z}}$, $n=0,\dots,N-1$, so that the combined sampling instances are $\{kT\}_{k\in\mathbb{Z}}$, which is equivalent to the output of one ADC that operates at the N times higher rate T^{-1}.

Many companies such as Maxim (http://www.maxim-ic.com/appnotes.cfm/appnote_number/2094), Agilent Technologies (http://www.agilent.com/labs/news/2003features/fea_adc03.html), and Analog Devices have been developing or are currently developing such time-interleaved ADCs.

To give another concrete current example for the need of time-interleaved ADCs consider the 10 Gigabit Ethernet over copper standard (which is part of the IEEE 802.3 standard; see http://www.ieee802.org). There 4 Cat6 copper pairs are used, so 2.5 Gigabits/sec are transmitted. Since a 12-PAM code with error correction is used, this means the baud rate is about 800 MHz. At least 8 bits precision is necessary. The fastest ADC with the desired precision runs about 1/2 of that. Consequently 2 or 4 time-interleaved ADC channels must be used to achieve the required precision.

While a time-interleaved ADC structure obviously has its merits, it does not come without caveats. The coordination of the N ADCs has to be done with high precision, but in practice timing errors between the individual ADCs result in sampling sets of the form $\{kNT+T_n\}_{k\in\mathbb{Z}}$, where the T_n are distinct random timing shifts. In other words, the combined sampling set does not form a uniform sampling set but consists of nonuniformly shifted unions of uniform sampling points, which is often referred to as *periodic nonuniform sampling* or *bunched sampling* [2, 18]. This poses two problems: (i) How can we estimate the unknown shifts T_n? (ii) How can we reconstruct quickly and stably the original signal from its periodically nonuniformly spaced samples? In this paper we focus on the second question, with shift estimation discussed in [23]. While there are several algorithms in the literature that deal with the reconstruction of bandlimited signals from periodic nonuniform samples (see, e.g., [14, 18, 2, 25, 24, 5, 12, 15, 20]), none of these algorithms provide high order accuracy with respect to truncation error.

Sometimes it can be advantageous to deliberately perform nonuniform periodic sampling in connection with analog-to-digital conversion. For instance, in [20] the use of periodic nonuniform sampling is proposed to avoid noise coupling in a mixed-signal integrated circuit, which contains analog and digital signal processing circuits, as is the case for an ADC. After the analog input signal has been sent through the ADC, the digital output is processed further by a DSP. However, switching of the digital circuits generates noise that can couple into the analog signal path through so-called parasite signal paths. Such noise coupling distorts the analog signal, which degrades the signal-to-noise ratio at the input of the ADC.

To avoid this noise coupling, it is proposed in [20] to have the ADC acquire a group (bunch) of samples at high rate while the digital signal processor is inactive, and allow digital processing of the ADC output during a second phase when the ADC is not sampling. This reduces the noise coupled to the analog signal, since the DSP operates only during the second phase. As the final step, one has to convert the bunched samples to uniformly spaced samples. Practical restrictions in terms of available memory and tolerable time delay imply an upper limit on the number of samples that can be processed during the conversion from bunched to uniform samples. This sampling pattern is obviously a special case of the periodic nonuniform sampling pattern described in the previous paragraph, with the simplification that all $T_n - T_{n-1}$ are (nearly) equal, but with the difficulty that we have a potentially large gap between two clusters of samples. This large gap may cause some instabilities; therefore it is vital to have a numerical reconstruction algorithm that is robust to such large gaps.

In this paper we develop the first method for reconstruction of a bandlimited signal from its periodic nonuniformly spaced samples that achieves root-exponential accuracy from a finite number of samples. The proposed method is numerically robust, and since its computationally most expensive steps consist of fast Fourier transforms (FFTs), it is numerically very efficient.

The paper is organized as follows. In section 2 we review some results on oversampling and localization of functions and their Fourier transforms. The Gevrey class arises as a natural candidate space for compactly supported smooth filter functions in connection with oversampled bandlimited signals. In section 3 we briefly describe how these smooth filters correspond to a localized reconstruction, resulting in a root-exponential accurate, fast algorithm for recovering a bandlimited signal from its uniformly spaced oversampled values. This simple observation is integral for the derivation of the main algorithm for the case of periodic nonuniformly spaced samples; cf. section 4. Numerical simulations that demonstrate the performance of the proposed method are presented in section 5. Finally, section 6 contains our conclusion and an outlook of future research.

2. Oversampling and localization. As mentioned in the introduction, the formulation in (1.2) is unsuitable for practical applications, where only a finite number of samples is available, $\{f(k/2\sigma)\}_{|k|\le L}$. For truncated samples the error, classically referred to as the truncation error, is controlled by the atom's localization

$$\epsilon(t,L,T) := \left| f(t) - \frac{\sqrt{2\pi}}{2\sigma} \sum_{|k|\le L} f\left(\frac{k}{2\sigma}\right) \psi\left(t - \frac{k}{2\sigma}\right) \right| \le \frac{\sqrt{2\pi}}{2\sigma} \cdot \|f\|_{L_\infty} \sum_{|k|>L} \left| \psi\left(t - \frac{k}{2\sigma}\right) \right|. \tag{2.1}$$

In the case of the classical Shannon sampling theorem, the atom, $\psi(\tau) := \mathrm{sinc}(\tau)$, suffers from an unacceptably slow decay, $\lim_{\tau\to\infty} \psi(\tau) \sim 1/\tau$, resulting in a first order convergence rate while moving from the sample boundaries, $\pm L/2\sigma$, to the interior. Moreover, if the samples $f(k/(2\sigma))$ are replaced by noisy samples $f(k/(2\sigma))+\epsilon_k$, then the corresponding approximation via the cardinal series in (1.2) may differ significantly from $f(t)$; cf. [4].

To remedy these problems in applications, one usually introduces oversampling. Sampling a function in the time domain introduces a periodization in the associated Fourier dual space, where sampling rate $T^{-1} = 2\sigma$ corresponds to a 2σ periodization. In (1.2) the reproducing atom (time domain), $\mathrm{sinc}(2\pi\sigma t - \pi k)$, removes the periodization introduced by sampling, through the action of its associated filter (Fourier dual space), $\chi_{[-\sigma,\sigma]}$. For this critical, Nyquist sampling rate, $\mathrm{sinc}(\cdot)$ is the unique atom that can be used to remove the periodization. However, if the bandlimited signal is sampled at a faster rate, $T^{-1} := 2\sigma/r$, where $r < 1$, then the dual space periodization is increased to $2\sigma/r$, allowing a large family of reproducing filters. Specifically, any function satisfying[2]

$$\Psi(w) = \begin{cases} 1, & |w| \le \sigma, \\ 0, & |w| > \sigma(2-r)/r =: \Omega \\ \text{anything} & \text{else} \end{cases} \qquad \Rightarrow \quad \psi \in B_\Omega, \tag{2.2}$$

gives rise to a Shannon-type series expansion

$$f(t) \equiv \sqrt{2\pi}T \sum_{k=-\infty}^{\infty} f\,(kT)\,\psi\,(t-kT)\,. \tag{2.3}$$

For $r = 1$ the above filter reduces to $\chi_{[-\sigma,\sigma]}$ and the classical Shannon's sampling theorem, whereas for $r < 1$ a gap[3] is introduced between σ and $\sigma(2-r)/r$, allowing for a host of other reproducing filters, including those with a high degree of regularity; see Figure 2.1.

Asymptotically the atom's localization is reflected in the filter's smoothness; consequently, the filter's regularity controls the convergence rate of the truncation error (2.1). By constructing infinitely differentiable filters with precise regularity estimates, we obtain root-exponential accuracy for the approximation of a bandlimited signal, as the point to be approximated moves from the sampling boundary, $\pm LT$, to the interior; see subsection 2.1. Unfortunately, unlike classical finite regularity filters, such as the raised cosine, which have a closed form expression for their corresponding atoms, to the authors' knowledge, there is no known infinitely differentiable compactly supported filter whose atom allows an explicit closed form expression. Alternatively to approximating the atoms as proposed in [16, 22], we introduce and analyze a direct Fourier domain implementation that does not adversely affect the high resolution

[2]It has been noted in [11] that the reproducing property is somewhat less strict than as stated in (2.2), in that the filter need not be zero for all $|w| \ge (2-r)/r$. Rather, the reproducing property is satisfied if the filter is one for $|w| \le \sigma$ and zero at the points where the periodic extension of the signal's dual representation is nonzero. However, this added flexibility cannot increase the regularity of the filter or decrease its regularity constants and, as such, cannot improve the asymptotic convergence rate. Although this added flexibility can be used to increase the atom's immediate localization about the origin, it introduces substantial peaks away from the origin [11], decreasing the overall convergence rate.

[3]Filters which are nonzero for $\sigma < |w| < \Omega$ necessarily decrease noise less than the characteristic filter, $\chi_{[-\sigma,\sigma]}$, but the vastly improved convergence rate more than makes up for this modestly lower denoising which is overcome in other ways.

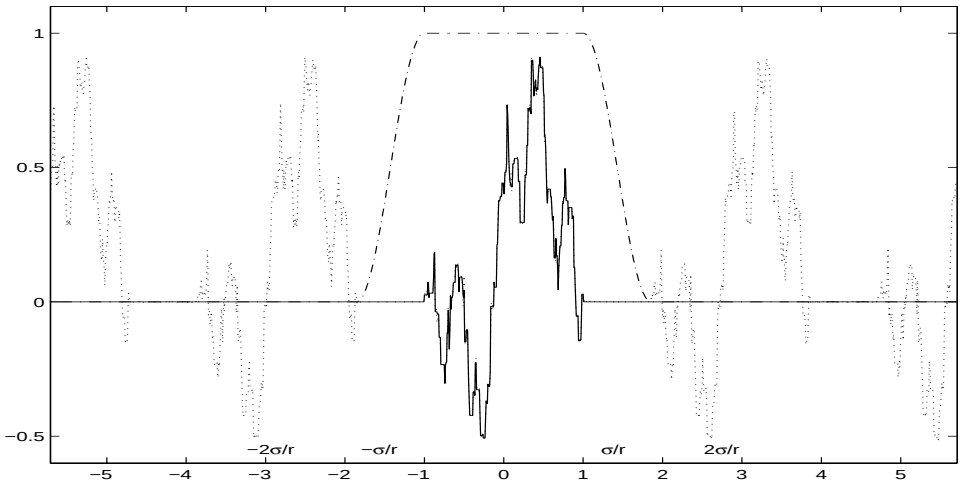


FIG. 2.1. *The signal's dual space representation (solid line) and the signal's periodization due to sampling at the rate, $T^{-1} := 2\sigma/r$, for $r < 1$ (dotted line). With the gap between periodization, a smooth filter (dot-dash line) can be used to remove the periodization introduced by sampling.*

achieved by smooth filters. The time domain localization of an atom, $\psi(\cdot)$, is reflected in the regularity[4] of its corresponding filter, $\Psi(\cdot)$,

$$|\psi(t)| \le (2\pi t)^{-s}\|\Psi\|_{C^s} \cdot \frac{2\Omega}{\sqrt{2\pi}} \quad \forall s, \qquad \psi \in B_\Omega, \tag{2.4}$$

where $\|\Psi\|_{C^s} := \|\Psi^{(s)}\|_{L^\infty}$.

Consequently, combining the bounds in (2.1) and (2.4), convergence is gained at the polynomial rate[5]

$$\epsilon(t, L, T) \le \text{Const} \cdot (LT - |t|)^{1-s}(2\pi)^{-s}\|\Psi\|_{C^s} \cdot \left(\frac{\Omega}{\sigma}\right), \quad s \ge 2, \tag{2.5}$$

as t passes from the boundary, $\pm LT$, to the interior, where $T := r/2\sigma$.

Rather than improving the atom's localization by increasing its corresponding filter's regularity, attempts have been made to construct highly localized atoms by maximizing the atom's local weight, $\int_{-R}^{R}\psi^2(t)dt/\int_{-\infty}^{\infty}\psi^2(t)dt$. However, such approaches have resulted in discontinuous filters [17] and atoms which do not decay globally [11]. A much more successful approach for polynomial order filters is to minimize the filter's regularity constant, $\|\Psi\|_{C^s}$. The classical raised cosine is such a filter [21]:

$$\Psi_{rc}(w) = \begin{cases} 1, & |w| \le \sigma, \\ 0, & |w| > \sigma(2-r)/r, \\ \frac{1}{2}(1 + \cos(\frac{\pi}{2}(\frac{r}{1-r})(\frac{w}{\sigma} - 1)), & \sigma < w < \sigma\frac{2-r}{r}, \\ \frac{1}{2}(1 + \cos(\frac{\pi}{2}(\frac{r}{1-r})(\frac{w}{\sigma} + 1)), & -\sigma > w > -\sigma\frac{2-r}{r}, \end{cases} \tag{2.6}$$

[4]This is achieved by s integrations by parts, where the derivatives are transferred onto the filter.

[5]If the underlying filter possesses $\|\Psi\|_{C^{s+1}} < \infty$, then the bound (2.5) can be tightened by one order of $(LT - |t|)$ to the rate $(LT - |t|)^{-s}$ [9].

where the bounded regularity constants are given explicitly by $\|\Psi\|_{C^1} = \frac{1}{2}(\frac{T}{2(1-r)})$ and $\|\Psi\|_{C^2} = \frac{1}{2}(\frac{T}{2(1-r)})^2$.

When a comparatively small number of sampling points is taken, low regularity constant polynomial order methods give extremely good reconstructions. However, when a larger number of samples is available, atoms with significantly improved asymptotic localization can be achieved by constructing infinitely regular filters, $\Psi \in C_0^\infty$. It should be noted that for infinitely differentiable functions, the optimal bound in (2.4) is not necessarily obtained for large s, as the regularity constant $\|\Psi\|_{C^s}$ grows rapidly in s. Rather, for functions where precise regularity estimates are known, the optimal s can be determined, resulting in an exponential decay without necessarily large constants. These claims will be realized in the numerical experiments presented in section 5, contrary to the assertion in [3], where it is claimed that the increased regularity does not improve numerical convergence. In a direct numerical comparison with the raised cosine filter, our infinitely differentiable filter (2.10) achieves dramatically superior convergence in the interior of the samples, and quantitatively similar errors near the sampling boundaries; see Figure 5.1.

2.1. Localization and Gevrey regularity. To achieve exponential accuracy and satisfy the reproducing condition, (2.2), requires a filter which is infinitely differentiable and compactly supported. The natural space for infinitely differentiable compactly supported functions is the Gevrey class which consists of functions satisfying the smoothness bound

$$\|h\|_{C^s} := \|h^{(s)}\|_{L^\infty} \leq \text{Const} \cdot \frac{(s!)^\alpha}{\eta_h^s} \Longleftrightarrow h \in G^\alpha, \tag{2.7}$$

where η_h is a constant independent of s. Incorporating the regularity information in the localization bound, (2.4), and minimizing over all admissible s, we conclude that Gevrey class filters satisfy a root-exponential localization decay,[6]

$$|\psi(t)| \leq \text{Const} \cdot \sqrt{|t|} \exp(-\alpha(2\pi\eta|t|)^{1/\alpha}), \quad \Psi \in G^\alpha, \tag{2.8}$$

and root-exponential truncation error

$$\epsilon(t, L, T) \leq \text{Const}_{\alpha,\eta} \exp(-(2\pi\eta(LT - |t|))^{1/\alpha}), \tag{2.9}$$

where $\text{Const}_{\alpha,\eta} \sim \eta^{-2} \sum_{l=0}^{q} q!\eta^{l/2}/(q-l)!$, with q the smallest integer greater than or equal to $(3\alpha - 2)/2$.

A similarly localized atom was constructed in [8, 16] by multiplying the sinc function with the inverse Fourier transform of an appropriately dilated G^2 function. Alternatively, such G^α filters can be expressed explicitly in the dual space, such as

$$\Psi_{G2}(w) = \begin{cases} 1, & |w| \leq \sigma, \\ 0, & |w| > \sigma(2-r)/r, \\ \rho\big(\frac{w-\sigma}{\sigma(2-r)/r}\big), & \sigma < w < \sigma\frac{2-r}{r}, \\ \rho\big(\frac{-w-\sigma}{\sigma(2-r)/r}\big), & -\sigma > w > -\sigma\frac{2-r}{r}, \end{cases} \tag{2.10}$$

where $\rho(w) := \exp[\beta(w-1)^{-1}\, e^{-1/w}] \ \in G^2$ [13].

[6] Compact support is inconsistent with analyticity, G^1, so reproducing atoms can at most be in the space G^α, for $\alpha > 1$, excluding true exponential decay, i.e., $\alpha = 1$, as was shown in the classical paper [1]. The Gevrey class of functions is essentially similar to ultradifferentiable functions [19].

Although the filter $\Psi_{G2}(\cdot)$ and the one in [16] result in rapid convergence while approaching the interior, $|t| \leq LT$, their associated atoms lack an explicit construction. As a result, to reconstruct a bandlimited signal at an arbitrary point has required the costly implementation of a quadrature evaluation, or a global approximation of the atom, such as the Padé and Gabor approximations proposed in [16] and [22], respectively. Alternatively, in the next section we introduce and analyze a direct Fourier domain implementation that does not adversely effect the high resolution achieved by smooth filters.

3. Dual space implementation for uniform oversampling. In this section we introduce an implementation which removes the sampling induced periodization through the direct action of the filter in the Fourier dual space. More specifically, if the bandlimited signal is sampled on the mesh $R := \{kT\}_{|k|\leq L}$, with $T := r/2\sigma$ and $r < 1$, we seek to compute an approximation to the signal on the refined mesh $P := \{kT/p\}_{|k|\leq pL}$, where $p \in \mathbb{N}/\{1\}$. This implementation is extremely efficient, as it only requires the FFT of the zero inserted signal, (3.2), and the pointwise multiplication in the dual space.

Define the approximation on the fine mesh as

$$Approx_\psi\ f\left(\frac{qT}{p}\right) := \sum_{|k|\leq L} f(kT)\psi\left[\left(\frac{q}{p}-k\right)T\right]. \tag{3.1}$$

We zero insert the samples from the coarse mesh to the fine mesh

$$f_o(x) := \begin{cases} f(x), & x \in R, \\ 0, & x \in P/R, \end{cases} \tag{3.2}$$

and note that the approximation in (3.1) is a discrete convolution,

$$Approx_\psi\ f(hq) = \sum_{|k|\leq pL} f_o(hk)\psi\left([q-k]\,h\right), \tag{3.3}$$

where $h := T/p$. To transfer the discrete convolution to pointwise multiplication in the dual space, we define the discrete, pseudo-Fourier transforms of a function as

$$\tilde{G}(w_j) := \frac{h}{\sqrt{2\pi}} \sum_{|k|\leq pL} g(hk)\exp(-2\pi i hkw_j), \tag{3.4}$$

$$g(hq) := \frac{\sqrt{2\pi}}{h(2pL+1)} \sum_{|j|\leq pL} \tilde{G}(w_j)\exp(2\pi i hqw_j), \tag{3.5}$$

where $w_j := j/h(2pL+1)$.

In this notation the time domain implementation can be expressed in the Fourier domain,

$$\begin{aligned} Approx_\psi\ f(hq) &= \sum_{|k|\leq pL} f_o(kh)\psi([q-k]h) \\ &= \frac{2\pi}{h^2(2pL+1)} \sum_{|j|\leq pL} \tilde{F}_o(w_j)\tilde{\Psi}(w_j)\exp(2\pi i hw_j q). \end{aligned} \tag{3.6}$$

Although this illustrates an efficient method for computing the approximation on a refined grid without increasing the overall error, it remains necessary to compute the atom in order to determine $\tilde{\Psi}(w_j)$. However, the pseudo-Fourier transform is the spectral projection (truncated Fourier series) of the true Fourier representation,

$$\begin{aligned}\tilde{\Psi}\left(\frac{w_j}{2\pi h}\right) &= \frac{h}{\sqrt{2\pi}}\sum_{|k|\le pL}\psi(hk)\exp(-ikw_j)\\ &= h\int_{-\sigma(2-r)/r}^{\sigma(2-r)/r}\Psi\left(\frac{w}{2\pi h}\right)D_{pL}(2\pi h(w-w_j))dw\\ &= S_{pL}\Psi\left(\frac{w_j}{2\pi h}\right), \qquad (3.7)\end{aligned}$$

where $D_R(x) := \sin((R+1/2)x)/2\pi\sin(x/2)$ is the Dirichlet kernel of order R, and $S_Rf(\cdot)$ is the R term truncated Fourier series projection of $f(\cdot)$, i.e., $S_Rf := D_R * f$. As such, for highly smooth filters, the two Fourier representations are root-exponentially close[7] for all w_j,

$$\begin{aligned}\left|\Psi(w_j)-\tilde{\Psi}(w_j)\right| &= |\Psi(w_j)-S_{pL}\Psi(w_j)|\\ &\le \text{Const}\cdot e^{-\alpha(\eta_\Psi LT)^{1/\alpha}}, \qquad \Psi(\cdot)\in G^\alpha. \qquad (3.8)\end{aligned}$$

The composite error is then composed of the traditional truncation error, (2.1), and the error in replacing the pseudo-Fourier transform with the exact dual space representation of the atom, i.e., the filter. We summarize the above results in the following theorem.

THEOREM 3.1. *Let $f(t)$ be a signal bandlimited to $[-\sigma,\sigma]$, and $\Psi(\cdot)$ a filter in G^α satisfying* (2.2). *From the function's oversampled values on the mesh, $R := \{kT\}_{|k|\le L}$, where $T := r/2\sigma$; its approximation on the fine mesh, $P := \{kh\}_{|k|\le PL}/R$, where $h := r/2\sigma p$ for $p \in \mathbb{N}/\{1\}$, can be computed by pointwise multiplying the pseudo-Fourier transform,* (3.4), *of the signal's zero insertion onto P and the filter, $\Psi(\cdot)$. The resulting error is bounded by*

$$\begin{aligned}&\left|f(hk)-\frac{2\pi}{h^2(2pL+1)}\sum_{|j|\le pL}\tilde{F}_o(w_j)\Psi(w_j)\exp(2\pi ihw_jk)\right|\\ &\le \text{Const}\cdot\frac{p^2L}{\sigma r}\|f\|_{L^\infty}\left(e^{-\alpha(\eta_\Psi LT)^{1/\alpha}}+e^{-\alpha(\eta_\Psi(LT-|t|))^{1/\alpha}}\right)\\ &\le \text{Const}\cdot\frac{p^2L}{\sigma r}\|f\|_{L^\infty}e^{-\alpha(\eta_\Psi(LT-|t|))^{1/\alpha}}.\end{aligned}$$

Remark. Although the approximation of an arbitrary point is not possible by the method put forth in Theorem 3.1, the filter can be modulated to give an approximation on a shifted submesh, $P+s_0 := \{kh+s_0\}_{|k|\le pL}$. This can be seen directly be replacing $\tilde{\Psi}(w_j)$ in (3.6) with its modulation $\tilde{\Psi}(w_j)\exp(2\pi is_0w_j)$ and absorbing the modulation into the exponential, resulting in $Approx_\psi f(hq+s_0)$. An arbitrary point in $|t| \le LT$ can be approximated for a small modulation, $|s_0| \le h/2$, and

[7] This is true for any $p \in \mathbb{N}/\{1\}$ as the trapezoidal quadrature is taken over the support of the filter, $[-\sigma(2-r)/r, \sigma(2-r)/r] \subset [w_{-pL}, w_{pL}]$.

as a result, the difference between the modulated pseudo-filter and the true modulated filter remain exponentially small, only modestly increasing the constant in (3.8). Consequently any point in the interior of the samples, $|t| \leq LT$, can be approximated with the exponential accuracy stated in Theorem 3.1 through the application of the appropriately modulated filter and a zero insertion of $p = 2$.

4. An algorithm for periodic nonuniform sampling. Although uniform oversampling achieves the optimal convergence rate for a given sampling rate, applications exist where for various reasons one is confronted with more general sampling geometries. In this section we devote our attention to the case of periodic nonuniform sampling (bunched sampling), i.e., when multiple uniform undersampled sets are combined to achieve an effective sampling density similar to the uniform sampling case. First order methods exist for the reconstruction of the bandlimited signal from periodic nonuniform sampling [18, 2, 25, 5]. Here we investigate the dual space structure induced by periodic nonuniform sampling, and derive direct high resolution reconstruction methods.

It is well known that for periodic nonuniform sampling a generalization of the cardinal series (1.2) holds by using N atoms in the series expansion (in combination with appropriate coefficients) instead of just one atom. Various essentially equivalent versions of this generalized sampling theorem have been derived, which all revolve around exploiting in the Fourier domain certain periodicities induced by the sampling geometry. Initially we follow a similar path, but unlike other approaches we pay special attention in our derivations to our goal of using highly localized atoms for the reconstruction of the sampled function.

Poisson's summation theorem relates the physical space sampling to the resulting dual space periodization. Specifically, a bandlimited signal, $f \in B_\sigma$, sampled at the rate $T^{-1} := 2\sigma/r$ causes a dual space periodization of $2\sigma/r$. For oversampling, $r < 1$, the signal's dual space representation is separated (Figure 2.1), and the introduced periodization can be removed in one step by applying a smooth filter which satisfies the reproducing condition (2.2). Alternatively, when a signal is undersampled ($r > 1$), the dual space periodizations overlap, and a general signal cannot be reconstructed from those samples alone. More precisely, if a signal is sampled at the points $\{lT + T_n\}_{l\in\mathbb{Z}}$, the dual space representation is given by

$$S_{T_n}(w) := e^{2\pi i T_n w} \sum_{l=-\infty}^{\infty} e^{-2\pi i l T_n T^{-1}} F(w - lT^{-1}). \tag{4.1}$$

If $N \geq \lceil r \rceil$ such undersampled sets are available, then an *effective sampling rate of* $2\sigma N/r$ is obtained and the overlapping can be removed for $|w| \leq \sigma$, allowing the recovery of the bandlimited signal. We now present general conditions, reminiscent of (2.2), for the recovery of a bandlimited signal from its bunched sampling. We then conclude this section with an algorithm for the construction of a family of filters which remove the sampling induced periodization.

For a given value of k, $F(w)$ can be recovered from the N bunched sampling sets of the signal in the interval $I_k := [(-N+k-1)T^{-1}+\sigma, kT^{-1}-\sigma]$ by multiplying each undersampled set's dual space representation, S_{T_n} by an undetermined coefficient, $c_{k,n}$, selected to remove the periodizations for $l = -N + k, \ldots, k - 1$, i.e.,

$$F_k(w) := \sum_{n=1}^{N} c_{k,n} e^{-2\pi i T_n w} S_{T_n}(w) \qquad \text{with } F_k(w) = F(w) \quad \text{for } w \in I_k. \tag{4.2}$$

The coefficients can then be determined by solving the resulting system[8]

$$AR(k)c(k) = e_{N-k+1} \quad \text{with} \quad A_{m,n} := \exp(2\pi i T_n T^{-1} m), \quad m, n = 1, \ldots, N, \tag{4.3}$$

$c(k) = (c_{k,1}\ c_{k,2}\ \cdots\ c_{k,N})^T$, $e_{N-k+1} = \delta_{j,N-k+1}$, and

$$R(k) := \operatorname{diag}(\gamma_1(k)\ \gamma_2(k)\ \cdots\ \gamma_N(k)), \text{where } \gamma_n(k) := \exp(2\pi i T_n T^{-1}(k - N - 1)).$$

Repeating this process for a sufficient set of intervals to cover the bandwidth of the signal, $[-\sigma, \sigma] \subseteq \cup_{j=1}^{\kappa} I_{k_j}$, removes the sampling induced periodization for the bandwidth of interest.[9] However, the overall domain is segmented into κ overlapping intervals which must be spliced together with partitioning functions $\Phi_{k_j}(w)$ constructed appropriately to recover $F(w)$,

$$F(w) = \sum_{j=1}^{\kappa} F_{k_j}(w)\Phi_{k_j}(w), \qquad |w| \le \sigma. \tag{4.4}$$

For partitions which are not dependent on the translates, $\{T_n\}_{n=1}^N$ requires $\Phi_k(w) = 0$ for $w \notin I_k$ and consequently $\sum_{j=1}^{\kappa} \Phi_{k_j}(w) = 1$ for $|w| \le \sigma$. In summary, the signal's dual space representation can be recovered from its bunched sampling if the following conditions on the intervals and partitioning functions are satisfied:

(4.5)

$$[-\sigma, \sigma] \subseteq \cup_{j=1}^{\kappa} I_{k_j}, \qquad \Phi_{k_j}(w) = 0 \ \text{ for } \ w \notin I_{k_j}, \qquad \sum_{j=1}^{\kappa} \Phi_{k_j}(w) = 1 \ \text{ for } |w| \le \sigma.$$

Accordingly, a bandlimited signal's ($f(\cdot) \in B_\sigma$) Fourier transform can be recovered from its overlapping induced periodization by the set of filters $\{\Psi_n(\cdot)\}_{n=1}^N$,

$$\begin{aligned} F(w) &= \sum_{j=1}^{\kappa} \Phi_{k_j}(w) \sum_{n=1}^{N} S_{T_n}(w) c_{k_j,n} e^{-2\pi i T_n w} \\ &= \sum_{n=1}^{N} \left(\sum_{j=1}^{\kappa} c_{n,k_j} \Phi_{k_j}(w) \right) \sum_{l=-\infty}^{\infty} e^{-2\pi i l T_n T^{-1}} F(w - lT^{-1}) \\ &= \sum_{n=1}^{N} \Psi_n(w) \sum_{l=-\infty}^{\infty} e^{-2\pi i l T_n T^{-1}} F(w - lT^{-1}) \\ &= T \sum_{n=1}^{N} \Psi_n(w) \sum_{l=-\infty}^{\infty} f(lT + T_n) e^{-2\pi i w (lT + T_n)}, \end{aligned} \tag{4.6}$$

where the last equality is due to the Poisson summation formula, and the filters are defined as

$$\Psi_n(w) := \sum_{j=1}^{\kappa} c_{k_j,n} \Phi_{k_j}(w), \tag{4.7}$$

[8] The matrix A is of Vandermonde type and is therefore invertible for distinct translates, $\{T_n\}_{n=1}^N$.

[9] The effective oversampling rate $N/r > 1$ guarantees the full set of intervals cover the signal's bandwidth, $[-\sigma, \sigma] \subseteq_{k=1}^{N} I_k$.

with I_{k_j} and Φ_{k_j} selected to satisfy conditions (4.5), and $c_{k_j,n}$ being the solutions of the system (4.3). Note that this N stage filtering is in contrast to the one-step filtering, $F(w) = \Psi(w)\sum_{l=-\infty}^{\infty} F(w-lT^{-1})$, used to remove the periodization induced by uniform oversampling.

Consequently, by taking the inverse Fourier transform of (4.6), the signal can be represented by its bunched samples and translated atoms:

$$f(t) = T\sum_{n=1}^{N}\sum_{l=-\infty}^{\infty} f(lT - T_n)\psi_n(t - (lT + T_n)), \tag{4.8}$$

where ψ_n, the inverse Fourier transform of Ψ_n, is the atom associated with the undersampled set $\{lT-T_n\}_{l\in\mathbb{Z}}$. This is the corresponding generalization of the oversampling representation (2.3) to the case of bunched sampling.

Similar to the case of uniform oversampling, the truncation error for bunched sampling, (4.8), is governed by the atom's localization:

$$\begin{aligned}\epsilon_b(t, L, T) &:= \left| f(t) - T\sum_{n=1}^{N}\sum_{|l|\le L} f\left(lT - T_n\right)\psi_n\left(t - (lT + T_n)\right)\right| \\ &\le T\|f\|_{L_\infty}\sum_{n=1}^{N}\sum_{|l|>L}\left|\psi_n\left(t - (lT + T_n)\right)\right|. \end{aligned} \tag{4.9}$$

In [5] the Nyquist sampling rate $N = r$ was considered, where the interval $|w| \le \sigma$ was necessarily partitioned with characteristic functions, $\Phi_k(w) = \chi_{I_k}$. However, similar to uniform Nyquist sampling the abrupt filtering results in atoms with first order decay, and consequently the convergence rate for truncated sets of samples is first order, making it an impractical method for real world applications.

For smooth G^α filters, the atoms, $\psi_n(t)$, possess root-exponential localization, (2.8), and consequently the truncation error satisfies

$$\epsilon_b(t, L, T) \le \mathrm{Const}_{\alpha,\eta} NT\|A^{-1}\|\|f\|_{L^\infty}\cdot\exp(-\alpha(2\pi\eta((L-1)T - |t|))^{1/\alpha}), \tag{4.10}$$

where $\mathrm{Const}_{\alpha,\eta}$ is as before and $\|A^{-1}\|$ is determined solely by the set of translates, $\{T_n\}_{n=1}^N$.

Remark. (i) The truncation error bound in (4.10) is overly pessimistic in the factor $\|A^{-1}\|$ due to inherent structure in the system of equations $AR(k)c(k) = e_{N-k+1}$. The $(N-k+1)$th row of this system simplifies to $\sum_{n=1}^{N} c_{k,n} = 1$ for each k, imposing the additional structure on the atoms that the sum of the filters is the sum of the partitions,

$$\sum_{n=1}^{N}\Psi_n(w) = \sum_{n=1}^{N}\sum_{j=1}^{\kappa} c_{n,k_j}\Phi_{k_j}(w) = \sum_{j=1}^{\kappa}\Phi_{k_j}(w), \tag{4.11}$$

which is by construction a smooth function satisfying condition (2.2) with a modified bandwidth Ω. Applying the Fourier transform to (4.11), this structure implies that the atoms sum to a fixed function, independent of the set of translates. As a result, even when the translates are such that the matrix A is ill conditioned, a substantial amount of cancellation between the

atoms significantly reduces its effect on the truncation error. To see this more quantitatively, first group those translates that are near one another, say $\{T_n\}_{n\in\lambda}$. Then let x_λ be the overall amount of e_{N-k+1} contained in the space spanned by the associated columns of $AR(k)$. We observe in (4.3) that due to the linear dependence of the columns in $AR(k)$ associated with $n \in \lambda$ the coefficients $\{c_{k,n}\}_{n\in\lambda}$ are often large in magnitude. Nonetheless, the sum of these coefficients is dictated by x_λ, which by construction is order one, not by the individual translates. For this reason, although the individual *atoms* associated with $n \in \lambda$ may have large magnitude, determined by $\{c_{k,n}\}_{n\in\lambda}$ rather than x_λ, their sum will not. Combined with the values of $f(lT - T_n)$ for $n \in \lambda$ being nearly equal results in a substantial amount of cancellation between the associated atoms. To capture the effect of this cancellation quantitatively, rather than pass the absolute value onto each element as in (4.9), the bound can be left as

$$\epsilon_b(t, L, T) \le T \sum_{|l|>L} \left| \sum_{n=1}^{N} f(lT - T_n)\psi_n(t - (lT + T_n)) \right|, \tag{4.12}$$

where for each l, the atoms corresponding to near translates possess substantial cancellation. The numerical example in Figure 5.3 illustrates this effect, where the error near the sampling boundaries does not increase substantially for highly ill conditioned matrices A, but rather roundoff error in the cancellations pollutes the high resolution near the origin.

(ii) To improve the robustness of the proposed method even further we could multiply each uniform sampling set by some weight, similar to the general nonuniform sampling case discussed in [7]. In fact, by introducing properly chosen weights we can obtain estimates for the condition number of A, since the Toeplitz matrix A^*A is of the same form as the Toeplitz matrix appearing in [10]. We leave the details to the reader.

We now turn our attention to constructing intervals and smooth G^α partitioning functions satisfying condition (4.5) where the number of undersampled sets is sufficient to achieve a density similar to oversampling, $N > r$, with an *effective oversampling rate of* $N/r > 1$.

For minimal oversampling, $N = \lceil r \rceil > r$, the recovered regions only overlap with their immediate neighbors, i.e., $I_k \cap I_j = \emptyset$ for $|k-j| > 1$, and the full set of recovered zones $\{I_k\}_{k=1}^N$ is required to cover the interval $[-\sigma, \sigma]$. Moreover, for each I_k there is a subset that is not contained in the other intervals; consequently, the condition $\sum_k \Phi_k(w) = 1$ for $|w| \le \sigma$ implies that the partitioning functions must satisfy

$$\Phi_k(w) = \begin{cases} 1, & w \in [\max(-\sigma, (k-1)T^{-1} - \sigma), \min(\sigma, (-N+k)T^{-1} + \sigma)], \\ 0, & w \notin I_k. \end{cases} \tag{4.13}$$

An example of G^α partitioning functions satisfying conditions (4.13) and $\sum_k \Phi_k(w) = 1$ for $|w| \le \sigma$ is

(4.14)
$$\Phi_{k_1}(w) := \begin{cases} 0, & w \le \sigma + (k_1 - 1)T^{-1}, \\ \rho\left(\frac{-w-\sigma}{-(k_1-1)T^{-1}-2\sigma}\right), & \sigma + (k_1-1)T^{-1} < w < -\sigma, \\ 1, & -\sigma \le w \le \sigma + (k_2-1)T^{-1}, \\ \rho\left(\frac{w-(\sigma+(k_2-1)T^{-1})}{(k_1-k_2+N+1)T^{-1}-2\sigma}\right), & \sigma + (k_2-1)T^{-1} < w < (k_1+N-1)T^{-1} - \sigma, \\ 0, & (k_1+N-1)T^{-1} - \sigma \le w, \end{cases}$$

for the leftmost partition,

(4.15)
$$\Phi_{k_j}(w) := \begin{cases} 0, & w \le \sigma + (k_j-1)T^{-1}, \\ 1 - \rho\left(\frac{w-(\sigma+(k_j-1)T^{-1})}{(k_{j-1}-k_j+N+1)T^{-1}-2\sigma}\right), & \sigma + (k_j-1)T^{-1} < w \\ & \quad < (k_{j-1}+N-1)T^{-1} - \sigma, \\ 1, & (k_{j-1}+N-1)T^{-1} - \sigma \le w \\ & \quad \le \sigma + (k_{j+1}-1)T^{-1}, \\ \rho\left(\frac{w-(\sigma+(k_{j+1}-1)T^{-1})}{(k_j-k_{j+1}+N+1)T^{-1}-2\sigma}\right), & \sigma + (k_{j+1}-1)T^{-1} < w \\ & \quad < (k_{j+1}+1)T^{-1} - \sigma, \\ 0, & (k_{j+1}+1)T^{-1} - \sigma \le w, \end{cases}$$

for interior regions $j = 2, 3, \ldots, \kappa - 1$, and

(4.16)
$$\Phi_{k_\kappa}(w) := \begin{cases} 0, & w \le \sigma + (k_\kappa - 1)T^{-1}, \\ 1 - \rho\left(\frac{w-(\sigma+(k_\kappa-1)T^{-1})}{(k_{\kappa-1}-k_\kappa+N+1)T^{-1}-2\sigma}\right), & \sigma + (k_\kappa-1)T^{-1} < w \\ & \quad < (k_{\kappa-1}+N)T^{-1} - \sigma, \\ 1, & (k_{\kappa-1}+N)T^{-1} - \sigma \le w \le \sigma, \\ \rho\left(\frac{w-\sigma}{(k_\kappa+N)T^{-1}-2\sigma}\right), & \sigma < w < (k_\kappa+N)T^{-1} - \sigma, \\ 0, & (k_\kappa+N)T^{-1} - \sigma \le w, \end{cases}$$

for the rightmost partition, where $k_j := j - N$ for $j = 1, 2, \ldots, N$ and $\rho(\cdot)$ is defined as in (2.10).

For more general effective oversampling rates, $N > r$, the recovered zones I_k often overlap many of their neighbors, and as such constructing the full set of N partitioning functions satisfying $\sum_k \Phi_k(w) = 1$ for $|w| \le \sigma$ becomes substantially more complicated. However, for such higher effective oversampling, a smaller number of partitions, $\kappa \le N$, is required to cover the support of $F(w)$. The ideal subset of intervals and partitioning functions selected through $\{k_j\}_{j=1}^{\kappa}$ possess minimal slope, requiring that the intervals have equal size of internal and boundary overlaps, i.e., $length(I_{k_j} \cap I_{k_{j+1}}) = length(I_{k_1}/[-\sigma, \sigma]) = length(I_{k_\kappa}/[-\sigma, \sigma])$. Combined with the fixed length of I_k the optimal subset for a given κ is selected as $k_j^* := j\frac{N+1}{\kappa+1} - N$ for $j = 1, 2, \ldots \kappa$. The overlap length for κ sets with k_j as defined before is $T^{-1}(1 - r - \frac{\kappa N - 1}{\kappa+1})$. The minimum number of partitions κ necessary to cover the bandwidth

$[-\sigma, \sigma]$ is then determined by requiring the overlap interval to be nonnegative, yielding $\kappa_{min} := \lceil \frac{r}{N+1-r} \rceil$.

For computational purposes using the minimum number of partitions, κ_{min} results in unnecessarily steep partitions. Alternatively, a reasonable balance between simplicity of construction and minimizing the partitions slope is achieved by using the maximum number of partitions subject to the constraint that the intervals only interact with their immediate neighbors, yielding $\kappa^* := \min(N, \lfloor \frac{N+1+r}{N+1-r} \rfloor)$. The above results for bunched sampling are summarized in the following theorem.

THEOREM 4.1. *A bandlimited signal $f \in B_\sigma$ can be expressed in terms of its samples on the $N > \lceil r \rceil$ uniform meshes $\{lT + T_n\}_{l \in \mathbb{Z}}$, where $\{T_n\}_{n=1}^N$ are distinct and $T := r/2\sigma$. Reminiscent of the classical Shannon sampling theorem, the signal is decomposed into the translates of N atoms, $\psi_n(\cdot)$, each of which are associated with a particular uniform sampling mesh,*

$$f(t) = T \sum_{n=1}^{N} \sum_{l=-\infty}^{\infty} f(lT - T_n) \psi_n(t - (lT + T_n)).$$

A particularly simple construction of filters is achieved by solving the system of equations (4.3) *for $k_j^* = round(j \frac{N+1}{\kappa^*+1} - N)$, where $\kappa^* = \min(N, \lfloor \frac{N+1+r}{N+1-r} \rfloor)$. The N filters are then given by*

$$\Psi_n(w) := \sum_{j=1}^{\kappa^*} c_{k_j^*, n} \Phi_{k_j^*}(w),$$

where the coefficient $c_{k_j^*, n}$ are determined by solving (4.3) *for $\{k_j^*\}_{j=1}^{\kappa^*}$, and the partitioning functions are given by* (4.14), (4.15), (4.16).

Before developing a dual space implementation for truncated bunched sampling, we illustrate partitions and representative atoms for bunched samples in the case of minimal oversampling, $N = \lceil r \rceil > r$.

EXAMPLE 1. *The partitions for $r = 2.4$ as expressed in* (4.14), (4.15), (4.16) *for the case $N = \lceil r \rceil > r$ are shown in Figure* 4.1, *where $\rho(w) := \exp[\beta(w-1)^{-1} e^{-1/w}] \in G^2$, with $\beta = e^2/3$. It should be noted that the partitioning does not depend on the translates $\{T_n\}_{n=1}^N$, rather solely on the number of undersampled sets, N.*

The atoms, $\{\psi_n\}_{n=1}^N$, associated with $r = 2.4$, $N = 3$, and random translates $T_n/T = \{-0.4484, 0.3419, -0.0984\}$ are given in Figure 4.2. *This distribution of shifts is near the distribution that would correspond to uniform oversampling, $T_n/T = \{-1/3, 0, 1/3\}$ at the rate $N/r = 1.25$, and as such the atoms for this bunched sampling are qualitatively similar to the atom associated with uniform oversampling at the rate $N/r = 1.25$. Figure* 4.3 *illustrates the atoms for the set of translates $T_n/T = \{-1/3, 0, 10^{-6}\}$, where $\mathrm{cond}(A) = 5.5 \times 10^5$. Having constructed the partitions Φ_k from $\rho(\cdot) \in G^2$, the atoms possess root-exponential localization.*

4.1. Direct dual space implementation for bunched sampling. Just as in the case of the single oversampled set, the approximation of $f(\cdot)$ on the zero inserted grid $\{hq\}_{|q| \le pL}$, where $h = T/p$, can be implemented directly in the dual space. Define

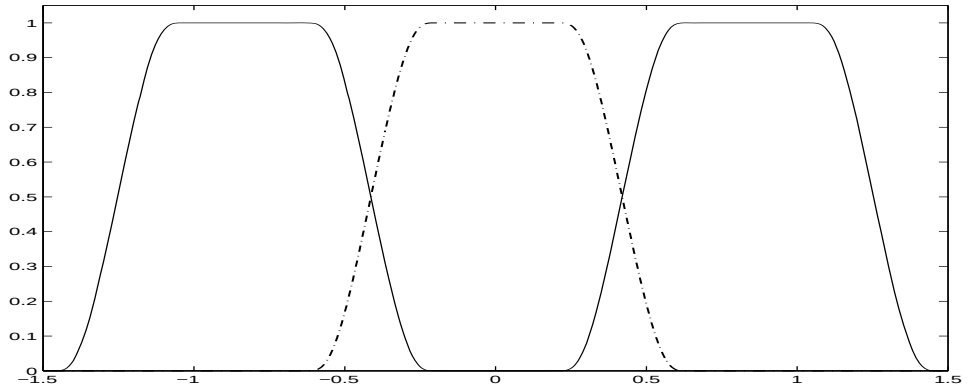


FIG. 4.1. *The partitions for* $r = 2.4$ *and* $N = 3$ *as described in* (4.14), (4.15), (4.16); *the end partitions* Φ_1, Φ_3 *(solid line) and the center partition* Φ_2 *(dot-dashed line).*

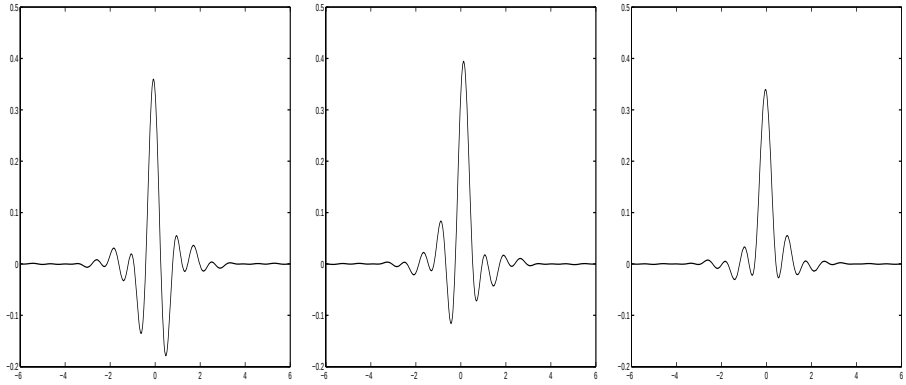


FIG. 4.2. *The atoms* $\{\psi_n\}_{n=1}^N$ *associated with* $r = 2.4$, $N = 3$, *and random translates* $T_n/T = \{-0.4484, 0.3419, -0.0984\}$ *shown from left to right, respectively. The atoms are qualitatively similar due to the low condition number* $\mathrm{cond}(A) = 1.8939$.

the approximation on this mesh from the bunched sampling as

$$Approx_{\psi,B} f(hq) := T \sum_{n=1}^{N} \sum_{|k| \le L} f(kT - T_n)\psi_n(hq - (kT + T_n))$$

$$= T \sum_{n=1}^{N} \sum_{|j| \le pL} f_0(jh - T_n)\psi_n(h(q-j)T_n), \tag{4.17}$$

where $f_0(x)$ is zero unless $x = kT - T_n$ for $k = -L, \ldots, L$. Replacing the point values in (4.17) with their pseudo-Fourier transform, we can express the evaluation in terms

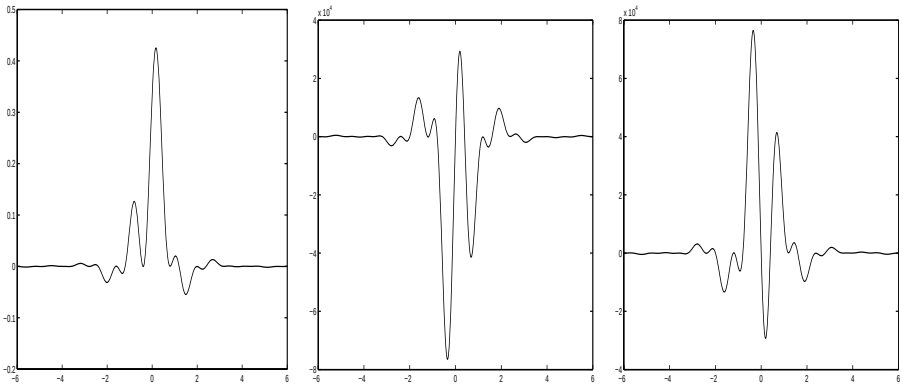


FIG. 4.3. *The atoms* $\{\psi_n\}_{n=1}^N$ *associated with* $r = 2.4$, $N = 3$, *and translates* $T_n/T = \{-1/3, 0, 10^{-6}\}$ *shown from left to right, respectively. The first atom is similar to those in Figure* 4.2, *and although the remaining atoms have significant amplitude, due to the relatively large condition number* $\mathrm{cond}(A) = 5.5 \times 10^5$, *they are nearly the negative of each other, allowing for significant cancellation.*

of the sampling sets pseudo-Fourier transforms,

$$\begin{aligned} Approx_{\psi,B} f(hq) &= T \sum_{n=1}^{N} \sum_{|j|\le pL} f_0(jh - T_n)\psi_n(h(q-j)T_n) \\ &= \frac{2\pi T}{h^2(2pL+1)^2} \sum_{n=1}^{N} \sum_{|j|\le pL} \left(\sum_{|l|\le pL} \tilde{F}_0(w_l) e^{2\pi i w_l (jh - T_n)} \right) \\ &\quad \times \left(\sum_{|k|\le pL} \tilde{\Psi}_n(w_k) e^{2\pi i w_k (h(q-j) - T_n)} \right) \\ &= T\sqrt{2\pi} \sum_{n=1}^{N} \sum_{|l|\le pL} \left(\frac{\sqrt{2\pi}}{h(2pL+1)} \tilde{F}_0(w_l) e^{2\pi i w_l (hq - T_n)} \right) \tilde{\Psi}_n(w_l) e^{-2\pi i T_n w_l}. \end{aligned} \tag{4.18}$$

The last line can be viewed as an algorithm, where first each undersampled set is zero inserted of order p, its pseudo-Fourier transform is computed, and it is multiplied by the appropriately modulated filter, $\tilde{\Psi}_n(\cdot)$, to remove the overlapping periodization. These filtered dual space representations are then summed and their inverse pseudo-Fourier transform computed to achieve an approximation of the bandlimited signal at the set of point $\{hq\}_{|q|\le pL}$. Just as in the case of the uniform oversampled dual space representation it is computationally advantageous to avoid the construction of $\psi_n(\cdot)$ in order to compute $\tilde{\Psi}_n$; rather for a fast algorithm the true filters $\{\Psi_n(\cdot)\}_{n=1}^N$ should be applied directly. Again we note that the atom's pseudo-Fourier transform is the pL order spectral projection of the atom's associated filter, (3.7). As such their difference is exponentially small,

$$\left| \Psi_n(w_l) - \tilde{\Psi}_n(w_l) \right| \le \mathrm{Const} \cdot \|A^{-1}\| e^{-\alpha(\eta_n LT)^{1/\alpha}}, \qquad \Psi_n(\cdot) \in G^\alpha, \tag{4.19}$$

if the zero padding is sufficient to expend the dual axis beyond the support of the filter, $p \ge 2N - r$.

In addition to the usual truncation error, this additional error gives a threshold below which the error does not fall, determined by the condition number of A as dictated by the set of translates. Unlike the truncation error, the structure in A does not result in cancellation to reduce the effects of $\|A^{-1}\|$; however, the bound does not depend on t and is well below the truncation error for all but the most ill conditioned sets of translates. We summarize the above results in the following theorem.

THEOREM 4.2. *Let $f(t) \in B_\sigma$ be sampled at the points $\{kT + T_n\}_{|k|\le L}$ with sampling rate $T^{-1} := 2\sigma/r$ and translate $|T_n| \le T/2$. From $N > r$ such distinct sampling sets, and filters $\{\Psi_n\}_{n=1}^N \in G^\alpha$ constructed as in Theorem* 4.1*, the signal can be approximated on the set $\{kT/N\}_{|k|\le pL}$, with $p \ge 2N - r$, within the bound*

$$|f(t) - Approx_B f(hk)| \le \text{Const}\,\frac{N^2}{\sigma}\|f\|_{L^\infty}\|A^{-1}\| \cdot e^{-\alpha(\eta((L-1)T-|t|))^{1/\alpha}},$$

where $Approx_B f(hk)$ is computed by the following algorithm:

1. *Zero insert each of the uniform sampling sets $\{kT+T_n\}_{|k|\le L}$ to the fine mesh $\{kh + T_n\}_{|k|\le pL}$, where $p \ge 2N - r$ and $h = T/p$, yielding $\{f_{o,n}\}_{n=1}^N$.*
2. *Compute the pseudo-Fourier transform, as defined in* (3.4)*, of each set of zero inserted samples from step* 1*, labeled $\{\tilde{F}_{o,n}(w_l)\}_{n=1}^N$, and pointwise multiply by $\exp(-2\pi i T_n w_l)$, respectively.*
3. *Pointwise multiply each of the pseudo-Fourier transforms from step* 2*, $\tilde{F}_{o,n}(w_l)$, by their corresponding filters, $\Psi_n(w_l)$, and sum over N, yielding an approximation of $F(w)$, $\tilde{F}_A(w_l) := \sum_{n=1}^N e^{-2\pi i T_n w_l}\tilde{F}_{o,n}(w_l)\Psi_n(w_l)$.*
4. *Compute the inverse pseudo-Fourier transform of $\tilde{F}_A$ formed in step* 4*, and multiply by p.*

Similar to the remark following Theorem 4.1, the signal can be recovered on a shifted mesh, $\{kT/N + s_0\}_{|k|\le pL}$, by multiplying $\tilde{F}_A(w_l)$ with $\exp(2\pi i s_0 w_l)$ between steps 3 and 4 of the algorithm in Theorem 4.2.

The direct implementation for uniform oversampling has a computational cost limited by the FFT, proportional to $L\log(L)$, where L is the number of samples used in the reconstruction. For the algorithm described in Theorem 4.2 for bunched sampling, the computational cost is again limited by the overall FFT evaluations. To allow for a direct comparison to the uniform oversampling let each of the N sampling sets contain L/N samples for a similar total number of samples being available for the algorithm. Each of the N FFTs then requires $\frac{L}{N}\log(pL/N)$, where p is the level of zero insertion required to be proportional to the number of sampling sets, $p \approx N$. The overall computational cost for the algorithm in Theorem 4.2 is then $NL\log(L)$, where L is the total number of samples used in the approximation. In principle we also have to take into account the costs for inverting the matrix A in (4.3). Since A is a Vandermonde matrix and A^*A is a Toeplitz matrix, there are plenty of fast standard algorithms for the solution of the system in (4.3) at our disposal. Moreover, in practice N is small compared to L, and thus the computational costs of this step have little impact on the overall complexity of the proposed method. In summary, the bunched sampling algorithm requires an additional factor of N in the total computational cost, when compared to uniform sampling at the same effective sampling rate and number of available samples.

5. Numerical examples. We illustrate the convergence rates and algorithms for the results presented in Theorems 3.1 and 4.2 for characteristic bandlimited signals. For the approximation of an arbitrary bandlimited signal, we form a test signal

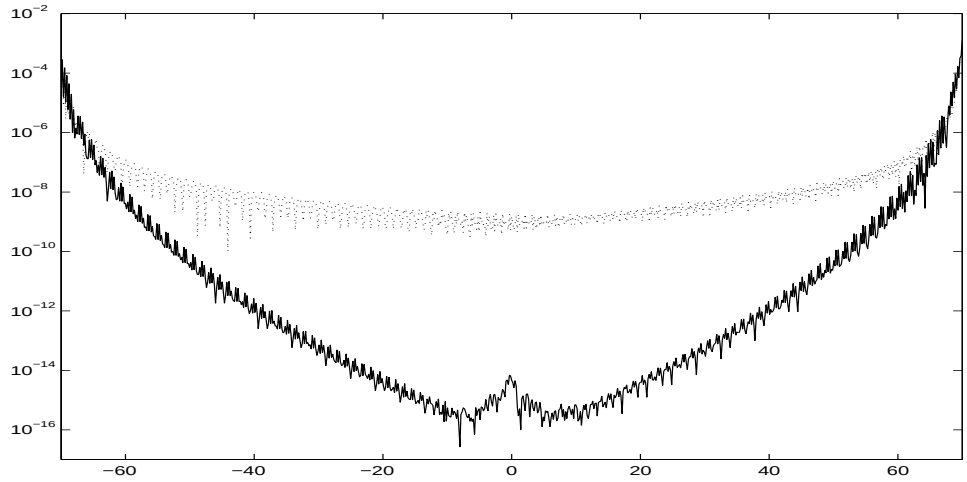


FIG. 5.1. *The error with oversample rate* $r^{-1} = 1.43$ *in recovering the signal in* B_1 *whose real portion is shown in Figure* 2.1. *With the log axis, note the log convergence exhibited by the reconstruction using the raised cosine (dotted line), as compared to the root convergence obtained by the dual space implementation using* Ψ_{G2} *(solid line). The approximations were computed with zero insertion rate* $p = 2$ *on the grid translated by* $s_0 = T/\sqrt{5}$.

whose Fourier transform is composed of one hundred characteristic functions with random complex valued amplitudes normalized to unit l_2 norm, and with random widths and centers, normalized so that the largest magnitude bandwidth is σ. The resulting numerics shown are characteristic of arbitrary complex valued bandlimited signals. The dual space representation, real portion, of such a function can be seen in Figure 2.1. Before illustrating the main results of Theorem 4.2 we briefly contrast the exponential convergence of Gevrey class filters with the polynomial order convergence of classical finite regular filters. To compare representative filters with finite and infinite regularity, we use the canonical raised cosine filter (2.6) and the Gevrey order two filter given in (2.10), respectively; see Figure 5.1.

Much of the success of the raised cosine filter is due to the optimally small first two regularity constants, $\|\Psi\|_{C^s}$ for $s = 1, 2$, which result in rapid initial localization. Infinitely regular filters possess bounded regularity constants for all s, but at the cost of necessarily larger regularity constants for small s. However, a great deal of freedom exists in the selection of G^α regular filters, for example the constant β used in the filter of (2.10). A good approximation of the β which minimizes the first regularity constant in Ψ_{G2} can be obtained by selecting β such that the filter's points of inflection are at the middle of the region connecting zero and one, i.e., $\Psi^{(2)}(\pm\sigma/r) = 0$. As such, for the numerical experiments involving the filter Ψ_{G2}, we use $\beta := \frac{1}{3}e^2$ which satisfies $\rho^{(2)}(\frac{1}{2}) = 0$.

Various properties of Theorem 4.2 are demonstrated in the following numerical examples. First we begin with the simplest case of bunched sampling where only just sufficiently many sampling sets are available for effective oversampling, $N := \lceil r \rceil > r$; in particular for $r = 2.4$ and a random set of well separated translates with corresponding atoms presented in Figure 4.2, and the exponentially small error shown in Figure 5.2. To illustrate the computational robustness, Figure 5.3 shows the error as two sampling sets approach one another, resulting in a poorly conditioned matrix A in system (4.3). However, as noted in the Remark following (4.10), the ill conditioning

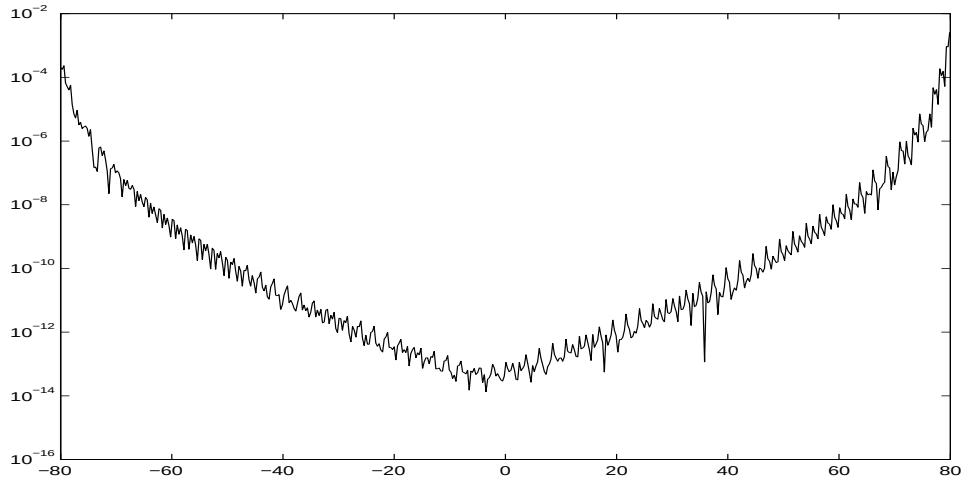


FIG. 5.2. *The error with $N = 3$ undersampled sets with effective sampling rate $N/r = 1.25$ in recovering the signal in B_1 whose real portion is shown in Figure* 2.1. *The translates were $T_n/T = \{-0.4484, -0.0984, 0.3419\}$ with $\mathrm{cond}(A) = 1.8939$ and atoms illustrated in Figure* 4.2.

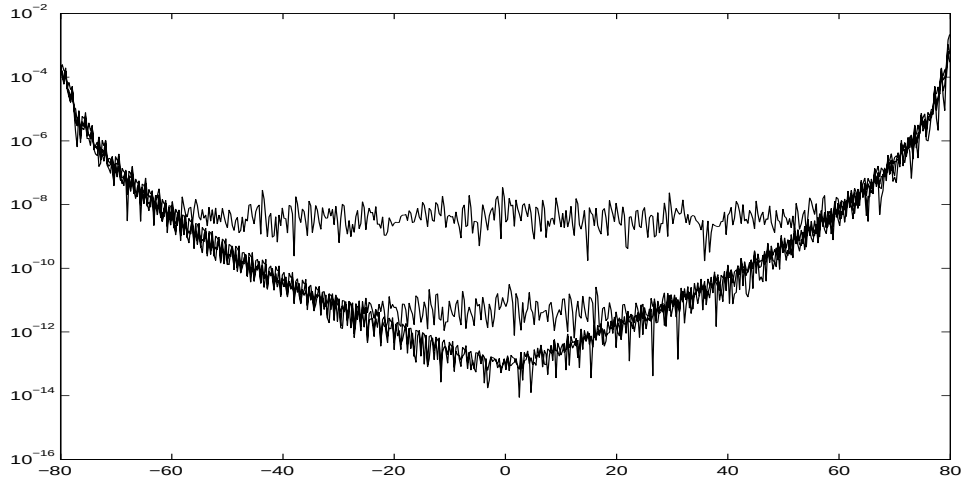


FIG. 5.3. *The error with $N = 3$ undersampled sets and effective sampling rate $N/r = 1.25$ in recovering the signal in B_1 whose real portion is shown in Figure* 2.1. *The translates used, $T_n/T = \{-1/3, 0, 10^{-3j}\}$ for $j = 1, 2, 3$, result in systems* (4.3) *with respective condition numbers $\mathrm{cond}(A) = 5.5 \times 10^{3j-1}$. Note that although the condition number becomes very large, the error near the boundaries does not suffer. Rather, the minimal error is increased due to roundoff errors. The atoms associated with the set of translations, $j = 2$, are shown in Figure* 4.3.

of the matrix does not increase the entire error by the factor $\|A^{-1}\|$ as stated in the pessimistic bound of Theorem 4.2; rather, the ill conditioning results in a rounding error that limits the achievable error for a given precision arithmetic.

A more general example of Theorem 4.2 is shown in Figure 5.4 where $N = 17$ uniform sampling sets with random translates undersampled at the rate $r = 12.4$ are given for an effective sampling rate of $N/r \approx 1.37$. The resulting error is typical for the algorithm of Theorem 4.2 when the system (4.3) has a relatively modest condition

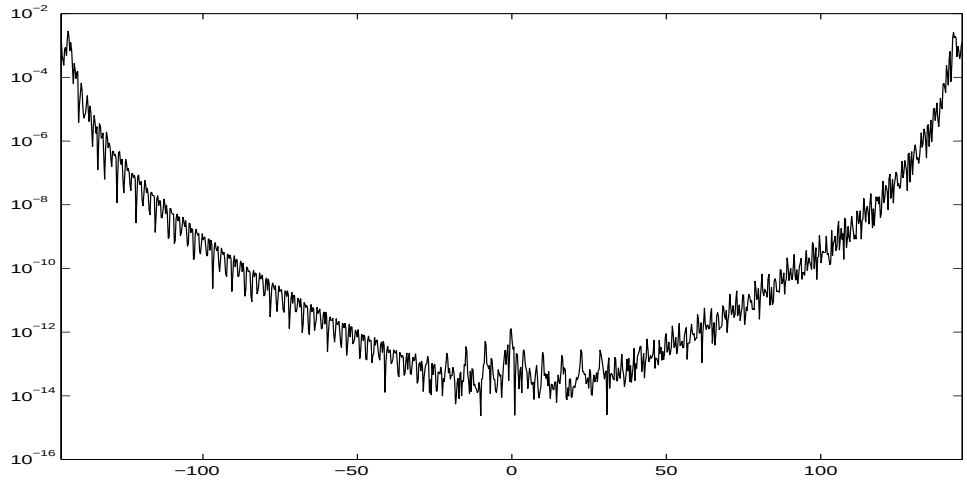


FIG. 5.4. *The error in recovering a random signal in* B_1 *from* $N = 17$ *undersampled sets of random translation with effective sampling rate* $N/r = 1.37$. *The filters were composed of five partitions,* $\kappa^* = 5$, *and the system of the random translates had condition number* $\mathrm{cond}(A) = 1.4 \times 10^4$.

number.

Before concluding the numerical examples we illustrate the algorithm's performance for a particular application discussed in the introduction. Due to feedback interactions between sampling and processing chips in close proximity, it is advantageous to use a sampling structure that includes relatively large sampling gaps, when the signal processing can be applied and not interact with the sampling. However, the introduced sampling gap could potentially introduce stability problems. The extraordinary robustness of the algorithm in Theorem 4.2 overcomes any stability issues, even for relatively large sampling gaps. For example, Figure 5.5 shows the approximation error with eight sets interleaved over a third of the effective sampling rate, i.e., $T_n := T(n/24)$ for $n = 1, 2, \ldots, 8$.

6. Final remarks. We have derived a fast algorithm for reconstructing a bandlimited signal from its periodic nonuniform samples that achieves root-exponential accuracy with respect to the given number of samples. Due to its high accuracy the method can be easily realized in practice via finite impulse response (FIR) filters. Furthermore, since the numerically most expensive steps are FFTs the proposed method lends itself to a simple implementation on standard DSP processors. Furthermore, the high accuracy provided by the algorithms derived in this paper will not be lost in the subsequent reconstruction of the signal from its quantized samples due to the recently developed highly accurate algorithms for recovering a quantized bandlimited signal; cf. [4].

Another application where periodic nonuniform sampling arises is image processing. For instance, in astronomical imaging one is confronted with images that are blurred and notoriously undersampled. The goal is to combine these blurred low-resolution images to one high-resolution image. This problem is also referred to as superresolution; see, e.g., [6]. The low-resolution images contain (blurred, noisy) samples of the high-resolution image where the sampling sets can be thought of as a union of arbitrarily shifted (and/or rotated) uniform sampling sets. One step in reconstruct-

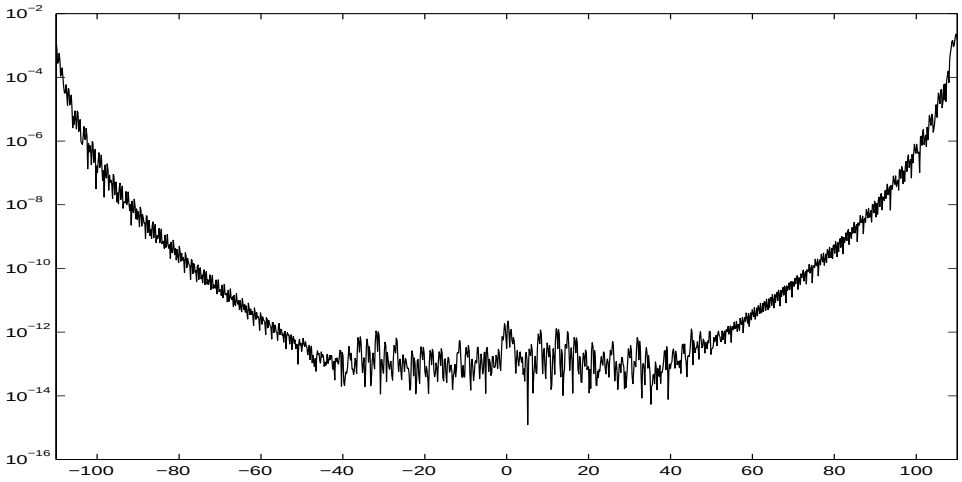


FIG. 5.5. *The error in recovering a random signal in B_1 from $N = 8$ undersampled sets with effective sampling rate $N/r = 1.82$. The filters were composed of two partitions, $\kappa^* = 2$, and the translates, $T_n := T(n/24)$ for $n = 1, 2, \ldots, 8$, had condition number* $\mathrm{cond}(A) = 3.1 \times 10^4$.

ing a high-resolution image is thus the conversion of the periodic nonuniform image samples to a uniform sampling set at a fine sampling grid. In our future research we will address two-dimensional reconstruction algorithms and their effect on deblurring and noise.

Acknowledgment. We are indebted to Professor Bernard Levy for insightful discussions on practical aspects of the research presented in this paper.

REFERENCES

[1] A. BEURLING AND P. MALLIAVIN, *On the closure of characters and the zeros of entire functions*, Acta Math., 118 (1967), pp. 79–93.

[2] P. L. BUTZER AND G. HINSEN, *Reconstruction of bounded signals from pseudo-periodic irregularly spaced samples*, Signal Process., 17 (1989), pp. 1–17.

[3] I. DAUBECHIES, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[4] I. DAUBECHIES AND R. DEVORE, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Ann. of Math. (2), 158 (2003), pp. 679–710.

[5] A. FARIDANI, *A generalized sampling theorem for locally compact abelian groups*, Math. Comp., 63 (1994), pp. 307–327.

[6] S. FARSIU, D. ROBINSON, M. ELAD, AND P. MILANFAR, *Advances and challenges in super-resolution*, Int. J. Imaging Systems and Technology, 14 (2004), pp. 47–57.

[7] H. G. FEICHTINGER AND K. H. GRÖCHENIG, *Theory and practice of irregular sampling*, in Wavelets: Mathematics and Applications, J. Benedetto and M. Frazier, eds., CRC Press, Boca Raton, FL, 1994, pp. 305–363.

[8] R. GERVAIS, Q. I. RAHMAN, AND G. SCHMEISSER, *A bandlimited function simulating a duration-limited one*, in Anniversary Volume on Approximation Theory and Functional Analysis (Oberwolfach, 1983), Internat. Schriftenreihe Numer. Math. 65, Birkhäuser-Verlag, Basel, Switzerland, 1984, pp. 355–362.

[9] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 26, SIAM, Philadelphia, 1977.

[10] K. GRÖCHENIG, *Irregular sampling of wavelet and short-time Fourier transforms*, Constr. Approx. Theory, 9 (1993), pp. 283–297.

[11] S. HUESTIS, *Optimum kernels for oversampled signals*, J. Acoust. Soc. Amer., 92 (1992), pp.

1172–1173.
[12] H. Johansson and P. Löwenborg, *Reconstruction of nonuniformly sampled bandlimited signals by means of digital fractional delay filters*, IEEE Trans. Acoust. Speech Sig. Proc., 50 (2002), pp. 2757–2767.
[13] F. John, *Partial Differential Equations*, 4th ed., Appl. Math. Sci. 1, Springer-Verlag, New York, 1991.
[14] A. Kohlenberg, *Exact interpolation of bandlimited functions*, J. Appl. Phys., 24 (1953), pp. 1432–1436.
[15] F. Marvasti, ed., *Theory and Practice of Nonuniform Sampling*, Kluwer Academic Publishers, Dordrecht, The Netherlands, Plenum, New York, 2001.
[16] F. Natterer, *Efficient evaluation of oversampled functions*, J. Comput. Appl. Math., 14 (1986), pp. 303–309.
[17] R. A. Niland, *Optimum oversampling*, J. Acoust. Soc. Amer., 86 (1989), pp. 1805–1812.
[18] A. Papoulis, *Signal Analysis*, McGraw–Hill, New York, 1977.
[19] S. Pilipović and N. Teofanov, *Wilson bases and ultramodulation spaces*, Math. Nachr., 242 (2002), pp. 179–196.
[20] R. S. Prendergast, B. C. Levy, and P. J. Hurst, *Reconstruction of band-limited periodic nonuniformly sampled signals from multirate filter banks*, IEEE Trans. Circuits Syst. I Regul. Pap., 51 (2004), pp. 1612–1622.
[21] T. S. Rappaport, *Wireless Communications: Principles & Practice*, Prentice–Hall, Upper Saddle River, NJ, 1996.
[22] T. Strohmer and J. Tanner, *Implementations of Shannon's sampling theorem, A time-frequency approach*, Sampl. Theory Signal Image Process., 4 (2005), pp. 1–17.
[23] T. Strohmer and J. Xu, *Blind calibration of multiple time-interleaved analog-to-digital converters*, IEEE Trans. Signal Process., submitted, 2006.
[24] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[25] P. P. Vaidyanathan and V. C. Liu, *Efficient reconstruction of band-limited sequences from nonuniformly decimated versions by use of polyphase filter banks*, IEEE Trans. Acoust. Speech Signal Process, 38 (1990), pp. 1927–1936.
[26] A. I. Zayed, *Advances in Shannon's Sampling Theory*, CRC Press, Boca Raton, FL, 1993.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1095–1121

C^1 INTERPOLATORY SUBDIVISION WITH SHAPE CONSTRAINTS FOR CURVES*

TOM LYCHE† AND JEAN-LOUIS MERRIEN‡

Abstract. We derive two reformulations of the C^1 Hermite subdivision scheme introduced by Merrien. One where we separate computation of values and derivatives and one based of refinement of a control polygon. We show that the latter leads to a subdivision matrix which is totally positive. Based on this we give algorithms for constructing subdivision curves that preserve positivity, monotonicity, and convexity.

Key words. interpolation, subdivision, corner cutting, total positivity, positivity, monotonicity, convexity

AMS subject classifications. 65D05, 65D17

DOI. 10.1137/040621302

1. Introduction. Subdivision is a technique for creating a smooth curve or surface out of a sequence of successive refinements of polygons; or grids see [2]. Subdivision has found applications in areas such as geometric design [6, 16], and in computer games and animation [4]. We consider here the two point Hermite scheme, the HC^1-algorithm, introduced in [11]. We start with values and derivatives at the endpoint of an interval and then compute values and derivatives at the midpoint. Repeating this on each subinterval we obtain in the limit a function with a certain smoothness. The scheme depends on two parameters α and β and it has been shown that the limit function is C^1 for a range C of these parameters. For more references to Hermite subdivision, see [5, 10, 12, 13].

The strong locality of the HC^1-algorithm was used in [13] to construct subdivision curves with shape constraints like positivity, monotonicity, and convexity. A notion of control points, control coefficients, and a Bernstein basis for two subfamilies of the HC^1-interpolant were introduced in [15].

In this paper we continue the study of subdivision with shape constraints initiated in [13, 15]. Before detailing our results let us first describe the shape preserving subdivision process and give an example. Suppose we have values $y_1, \dots, y_n$ and derivatives $y'_1, \dots, y'_n$ at some abscissae $t_1 < t_2 < \cdots < t_n$. With each subinterval $[t_i, t_{i+1}]$ we associate parameters $(\alpha_i, \beta_i) \in C$ chosen so that the HC^1-interpolant using data $(y_i, y'_i, y_{i+1}, y'_{i+1})$ has the required shape on $[t_i, t_{i+1}]$. We then obtain a C^1-function on $[t_1, t_n]$. As an illustration consider the function in Figure 1.1.

This function is defined on the interval $[0, 4]$. It is positive on $[0, 1]$, strictly increasing on $[1, 2]$, constant on $[2, 3]$, and concave on $[3, 4]$. Suppose we want to use subdivision to construct a C^1-approximation to this function with the same shape characteristics and that all we know about the function is the function values $y_1, \dots, y_n$ at some points $t_1 < \cdots < t_n$. We can achieve this with the HC^1-algorithm using only

*Received by the editors December 22, 2004; accepted for publication (in revised form) December 5, 2005; published electronically June 2, 2006.
http://www.siam.org/journals/sinum/44-3/62130.html
†CMA and Institute of Informatics, University of Oslo, PO Box 1053, Blindern, 0316 Oslo, Norway (tom@ifi.uio.no).
‡INSA de Rennes, 20 av. des Buttes de Coesmes, CS 14315, 35043 RENNES CEDEX, France (Jean-Louis.Merrien@insa-rennes.fr).

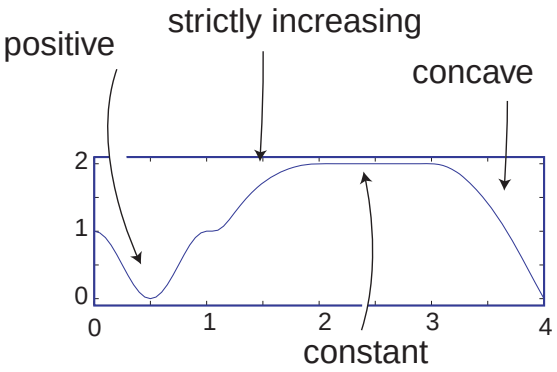


FIG. 1.1. *A given function.*

crude estimates for the derivatives $y'_1, \ldots, y'_n$ as long as the transition points $1, 2, 3$ are among the abscissae and the chosen derivatives are consistent with the required shapes; see section 6 for details. For classical curve based shape preserving algorithms we refer to [7, 8, 9] and references therein.

As for Bézier and spline curves we introduce a control polygon for the HC^1-interpolant. The originality of this family of interpolants is that monotonicity or convexity of the control polygon and of the function are equivalent (see Propositions 5.3 and 5.6). We observe that this is only true in one direction for Bezier or spline curves.

Our paper can be detailed as follows. In section 2, we recall the HC^1-algorithm and some properties which were proved in [13]. We give a new formulation of the HC^1-algorithm where we separate the computation of function values and derivatives. This formulation is useful for proving shape preserving properties and with the aid of this formulation we simplify the proofs of the main results in [13]. The new formulation also shows why the one parameter family given by $\alpha = \beta/(4(1-\beta))$ and $\beta \in [-1, 0)$ considered in [13, 15] really is an extension of the quadratic spline case. We will refer to this family as the *EQS-case* of the HC^1-algorithm. We also give a new domain C for C^1-convergence of the algorithm. In section 3 we use the control points introduced in [15] to reformulate the HC^1-algorithm as a stationary subdivision algorithm called SC^1. The control points depend on a third parameter $\lambda \geq 2$ and we show convergence of the SC^1-algorithm for $(\alpha, \beta) \in C$ and $\lambda \geq 2$. Starting in section 4, we restrict our attention to the EQS-case. By formulating the SC^1-algorithm as a corner cutting scheme we show that the subdivision matrix $\mathbf{S}$ is totally positive. We show this for an extended range of β and λ and also prove the total positivity of the HC^1-Bernstein basis. With this last property, the interpolant inherits shape properties of the control polygon such as nonnegativity, monotonicity, or convexity. In section 5, we give algorithms for interpolation with any of the previous shape constraints. An example based on Figure 1.1 is given in section 6.

We also point out that Proposition 2.1 on one hand and Proposition 3.1 with Theorem 3.4 on the other hand show that we obtain two Lagrange subdivision schemes from the HC^1-Hermite subdivision scheme.

2. The HC^1-algorithm. We recall the univariate version of the Hermite subdivision scheme for C^1 interpolation given by Merrien [11], which we call here HC^1. We start with values $(f(a), p(a))$ and $(f(b), p(b))$ of a function f and of its first derivative

$p = f'$ at the endpoints a, b of a bounded interval $I := [a, b]$ of $\mathbb{R}$. To build f and p on I, we proceed recursively. At step n ($n \geq 0$), let us denote by $\mathcal{P}_n$ the regular partition of I in 2^n subintervals and let us write $h_n := (b-a)/2^n$. If c and d are two consecutive points of $\mathcal{P}_n$, then we compute f and p at the midpoint $(c+d)/2$ according to the following scheme, which depends on two parameters α and β,

$$\begin{aligned} f\left(\frac{c+d}{2}\right) &:= \frac{f(d)+f(c)}{2} + \alpha h_n[p(d)-p(c)], \\ p\left(\frac{c+d}{2}\right) &:= (1-\beta)\frac{f(d)-f(c)}{h_n} + \beta\frac{p(d)+p(c)}{2}. \end{aligned} \tag{2.1}$$

By applying these formulae on ever finer partitions, we define f and p on $\mathcal{P} = \cup\mathcal{P}_n$ which is a dense subset of I. We say that the scheme is C^1-*convergent* if, for any initial data, f and p can be extended from $\mathcal{P}$ to continuous functions on I with $p = f'$. We call f defined either on I or on $\mathcal{P}$ the HC^1-*interpolant* to the data.

The HC^1-algorithm can also be formulated as follows. We start with Hermite data f_0, p_0, f_1, p_1 at the endpoints of a finite interval $[a, b]$ and set $f_0^0 = f_0$, $p_0^0 = p_0$, $f_1^0 = f_1$, and $p_1^0 = p_1$. For $n = 0, 1, 2, \ldots$, $h_n = 2^{-n}(b-a)$, and $k = 0, 1, \ldots, 2^n - 1$

$$f_{2k}^{n+1} := f_k^n, \quad f_{2k+1}^{n+1} := \frac{f_{k+1}^n + f_k^n}{2} + \alpha h_n\big(p_{k+1}^n - p_k^n\big), \tag{2.2}$$

$$p_{2k}^{n+1} := p_k^n, \quad p_{2k+1}^{n+1} := (1-\beta)\frac{f_{k+1}^n - f_k^n}{h_n} + \beta\frac{p_{k+1}^n + p_k^n}{2}, \tag{2.3}$$

and $f_{2^{n+1}}^{n+1} := f_{2^n}^n$, $p_{2^{n+1}}^{n+1} := p_{2^n}^n$. If the scheme is C^1-convergent with limit functions f and p, then

$$f(t_k^n) = f_k^n, \ f'(t_k^n) = p(t_k^n) = p_k^n, \ t_k^n := a + kh_n, \ k = 0, 1, \ldots, 2^n. \tag{2.4}$$

2.1. The vector space of HC^1-interpolants. To each choice of (α, β) there is a vector space

$$VC^1_{\alpha,\beta}(\mathcal{P}) := \{f : \mathcal{P} \to \mathbb{R} : f, p \text{ computed by (2.2)–(2.4)}\}$$

of HC^1-interpolants. If the scheme is C^1-convergent we define

$$VC^1_{\alpha,\beta}(I) := \{f : I \to \mathbb{R} : f|_{\mathcal{P}} \in VC^1_{\alpha,\beta}(\mathcal{P})\}.$$

The HC^1-*Hermite basis functions* $\{\phi_0, \psi_0, \phi_1, \psi_1\}$ are defined by taking as initial data the four unit vectors $e_j = (\delta_{i,j})_{i=1}^4$, respectively. They are always defined on $\mathcal{P}$ and the HC^1-interpolant corresponding to initial data (f_0, p_0, f_1, p_1) can be written $f = f_0\phi_0 + p_0\psi_0 + f_1\phi_1 + p_1\psi_1$. Since the Hermite basis functions are clearly linearly independent on $\mathcal{P}$ they form a basis for $VC^1_{\alpha,\beta}(\mathcal{P})$. Thus $VC^1_{\alpha,\beta}(\mathcal{P})$ and $VC^1_{\alpha,\beta}(I)$ are vector spaces of dimension 4.

Let us denote the HC^1-interpolant to initial data sampled from a function g by $f = Hg$. By induction it is easy to see that for any (α, β) we have $g = Hg$ for all polynomials g of degree at most one, while $g = Hg$ for all quadratic polynomials if and only if $\alpha = -1/8$. We also have $g = Hg$ for all cubic polynomials if and only if $\alpha = -1/8$ and $\beta = -1/2$ and it can be shown that $x^k \neq Hx^k$ for any integer $k \geq 4$. The fact that the scheme reproduces polynomials up to a certain degree can be used to give error bounds; see [13, section 5]. Assume (α, β) are chosen so that the scheme

is C^1-convergent. Then there is a constant $C(\alpha,\beta)$ such that for all intervals $I=[a,b]$ and all $g\in C^k(I)$ we have

$$\|g-Hg\|_{L_\infty(I)}\le C(\alpha,\beta)h^k\|g^{(k)}\|_{L_\infty(I)}, \tag{2.5}$$

where $h:=b-a$ and $k=2$ for most choices of α and β.

Notice some important choices of (α,β):

1. If $\alpha=-1/8$, $\beta=-1/2$, then f is the cubic polynomial known as the Hermite cubic interpolant. For this choice of parameters, (2.5) holds with $k=4$ and $C(\alpha,\beta)=1/384$.
2. If $\alpha=-1/8$, $\beta=-1$, then f is the Hermite quadratic interpolant, i.e., the quadratic C^1 spline interpolant with one knot at the midpoint of the initial interval. In this case (2.5) holds with $k=3$ and $C(\alpha,\beta)=1/96$; see [13].
3. The EQS-case $\alpha=\frac{\beta}{4(1-\beta)}$ with $\beta\in[-1,0)$ is a one parameter extension of the quadratic spline case; it was introduced and studied in [13], see also [15]. In this case (2.5) only holds with $k=2$ and $C(\alpha,\beta)\le 1/48$ unless $\beta=-1$, but as we will see this scheme has important shape preserving properties.

2.2. Direct computation of the function or the derivative. We can reformulate (2.2), (2.3) so that only values of p are involved and similarly for f.

PROPOSITION 2.1. *For $\alpha,\beta\in\mathbb{R}$, the function f and the derivative p of the HC^1-interpolant satisfy the following relations:*
For $n=1,2,\dots$ and $i=0,1,\dots,2^{n-1}-1$,

$$\begin{bmatrix} p_{4i}^{n+1}\\ p_{4i+1}^{n+1}\\ p_{4i+2}^{n+1}\\ p_{4i+3}^{n+1}\end{bmatrix}=\begin{bmatrix} 1 & 0 & 0\\ \mu & 1+\beta/2 & -\nu\\ 0 & 1 & 0\\ -\nu & 1+\beta/2 & \mu\end{bmatrix}\begin{bmatrix} p_{2i}^{n}\\ p_{2i+1}^{n}\\ p_{2i+2}^{n},\end{bmatrix}. \tag{2.6}$$

For $n\ge 2$ and $i=0,1,\dots,2^{n-2}-1$

$$\begin{bmatrix} f_{8i}^{n+1}\\ f_{8i+1}^{n+1}\\ f_{8i+2}^{n+1}\\ f_{8i+3}^{n+1}\\ f_{8i+4}^{n+1}\\ f_{8i+5}^{n+1}\\ f_{8i+6}^{n+1}\\ f_{8i+7}^{n+1}\end{bmatrix}=\frac14\begin{bmatrix} 4 & 0 & 0 & 0 & 0\\ 1+\mu & 2(2-\mu) & \mu+\nu-1 & -2\nu & \nu\\ 0 & 4 & 0 & 0 & 0\\ -\mu & 2(1+\mu) & 2-\mu-\nu & 2\nu & -\nu\\ 0 & 0 & 4 & 0 & 0\\ -\nu & 2\nu & 2-\mu-\nu & 2(1+\mu) & -\mu\\ 0 & 0 & 0 & 4 & 0\\ \nu & -2\nu & \mu+\nu-1 & 2(2-\mu) & 1+\mu\end{bmatrix}\begin{bmatrix} f_{4i}^{n}\\ f_{4i+1}^{n}\\ f_{4i+2}^{n}\\ f_{4i+3}^{n}\\ f_{4i+4}^{n},\end{bmatrix}, \tag{2.7}$$

where $\mu:=-2\alpha(1-\beta)$ and $\nu=\mu+\beta/2$.

Proof. The result is clear for equations corresponding to even subscripts of p and f since the scheme is interpolating. Consider therefore the odd subscript equations. We will use the notation $\Delta p_k^n=p_{k+1}^n-p_k^n$, $\Delta f_k^n=f_{k+1}^n-f_k^n$ and $\Delta^2 f_k^n=\Delta(\Delta f_k^n)=f_{k+2}^n-2f_{k+1}^n+f_k^n$.

Let us start by proving (2.6). Using (2.3) with $k=2i$ and $k=2i+1$

$$\begin{aligned} p_{4i+1}^{n+1}&=(1-\beta)\frac{\Delta f_{2i}^n}{h_n}+\frac{\beta}{2}\big(p_{2i+1}^n+p_{2i}^n\big)\\ p_{4i+3}^{n+1}&=(1-\beta)\frac{\Delta f_{2i+1}^n}{h_n}+\frac{\beta}{2}\big(p_{2i+2}^n+p_{2i+1}^n\big).\end{aligned} \tag{2.8}$$

From (2.2) we obtain

$$\begin{aligned}\frac{\Delta f_{2i}^n}{h_n} &= \frac{\Delta f_i^{n-1}}{h_{n-1}} + 2\alpha\Delta p_i^{n-1},\\ \frac{\Delta f_{2i+1}^n}{h_n} &= \frac{\Delta f_i^{n-1}}{h_{n-1}} - 2\alpha\Delta p_i^{n-1}.\end{aligned} \tag{2.9}$$

The f difference on the right can be eliminated by a reordering of (2.3) with $k = i$ and $n \to n-1$

$$(1-\beta)\frac{\Delta f_i^{n-1}}{h_{n-1}} = p_{2i+1}^n - \frac{\beta}{2}\big(p_{i+1}^{n-1} + p_i^{n-1}\big). \tag{2.10}$$

Combining (2.8)–(2.10), we find

$$\begin{aligned}p_{4i+1}^{n+1} &= p_{2i+1}^n + \frac{\beta}{2}\big(p_{2i+1}^n - p_{i+1}^{n-1}\big) - \mu\Delta p_i^{n-1},\\ p_{4i+3}^{n+1} &= p_{2i+1}^n + \frac{\beta}{2}\big(p_{2i+1}^n - p_i^{n-1}\big) + \mu\Delta p_i^{n-1},\end{aligned}$$

and we obtain (2.6).

In terms of differences (2.6) takes the form

$$\begin{aligned}\Delta p_{4i}^n &= (1-\mu)\Delta p_{2i}^{n-1} - \nu\Delta p_{2i+1}^{n-1},\\ \Delta p_{4i+1}^n &= \mu\Delta p_{2i}^{n-1} + \nu\Delta p_{2i+1}^{n-1},\\ \Delta p_{4i+2}^n &= \nu\Delta p_{2i}^{n-1} + \mu\Delta p_{2i+1}^{n-1},\\ \Delta p_{4i+3}^n &= -\nu\Delta p_{2i}^{n-1} + (1-\mu)\Delta p_{2i+1}^{n-1}.\end{aligned} \tag{2.11}$$

Notice that an equivalent formulation of (2.2) is

$$\Delta^2 f_{2k}^{n+1} = -2\alpha h_n \Delta p_k^n$$

and (2.11) can be written

$$\begin{aligned}2\Delta^2 f_{8i}^{n+1} &= (1-\mu)\Delta^2 f_{4i}^n - \nu\Delta^2 f_{4i+2}^n,\\ 2\Delta^2 f_{8i+2}^{n+1} &= \mu\Delta^2 f_{4i}^n + \nu\Delta^2 f_{4i+2}^n,\\ 2\Delta^2 f_{8i+4}^{n+1} &= \nu\Delta^2 f_{4i}^n + \mu\Delta^2 f_{4i+2}^n,\\ 2\Delta^2 f_{8i+6}^{n+1} &= -\nu\Delta^2 f_{4i}^n + (1-\mu)\Delta^2 f_{4i+2}^n.\end{aligned} \tag{2.12}$$

It remains to extract the values f_{8i+j}^{n+1}, $j = 1, 3, 5, 7$ from the previous formulae to obtain (2.7). □

From (2.6) it follows that the new p-values on level $n+1$ ($n \geq 1$) can be formed by an affine combination of three p values on the previous level n. This can especially be used to simplify the proofs of two results in [13] on monotonicity and convexity of the HC^1-interpolant.

For monotonicity the HC^1-algorithm is applied in [13, section 3] to test data $(f_0, p_0, f_1, p_1) = (0, x, 1, y)$ computing the corresponding HC^1-interpolant f and its derivative p. For fixed (α, β) the authors determine the set of slopes (x, y) giving $p \geq 0$. Theorem 11 in [13] states that if $-1 < \beta < 0$ and $0 > \alpha \geq \beta/(4(1-\beta))$, then

$$M(\alpha,\beta) := \{(x,y) \in \mathbb{R}_+^2 : p \geq 0\} = \{(x,y) \in \mathbb{R}_+^2 : x + y \leq \gamma\} =: T(\gamma),$$

where $\gamma := \frac{2(\beta-1)}{\beta}$ and $\mathbb{R}^2_+ = \{(x,y) \in \mathbb{R} : x > 0,\ y > 0\}$. Note that any point in $\mathbb{R}^2_+$ belongs to $T(\gamma)$ for some $\beta < 0$. Thus we can obtain an increasing interpolant for any nonnegative initial slopes x, y by choosing β suitably close to zero. For arbitrary initial data (f_0, p_0, f_1, p_1) on $[a,b]$ one can use the change of variables $g(t) := \big(f(a + t(b-a)) - f_0\big)/(f_1 - f_0)$ to show that the HC^1-interpolant f is increasing if and only if $\big(p_0/\Delta, p_1/\Delta\big) \in M(\alpha,\beta)$, where $\Delta := (f_1 - f_0)/(b-a)$.

The proof of Theorem 11 follows immediately from (2.6). Indeed for the assumed range of (α,β) the elements in the matrix in (2.6) are all nonnegative. Thus if p is nonnegative on level $n-1$ it is nonnegative on level n. Moreover, if $(x,y) \in T(\gamma)$, then $p((a+b)/2) = 1 - \beta + \beta(x+y)/2 \geq 1 - \beta + \frac{\beta}{2}\frac{2}{\beta}(\beta - 1) = 0$ and the theorem follows. In fact the theorem holds for $-2 \leq \beta < 0$ as long as we have C^1-convergence of the HC^1-interpolant; see the next subsection for convergence results.

For convexity the HC^1-algorithm is applied to the test data $(f_0, p_0, f_1, p_1) = (0, -x, 0, y)$ with $(x,y) \in \mathbb{R}^2_+$. For fixed (α,β) the set of slopes (x,y) giving an increasing p, is determined. Theorem 18 in [13] states that if $-1 \leq \beta < 0$ and $\gamma := (\beta - 2)/\beta$, then

$$C(\alpha,\beta) := \{(x,y) \in \mathbb{R}^2_+ : p \text{ is increasing}\} = \{(x,y) \in \mathbb{R}^2_+ : x/\gamma \leq y \leq x\gamma\} =: C^*(\gamma)$$

if and only if $\alpha = \beta/(4(1-\beta))$. Since any point in $\mathbb{R}^2_+$ belongs to $C^*(\gamma)$ for β sufficiently close to zero, this result implies that we can obtain a convex HC^1-interpolant in the EQS-case by using any nonnegative values (x,y) as initial data. For arbitrary initial data (f_0, p_0, f_1, p_1) on $[a,b]$ with $h = b-a$, one can for convexity use the change of variables $g(t) := f(a+th) - (1-t)f_0 - tf_1$ to show that the HC^1-interpolant f is convex if and only if $h * \big(p_1 - \Delta, \Delta - p_0\big) \in C(\alpha,\beta)$, where as before $\Delta := (f_1 - f_0)/h$.

To prove Theorem 18 we use (2.11). For the given value of α we have $\nu = 0$ and moreover $0 \leq \mu \leq 1$. Thus p is increasing on level n if it is increasing on level $n-1$. Since $-x = \beta(\gamma - 1)x \leq \beta(y-x)/2 = p((a+b)/2) \leq \beta(y - \gamma y)/2 = y$, p is increasing on level 1 and the if part of the theorem follows. The only if part is easy; see [13, p. 293].

In the EQS-case we only need two of the three p-values on the right of (2.6). Moreover, the derivatives will be sampled from a piecewise linear curve.

COROLLARY 2.2. *In the EQS-case* $\alpha = \frac{\beta}{4(1-\beta)}$ *we have*

$$\begin{aligned} p^{n+1}_{4i+1} &= -\frac{\beta}{2}p^n_{2i} + \left(1 + \frac{\beta}{2}\right)p^n_{2i+1}, \\ p^{n+1}_{4i+3} &= \left(1 + \frac{\beta}{2}\right)p^n_{2i+1} - \frac{\beta}{2}p^n_{2i+2}. \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} 4f^{n+1}_{8i+1} &= \left(1 - \frac{\beta}{2}\right)f^n_{4i} + (4+\beta)f^n_{4i+1} - \left(1 + \frac{\beta}{2}\right)f^n_{4i+2} \\ 4f^{n+1}_{8i+3} &= \frac{\beta}{2}f^n_{4i} + (2-\beta)f^n_{4i+1} + \left(2 + \frac{\beta}{2}\right)f^n_{4i+2} \\ 4f^{n+1}_{8i+5} &= \left(2 + \frac{\beta}{2}\right)f^n_{4i+2} + (2-\beta)f^n_{4i+3} + \frac{\beta}{2}f^n_{4i+4} \\ 4f^{n+1}_{8i+7} &= -\left(1 + \frac{\beta}{2}\right)f^n_{4i+2} + (4+\beta)f^n_{4i+3} + \left(1 - \frac{\beta}{2}\right)f^n_{4i+4}. \end{aligned} \tag{2.14}$$

If in addition $\beta \in (-2, 0)$, then there exist

$$a = \tau_0^n < \tau_1^n < \cdots < \tau_{2^n}^n = b, \tag{2.15}$$

with $\tau_{2^{n-1}}^n = \frac{a+b}{2}$ for $n \geq 1$, such that

$$p_i^n = L(\tau_i^n), \quad i = 0, 1, \ldots, 2^n, \quad n = 0, 1, \ldots, \tag{2.16}$$

where L is the piecewise linear curve connecting the three points $(a, p(a))$, $(\frac{a+b}{2}, p(\frac{a+b}{2}))$, $(b, p(b))$.

Proof. If $\alpha = \frac{\beta}{4(1-\beta)}$, then $\mu = -\beta/2$ and (2.13) follows from (2.6). Similarly, we obtain (2.14).

We claim that (2.16) holds with

$$\begin{aligned} \tau_{4i}^{n+1} &= \tau_{2i}^n, \quad \tau_{4i+1}^{n+1} = -\frac{\beta}{2}\tau_{2i}^n + \left(1 + \frac{\beta}{2}\right)\tau_{2i+1}^n, \\ \tau_{4i+2}^{n+1} &= \tau_{2i+1}^n, \quad \tau_{4i+3}^{n+1} = -\frac{\beta}{2}\tau_{2i+2}^n + \left(1 + \frac{\beta}{2}\right)\tau_{2i+1}^n. \end{aligned} \tag{2.17}$$

Since $p_0^n = p(a)$ and $p_{2^n}^n = p(b)$, we have $\tau_0^n = a$ and $\tau_{2^n}^n = b$ for all $n \geq 0$. Moreover, since $p_{2^{n-1}}^n = p(\frac{a+b}{2})$, we see that $\tau_{2^{n-1}}^n = \frac{a+b}{2}$ for all $n \geq 1$. Thus (2.15) will follow from (2.17) since the latter involves convex combinations for $\beta \in (-2, 0)$; (2.17) follows from (2.13) by induction. Suppose (2.16) holds for some n. Since L is linear on the actual segment we obtain

$$p_{4i+1}^{n+1} = -\frac{\beta}{2}L(\tau_{2i}^n) + \left(1 + \frac{\beta}{2}\right)L(\tau_{2i+1}^n) = L(\tau_{4i+1}^{n+1}),$$

where τ_{4i+1}^{n+1} is given by (2.17). The proof of the other τ-relation is similar. ∎

2.3. C^1-convergence. To study convergence we observe that it is enough to consider the interval $[0, 1]$. Indeed, if $I := [a, b]$ and $h := b - a$, defining the initial data $g(u) = f(a + hu)$, $g'(u) = hf'(a + hu)$, for $u \in \{0, 1\}$, the construction of f on $[a, b]$ or g on $[0, 1]$ by (2.1) are equivalent and at step n, $g(u) = f(a + uh)$ and $g'(u) = hf'(a + hu)$ for $u \in \{0, 1/2^n, \ldots, \ell/2^n, \ldots, 1\}$.

In [12] it was shown that if there exist positive constants c, ρ with $\rho < 1$ such that for each integer $n \geq 0$ we have $|\Delta p_i^n| \leq c\rho^n$ for $i = 0, 1, \ldots, 2^n - 1$, where

$$\Delta p_i^n := p\left(\frac{i+1}{2^n}\right) - p\left(\frac{i}{2^n}\right), \quad i = 0, 1, \ldots, 2^n - 1, \tag{2.18}$$

then p has a unique continuous extension to I. Moreover, there is a positive constant c_1 such that for all $(x, y) \in [0, 1]^2$

$$|p(x) - p(y)| \leq c_1 |x - y|^{-\log_2 \rho},$$

i.e., p is Hölder continuous with exponent $-\log_2 \rho$.

Suppose p is continuous and $\lim_{n\to\infty} \max_{0 \leq i < 2^n - 1} |\Delta(f, p)_i^n| = 0$, where

$$\Delta(f, p)_i^n := 2^n \Delta f_i^n - \sigma p_i^n, \quad \sigma p_i^n := \frac{1}{2}\left(p\left(\frac{i+1}{2^n}\right) + p\left(\frac{i}{2^n}\right)\right), \tag{2.19}$$

and $\Delta f_i^n = f\left(\frac{i+1}{2^n}\right) - f\left(\frac{i}{2^n}\right)$. Then [12] f has a unique continuous extension to $I := [0,1]$. Moreover, $f \in C^1([0,1])$ with $f' = p$. From this discussion we have the following lemma.

LEMMA 2.3. *Let* $U_i^n := [\Delta p_i^n, \Delta(f,p)_i^n]^T$ *for* $i = 0,1,\ldots,2^n-1$ *and* $n = 0,1,2,\ldots$. *If we can find a vector norm* $\|\cdot\|$ *on* $\mathbb{R}^2$ *and positive constants* c, ρ *with* $\rho < 1$ *such that*

$$\|U_i^n\| \le c\rho^n,\ i = 0,1,\ldots,2^n-1 \text{ and } n = 0,1,\ldots,$$

then the HC^1*-algorithm is* C^1*-convergent and* $f' = p$ *is Hölder continuous with exponent* $-\log_2 \rho$.

We can now show the following proposition.

PROPOSITION 2.4. *Algorithm* HC^1 *is* C^1*-convergent for* $(\alpha,\beta) \in [-1/8,0) \times [-2,1)$.

Proof. An immediate evaluation gives

$$U_{2i}^{n+1} = \Lambda_1 U_i^n \text{ and } U_{2i+1}^{n+1} = \Lambda_{-1} U_i^n \text{ for } i = 0,1,\ldots,2^n, n = 0,1,\ldots,$$

where

$$\Lambda_\epsilon = \begin{bmatrix} \frac{1}{2} & \epsilon(1-\beta) \\ \epsilon\frac{8\alpha+1}{4} & \frac{1+\beta}{2} \end{bmatrix}, \quad \varepsilon = \pm 1. \tag{2.20}$$

Note that the off-diagonal elements of Λ_ε have the same sign for $\alpha \ge -1/8$ and $\beta \le 1$. We define a vector norm by $\|v\| := \|P^{-1}v\|_2$, where $\|\cdot\|_2$ is the usual Euclidian vector norm and $P := \begin{bmatrix} 2\sqrt{1-\beta} & 0 \\ 0 & \sqrt{8\alpha+1} \end{bmatrix}$. Then $P^{-1}\Lambda_\varepsilon P$ is symmetric and the corresponding matrix operator norm is given by $\|\Lambda_\varepsilon\| := \|P^{-1}\Lambda_\varepsilon P\|_2$, where $\|A\|_2 := \sqrt{\rho(A^TA)}$ is the spectral norm of a matrix A. The eigenvalues of Λ_ε or of $P^{-1}\Lambda_\varepsilon P$ are

$$\lambda_1 = \frac{1}{4}\big(2+\beta+\sqrt{(2-\beta)^2+32\alpha(1-\beta)}\big), \quad \lambda_2 = \frac{1}{4}\big(2+\beta-\sqrt{(2-\beta)^2+32\alpha(1-\beta)}\big).$$

Since $P^{-1}\Lambda_\varepsilon P$ is symmetric the eigenvalues are real with $\lambda_2 < \lambda_1$. Now for $\beta \in [-2,1)$ and $\alpha \in [-1/8,0)$ we find $\lambda_1 < (2+\beta+\sqrt{(2-\beta)^2})/4 = 1$ and $\lambda_2 > (2+\beta-\sqrt{(2-\beta)^2})/4 = \beta/2 \ge -1$. Thus $\rho := \|\Lambda_\varepsilon\| = \max\{|\lambda_1|,|\lambda_2|\} < 1$ for $\varepsilon = \pm 1$ and we have shown that $\max\{\|U_{2i}^{n+1}\|, \|U_{2i+1}^{n+1}\|\} \le \rho\|U_i^n\|$ so that $\|U_i^n\| \le \rho^n\|U_0^0\|$ for $i = 0,1,\ldots,2^n-1$ and $n = 0,1,2,\ldots$. The C^1-convergence for $(\alpha,\beta) \in [-1/8,0)\times[-2,1)$ now follows from Lemma 2.3. ∎

By Proposition 2.4, the HC^1-algorithm converges for $\beta \in [-1,0)$ if $\alpha = \frac{\beta}{4(1-\beta)}$. We can now extend this result.

PROPOSITION 2.5. *If* $\alpha = \frac{\beta}{4(1-\beta)}$, *then the* HC^1*-algorithm is* C^1*-convergent for* $\beta \in (-2,0)$.

Proof. For $\varepsilon = \pm 1$ the matrices Λ_ε in (2.20) take the form

$$\Lambda_\varepsilon = \begin{pmatrix} \frac{1}{2} & \varepsilon(1-\beta) \\ \varepsilon\frac{\beta+1}{4(1-\beta)} & \frac{1+\beta}{2} \end{pmatrix}.$$

Now, for any positive real number θ, we define the norm $\|\cdot\|_\theta$ on $\mathbb{R}^2$ by $\|(x,y)\|_\theta = |x|+\theta|y|$. It is easy to prove that for any matrix $M = (m_{ij}) \in \mathbb{R}^{2\times 2}$, the corresponding matrix operator norm is given by $\|M\|_\theta := \max(|m_{11}| + \theta|m_{21}|, |m_{12}|/\theta + |m_{22}|)$. Choosing $\theta = 2(1-\beta)$ we find $\|\Lambda_1\|_\theta = \|\Lambda_{-1}\|_\theta = 1/2(1+|1+\beta|)$, which is stricly less than one for $-2 < \beta < 0$. Lemma 2.3 now gives the convergence. ∎

We define the convergence region C by

$$C := \{(\alpha,\beta) : \text{ the scheme } HC^1 \text{ is } C^1\text{-convergent}\}. \tag{2.21}$$

We have shown that $[-1/8, 0) \times [-2, 1) \subset C$ and also that $\{(\frac{\beta}{4(1-\beta)}, \beta) : -2 < \beta < 0\} \subset C$.

The function $f' = p$ is Hölder continuous with exponent $-\log_2 \rho$. In the case where $\alpha = \frac{\beta}{4(1-\beta)}$ we have $\|\Lambda_1\|_\theta = \|\Lambda_{-1}\|_\theta = \rho = \rho(\beta) = 1/2(1+|1+\beta|)$ which is piecewise linear with a minimum for $\beta = -1$ and we obtain the best regularity of the interpolant for $\beta = -1$ when f is a quadratic spline.

To illustrate the smoothness properties of a HC^1-interpolant we show the Hermite basis with $\beta = -3/5$ and $\alpha = \frac{\beta}{4(1-\beta)} = -3/32$ in Figure 2.1. The spectral radius of the matrices Λ_ε is $7/10$ and hence the derivatives of the Hermite basis functions are Hölder continuous with exponent $\rho = -\log_2(7/10) \approx 0.5146$.

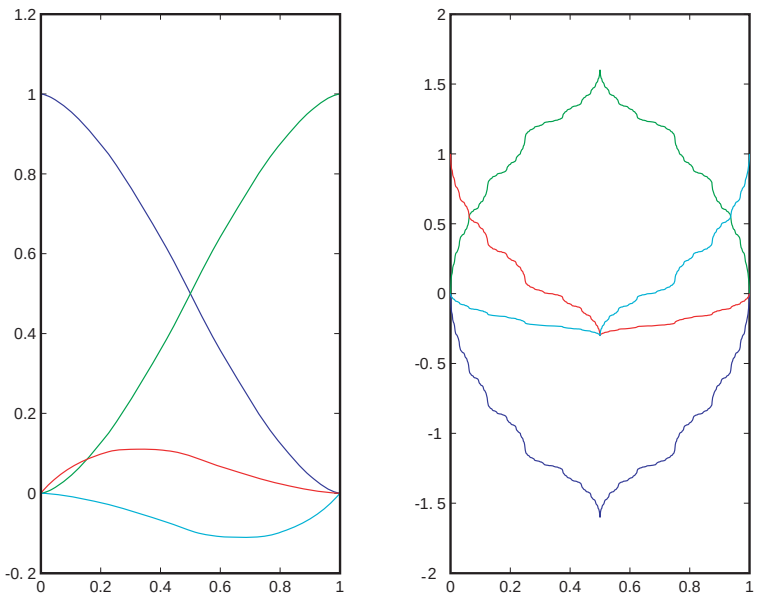


FIG. 2.1. *Hermite basis and derivatives, corresponding to* $\alpha = -3/32$ *and* $\beta = -3/5$.

Remark. The data $f(a), p(a), f(b), p(b)$ can either have real values or vector values in $\mathbb{R}^s$, $s \geq 2$. In this second case, we look for vector continuous functions f and p with $f' = p$ from $I = [a,b]$ to $\mathbb{R}^s$. The C^1-convergence is guaranteed for all (α,β) in the convergence region C since it suffices to study the convergence independently for each component of f and p.

3. Control polygons and subdivision algorithm.

3.1. Control coefficients and control polygons. Suppose we apply the subdivision scheme HC^1 to some real valued data $f(a)$, $p(a)$, $f(b)$, $p(b)$. In order to obtain a geometric formulation of the scheme we define *control coefficients* relative to the interval $[a,b]$ by

$$a_0 = f(a), \quad a_1 = f(a) + \frac{h}{\lambda}p(a), \quad a_2 = f(b) - \frac{h}{\lambda}p(b), \quad a_3 = f(b), \tag{3.1}$$

where $h := b - a$ and $\lambda \geq 2$ is a real number to be chosen. We define the *control points* $(A_0,\ A_1,\ A_2,\ A_3)$ on $[a,b]$ by

$$A_0 = (a, a_0), \quad A_1 = \left(a + \frac{h}{\lambda}, a_1\right), \quad A_2 = \left(b - \frac{h}{\lambda}, a_2\right), \quad A_3 = (b, a_3), \tag{3.2}$$

and the *control polygon* $\{A_0,\ A_1,\ A_2,\ A_3\}$ on $[a,b]$ by connecting the four control points by straight line segments. If f is the HC^1-interpolant, then the parametric curve $(x, f(x))$ with $x \in [a,b]$ passes through A_0 with tangent directions $A_1 - A_0$ and A_3 with tangent direction $A_3 - A_2$; see Figure 3.1.

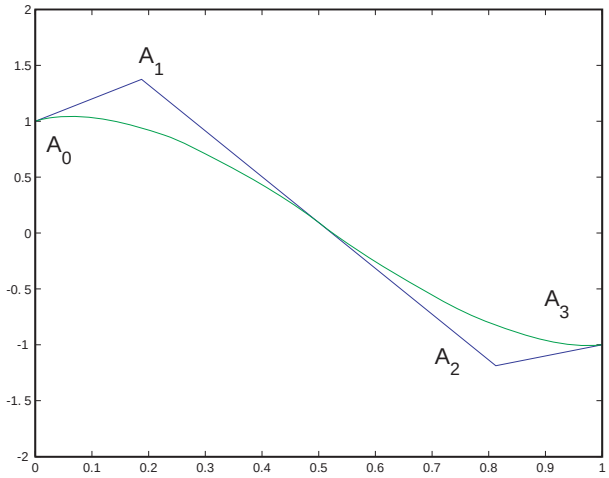


FIG. 3.1. *A HC^1-interpolant and its control polygon, $\beta = -3/5$, $\alpha = -3/32$, $\lambda = 16/3$.*

We can also apply the subdivision scheme HC^1 to vector valued data f_0, p_0, f_1, p_1 in $\mathbb{R}^s$ for some $s \geq 2$. We pick an interval $[a,b]$ and use the HC^1-algorithm on each component of f and p. To obtain a geometric formulation of this process we define control coefficients relative to $[a,b]$ by (3.1) and we define the control points to be the same as the control coefficients. The computed curve interpolates the first and last control coefficient and its tangent direction at a_0 is $a_1 - a_0$, and at a_3 the tangent direction is $a_3 - a_2$.

Note that if 4 points a_0, a_1, a_2, a_3 in $\mathbb{R}^s$ for $s \geq 1$ are given we can think of these as control coefficients of a HC^1-interpolant on some finite interval $[a,b]$ and apply the HC^1-algorithm to the data given by

$$f(a) := a_0,\ p(a) := \frac{\lambda}{h}(a_1 - a_0),\ f(b) := a_3,\ p(b) := \frac{\lambda}{h}(a_3 - a_2), \tag{3.3}$$

where $h := b-a$. We now derive a parameter independent formulation of this scheme. In particular suppose (a_0, a_1, a_2, a_3) are points in $\mathbb{R}^s$ for some $s \geq 1$ which are distinct if $s \geq 2$ and let $[a, b]$ be any finite interval.

Using (3.1) and (3.3) we can compute new control coefficients $(\bar{a}_0, \bar{a}_1, \bar{a}_2, \bar{a}_3)$ for the interval I_1 and new control coefficients $(\bar{a}_3, \bar{a}_4, \bar{a}_5, \bar{a}_6)$ for I_2, and then join them into control coefficients $(\bar{a}_0, \bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4, \bar{a}_5, \bar{a}_6)$ on $[a, b]$. In the following geometric formulation of the subdivision scheme we do this computation directly without picking an underlying interval $[a, b]$. The proposition is a generalization of [15, Theorem 10].

PROPOSITION 3.1. *Suppose $a_i \in \mathbb{R}^s$ for $i = 0, 1, 2, 3$ and some $s \geq 1$. After one subdivision of the control coefficients (a_0, a_1, a_2, a_3) we obtain new control coefficients $(\bar{a}_0, \bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4, \bar{a}_5, \bar{a}_6)$ given by*

$$\begin{bmatrix} \bar{a}_0 \\ \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ \bar{a}_4 \\ \bar{a}_5 \\ \bar{a}_6 \end{bmatrix} = \mathbf{S} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} := \frac{1}{4} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ \gamma & v-\beta & v+\beta & \delta \\ 2-v & v & v & 2-v \\ \delta & v+\beta & v-\beta & \gamma \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}, \tag{3.4}$$

where

$$\begin{aligned} v &= -4\alpha\lambda, \\ \gamma &= 2 - v + (2 + \beta(\lambda - 2))/\lambda, \\ \delta &= 2 - v - (2 + \beta(\lambda - 2))/\lambda. \end{aligned} \tag{3.5}$$

Moreover,

$$\bar{a}_3 = \frac{1}{2}(\bar{a}_2 + \bar{a}_4). \tag{3.6}$$

Proof. Pick any interval $[a, b]$ and let $h := b - a$. By (3.1)

$$\bar{a}_0 = f(a),\ \bar{a}_1 = f(a) + \frac{h}{2\lambda}p(a),\ \bar{a}_2 = f\left(\frac{a+b}{2}\right) - \frac{h}{2\lambda}p\left(\frac{a+b}{2}\right),\ \bar{a}_3 = f\left(\frac{a+b}{2}\right),$$
$$\bar{a}_6 = f(b),\ \bar{a}_5 = f(b) - \frac{h}{2\lambda}p(b),\ \bar{a}_4 = f\left(\frac{a+b}{2}\right) + \frac{h}{2\lambda}p\left(\frac{a+b}{2}\right).$$

From (2.1) and (3.3) we obtain on an interval $[a, b]$ the inverse relations

$$\begin{aligned} f(a) = a_0, \quad p(a) &= \frac{\lambda}{h}(a_1 - a_0), \\ f(b) = a_3, \quad p(b) &= \frac{\lambda}{h}(a_3 - a_2), \\ f\left(\frac{a+b}{2}\right) &= \frac{a_0 + a_3}{2} - \frac{v}{4}(a_3 - a_2 - a_1 + a_0), \\ \frac{h}{2\lambda}p\left(\frac{a+b}{2}\right) &= \frac{1-\beta}{2\lambda}(a_3 - a_0) + \frac{\beta}{4}(a_3 - a_2 + a_1 - a_0), \\ &= \frac{2 + \beta(\lambda - 2)}{\lambda}(a_3 - a_0) + \frac{\beta}{4}(a_1 - a_2). \end{aligned} \tag{3.7}$$

But then we see that $(\bar a_0, \bar a_1, \bar a_2, \bar a_3, \bar a_4, \bar a_5, \bar a_6)^T = \mathbf{S}(a_0, \dots, a_3)^T$, where $\mathbf{S}$ is the matrix in (3.4). Since the sum of rows three and five in the matrix $\mathbf{S}$ equals twice row four the relation (3.6) follows. □

For $s \geq 2$ the control coefficients and control points are the same and the proposition also gives rules for subdividing the control polygon. The following corollary holds in general.

COROLLARY 3.2. *Suppose $(a_0, a_1, a_2, a_3) \in \mathbb{R}^s$ for some $s \geq 1$. After one subdivision of the corresponding control polygon $\{A_0, A_1, A_2, A_3\}$ we obtain a new control polygon $\{\bar A_0, \bar A_1, \bar A_2, \bar A_3, \bar A_4, \bar A_5, \bar A_6\}$ given by*

$$\begin{bmatrix}\bar A_0 & \bar A_1 & \bar A_2 & \bar A_3 & \bar A_4 & \bar A_5 & \bar A_6\end{bmatrix}^T = \mathbf{S}\begin{bmatrix}A_0 & A_1 & A_2 & A_3\end{bmatrix}^T, \tag{3.8}$$

where $\mathbf{S}$ is given by (3.4). *Moreover,*

$$\bar A_3 = \frac{1}{2}(\bar A_2 + \bar A_4), \tag{3.9}$$

which means that these control points always lie on a straight line.

Proof. This has already been shown for $s \geq 2$ and for the control coefficients for $s = 1$. For the control point abscissas we obtain the relation $(a, a + h/(2\lambda), \bar a - h/(2\lambda), \bar a, \bar a + h/(2\lambda), b - h/(2\lambda), d)^T = \mathbf{S}(a, a + h/\lambda, b - h/\lambda, d)^T$, where $\bar a = (a+b)/2$, since the scheme HC^1 reproduces linear functions. Thus (3.8) and (3.9) also holds for $s = 1$. □

3.2. A stationary subdivision algorithm. By applying (3.4), we can reformulate the Hermite subdivision scheme HC^1 as a stationary subdivision scheme working on points in $\mathbb{R}^s$.

Starting with 4 points a_0, a_1, a_2, a_3 in $\mathbb{R}^s$, $s \geq 1$, (α, β) in the convergence region C, and $\lambda \geq 2$, we define Algorithm SC^1 as follows.

At step $n = 0$, we set $a_0^0 = a_0$, $a_1^0 = a_1$, $a_2^0 = a_2$, $a_3^0 = a_3$.

At step $n + 1$, $n \geq 0$, we define

$$\begin{bmatrix} a_{6i}^{n+1} \\ a_{6i+1}^{n+1} \\ a_{6i+2}^{n+1} \\ a_{6i+3}^{n+1} \\ a_{6i+4}^{n+1} \\ a_{6i+5}^{n+1} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ \gamma & v-\beta & v+\beta & \delta \\ 2-v & v & v & 2-v \\ \delta & v+\beta & v-\beta & \gamma \\ 0 & 0 & 2 & 2 \end{bmatrix}\begin{bmatrix} a_{3i}^n \\ a_{3i+1}^n \\ a_{3i+2}^n \\ a_{3i+3}^n \end{bmatrix}, \quad i = 0, 1, \dots, 2^n - 1 \tag{3.10}$$

and $a_{3.2^{n+1}}^{n+1} = a_{3.2^n}^n$. Here v, γ, δ are given by (3.5). The matrix $(s_{\ell,k})_{\ell=0,\dots,5,k=0,\dots,3}$ in (3.10) is formed from the first 6 rows of $\mathbf{S}$ given by (3.4).

LEMMA 3.3. *For all $n \geq 1$ and for all $i = 1, \dots, 2^n - 1$, we have*

$$\begin{aligned} a_{6i}^{n+1} &= a_{3i}^n, \ i = 1, \dots, 2^{n-1}, \\ a_{6i+1}^{n+1} - a_{6i}^{n+1} &= \frac{1}{2}\big(a_{3i+1}^n - a_{3i}^n\big), \ i = 1, \dots, 2^{n-1} - 1, \\ a_{3i+1}^n + a_{3i-1}^n &= 2a_{3i}^n, \ i = 1, \dots, 2^n - 1. \end{aligned} \tag{3.11}$$

Proof. The first two equations follow immediately from (3.10). As in the proof of (3.6) it is clear that

$$a_{6i+2}^{n+1} + a_{6i+4}^{n+1} = 2a_{6i+3}^{n+1}, \quad i = 0, \dots, 2^n - 1, \quad n = 0, 1, \dots,$$

and in particular $a_2^1 + a_4^1 = 2a_3^1$. By (3.11) and induction on n

$$a_{6i+1}^{n+1} + a_{6i-1}^{n+1} = \frac{1}{2}(a_{3i}^n + a_{3i+1}^n) + \frac{1}{2}(a_{3i-1}^n + a_{3i}^n) = 2a_{3i}^n = 2a_{6i}^{n+1}. \qquad \square$$

If we define a_i^0 for $i < 0$ and $i > 3$ in any way, the subdivision scheme can be written $a_\ell^{n+1} = \sum_{k\in\mathbb{Z}} \sigma_{\ell,k} a_k^n$, $\ell \in \mathbb{Z}$, where $\sigma_{6i+\ell,3i+k} = s_{\ell,k}$ for $i \in \mathbb{Z}$, $\ell = 0, \ldots, 5$, $k = 0, \ldots, 3$ and $\sigma_{i,j} = 0$ otherwise. With the definitions recalled in [3], the scheme is *local* since $\sigma_{\ell,k} = 0$ for $|\ell - 2k| > 4$. Since $\sum_{k\in\mathbb{Z}} \sigma_{\ell,k} = 1$, it is *affine* but is not interpolating in a classical sense since we generally have $a_{6i+2}^{n+1} \neq a_{3i+1}^n$.

3.3. Convergence of SC^1. The convergence of the subdivision schemes are usually established by studying the difference sequence. Alternatively convergence follows since SC^1 was derived from HC^1. Here are the details.

THEOREM 3.4. *Let $s \geq 1$ and a_0, a_1, a_2, a_3 be 4 points in $\mathbb{R}^s$. Suppose $\lambda \geq 2$ and that (α, β) is in the convergence region C given by (2.21). We build the sequence of points $\{a_i^n\}_{n\in\mathbb{N}, i=0,\ldots,3.2^n}$ by (3.10). Choose any interval $I := [a, b]$ with $h := b - a > 0$ and define $t_i^n := a + ih_n$, where $h_n := h2^{-n}$ for $n \in \mathbb{N}$ and $i = 0, \ldots, 2^n$. Then, there exists a C^1 function $f : I \to \mathbb{R}^s$ such that for all $n \in \mathbb{N}$:*

$$\begin{aligned} a_{3i}^n &= f(t_i^n), \quad i = 0, \ldots, 2^n, \\ a_{3i+1}^n - a_{3i}^n &= \frac{h_n}{\lambda} f'(t_i^n), \quad i = 0, \ldots, 2^n - 1, \\ a_{3i}^n - a_{3i-1}^n &= \frac{h_n}{\lambda} f'(t_i^n), \quad i = 1, \ldots, 2^n. \end{aligned}$$

For $s \geq 2$, let $A_i^n = a_i^n$, $i = 0, 1, \ldots, 3 \times 2^n$ and for $s = 1$, let $A_{3i}^n = (t_i^n, a_{3i}^n)$, $A_{3i+1}^n = (t_i^n + \frac{h_n}{\lambda}, a_{3i+1}^n)$, $A_{3i+2}^n = (t_{i+1}^n - \frac{h_n}{\lambda}, a_{3i+2}^n)$, $i = 0, 1, \ldots, 2^n - 1$, and $A_{3\times 2^n}^n = (b, a_{3\times 2^n}^n)$.

Then the sequence of polygons $\{A_0^n, \ldots, A_{3\times 2^n}^n\}$ converges to the curve $\{f(t), t \in I\}$.

Proof. We will show that the scheme SC^1 generates sequences $\{f^n\}$ and $\{p^n\}$ of piecewise linear vector functions which interpolate values and derivatives at the points of $\mathcal{P}_n = \{t_0^n, \ldots, t_{2^n}^n\}$.

We define f^n and p^n to be linear on $[t_i^n, t_{i+1}^n]$, $i = 0, \ldots, 2^n - 1$, and to interpolate the following values:

$$\begin{aligned} f^n(t_i^n) = a_{3i}^n, \quad p^n(t_i^n) &= \frac{\lambda}{h_n}(a_{3i+1}^n - a_{3i}^n), \quad i = 0, \ldots, 2^n - 1, \\ f^n(b) = a_{3.2^n}^n, \quad p^n(b) &= \frac{\lambda}{h_n}(a_{3\times 2^n}^n - a_{3\times 2^n-1}^n). \end{aligned} \tag{3.12}$$

Since $t_i^n = t_{2i}^{n+1}$ we find from (3.11) and (3.12)

$$f^{n+1}(t_i^n) = f^n(t_i^n), \quad p^{n+1}(t_i^n) = p^n(t_i^n), \ i = 0, \ldots, 2^n. \tag{3.13}$$

Below we prove that, for $i = 0, \ldots, 2^n - 1$,

$$f^{n+1}(t_{2i+1}^{n+1}) = \frac{f^n(t_{i+1}^n) + f^n(t_i^n)}{2} + \alpha h_n (p^n(t_{i+1}^n) - p^n(t_i^n)), \tag{3.14}$$

$$p^{n+1}(t_{2i+1}^{n+1}) = (1 - \beta)\frac{f^n(t_{i+1}^n) - f^n(t_i^n)}{h_n} + \beta \frac{p^n(t_{i+1}^n) + p^n(t_i^n)}{2}. \tag{3.15}$$

Comparing (3.13), (3.14), and (3.15) with (2.2)–(2.3) we conclude that $f^n = f$ and $p^n = p$ on $\mathcal{P}_n$, where f and p are the functions built on $\cup\mathcal{P}_n$ by HC^1 defined by (2.2)–(2.4) from the initial data $f(a) = a_0$, $p(a) = \frac{\lambda}{h}(a_1 - a_0)$, $f(b) = a_3$, and $p(b) = \frac{\lambda}{h}(a_3 - a_2)$, and then extended to $[a, b]$. So that if $(\alpha, \beta) \in C$, then the sequences f^n and p^n defined from SC^1 by (3.12) converge uniformly to continuous vector functions f and p defined on $[a, b]$. Moreover, $f \in C^1([a, b])$ and $f' = p$.

Now since f' is bounded and $a^n_{3i+1} - a^n_{3i} = \frac{h}{\lambda 2^n} f'(t^n_i)$, $i = 0, \ldots, 2^n - 1$, we deduce that $a^n_{3i+1} - a^n_{3i}$ tends uniformly to 0. We conclude that the sequence of polygons $\{A_0, \ldots, A_{3\times 2^n}\}$ tends to the curve $\{f(t),\, t \in I\}$ since $a^n_{3i} = f(t^n_i)$ for $i = 0, \ldots, 2^n$.

It remains to prove (3.14) and (3.15). Since $\alpha = -v/4\lambda$, for $i = 0, \ldots, 2^n - 1$ and using (3.12) and (3.10),

$$\begin{aligned}
&\frac{1}{2}(f^n(t^n_{i+1}) + f^n(t^n_i)) + \alpha h_n(p^n(t^n_{i+1}) - p^n(t^n_i)) \\
&\quad = \frac{1}{2}(a^n_{3i+3} + a^n_{3i}) - \frac{v}{4}(a^n_{3i+3} - a^n_{3i+2} - a^n_{3i+1} + a^n_{3i}) \\
&\quad = a^{n+1}_{6i+3} = f^{n+1}(t^{n+1}_{2i+1})
\end{aligned}$$

so that (3.14) is proved.

Similarly, for (3.15), let $i \in \{0, \ldots, 2^n - 1\}$. With the definitions of γ and δ in (3.5) we find

$$\begin{aligned}
&\frac{1-\beta}{h_n}(f^n(t^n_{i+1}) - f^n(t^n_i)) + \frac{\beta}{2}(p^n(t^n_{i+1}) + p^n(t^n_i)) \\
&\qquad = \frac{1-\beta}{h_n}(a^n_{3i+3} - a^n_{3i}) + \frac{\beta\lambda}{2h_n}(a^n_{3i+3} - a^n_{3i+2} + a^n_{3i+1} - a^n_{3i}) \\
&\qquad = \frac{\lambda}{h_{n+1}}\left(-\left(\frac{1-\beta}{2\lambda} + \frac{\beta}{4}\right)a^n_{3i} + \frac{\beta}{4}a^n_{3i+1} - \frac{\beta}{4}a^n_{3i+2} + \left(\frac{1-\beta}{2\lambda} + \frac{\beta}{4}\right)a^n_{3i+3}\right) \\
&\qquad = \frac{1}{4}\frac{\lambda}{h_{n+1}}\big((\delta - 2 + v)a^n_{3i} + \beta a^n_{3i+1} - \beta a^n_{3i+2} + (\gamma - 2 + v)a^n_{3i+3}\big) \\
&\qquad = \frac{\lambda}{h_{n+1}}\big(a^{n+1}_{6i+4} - a^{n+1}_{6i+3}\big) = p^{n+1}(t^{n+1}_{2i+1}). \qquad \square
\end{aligned}$$

4. Total positivity and consequences.

4.1. Corner cutting and total positivity of the subdivision matrix. Consider now the subdivision process in the EQS-case when $\alpha = \frac{\beta}{4(1-\beta)}$ with $\beta \in (-2, 0)$. Since $v = -4\alpha\lambda = \frac{\beta}{\beta-1}\lambda$, or $\lambda = \frac{\beta-1}{\beta}v$ we find from (3.5)

$$\gamma = 2 - v + \frac{2 + \beta\lambda - 2\beta}{\lambda} = 2 - v + \frac{2 + (\beta-1)v - 2\beta}{(\beta-1)v}\beta = \frac{(2-v)(v-\beta)}{v},$$

and similarly

$$\delta = \frac{(2-v)(v+\beta)}{v}.$$

Thus the subdivision matrix (3.4) can be written

$$\mathbf{S} = \frac{1}{4}\begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ \frac{(2-v)(v-\beta)}{v} & v-\beta & v+\beta & \frac{(2-v)(v+\beta)}{v} \\ 2-v & v & v & 2-v \\ \frac{(2-v)(v+\beta)}{v} & v+\beta & v-\beta & \frac{(2-v)(v-\beta)}{v} \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}. \tag{4.1}$$

In this case, as soon as $1 \le v \le 2$ and $v \ge -\beta$, we can compute the subdivided control points

$$(\bar{A}_0, \bar{A}_1, \bar{A}_2, \bar{A}_3, \bar{A}_4, \bar{A}_5, \bar{A}_6)^T = \mathbf{S}(A_0, A_1, A_2, A_3)^T$$

by successive convex combinations starting with the polygon defined by (A_0, A_1, A_2, A_3). With 2 intermediate quantities B and C we have (see Figure 4.1)

$$\begin{aligned} &\bar{A}_0 = A_0, \quad \bar{A}_1 = \frac{1}{2}A_0 + \frac{1}{2}A_1, \quad \bar{A}_5 = \frac{1}{2}A_2 + \frac{1}{2}A_3, \quad \bar{A}_6 = A_3, \\ &B = (1 - \frac{v}{2})A_0 + \frac{v}{2}A_1, \\ &C = (1 - \frac{v}{2})A_3 + \frac{v}{2}A_2, \\ &\bar{A}_2 = \frac{v-\beta}{2v}B + \frac{v+\beta}{2v}C, \\ &\bar{A}_4 = \frac{v+\beta}{2v}B + \frac{v-\beta}{2v}C, \\ &\bar{A}_3 = \frac{1}{2}\bar{A}_2 + \frac{1}{2}\bar{A}_4. \end{aligned} \tag{4.2}$$

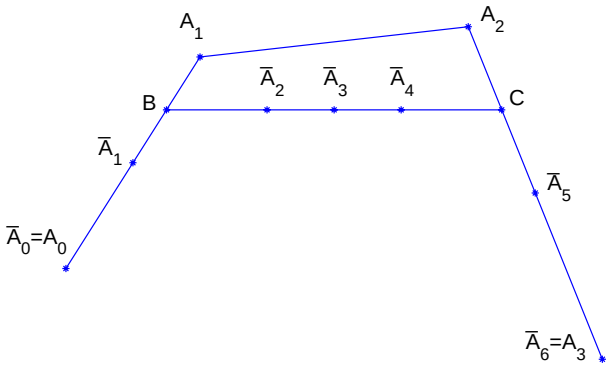


FIG. 4.1. *Corner cutting with* $\alpha = -3/32$, $\beta = -3/5$ *and* $v = 1.5$.

For $v = 2$ we obtain $B = A_1$ and $C = A_2$. The value of λ corresponding to $v = 2$ was considered in [15, Theorem 10], where formulae similar to (4.2) were given.

The equations (4.2) can be formulated as a corner cutting scheme in the following way. We start with the polygon $\{A_0, A_1, A_2, A_3\}$ and then either cut one of the previous corners or break an edge in the following sequence of convex combinations:

1. $B = (1-\frac{v}{2})A_0 + \frac{v}{2}A_1$ (replace A_1 by B to obtain $\{A_0, B, A_2, A_3\}$).
2. $C = (1-\frac{v}{2})A_3 + \frac{v}{2}A_2$ (replace A_2 by C to obtain $\{A_0, B, C, A_3\}$).
3. $\bar{A}_1 = (1-\frac{1}{v})A_0 + \frac{1}{v}B$ (break $[A_0, B]$ to obtain $\{A_0, \bar{A}_1, B, C, A_3\}$).
4. $\bar{A}_5 = \frac{1}{v}C + (1-\frac{1}{v})A_3$ (break $[C, A_3]$ to obtain $\{A_0, \bar{A}_1, B, C, \bar{A}_5, A_3\}$).
5. $\bar{A}_2 = \frac{v-\beta}{2v}B + \frac{v+\beta}{2v}C$ (replace B by $\bar{A}_2$ to obtain $\{A_0, \bar{A}_1, \bar{A}_2, C, \bar{A}_5, A_3\}$).
6. $\bar{A}_4 = \frac{v+\beta}{v-\beta}\bar{A}_2 - \frac{2\beta}{v-\beta}C$ (replace C by $\bar{A}_4$ to obtain $\{A_0, \bar{A}_1, \bar{A}_2, \bar{A}_4, \bar{A}_5, A_3\}$).
7. $\bar{A}_3 = (\bar{A}_2 + \bar{A}_4)/2$ (break $[\bar{A}_2, \bar{A}_4]$ to obtain $\{A_0, \bar{A}_1, \bar{A}_2, \bar{A}_3, \bar{A}_4, \bar{A}_5, A_3\}$).

Since $\bar{A}_0 = A_0$ and $\bar{A}_6 = A_3$ we have obtained the subdivided polygon $\{\bar{A}_0, \bar{A}_1, \bar{A}_2, \bar{A}_3, \bar{A}_4, \bar{A}_5, \bar{A}_6\}$ by carrying out a sequence of simple corner cuts (see, for example, [14, 8]) on the polygon defined by $\{A_0, A_1, A_2, A_3\}$.

We recall that a matrix is totally positive if all minors are nonnegative [1]. Then we obtain the following theorem.

THEOREM 4.1. *Suppose* $-2 < \beta < 0$, $1 \le v := \frac{\lambda\beta}{\beta-1} \le 2$, *and* $\lambda \ge 1-\beta$. *Then the matrix* $\mathbf{S}$ *given by* (4.1) *is totally positive. For each* $v \notin [1,2]$ *there is a* $\beta \in [-1,0[$ *such that* $\mathbf{S}$ *is not totally positive.*

Proof. The sequence of simple corner cuts corresponds to a factorization of $\mathbf{S}$ into a product of 7 matrices as follows:

$$\mathbf{S} = \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&\frac{1}{2}&\frac{1}{2}&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&\frac{v+\beta}{v-\beta}&\frac{-2\beta}{v-\beta}&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&\frac{v-\beta}{2v}&\frac{v+\beta}{2v}&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{bmatrix}$$

$$\begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&\frac{1}{v}&1-\frac{1}{v}\\ 0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0\\ 1-\frac{1}{v}&\frac{1}{v}&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&\frac{v}{2}&1-\frac{v}{2}\\ 0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0\\ 1-\frac{v}{2}&\frac{v}{2}&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix}.$$

Since these matrices are bidiagonal and the entries are nonnegative for the indicated values of the parameters it is well known that each of the 7 matrices are totaly positive (see, for example, [8]). Since a product of totally positive matrices is totally positive we conclude that $\mathbf{S}$ is totally positive.

If $v \notin [1,2]$, then we can find $\beta \in [-1,0)$ such that $\mathbf{S}$ has at least one negative entry. Hence $\mathbf{S}$ is not totally positive for these v, β. ∎

4.2. The HC^1-Bernstein basis. Let a, b be 2 real numbers with $a < b$ and let us define $h := b - a$. Recall that the HC^1-Hermite basis $\{\phi_0, \psi_0, \phi_1, \psi_1\}$ on $I := [a, b]$ forms a basis for the space $VC^1_{\alpha,\beta}(I)$ of all possible HC^1 interpolants on I. The *HC^1-Bernstein basis* $\{b_0, b_1, b_2, b_3\}$ on I are defined as in [15] from the Hermite basis on I by

$$b_0 := \phi_0 - \frac{\lambda}{h}\psi_0, \quad b_1 := \frac{\lambda}{h}\psi_0, \quad b_2 := -\frac{\lambda}{h}\psi_1, \quad b_3 := \phi_1 + \frac{\lambda}{h}\psi_1, \tag{4.3}$$

where $\lambda \ge 2$ is the parameter used to define the control points; see Figure 4.2. These functions are clearly linearly independent and so they form a basis for $VC^1_{\alpha,\beta}(I)$. The

coefficients in terms of this basis are the control coefficients of f. This follows since

$$f := f(a)\phi_0 + p(a)\psi_0 + f(b)\phi_1 + p(b)\psi_1, \quad \Leftrightarrow \quad f = a_0 b_0 + a_1 b_1 + a_2 b_2 + a_3 b_3,$$

where a_0, a_1, a_2, a_3 are the control coefficients of f on I given by (3.1).

We note that $b_j(0) = \delta_{j,0}$ and $b_j(1) = \delta_{j,3}$.

For certain values of the parameters the HC^1-Benstein basis is totally positive.

THEOREM 4.2. *Suppose* $-2 < \beta < 0$, $1 \le v := \frac{\lambda\beta}{\beta - 1} \le 2$, *and* $\lambda \ge 1 - \beta$. *Then the* HC^1*-Bernstein basis is totally positive.*

Proof. It is enough to prove the result for the interval $[0, 1]$. Consider for some integers n, k with $n \ge 0$ and $0 \le k \le 2^n - 1$ the interval $I_k^n := [k/2^n, (k+1)/2^n]$.

On I_k^n the HC^1-Hermite basis $\{\phi_{0,k}^n, \psi_{0,k}^n, \phi_{1,k}^n, \psi_{1,k}^n\}$ can be expressed as

$$\begin{aligned} \phi_{0,k}^n(t) &= \phi_0(2^n t - k), \quad \psi_{0,k}^n(t) = 2^{-n}\psi_0(2^n t - k), \\ \phi_{1,k}^n(t) &= \phi_1(2^n t - k), \quad \psi_{1,k}^n(t) = 2^{-n}\psi_1(2^n t - k), \end{aligned}$$

where $\{\phi_0, \psi_0, \phi_1, \psi_1\}$ is the HC^1-Hermite basis on $[0, 1]$. From (4.3) with $h := 2^{-n}$, it then follows that the HC^1-Bernstein basis $\{b_{4k}^n, b_{4k+1}^n, b_{4k+2}^n, b_{4k+3}^n\}$ on I_k^n can be expressed in terms of the HC^1-Bernstein basis $\{b_0, b_1, b_2, b_3\}$ on $[0, 1]$ as

$$b_{4k+j}^n(t) = \begin{cases} b_j(2^n t - k), & \text{if } t \in I_k^n \text{ and } j = 0, 1, 2, 3 \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

We note that

$$b_{4k+j}^n(k/2^n) = \delta_{j,0}, \quad b_{4k+j}^n((k+1)/2^n) = \delta_{j,3} \quad \text{for} \quad j = 0, 1, 2, 3. \tag{4.5}$$

Let $f \in C^1[0, 1]$ be a HC^1-interpolant to some initial data. We can then write

$$f = \sum_{i=0}^{m} a_i^n b_i^n,$$

where $m := 4 \times 2^n - 1$ and where for $k = 0, \ldots, 2^n - 1$ the numbers a_{4k}^n, a_{4k+1}^n, a_{4k+2}^n, a_{4k+3}^n are the control points of f on I_k^n. In vector form, we have $f = b^n a^n$, where $b^n = (b_0^n, \ldots, b_m^n)$ is a row vector and $a^n = (a_0^n, \ldots, a_m^n)^T$ a column vector. Note that b^n is a vector of linearly independent functions on $[0, 1]$. They span a space containing $VC_{\alpha,\beta}^1[0, 1]$ as a 4-dimensional subspace. On level $n + 1$ we have $f = b^{n+1}a^{n+1}$, where from Proposition 3.1 it follows that $a^{n+1} = \mathbf{A}_n a^n$ for some matrix $\mathbf{A}_n$. The matrix $\mathbf{A}_n$ is a block diagonal with 2^n diagonal blocks $\hat{\mathbf{S}}$ of order 8×4. Indeed, $\hat{\mathbf{S}}$ is obtained from the matrix $\mathbf{S}$ in (3.4) by adding a copy of row 4 as a new row 5. But then $f = b^{n+1}a^{n+1} = b^{n+1}\mathbf{A}_n a^n = b^n a^n$ and by linear independence, it follows that $b^n = b^{n+1}\mathbf{A}_n$. Thus we obtain

$$b^0 = b^n \mathbf{A}_{n-1} \cdots \mathbf{A}_0, \quad n \ge 1. \tag{4.6}$$

For distinct points $y_0, \ldots, y_p$ and functions $f_0, \ldots, f_q$ defined on the y's, we use the standard notation

$$M \begin{bmatrix} y_0, \ldots, y_p \\ f_0, \ldots, f_q \end{bmatrix} := \begin{bmatrix} f_0(y_0) & \cdots & f_q(y_0) \\ \vdots & & \vdots \\ f_0(y_p) & \cdots & f_q(y_p) \end{bmatrix}$$

for a collocation matrix of order $(p+1)\times(q+1)$. In order to show total positivity of $b=b^0$ we choose $0\le x_0<x_1<x_2<x_3\le 1$ and consider the collocation matrix $M\begin{bmatrix}x_0,x_1,x_2,x_3\\ b_0,b_1,b_2,b_3\end{bmatrix}$. From (4.6) we immediatly obtain

$$M\begin{bmatrix}x_0,\dots,x_3\\ b_0,\dots,b_3\end{bmatrix}=M\begin{bmatrix}x_0,\dots,x_3\\ b_0^n,\dots,b_m^n\end{bmatrix}\mathbf{A}_{n-1}\cdots\mathbf{A}_0,\quad n\ge 1. \tag{4.7}$$

Since the matrix $\mathbf{S}$ is totally positive, it follows that $\hat{\mathbf{S}}$ and hence each $\mathbf{A}_k$ is totally positive. We now show that the first matrix on the right of (4.7) is totally positive provided $x_j\in\mathcal{P}_n$ for $j=0,1,2,3$. For this, with $m=2^{n-1}-1$, we consider the bigger matrix

$$\mathbf{B}=M\begin{bmatrix}y_0,\dots,y_{m+1}\\ b_0^n,\dots,b_m^n\end{bmatrix}$$

using all points $y_i=i/2^n$, $i=0,1,\dots,2^n$ in $\mathcal{P}_n$. From (4.5) it follows that $b_{4k-1}(y_k)=1$ for $k=1,\dots,2^n$, $b_{4k}(y_k)=1$ for $k=0,\dots,2^n-1$ and $b_i^n(y_j)=0$ otherwise. Thus the columns of $\mathbf{B}$ have the following form:

$$\mathbf{B}=[e_1,0,0,e_2,e_2,0,0,e_3,e_3,0,0,e_4\dots,e_m,0,0,e_{m+1}],$$

where $e_j=(\delta_{i,j})_{i=0}^m$ is the jth unit vector in $\mathbb{R}^{m+1}$. From this explicit form we see that $\mathbf{B}$ is totally positive since each nonzero minor must be the determinant of the identity matrix. But then all matrices on the right in (4.7) are totally positive and we conclude that $M\begin{bmatrix}x_0,\dots,x_3\\ b_0,\dots,b_3\end{bmatrix}$ is totally positive provided $x_j\in\mathcal{P}_n$ for $j=0,1,2,3$. Since $\cup\mathcal{P}_n$ is dense in $[0,1]$ we conclude that the HC^1-Bernstein basis is totally positive. ∎

COROLLARY 4.3. *For $p\ge 0$ and $m=4\cdot 2^p-1$, the basis $b^p=(b_0^p,\dots,b_m^p)$ for the space span(b^p) is totally positive on $[0,1]$.*

Proof. Instead of (4.6) we use for $n>p$ the equation

$$b^p=b^n\mathbf{A}_{n-1}\cdots\mathbf{A}_p.$$

The argument now proceeds as in the proof of Theorem 4.2 replacing $x_0,\dots,x_3$ by suitable $x_0,\dots,x_m$. ∎

It is well known that total positivity of the HC^1-Bernstein basis on $[0,1]$ implies that the HC^1-interpolant f inherits properties of the control polygon P^0 defined by $\{a_0,a_1,a_2,a_3\}$; see, for example, [8]. In particular if P_0 is positive (monotone, convex) then f is positive (monotone, convex). We can use this to generalize Theorem 4 in [15].

COROLLARY 4.4. *Let b_0,b_1,b_2,b_3 be the HC^1-Bernstein basis on $[0,1]$ given by (4.3) with $\lambda=v(\beta-1)/\beta>2$. Suppose also that $\alpha=\frac{\beta}{4(1-\beta)}$, $-1\le\beta<0$, and $1\le v\le 2$. Then*

1. *b_0 is nonnegative, decreasing, and convex on $[0,1]$. If $v=2$, then $b_0(t)=0$ for $t\in[1/2,1]$.*
2. *b_1 is nonnegative and concave on $[0,1/2]$ and nonnegative, decreasing, and convex on $[1/2,1]$.*
3. *b_2 is nonnegative, increasing, and convex on $[0,1/2]$ and nonnegative and concave on $[1/2,1]$.*
4. *b_3 is nonnegative, increasing, and convex on $[0,1]$. If $v=2$, then $b_3(t)=0$ for $t\in[0,1/2]$.*

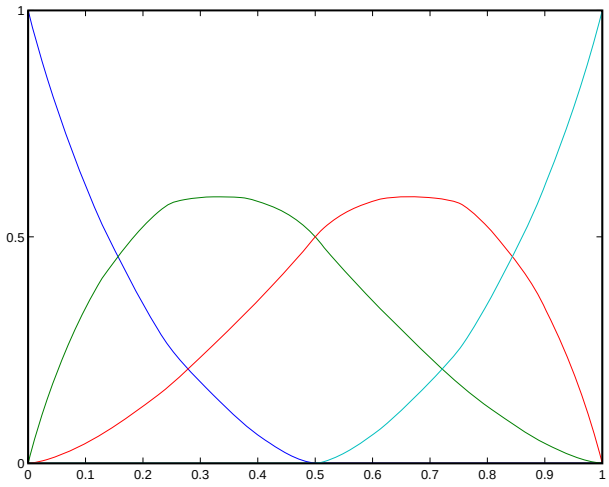


FIG. 4.2. *Bernstein basis,* $\beta = -3/5$, $\alpha = -3/32$, $\lambda = 16/3$.

5. $\sum_{j=0}^{3} b_j(t) = 1$ *for* $t \in [0,1]$.

Proof. From (4.3) it follows that the control points of the function b_j is the jth unit vector e_{j+1} for $j = 0,1,2,3$. Thus nonnegativity of b_j follows from the nonnegativity of e_{j+1} for $j = 0,1,2,3$. Moreover, the monotonicity and convexity properties of b_0 and b_3 follow. For the remaining properties of b_1 and b_2, we carry out one subdivision, then the proof is similar.

The refined points are given as the columns of the matrix $\mathbf{S}$ given by (4.1). When $v = 2$ the first column is given by $[1, 1/2, 0, 0, 0, 0, 0]$. Since the last four entries are zero it follows that $b_0(t) = 0$ for $t \in [1/2, 1]$. Similarly $b_3(t) = 0$ for $t \in [0, 1/2]$.

The interpolation of the constant function $f = 1$ with $p = f' = 0$ gives $a_0 = a_1 = a_2 = a_3 = 1$ in (3.1) so that 5. holds. ∎

5. Algorithms for local shape constraints. We base shape preserving algorithms on the extended quadratic spline case given by $\alpha = \frac{\beta}{4(1-\beta)}$. The control point subdivision matrix for this case is given by (4.1), where we have both β and λ as free parameters. The matrix simplifies when $v = \frac{\beta\lambda}{\beta - 1} = 2$ and we will use this one parameter family of schemes in our algorithms. Using the parameter λ to control the shape we thus have

$$\alpha = \frac{\beta}{4(1-\beta)} = -\frac{1}{2\lambda}, \quad \beta = \frac{2}{2-\lambda}. \tag{5.1}$$

We restrict our attention to $\lambda \geq 4$. We then have $\beta \in [-1, 0)$ and both algorithms HC^1 and SC^1 are convergent. In the limit when $n \to \infty$ we obtain a function $f \in C^1(I)$. This function is the quadratic spline interpolant with a knot at the midpoint of I when $\lambda = 4$, while $p = f'$ is Hölder continuous on I with exponent

$$\log_2\left(1 + \frac{1}{\lambda - 3}\right) \approx \frac{1.44}{\lambda - 3}, \quad \lambda \to \infty.$$

Thus the derivative becomes less regular when λ increases, but it is always C^1.

Given $s \geq 1$, points $a_j^0 = a_j \in \mathbb{R}^s$ for $j = 0,1,2,3$, and $\lambda \geq 4$, the following algorithm computes sequences $\{a^n\}$ of control coefficients $a^n = (a_0^n, a_1^n, \ldots, a_{3\times 2^n}^n)$ in $\mathbb{R}^s$.

ALGORITHM 5.1 (CC^1).

1. $\beta = 2/(2-\lambda)$.
2. *For* $n = 0, 1, 2, 3, \ldots,$
   For $i = 0, 1, \ldots 2^n - 1,$
   (a) $a_{6i}^{n+1} = a_{3i}^n,$
   (b) $a_{6i+1}^{n+1} = \frac{1}{2}(a_{3i}^n + a_{3i+1}^n),$
   (c) $a_{6i+2}^{n+1} = (\frac{1}{2} - \frac{\beta}{4})a_{3i+1}^n + (\frac{1}{2} + \frac{\beta}{4})a_{3i+2}^n,$
   (d) $a_{6i+3}^{n+1} = \frac{1}{2}(a_{3i+1}^n + a_{3i+2}^n),$
   (e) $a_{6i+4}^{n+1} = (\frac{1}{2} + \frac{\beta}{4})a_{3i+1}^n + (\frac{1}{2} - \frac{\beta}{4})a_{3i+2}^n,$
   (f) $a_{6i+5}^{n+1} = \frac{1}{2}(a_{3i+2}^n + a_{3i+3}^n),$
   $a_{3\cdot 2^{n+1}} = a_{3\cdot 2^n}.$

The control points corresponding to the computed control coefficients converges to a C^1-curve. More specifically, pick any finite closed interval $[a, b]$ and define $h_n := (b-a)/2^n$ and $t_k^n := a + kh_n$ for $k = 0, \ldots, 2^n$, $n \geq 0$. By Theorem 3.4 the computed control points converge uniformly to a C^1-curve $f : [a, b] \to \mathbb{R}^s$. Moreover,

$$
\begin{aligned}
a_{3i}^n &= f(t_i^n), \quad i = 0, \ldots, 2^n, \\
a_{3i+1}^n - a_{3i}^n &= \frac{h_n}{\lambda} f'(t_i^n), \quad i = 0, \ldots, 2^n - 1, \\
a_{3i}^n - a_{3i-1}^n &= \frac{h_n}{\lambda} f'(t_i^n), \quad i = 1, \ldots, 2^n.
\end{aligned}
$$

We now discuss shape preservation in the scalar case $s = 1$ in more detail. We start by noting that if the initial control polygon is nonnegative (respectively, increasing, convex) on an interval $I = [a, b]$, then the HC^1-interpolant computed in Algorithm 5.1 will be nonnegative (respectively, increasing, convex) on the same interval I. This follows from the total positivity of the Bernstein basis. In addition to total positivity the main tool will be Corollary 2.2 which says that the p-values of the interpolant are located on the piecewise linear curve connecting the three points $(a, p(a))$, $(\frac{a+b}{2}, p(\frac{a+b}{2}))$, $(b, p(b))$.

5.1. Nonnegative interpolants. We already remarked that if the initial control coefficients are nonnegative, then the HC^1-interpolant will be nonnegative. Notice that the converse is false. For example, the HC^1-interpolant to the function f given on $[0, 1]$ by $f(x) := 16(x - 1/4)^2$ and using $\lambda = 4$ is f itself. Note that f is nonnegative, but the initial control coefficient $a_1 = -1$ is negative.

To give an algorithm for constructing a nonnegative interpolant we assume that

$$
f(a) \geq 0, \ f(b) \geq 0, \ p(a) \geq 0 \text{ if } f(a) = 0, \text{ and } p(b) \leq 0 \text{ if } f(b) = 0. \tag{5.2}
$$

Under these weak assumptions nonnegative initial control coefficients $a_0, \ldots, a_3$ can always be obtained by choosing λ sufficiently large. Indeed, since $a_0 = f(a) \geq 0$ and $a_3 = f(b) \geq 0$ we only need to make sure that $a_1 = f(a) + hp(a)/\lambda \geq 0$ and $a_2 = f(b) - hp(b)/\lambda \geq 0$. If $f(a) = 0$, then $p(a) \geq 0$ and $a_1 \geq 0$ whenever $\lambda > 0$. Similarly $a_2 \geq 0$ if $f(b) = 0$. But then we can choose $\lambda = 4$ except possibly in the two cases $f(a) > 0$, $p(a) < 0$ and $f(b) > 0$, $p(b) > 0$. If (5.2) holds, then the following algorithm will compute a nonnegative HC^1-interpolant on $[a, b]$.

ALGORITHM 5.2 (nonnegative interpolant).

1. *Compute* λ*:*
   (a) $\lambda = 4$,
   (b) *if* $(f(a) > 0)$ *and* $(p(a) < 0)$, *then* $\lambda = \max(\lambda, -hp(a)/f(a))$,

(c) *if* $(f(b) > 0)$ *and* $(p(b) > 0)$, *then* $\lambda = \max(\lambda, hp(b)/f(b))$.

2. *Compute initial control coefficients using* (3.1).
3. *Apply Algorithm* 5.1 *or Algorithm* HC^1 *with* $\alpha = -\frac{1}{2\lambda}$, $\beta = \frac{2}{2-\lambda}$.

5.2. Monotone interpolants. The monotonicity of the HC^1-interpolant is completely determined by the monotonicity of the initial control polygon. If f is decreasing, then $-f$ is increasing and we restrict our discussion to increasing interpolants.

PROPOSITION 5.3. *Suppose that the parameters are chosen according to* (5.1). *Then the* HC^1*-interpolant* f *is nondecreasing on an interval* $I = [a, b]$ *if and only if the control polygon on* I *is nondecreasing.*

Proof. By Theorem 4.2 the Bernstein basis is totally positive and it follows that the HC^1-interpolant is nondecreasing if the control polygon is nondecreasing; see [8]. Conversely, suppose the HC^1-interpolant f is nondecreasing. Since $\beta = 2/(2-\lambda)$, we obtain from (2.1)

$$p\left(\frac{a+b}{2}\right) = \frac{1}{\lambda - 2}\left(\lambda \frac{f(b) - f(a)}{h} - (p(a) + p(b))\right). \tag{5.3}$$

From (3.1), we then find

$$a_1 - a_0 = \frac{h}{\lambda}p(a), \quad a_2 - a_1 = \frac{\lambda - 2}{\lambda}hp\left(\frac{a+b}{2}\right), \quad a_3 - a_2 = \frac{h}{\lambda}p(b). \tag{5.4}$$

Now $p \geq 0$ at all points if f is nondecreasing. It follows that the control coefficients, and hence the control polygon is nondecreasing. □

Consider next the case of a strictly increasing interpolant.

PROPOSITION 5.4. *Suppose that the parameters are chosen according to* (5.1) *and that the* HC^1*-interpolant* f *is nondecreasing on an interval* $I = [a, b]$. *Then* f *is strictly increasing on* $[a, b]$ *if and only if the two middle control coefficients on* I *satisfy* $a_2 > a_1$.

Proof. Since f is nondecreasing, we have $p(a) \geq 0$, $p(\frac{a+b}{2}) \geq 0$ and $p(b) \geq 0$. By Corollary 2.2, it follows that f is strictly increasing on $[a, b]$ if and only if $p(\frac{a+b}{2}) > 0$. By (5.4), this happens if and only if $a_2 > a_1$. □

To give an algorithm to construct a nondecreasing interpolant we assume that

$$f(a) \leq f(b),\ p(a) \geq 0,\ p(b) \geq 0 \text{ and } p(a) = p(b) = 0 \text{ if } f(a) = f(b). \tag{5.5}$$

In the latter case the HC^1-interpolant is constant and we can set $\lambda = 4$.

Suppose $f(b) > f(a)$. With $h := b - a$ we then have

$$a_0 = f(a) \leq a_1 = f(a) + \frac{h}{\lambda}p(a) \leq a_2 = f(b) - \frac{h}{\lambda}p(b) \leq a_3 = f(b)$$

provided

$$a_2 - a_1 = f(b) - f(a) - \frac{h}{\lambda}\big(p(b) + p(a)\big) \geq 0$$

or

$$\lambda \geq \frac{(p(a) + p(b))h}{f(b) - f(a)}. \tag{5.6}$$

If (5.5) holds, then the following algorithm will compute a nondecreasing HC^1-interpolant on $[a,b]$. It will be strictly increasing if $f(b) > f(a)$ and (5.6) holds with strict inequality.

ALGORITHM 5.5 (nondecreasing or strictly increasing interpolant).

1. *Compute* λ*:*
   (a) $\lambda = 4$,
   (b) *If* $f(a) < f(b)$*, then*
      (i.) $\lambda_1 \geq \frac{(p(a)+p(b))h}{f(b)-f(a)}$,
      (ii.) $\lambda = \max(4, \lambda_1)$.
2. *Compute initial control coefficients using* (3.1).
3. *Apply Algorithm* 5.1 *or Algorithm* HC^1 *with* $\alpha = -\frac{1}{2\lambda}$, $\beta = \frac{2}{2-\lambda}$.

Note that if the initial control points are located on a straight line then the HC^1-interpolant is the line segment connecting the first and last control point. For if the initial control points are located on a straight line, then

$$\frac{\lambda}{h}(a_1 - a_0) = \frac{\lambda}{(\lambda-2)h}(a_2 - a_1) = \frac{\lambda}{h}(a_3 - a_2)$$

and by (5.4) the three slopes $p(a), p(\frac{a+b}{2}), p(b)$ are all equal. By Corollary 2.2, all slopes are equal and the function f is a straight line.

In Figure 5.1 we interpolate three sets of data on $[0,1]$. In all cases $f(0) = -1$ and $f(1) = 1$. In the first case, with $p(0) = 3$ and $p(1) = 4$ we find $\frac{p(0)+p(1)}{f(1)-f(0)} = 7/2 < 4$. Suppose in Algorithm 5.5 we choose $7/2 \leq \lambda_1 \leq 4$ in statement (b)(i). and apply Algorithm 5.1 with $\lambda = 4$. Then the HC^1-interpolant is the quadratic spline and it is strictly increasing since $\lambda > 7/2$. In the two other cases we use $p(0) = 8$ and $p(1) = 4$ giving $\frac{p(0)+p(1)}{f(1)-f(0)} = 6$. With $\lambda = 6$ we have $p(1/2) = 0$ and the interpolant is increasing, but not strictly increasing. We obtain a strictly increasing interpolant by using $\lambda = 10$. Note that choosing a bigger λ decreases the regularity of the interpolant. In both cases the first derivative is Hölder continuous, but the exponent is $\log_2(4/3) \approx 0.415$ when $\lambda = 6$ and $\log_2(4/3) \approx 0.193$ when $\lambda = 10$.

5.3. Convex interpolants. The convexity of the HC^1-interpolant is also completely determined by the convexity of the initial control polygon.

PROPOSITION 5.6. *Suppose that the parameters are chosen according to* (5.1). *Then* f *is convex (concave) on an interval* $I = [a,b]$ *if and only if the control polygon on* I *is convex (concave).*

Proof. Again by total positivity of the Bernstein basis the HC^1-interpolant is convex (concave) if the control polygon is convex (concave); see [8]. Conversely, suppose the HC^1-interpolant f is convex (concave). Now the control polygon is convex if and only if the conditions

$$\frac{a_1 - a_0}{h/\lambda} \leq \frac{a_2 - a_1}{h - 2h/\lambda} \leq \frac{a_3 - a_2}{h/\lambda}$$

hold. But from (5.4) we find

$$\frac{a_1 - a_0}{h/\lambda} = p(a), \quad \frac{a_2 - a_1}{h - 2h/\lambda} = p(\frac{a+b}{2}), \quad \frac{a_3 - a_2}{h/\lambda} = p(b).$$

Since f is convex (concave) the function p is nondecreasing (nonincreasing) and hence the control polygon is convex (concave). ∎

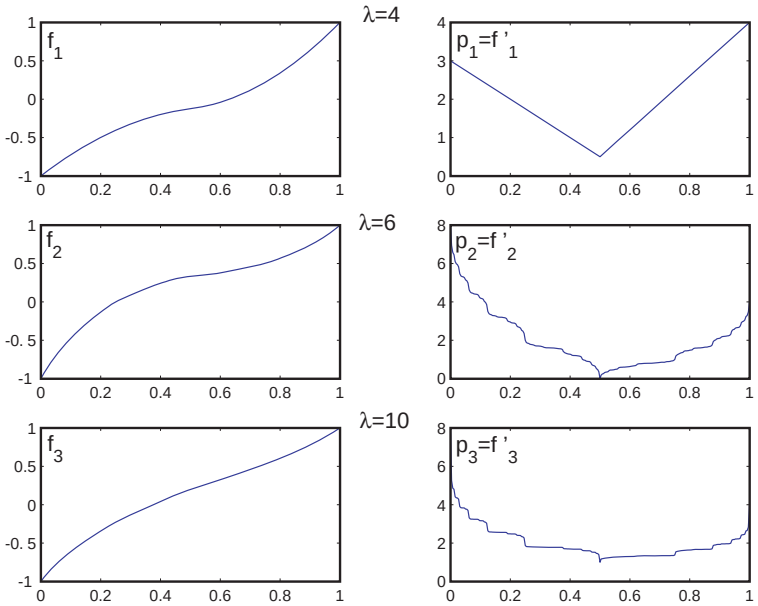


FIG. 5.1. *Monotone interpolants.*

To give an algorithm for constructing a convex (concave) HC^1-interpolant on an interval $I = [a, b]$ we first assume that

$$p(a) < \frac{f(b) - f(a)}{h} < p(b) \quad \big(p(a) > \frac{f(b) - f(a)}{h} > p(b)\big), \tag{5.7}$$

where $h := b - a$. We define

$$\lambda_1 := \frac{p(b) - p(a)}{p(b) - \frac{f(b)-f(a)}{h}}, \quad \lambda_2 := \frac{p(b) - p(a)}{\frac{f(b)-f(a)}{h} - p(a)} \tag{5.8}$$

and note that the tangents

$$t_c(x) := f(a) + (x - a)p(a), \quad t_d(x) := f(b) + (x - b)p(b)$$

of f at a and b intersect at the point $(\bar{x}, \bar{y})$ given by

$$\frac{\bar{x} - a}{h} = \frac{1}{\lambda_1}, \quad \text{and} \quad \frac{b - \bar{x}}{h} = \frac{1}{\lambda_2}.$$

Moreover, the hypothesis (5.7) is equivalent to $a < \bar{x} < b$.

Under the assumption

$$p(a) \le \frac{f(b) - f(a)}{h} \le p(b) \quad \big(p(a) \ge \frac{f(b) - f(a)}{h} \ge p(b)\big), \tag{5.9}$$

the following algorithm will compute a convex (concave) interpolant.

ALGORITHM 5.7 (convex or concave interpolant).

1. (a) *If* $p(a) = \frac{f(b)-f(a)}{h} \ne p(b)$, *choose* $\lambda \ge \max(4, \lambda_1)$,
   (b) *If* $p(a) \ne \frac{f(b)-f(a)}{h} = p(b)$, *choose* $\lambda \ge \max(4, \lambda_2)$,

(c) *If* $p(a) \neq \frac{f(b)-f(a)}{h} \neq p(b)$, *choose* $\lambda \geq \max(4, \lambda_1, \lambda_2)$.

2. *Compute initial control points using* (3.1).
3. *Apply Algorithm* 5.1 *or Algorithm* HC^1 *with* $\alpha = -\frac{1}{2\lambda}$, $\beta = \frac{2}{2-\lambda}$.

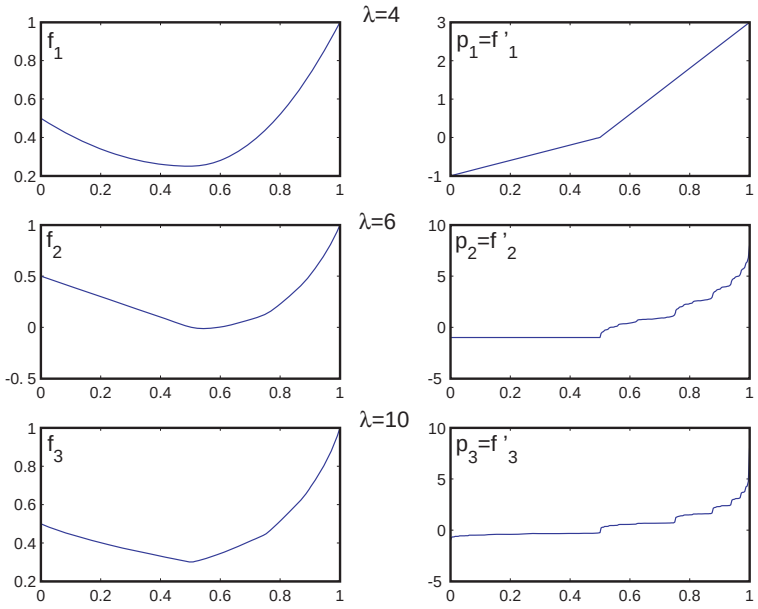


FIG. 5.2. *Convex interpolants.*

In Figure 5.2, we have interpolated three sets of data on $[0, 1]$. In all cases $f(0) = 0.5$ and $f(1) = 1$.

In the first case, $p(0) = -1$ and $p(1) = 3$ so that $\lambda_1 = 8/5$ and $\lambda_2 = 8/3$. Then $\max(4, \lambda_1, \lambda_2) = 4$ and we have chosen $\lambda = 4$. In this case, the interpolant is the quadratic spline.

In the two other cases $p(0) = -1$ and $p(1) = 8$ so that $\lambda_1 = 18/5$ and $\lambda_2 = 6$. Then $\max(4, \lambda_1, \lambda_2) = 6$. With $\lambda = 6$ we have $p = -1$ on $[0, 1/2]$, while we obtain a strictly convex interpolant by using $\lambda = 10$. Recall that choosing a bigger λ decreases the regularity of the interpolant.

6. Example. Given data (t_i, y_i, y_i') for $i = 1, \ldots, n$, where $t_1 < \cdots < t_n$ and the y's are real numbers, we look for a function $f \in C^1([t_1, t_n])$ that satisfies

$$f(t_i) = y_i, \;\; f'(t_i) = y_i' \;\; \text{for } i = 1, \ldots, n. \tag{6.1}$$

In addition we would like f to be positive, monotone, linear, or convex on some or all of the subintervals $I_i = [t_i, t_{i+1}], i = 1, \ldots, n-1$. We assume that

(P) (5.2) holds for the subintervals where we want nonnegativity or positivity,
(M) (5.5) holds for the subintervals where we want a nondecreasing or a strictly increasing interpolant,
(L) $y_i' = y_{i+1}' = \frac{y_{i+1}-y_i}{t_{i+1}-t_i}$ for the subintervals where the interpolant should be linear,
(C) (5.9) holds for the subintervals where the interpolant should be convex or concave.

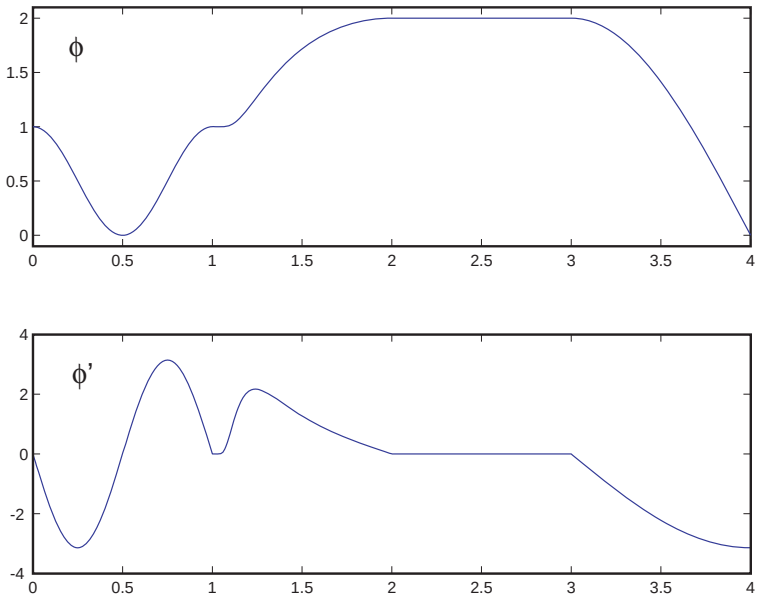


FIG. 5.3. *The function ϕ and its derivative.*

We also require that the given data is consistent with these shape requirements. We can compute f locally by applying the HC^1-algorithm with parameters given by (5.1) on each subinterval $I_i = [t_i, t_{i+1}]$, $i = 1, \ldots, n-1$ using initial data $f(t_i) = y_i$, $f(t_{i+1}) = y_{i+1}$, $p(t_i) = y'_i$, and $p(t_{i+1}) = y'_{i+1}$. We obtain C^1-convergence and the desired shape locally by choosing the parameter λ_i for the interval I_i sufficiently large.

Consider now (6.1) for the example illustrated in Figure 1.1. The data are sampled from the function $\phi \in C^1([0,4])$ given by

$$\phi(t) = \begin{cases} \frac{1}{2}\sin(2\pi t + \pi/2) + \frac{1}{2}, & 0 \le t \le 1, \\ 1 + \exp(-\frac{1}{1-(t-2)^2} + 1), & 1 < t \le 2, \\ 2, & 2 < t \le 3, \\ 2\cos(\pi \frac{t-3}{2}), & 3 \le t \le 4. \end{cases} \tag{6.2}$$

The function and its first derivative are displayed in Figure 5.3 and it can be shown that ϕ is positive on $[0,1]$, strictly increasing on $[1,2]$, constant on $[2,3]$, and concave on $[3,4]$; given n and let $(t_1, \ldots, t_n)$ be a partition of $[0,4]$. The points $(t_2, \ldots, t_{n-1})$ are chosen randomly except that $1, 2, 3$, are among them. In the example, we used $t_1 = 0$, $t_{n_1} = t_5 = 1$, $t_{n_2} = t_9 = 2$, $t_{n_3} = t_{13} = 3$, and $t_n = t_{17} = 4$. We want an interpolant f which is positive on $[t_1, t_{n_1}] = [0,1]$, strictly increasing on $[t_{n_1}, t_{n_2}] = [1,2]$, constant on $[t_{n_2}, t_{n_3} = [2,3]$, and concave on $[t_{n_3}, t_n] = [3,4]$.

In the first test we use $y_i = \phi(t_i)$ and exact derivatives $y'_i = \phi'(t_i), i = 1, \ldots, n$. In this case all λ's become equal to 4 and the quadratic spline interpolant f_1 does the job. Plots of this function and its first derivative are shown in Figure 6.1. The first derivative appears continuous and piecewise linear.

For the second test shown in Figure 6.2, we kept the previous data t_i and $y_i = \phi(t_i)$ for $i = 1, \ldots, n = 17$, but we used inexact derivatives given by crosses in the lower part of the figure. However, the derivatives were chosen so that the relevant requirement (P), (M), (L), and (C) above are satisfied on each subinterval $[t_i, t_{i+1}]$. We obtain

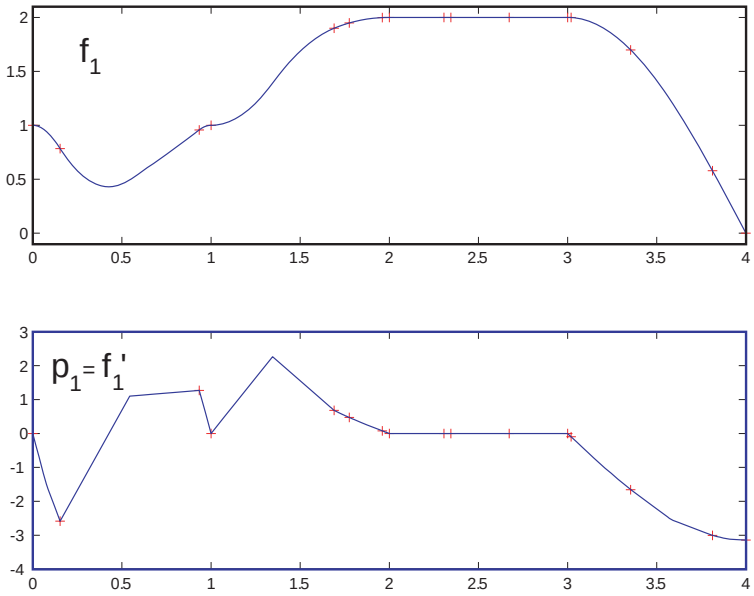


FIG. 6.1. *Interpolation with exact derivatives.*

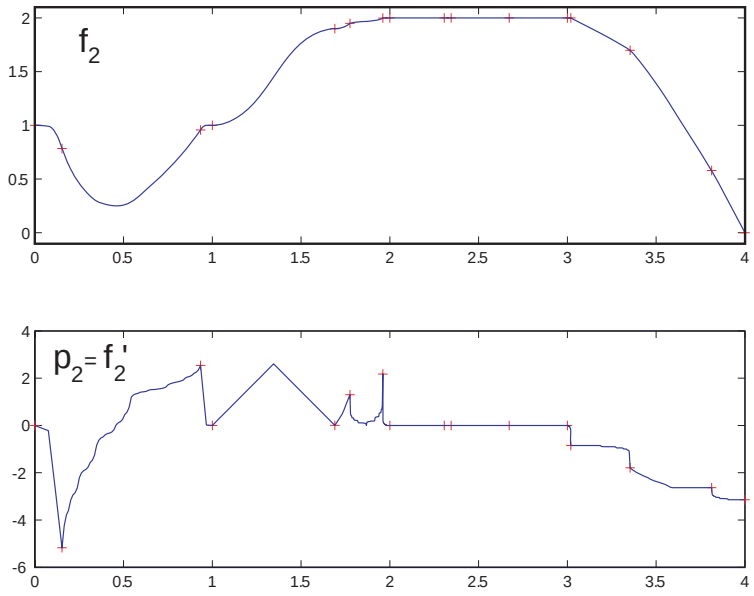


FIG. 6.2. *Interpolation with modified derivatives.*

a C^1-interpolant f_2 satisfying the required shape constraints. The computed values of λ_i are successively $(4, 5.1425, 4, 4, 4, 12.8631, 55.8239, 4, 4, 4, 4, 17.6767, 20.0216, 4.4087, 11.3544)$. These are the smallest values on each interval. Any larger value of λ_i is possible without losing the shape, but then the curve is less regular (smaller Hölder exponent) on the corresponding interval. This example shows that we can obtain a desired shape even with more or less random derivative values.

7. Conclusion. We have shown that the Hermite subdivision scheme introduced by Merrien in 1992 [11] has many desirable properties. It gives a C^1 limit curve for a wide range of parameters. A one parameter subfamily called the extended quadratic spline-scheme is particularly interesting. This family can be formulated as a scheme, the SC^1 algorithm, with a totally positive subdivision matrix. When applied in a piecewise fashion its local nature makes it easy to control the final shape of the subdivision curve. In many cases a desired shape can be obtained even without accurate derivative estimates.

The SC^1 algorithm can also be used in the parametric case, but a discussion of this will be deferred to a future paper. We also defer the construction of interpolating C^1 surfaces with shape preserving properties.

REFERENCES

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (2001), pp. 165–219.

[2] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc., 93 (453) 1991, pp. 1–186.

[3] I. Daubechies, I. Guskov, and W. Sweldens, *Commutation for irregular subdivision*, Constr. Approx., 17 (2001), pp. 479–514.

[4] T. D. DeRose, *Subdivision surfaces in feature films*, in Mathematical Methods for Curves and Surfaces (Oslo, 2000), T. Lyche and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 73–79.

[5] N. Dyn and D. Levin, *Analysis of Hermite-interpolatory subdivision schemes*, in Spline Functions and the Theory of Wavelets (Montreal, 1996), S. Dubuc and G. Deslauriers, eds., Amer. Math. Soc., Providence, RI, 1999, pp. 105–113.

[6] N. Dyn and D. Levin, *Subdivision schemes in geometric modelling*, Acta Numer., 11 (2002), pp. 73–144.

[7] T. A. Foley, T. N. T. Goodman, and K. Unsworth, *An algorithm for shape preserving parametric interpolating curves with G^2 continuity*, in Mathematical Methods in Computer Aided Geometric Design (Oslo, 1998), T. Lyche and L. L. Schumaker, eds., Academic Press, Boston, 1989, pp. 249–259.

[8] T. N. T. Goodman, *Shape preserving representations*, in Mathematical Methods in Computer Aided Geometric Design (Oslo, 1998), T. Lyche and L. L. Schumaker, eds., Academic Press, Boston, 1989, pp. 333–351.

[9] T. N. T. Goodman, *Shape preserving interpolation by curves*, in Algorithms for Approximation IV, J. Levesley, I. J. Anderson, and J. C. Mason, eds., University of Huddersfield, 2002, Huddersfield, UK, pp. 24–35.

[10] B. Han, T. P.-Y. Yu, and B. Piper, *Multivariate refinable Hermite interpolants*, Math. Comp., 73 (2004), pp. 1913–1935.

[11] J.-L. Merrien, *A family of Hermite interpolants by bisection algorithms*, Numer. Algorithms, 2 (1992), pp. 187–200.

[12] J.-L. Merrien, *Interpolants d'Hermite C^2 obtenus par subdivision*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 55–65.

[13] J.-L. Merrien and P. Sablonnière, *Monotone and convex C^1 Hermite interpolants generated by a subdivision scheme*, Constr. Approx., 19 (2002), pp. 279–298.

[14] C. A. Micchelli, *Mathematical Aspects of Geometric Modeling*, SIAM, Philadelphia, 1995.

[15] P. Sablonnière, *Bernstein-type bases and corner cutting algorithms for C^1 Merrien's curves*, Adv. Comput. Math., 20 (2004), pp. 229–246.

[16] J. Warren and H. Weimer, *Subdivision Methods for Geometric Design: A Constructive Approach*, Morgan Kaufmann, San Francisco, 2002.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1122–1149

ASYMPTOTIC EXPANSIONS AND RICHARDSON EXTRAPOLATION OF APPROXIMATE SOLUTIONS FOR SECOND ORDER ELLIPTIC PROBLEMS ON RECTANGULAR DOMAINS BY MIXED FINITE ELEMENT METHODS*

GRAEME FAIRWEATHER†, QUN LIN‡, YANPING LIN§, JUNPING WANG¶, AND SHUHUA ZHANG‖

Abstract. In this paper asymptotic error expansions for mixed finite element approximations of general second order elliptic problems are derived under rectangular meshes, and the Richardson extrapolation is applied to improve the accuracy of the approximations by two different schemes with the help of an interpolation postprocessing technique. The results of this paper provide new asymptotic expansions and new approximate solutions which are one-order and a half-order higher in accuracy than those obtained in [J. Wang, *Math Comp.*, 56 (1991), pp. 477–503] and [H. Chen, R. E. Ewing, and R. Lazarov, *Asymptotic Error Expansion for the Lowest Order Raviart–Thomas Rectangular Mixed Finite Elements*, Technical report ISC-97-01, 1997], respectively. As a by-product, we illustrate that all the approximations of higher accuracy can be used to form a class of a posteriori error estimators for the mixed finite element method.

Key words. second order elliptic problems, mixed finite element methods, asymptotic expansions, interpolation postprocessing, a posteriori error estimators

AMS subject classifications. 76S05, 45K05, 65M12, 65M60, 65R20

DOI. 10.1137/040614293

1. Introduction.

We are concerned with approximate solutions for the following system of linear equations:

$$\mathbf{u} = -A\nabla p, \quad \nabla \cdot \mathbf{u} + cp = f \quad \text{in } \Omega \tag{1.1}$$

subject to the Neumann boundary condition

$$\nabla p \cdot \mathbf{n} = 0 \quad \text{on } \Gamma. \tag{1.2}$$

Here ∇ is the gradient operator; $\Omega \subset \mathbf{R}^2$ is a rectangular domain; $\mathbf{n}$ indicates the outward unit normal vector along Γ; and $A = (a_{ij})_{2\times 2}$ is a positive definite matrix uniformly in Ω. Mixed finite element methods [4] shall be employed to discretize the system (1.1).

*Received by the editors September 1, 2004; accepted for publication (in revised form) December 20, 2005; published electronically June 21, 2006. This work is supported in part by NSERC, NSF, National Natural Science Foundation of China (10471103), Liu Hui Center for Applied Mathematics of Nankai University and Tianjin University, Tianjin Education Committee, and Tianjin University of Finance and Economics.

http://www.siam.org/journals/sinum/44-3/61429.html

†Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (gfairwea@glenclova.mines.edu).

‡Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China (qlin@lsec.cc.ac.cn).

§Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada (ylin@math.ualberta.ca).

¶Division of Mathematical Science, National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230 (jwang@nsf.gov).

‖Department of Mathematics, University of Finance and Economics, Tianjin 300222, China, and Liu Hui Center for Applied Mathematics of Nankai University and Tianjin University, Tianjin 30072, China (szhang@tjufe.edu.cn).

There is extensive literature on the numerical methods for problem (1.1)–(1.2). For example, see [5, 9, 18, 26] and the references therein for mixed finite element methods, finite element methods, and finite difference methods.

Our objective in this paper is to present an analysis for the Richardson extrapolation of two different forms for the mixed finite element approximations. It is well known that the extrapolation method, which was established by Richardson in 1926, is an efficient procedure for increasing the accuracy of approximation of many problems in numerical analysis. The effectiveness of this technique relies heavily on the existence of an asymptotic expansion for the error. The application of this approach in finite difference methods can be found in the book of Marchuk and Shaidurov [26]. Also, this technique has been well demonstrated in its applications to the finite element and the mixed finite element methods for elliptic partial differential equations [2, 9, 10, 18, 28, 29], parabolic partial differential equations [15], Sobolev- and viscoelasticity-type equations [19], partial integro-differential equations [18, 20, 21], Fredholm and Volterra integral equations of the second kind [23, 6], and Volterra integro-differential equations [24, 32], and to boundary element methods and collocation methods in [31] and [17], respectively. Recently, multiparameter parallel algorithms have been considered to accelerate its computational speed [33].

In [28, 29] Richardson extrapolation methods of mixed finite elements have been investigated for a special case of the problem (1.1)–(1.2): Seek $u \in H^1(\Omega)$ such that

$$\begin{aligned} -\mathrm{div}(\beta(\mathbf{x})\nabla u(\mathbf{x})) &= f(\mathbf{x}) \qquad \text{in } \Omega, \\ \beta(\mathbf{x})\nabla u \cdot \mathbf{n} &= 0 \qquad\quad \text{on } \Gamma, \end{aligned} \tag{1.3}$$

where $\beta(\mathbf{x})$ is a positive, continuous function on the closure of the rectangular domain Ω, and the approximate solution of accuracy $O(h^3|\log h|)$ in the L^∞-norm has been obtained for the lowest triangular and rectangular elements, respectively.

In particular, when $\beta(\mathbf{x}) = 1$ in (1.3), Chen, Ewing, and Lazarov [10] obtained an asymptotic error expansion of the lowest order Raviart–Thomas rectangular mixed finite elements for the velocity, such that an approximate solution of accuracy $O(h^{3+1/2})$ was gained by the Richardson extrapolation method.

To our best knowledge, there is no analysis for the Richardson extrapolation in mixed finite element methods for the problem (1.1)–(1.2) because $A(\mathbf{x})$ is a full matrix, such that the high accuracy analysis of mixed finite elements and finite elements, such as Richardson extrapolation and superconvergence, is much more difficult than the case (1.3). In this paper, we employ the sharp integral estimates, which were first proposed by Lin in 1990 (see, for example, [18, 19]), to establish an asymptotic expansion for the error between the mixed finite element solution and the corresponding interpolation function of the exact solution to (1.1). These analysis techniques are quite different from those used in [10] and [28]. Furthermore, by virtue of an interpolation postprocessing method we will obtain an asymptotic expansion of the error in a mixed finite element solution, so that the Richardson extrapolation of two different types can be applied to yield mixed finite element approximations of accuracy $O(h^4)$ in the L^2-norm which are a half-order higher than those obtained in [10]. In another paper, we will demonstrate that our new method in this paper can also provide the approximations of accuracy $O(h^4|\log h|)$ in the L^∞-norm which are one-order higher than those gained in [28]. In addition, based on the high accuracy approximations, a class of a posteriori error estimators are constructed for this mixed finite element

method. Also, other types of superconvergence and a posteriori estimates have been obtained; see, for example, [1, 3, 7, 8, 16, 25, 27, 30], where the authors studied these problems on mixed finite element methods for optimal control, quadratic control, and elliptic problems by using conforming and nonconforming elements via structured and unstructured meshes.

This paper is organized in the following way. In section 2, we give the approximate subspace and the variational formula for the problem (1.1)–(1.2) as well as the Raviart–Thomas projection. The asymptotic expansion for the Raviart–Thomas projection is derived in section 3. Section 4 is devoted to investigating the asymptotic expansion of the error between the mixed finite element solution and the Raviart–Thomas projection of the exact solution to (1.1)–(1.2), on the basis of which the asymptotic expansion of the mixed finite element approximation is demonstrated by an interpolation postprocessing method. Hence, the Richardson extrapolation of two schemes can be used to improve the mixed finite element solution. Moreover, in this section suggestions are presented on how to form a posteriori error estimators by those approximations with high convergence rates.

2. The mixed finite element method. In this section we first formulate the mixed finite element method for the second order elliptic partial differential equation (1.1)–(1.2).

Let

$$W := L^2(\Omega) \quad \text{and} \quad \mathbf{V} := H(\mathrm{div}, \Omega) = \left\{ \mathbf{v} \in (L^2(\Omega))^2 : \ \nabla \cdot \mathbf{v} \in L^2(\Omega) \right\}$$

be the standard L^2-space on Ω with norm $\|\cdot\|_0$ and the Hilbert space equipped with the norm

$$\|\mathbf{v}\|_{\mathbf{V}} := \left(\|\mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2 \right)^{\frac{1}{2}},$$

respectively. In addition, set

$$\mathbf{V}_0 := \{ \mathbf{v} \in \mathbf{V} : \ \ \mathbf{v} \cdot \mathbf{n} = 0, \ \ \mathbf{x} \in \Gamma \}.$$

Using Green's formula and the boundary condition $\mathbf{v} \cdot \mathbf{n} = 0$, one finds

$$(\nabla p, \mathbf{v}) = -(\nabla \cdot \mathbf{v}, p) \tag{2.1}$$

for all $\mathbf{v} \in \mathbf{V}_0$. Moreover, from (1.1) and (1.2) we know that the vector-valued function $\mathbf{u}$ satisfies

$$\mathbf{u}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = 0 \quad \text{on} \quad \Gamma. \tag{2.2}$$

Thus, noticing (2.1) and (2.2), the weak mixed formulation for the problem (1.1)–(1.2) is to seek $(p, \mathbf{u}) \in W \times \mathbf{V}_0$ such that

$$\begin{aligned} &a(\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) = 0 && \forall\, \mathbf{v} \in \mathbf{V}_0, \\ &(\nabla \cdot \mathbf{u}, w) + (cp, w) = (f, w) && \forall\, w \in W, \end{aligned} \tag{2.3}$$

where $a(\cdot,\cdot)$ is a bilinear form defined by

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega A^{-1}\mathbf{u}\cdot\mathbf{v}d\Omega,$$

and $(\cdot,\cdot)$ denotes the standard L^2-inner product.

Let $\mathcal{T}_{h_1,h_2}$ be a finite element partition of Ω into uniform rectangles and $\mathbf{V}_{h_1,h_2}\times W_{h_1,h_2}\subset\mathbf{V}\times W$ denote a pair of finite element spaces satisfying the Babuška–Brezzi condition, where h_1 and h_2 are the mesh sizes in the x- and y-axis, respectively. Even if there are now several choices for $\mathbf{V}_{h_1,h_2}$ and W_{h_1,h_2}, here we will consider only the Raviart–Thomas space of the lowest order; i.e.,

$$\begin{aligned}\mathbf{V}_{h_1,h_2} &:= \{\mathbf{v}_{h_1,h_2}\in\mathbf{V}:\ \ \mathbf{v}_{h_1,h_2}|_e\in Q_{1,0}(e)\times Q_{0,1}(e),\ \ e\in\mathcal{T}_{h_1,h_2}\},\\ W_{h_1,h_2} &:= \{w_{h_1,h_2}\in W:\ \ w_{h_1,h_2}|_e\in Q_{0,0}(e),\ \ e\in\mathcal{T}_{h_1,h_2}\},\end{aligned}\tag{2.4}$$

where $Q_{m,n}(e)$ indicates the space of polynomials of degree no more than m and n in x and y on e, respectively. Following the steps of the paper, the extension to other stable rectangular element spaces can be also made. Hence, the corresponding discrete mixed finite element version of (2.3) seeks a pair $(p_{h_1,h_2},\mathbf{u}_{h_1,h_2})\in W_{h_1,h_2}\times\mathbf{V}_{0,h_1,h_2}\subset W\times\mathbf{V}_0$ such that

$$\begin{aligned}&a(\mathbf{u}_{h_1,h_2},\mathbf{v})-(\nabla\cdot\mathbf{v},p_{h_1,h_2})=0 &&\forall\ \mathbf{v}\in\mathbf{V}_{0,h_1,h_2},\\ &(\nabla\cdot\mathbf{u}_{h_1,h_2},w)+(cp_{h_1,h_2},w)=(f,w) &&\forall\ w\in W_{h_1,h_2}.\end{aligned}\tag{2.5}$$

Moreover, from (2.3) and (2.5) one derives the following mixed finite element error equation:

$$\begin{aligned}&a(\mathbf{u}-\mathbf{u}_{h_1,h_2},\mathbf{v})-(\nabla\cdot\mathbf{v},p-p_{h_1,h_2})=0 &&\forall\ \mathbf{v}\in\mathbf{V}_{0,h_1,h_2},\\ &(\nabla\cdot(\mathbf{u}-\mathbf{u}_{h_1,h_2}),w)+(c(p-p_{h_1,h_2}),w)=0 &&\forall\ w\in W_{h_1,h_2}.\end{aligned}\tag{2.6}$$

Let us recall that the Raviart–Thomas projection

$$\Pi^0_{h_1,h_2}\times P^0_{h_1,h_2}:\ \ \mathbf{V}\times W\to\mathbf{V}_{0,h_1,h_2}\times W_{h_1,h_2}$$

is defined by the following conditions:

$$\begin{aligned}&\int_{s_i}(\mathbf{u}-\Pi^0_{h_1,h_2}\mathbf{u})\cdot\mathbf{n}_i ds=0,\quad i=1,2,3,4,\\ &\int_e(p-P^0_{h_1,h_2}p)de=0,\end{aligned}\tag{2.7}$$

where s_i $(i=1,2,3,4)$ are the four edges of the rectangle e and $\mathbf{n}_i$ is the outward normal direction on s_i (see the left-hand side of Figure 2.1).

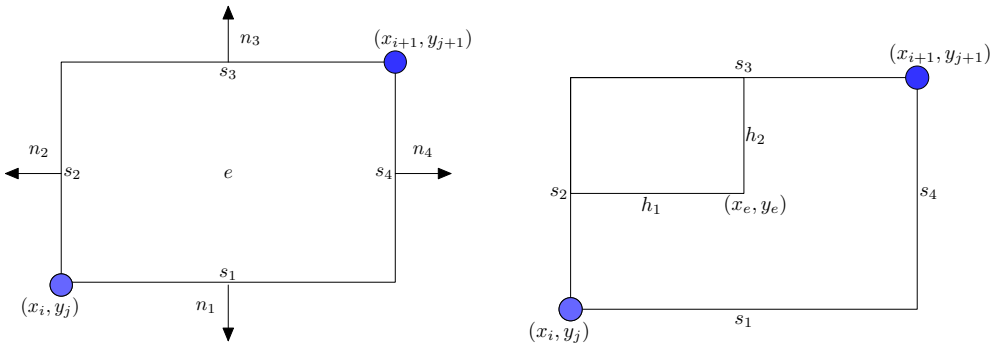


FIG. 2.1. *The rectangular element e and its four edges.*

This projection enjoys the following properties [11]:
(i) $P^0_{h_1,h_2}$ is the local $L^2(\Omega)$ projection;
(ii) $\Pi^0_{h_1,h_2}$ and $P^0_{h_1,h_2}$ satisfy

$$\begin{aligned} &(\nabla \cdot (\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), w) = 0, \quad && w \in W_{h_1,h_2}, \\ &(\nabla \cdot \mathbf{v}, p - P^0_{h_1,h_2} p) = 0, && \mathbf{v} \in \mathbf{V}_{h_1,h_2}; \end{aligned} \tag{2.8}$$

(iii) the following approximation properties hold:

$$\begin{aligned} &||\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}||_0 \le Ch||\mathbf{u}||_1, \\ &||\nabla \cdot (\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u})||_{-s} \le Ch^{1+s}||\nabla \cdot \mathbf{u}||_1, && 0 \le s \le 1, \\ &||p - P^0_{h_1,h_2} p||_{-s} \le Ch^{1+s}||p||_1, && 0 \le s \le 1, \end{aligned} \tag{2.9}$$

where $h := \max\{h_1, h_2\}$.

3. The asymptotic expansion. The aim of this section is to give an asymptotic error expansion for the Raviart–Thomas projection. To this end, we first introduce some notation for future use.

For any element $e \in \mathcal{T}_{h_1,h_2}$, let (x_e, y_e) stand for the center of e and let $2h_1$ and $2h_2$ stand for the side lengths of e in the x- and y-direction, respectively. Define two error functions for x and y as follows [18, 19]:

$$E(x) := \frac{1}{2}[(x - x_e)^2 - h_1^2] \quad \text{and} \quad F(y) := \frac{1}{2}[(y - y_e)^2 - h_2^2].$$

Then, we have

$$(E^m)^{(r)}|_{s_i} = 0 \ (i = 2, 4) \ \text{ and } \ (F^m)^{(r)}|_{s_i} = 0 \ (i = 1, 3) \ \text{ when } \ r \le m - 1, \tag{3.1}$$

where s_2, s_4 and s_1, s_3 are the sides of e parallel to the y-axis and x-axis, respectively (see the right-hand side of Figure 2.1).

In addition, it is easy to check that

$$
\begin{aligned}
&E = \frac{1}{6}(E^2)_{xx} - \frac{1}{3}h_1^2, && F = \frac{1}{6}(F^2)_{yy} - \frac{1}{3}h_2^2,\\
&(x-x_e)^2 = \frac{1}{3}(E^2)_{xx} + \frac{1}{3}h_1^2, && (y-y_e)^2 = \frac{1}{3}(F^2)_{yy} + \frac{1}{3}h_2^2,\\
&x - x_e = \frac{1}{6}(E^2)_{xxx}, && y - y_e = \frac{1}{6}(F^2)_{yyy},\\
&E^2 = \frac{1}{420}(E^4)_{xxxx} - \frac{2}{21}h_1^2(E^2)_{xx} + \frac{2}{15}h_1^4,\\
&F^2 = \frac{1}{420}(F^4)_{yyyy} - \frac{2}{21}h_2^2(F^2)_{yy} + \frac{2}{15}h_2^4,
\end{aligned}
\tag{3.2}
$$

where $f_{xx} = \frac{\partial^2 f}{\partial x^2}$, $f_{xxx} = \frac{\partial^3 f}{\partial x^3}$, and so on.

We are now in a position to derive an asymptotic expansion for the error of the Raviart–Thomas projection in the Raviart–Thomas space of the lowest order.

THEOREM 3.1. *Assume that* $\mathbf{u} \in \mathbf{V} \cap (H^4(\Omega))^2$ *and* $\alpha_{ij} \in H^4(\Omega)$ $(1 \le i, j \le 2)$. *Then we have*

$$
\begin{aligned}
(\alpha \cdot (\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), \mathbf{v}) = &-\frac{h_1^2}{3}\int_\Omega [\alpha_{11}(u_1)_{xx} + \alpha_{12}(u_2)_{xx}]v_1 d\Omega\\
&+ \frac{h_1^2}{3}\int_\Omega [(\alpha_{22})_x(u_2)_x - \alpha_{21}(u_1)_{xx}]v_2 d\Omega\\
&+ \frac{h_2^2}{3}\int_\Omega [(\alpha_{11})_y(u_1)_y - \alpha_{12}(u_2)_{yy}]v_1 d\Omega\\
&- \frac{h_2^2}{3}\int_\Omega [\alpha_{22}(u_2)_{yy} + \alpha_{21}(u_1)_{yy}]v_2 d\Omega\\
&+ O(h^4)||\mathbf{u}||_4||\mathbf{v}||_0, \quad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2},
\end{aligned}
$$

where u_1, u_2 *and* v_1, v_2 *are the first components and the second components of the vector-valued functions* $\mathbf{u}$ *and* $\mathbf{v}$, *respectively, and* $\alpha = (\alpha_{ij})_{2\times 2}$ *is the inverse of the matrix* $A : \alpha = A^{-1}$.

Proof. Set

$$
\begin{aligned}
&(\alpha \cdot (\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), \mathbf{v})\\
&= \sum_{e\in\mathcal{T}_{h_1,h_2}} \int_e [\alpha_{11}(u_1 - \Pi^0_{1,h_1,h_2}u_1) + \alpha_{12}(u_2 - \Pi^0_{2,h_1,h_2}u_2)]v_1 de\\
&\quad + \sum_{e\in\mathcal{T}_{h_1,h_2}} \int_e [\alpha_{21}(u_1 - \Pi^0_{1,h_1,h_2}u_1) + \alpha_{22}(u_2 - \Pi^0_{2,h_1,h_2}u_2)]v_2 de\\
&:= \sum_{e\in\mathcal{T}_{h_1,h_2}} (I + II).
\end{aligned}
\tag{3.3}
$$

Since the treatment for II is the same as that for I, we deal only with I in Lemmas 3.2 and 3.3. ∎

LEMMA 3.2. *We have under the conditions of Theorem* 3.1 *that*

$$\int_\Omega \alpha_{11}(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1 d\Omega = -\frac{h_1^2}{3}\int_\Omega \alpha_{11}(u_1)_{xx}v_1 d\Omega + \frac{h_2^2}{3}\int_\Omega (\alpha_{11})_y(u_1)_y v_1 d\Omega + O(h^4)||u_1||_4||v_1||_0.$$

Proof. Denoting the first term of I in (3.3) by I_1, we obtain from the Taylor expansion of α_{11} and (2.9) that

$$\begin{aligned} I_1 &= \int_e [\alpha_{11}(x,y_e) + (y-y_e)(\alpha_{11})_y(x,y_e)](u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1 de \\ &+ \frac{1}{2}\int_e (y-y_e)^2(\alpha_{11})_{yy}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1 de + O(h^4)||u_1||_{1,e}||v_1||_{0,e} \\ &:= I_{11} + I_{12} + I_{13} + O(h^4)||u_1||_{1,e}||v_1||_{0,e}. \end{aligned} \tag{3.4}$$

Next we shall deal with I_{11}, I_{12}, and I_{13}, respectively.

It follows from the Taylor expansion of v_1 that

$$\begin{aligned} I_{11} &= \int_e \alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)[v_1(x_e,y) + (x-x_e)(v_1)_x]de \\ &:= I_{111} + I_{112}. \end{aligned} \tag{3.5}$$

From (2.7) we find that

$$\int_{s_i} (u_1 - \Pi^0_{1,h_1,h_2}u_1)ds = 0, \quad i = 2,4,$$

where s_2 and s_4 are the two sides of e parallel to the y-axis. Thus, we further obtain for I_{111} according to the definition of the error function $E(x)$ and integration by parts as well as (3.1) that

$$\begin{aligned} I_{111} &= \int_e \alpha_{11}(x,y_e)E_{xx}(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1(x_e,y)de \\ &= \int_e E[\alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)]_{xx}v_1(x_e,y)de \\ &= \int_e E[(u_1)_{xx}\alpha_{11}(x,y_e) + 2(u_1 - \Pi^0_{1,h_1,h_2}u_1)_x(\alpha_{11})_x(x,y_e)]v_1(x_e,y)de \\ &+ \int_e E(u_1 - \Pi^0_{1,h_1,h_2}u_1)(\alpha_{11})_{xx}(x,y_e)v_1(x_e,y)de \\ &:= I^1_{111} + I^2_{111} + I^3_{111}. \end{aligned} \tag{3.6}$$

For I^1_{111} it follows from (3.2), integration by parts, and the Taylor expansion of $\alpha_{11}(x,y_e)$,

$$\alpha_{11}(x,y_e) = \alpha_{11}(x,y) - F_y(\alpha_{11})_y(x,\xi_1),$$

where ξ_1 is between y_e and y, that

$$\begin{aligned}
I^1_{111} &= \frac{1}{6}\int_e (E^2)_{xx}(u_1)_{xx}\alpha_{11}(x,y_e)v_1(x_e,y)de - \frac{h_1^2}{3}\int_e (u_1)_{xx}\alpha_{11}(x,y_e)v_1(x_e,y)de \\
&= \frac{1}{6}\int_e E^2[(u_1)_{xx}\alpha_{11}(x,y_e)]_{xx}v_1(x_e,y)de - \frac{h_1^2}{3}\int_e (u_1)_{xx}\alpha_{11}v_1(x_e,y)de \\
&\quad - \frac{h_1^2}{3}\int_e (u_1)_{xx}F(\alpha_{11})_y(x,\xi_1)v_1(x_e,y)de \\
&= O(h^4)||u_1||_{4,e}||v_1||_{0,e} - \frac{h_1^2}{3}\int_e (u_1)_{xx}\alpha_{11}v_1(x_e,y)de \\
&= -\frac{h_1^2}{3}\int_e (u_1)_{xx}\alpha_{11}v_1(x,y)de - \frac{h_1^2}{3}\int_e E[(u_1)_{xx}\alpha_{11}]_x(v_1)_x de \\
&\quad + O(h^4)||u_1||_{4,e}||v_1||_{0,e} \\
&= -\frac{h_1^2}{3}\int_e \alpha_{11}(u_1)_{xx}v_1 de + \frac{h_1^4}{9}\int_e [(u_1)_{xx}\alpha_{11}]_x(v_1)_x de \\
&\quad - \frac{h_1^2}{18}\int_e (E^2)_{xx}[(u_1)_{xx}\alpha_{11}]_x(v_1)_x de + O(h^4)||u_1||_{4,e}||v_1||_{0,e} \\
&= -\frac{h_1^2}{3}\int_e \alpha_{11}(u_1)_{xx}v_1 de + \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[(u_1)_{xx}\alpha_{11}]_x v_1 ds \\
&\quad + O(h^4)||u_1||_{4,e}||v_1||_{0,e}. \qquad (3.7)
\end{aligned}$$

Here we have used (3.1) and the standard inverse inequality for finite element functions,

$$h||v_1||_{1,e} \le C||v_1||_{0,e}.$$

Similarly, we have for I^2_{111} by (3.2), integration by parts, and the finite element inverse inequality as well as (2.7) that

$$\begin{aligned}
I^2_{111} &= -\frac{2}{3}h_1^2\int_e (u_1 - \Pi^0_{1,h_1,h_2}u_1)_x(\alpha_{11})_x(x,y_e)v_1(x_e,y)de \\
&\quad + \frac{1}{3}\int_e (E^2)_{xx}(u_1 - \Pi^0_{1,h_1,h_2}u_1)_x(\alpha_{11})_x(x,y_e)v_1(x,y_e)de \\
&= -\frac{2}{3}h_1^2\left(\int_{s_4} - \int_{s_2}\right)(u_1 - \Pi^0_{1,h_1,h_2}u_1)(\alpha_{11})_x(x,y_e)v_1(x_e,y)ds \\
&\quad + \frac{2}{3}h_1^2\int_e (u_1 - \Pi^0_{1,h_1,h_2}u_1)(\alpha_{11})_{xx}(x,y_e)v_1 de + O(h_1^4)||u_1||_{3,e}||v_1||_{0,e} \qquad (3.8) \\
&= \frac{2}{3}h_1^2\int_e E_{xx}(u_1 - \Pi^0_{1,h_1,h_2}u_1)(\alpha_{11})_{xx}(x,y_e)v_1(x_e,y)de \\
&\quad + O(h_1^4)||u_1||_{3,e}||v_1||_{0,e} \\
&= O(h_1^4)||u_1||_{3,e}||v_1||_{0,e}.
\end{aligned}$$

Also, we obtain for I^3_{111} that

$$I^3_{111} = O(h_1^4)||u_1||_{2,e}||v_1||_{0,e}. \tag{3.9}$$

Thus, combining (3.7)–(3.9) with (3.6) yields

$$\begin{aligned} I_{111} = &-\frac{h_1^2}{3}\int_e \alpha_{11}(u_1)_{xx}v_1 de + \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[(u_1)_{xx}\alpha_{11}]_x v_1 ds \\ &+ O(h^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned} \tag{3.10}$$

Next we turn our attention to I_{112} in (3.5).

Again, it follows from (3.2), integration by parts, and (2.7) that

$$\begin{aligned} I_{112} &= \frac{1}{6}\int_e (E^2)_{xxx}\alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)(v_1)_x de \\ &= -\frac{1}{6}\int_e E^2[\alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)]_{xxx}(v_1)_x de \\ &= -\frac{1}{6}\int_e \left[\frac{2}{15}h_1^4 + \frac{1}{420}(E^4)_{xxxx}\right]\left[\alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)\right]_{xxx}(v_1)_x de \\ &\quad + \frac{1}{3}\int_e \frac{1}{21}h_1^2(E^2)_{xx}\left[\alpha_{11}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)\right]_{xxx}(v_1)_x de \\ &= -\frac{h_1^4}{45}\int_e [\alpha_{11}(x,y_e)(u_1)_{xxx} + 3(\alpha_{11})_x(x,y_e)(u_1)_{xx}](v_1)_x de \\ &\quad - \frac{h_1^4}{45}\int_e 3(\alpha_{11})_{xx}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)_x(v_1)_x de \\ &\quad - \frac{h_1^4}{45}\int_e (\alpha_{11})_{xxx}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)(v_1)_x de \\ &\quad + O(h_1^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned}$$

For the second and third terms of I_{112}, we have from the Cauchy–Schwartz inequality, (2.9), and the inverse estimate for finite element functions that

$$\begin{aligned} &\left|-\frac{h_1^4}{45}\int_e [3(\alpha_{11})_{xx}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)_x \right. \\ &\qquad \left. + (\alpha_{11})_{xxx}(x,y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)](v_1)_x de\right| \\ &\le Ch_1^4(||u_1 - \Pi^0_{1,h_1,h_2}u_1||_{1,e} + ||u_1 - \Pi^0_{1,h_1,h_2}u_1||_{0,e})||v_1||_{1,e} \\ &\le Ch_1^4(Ch_1||u_1||_{2,e} + Ch_1||u_1||_{1,e})||v_1||_{1,e} \\ &\le Ch_1^5||u_1||_{2,e}||v_1||_{1,e} \le Ch_1^4||u_1||_{2,e}||v_1||_{0,e}, \end{aligned}$$

which yields

$$I_{112} = -\frac{h_1^4}{45}\int_e [\alpha_{11}(x,y_e)(u_1)_{xxx}+3(\alpha_{11})_x(x,y_e)(u_1)_{xx}](v_1)_x de+O(h_1^4)||u_1||_{4,e}||v_1||_{0,e}.$$

Furthermore, we know from the Taylor expansions of $\alpha_{11}(x,y_e)$ and $(\alpha_{11})_x(x,y_e)$,

$$\alpha_{11}(x,y_e) = \alpha_{11}(x,y) + O(h_2), \quad (\alpha_{11})_x(x,y_e) = (\alpha_{11})_x(x,y) + O(h_2),$$

and the inverse inequality for finite element functions that

$$\begin{aligned} I_{112} =& -\frac{h_1^4}{45}\int_e [\alpha_{11}(x,y)(u_1)_{xxx} + 3(\alpha_{11})_x(x,y)(u_1)_{xx}](v_1)_x de \\ & -\frac{h_1^4}{45}\int_e [O(h_2)(u_1)_{xxx} + 3O(h_2)(u_1)_{xx}](v_1)_x de + O(h_1^4)||u_1||_{4,e}||v_1||_{0,e} \\ =& -\frac{h_1^4}{45}\int_e [\alpha_{11}(u_1)_{xxx} + 3(\alpha_{11})_x(u_1)_{xx}](v_1)_x de + O(h_1^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned}$$

This, together with integration by parts with respect to the variable x, implies

$$\begin{aligned} & I_{112} \\ =& -\frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{11}(u_1)_{xxx} + 3(\alpha_{11})_x(u_1)_{xx}]v_1 ds \\ & + \frac{h_1^4}{45}\int_e [\alpha_{11}(u_1)_{xxx} + 3(\alpha_{11})_x(u_1)_{xx}]_x v_1 de + O(h_1^4)||u_1||_{4,e}||v_1||_{0,e} \\ =& -\frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{11}(u_1)_{xxx} + 3(\alpha_{11})_x(u_1)_{xx}]v_1 ds + O(h_1^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned}$$

Thus, by means of (3.5) and (3.10), we have

$$\begin{aligned} I_{11} =& -\frac{h_1^2}{3}\int_e \alpha_{11}(u_1)_{xx}v_1 de + \frac{h_1^2}{9}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{11}(u_1)_{xx}]_x v_1 ds \\ & - \frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{11}(u_1)_{xxx} + 3(\alpha_{11})_x(u_1)_{xx}]v_1 ds \\ & + O(h^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned} \tag{3.11}$$

We shall handle I_{12} in (3.4) as follows.

Noticing $v_1|_e \in Q_{1,0}(e)$ and (3.1), we get from the definition of the error function $F(y)$ and integration by parts as well as (3.2) that

$$\begin{aligned} I_{12} =& -\int_e F(\alpha_{11})_y(x,y_e)(u_1)_y v_1 de \\ =& \frac{h_2^2}{3}\int_e (\alpha_{11})_y(x,y_e)(u_1)_y v_1 de + O(h_2^4)||u_1||_{3,e}||v_1||_{0,e}. \end{aligned}$$

Therefore, it follows from the Taylor expansion of $(\alpha_{11})_y(x, y_e)$,

$$(\alpha_{11})_y(x, y_e) = (\alpha_{11})_y(x, y) + (y_e - y)(\alpha_{11})_{yy}(x, y) + O(h_2^2),$$

integration by parts, and (3.1) that

$$\begin{aligned} I_{12} &= \frac{h_2^2}{3}\int_e (\alpha_{11})_y (u_1)_y v_1 de - \frac{h_2^2}{3}\int_e F_y (\alpha_{11})_{yy}(u_1)_y v_1 de \\ &\quad + O(h_2^4)||u_1||_{3,e}||v_1||_{0,e} \\ &= \frac{h_2^2}{3}\int_e (\alpha_{11})_y (u_1)_y v_1 de - \frac{h_2^2}{3}\left(\int_{s_3} - \int_{s_1}\right) F(\alpha_{11})_{yy}(u_1)_y v_1 ds \\ &\quad + \frac{h_2^2}{3}\int_e F[(\alpha_{11})_{yy}(u_1)_y]_y v_1 de + O(h_2^4)||u_1||_{3,e}||v_1||_{0,e} \\ &= \frac{h_2^2}{3}\int_e (\alpha_{11})_y (u_1)_y v_1 de + O(h_2^4)||u_1||_{3,e}||v_1||_{0,e}. \end{aligned} \tag{3.12}$$

As to I_{13}, one can find from (3.2), the Taylor expansion of v_1,

$$v_1(x, y) = v_1(x_e, y) + (x - x_e)(v_1)_x,$$

and integration by parts that

$$\begin{aligned} I_{13} &= \frac{h_2^2}{6}\int_e (\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1(x_e, y) de \\ &\quad + \frac{h_2^2}{6}\int_e (\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)(x - x_e)(v_1)_x de \\ &\quad + \frac{1}{6}\int_e (F^2)_{yy}(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1 de \\ &:= I_{13}^1 + I_{13}^2 + O(h_2^4)||u_1||_{2,e}||v_1||_{0,e}. \end{aligned} \tag{3.13}$$

It follows from (2.7), (3.1), and integration by parts that

$$\begin{aligned} I_{13}^1 &= \frac{h_2^2}{6}\int_e E_{xx}(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1(x_e, y) de \\ &= \frac{h_2^2}{6}\left(\int_{s_4} - \int_{s_2}\right) E_x(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)v_1(x_e, y) ds \\ &\quad - \frac{h_2^2}{6}\int_e E_x[(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)]_x v_1(x_e, y) de \\ &= -\frac{h_2^2}{6}\int_e E_x[(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)]_x v_1(x_e, y) de \\ &= \frac{h_2^2}{6}\int_e E[(\alpha_{11})_{yy}(x, y_e)(u_1 - \Pi^0_{1,h_1,h_2}u_1)]_{xx} v_1(x_e, y) de \\ &= O(h^4)||u_1||_{2,e}||v_1||_{0,e}. \end{aligned} \tag{3.14}$$

Analogously, we have via (2.7), (3.1)–(3.2), and integration by parts as well as the finite element inverse inequality that

$$\begin{aligned} I_{13}^2 &= \frac{h_2^2}{36}\int_e (E^2)_{xxx}(\alpha_{11})_{yy}(x,y_e)(u_1-\Pi^0_{1,h_1,h_2}u_1)(v_1)_x de \\ &= \frac{h_2^2}{36}\int_e (E^2)_x[(\alpha_{11})_{yy}(x,y_e)(u_1-\Pi^0_{1,h_1,h_2}u_1)]_{xx}(v_1)_x de \\ &= O(h^4)||u_1||_{2,e}||v_1||_{0,e}, \end{aligned}$$

which, together with (3.13) and (3.14), implies

$$I_{13} = O(h^4)||u_1||_{2,e}||v_1||_{0,e}. \tag{3.15}$$

Finally, combining (3.11), (3.12), and (3.15) with (3.4) results in

$$\begin{aligned} I_1 = &-\frac{h_1^2}{3}\int_e \alpha_{11}(u_1)_{xx}v_1 de + \frac{h_2^2}{3}\int_e (\alpha_{11})_y(u_1)_y v_1 de \\ &+ \frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)[4\alpha_{11}(u_1)_{xxx} + 2(\alpha_{11})_x(u_1)_{xx}]v_1 ds \\ &+ O(h^4)||u_1||_{4,e}||v_1||_{0,e}. \end{aligned}$$

The line integrals in the above can be canceled if we sum up I_1 over all the elements $e \in \mathcal{T}_{h_1,h_2}$. In fact, the integral on s_2 will be canceled from a similar contribution from the element to the immediate left of the element e if s_2 is an interior edge. In the case that s_2 is a boundary edge, the trace of v_1 on s_2 is vanishing. To summarize, we have obtained the expansion stated in Lemma 3.2 □

LEMMA 3.3. *Assume that the conditions of Theorem* 3.1 *hold. Then, there exists the following asymptotic expansion:*

$$\begin{aligned} \int_\Omega \alpha_{12}(u_2-\Pi^0_{2,h_1,h_2}u_2)v_1 d\Omega = &-\frac{h_1^2}{3}\int_\Omega \alpha_{12}(u_2)_{xx}v_1 d\Omega - \frac{h_2^2}{3}\int_\Omega \alpha_{12}(u_2)_{yy}v_1 d\Omega \\ &+ O(h^4)||u_2||_4||v_1||_0. \end{aligned}$$

Proof. By letting I_2 stand for the second term of I in (3.3), one finds from the Taylor expansion of v_1 with respect to x_e that

$$\begin{aligned} I_2 &= \int_e \alpha_{12}(u_2-\Pi^0_{2,h_1,h_2}u_2)[v_1(x_e,y) + (v_1)_x(x-x_e)]de \\ &:= I_{21} + I_{22}. \end{aligned} \tag{3.16}$$

For I_{21} we derive from the Taylor expansion of α_{12} with respect to x_e and (2.9) that

$$\begin{aligned} I_{21} = &\int_e [\alpha_{12}(x_e,y) + (x-x_e)(\alpha_{12})_x(x_e,y)](u_2-\Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de \\ &+ \frac{1}{2}\int_e (x-x_e)^2(\alpha_{12})_{xx}(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de \\ &+ O(h^4)||u_2||_{2,e}||v_1||_{0,e} := I_{211} + I_{212} + I_{213} + O(h^4)||u_2||_{2,e}||v_1||_{0,e}. \end{aligned} \tag{3.17}$$

Next we shall estimate I_{211}, I_{212}, and I_{213}, respectively.

Similar to I_{111} we have by means of (2.7), (3.1)–(3.2), and integration by parts that

(3.18)
$$\begin{aligned} I_{211} &= \int_e F[\alpha_{12}(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)]_{yy}v_1(x_e,y)de \\ &= -\frac{h_2^2}{3}\int_e [\alpha_{12}(x_e,y)(u_2)_{yy}+2(\alpha_{12})_y(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)_y]v_1(x_e,y)de \\ &\quad -\frac{h_2^2}{3}\int_e (\alpha_{12})_{yy}(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de + O(h_2^4)||u_2||_{4,e}||v_1||_{0,e} \\ &:= I^1_{211}+I^2_{211}+I^3_{211}+O(h_2^4)||u_2||_{4,e}||v_1||_{0,e}. \end{aligned}$$

From the Taylor expansion of $\alpha_{12}(x_e,y)$ we know that there exists ξ_2 between x_e and x such that

$$\alpha_{12}(x_e,y)=\alpha_{12}(x,y)-E_x(\alpha_{12})_x(\xi_2,y),$$

which, together with the Taylor expansion of $v_1(x_e,y)$ and (3.2), leads to

$$\begin{aligned} I^1_{211} &= -\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}v_1de+\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}E_x(v_1)_xde \\ &\quad -\frac{h_2^2}{3}\int_e E[(\alpha_{12})_x(\xi_2,y)(u_2)_{yy}]_xv_1(x_e,y)de \\ &= -\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}v_1de-\frac{h_2^2}{3}\int_e E[\alpha_{12}(u_2)_{yy}]_x(v_1)_xde \\ &\quad +O(h^4)||u_2||_{3,e}||v_1||_{0,e} \\ &= -\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}v_1de+\frac{(h_2h_1)^2}{9}\left(\int_{s_4}-\int_{s_2}\right)[\alpha_{12}(u_2)_{yy}]_xv_1ds \\ &\quad -\frac{(h_2h_1)^2}{9}\int_e[\alpha_{12}(u_2)_{yy}]_{xx}v_1de+\frac{h_2^2}{18}\int_e(E^2)_x[\alpha_{12}(u_2)_{yy}]_{xx}(v_1)_xde \\ &\quad +O(h^4)||u_2||_{3,e}||v_1||_{0,e} \\ &= -\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}v_1de+\frac{(h_2h_1)^2}{9}\left(\int_{s_4}-\int_{s_2}\right)[\alpha_{12}(u_2)_{yy}]_xv_1ds \\ &\quad +O(h^4)||u_2||_{4,e}||v_1||_{0,e}. \end{aligned} \tag{3.19}$$

For u_2, we have by (2.7) that

$$\int_{s_i}(u_2-\Pi^0_{2,h_1,h_2}u_2)ds=0,\quad i=1,3,$$

where s_1 and s_3 are the two sides of e parallel to the x-axis. Thus, analogous to I^2_{111} we can also obtain via integration by parts with respect to y and (3.2) that

$$I^2_{211}=O(h_2^4)||u_2||_{2,e}||v_1||_{0,e}. \tag{3.20}$$

Similarly, we have

$$I_{211}^3 = O(h_2^4)||u_2||_{2,e}||v_1||_{0,e},$$

which, together with (3.18)–(3.20), implies

$$\begin{aligned} I_{211} &= -\frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy}v_1 de + \frac{(h_2h_1)^2}{9}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{12}(u_2)_{yy}]_x v_1 ds \\ &\quad + O(h^4)||u_2||_{4,e}||v_1||_{0,e}. \end{aligned} \tag{3.21}$$

Also, we can obtain from integration by parts and (3.2) that

$$\begin{aligned} I_{212} &= -\int_e E(\alpha_{12})_x(x_e,y)(u_2)_x v_1(x_e,y)de \\ &= \frac{h_1^2}{3}\int_e [(\alpha_{12})_x + (x_e - x)(\alpha_{12})_{xx}](u_2)_x v_1(x_e,y)de \\ &\quad + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e} \\ &= \frac{h_1^2}{3}\int_e (\alpha_{12})_x(u_2)_x v_1 de - \frac{h_1^2}{3}\int_e E_x(\alpha_{12})_x(u_2)_x(v_1)_x de \\ &\quad + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e} \\ &= \frac{h_1^2}{3}\int_e (\alpha_{12})_x(u_2)_x v_1 de - \frac{h_1^4}{9}\int_e [(\alpha_{12})_x(u_2)_x]_x(v_1)_x de \\ &\quad + \frac{h_1^2}{18}\int_e (E^2)_{xx}[(\alpha_{12})_x(u_2)_x]_x(v_1)_x de + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e} \\ &= \frac{h_1^2}{3}\int_e (\alpha_{12})_x(u_2)_x v_1 de - \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[(\alpha_{12})_x(u_2)_x]_x v_1 ds \\ &\quad + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e}. \end{aligned} \tag{3.22}$$

Using a similar argument, we obtain

$$\begin{aligned} I_{213} &= \frac{h_1^2}{6}\int_e (\alpha_{12})_{xx}(x_e,y)(u_2 - \Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de \\ &\quad + \frac{1}{6}\int_e (E^2)_{xx}(\alpha_{12})_{xx}(x_e,y)(u_2 - \Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de \\ &= \frac{h_1^2}{6}\int_e F_{yy}(\alpha_{12})_{xx}(x_e,y)(u_2 - \Pi^0_{2,h_1,h_2}u_2)v_1(x_e,y)de \\ &\qquad + O(h_1^4)||u_2||_{2,e}||v_1||_{0,e} \\ &= \frac{h_1^2}{6}\int_e F[(\alpha_{12})_{xx}(x_e,y)(u_2 - \Pi^0_{2,h_1,h_2}u_2)]_{yy}v_1(x_e,y)de \\ &\qquad + O(h_1^4)||u_2||_{2,e}||v_1||_{0,e} \\ &= O(h^4)||u_2||_{2,e}||v_1||_{0,e}. \end{aligned} \tag{3.23}$$

Combining (3.21)–(3.23) with (3.17) implies

$$
\begin{aligned}
I_{21} &= \frac{h_1^2}{3}\int_e (\alpha_{12})_x (u_2)_x v_1 de - \frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy} v_1 de \\
&\quad - \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[(\alpha_{12})_x (u_2)_x]_x v_1 ds \\
&\quad + \frac{(h_1 h_2)^2}{9}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{12}(u_2)_{yy}]_x v_1 ds \\
&\quad + O(h^4)||u_2||_{4,e}||v_1||_{0,e}.
\end{aligned} \tag{3.24}
$$

It remains to deal with I_{22} in (3.16). To this end, we notice that from the Taylor expansion for α_{12} we have

$$
\begin{aligned}
&I_{22} \\
&= \int_e [\alpha_{12}(x_e, y) + (x - x_e)(\alpha_{12})_x (x_e, y)](u_2 - \Pi^0_{2,h_1,h_2} u_2)(x - x_e)(v_1)_x de \\
&\quad + \frac{1}{2}\int_e (x - x_e)^2 (\alpha_{12})_{xx}(x_e, y)(u_2 - \Pi^0_{2,h_1,h_2} u_2)(x - x_e)(v_1)_x de \\
&\quad + O(h_1^4)||u_2||_{1,e}||v_1||_{0,e} := I_{221} + I_{222} + I_{223} + O(h_1^4)||u_2||_{1,e}||v_1||_{0,e}.
\end{aligned} \tag{3.25}
$$

The above terms can be estimated as follows. First, we have

$$
\begin{aligned}
&I_{221} \\
&= -\int_e E\alpha_{12}(x_e, y)(u_2)_x (v_1)_x de \\
&= \frac{h_1^2}{3}\int_e \alpha_{12}(u_2)_x (v_1)_x de - \frac{h_1^2}{3}\int_e E_x \alpha_{12}(\xi_3, y)(u_2)_x (v_1)_x de \\
&\quad - \frac{1}{6}\int_e E^2 \alpha_{12}(x_e, y)(u_2)_{xxx}(v_1)_x de \\
&= \frac{h_1^2}{3}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_x v_1 ds - \frac{h_1^2}{3}\int_e [\alpha_{12}(u_2)_x]_x v_1 de \\
&\quad + \frac{h_1^2}{3}\int_e E[\alpha_{12}(\xi_3, y)(u_2)_x]_x (v_1)_x de \\
&\quad - \frac{1}{6}\int_e \left[\frac{1}{420}(E^4)_{xxxx} - \frac{2}{21}h_1^2 (E^2)_{xx} + \frac{2}{15}h_1^4\right]\alpha_{12}(x_e, y)(u_2)_{xxx}(v_1)_x de \\
&= -\frac{h_1^2}{3}\int_e [\alpha_{12}(u_2)_x]_x v_1 de + \frac{h_1^2}{3}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_x v_1 ds \\
&\quad - \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{12}(u_2)_x]_x v_1 ds - \frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_{xxx} v_1 ds \\
&\quad + O(h_1^4)||u_2||_{4,e}||v_1||_{0,e},
\end{aligned} \tag{3.26}
$$

where ξ_3 is between x_e and x. Also, we have

$$
\begin{aligned}
I_{222} &= \frac{h_1^2}{3}\int_e F_{yy}(\alpha_{12})_x(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)(v_1)_x de \\
&\quad + \frac{1}{3}\int_e E^2(\alpha_{12})_x(x_e,y)(u_2)_{xx}(v_1)_x de \\
&= \frac{h_1^2}{3}\left(\int_{s_3}-\int_{s_1}\right) F_y(\alpha_{12})_x(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)(v_1)_x ds \\
&\quad - \frac{h_1^2}{3}\int_e F_y[(\alpha_{12})_x(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)]_y(v_1)_x de \\
&\quad + \frac{1}{3}\int_e \left[\frac{2}{15}h_1^4+\frac{1}{420}(E^4)_{xxxx}-\frac{2}{21}h_1^2(E^2)_{xx}\right](\alpha_{12})_x(x_e,y)(u_2)_{xx}(v_1)_x de \qquad (3.27)\\
&= -\frac{(h_1h_2)^2}{9}\int_e (\alpha_{12})_x(u_2)_{yy}(v_1)_x de \\
&\quad - \frac{2(h_1h_2)^2}{9}\left(\int_{s_3}-\int_{s_1}\right)(\alpha_{12})_{xy}(x_e,y)(u_2-\Pi^0_{2,h_1,h_2}u_2)(v_1)_x ds \\
&\quad + \frac{2}{45}h_1^4\left(\int_{s_4}-\int_{s_2}\right)(\alpha_{12})_x(u_2)_{xx}v_1 ds + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e} \\
&= -\frac{(h_1h_2)^2}{9}\left(\int_{s_4}-\int_{s_2}\right)(\alpha_{12})_x(u_2)_{yy}v_1 ds \\
&\quad + \frac{2}{45}h_1^4\left(\int_{s_4}-\int_{s_2}\right)(\alpha_{12})_x(u_2)_{xx}v_1 ds + O(h_1^4)||u_2||_{3,e}||v_1||_{0,e}.
\end{aligned}
$$

Since

$$(x-x_e)^3 = \frac{1}{15}(E^3)_{xxx}+\frac{9}{15}h_1^2E_x,$$

we have

$$
\begin{aligned}
I_{223} &= -\frac{3h_1^2}{10}\int_e E(\alpha_{12})_{xx}(x_e,y)(u_2)_x(v_1)_x de + O(h_1^4)||u_2||_{2,e}||v_1||_{0,e} \\
&= \frac{h_1^4}{10}\int_e (\alpha_{12})_{xx}(u_2)_x(v_1)_x de + O(h_1^4)||u_2||_{2,e}||v_1||_{0,e} \\
&= \frac{h_1^4}{10}\left(\int_{s_4}-\int_{s_2}\right)(\alpha_{12})_{xx}(u_2)_x v_1 ds + O(h_1^4)||u_2||_{2,e}||v_1||_{0,e},
\end{aligned}
$$

which, together with (3.25)–(3.27), leads to

$$
\begin{aligned}
I_{22} = &-\frac{h_1^2}{3}\int_e [\alpha_{12}(u_2)_x]_x v_1 de + \frac{h_1^2}{3}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_x v_1 ds \\
&- \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)[\alpha_{12}(u_2)_x]_x v_1 ds - \frac{h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_{xxx} v_1 ds \\
&- \frac{(h_1 h_2)^2}{9}\left(\int_{s_4} - \int_{s_2}\right)(\alpha_{12})_x (u_2)_{yy} v_1 ds \\
&+ \frac{2h_1^4}{45}\left(\int_{s_4} - \int_{s_2}\right)(\alpha_{12})_x (u_2)_{xx} v_1 ds \\
&+ \frac{h_1^4}{10}\left(\int_{s_4} - \int_{s_2}\right)(\alpha_{12})_{xx} (u_2)_x v_1 ds + O(h^4)||u_2||_{4,e}||v_1||_{0,e}.
\end{aligned} \tag{3.28}
$$

Hence, combining (3.24) and (3.28) with (3.16) gives rise to

$$
\begin{aligned}
I_2 = &-\frac{h_1^2}{3}\int_e \alpha_{12}(u_2)_{xx} v_1 de - \frac{h_2^2}{3}\int_e \alpha_{12}(u_2)_{yy} v_1 de \\
&+ \frac{h_1^2}{3}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_x v_1 ds \\
&- \frac{h_1^4}{9}\left(\int_{s_4} - \int_{s_2}\right)\left\{[(\alpha_{12})_x (u_2)_x]_x + [\alpha_{12}(u_2)_x]_x + \frac{1}{5}\alpha_{12}(u_2)_{xxx}\right\} v_1 ds \\
&+ \frac{h_1^4}{5}\left(\int_{s_4} - \int_{s_2}\right)\left[\frac{2}{9}(\alpha_{12})_x (u_2)_{xx} + \frac{1}{2}(\alpha_{12})_{xx}(u_2)_x\right] v_1 ds \\
&+ \frac{(h_1 h_2)^2}{9}\left(\int_{s_4} - \int_{s_2}\right)\alpha_{12}(u_2)_{yyx} v_1 ds + O(h^4)||u_2||_{4,e}||v_1||_{0,e}.
\end{aligned}
$$

Furthermore, summing up I_2 over all the elements $e \in \mathcal{T}_{h,k}$, we can obtain the desired result. ∎

THEOREM 3.4. *Assume that p, $c \in H^3(\Omega)$. Then we have the following asymptotic expansion:*

$$
(c(p - P^0_{h_1,h_2}p), w) = \frac{h_1^2}{3}\int_\Omega c_x p_x w d\Omega + \frac{h_2^2}{3}\int_\Omega c_y p_y w d\Omega + O(h^4)||p||_3||w||_0, \quad w \in W_{h_1,h_2}.
$$

Proof. It follows from (2.7), (2.9), and the Taylor expansion of $c(x, y)$ at x_e that

$$
\begin{aligned}
&\int_e c(x,y)(p - P^0_{h_1,h_2}p) w de \\
&= \int_e [c(x_e, y) + c_x(x_e, y)(x - x_e)](p - P^0_{h_1,h_2}p) w de \\
&\quad + \frac{1}{2}\int_e (x - x_e)^2 c_{xx}(x_e, y)(p - P^0_{h_1,h_2}p) w de + O(h_1^4)||p||_{1,e}||w||_{0,e} \\
&:= III_1 + III_2 + III_3 + O(h_1^4)||p||_{1,e}||w||_{0,e}.
\end{aligned} \tag{3.29}
$$

For III_1 we know from (2.7), (2.9), (3.1)–(3.2), integration by parts, and the Taylor expansion of $c(x_e, y)$ with respect to y_e that

$$\begin{aligned}
III_1 &= \int_e [c(x_e, y_e) + (y - y_e)c_y(x_e, y_e)](p - P^0_{h_1,h_2}p)wde \\
&\quad + \frac{1}{2}\int_e (y - y_e)^2 c_{yy}(x_e, y_e)(p - P^0_{h_1,h_2}p)wde + O(h_2^4)||p||_{1,e}||w||_{0,e} \\
&= -\int_e F c_y(x_e, y_e)p_y wde + \frac{h_2^2}{6}\int_e c_{yy}(x_e, y_e)(p - P^0_{h_1,h_2}p)wde \\
&\quad + \frac{1}{6}\int_e (F^2)_{yy}c_{yy}(x_e, y_e)(p - P^0_{h_1,h_2}p)wde + O(h_2^4)||p||_{1,e}||w||_{0,e} \\
&= \frac{h_2^2}{3}\int_e c_y(x_e, y_e)p_y wde - \frac{1}{6}\int_e (F^2)_{yy}c_y(x_e, y_e)p_y wde + O(h_2^4)||p||_{2,e}||w||_{0,e} \\
&= \frac{h_2^2}{3}\int_e c_y p_y wde - \frac{h_2^2}{3}\int_e E_x c_{yx}p_y wde \\
&\quad - \frac{h_2^2}{3}\int_e F_y c_{yy}p_y wde + O(h_2^4)||p||_{3,e}||w||_{0,e} \\
&= \frac{h_2^2}{3}\int_e c_y u_y wde + O(h^4)||p||_{3,e}||w||_{0,e}. \qquad (3.30)
\end{aligned}$$

We can also get for III_2 that

$$\begin{aligned}
III_2 &= -\int_e E c_x(x_e, y)p_x wde \\
&= \frac{h_1^2}{3}\int_e (c_x - E_x c_x)p_x wde + O(h_1^4)||p||_{3,e}||w||_{0,e} \qquad (3.31) \\
&= \frac{h_1^2}{3}\int_e c_x p_x wde + O(h_1^4)||p||_{3,e}||w||_{0,e}.
\end{aligned}$$

Analogously, we have

$$\begin{aligned}
III_3 &= \frac{h_1^2}{6}\int_e c_{xx}(x_e, y)(p - P^0_{h_1,h_2}p)wde + \frac{1}{6}\int_e c_{xx}(x_e, y)(E^2)_{xx}(p - P^0_{h_1,h_2}p)wde \\
&= \frac{h_1^2}{6}\int_e c_{xxy}(x_e, y_e)F_y(p - P^0_{h_1,h_2}p)wde + O(h_1^4)||p||_{2,e}||w||_{0,e} \\
&= O(h^4)||p||_{2,e}||w||_{0,e},
\end{aligned}$$

which, together with (3.29)–(3.31), implies

$$\int_e c(p - P^0_{h_1,h_2}p)wde = \frac{h_1^2}{3}\int_e c_x p_x wde + \frac{h_2^2}{3}\int_e c_y p_y wde + O(h^4)||p||_{3,e}||w||_{0,e}.$$

Thus,

$$\int_\Omega c(p - P^0_{h_1,h_2}p)wd\Omega = \frac{h_1^2}{3}\int_\Omega c_x p_x w d\Omega + \frac{h_2^2}{3}\int_\Omega c_y p_y w d\Omega + O(h^4)||p||_3||w||_0. \qquad \square$$

From Theorems 3.1 and 3.4 we immediately obtain the following corollaries.

COROLLARY 3.5. *If* $\mathbf{u} \in \mathbf{V}\bigcap(H^2(\Omega))^2$ *and* $\alpha_{ij} \in H^2(\Omega)$ $(1 \le i, j \le 2)$, *we have*

$$|(\alpha \cdot (\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), \mathbf{v})| \le Ch^2||\mathbf{u}||_2||\mathbf{v}||_0, \quad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2}.$$

COROLLARY 3.6. *If* p, $c \in H^1(\Omega)$, *we have*

$$|(c(p - P^0_{h_1,h_2}p), w)| \le Ch^2||p||_1||w||_0, \quad w \in W_{h_1,h_2}.$$

4. The global Richardson extrapolation. In this section we turn to the asymptotic expansions between the mixed finite element solution and the interpolant of the exact solution of the problem (1.1)–(1.2), from which asymptotic expansions between the exact solution and the postprocessed mixed finite element solution by interpolation are further obtained. The Richardson extrapolations of two different schemes will be performed to generate high order approximations to the exact solution of (1.1)–(1.2). First, we recall from [13, 14] the following lemma.

LEMMA 4.1. *Assume that the matrix* A *is positive definite. Then, the norms* $||\mathbf{u}||_0^2 := (\mathbf{u}, \mathbf{u})$ *and* $||\mathbf{u}||_{A^{-1}}^2 := (A^{-1}\mathbf{u}, \mathbf{u})$ *are equivalent.*

4.1. The global Richardson extrapolation in two directions. We first discuss the extrapolation method of mixed finite element approximation for (1.1)–(1.2) in both the x and y directions as follows.

THEOREM 4.2. *Suppose that* $(p, \mathbf{u})$ *and* $(p_{h_1,h_2}, \mathbf{u}_{h_1,h_2})$ *are the exact solution of* (2.3) *and its mixed finite element solution, respectively. Then we have the following asymptotic expansions under the conditions that* p, $c \in H^3(\Omega)$, $\mathbf{u} \in \mathbf{V} \cap (H^4(\Omega))^2$, *and* $\alpha_{ij} \in H^4(\Omega)$ $(1 \le i, j \le 2)$*:*

$$p_{h_1,h_2} - P^0_{h_1,h_2}p = h^2\xi_{h_1,h_2} + r_{h_1,h_2}, \qquad ||r_{h_1,h_2}||_0 \le Ch^4,$$

$$\mathbf{u}_{h_1,h_2} - \Pi^0_{h_1,h_2}\mathbf{u} = h^2\eta_{h_1,h_2} + \mathbf{r}_{h_1,h_2}, \quad ||\mathbf{r}_{h_1,h_2}||_{\mathbf{V}} \le Ch^4,$$

where $(\xi_{h_1,h_2}, \eta_{h_1,h_2}) \in W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$ *and* $P^0_{h_1,h_2} \times \Pi^0_{h_1,h_2} : W \times \mathbf{V}_0 \to W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$ *is the Raviart–Thomas projection operator.*

Proof. Let

$$\rho_{h_1,h_2} := p_{h_1,h_2} - P^0_{h_1,h_2}p, \quad \theta_{h_1,h_2} := \mathbf{u}_{h_1,h_2} - \Pi^0_{h_1,h_2}\mathbf{u}.$$

Then, it follows from (2.6) and (2.8) that

$$\begin{aligned}
&(\alpha\theta_{h_1,h_2}, \mathbf{v}) - (\rho_{h_1,h_2}, \nabla \cdot \mathbf{v}) = (\alpha(\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), \mathbf{v}), \qquad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2}, \\
&(\nabla \cdot \theta_{h_1,h_2}, w) + (c\rho_{h_1,h_2}, w) = (c(p - P^0_{h_1,h_2}p), w), \qquad w \in W_{h_1,h_2},
\end{aligned} \tag{4.1}$$

where $\alpha = A^{-1}$. From Theorems 3.1 and 3.4 we derive that

$$
\begin{aligned}
&(\alpha(\mathbf{u} - \Pi^0_{h_1,h_2}\mathbf{u}), \mathbf{v}) = h^2 L_{h_1,h_2}(\mathbf{v}) + O(h^4)||\mathbf{v}||_0, \quad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2},\\
&(c(p - P^0_{h_1,h_2}p), w) = h^2 G_{h_1,h_2}(w) + O(h^4)||w||_0, \quad w \in W_{h_1,h_2},
\end{aligned} \tag{4.2}
$$

where

$$
\begin{aligned}
G_{h_1,h_2}(\phi) &= \frac{1}{3}\left(\frac{h_1}{h}\right)^2 \int_\Omega c_x p_x \phi d\Omega + \frac{1}{3}\left(\frac{h_2}{h}\right)^2 \int_\Omega c_y p_y \phi d\Omega,\\
L_{h_1,h_2}(\psi) &= -\frac{1}{3}\left(\frac{h_1}{h}\right)^2 \int_\Omega [\alpha_{11}(u_1)_{xx} + \alpha_{12}(u_2)_{xx}]\psi_1 d\Omega\\
&\quad + \frac{1}{3}\left(\frac{h_1}{h}\right)^2 \int_\Omega [(\alpha_{22})_x (u_2)_x - \alpha_{21}(u_1)_{xx}]\psi_2 d\Omega\\
&\quad + \frac{1}{3}\left(\frac{h_2}{h}\right)^2 \int_\Omega [(\alpha_{11})_y (u_1)_y - \alpha_{12}(u_2)_{yy}]\psi_1 d\Omega\\
&\quad - \frac{1}{3}\left(\frac{h_2}{h}\right)^2 \int_\Omega [\alpha_{22}(u_2)_{yy} + \alpha_{21}(u_1)_{yy}]\psi_2 d\Omega.
\end{aligned}
$$

Here, $\psi = (\psi_1, \psi_2)$ is a vector-valued function. Obviously,

$$
L_{h_1/2,h_2/2}(\psi) = L_{h_1,h_2}(\psi) \quad \text{and} \quad G_{h_1/2,h_2/2}(\phi) = G_{h_1,h_2}(\phi). \tag{4.3}
$$

Let $(\xi, \eta) \in W \times \mathbf{V}_0$ and $(\xi_{h_1,h_2}, \eta_{h_1,h_2}) \in W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$ be the exact solution and the mixed finite element solution, respectively, of the following auxiliary problem:

$$
\begin{aligned}
&(\alpha\eta, \mathbf{v}) - (\xi, \nabla \cdot \mathbf{v}) = L_{h_1,h_2}(\mathbf{v}), \qquad \mathbf{v} \in \mathbf{V}_0,\\
&(\nabla \cdot \eta, w) + (c\xi, w) = G_{h_1,h_2}(w), \qquad w \in W.
\end{aligned} \tag{4.4}
$$

Then, from (4.1), (4.2), and (4.4) one finds that

$$
\begin{aligned}
&(\alpha(\theta_{h_1,h_2} - h^2\eta_{h_1,h_2}), \mathbf{v}) - (\rho_{h_1,h_2} - h^2\xi_{h_1,h_2}, \nabla \cdot \mathbf{v}) = O(h^4)||\mathbf{v}||_0, \qquad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2},\\
&(\nabla \cdot (\theta_{h_1,h_2} - h^2\eta_{h_1,h_2}), w) + (c(\rho_{h_1,h_2} - h^2\xi_{h_1,h_2}), w) = O(h^4)||w||_0, \quad w \in W_{h_1,h_2}.
\end{aligned}
$$

Set

$$
\theta^*_{h_1,h_2} := \theta_{h_1,h_2} - h^2\eta_{h_1,h_2} \quad \text{and} \quad \rho^*_{h_1,h_2} := \rho_{h_1,h_2} - h^2\xi_{h_1,h_2}.
$$

Thus, we have

$$
\begin{aligned}
&(\alpha\theta^*_{h_1,h_2}, \mathbf{v}) - (\rho^*_{h_1,h_2}, \nabla \cdot \mathbf{v}) = O(h^4)||\mathbf{v}||_0, \quad \mathbf{v} \in \mathbf{V}_{0,h_1,h_2},\\
&(\nabla \cdot \theta^*_{h_1,h_2}, w) + (c\rho^*_{h_1,h_2}, w) = O(h^4)||w||_0, \quad w \in W_{h_1,h_2},
\end{aligned} \tag{4.5}
$$

which implies by using the standard stability argument [5] that

$$||\theta^*_{h_1,h_2}||_{\mathbf{V}} + ||\rho^*_{h_1,h_2}||_0 \leq C\left(\sup_{\mathbf{v}\in\mathbf{V}_{0,h_1,h_2}} \frac{|O(h^4)|\,||\mathbf{v}||_0}{||\mathbf{v}||_{\mathbf{V}}} + \sup_{w\in W_{h,k}} \frac{|O(h^4)|\,||w||_0}{||w||_0}\right)$$

$$\leq Ch^4.$$

Thus, the proof of Theorem 4.2 is complete. □

Remark 4.1. In another paper we will discuss asymptotic expansions in L^∞-norm that are similar to those in Theorem 4.2 above.

Following the procedure for Theorem 4.2 and utilizing Corollaries 3.5 and 3.6 we can also prove the following result.

LEMMA 4.3. *If* $(\xi,\eta)\in W\times\mathbf{V}_0$ *and* $(\xi_{h_1,h_2},\eta_{h_1,h_2})\in W_{h_1,h_2}\times\mathbf{V}_{0,h_1,h_2}$ *are the variational solution and the mixed finite element solution of* (4.4)*, respectively, then we have the superconvergent estimate*

$$||\xi_{h_1,h_2} - P^0_{h_1,h_2}\xi||_0 + ||\eta_{h_1,h_2} - \Pi^0_{h_1,h_2}\eta||_{\mathbf{V}} \leq Ch^2(||\xi||_1 + ||\eta||_2).$$

Now we use the interpolation postprocessing technique to get a global extrapolation approximation of high accuracy for the pressure and the velocity fields. Analogous to [18, 13, 14] we need to define two postprocessing interpolation operators $\Pi^3_{4h_1,4h_2}$ and $P^3_{4h_1,4h_2}$ to satisfy

$$\begin{aligned}
&\Pi^3_{4h_1,4h_2}\Pi^0_{h_1,h_2} = \Pi^3_{4h_1,4h_2},\\
&||\Pi^3_{4h_1,4h_2}\mathbf{v}||_0 \leq C||\mathbf{v}||_0 \qquad &&\forall\mathbf{v}\in\mathbf{V}_{0,h_1,h_2},\\
&||\Pi^3_{4h_1,4h_2}\mathbf{u}-\mathbf{u}||_0 \leq Ch^4||\mathbf{u}||_4 \qquad &&\forall\mathbf{u}\in(H^4(\Omega))^2,\\
&P^3_{4h_1,4h_2}P^0_{h_1,h_2} = P^3_{4h_1,4h_2},\\
&||P^3_{4h_1,4h_2}w||_0 \leq C||w||_0 \qquad &&\forall w\in W_{h_1,h_2},\\
&||P^3_{4h_1,4h_2}p-p||_0 \leq Ch^4||p||_4 \qquad &&\forall p\in H^4(\Omega).
\end{aligned} \tag{4.6}$$

To this end, assume that the rectangular partition $\mathcal{T}_{h_1,h_2}$ has been obtained from $\mathcal{T}_{4h_1,4h_2}$ with mesh size $4h$ by subdividing each element of $\mathcal{T}_{4h_1,4h_2}$ into sixteen small congruent rectangles. Let $\tau := \bigcup_{i=1}^{16} e_i$ with $e_i\in\mathcal{T}_{h_1,h_2}$. We define two projection operators $\Pi^3_{4h_1,4h_2}$ and $P^3_{4h_1,4h_2}$ associated with $\mathcal{T}_{4h_1,4h_2}$ of degree at most 3 in x and y on τ, respectively, according to the following conditions:

$$\begin{aligned}
&\Pi^3_{4h_1,4h_2}\mathbf{u}|_\tau \in Q_{4,3}(\tau)\times Q_{3,4}(\tau), \qquad &&P^3_{4h_1,4h_2}u|_\tau \in Q_{3,3}(\tau),\\
&\int_{s_i}(\mathbf{u}-\Pi^3_{4h_1,4h_2}\mathbf{u})\cdot\mathbf{n}ds = 0, \qquad &&i=1,2,\ldots,40, \text{ and}\\
&\int_{e_i}(p-P^3_{4h_1,4h_2}p) = 0, \qquad &&i=1,2,\ldots,16,
\end{aligned} \tag{4.7}$$

where s_i $(i=1,2,\ldots,40)$ is one of the forty sides of the sixteen small elements e_i $(i=1,2,\ldots,16)$. It is easy to check that the two operators $\Pi^3_{4h_1,4h_2}$ and $P^3_{4h_1,4h_2}$ defined by (4.7) satisfy the properties described in (4.6).

THEOREM 4.4. *We have under the conditions of Theorem* 4.2 *that*

$$P^3_{4h_1,4h_2}p_{h_1,h_2} - p = h^2\xi + r^*_{h_1,h_2}, \quad ||r^*_{h_1,h_2}||_0 \le Ch^4,$$

$$\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2} - \mathbf{u} = h^2\eta + \mathbf{r}^*_{h_1,h_2}, \quad ||\mathbf{r}^*_{h_1,h_2}||_0 \le Ch^4,$$

where $(\xi,\eta) \in W \times \mathbf{V}_0$ *is the variational solution of* (4.4).

Proof. Let

$$\overline{r}_{h_1,h_2} := p_{h_1,h_2} - P^0_{h_1,h_2}p - h^2P^0_{h_1,h_2}\xi.$$

Then, it follows from Theorem 4.2 and Lemma 4.3 that

$$||\overline{r}_{h_1,h_2}||_0 \le Ch^4.$$

Thus, we find from (4.6) that

$$\begin{aligned}
&P^3_{4h_1,4h_2}p_{h_1,h_2} - p \\
&= P^3_{4h_1,4h_2}(p_{h_1,h_2} - P^0_{h_1,h_2}p) + (P^3_{4h_1,4h_2}p - p) \\
&= P^3_{4h_1,4h_2}(h^2P^0_{h_1,h_2}\xi + \overline{r}_{h_1,h_2}) + (P^3_{4h_1,4h_2}p - p) \\
&= h^2P^3_{4h_1,4h_2}\xi + P^3_{4h_1,4h_2}\overline{r}_{h_1,h_2} + (P^3_{4h_1,4h_2}p - p) \\
&= h^2\xi + h^2(P^3_{4h_1,4h_2}\xi - \xi) + P^3_{4h_1,4h_2}\overline{r}_{h_1,h_2} + (P^3_{4h_1,4h_2}p - p) \\
&= h^2\xi + r^*_{h_1,h_2},
\end{aligned}$$

where

$$r^*_{h_1,h_2} := h^2(P^3_{4h_1,4h_2}\xi - \xi) + P^3_{4h_1,4h_2}\overline{r}_{h_1,h_2} + (P^3_{4h_1,4h_2}p - p)$$

with $||r^*_{h_1,h_2}||_0 \le Ch^4$.

Analogously, we can also get the second equality in the theorem. ∎

Theorem 4.4 guarantees that we can use low order mixed finite element solutions to generate high order approximations by the Richardson extrapolation. And thus, we employ, in addition to $W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$, the Raviart–Thomas mixed finite element space $W_{h_1/2,h_2/2} \times \mathbf{V}_{0,h_1/2,h_2/2}$ of the lowest order gained by subdividing each element $e_i \in \mathcal{T}_{h_1,h_2}$ into four small congruent elements $\hat{e}_{i,j} \in \mathcal{T}_{h_1/2,h_2/2}$ $(j = 1,2,3,4)$. Denote by $(p_{h_1/2,h_2/2}, \mathbf{u}_{h_1/2,h_2/2}) \in W_{h_1/2,h_2/2} \times \mathbf{V}_{0,h_1/2,h_2/2}$ and $P^3_{2h_1,2h_2} \times \Pi^3_{2h_1,2h_2}$ the mixed finite element approximation and the Raviart–Thomas projection of degree at most 3 in x and y with respect to this new partition. From Theorem 4.4 we know under the L^2-norm that

$$P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - p = \left(\frac{h}{2}\right)^2 \xi + O(h^4),$$

which produces by applying the Richardson extrapolation that under the L^2-norm

$$\frac{4P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - P^3_{4h_1,4h_2}p_{h_1,h_2}}{3} = p + O(h^4). \tag{4.8}$$

Similarly, we have under the L^2-norm that

$$\frac{4\Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2} - \Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}}{3} = \mathbf{u} + O(h^4). \tag{4.9}$$

It is very important for a mixed finite element method to have a computable a posteriori error estimator such that we can assess the accuracy of the approximate solutions by means of the error estimator in applications. The superconvergent approximations generated above in (4.8) and (4.9) can be used naturally to produce efficient a posteriori error estimators. In fact, we have by the same way as in Theorem 5.3 in [13] that the following theorem holds.

THEOREM 4.5. *Under the assumptions of Theorem* 4.4*, we have*

$$\begin{aligned} &||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0 \\ &\quad = \frac{1}{3}||P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - P^3_{4h_1,4h_2}p_{h_1,h_2}||_0 + O(h^4), \end{aligned} \tag{4.10}$$

$$\begin{aligned} &||\mathbf{u} - \Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2}||_0 \\ &\quad = \frac{1}{3}||\Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2} - \Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0 + O(h^4). \end{aligned} \tag{4.11}$$

In addition, if there exist positive constants C_1, C_2 *and* ϵ_1, $\epsilon_2 \in (0,1)$ *such that*

$$||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0 \geq C_1 h^{4-\epsilon_1}, \tag{4.12}$$

$$||\mathbf{u} - \Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2}||_0 \geq C_2 h^{4-\epsilon_2}, \tag{4.13}$$

then we have

$$\lim_{h\to 0} \frac{3||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0}{||P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - P^3_{4h_1,4h_2}p_{h_1,h_2}||_0} = 1, \tag{4.14}$$

$$\lim_{h\to 0} \frac{3||\mathbf{u} - \Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2}||_0}{||\Pi^3_{2h_1,2h_2}\mathbf{u}_{h_1/2,h_2/2} - \Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0} = 1. \tag{4.15}$$

From (4.10) we see that the computable error estimator $\frac{1}{3}||P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - P^3_{4h_1,4h_2}p_{h_1,h_2}||_0$ is the principal part of the error $||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0$, and can be used as an a posteriori error indicator to assess the accuracy of the pressure error $||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0$. Meanwhile, the condition (4.12) seems to be a reasonable assumption because $O(h^2)$ is the optimal convergence rate of $||p - P^3_{2h_1,2h_2}p_{h_1/2,h_2/2}||_0$ according to Theorem 4.4. Also, it can be further seen from (4.14) that the a posteriori error estimator $\frac{1}{3}||P^3_{2h_1,2h_2}p_{h_1/2,h_2/2} - P^3_{4h_1,4h_2}p_{h_1,h_2}||_0$ is quite reliable. The same comments are also valid for (4.11), (4.13), and (4.15).

4.2. The global Richardson extrapolation in one direction. The approach introduced in the last subsection has a limitation in that it requires a global and uniform refinement in both the x- and y-directions, and hence, it wastes computing time and memory. To overcome this shortcoming, here we propose an extrapolation method of a partial refinement [18], in which the meshes are refined just in either the x- or y-direction. Thus, this method is more efficient and is also more suitable for parallel computations.

THEOREM 4.6. *Under the conditions of Theorem* 4.4 *we have*

$$p_{h_1,h_2} - P^0_{h_1,h_2}p = h_1^2\xi^1_{h_1,h_2} + h_2^2\xi^2_{h_1,h_2} + \hat{r}_{h_1,h_2}, \qquad ||\hat{r}_{h_1,h_2}||_0 \le Ch^4,$$

$$\mathbf{u}_{h_1,h_2} - \Pi^0_{h_1,h_2}\mathbf{u} = h_1^2\eta^1_{h_1,h_2} + h_2^2\eta^2_{h_1,h_2} + \hat{\mathbf{r}}_{h_1,h_2}, \qquad ||\hat{\mathbf{r}}_{h_1,h_2}||_0 \le Ch^4,$$

where $(\xi^1_{h_1,h_2}, \eta^1_{h_1,h_2})$, $(\xi^2_{h_1,h_2}, \eta^2_{h_1,h_2}) \in W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$.

Proof. Let (ξ^1, η^1), $(\xi^2, \eta^2) \in W \times \mathbf{V}_0$ and $(\xi^1_{h_1,h_2}, \eta^1_{h_1,h_2})$, $(\xi^2_{h_1,h_2}, \eta^2_{h_1,h_2}) \in W_{h_1,h_2} \times \mathbf{V}_{0,h_1,h_2}$ be the exact solutions and the mixed finite element solutions, respectively, of the following two auxiliary variational problems:

$$\begin{aligned} &(\alpha\eta^1, \mathbf{v}) - (\xi^1, \nabla\cdot\mathbf{v}) = L_1(\mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_0, \\ &(\nabla\cdot\eta^1, w) + (c\xi^1, w) = L_3(w), \quad w \in W, \end{aligned} \tag{4.16}$$

and

$$\begin{aligned} &(\alpha\eta^2, \mathbf{v}) - (\xi^2, \nabla\cdot\mathbf{v}) = L_2(\mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_0, \\ &(\nabla\cdot\eta^2, w) + (c\xi^2, w) = L_4(w), \quad w \in W, \end{aligned} \tag{4.17}$$

where

$$\begin{aligned} L_1(\mathbf{v}) &= -\frac{1}{3}\int_\Omega [\alpha_{11}(u_1)_{xx} + \alpha_{12}(u_2)_{xx}]v_1 d\Omega \\ &\quad + \frac{1}{3}\int_\Omega [(\alpha_{22})_x(u_2)_x - \alpha_{21}(u_1)_{xx}]v_2 d\Omega, \\ L_2(\mathbf{v}) &= \frac{1}{3}\int_\Omega [(\alpha_{11})_y(u_1)_y - \alpha_{12}(u_2)_{yy}]v_1 d\Omega \\ &\quad - \frac{1}{3}\int_\Omega [\alpha_{22}(u_2)_{yy} + \alpha_{21}(u_1)_{yy}]v_2 d\Omega, \\ L_3(w) &= \frac{1}{3}\int_\Omega c_x p_x w d\Omega, \quad L_4(w) = \frac{1}{3}\int_\Omega c_y p_y w d\Omega. \end{aligned}$$

Then, it follows from (4.1) and Theorems 3.1 and 3.4 that

$$\begin{aligned}&(\alpha\theta_{h_1,h_2},\mathbf{v})-(\rho_{h_1,h_2},\nabla\cdot\mathbf{v})=h_1^2L_1(\mathbf{v})+h_2^2L_2(\mathbf{v})+O(h^4)||\mathbf{v}||_0,\quad \mathbf{v}\in\mathbf{V}_{0,h_1,h_2},\\&(\nabla\cdot\theta_{h_1,h_2},w)+(c\rho_{h_1,h_2},w)=h_1^2L_3(w)+h_2^2L_4(w)+O(h^4)||w||_0,\quad w\in W_{h_1,h_2}.\end{aligned}\tag{4.18}$$

Set

$$\hat{\theta}_{h_1,h_2}:=\theta_{h_1,h_2}-h_1^2\eta^1_{h_1,h_2}-h_2^2\eta^2_{h_1,h_2},\quad \hat{\rho}_{h_1,h_2}:=\rho_{h_1,h_2}-h_1^2\xi^1_{h_1,h_2}-h_2^2\xi^2_{h_1,h_2}.$$

Thus, we know from (4.16)-(4.18) that

$$\begin{aligned}&(\alpha\hat{\theta}_{h_1,h_2},\mathbf{v})-(\hat{\rho}_{h,k},\nabla\cdot\mathbf{v})=O(h^4)||\mathbf{v}||_0,\quad \mathbf{v}\in\mathbf{V}_{0,h,k},\\&(\nabla\cdot\hat{\theta}_{h,k},w)+(c\hat{\rho}_{h,k},w)=O(h^4)||w||_0,\qquad w\in W_{h,k}.\end{aligned}\tag{4.19}$$

Following the steps for the estimates of $\theta^*_{h_1,h_2}$ and $\rho^*_{h_1,h_2}$ in the proof of Theorem 4.2 yields by means of (4.19) that

$$||\hat{\rho}_{h_1,h_2}||_0\le Ch^4\quad\text{and}\quad||\hat{\theta}_{h_1,h_2}||_0\le Ch^4.\qquad\square$$

By the same argument as that for Theorem 4.4, we can establish the following result.

THEOREM 4.7. *We have under the conditions of Theorem* 4.6 *that*

$$\begin{aligned}&P^3_{4h_1,4h_2}p_{h_1,h_2}-p=h_1^2\xi^1+h_2^2\xi^2+\tilde{r}_{h_1,h_2},\qquad ||\tilde{r}_{h_1,h_2}||_0\le Ch^4,\\&\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}-\mathbf{u}=h_1^2\eta^1+h_2^2\eta^2+\tilde{\mathbf{r}}_{h_1,h_2},\qquad ||\tilde{\mathbf{r}}_{h_1,h_2}||_0\le Ch^4,\end{aligned}$$

where (ξ^1,η^1), $(\xi^2,\eta^2)\in W\times\mathbf{V}_0$.

From Theorem 4.7 one can obtain the following unidirectional Richardson extrapolation results under the L^2-norm:

$$\begin{aligned}&\frac{4(\Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2}+\Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2})-5\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}}{3}=\mathbf{u}+O(h^4),\\&\frac{4(P^3_{2h_1,4h_2}p_{h_1/2,h_2}+P^3_{4h_1,2h_2}p_{h_1,h_2/2})-5P^3_{4h_1,4h_2}p_{h_1,h_2}}{3}=p+O(h^4),\end{aligned}\tag{4.20}$$

where $(p_{h_1/2,h_2},\mathbf{u}_{h_1/2,h_2})$, $(p_{h_1,h_2/2},\mathbf{u}_{h_1,h_2/2})$, and $(p_{h_1,h_2},\mathbf{u}_{h_1,h_2})$ are the mixed finite element solutions corresponding to the meshes $\mathcal{T}_{h_1/2,h_2}$, $\mathcal{T}_{h_1,h_2/2}$, and $\mathcal{T}_{h_1,h_2}$, respectively, and $\mathcal{T}_{h_1/2,h_2}$ as well as $\mathcal{T}_{h_1,h_2/2}$ are gained by subdividing each element of $\mathcal{T}_{h_1,h_2}$ into two small congruent rectangles in the x-direction and y-direction, respectively.

In order to fulfill the Richardson extrapolation, in (4.8) and (4.9) we have to compute $p_{h_1/2,h_2/2}$, $\mathbf{u}_{h_1/2,h_2/2}$, and p_{h_1,h_2}, $\mathbf{u}_{h_1,h_2}$ in advance. The numbers of the unknown variables for $p_{h_1/2,h_2/2}$ (or $\mathbf{u}_{h_1/2,h_2/2}$), and p_{h_1,h_2} (or $\mathbf{u}_{h_1,h_2}$) are $O(4h^{-2})$ and $O(h^{-2})$, respectively. However, in (4.20) we need only compute $p_{h_1/2,h_2}$, $p_{h_1,h_2/2}$, p_{h_1,h_2} and $\mathbf{u}_{h_1/2,h_2}$, $\mathbf{u}_{h_1,h_2/2}$, $\mathbf{u}_{h_1,h_2}$. The numbers of the unknown variables for $p_{h_1/2,h_2}$ (or $\mathbf{u}_{h_1/2,h_2}$), $p_{h_1,h_2/2}$ (or $\mathbf{u}_{h_1,h_2/2}$), and p_{h_1,h_2} (or $\mathbf{u}_{h_1,h_2}$) are $O(2h^{-2})$, $O(2h^{-2})$, and

$O(h^{-2})$, respectively. Generally speaking, the scale of computing $p_{h_1/2,h_2}$ (or $p_{h_1,h_2/2}$) and $\mathbf{u}_{h_1/2,h_2}$ (or $\mathbf{u}_{h_1,h_2/2}$) is smaller than that of computing $p_{h_1/2,h_2/2}$ and $\mathbf{u}_{h_1/2,h_2/2}$. Hence, the computation and storage can be saved. In addition, compared with (4.8) and (4.9), (4.20) is easy to compute in a parallel manner, so that this method is more efficient than the normal method described in (4.8) and (4.9), especially for three-dimensional problems.

Similarly to (4.8) and (4.9), we can also construct a posteriori error estimators by virtue of (4.20).

THEOREM 4.8. *Under the conditions of Theorem* 4.7 *we have*

$$\begin{aligned}
&||p - P^3_{2h_1,4h_2}p_{h_1/2,h_2}||_0\\
&\quad = \frac{1}{3}||P^3_{2h_1,4h_2}p_{h_1/2,h_2} + 4P^3_{4h_1,2h_2}p_{h_1,h_2/2} - 5P^3_{4h_1,4h_2}p_{h_1,h_2}||_0 + O(h^4),\\
&\qquad ||\mathbf{u} - \Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2}||_0\\
&\quad = \frac{1}{3}||\Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2} + 4\Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2} - 5\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0 + O(h^4),\\
&\qquad ||p - P^3_{4h_1,2h_2}p_{h_1,h_2/2}||_0\\
&\quad = \frac{1}{3}||4P^3_{2h_1,4h_2}p_{h_1/2,h_2} + P^3_{4h_1,2h_2}p_{h_1,h_2/2} - 5P^3_{4h_1,4h_2}p_{h_1,h_2}||_0 + O(h^4),\\
&\qquad ||\mathbf{u} - \Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2}||_0\\
&\quad = \frac{1}{3}||4\Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2} + \Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2} - 5\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0 + O(h^4).
\end{aligned}$$

Moreover, if there exist positive constants C_1, C_2, C_3, C_4 *and* ϵ_1, ϵ_2, ϵ_3, $\epsilon_4 \in (0,1)$ *such that*

$$||p - P^3_{2h_1,4h_2}p_{h_1/2,h_2}||_0 \geq C_1h^{4-\epsilon_1},$$

$$||\mathbf{u} - \Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2}||_0 \geq C_2h^{4-\epsilon_2},$$

$$||p - P^3_{4h_1,2h_2}p_{h_1,h_2/2}||_0 \geq C_3h^{4-\epsilon_3},$$

$$||\mathbf{u} - \Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2}||_0 \geq C_4h^{4-\epsilon_4},$$

then we have

$$\lim_{h\to 0}\frac{3||p - P^3_{2h_1,4h_2}p_{h_1/2,h_2}||_0}{||P^3_{2h_1,4h_2}p_{h_1/2,h_2} + 4P^3_{4h_1,2h_2}p_{h_1,h_2/2} - 5P^3_{4h_1,4h_2}p_{h_1,h_2}||_0} = 1,$$

$$\lim_{h\to 0}\frac{3||\mathbf{u} - \Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2}||_0}{||\Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2} + 4\Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2} - 5\Pi_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0} = 1,$$

$$\lim_{h\to 0}\frac{3||p - P^3_{4h_1,2h_2}p_{h_1,h_2/2}||_0}{||4P^3_{2h_1,4h_2}p_{h_1/2,h_2} + P^3_{4h_1,2h_2}p_{h_1,h_2/2} - 5P^3_{4h_1,4h_2}p_{h_1,h_2}||_0} = 1,$$

$$\lim_{h\to 0}\frac{3||\mathbf{u} - \Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2}||_0}{||4\Pi^3_{2h_1,4h_2}\mathbf{u}_{h_1/2,h_2} + \Pi^3_{4h_1,2h_2}\mathbf{u}_{h_1,h_2/2} - 5\Pi^3_{4h_1,4h_2}\mathbf{u}_{h_1,h_2}||_0} = 1.$$

Acknowledgments. The authors would like to thank Professor Michal Křížek for his many helpful suggestions during the preparation of the paper, Professor Raytcho Lazarov for the reference on his early work [10], and the anonymous referees' comments and suggestions, which improved the presentation.

REFERENCES

[1] M. Ainsworth, *A posteriori error estimation for non-conforming quadrilateral finite elements*, Int. J. Numer. Anal. Model., 2 (2005), pp. 1–18.

[2] H. Blum, Q. Lin, and R. Rannacher, *Asymptotic error expansion and Richardson extrapolation for linear finite elements*, Numer. Math., 49 (1986), pp. 11–38.

[3] J. Brandts and Y. Chen, *Superconvergence of least-squares mixed finite elements*, Int. J. Numer. Anal. Model., 3 (2006), pp. 303–310.

[4] F. Brezzi, J. Douglas, and L. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[6] H. Brunner, Y. Lin, and S. Zhang, *Higher accuracy methods for second-kind Volterra integral equations based on asymptotic expansions of iterated Galerkin methods*, J. Integral Equations Appl., 10 (1998), pp. 375–396.

[7] C. Carstensen, *Reliable and efficient averaging techniques as universal tool for a posteriori finite element error control on unstructured grids*, Int. J. Numer. Anal. Model., 3 (2006), pp. 333–347.

[8] Y. Chen and W. Liu, *Error estimates and superconvergence of mixed finite element for quadratic optimal control*, Int. J. Numer. Anal. Model., 3 (2006), pp. 311–321.

[9] C. Chen and Y. Huang, *Higher Accuracy Theory of FEM*, Hunan Science Press, Changsha, China, 1995.

[10] H. Chen, R. E. Ewing, and R. Lazarov, *Asymptotic Error Expansion for the Lowest Order Raviart–Thomas Rectangular Mixed Finite Elements*, Technical report ISC-97-01, Institute for Scientific Computation, Texas A & M University, College Station, TX, 1997.

[11] J. Douglas and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[12] J. Douglas and J. Wang, *Superconvergence of mixed finite element spaces on rectangular domains*, Calcolo, 26 (1989), pp. 121–134.

[13] R. E. Ewing, Y. Lin, T. Sun, J. Wang, and S. Zhang, *Sharp L^2-error estimates and superconvergence of mixed finite element methods for non-Fickian flows in porous media*, SIAM J. Numer. Anal., 40 (2002), pp. 1538–1560.

[14] R. E. Ewing, Y. Lin, J. Wang, and S. Zhang, *L^∞-error estimates and superconvergence in maximum norm of mixed finite element methods for non-Fickian flows in porous media*, Int. J. Numer. Anal. Model., 2 (2005), pp. 301–328.

[15] P. Helfrich, *Asymptotic expansion for the finite element approximations of parabolic problems*, Bonner Math. Schriften, 158 (1983), pp. 11–30.

[16] M. Křížek, *Superconvergence phenomena on three-dimensional meshes*, Int. J. Numer. Anal. Model., 2 (2005), pp. 43–56.

[17] Q. Lin, I. H. Sloan, and R. Xie, *Extrapolation of the iterated-collocation method for integral equations of the second kind*, SIAM J. Numer. Anal., 27 (1990), pp. 1535–1541.

[18] Q. Lin and N. Yan, *The Construction and Analysis of High Efficiency Finite Element Methods*, Hebei University Press, Baoding, China, 1996.

[19] Q. Lin, S. Zhang, and N. Yan, *Asymptotic error expansion and defect correction for Sobolev and viscoelasticity type equations*, J. Comput. Math., 16 (1998), pp. 57–62.

[20] Q. Lin, S. Zhang, and N. Yan, *High accuracy analysis for integrodifferential equations*, Acta Math. Appl. Sinica, 14 (1998), pp. 202–211.

[21] Q. Lin, S. Zhang, and N. Yan, *Methods for improving approximate accuracy for hyperbolic integro-differential equations*, Systems Sci. Math. Sci., 10 (1997), pp. 282–288.

[22] Q. Lin, S. Zhang, and N. Yan, *Extrapolation and defect correction for diffusion equations with boundary integral conditions*, Acta Math. Sci. (English Ed.), 17 (1997), pp. 405–412.

[23] Q. Lin, S. Zhang, and N. Yan, *An acceleration method for integral equations by using interpolation post-processing*, Adv. Comput. Math., 9 (1998), pp. 117–129.

[24] T. Lin, Y. Lin, M. Rao, and S. Zhang, *Petrov–Galerkin methods for linear Volterra integro-differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 937–963.

[25] H. Liu and N. Yan, *Global superconvergence for optimal control problems governed by Stokes equations*, Int. J. Numer. Anal. Model., 3 (2006), pp. 283–302.
[26] G. Marchuk and V. Shaidurov, *Difference Methods and Their Extrapolation*, Springer-Verlag, New York, 1983.
[27] P. Ming, Z.-C. Shi, and Y. Xu, *Superconvergence studies of quadrilateral nonconforming rotated Q_1 elements*, Int. J. Numer. Anal. Model., 3 (2006), pp. 322–332.
[28] J. Wang, *Superconvergence and extrapolation for mixed finite element methods on rectangular domains*, Math. Comp., 56 (1991), pp. 477–503.
[29] J. Wang, *Asymptotic expansions and L^∞-error estimates for mixed finite element methods for second order elliptic problems*, Numer. Math., 55 (1989), pp. 401–430.
[30] J. Wang and X. Ye, *A superconvergent finite element scheme for the Reissner-Mindlin plate by projection methods*, Int. J. Numer. Anal. Model., 1 (2004), pp. 99–110.
[31] N. Yan and K. Li, *An extrapolation method for BEM*, J. Comput. Math., 2 (1989), pp. 217–224.
[32] S. Zhang, T. Lin, Y. Lin, and M. Rao, *Extrapolation and a-posteriori error estimators of Petrov-Galerkin methods for non-linear Volterra integro-differential equations*, J. Comput. Math., 19 (2001), pp. 407–422.
[33] A. Zhou, C. B. Liem, T. M. Shih, and T. Lü, *A multi-parameter splitting extrapolation and a parallel algorithm*, Systems Sci. Math. Sci., 10 (1997), pp. 253–260.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1150–1171

SUBGRID UPSCALING AND MIXED MULTISCALE FINITE ELEMENTS*

TODD ARBOGAST† AND KIRSTEN J. BOYD‡

Abstract. Second order elliptic problems in divergence form with a highly varying leading order coefficient on the scale ϵ can be approximated on coarse meshes of spacing $H \gg \epsilon$ only if one uses special techniques. The mixed variational multiscale method, also called subgrid upscaling, can be used, and this method is extended to allow oversampling of the local subgrid problems. The method is shown to be equivalent to the multiscale finite element method when one uses the lowest order Raviart–Thomas spaces and provided that there are no fine scale components in the source function f. In the periodic setting, a multiscale error analysis based on homogenization theory of the more general subgrid upscaling method shows that the error is $O(\epsilon+H^m+\sqrt{\epsilon/H})$, where $m = 1$. Moreover, $m = 2$ if one uses the second order Brezzi–Douglas–Marini or Brezzi–Douglas–Durán–Fortin spaces and no oversampling. The error bounding constant depends only on the H^{m-1}-norm of f and so is independent of small scales when $m = 1$. When oversampling is not used, a superconvergence result for the pressure approximation is shown.

Key words. mixed method, multiscale finite element, subgrid upscaling, variational multiscale

AMS subject classifications. 65N15, 65N30, 35J20

DOI. 10.1137/050631811

1. Introduction.

Many physical problems can be modeled by a second order elliptic partial differential equation in space. In many cases, the coefficients of the equation are highly heterogeneous, which induces fine scale variability in the solution. Thus the difficulty in approximating the solution on a coarse finite element mesh $\mathcal{T}_H$ is that the solution is not fully resolved on this scale. Traditional finite element analysis fails, and we require some multiscale approximation techniques.

Babuška and Osborn [10, 9] proposed using special finite elements to approximate the solution. Hughes et al. [23, 24] (see also [13]) developed a more formal framework, which they called the variational multiscale method. A mixed variant, described as *subgrid upscaling,* was developed by Arbogast et al. [7, 3, 4, 6, 5]. Hou and Wu [21] and Hou, Wu, and Cai [22] took a more direct approach and simply proposed finding a special finite element basis by solving the problem locally. They called this approach the multiscale finite element method. A mixed form was developed later by Chen and Hou [17].

To be more precise, consider a connected, convex polygonal domain $\Omega \subseteq \mathbb{R}^d$, where $d = 2$ or 3, and a second order, uniformly positive-definite symmetric tensor a, so that both a and a^{-1} are uniformly elliptic and uniformly bounded. Suppose we are also given vectors $\mathbf{b}$ and $\mathbf{v}_g$. For a set S, let ν^S be the outward unit normal to ∂S, and define the function g on $\partial\Omega$ by $g = \mathbf{v}_g \cdot \nu$, where $\nu = \nu^\Omega$. The problem under

*Received by the editors May 18, 2005; accepted for publication (in revised form) February 3, 2006; published electronically June 21, 2006.
http://www.siam.org/journals/sinum/44-3/63181.html
†Department of Mathematics, University of Texas, 1 University Station C1200, Austin, TX 78712, and Institute for Computational Engineering and Sciences, University of Texas, 1 University Station C0200, Austin, TX 78712 (arbogast@ices.utexas.edu). The work of this author was supported by the U.S. National Science Foundation under grant DMS-0408489.
‡Eureka College, 300 E. College Ave., Eureka, IL 61530 (kboyd@eureka.edu).

consideration is to find $\mathbf{u}$ and p such that

$$\nabla \cdot \mathbf{u} = f \quad \text{in } \Omega, \tag{1.1}$$

$$\mathbf{u} = -a(\nabla p - \mathbf{b}) \quad \text{in } \Omega, \tag{1.2}$$

$$\mathbf{u} \cdot \nu = g \quad \text{on } \partial\Omega. \tag{1.3}$$

The above system of two first order differential equations is described as a *mixed* formulation, and it is preferable to a single second order differential equation for p because it allows one to enforce the conservation property for the flux (1.1) locally [16]. An example governed by this system is fluid flow in porous media, where the permeability (divided by fluid viscosity) a can vary by many orders of magnitude over a small spatial displacement, $\mathbf{u}$ is the Darcy velocity, p is the fluid pressure, and f models sources and sinks, i.e., wells, which themselves may be quite small scale features in the problem.

To approximate the velocity $\mathbf{u}$ and pressure p on the coarse mesh $\mathcal{T}_H$ requires meeting two competing objectives. First, the approximating spaces must be rich enough to follow the variability in the solution. While a fully fine scale approximating space fulfills this objective, it is not computationally efficient. The second objective is to somehow reduce the problem to the size and complexity of an ordinary coarse scale approximation. The natural approach is to simplify the representation of the solution on the coarse element edges in two dimensions, or faces in three dimensions.

In the variational point of view taken by Arbogast et al., the solution space is decomposed into coarse and fine scale components. This also splits the trial space, and therefore the equations, into coarse and fine scale parts. The fine scale equations are local, and thus solvable, and allow one to compute the fine scale part of the solution from the coarse scale part. The problem then reduces to solving a coarse scale problem for the coarse part of the solution. Any of the usual mixed finite element spaces can be used on the coarse scale. To obtain good approximation on the coarse element edges or faces in this context, it was found that one should use at least second order accurate velocities on the coarse scale.

The multiscale finite element approach of Hou, Wu, and Cai is based on using the lowest order Raviart–Thomas (RT0) spaces [27] on the coarse scale. One modifies the usual coarse basis to incorporate the microstructure in a by solving the system (1.1)–(1.2) locally. This produces finite elements that vary much like the solution itself. One simply solves a coarse mesh mixed finite element method using these perturbed elements. However, because the RT0 spaces are only first order accurate, they do not give good approximation on the coarse element boundaries. To alleviate this problem, Hou, Wu, and Cai propose an oversampling technique, in which they modify each local basis function by sampling the microstructure over a domain larger than its support. This induces variability in the velocity across coarse element edges or faces and improves the quality of the solution. Several interesting and important advances in the design of the mixed multiscale finite elements have been proposed by Aarnes [1] and Aarnes, Krogstad, and Lie [2].

In this paper, we obtain a connection between the two frameworks. Even though they appear very different, we show that they are in fact equivalent under mild restrictions. We first extend the subgrid upscaling approach to allow oversampling. Then the two frameworks are equivalent provided that one uses the RT0 spaces, and provided that there are no fine scale components in f. This last is a subtle point, but important in porous media applications, since wells are so small in two of their three dimensions. The variational framework picks up additional terms related to fine scale

components of the wells that are overlooked in the multiscale finite element approach, since the latter emphasizes only heterogeneity in a (unless perhaps one supplements the finite element space with special well elements).

We also show that the multiscale error analysis of Chen and Hou [17] extends to the variational multiscale framework. In this analysis, one considers $a(x)$ to be locally periodic of period ϵ; that is, the scale of the heterogeneity is well defined as ϵ. In the case considered in [17], RT0 on simplices, our results are similar and give an $O(\epsilon + H + \sqrt{\epsilon/H})$ error bound, wherein the bounding constant depends on Sobolev norms of the smooth homogenized solution but not on the solution itself. Moreover, the proof is elucidated by the application of variational upscaling ideas and results in improved error estimates with regard to f, requiring its L^2-norm to be bounded rather than its H^1-norm. When oversampling is not used, we obtain error bounds of $O(\epsilon + H^m + \sqrt{\epsilon/H})$ for RT0 on nonsimplicial elements ($m = 1$) and the second order accurate ($m \le 2$) Brezzi–Douglas–Marini (BDM1) [15] spaces in two dimensions or the Brezzi–Douglas–Durán–Fortin (BDDF1) [14] spaces in three dimensions. Furthermore, when oversampling is not used, we obtain an important superconvergence result for the pressure approximation, showing that it is $O\big((\epsilon + H + (\epsilon/H)^{1/d-\eta})(\epsilon + H^m + \sqrt{\epsilon/H})\big)$, where $\eta > 0$ if $d = 2$ and $\eta = 0$ if $d = 3$.

The outline of the paper follows. In section 2, we apply the construction in [5] to obtain equations upscaled to the coarse level. We show that the upscaling correction terms are antidiffusive and nonlocal in character. We also extend the method to allow oversampling. In section 3, we extract the multiscale finite elements that are implicit in the construction and show when the method is equivalent to that of Chen and Hou [17]. In section 4, we discuss the fundamental inf-sup lemma regarding solvability and approximability. In the next section, section 5, we state certain homogenization results that we need for section 6, in which our multiscale convergence result for the velocity is stated and proved. Due to the structure of the inf-sup lemma, the error has two components, the optimal velocity error and an error due to the use of nonconforming spaces. Finally, in section 7, we treat the pressure error and show that it is superconvergent in the multiscale setting.

We close the introduction by recasting (1.1)–(1.3) in variational form. Let

$$H(\mathrm{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$$

with inner product

$$(\mathbf{v}_1, \mathbf{v}_2)_{H(\mathrm{div};\Omega)} = (\mathbf{v}_1, \mathbf{v}_2)_{(L^2(\Omega))^d} + (\nabla \cdot \mathbf{v}_1, \nabla \cdot \mathbf{v}_2)_{L^2(\Omega)}$$

and norm $\|\mathbf{v}\|_{H(\mathrm{div};\Omega)} = (\mathbf{v}, \mathbf{v})^{1/2}_{H(\mathrm{div};\Omega)}$. We set

$$\mathbf{V} = H_0(\mathrm{div};\Omega) = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v} \cdot \nu = 0 \text{ on } \partial\Omega\}$$

and $W = L^2(\Omega)/\mathbb{R}$ with the $L^2(\Omega)$-norm, so that $\nabla \cdot \mathbf{V} = W$. We wish to find $\mathbf{u} \in \mathbf{V} + \mathbf{v}_g$ and $p \in W$ such that

$$(\nabla \cdot \mathbf{u}, w) = (f, w) \qquad \forall\, w \in W, \tag{1.4}$$

$$(\alpha\mathbf{u}, \mathbf{v}) = (p, \nabla \cdot \mathbf{v}) + (\mathbf{b}, \mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{V}, \tag{1.5}$$

where $\alpha = a^{-1}$ and we denote the $L^2(S)$ inner product by $(\cdot,\cdot)_S$ for set S and omit S from the notation when it is Ω. We assume that $a \in (L^\infty(\Omega))^{d\times d}$, $\mathbf{b} \in \big(L^2(\Omega)\big)^d$,

$f \in L^2(\Omega)$, and $\mathbf{v}_g \in H^1(\Omega)$. Provided that we have the compatibility condition

$$\int_\Omega f(x)\,dx = \int_{\partial\Omega} g(x)\,ds,$$

it follows from the standard inf-sup theory of saddle point problems [8, 12, 11, 16, 17] that (1.4)–(1.5) is indeed uniquely solvable.

2. Approximation by the variational multiscale method. Let $\mathcal{T}_H$ be a regular and quasi-uniform partition of Ω into simplices and/or bricks having maximum diameter H, satisfying the condition that the minimum angle of each E is bounded below by some positive constant independent of H. Consider the orthogonal direct sum decomposition

$$W = \bar{W} \oplus W',$$

where the coarse space is

$$\bar{W} = \{\bar{w} \in W : \bar{w} \text{ is constant on each } E \in \mathcal{T}_H\}$$

and the "subgrid" space is the orthogonal complement

$$W' = \bar{W}^\perp = \left\{ w' \in W : \int_E w'(x)\,dx = 0 \ \ \forall\, E \in \mathcal{T}_H \right\}.$$

Following [5], we can find a (nonorthogonal) direct sum decomposition of $\mathbf{V}$ into closed subspaces $\bar{\mathbf{V}}$ and $\mathbf{V}'$ such that

$$\begin{aligned}
&\mathbf{V} = \bar{\mathbf{V}} \oplus \mathbf{V}',\\
&\bar{\mathbf{V}} \subseteq \left\{\bar{\mathbf{v}} \in \mathbf{V} : \nabla\cdot\bar{\mathbf{v}} \in \bar{W}\right\},\\
&\mathbf{V}' = \left\{\mathbf{v}' \in \mathbf{V} : \nabla\cdot\mathbf{v}' \in W' \text{ and } \mathbf{v}'\cdot\nu^E = 0 \text{ on } \partial E \ \ \forall E \in \mathcal{T}_H\right\};
\end{aligned}$$

moreover, $\nabla\cdot\bar{\mathbf{V}} = \bar{W}$ and $\nabla\cdot\mathbf{V}' = W'$. Thus we can uniquely decompose the solution $(\mathbf{u}, p) \in (\mathbf{V} + \mathbf{v}_g) \times W$ of (1.4)–(1.5) as

$$\mathbf{u} = \bar{\mathbf{u}} + \mathbf{u}' + \mathbf{v}_g, \tag{2.1}$$

$$p = \bar{p} + p', \tag{2.2}$$

where $\bar{\mathbf{u}} \in \bar{\mathbf{V}}$, $\mathbf{u}' \in \mathbf{V}'$, $\bar{p} \in \bar{W}$, and $p' \in W'$.

2.1. Subgrid closure operators. By using the above decompositions and restricting the test functions in (1.4)–(1.5) to $(\mathbf{v}', w') \in \mathbf{V}' \times W'$, we obtain the *subgrid equation*

$$(\nabla\cdot\mathbf{u}', w') = (f - \nabla\cdot\mathbf{v}_g, w') \qquad \forall\, w' \in W', \tag{2.3}$$

$$(\alpha(\bar{\mathbf{u}} + \mathbf{u}'), \mathbf{v}') = (p', \nabla\cdot\mathbf{v}') + (\mathbf{b} - \alpha\mathbf{v}_g, \mathbf{v}') \quad \forall\, \mathbf{v}' \in \mathbf{V}', \tag{2.4}$$

where certain terms have vanished due to the orthogonality of $\bar{W}$ and W' and the property that $\nabla\cdot\bar{\mathbf{V}} = \bar{W}$. Note that for our problem, $\bar{p}$ does not appear in the above equation (see [5] for handling the general case).

We now define the *subgrid closure operators* mapping each $\bar{\mathbf{u}} \in \bar{\mathbf{V}}$ to some $\mathbf{u}' \in \mathbf{V}'$ and $p' \in W'$. Each is an affine operator consisting of a linear and a constant part depending on $\bar{\mathbf{u}}$, the coarse part of $\mathbf{u}$, so we write

$$\mathbf{u}' = \mathbf{u}'(\bar{\mathbf{u}}) = \hat{\mathbf{u}}'(\bar{\mathbf{u}}) + \tilde{\mathbf{u}}', \tag{2.5}$$

$$p' = p'(\bar{\mathbf{u}}) = \hat{p}'(\bar{\mathbf{u}}) + \tilde{p}'. \tag{2.6}$$

More generally, for each $\mathbf{v} \in H(\mathrm{div};\Omega)$, by (2.3)–(2.4), $(\hat{\mathbf{u}}'(\mathbf{v}), \hat{p}'(\mathbf{v})) \in \mathbf{V}' \times W'$ is defined by

$$(\nabla \cdot \hat{\mathbf{u}}'(\mathbf{v}), w') = 0 \qquad \forall\, w' \in W', \tag{2.7}$$

$$(\alpha\,(\mathbf{v} + \hat{\mathbf{u}}'(\mathbf{v}))\,, \mathbf{v}') = (\hat{p}'(\mathbf{v}), \nabla \cdot \mathbf{v}') \quad \forall\, \mathbf{v}' \in \mathbf{V}', \tag{2.8}$$

and $(\tilde{\mathbf{u}}', \tilde{p}') \in \mathbf{V}' \times W'$ is defined by

$$(\nabla \cdot \tilde{\mathbf{u}}', w') = (f - \nabla \cdot \mathbf{v}_g, w') \qquad \forall\, w' \in W', \tag{2.9}$$

$$(\alpha\tilde{\mathbf{u}}', \mathbf{v}') = (\tilde{p}', \nabla \cdot \mathbf{v}') + (\mathbf{b} - \alpha\mathbf{v}_g, \mathbf{v}') \quad \forall\, \mathbf{v}' \in \mathbf{V}'. \tag{2.10}$$

These equations are well-posed on each $E \in \mathcal{T}_H$ [5].

For future reference, we note that on each $E \in \mathcal{T}_H$,

$$-a\nabla\hat{p}'(\mathbf{v}) = \mathbf{v} + \hat{\mathbf{u}}'(\mathbf{v}), \tag{2.11}$$

$$-a\nabla\tilde{p}' = \tilde{\mathbf{u}}' - a\mathbf{b} + \mathbf{v}_g, \tag{2.12}$$

because $\mathbf{V}'|_E = H_0(\mathrm{div};E)$ is the full space. Moreover, $W'|_E = L^2(E)/\mathbb{R}$, so

$$\nabla \cdot \hat{\mathbf{u}}'(\mathbf{v}) = 0. \tag{2.13}$$

2.2. The upscaled equation. We now define a vector space $\hat{\mathbf{V}} \subseteq \mathbf{V}$ by

$$\hat{\mathbf{V}} = \left\{\hat{\mathbf{v}} \in \mathbf{V} : \hat{\mathbf{v}} = \bar{\mathbf{v}} + \hat{\mathbf{u}}'(\bar{\mathbf{v}}) \text{ for some } \bar{\mathbf{v}} \in \bar{\mathbf{V}}\right\},$$

restrict the test functions in (1.4)–(1.5) to be in $\hat{\mathbf{V}} \times \bar{W}$, use the various decompositions, and introduce the notation

$$\hat{f} = f - \nabla \cdot \mathbf{v}_g \quad \text{and} \quad \hat{\mathbf{b}} = \mathbf{b} - \alpha(\mathbf{v}_g + \tilde{\mathbf{u}}').$$

Thus we rewrite (1.4)–(1.5) in *upscaled form* as the problem of finding $(\hat{\mathbf{u}}, \bar{p}) \in \hat{\mathbf{V}} \times \bar{W}$ such that

$$(\nabla \cdot \hat{\mathbf{u}}, \bar{w}) = (\hat{f}, \bar{w}) \qquad \forall\, \bar{w} \in \bar{W}, \tag{2.14}$$

$$(\alpha\hat{\mathbf{u}}, \hat{\mathbf{v}}) = (\bar{p}, \nabla \cdot \hat{\mathbf{v}}) + (\hat{\mathbf{b}}, \hat{\mathbf{v}}) \quad \forall\, \hat{\mathbf{v}} \in \hat{\mathbf{V}}, \tag{2.15}$$

where now

$$\mathbf{u} = \bar{\mathbf{u}} + \hat{\mathbf{u}}'(\bar{\mathbf{u}}) + \tilde{\mathbf{u}}' + \mathbf{v}_g = \hat{\mathbf{u}} + \tilde{\mathbf{u}}' + \mathbf{v}_g. \tag{2.16}$$

By [5, Theorem 4.6], the above problem has a unique solution.

2.3. Character of the upscaled operator. With $\mathbf{v}' = \hat{\mathbf{u}}'(\bar{\mathbf{u}})$ in (2.8), we note that

$$(\alpha\hat{\mathbf{u}}'(\bar{\mathbf{u}}), \bar{\mathbf{v}}) = -(\alpha\hat{\mathbf{u}}'(\bar{\mathbf{u}}), \hat{\mathbf{u}}'(\bar{\mathbf{v}})),$$

so (1.5) with $\mathbf{v} = \bar{\mathbf{v}} \in \bar{\mathbf{V}}$ enables us to rewrite the upscaled equation (2.15) as

$$(\alpha\bar{\mathbf{u}}, \bar{\mathbf{v}}) - (\alpha\hat{\mathbf{u}}'(\bar{\mathbf{u}}), \hat{\mathbf{u}}'(\bar{\mathbf{v}})) = (\bar{p}, \nabla \cdot \bar{\mathbf{v}}) + (\hat{\mathbf{b}}, \bar{\mathbf{v}}) \quad \forall\, \bar{\mathbf{v}} \in \bar{\mathbf{V}}. \tag{2.17}$$

Thus the second term on the left-hand side, the primary subscale correction, is purely antidiffusive on the coarse scale, as we should expect. Moreover, there is an affine correction term related to subscales of $\mathbf{b}$, f, and $\mathbf{v}_g$ through $\tilde{\mathbf{u}}'$.

Next let $G_x(y)$ be the Green's function on a coarse element E, defined by

$$\begin{aligned} -\nabla \cdot a\nabla G_x &= \delta_x - 1/|E| \quad \text{in } E, \\ -a\nabla G_x \cdot \nu^E &= 0 \qquad\qquad\quad \text{on } \partial E, \end{aligned}$$

where δ_x is the Dirac mass at $x \in E$ and the average of G_x vanishes, and where vertical bars around a set in $\mathbb{R}^d$ denotes d-dimensional or $(d-1)$-dimensional Lebesgue measure, as appropriate. Then on E,

$$\begin{aligned} p(x) &= (-\nabla \cdot a\nabla G_x + 1/|E|, p)_E \\ &= (a\nabla G_x, \nabla p)_E + \bar{p} \\ &= -(\nabla G_x, \mathbf{u} - a\mathbf{b})_E + \bar{p} \\ &= -(\nabla G_x, \bar{\mathbf{u}} - a\mathbf{b})_E - (\nabla G_x, \mathbf{u}' + \mathbf{v}_g)_E + \bar{p} \\ &= -(\nabla G_x, \bar{\mathbf{u}} - a\mathbf{b})_E + (G_x, \nabla \cdot (\mathbf{u}' + \mathbf{v}_g))_E - (G_x, \mathbf{v}_g \cdot \nu^E)_{\partial E} + \bar{p} \\ &= -(\nabla G_x, \bar{\mathbf{u}} - a\mathbf{b})_E + (G_x, f')_E - (G_x, \mathbf{v}_g \cdot \nu^E)_{\partial E} + \bar{p}, \end{aligned}$$

where f' is defined by the decomposition $f = \bar{f} + f' \in \bar{W} \oplus W'$ and we use that $G_x \in W'$ to replace $(G_x, \nabla \cdot (\mathbf{u}' + \mathbf{v}_g))_E$ by $(G_x, \nabla \cdot \mathbf{u})_E = (G_x, f')$. Now

$$\alpha(x)\mathbf{u}(x) - \mathbf{b} = -\nabla p = (\nabla_x \nabla_y G_x, \bar{\mathbf{u}} - a\mathbf{b})_E - (\nabla_x G_x, f')_E + (\nabla_x G_x, \mathbf{v}_g \cdot \nu^E)_{\partial E},$$

so the diffusive and $\mathbf{b}$ terms of (1.5), tested on the coarse scale, are

$$\begin{aligned} (\alpha\mathbf{u} - \mathbf{b}, \bar{\mathbf{v}})_E = &\int_E \int_E \bar{\mathbf{u}}(y) \cdot \nabla_x \nabla_y G_x(x,y) \cdot \bar{\mathbf{v}}(x)\, dy\, dx \\ &- \int_E \int_E \mathbf{b}(y) \cdot a(y) \nabla_x \nabla_y G_x(x,y) \cdot \bar{\mathbf{v}}(x)\, dy\, dx \\ &- \int_E \int_E f'(y) \nabla_x G_x(x,y) \cdot \bar{\mathbf{v}}(x)\, dy\, dx \\ &+ \int_E \int_{\partial E} \mathbf{v}_g \cdot \nu^E(y) \nabla_x G_x(x,y) \cdot \bar{\mathbf{v}}(x)\, ds(y)\, dx, \end{aligned} \tag{2.18}$$

so the upscaled inverse permeability tensor is a nonlocal operator (confined to E) related to $a(y)\nabla_x \nabla_y G_x(y)$.

2.4. Oversampling. For each element $E \in \mathcal{T}_H$, choose some larger set $E_* \supseteq E$ such that $E_* \subseteq \Omega$, E_* is the same shape as E (i.e., a simplex or brick, again such that the minimum angle is bounded below by some positive constant independent of H and E), and, for some $C > 0$ independent of H and E, $\operatorname{diam}(E_*) \leq C \operatorname{diam}(E)$. Locally on each E_*, recalling the definition of W' and properties of $\mathbf{V}'$, we define function spaces $W'_*(E_*) = L^2(E_*)/\mathbb{R}$ and

$$\mathbf{V}'_*(E_*) = \left\{ \mathbf{v}'_* \in \mathbf{V} : \nabla \cdot \mathbf{v}'_* \in W'_*(E_*) \text{ and } \mathbf{v}'_* \cdot \nu^{E_*} = 0 \text{ on } \partial E_* \right\}. \tag{2.19}$$

By analogy to (2.7)–(2.8), we now define the linear part of the *oversampled* subgrid closure operators mapping any $\mathbf{v} \in \mathbf{V}$ to some $(\hat{\mathbf{u}}'_*(\mathbf{v}), \hat{p}'_*(\mathbf{v})) \in \mathbf{V}'_*(E_*) \times W'_*(E_*)$ defined by

$$(\nabla \cdot \hat{\mathbf{u}}'_*(\mathbf{v}), w'_*)_{E_*} = 0 \qquad \forall\, w'_* \in W'_*(E_*), \tag{2.20}$$

$$\left(\alpha\left(\mathbf{v} + \hat{\mathbf{u}}'_*(\mathbf{v})\right), \mathbf{v}'_*\right)_{E_*} = (\hat{p}'_*(\mathbf{v}), \nabla \cdot \mathbf{v}'_*)_{E_*} \quad \forall\, \mathbf{v}'_* \in \mathbf{V}'_*(E_*). \tag{2.21}$$

Note that if $E_* = E$, then the operators $\hat{\mathbf{u}}'_*(\cdot)$ and $\hat{\mathbf{u}}'(\cdot)$ coincide. We also define the *oversampled* constant parts of the subgrid closure operators corresponding to (2.9)–(2.10) as $(\tilde{\mathbf{u}}'_*, \tilde{p}'_*) \in \mathbf{V}'_*(E_*) \times W'_*(E_*)$ defined by

$$(\nabla \cdot \tilde{\mathbf{u}}'_*, w'_*)_{E_*} = (f - \nabla \cdot \mathbf{v}_g, w'_*)_{E_*} \qquad \forall\, w'_* \in W'_*(E_*), \tag{2.22}$$

$$(\alpha \tilde{\mathbf{u}}'_*, \mathbf{v}'_*)_{E_*} = (\tilde{p}'_*, \nabla \cdot \mathbf{v}'_*)_{E_*} + (\mathbf{b} - \alpha \mathbf{v}_g, \mathbf{v}'_*)_{E_*} \quad \forall\, \mathbf{v}'_* \in \mathbf{V}'_*(E_*). \tag{2.23}$$

Usually we consider these quantities only locally on E, so we need not concern ourselves with the overlap of the E_*'s. Also note that $\nabla \cdot \hat{\mathbf{u}}'_*(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathbf{V}$.

2.5. Discretization. In practice, we must approximate the solution to the subgrid problems (2.20)–(2.21) and (2.22)–(2.23). Since these problems are small (i.e., localized to E_*), we assume that we can fully resolve the fine scales in these problems on a fine subgrid mesh and thereby obtain a sufficiently accurate approximation (see also [5]). Thus, we will discuss only approximation of the coarse space in this paper, and we assume that the subgrid is solved exactly.

Let $\bar{\mathbf{V}}_H \times \bar{W}_H \subseteq \mathbf{V} \times W$ be the lowest order Raviart–Thomas (RT0) [27] space or the lowest order Brezzi–Douglas–Marini (BDM1) [15] space in two dimensions or the Brezzi–Douglas–Durán–Fortin (BDDF1) [14] space in three dimensions. In each case, the pressure approximation space is the space of piecewise constants, so we have $\bar{W}_H = \bar{W}$. Let $\mathcal{E}_E$ be the analytic extension operator from E to E_*, and define the function space

$$\hat{\mathbf{V}}_{H,*} = \left\{ \hat{\mathbf{v}}_{H,*} : \hat{\mathbf{v}}_{H,*} = \bar{\mathbf{v}}_H + \sum_{E \in \mathcal{T}_H} \hat{\mathbf{u}}'_*(\mathcal{E}_E \bar{\mathbf{v}}_H)|_E \text{ for some } \bar{\mathbf{v}}_H \in \bar{\mathbf{V}}_H \right\}, \tag{2.24}$$

wherein, technically, $\mathcal{E}_E \bar{\mathbf{v}}_H = \mathcal{E}_E(\bar{\mathbf{v}}_H|_E)$. Now $\hat{\mathbf{V}}_{H,*} \subseteq X$, where

$$X = \bigoplus_{E \in \mathcal{T}_H} H(\mathrm{div}; E) \tag{2.25}$$

is a Banach space with the norm $\|\mathbf{v}\|_X = \left(\sum_{E \in \mathcal{T}_H} \|\mathbf{v}\|^2_{H(\mathrm{div};E)} \right)^{1/2}$. Clearly $\mathbf{V} = H_0(\mathrm{div}; \Omega) \subseteq X$, but if $E_* \neq E$ for any $E \in \mathcal{T}_H$, then $\hat{\mathbf{V}}_{H,*} \not\subseteq \mathbf{V}$ and we have a nonconforming finite element space.

We approximate (2.14)–(2.15) by the problem of finding $(\hat{\mathbf{u}}_H, \bar{p}_H) \in \hat{\mathbf{V}}_{H,*} \times \bar{W}_H$ such that

$$\sum_{E \in \mathcal{T}_H} (\nabla \cdot \hat{\mathbf{u}}_H, \bar{w}_H)_E = (\hat{f}, \bar{w}_H) \qquad \forall\, \bar{w}_H \in \bar{W}_H, \tag{2.26}$$

$$(\alpha \hat{\mathbf{u}}_H, \hat{\mathbf{v}}_H) = \sum_{E \in \mathcal{T}_H} (\bar{p}_H, \nabla \cdot \hat{\mathbf{v}}_H)_E + (\hat{\mathbf{b}}_*, \hat{\mathbf{v}}_H) \quad \forall\, \hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}, \tag{2.27}$$

where

$$\hat{\mathbf{b}}_* = \mathbf{b} - \alpha \left(\mathbf{v}_g + \sum_{E \in \mathcal{T}_H} \tilde{\mathbf{u}}'_*|_E \right).$$

Define the affine space

$$\mathbf{V}_{H,*} = \hat{\mathbf{V}}_{H,*} + \sum_{E \in \mathcal{T}_H} \tilde{\mathbf{u}}'_*|_E + \mathbf{v}_g \tag{2.28}$$

and the discrete oversampled approximation

$$\mathbf{u} \approx \mathbf{u}_H = \hat{\mathbf{u}}_H + \sum_{E\in\mathcal{T}_H} \tilde{\mathbf{u}}'_*|_E + \mathbf{v}_g \in \mathbf{V}_{H,*}, \tag{2.29}$$

$$p \approx p_H = \bar{p}_H + \sum_{E\in\mathcal{T}_H} \left(\hat{p}'_*(\mathcal{E}_E\bar{\mathbf{u}}_H) + \tilde{p}'_*\right)\big|_E \in W. \tag{2.30}$$

The full approximation satisfies

$$\sum_{E\in\mathcal{T}_H} (\nabla\cdot\mathbf{u}_H, w)_E = (f,w) \qquad \forall\, w\in W, \tag{2.31}$$

$$(\alpha\mathbf{u}_H, \hat{\mathbf{v}}_H) = \sum_{E\in\mathcal{T}_H} (\bar{p}_H, \nabla\cdot\hat{\mathbf{v}}_H)_E + (\mathbf{b},\hat{\mathbf{v}}_H) \quad \forall\, \hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}, \tag{2.32}$$

which corresponds to the original system (1.4)–(1.5). The systems (2.26)–(2.27) and (2.31)–(2.32) are equivalent; the former is suitable for computation and the latter for analysis. In section 4, it will be shown using the abstract inf-sup lemma [8, 12, 11, 16, 17] that this problem has a unique solution.

If oversampling is *not* used, this is the same discrete approximation considered in [5], except that there the subgrid operators are also approximated on a finer mesh than $\mathcal{T}_H$. Since our concern in this paper is to relate ϵ, the scale of the heterogeneity, to H, the size of the coarse mesh, we have assumed that the subgrid operators are fully resolved (as was done in [17]).

3. Partial equivalence with the multiscale finite element method. In [17], Chen and Hou give a mixed finite element method for the equations making use of their multiscale finite element basis functions. As we show in this section, their method is fundamentally equivalent to that described in this paper in the case where $\bar{\mathbf{V}}_H$ is the vector variable part of the RT0 space and $\tilde{\mathbf{u}}'$ and $\tilde{p}'$ vanish. Note that from (2.9)–(2.10), $\tilde{\mathbf{u}}'$ and $\tilde{p}'$ vanish exactly when $(f-\nabla\cdot\mathbf{v}_g, w') = 0$ for all $w'\in W'$ and $(\mathbf{b}-\alpha\mathbf{v}_g, \mathbf{v}') = 0$ for all $\mathbf{v}'\in\mathbf{V}'$; that is, the subscales of $f-\nabla\cdot\mathbf{v}_g$ and $\mathbf{b}-\alpha\mathbf{v}_g$ vanish.

Let $E\in\mathcal{T}_H$ be given, and let e_i^E represent the ith edge in two dimensions or face in three dimensions of E. We begin by recalling a standard basis $\{R_i^E\}$ for $\mathrm{RT}_0(E)$, the vector part of the RT0 space [27] on $E\in\mathcal{T}_H$, which satisfies

$$\begin{aligned}
&\nabla\cdot R_i^E = 1/|E| && \text{in } E,\\
&R_i^E = -\nabla\omega_i^E && \text{in } E,\\
&R_i^E\cdot\nu^E = \begin{cases} 1/|e_i^E| & \text{on } e_i^E,\\ 0 & \text{on } e_j^E,\ j\neq i. \end{cases}
\end{aligned}$$

That is, R_i^E is linear, has a constant divergence, and has constant fluxes over the edges or faces of E.

Chen and Hou [17] construct the multiscale finite element space in the following way. Let $\{R_i^{E_*}\}$ be the basis of $\mathrm{RT}_0(E_*)$, the vector part of the RT0 space on E_*, which satisfies

$$R_i^{E_*}\cdot\nu^{E^*} = \begin{cases} 1/|e_i^{E_*}| & \text{on } e_i^{E_*},\\ 0 & \text{on } e_j^{E_*},\ j\neq i, \end{cases}$$

where $e_i^{E_*}$ represents an edge or face of E_*. Since the RT0 basis functions on E_* also span $\mathrm{RT}_0(E)$, there must exist, for each i and j, constants c_{ij}^E such that

$$R_i^E = \sum_j c_{ij}^E R_j^{E_*}|_E.$$

(If $E_* = E$, we simply have $c_{ij}^E = \delta_{ij}$, where δ_{ij} is the Kronecker delta.) Now for each j, let $w_j^{E_*}$ be the unique solution in $L^2(E_*)/\mathbb{R} = W_*'(E_*)$ of the Neumann problem

$$\int_{E_*} a\nabla w_j^{E_*} \cdot \nabla\varphi\,dx = \frac{1}{|E_*|}\int_{E_*} \varphi\,dx - \frac{1}{|e_j^{E_*}|}\int_{e_j^{E_*}} \varphi\,ds \quad \forall \varphi \in H^1(E_*), \tag{3.1}$$

which is equivalent to

$$\nabla\cdot\hat{\psi}_{H,i} = \nabla\cdot R_i^{E_*} \qquad \text{in } E_*, \tag{3.2}$$

$$\hat{\psi}_{H,i} = -a\nabla w_i^{E_*} \qquad \text{in } E_*, \tag{3.3}$$

$$\hat{\psi}_{H,i}\cdot\nu^{E_*} = R_i^{E_*}\cdot\nu^{E_*} \quad \text{on } \partial E_*. \tag{3.4}$$

For each i, set

$$\tilde{w}_i^{E_*} = \sum_j c_{ij}^E w_j^{E_*}.$$

Now let

$$\mathrm{MS}_*(E) = \mathrm{span}\{-a\nabla\tilde{w}_i^{E_*}|_E\}$$

and $\tilde{X}_{H,*} = \{\mathbf{v}\in X : \mathbf{v}|_E \in \mathrm{MS}_*(E) \text{ for all } E\in\mathcal{T}_H\}$. Define

$$\Pi_H : \tilde{X}_{H,*} \to \bigoplus_{E\in\mathcal{T}_H} \mathrm{RT}_0(E)$$

to be the natural projection defined locally for $\mathbf{v}|_E = -\sum_i b_i a\nabla\tilde{w}_i^E$ by $\Pi_H(\mathbf{v})|_E = \sum_i b_i R_i^E$. The *oversampled multiscale finite element* space $\hat{X}_{H,*}\subseteq X$ is then defined by

$$\hat{X}_{H,*} = \big\{\mathbf{v}\in\tilde{X}_{H,*} : \Pi_H\mathbf{v}\in\bar{\mathbf{V}}_H\big\},$$

wherein $\bar{\mathbf{V}}_H$ is RT_0 in this section. Note that again the subgrid problems have been assumed to be solved exactly, since the fine scales can be fully resolved.

To see the equivalence with the construction in this paper, first note that the problems (3.1), i.e., (3.2)–(3.4), and (2.20)–(2.21) are closely related, so that

$$-a\nabla w_j^{E_*} = R_j^{E_*} + \hat{\mathbf{u}}_*'(R_j^{E_*}) = -a\nabla\hat{p}_*'(R_j^{E_*});$$

that is, $w_j^{E_*} = \hat{p}_*'(R_j^{E_*})$. Now clearly $\mathcal{E}_E R_i^E = \sum_j c_{ij}^E R_j^{E_*}$, so

$$\mathcal{E}_E R_i^E + \hat{\mathbf{u}}_*'(\mathcal{E}_E R_i^E) = -a\nabla\hat{p}_*'(\mathcal{E}_E R_i^E) = -\sum_j c_{ij}^E a\nabla\tilde{w}_j^{E_*}.$$

Since the matrix c_{ij}^E is invertible,

$$\mathrm{MS}_*(E) = \mathrm{span}\{R_i^E + \hat{\mathbf{u}}_*'(\mathcal{E}_E R_i^E)|_E\}.$$

Now if $\mathbf{v} \in \tilde{X}_{H,*}$, then $\mathbf{v}|_E = \sum_i b_i\big(R_i^E + \hat{\mathbf{u}}'_*(\mathcal{E}_E R_i^E)|_E\big)$, and so $\Pi_H \mathbf{v}|_E = \sum_i b_i R_i^E$. The condition that $\Pi_H \mathbf{v} \in \bar{\mathbf{V}}_H$ merely says that the local RT0 basis functions fit together globally in $H(\mathrm{div};\Omega)$. Thus $\hat{X}_{H,*}$ is the span of $\bar{\mathbf{v}}_H + \sum_E \hat{\mathbf{u}}'_*(\mathcal{E}_E \bar{\mathbf{v}}_H)|_E$ for $\bar{\mathbf{v}}_H \in \bar{\mathbf{V}}_H$; that is,

$$\hat{X}_{H,*} = \hat{\mathbf{V}}_{H,*},$$

and our construction agrees with that in [17], up to the treatment of $\tilde{\mathbf{u}}'$ and $\tilde{p}'$. Moreover, the mixed multiscale finite element method obtains only $\bar{p}_H \in \bar{W}_H$ (not p_H) from (2.30).

4. Analysis of the saddle point variational problem. In this paper, we use the notation $\|\cdot\|_{j,S}$ for the norm of the Sobolev space $H^j(S)$, and $\|\cdot\|_{j,p,S}$ for the norm of the Sobolev space $W^{j,p}(S)$ when $p \neq 2$. We proceed through a series of lemmas.

LEMMA 4.1. *There exists $C > 0$, independent of ϵ and H, such that for each $E \in \mathcal{T}_H$ and $\mathbf{v} \in \mathbf{V}$,*

$$\|\hat{\mathbf{u}}'_*(\mathbf{v})\|_{0,E_*} + \|\nabla \hat{p}'_*(\mathbf{v})\|_{0,E_*} \leq C\|\mathbf{v}\|_{0,E_*}.$$

Moreover, if $\mathbf{v} \in \mathbf{V}'_*(E_*)$ has vanishing divergence, then $\hat{\mathbf{u}}'_*(\mathbf{v}) = -\mathbf{v}$ and $\hat{p}'_*(\mathbf{v}) = 0$.

Proof. This is the standard energy estimate for the differential system (2.20)–(2.21), and the bound depends only on the ellipticity and continuity constants for a and so is independent of ϵ and H. The final remark is obvious from the definition of the operator. ∎

LEMMA 4.2. *There exists $C > 0$, independent of ϵ and H, such that for any $\hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}$, if $\bar{\mathbf{v}}_H \in \bar{\mathbf{V}}_H$ is any element corresponding to $\hat{\mathbf{v}}_H$ in the sense that*

$$\hat{\mathbf{v}}_H = \bar{\mathbf{v}}_H + \sum_{E \in \mathcal{T}_H} \hat{\mathbf{u}}'_*(\mathcal{E}_E \bar{\mathbf{v}}_H)|_E, \tag{4.1}$$

then on any E, $\nabla \cdot \hat{\mathbf{v}}_H|_E = \nabla \cdot \bar{\mathbf{v}}_H|_E$ and

$$\|\hat{\mathbf{v}}_H\|_{H(\mathrm{div};E)} \leq C\|\bar{\mathbf{v}}_H\|_{H(\mathrm{div};E)}.$$

Proof. By the definition (2.24), each $\hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}$ has at least one $\bar{\mathbf{v}}_H$ satisfying (4.1). Since the operator norm of $\mathcal{E}_E$ (as applied to low order polynomials and with respect to the $L^2(E)$- and $L^2(E_*)$-norms) is bounded uniformly in E and H under our assumptions on the shape regularity of E and E_*, we have

$$\|\mathcal{E}_E \bar{\mathbf{v}}_H + \hat{\mathbf{u}}'_*(\mathcal{E}_E \bar{\mathbf{v}}_H)\|_{0,E_*} \leq C\|\mathcal{E}_E \bar{\mathbf{v}}_H\|_{0,E_*} \leq C\|\bar{\mathbf{v}}_H\|_{0,E}.$$

Note that $\mathcal{E}_E \bar{\mathbf{v}}_H + \hat{\mathbf{u}}'_*(\mathcal{E}_E \bar{\mathbf{v}}_H)$ agrees with $\hat{\mathbf{v}}_H$ on E, so we have

$$\|\hat{\mathbf{v}}_H\|_{0,E} \leq C\|\bar{\mathbf{v}}_H\|_{0,E}.$$

The above inequality holds for the $H(\mathrm{div};E)$-norm as well, because $\nabla \cdot \hat{\mathbf{u}}'_*(\mathbf{v})$ is identically zero for all $\mathbf{v} \in \mathbf{V}$, which implies $\nabla \cdot \hat{\mathbf{v}}_H = \nabla \cdot \bar{\mathbf{v}}_H$. ∎

LEMMA 4.3. *There exists a constant $\beta > 0$, independent of ϵ and H, such that for any $\bar{q}_H \in \bar{W}_H$, the following inf-sup condition holds:*

$$\sup_{0 \neq \hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}} \frac{\sum_{E \in \mathcal{T}_H} (\bar{q}_H, \nabla \cdot \hat{\mathbf{v}}_H)_E}{\|\hat{\mathbf{v}}_H\|_X} \geq \beta \|\bar{q}_H\|_{0,\Omega}.$$

Proof. It is known that the inf-sup condition holds for all the usual mixed finite element spaces, such as $\bar{W}_H \times \bar{\mathbf{V}}_H$. Because $\nabla \cdot \hat{\mathbf{u}}'_*(\cdot) = 0$, and by Lemma 4.2, we have

$$\sup_{0 \neq \hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_{H,*}} \frac{\sum_{E \in \mathcal{T}_H} (\bar{q}_H, \nabla \cdot \hat{\mathbf{v}}_H)_E}{\|\hat{\mathbf{v}}_H\|_X} \geq C \sup_{0 \neq \bar{\mathbf{v}}_H \in \bar{\mathbf{V}}_H} \frac{(\bar{q}_H, \nabla \cdot \bar{\mathbf{v}}_H)}{\|\bar{\mathbf{v}}_H\|_{H(\mathrm{div};\Omega)}} \geq C\bar{\beta} \|\bar{q}_H\|_{0,\Omega},$$

where $\bar{\beta} > 0$ is the inf-sup condition constant for $\bar{W}_H \times \bar{\mathbf{V}}_H$. ∎

To obtain a unique solution of the discrete approximation (2.31)–(2.32) of (1.4)–(1.5), we can now apply the abstract inf-sup theory given in [16], for example. We can also obtain a bound on the approximation error, but it involves the approximation of p in $\bar{W}_H$, which is only first order accurate. This is acceptable for RT0, but suboptimal for the higher order spaces.

THEOREM 4.4. *There exists a unique solution $(\mathbf{u}_H, \bar{p}_H) \in \mathbf{V}_{H,*} \times \bar{W}_H$ to (2.31)–(2.32). Moreover, there exists $C > 0$, independent of ϵ and H, such that if $(\mathbf{u}, p)$ is the solution of (1.4)–(1.5), then*

$$\nabla \cdot \mathbf{u}_H = \nabla \cdot \mathbf{u} = f, \tag{4.2}$$

$$\|\mathbf{u} - \mathbf{u}_H\|_{0,\Omega} \leq C \Bigg\{ \inf_{\mathbf{v}_H \in \mathbf{V}_{H,*}, \nabla \cdot \mathbf{v}_H = \nabla \cdot \mathbf{u}} \|\mathbf{u} - \mathbf{v}_H\|_{0,\Omega} + \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}, \nabla \cdot \psi_H = 0} \frac{|(\alpha \mathbf{u} - \mathbf{b}, \psi_H)|}{\|\psi_H\|_{0,\Omega}} \Bigg\}, \tag{4.3}$$

$$\|\bar{p} - \bar{p}_H\|_{0,\Omega} \leq C \Bigg\{ \|\mathbf{u} - \mathbf{u}_H\|_{0,\Omega} + \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}} \frac{\left| (\alpha \mathbf{u} - \mathbf{b}, \psi_H) - \sum_{E \in \mathcal{T}_H} (p, \nabla \cdot \psi_H)_E \right|}{\|\psi_H\|_X} \Bigg\}. \tag{4.4}$$

Proof. The first equality follows from (2.31). The inf-sup condition of the previous lemma and (2.32) allow us to estimate directly that

$$\begin{aligned}
&\beta \|\bar{p} - \bar{p}_H\|_{0,\Omega} \\
&\quad \leq \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}} \frac{\sum_{E \in \mathcal{T}_H} (\bar{p} - \bar{p}_H, \nabla \cdot \psi_H)_E}{\|\psi_H\|_X} \\
&\quad = \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}} \frac{\sum_{E \in \mathcal{T}_H} \left[(p, \nabla \cdot \psi_H)_E - (\alpha \mathbf{u}_H - \mathbf{b}, \psi_H)_E \right]}{\|\psi_H\|_X} \\
&\quad \leq \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}} \frac{\sum_{E \in \mathcal{T}_H} \left[(p, \nabla \cdot \psi_H)_E - (\alpha \mathbf{u} - \mathbf{b}, \psi_H)_E \right]}{\|\psi_H\|_X} + \|\alpha(\mathbf{u} - \mathbf{u}_H)\|_{0,\Omega},
\end{aligned}$$

and the third result (4.3) follows.

The statement $\mathbf{v}_H \in \mathbf{V}_{H,*}$ such that $\nabla \cdot \mathbf{v}_H = \nabla \cdot \mathbf{u} = \nabla \cdot \mathbf{u}_H$ merely says that $\mathbf{v}_H - \mathbf{u}_H \in \hat{\mathbf{V}}_{H,*}$ and has vanishing divergence. For any such $\mathbf{v}_H$, we compute

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}_H\|_{0,\Omega} &\le \|\mathbf{u} - \mathbf{v}_H\|_{0,\Omega} + \|\mathbf{v}_H - \mathbf{u}_H\|_{0,\Omega} \\
&\le \|\mathbf{u} - \mathbf{v}_H\|_{0,\Omega} + C \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}, \nabla \cdot \psi_H = 0} \frac{|(\alpha(\mathbf{v}_H - \mathbf{u}_H), \psi_H)|}{\|\psi_H\|_{0,\Omega}} \\
&\le \|\mathbf{u} - \mathbf{v}_H\|_{0,\Omega} + C \Bigg\{ \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}, \nabla \cdot \psi_H = 0} \frac{|(\alpha(\mathbf{u} - \mathbf{u}_H), \psi_H)|}{\|\psi_H\|_{0,\Omega}} \\
&\quad + \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}, \nabla \cdot \psi_H = 0} \frac{|(\alpha(\mathbf{v}_H - \mathbf{u}), \psi_H)|}{\|\psi_H\|_{0,\Omega}} \Bigg\} \\
&\le C \Bigg\{ \|\mathbf{u} - \mathbf{v}_H\|_{0,\Omega} + \sup_{0 \neq \psi_H \in \hat{\mathbf{V}}_{H,*}, \nabla \cdot \psi_H = 0} \frac{|(\alpha \mathbf{u} - \mathbf{b}, \psi_H)|}{\|\psi_H\|_{0,\Omega}} \Bigg\},
\end{aligned}$$

since $(\alpha \mathbf{u}_H, \psi_H) = (\mathbf{b}, \psi_H)$, and the second result (4.4) follows.

Finally, we obtain that the discrete solution must be unique by setting all the data to zero (f, $\mathbf{b}$, $\mathbf{v}_g$, which implies that $\mathbf{u}$ and p also vanish). We then also obtain existence of a solution, since the system has finite dimensions and is square. ∎

5. Some homogenization theory needed for multiscale error analysis. We give a multiscale analysis of the error similar to that given by Hou et al. [22, 17]. This analysis determines the behavior of the error as a function both of H and the scale of the heterogeneity in a, which we denote by ϵ. If $H \sim \epsilon$, the system is well resolved, there is no need for oversampling, and the scheme converges with optimal order of approximation [5]. Thus we tacitly assume the underresolved case where $\epsilon \ll H$. The difficulty, then, with standard approximation theory is that the error is bounded in terms of H and derivatives of the solution. However, we expect that each derivative of the solution is proportional to ϵ^{-1}, and H/ϵ is not small. The two exceptions are given by the standard energy estimates for our problem, which are stated in the following lemma.

LEMMA 5.1. *Let $(\mathbf{u}, p) \in (\mathbf{V} + \mathbf{v}_g) \times W$ be the solution of* (1.4)–(1.5). *Then there is a constant $C > 0$, depending only on the ellipticity bounds for a, such that*

$$\begin{aligned}
&\|\mathbf{u}\|_{0,\Omega} + \|\nabla p\|_{0,\Omega} \le C\big\{\|f - \nabla \cdot \mathbf{v}_g\|_{0,\Omega} + \|\mathbf{v}_g\|_{0,\Omega} + \|\mathbf{b}\|_{0,\Omega}\big\}, \\
&\|\nabla \cdot \mathbf{u}\|_{0,\Omega} = \|f\|_{0,\Omega}.
\end{aligned}$$

In order to quantify the scale of the heterogeneity, we use homogenization theory. Thus we assume that the permeability has "locally periodic" oscillations whose scale is on the order of some $\epsilon > 0$. That is, let $C^1_{\mathrm{per}}(\mathbb{R}^d)$ denote the space of all $C^1(\mathbb{R}^d)$ functions that are periodic with respect to the unit cube $Y \subseteq \mathbb{R}^d$, and assume that $a = a(x, x/\epsilon)$, where for each $i, j = 1, \ldots, d$, $a_{ij}(x, y) \in C^1(\bar{\Omega}; C^1_{\mathrm{per}}(\mathbb{R}^d))$, and a varies slowly in its first argument on a scale resolved by H. Moreover, suppose $\mathbf{B} \in (L^2(D))^d$.

Following Chen and Hou [17], we now review the relevant aspects of homogenization theory. Let $D \subseteq \Omega$ be a Lipschitz domain in $\mathbb{R}^d$ and suppose $F \in L^2(D)$ and $G \in L^2(\partial D)$ satisfy the compatibility condition

$$\int_D F(x)\, dx = \int_{\partial D} G(x)\, ds.$$

For each $\epsilon > 0$, we let $q^\epsilon \in H^1(D)/\mathbb{R}$ be the unique solution of the Neumann problem

$$\int_D a(x, x/\epsilon)\,(\nabla q^\epsilon - \mathbf{B}) \cdot \nabla\varphi\, dx = \int_D F\varphi\, dx - \int_{\partial D} G\varphi\, ds \quad \forall\, \varphi \in H^1(D). \tag{5.1}$$

The homogenized coefficient matrix $a^0(x) = \big(a^0_{ij}(x)\big)_{ij}$ of $a(x, x/\epsilon)$ is given by

$$a^0_{ij}(x) = \frac{1}{|Y|} \sum_{k=1}^{d} \int_Y a_{ik}(x, y) \left(\delta_{kj} - \frac{\partial \chi^j}{\partial y_k}(x, y)\right) dy, \quad x \in \Omega,\ y \in Y, \tag{5.2}$$

where δ_{kj} is the Kronecker delta and $\chi^j(x, y)$ is the Y-periodic (in y) solution of the jth cell problem

$$\sum_{i=1}^{d} \sum_{k=1}^{d} \frac{\partial}{\partial y_i} \left(a_{ik}(x, y) \frac{\partial \chi^j}{\partial y_k}(x, y)\right) = \sum_{i=1}^{d} \frac{\partial}{\partial y_i} a_{ij}(x, y),$$

with $\int_Y \chi^j(x, y)\, dy = 0$. Now, we let $q^0 \in H^1(D)/\mathbb{R}$ be the unique solution of the homogenized counterpart of (5.1), namely,

$$\int_D a^0(x)(\nabla q^0 - \mathbf{B}) \cdot \nabla\varphi\, dx = \int_D F\varphi\, dx - \int_{\partial D} G\varphi\, ds \ \ \forall\, \varphi \in H^1(D). \tag{5.3}$$

In the usual way, we define the first order corrector of q^ϵ by

$$(q^0)^\epsilon_1(x) = q^0(x) - \epsilon \sum_{k=1}^{d} \chi^k(x, x/\epsilon) \left(\frac{\partial q^0}{\partial x_k} - B_k\right). \tag{5.4}$$

Recall that we use the notation $\|\cdot\|_{j,p,S}$ for the norm of the Sobolev space $W^{j,p}(S)$, and simply $\|\cdot\|_{j,S}$ if $p = 2$.

THEOREM 5.2. *Suppose that $q^0 \in H^2(D) \cap W^{1,\infty}(D)$ and $D' \subset D$. Let the fluxes be denoted by*

$$\mathbf{U}^\epsilon = -a_\epsilon(\nabla q^\epsilon - \mathbf{B}) \quad \textit{and} \quad \mathbf{U}^0 = -a^0(\nabla q^0 - \mathbf{B}).$$

There exist a constant C, independent of ϵ, the size of the domains D and D', and the terms $\mathbf{B}$, F, and G, and there exists a boundary corrector $\theta^S_\epsilon \in H^1(S)/\mathbb{R}$, defined below for $S \subset D$ in (5.11)–(5.12) and (5.9)–(5.10), such that

$$\|\nabla[q^\epsilon - (q^0)^\epsilon_1 - \epsilon\theta^D_\epsilon]\|_{0,D} \le C\epsilon \|\nabla q^0 - \mathbf{B}\|_{1,D}, \tag{5.5}$$

$$\|\mathbf{U}^\epsilon - (\mathbf{U}^0 + \epsilon a_\epsilon \nabla\theta^{D'}_\epsilon - \epsilon a_\epsilon \nabla\theta^D_\epsilon + \psi^{D'}_{\mathrm{Sol}})\|_{0,D'} \le C\epsilon \|\nabla q^0 - \mathbf{B}\|_{1,D}, \tag{5.6}$$

where $\psi^{D'}_{\mathrm{Sol}}$ is a solenoidal vector, i.e., $\nabla\cdot\psi^{D'}_{\mathrm{Sol}} = 0$ and $\psi^{D'}_{\mathrm{Sol}}\cdot\nu^{D'} = 0$ on $\partial D'$. Moreover,

$$\|\epsilon\nabla\theta^S_\epsilon\|_{0,S} \le C\big\{\epsilon\|\nabla q^0 - \mathbf{B}\|_{1,S} + \sqrt{\epsilon|\partial S|}\,\|\nabla q^0 - \mathbf{B}\|_{0,\infty,S}\big\}, \tag{5.7}$$

$$\|\epsilon\nabla\theta^S_\epsilon\|_{0,S} \le C\big(\epsilon + H_S^{-1}(\epsilon|\partial S|)^{1/d-\eta}\big)\|\nabla q^0 - \mathbf{B}\|_{1,S}, \tag{5.8}$$

where S is D or D', $H_S = \operatorname{diam}(S)$, and $\eta = 0$ if $d = 3$ and η is any fixed positive number if $d = 2$.

REMARK 5.1. *In [17], it is conjectured, but not proven, that estimate (5.7) can be improved by replacing $\sqrt{\epsilon|\partial S|}$ by $\epsilon\sqrt{|S|}/H_S$ in the oversampled case, i.e., with S*

replaced by $S' \subset S$ on the left-hand side. If this turns out to be the case, the estimates we derive can be correspondingly improved.

In [25], (5.5) and (5.7) were derived in the case where the coefficient $a = a(x)$ is periodic and $\mathbf{B} = 0$. In [17], the proof was elucidated and extended to the case in point, where $a = a(x, x/\epsilon)$ is locally periodic. The proof easily modifies to handle the extra term related to $\mathbf{B}$, and we reproduce it here in brief so that we can extract the estimates (5.6) and (5.8).

Proof. We use the Einstein summation convention for repeated indices, and the more concise notation $\partial_j = \partial/\partial x_j$, and $\partial_j^x = \partial/\partial x_j$ and $\partial_j^y = \partial/\partial y_j$ if we are dealing with a function of (x, y). The key to the proof is to note that

$$a_{ik}^0(x) - a_{\epsilon,ij}(x,y)\big(\delta_{jk} - \partial_j^y \chi^k(x,y)\big) = \partial_j^y A_{ij}^k(x,y), \tag{5.9}$$

where $A_{ij}^k(x,y)$ is skew-symmetric for each k [25, p. 6]. Let us denote

$$\gamma_i(x) = \partial_j\{A_{ij}^k(x, x/\epsilon)\,[\partial_k q^0(x) - B_k(x)]\}, \tag{5.10}$$

so that

$$\partial_j^y A_{ij}^k(x,x/\epsilon)\,(\partial_k q^0 - B_k) = \epsilon\gamma_i - \epsilon\partial_j^x A_{ij}^k(\partial_k q^0 - B_k) - \epsilon A_{ij}^k \partial_j(\partial_k q^0 - B_k).$$

After some manipulation

$$-a_{\epsilon,ij}[\partial_j (q^0)_1^\epsilon - B_j] = -a_{ij}^0(\partial_j q^0 - B_j) + \epsilon\gamma_i + \epsilon\psi_{1,i},$$

where

$$\begin{aligned}\psi_{1,i} = {} & -\partial_j^x A_{ij}^k(\partial_k q^0 - B_k) - A_{ij}^k \partial_j(\partial_k q^0 - B_k) \\ & + a_{\epsilon,ij}\partial_j^x \chi^k(\partial_k q^0 - B_k) + a_{\epsilon,ij}\chi^k \partial_j(\partial_k q^0 - B_k).\end{aligned}$$

Now we let $\theta_\epsilon^S \in H^1(S)/\mathbb{R}$ be defined by

$$\nabla \cdot (a_\epsilon \nabla \theta_\epsilon^S) = 0 \qquad \text{in } S, \tag{5.11}$$

$$a_\epsilon \nabla \theta_\epsilon^S \cdot \nu^S = \gamma \cdot \nu^S \quad \text{on } \partial S, \tag{5.12}$$

so that $\psi_{\mathrm{Sol}}^S = \epsilon(\gamma - a_\epsilon \nabla \theta_\epsilon^S)$ has the requisite properties and

$$-a_\epsilon[\nabla (q^0)_1^\epsilon - \mathbf{B}] = -a^0(\nabla q^0 - \mathbf{B}) + \epsilon a_\epsilon \nabla \theta_\epsilon^S + \psi_{\mathrm{Sol}}^S + \epsilon\psi_1. \tag{5.13}$$

It is now a simple consequence of the governing equations (5.1) and (5.3) and the properties of ψ_{Sol}^D that

$$(a_\epsilon \nabla[q_\epsilon - (q^0)_1^\epsilon - \epsilon\theta_\epsilon^D], \nabla\varphi)_D = \epsilon(\psi_1, \nabla\varphi)_D,$$

and the first result (5.5) follows easily. The second result (5.6) follows from (5.13) and the previous result.

To obtain bounds on the boundary corrector, we use a smooth cut-off function $\zeta_\epsilon(x) \in [0,1]$ with compact support that is one except near ∂S, where it tends to zero in a narrow region of width ϵ with gradient bounded by C/ϵ. Now let

$$\begin{aligned}\gamma_{I,i}(x) &= \partial_j\{A_{ij}^k(x, x/\epsilon)\,[\partial_k q^0(x) - B_k(x)]\zeta_\epsilon(x)\}, \\ \gamma_{B,i}(x) &= \partial_j\{A_{ij}^k(x, x/\epsilon)\,[\partial_k q^0(x) - B_k(x)][1 - \zeta_\epsilon(x)]\}\end{aligned}$$

(i.e., $\gamma = \gamma_B + \gamma_I$), and note that (5.11)–(5.12) imply that

$$\|\nabla\theta_\epsilon\|_{0,S} \le C\|\gamma_B\|_{0,S},$$

since γ_B is divergence free. Finally,

$$\begin{aligned}
\epsilon\|\gamma_{B,i}\|_{0,S} &= \epsilon\|\partial_j\{A^k_{ij}[\partial_k q^0 - B_k][1-\zeta_\epsilon]\}\|_{0,S} \\
&\le \epsilon\|\partial^x_j A^k_{ij}[\partial_k q^0 - B_k][1-\zeta_\epsilon]\|_{0,S} + \epsilon\|A^k_{ij}\partial_j[\partial_k q^0 - B_k][1-\zeta_\epsilon]\|_{0,S} \\
&\quad + \epsilon\|A^k_{ij}[\partial_k q^0 - B_k]\partial_j[1-\zeta_\epsilon]\|_{0,S} + \|\partial^y_j A^k_{ij}[\partial_k q^0 - B_k][1-\zeta_\epsilon]\|_{0,S} \\
&\le C\big\{\epsilon\|\nabla q^0 - \mathbf{B}\|_{1,S} + \|\nabla q^0 - \mathbf{B}\|_{0,S^\zeta_\epsilon}\big\},
\end{aligned}$$

where S^ζ_ϵ is the support of $1-\zeta_\epsilon$. Since the measure of S^ζ_ϵ is proportional to $\epsilon|\partial S|$, we have

$$\|\nabla q^0 - \mathbf{B}\|_{0,S^\zeta_\epsilon} \le \sqrt{\big|S^\zeta_\epsilon\big|}\,\|\nabla q^0 - \mathbf{B}\|_{0,\infty,S} \le C\sqrt{\epsilon|\partial S|}\,\|\nabla q^0 - \mathbf{B}\|_{0,\infty,S},$$

and (5.7) follows. To show (5.8), we instead use Hölder's inequality with $r = d/(d-2)$ (or large but finite if $d = 2$) and the Sobolev imbedding theorem to show that

$$\begin{aligned}
\|\nabla q^0 - \mathbf{B}\|_{0,S^\zeta_\epsilon} &\le C(\epsilon|\partial S|)^{(r-1)/2r}\|\nabla q^0 - \mathbf{B}\|_{0,2r,S^\zeta_\epsilon} \\
&\le CH_S^{-1}(\epsilon|\partial S|)^{1/d-\eta}\|\nabla q^0 - \mathbf{B}\|_{1,S},
\end{aligned}$$

wherein $\eta = 0$ if $d = 3$ and $\eta > 0$ if $d = 2$ (the factor H_S^{-1} comes from a scaling argument on the size of the domain S). The proof is complete. ∎

We will apply Theorem 5.2 several times, with D being one of Ω, E, or E_*. Since these are convex polygonal domains, the hypothesis $q^0 \in H^2(D) \cap W^{1,\infty}(D)$ will hold provided that, for some $r > d$, $F \in L^r(D)$ and $G = \mathbf{v}_g \cdot \nu^D$ on ∂D for some $\mathbf{v}_g \in (W^{1,r}(D))^d$ [17, 20, 26].

6. Multiscale estimation of the errors. In this section, we estimate the terms in the basic estimates of Theorem 4.4 for the velocity and pressure errors. We obtain the following estimates which isolate the dependence on both H and ϵ.

THEOREM 6.1. *For each $\epsilon > 0$, let $(\mathbf{u}^\epsilon, p^\epsilon) \in (\mathbf{V}+\mathbf{v}_g)\times W$ be the solution of* (1.4)–(1.5) *with the coefficient $a_\epsilon = a(x, x/\epsilon)$ and $\alpha_\epsilon = a_\epsilon^{-1}$. Let $(\mathbf{u}^0, p^0) \in (\mathbf{V} + \mathbf{v}_g) \times W$ satisfy* (1.4)–(1.5) *with the homogenized coefficient a^0 defined by* (5.2) *in place of a, and $\alpha^0 = (a^0)^{-1}$. For $H > 0$, let $(\hat{\mathbf{u}}^\epsilon_H, \bar{p}^\epsilon_H) \in \hat{\mathbf{V}}_{H,*} \times \bar{W}_H$ be the solution of the discrete upscaled equation* (2.26)–(2.27), *and define $\mathbf{u}^\epsilon_H$ by* (2.29).

(a) Oversampling. *Assume that the partition $\mathcal{T}_H$ consists only of simplices and $\bar{\mathbf{V}}_H$ is RT0. Then*

$$\begin{aligned}
(6.1)\qquad &\|\mathbf{u}^\epsilon - \mathbf{u}^\epsilon_H\|_{H(\mathrm{div};\Omega)} + \|p^\epsilon - \bar{p}^\epsilon_H\|_{0,\Omega} \\
&\quad\le C\big\{(\epsilon + \sqrt{\epsilon/H} + H)\big[\|f - \nabla\cdot\mathbf{v}_g\|_{0,\Omega} + \|\mathbf{v}_g\|_{0,\Omega} + \|\mathbf{b}\|_{0,\Omega}\big] \\
&\qquad + (\epsilon/H)\,\|\nabla p^0\|_{0,\infty,\Omega} + \sqrt{\epsilon/H}\,\|\nabla p^0 - \mathbf{b}\|_{0,\infty,\Omega} \\
&\qquad + (\epsilon + H)\big[\|\nabla p^0\|_{1,\Omega} + \|\nabla p^0 - \mathbf{b}\|_{1,\Omega}\big] + H\|\mathbf{u}^0 - \mathbf{v}_g\|_{1,\Omega}\big\}.
\end{aligned}$$

If the oversampling conjecture of Chen and Hou [17] *holds (Remark* 5.1*), then $\sqrt{\epsilon/H}$ may be replaced by ϵ/H above.*

(b) Nonoversampling. *Assume that oversampling is* not *used. Let* $m = 1$ *when* $\bar{\mathbf{V}}_H$ *is RT*0 *and* $m = 1$ *or* 2 *when* $\bar{\mathbf{V}}_H$ *is BDM*1 *or BDDF*1. *Then*

$$\begin{aligned}(6.2)\qquad &\|\mathbf{u}^\epsilon - \mathbf{u}_H^\epsilon\|_{H(\mathrm{div};\Omega)} + \|\bar{p}^\epsilon - \bar{p}_H^\epsilon\|_{0,\Omega}\\ &\quad\le C\big\{\epsilon\|\nabla p^0 - \mathbf{b}\|_{1,\Omega} + \sqrt{\epsilon/H}\,\|\nabla p^0 - \mathbf{b}\|_{0,\infty,\Omega}\\ &\qquad + H^m\big(\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,\Omega} + \|f - \nabla\cdot\mathbf{v}_g\|_{m-1,\Omega}\big)\big\}.\end{aligned}$$

Moreover, with $\eta = 0$, *if* $d = 3$, *and* η *any fixed positive number, if* $d = 2$,

$$\begin{aligned}(6.3)\qquad \|\mathbf{u}^\epsilon - \mathbf{u}_H^\epsilon\|_{H(\mathrm{div};\Omega)} &\le C\big\{\big(\epsilon + (\epsilon/H)^{1/d-\eta}\big)\|\nabla p^0 - \mathbf{b}\|_{1,\Omega}\\ &\qquad + H^m\big(\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,\Omega} + \|f - \nabla\cdot\mathbf{v}_g\|_{m-1,\Omega}\big)\big\}.\end{aligned}$$

We remark that (a) is a small improvement over the result in [17, Theorem 2.2]. Assuming the more pessimistic but proven bound on the boundary corrector, and with $\mathbf{v}_g = \mathbf{b} = 0$, this previous result is

$$\begin{aligned}&\|\mathbf{u}^\epsilon - \mathbf{u}_H^\epsilon\|_{H(\mathrm{div};\Omega)} + \|p^\epsilon - \bar{p}_H^\epsilon\|_{0,\Omega}\\ &\quad\le C\big\{(\epsilon + H)\big(\|p^0\|_{2,\Omega} + \|f\|_{1,\Omega} + \|\mathbf{u}^0\|_{H(\mathrm{div};\Omega)}\big)\\ &\qquad + \sqrt{\epsilon/H}\,\big(\|p^0\|_{1,\infty,\Omega} + \|f\|_{0,\Omega} + \|\mathbf{u}^0\|_{H(\mathrm{div};\Omega)}\big)\big\}.\end{aligned}$$

The small improvement is in the norm on f, which as noted in the introduction can have small scale aspects in some applications such as flow in porous media. Result (b) is new for the BDM1 and BDDF1 spaces, and for RT0 with nonsimplicial elements.

Concerning the proof of this theorem, by Theorem 4.4, for (b), we need only to bound the optimal velocity error, which is done in section 6.1. For (a), we need this, the oversampling error, handled in section 6.2, and the following simple estimate for the pressure. Note that in (6.1), we have p^ϵ rather than $\bar{p}^\epsilon$. This is allowed by the estimate

$$\|p - \bar{p}_H\|_{0,\Omega} \le \|p - \bar{p}\|_{0,\Omega} + \|\bar{p} - \bar{p}_H\|_{0,\Omega} \le CH\|\nabla p\|_{0,\Omega} + \|\bar{p} - \bar{p}_H\|_{0,\Omega}$$

and the bound on $\|\nabla p\|_{0,\Omega}$ in Lemma 5.1. We will improve the pressure estimate of (b) in section 7.

6.1. The optimal velocity error. In this subsection, we assume that oversampling may be used, so as to handle cases (a) and (b) of Theorem 6.1 simultaneously. Let $\bar{\pi}_H : \mathbf{V} \cap L^r(\Omega) \to \bar{\mathbf{V}}_H$ (for some $r > 2$) be the standard mixed finite element interpolation operator [27, 18, 15, 14, 16]. It has the property that

$$\nabla\cdot\bar{\pi}_H\mathbf{v} = \mathcal{P}_{\bar{W}_H}\nabla\cdot\mathbf{v} \tag{6.4}$$

for all $\mathbf{v} \in \mathbf{V} \cap L^r(\Omega)$, where $\mathcal{P}_{\bar{W}_H}$ is the L^2-projection onto $\bar{W}_H$. We also have the approximation property

$$\|\mathbf{v} - \bar{\pi}_H\mathbf{v}\|_{0,\Omega} \le CH^m\|\mathbf{v}\|_{m,\Omega}, \tag{6.5}$$

where $m = 1$ when $\bar{\mathbf{V}}_H$ is RT0 and $m = 1$ or 2 when $\bar{\mathbf{V}}_H$ is BDM1 or BDDF1.

We now note a lemma on the difference between nonoversampled and oversampled quantities.

LEMMA 6.2. *If* $E \in \mathcal{T}_H$ *and* $w \in H^1(E_*)$, *then*

$$\nabla\hat{p}_*'(a\nabla w) = -\nabla w \quad \textit{and} \quad \hat{\mathbf{u}}_*'(a\nabla w) = 0. \tag{6.6}$$

In fact, $\hat{p}'_*(a\nabla w) = -w$ *provided* $w \in W'_*(E_*)$. *Moreover, on* E,

$$\nabla \hat{p}'(a\nabla w) = \nabla \hat{p}'_*(a\nabla w)|_E = -\nabla w. \tag{6.7}$$

Proof. We simply observe that (6.6) provides the unique solution to the equations defining the subgrid operator (2.20)–(2.21). The remark for $w \in W'_*(E_*) \cap H^1(E_*)$ is then trivial, since w is correctly normalized. Similar results hold for $\hat{p}'$, so (6.7) follows. ∎

Our main result in this subsection follows.

LEMMA 6.3. *Let*

$$\hat{\mathbf{v}}_H^\epsilon = \bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g) + \sum_{E\in\mathcal{T}_H} \hat{\mathbf{u}}'_*(\mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g))|_E \in \hat{\mathbf{V}}_{H,*},$$
$$\mathbf{v}_H^\epsilon = \hat{\mathbf{v}}_H^\epsilon + \sum_{E\in\mathcal{T}_H} \tilde{\mathbf{u}}'_*|_E + \mathbf{v}_g \in \mathbf{V}_{H,*}.$$

Then there is $C > 0$, *independent of* ϵ *and* H, *such that*

$$\begin{aligned}
\nabla\cdot\mathbf{u}^\epsilon &= \nabla\cdot\mathbf{v}_H^\epsilon,\\
\|\mathbf{u}^\epsilon - \mathbf{v}_H^\epsilon\|_{0,\Omega} &\le C\big\{\epsilon\|\nabla p^0 - \mathbf{b}\|_{1,\Omega} + \sqrt{\epsilon/H}\,\|\nabla p^0 - \mathbf{b}\|_{0,\infty,\Omega}\\
&\qquad + H^m\big(\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,\Omega} + \|f - \nabla\cdot\mathbf{v}_g\|_{m-1,\Omega}\big)\big\},\\
\|\mathbf{u}^\epsilon - \mathbf{v}_H^\epsilon\|_{0,\Omega} &\le C\big\{\big(\epsilon + (\epsilon/H)^{1/d-\eta}\big)\|\nabla p^0 - \mathbf{b}\|_{1,\Omega}\\
&\qquad + H^m\big(\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,\Omega} + \|f - \nabla\cdot\mathbf{v}_g\|_{m-1,\Omega}\big)\big\},
\end{aligned}$$

where m *is* 1 *or* 2 *and* $\eta \ge 0$ *as in Theorem* 6.1.

Proof. The divergence result is easy to see from (6.4). For the other result, we work locally on $E_* \supset E \in \mathcal{T}_H$. We have an expansion over E_* similar to the one over E, so on E_* we can write

$$p^\epsilon = \bar{p}_* + \hat{p}'_* + \tilde{p}'_*,$$

where $\bar{p}_*$ is the average of p^ϵ over E_* and $\hat{p}'_*, \tilde{p}'_* \in W'_*(E_*)$ are defined in (2.20)–(2.23) above. (To see this fact, simply consider an expansion as in section 2 on a perturbed coarse mesh containing E_*, and discard the expansion outside E_*.) In fact, we have $\mathbf{u}^\epsilon = \bar{\mathbf{u}}_* + \hat{\mathbf{u}}'_* + \tilde{\mathbf{u}}'_* + \mathbf{v}_g$ and the functional relationship

$$\hat{p}'_* = \hat{p}'_*(\bar{\mathbf{u}}_*) = \hat{p}'_*(\mathbf{u}^\epsilon - \tilde{\mathbf{u}}'_* - \mathbf{v}_g),$$

using Lemma 4.1 to avoid further discussion of $\bar{\mathbf{u}}_*$ and $\hat{\mathbf{u}}'_*$. Thus we have on E that

$$\begin{aligned}
\mathbf{u}^\epsilon &= -a_\epsilon(\nabla p^\epsilon - \mathbf{b})\\
&= -a_\epsilon\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \tilde{\mathbf{u}}'_* - \mathbf{v}_g) - a_\epsilon(\nabla\tilde{p}'_* - \mathbf{b})\\
&= -a_\epsilon[\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0) + \nabla\hat{p}'_*(\mathbf{u}^0 - \mathbf{v}_g) - \nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*)] + \tilde{\mathbf{u}}'_* + \mathbf{v}_g,
\end{aligned}$$

using (2.23) in the last step.

Note that on E,

$$\hat{\mathbf{v}}_H^\epsilon|_E = \bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g) + \hat{\mathbf{u}}'_*(\mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g))|_E = -a_\epsilon\nabla\hat{p}'_*(\mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g))|_E, \tag{6.8}$$

so

$$\begin{aligned}\mathbf{u}^\epsilon - \mathbf{v}_H^\epsilon &= \mathbf{u}^\epsilon - (\hat{\mathbf{v}}_H^\epsilon + \tilde{\mathbf{u}}'_* + \mathbf{v}_g)\\ &= -a_\epsilon[\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0) + \nabla\hat{p}'_*(\mathbf{u}^0 - \mathbf{v}_g) - \nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*) - \nabla\hat{p}'_*(\mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g))].\end{aligned}$$

Now we estimate

$$\begin{aligned}\|\mathbf{u}^\epsilon - \mathbf{v}_H^\epsilon\|_{0,E} \le C\{&\|\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0)\|_{0,E} + \|\mathbf{u}^0 - \mathbf{v}_g - \mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g)\|_{0,E_*}\\ &+ \|\nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*)\|_{0,E}\},\end{aligned} \tag{6.9}$$

using Lemma 4.1 again, this time to bound the operator. The second term on the right is bounded as

$$\|\mathbf{u}^0 - \mathbf{v}_g - \mathcal{E}_E\bar{\pi}_H(\mathbf{u}^0 - \mathbf{v}_g)\|_{0,E_*} \le C\,H^m\,\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,E_*},$$

using the approximation property (6.5) of $\bar{\pi}_H$ (actually, a slight extension to E_*, but the approximation result continues to hold since the operator $\mathcal{E}_E\bar{\pi}_H$ preserves low order polynomials).

For the last term on the far right side of (6.9), since $\tilde{\mathbf{u}}'_* \in \mathbf{V}'_*(E_*)$, note that we have the differential system

$$\begin{aligned}-\nabla\cdot a_\epsilon\nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*) &= \mathcal{P}_{W'_*}(f - \nabla\cdot\mathbf{v}_g) \quad \text{in } E_*,\\ -a_\epsilon\nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*)\cdot\nu^{E_*} &= 0 \qquad\qquad\qquad\quad \text{on } \partial E_*,\end{aligned}$$

where $\mathcal{P}_{W'_*}$ is the L^2-projection onto $W'_*(E_*)$. The standard energy estimate is

$$\|\nabla\hat{p}'_*(\tilde{\mathbf{u}}'_*)\|_{0,E_*} \le C\|\mathcal{P}_{W'_*}(f - \nabla\cdot\mathbf{v}_g)\|_{(H^1(E_*))^*} \le C\,H^m\,\|f - \nabla\cdot\mathbf{v}_g\|_{m-1,E_*}, \tag{6.10}$$

where $(H^1(E_*))^*$ is the dual space of $H^1(E_*)$, using standard negative norm estimates for approximation of a function with vanishing average.

Finally, we estimate the first term on the far right side of (6.9), using Theorem 5.2, specifically the expansion in (5.6). By Lemma 4.1 we can introduce the local solenoidal term $\psi_{\mathrm{Sol}}^{E_*}$, so we have

$$\begin{aligned}&\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0)\\ &\quad= \nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0 - \epsilon a_\epsilon\nabla\theta_\epsilon^{E_*} + \epsilon a_\epsilon\nabla\theta_\epsilon^{\Omega}) + \epsilon\nabla\hat{p}'_*(a_\epsilon\nabla\theta_\epsilon^{E_*}) - \epsilon\nabla\hat{p}'_*(a_\epsilon\nabla\theta_\epsilon^{\Omega})\\ &\quad= \nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0 - \epsilon a_\epsilon\nabla\theta_\epsilon^{E_*} + \epsilon a_\epsilon\nabla\theta_\epsilon^{\Omega} + \psi_{\mathrm{Sol}}^{E_*}) - \epsilon\nabla\theta_\epsilon^{E_*} + \epsilon\nabla\theta_\epsilon^{\Omega},\end{aligned}$$

using Lemma 6.2. Thus Theorem 5.2 gives the two bounds

$$\begin{aligned}&\|\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0)\|_{0,E}\\ &\quad\le C\big\{\epsilon\|\nabla p^0 - \mathbf{b}\|_{1,E_*} + \sqrt{\epsilon|\partial E_*|}\,\|\nabla p^0 - \mathbf{b}\|_{0,\infty,E_*} + \|\epsilon\nabla\theta_\epsilon^{\Omega}\|_{0,E_*}\big\},\\ &\|\nabla\hat{p}'_*(\mathbf{u}^\epsilon - \mathbf{u}^0)\|_{0,E} \le C\big\{\big(\epsilon + H^{-1}(\epsilon|\partial E_*|)^{1/d-\eta}\big)\|\nabla p^0 - \mathbf{b}\|_{1,E_*} + \|\epsilon\nabla\theta_\epsilon^{\Omega}\|_{0,E_*}\big\}\end{aligned}$$

(with the first bound improved if the oversampling conjecture holds).

Combining terms, summing over $E \in \mathcal{T}_H$, and using that the number of overlaps of the E_* are bounded yield

$$\begin{aligned}\|\mathbf{u}^\epsilon - \mathbf{v}_H^\epsilon\|_{0,\Omega} \le C\big\{&\epsilon\|\nabla p^0 - \mathbf{b}\|_{1,\Omega} + \sqrt{\epsilon/H}\,\|\nabla p^0 - \mathbf{b}\|_{0,\infty,\Omega}\\ &+ \|\epsilon\nabla\theta_\epsilon^{\Omega}\|_{0,\Omega} + H^m\left(\|\mathbf{u}^0 - \mathbf{v}_g\|_{m,\Omega} + \|f - \nabla\cdot\mathbf{v}_g\|_{m-1,\Omega}\right)\big\},\end{aligned}$$

wherein $\sqrt{\epsilon/H}\,\|\nabla p^0-\mathbf{b}\|_{0,\infty,\Omega}$ can be replaced by $(\epsilon/H)^{1/d-\eta}\|\nabla p^0-\mathbf{b}\|_{1,\Omega}$. The proof is completed by Theorem 5.2 to bound the global boundary corrector term. ∎

We have an abstract result, analogous to Theorem 5.2, relating $\mathbf{u}^\epsilon$ to a correction of the homogenized solution $\mathbf{u}^0$.

COROLLARY 6.4. *If $\tilde{\mathbf{u}}'_*(\mathbf{u}^0)$ is defined by* (2.22)–(2.23) *with $\mathbf{v}_g$ replaced by $\mathbf{u}^0$, then*

$$\left\|\mathbf{u}^\epsilon-\left(\mathbf{u}^0+\sum_{E\in\mathcal{T}_H}\tilde{\mathbf{u}}'_*(\mathbf{u}^0)|_E\right)\right\|_{0,\Omega}\le C\big\{\epsilon\|\nabla p^0-\mathbf{b}\|_{1,\Omega}+\sqrt{\epsilon/H}\,\|\nabla p^0-\mathbf{b}\|_{0,\infty,\Omega}\big\},$$

$$\left\|\mathbf{u}^\epsilon-\left(\mathbf{u}^0+\sum_{E\in\mathcal{T}_H}\tilde{\mathbf{u}}'_*(\mathbf{u}^0)|_E\right)\right\|_{0,\Omega}\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}\big)\|\nabla p^0-\mathbf{b}\|_{1,\Omega}.$$

Proof. Simply take $\mathbf{v}_g=\mathbf{u}^0$. Then $\hat{\mathbf{v}}_H^\epsilon=0$, and we can remove the term involving f from the estimate since $\nabla\cdot\mathbf{v}_g=\nabla\cdot\mathbf{u}^0=f$. ∎

6.2. The oversampled nonconforming error. Chen and Hou [17, pp. 559–563] bounded the nonconforming error terms in Theorem 4.4 when $\mathbf{b}=0$. The key features needed in the analysis are (1) that the vector variable of the RT0 spaces, when restricted to an element and multiplied by a constant matrix, is a pure potential (i.e., a gradient of a scalar function), and (2) a vector variable $\bar{\mathbf{v}}_H$ in RT0 satisfies the estimate

$$\|\bar{\mathbf{v}}_H\|_{1,E}\le C\|\bar{\mathbf{v}}_H\|_{H(\mathrm{div};E)} \tag{6.11}$$

(see [17, (4.26)]). These properties hold only for RT0 on simplices.

The extension of their result to nonzero $\mathbf{b}$ is not difficult, and, again using the more pessimistic but proven bound on the homogenization boundary corrector terms (see Remark 5.1), the extended result follows.

LEMMA 6.5. *There is a constant $C>0$, independent of H and ϵ, such that for any $\psi_H\in\hat{\mathbf{V}}_{H,*}$,*

$$\begin{aligned}&\left|(\alpha\mathbf{u}-\mathbf{b},\psi_H)-\sum_{E\in\mathcal{T}_H}(p,\nabla\cdot\psi_H)_E\right|\\&\quad\le C\Big\{(\epsilon+\sqrt{\epsilon/H}+H)\big[\|f-\nabla\cdot\mathbf{v}_g\|_{0,\Omega}+\|\mathbf{v}_g\|_{0,\Omega}+\|\mathbf{b}\|_{0,\Omega}\big]\\&\qquad+(\epsilon+H)\big[\|\nabla p^0\|_{1,\Omega}+\|\nabla p^0-\mathbf{b}\|_{1,\Omega}\big]\\&\qquad+(\epsilon/H)\,\|\nabla p^0\|_{0,\infty,\Omega}+\sqrt{\epsilon/H}\,\|\nabla p^0-\mathbf{b}\|_{0,\infty,\Omega}\Big\}\|\psi_H\|_X.\end{aligned}$$

This completes the proof of Theorem 6.1.

7. Superconvergent multiscale estimation of the pressure error. In this section, we assume that oversampling is not used. In this case, we can significantly improve the estimate of the pressure error over that obtained in Theorem 6.1 above.

THEOREM 7.1. *For each $\epsilon>0$, let $(\mathbf{u}^\epsilon,p^\epsilon)\in(\mathbf{V}+\mathbf{v}_g)\times W$ be the solution of* (1.4)–(1.5)*, with the coefficient $a_\epsilon=a(x,x/\epsilon)$ and $\alpha_\epsilon=a_\epsilon^{-1}$. For each $H>0$, let $(\hat{\mathbf{u}}_H^\epsilon,\bar{p}_H^\epsilon)\in\hat{\mathbf{V}}_H\times\bar{W}_H$ be the solution of the nonoversampled ($E_*=E\ \forall E\in\mathcal{T}_H$) discrete upscaled equation* (2.26)–(2.27)*, and define $(\mathbf{u}_H^\epsilon,p_H^\epsilon)$ by* (2.29)–(2.30)*. Let $m=1$ when $\bar{\mathbf{V}}_H$ is the RT0 space and $m=1$ or 2 when $\bar{\mathbf{V}}_H$ is BDM1 or BDDF1. Assume that the domain Ω is k-regular, in the sense of* (7.3) *below. If $k=2$, then*

$$\|p^\epsilon-p_H^\epsilon\|_{0,\Omega}\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H\big)\|\mathbf{u}^\epsilon-\mathbf{u}_H^\epsilon\|_{0,\Omega},$$

and if $k = 3$, then

$$\|p^\epsilon - p_H^\epsilon\|_{-1,\Omega} \le C\big(\epsilon + (\epsilon/H)^{1/d-\eta} + H^m\big)\|\mathbf{u}^\epsilon - \mathbf{u}_H^\epsilon\|_{0,\Omega}.$$

These results display superconvergence, in that the pressure converges at a rate better that we would normally expect from approximation theory. Combining Theorems 7.1 and 6.1(b), we obtain for $d = 2$ that

$$\|p^\epsilon - p_H^\epsilon\|_{0,\Omega} \le C\big(\epsilon^2 + \epsilon/H + H^{m+1}\big). \tag{7.1}$$

One should compare this estimate to the L^2-estimate of Efendiev, Hou, and Wu [19] for the (nonmixed) multiscale finite element method:

$$\|p^\epsilon - p_H^\epsilon\|_{0,\Omega} \le C\big(\epsilon + \epsilon|\ln h| + (\epsilon/H)^2 + C_\theta \epsilon/H + H^2\big),$$

although numerical results suggest that C_θ is negligible.

Proof. The difference between (1.5) and the conforming, nonoversampled method (2.32) is

$$\big(\alpha(\mathbf{u} - \mathbf{u}_H), \hat{\mathbf{v}}_H\big) = (\bar{p} - \bar{p}_H, \nabla \cdot \bar{\mathbf{v}}_H) \quad \forall\ \hat{\mathbf{v}}_H \in \hat{\mathbf{V}}_H \tag{7.2}$$

(wherein we suppress the superscript ϵ on the solutions and the subscript ϵ on α). For $\varphi \in H^{k-2}(\Omega)$, we construct a test function from the solution $\mathbf{U} \in \mathbf{V}$ to the problem

$$\begin{aligned} \nabla \cdot \mathbf{U} &= \varphi \quad && \text{in } \Omega, \\ \mathbf{U} &= -a\nabla q \quad && \text{in } \Omega, \\ \mathbf{U} \cdot \nu &= 0 \quad && \text{on } \partial\Omega. \end{aligned}$$

This is the same problem as (1.1)–(1.3), with $f = \varphi$, $\mathbf{b} = 0$, and $g = 0$ (i.e., $\mathbf{v}_g = 0$). We solve this problem approximately with the variational multiscale method (2.31)–(2.32) of section 2 for $\mathbf{U}_H = \hat{\mathbf{U}}_H + \tilde{\mathbf{U}}' \in \hat{\mathbf{V}}_H + \tilde{\mathbf{U}}'$. Note that $\nabla \cdot \hat{\mathbf{U}}_H = \bar{\varphi}$, $\tilde{\mathbf{U}}' = -a\nabla\tilde{q}'$, and Theorem 6.1 imply that

$$\|\mathbf{U} - \mathbf{U}_H\|_{0,\Omega} \le C\big\{\big(\epsilon + (\epsilon/H)^{1/d-\eta}\big)\|\nabla q^0\|_{1,\Omega} + H^m\big(\|\mathbf{U}^0\|_{m,\Omega} + \|\varphi\|_{m-1,\Omega}\big)\big\},$$

where $(\mathbf{U}^0, q^0) \in \mathbf{V} \times W$ satisfies the corresponding homogenized problem (i.e., with a^0 replacing a). The k-regularity assumption means that there is some constant $C > 0$, independent of H and ϵ, such that

$$\|\mathbf{U}^0\|_{k-1,\Omega} + \|q^0\|_{k,\Omega} \le C\|\varphi\|_{k-2,\Omega}, \tag{7.3}$$

so

$$\|\mathbf{U} - \mathbf{U}_H\|_{0,\Omega} \le C\big(\epsilon + (\epsilon/H)^{1/d-\eta} + H^m\big)\|\varphi\|_{k-2,\Omega},$$

wherein $m = 1$ if $k = 2$.

Using the test function $\hat{\mathbf{U}}_H \in \hat{\mathbf{V}}_H$ in (7.2), we obtain that

$$\begin{aligned} (\bar{p} - \bar{p}_H, \bar{\varphi}) &= (\bar{p} - \bar{p}_H, \varphi) = \big(\alpha(\mathbf{u} - \mathbf{u}_H), \hat{\mathbf{U}}_H\big) \\ &= \big(\alpha(\mathbf{u} - \mathbf{u}_H), \mathbf{U}_H - \mathbf{U}\big) + \big(\alpha(\mathbf{u} - \mathbf{u}_H), \mathbf{U} - \tilde{\mathbf{U}}'\big). \end{aligned} \tag{7.4}$$

Now, by the divergence theorem,

$$\begin{aligned}\big(\alpha(\mathbf{u}-\mathbf{u}_H),\mathbf{U}-\tilde{\mathbf{U}}'\big) &= -\big(\alpha(\mathbf{u}-\mathbf{u}_H),a\nabla(q-\tilde{p}')\big)\\ &= -\big(\mathbf{u}-\mathbf{u}_H,\nabla(q-\tilde{p}')\big) = \big(\nabla\cdot(\mathbf{u}-\mathbf{u}_H),q-\tilde{p}'\big) = 0,\end{aligned}$$

so

$$\begin{aligned}(\bar{p}-\bar{p}_H,\varphi) &= \big(\alpha(\mathbf{u}-\mathbf{u}_H),\mathbf{U}_H-\mathbf{U}\big)\\ &\le C\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega}\|\mathbf{U}-\mathbf{U}_H\|_{0,\Omega}\\ &\le C\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega}\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H^m\big)\|\varphi\|_{k-2,\Omega},\end{aligned}$$

wherein $m=1$ if $k=2$.

Taking $k=2$ and the supremum over $\varphi\in L^2(\Omega)$, we see the estimate

$$\|\bar{p}-\bar{p}_H\|_{0,\Omega}\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H\big)\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega}.$$

If instead $k=3$ and $\varphi\in H_0^1(\Omega)$, we obtain

$$\|\bar{p}-\bar{p}_H\|_{-1,\Omega}\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H^m\big)\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega}.$$

Now by Lemma 4.1, we see that

$$\hat{p}'(\bar{\mathbf{u}}-\bar{\mathbf{u}}_H)=\hat{p}'\big(\bar{\mathbf{u}}+\hat{\mathbf{u}}'(\bar{\mathbf{u}})-\bar{\mathbf{u}}_H-\hat{\mathbf{u}}'(\bar{\mathbf{u}}_H)\big)=\hat{p}'(\mathbf{u}-\mathbf{u}_H),$$

so we conclude that

$$\begin{aligned}\|p-p_H\|_{0,\Omega} &\le \|\bar{p}-\bar{p}_H\|_{0,\Omega}+\|\hat{p}'(\mathbf{u}-\mathbf{u}_H)\|_{0,\Omega}\\ &\le \|\bar{p}-\bar{p}_H\|_{0,\Omega}+CH\|\nabla\hat{p}'(\mathbf{u}-\mathbf{u}_H)\|_{0,\Omega}\\ &\le \|\bar{p}-\bar{p}_H\|_{0,\Omega}+CH\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega}\\ &\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H\big)\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega},\end{aligned}$$

which is our first estimate. For the negative norm estimate,

$$\|\hat{p}'(\mathbf{u}-\mathbf{u}_H)\|_{-1,\Omega}\le CH^2\|\nabla\hat{p}'(\mathbf{u}-\mathbf{u}_H)\|_{0,\Omega}\le CH^2\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega},$$

so if $k=3$,

$$\begin{aligned}\|p-p_H\|_{-1,\Omega} &\le \|\bar{p}-\bar{p}_H\|_{-1,\Omega}+\|\hat{p}'(\mathbf{u}-\mathbf{u}_H)\|_{-1,\Omega}\\ &\le C\big(\epsilon+(\epsilon/H)^{1/d-\eta}+H^m\big)\|\mathbf{u}-\mathbf{u}_H\|_{0,\Omega},\end{aligned}$$

completing the proof. ∎

REFERENCES

[1] J. E. Aarnes, *On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation*, Multiscale Model. Simul., 2 (2004), pp. 421–439.

[2] J. E. Aarnes, S. Krogstad, and K.-A. Lie, *A hierarchical multiscale method for two-phase flow based upon mixed finite elements and nonuniform coarse grids*, Multiscale Model. Simul., 5 (2006), pp. 337–363.

[3] T. Arbogast, *Numerical subgrid upscaling of two-phase flow in porous media*, in Numerical Treatment of Multiphase Flows in Porous Media, Z. Chen, R. E. Ewing, and Z.-C. Shi, eds., Lecture Notes in Phys. 552, Springer-Verlag, Berlin, 2000, pp. 35–49.

[4] T. Arbogast, *Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase Darcy flow*, Comput. Geosci., 6 (2002), pp. 453–481.
[5] T. Arbogast, *Analysis of a two-scale, locally conservative subgrid upscaling for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 576–598.
[6] T. Arbogast and S. L. Bryant, *A two-scale numerical subgrid technique for waterflood simulations*, SPE J., 7 (2002), pp. 446–457.
[7] T. Arbogast, S. E. Minkoff, and P. T. Keenan, *An operator-based approach to upscaling the pressure equation*, in Computational Methods in Water Resources XII, Vol. 1: Computational Methods in Contamination and Remediation of Water Resources, V. N. Burganos et al., eds., Computational Mechanics Publications, Southampton, U.K., 1998, pp. 405–412.
[8] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.
[9] I. Babuška, G. Caloz, and J. E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.
[10] I. Babuška and J. E. Osborn, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.
[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[12] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationelle Sér. Rouge, 8 (1974), pp. 129–151.
[13] F. Brezzi, *Interacting with the subgrid world*, in Numerical Analysis 1999, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp 69–82.
[14] F. Brezzi, J. Douglas, Jr., R. Durán, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.
[15] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[17] Z. Chen and T. Y. Hou, *A mixed multiscale finite element method for elliptic problems with oscillating coefficients*, Math. Comp., 72 (2003), pp. 541–576.
[18] J. Douglas, Jr. and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.
[19] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.
[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[21] T. Y. Hou and X. H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[22] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[23] T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.
[24] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.
[25] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functions*, Springer-Verlag, New York, 1994.
[26] G. M. Lieberman, *Oblique derivative problems in Lipschitz domains* II. *Discontinuous boundary data*, J. Reine Angew. Math., 389 (1988), pp. 1–21.
[27] R. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, I. Galligani and E. Magenes, eds., Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1172–1190

SEMI-IMPLICIT EULER SCHEME FOR GENERALIZED NEWTONIAN FLUIDS*

LARS DIENING†, ANDREAS PROHL‡, AND MICHAEL RŮŽIČKA†

Abstract. Rheological behavior of certain non-Newtonian fluids in engineering sciences is often modeled by power law ansatzes with $p \leq 2$. So far, existing numerical analysis for local strong solutions studies a fully implicit time discretization and find only restricted ranges of admissible p's for corresponding error estimates [A. Prohl and M. Růžička, *SIAM J. Numer. Anal.*, 39 (2001), pp. 214–249]; different nonlinear stabilization strategies which allow a corresponding error analysis for smaller p's are examined in [L. Diening, *Theoretical and Numerical Results for Electrorheological Fluids*, Ph.D. thesis, University of Freiburg, Freiburg, Germany, 2002] and [L. Diening, A. Prohl, and M. Růžička, in *Nonlinear Problems in Mathematical Physics and Related Topics*, II, Kluwer/Plenum, New York, 2002, pp. 89–118]. In the present paper, a semi-implicit time discretization scheme is proposed, and error estimates apply to the extended range $p \in (\frac{3}{2}, 2]$. The key analytical tool is a new Gronwall-type inequality.

Key words. non-Newtonian fluid flow, degenerate parabolic system, time discretization, weak and strong solution, shear-dependent viscosity, error analysis

AMS subject classifications. 76A05, 35K65, 35B65, 65M06, 65M12, 65M15,

DOI. 10.1137/050634335

1. Introduction and main results.

Viscous fluids that cannot be adequately described by the classical linearly viscous fluid model are usually called non-Newtonian fluids. There are many fluids which differ from a Newtonian fluid only in that their viscosity depends on the shear rate, i.e., on the modulus of the symmetric part of the velocity gradient. Such fluids are called *fluids with shear-dependent viscosity* or *generalized Newtonian fluids*. We refer to [2], [4], [10], [12], [14], and [17] for a detailed discussion of the modeling and the engineering relevance of such fluids.

A typical example of a constitutive relation for the extra stress tensor $\mathbf{S}$ of a *generalized Newtonian fluid* is

$$\mathbf{S}(\mathbf{D}) = \mu \left(\kappa + |\mathbf{D}|^2\right)^{\frac{p-2}{2}} \mathbf{D},$$

where $\mu > 0$, $\kappa \geq 0$, and $p \in (1, \infty)$ are some given material constants. Note that in the case $p \in (1, 2)$ the model reflects shear-thinning behavior, while $p \in (2, \infty)$ corresponds to shear-thickening behavior. For $p = 2$ the model reduces to the Newtonian one.

In this paper we abstract from the specific form of the constitutive relation of the extra stress tensor $\mathbf{S}$ but make the following assumptions: We assume that the extra stress tensor $\mathbf{S}$ has *p-structure*; i.e., for the constitutive function $\mathbf{S} : \mathbb{R}^{3\times3}_{\text{sym}} \to \mathbb{R}^{3\times3}_{\text{sym}}$, with $\mathbb{R}^{3\times3}_{\text{sym}} := \{\mathbf{D} \in \mathbb{R}^{3\times3}; \mathbf{D} = \mathbf{D}^\top\}$, there exist $p > 1$ and $C_1, C_2 > 0$ such that for

*Received by the editors June 23, 2005; accepted for publication (in revised form) February 3, 2006; published electronically June 21, 2006.

http://www.siam.org/journals/sinum/44-3/63433.html

†Mathematisches Institut, Abteilung für Angewandte Mathematik, Universität Freiburg, Eckerstraße 1, D-79104 Freiburg, Germany (diening@mathematik.uni-freiburg.de, rose@mathematik.uni-freiburg.de). The first and third authors were supported by the DFG–Forschergruppe "Nonlinear Partial Differential Equations: Theoretical and Numerical Analysis."

‡Department of Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland (apr@math.ethz.ch).

all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ it holds that

$$\sum_{i,j,k,l=1}^{3} \frac{\partial S_{ij}(\mathbf{A})}{\partial A_{kl}} B_{ij} B_{kl} \geq C_1 \left(1 + |\mathbf{A}|\right)^{p-2} |\mathbf{B}|^2 \tag{1.1}$$

and for all $i, j, k, l = 1, 2, 3$

$$\left| \frac{\partial S_{ij}(\mathbf{A})}{\partial A_{kl}} \right| \leq C_2 \left(1 + |\mathbf{A}|\right)^{p-2}. \tag{1.2}$$

We consider only the case $p \in (1, 2]$, i.e., the shear-thinning and the Newtonian model. Moreover, in our investigation of the governing system for such fluids we restrict ourselves to space-periodic boundary conditions. From the physical point of view this can be viewed only as a model case. However, in the case of Dirichlet boundary conditions in three-dimensional bounded domains, appropriate existence, uniqueness, and regularity results are only partially available. For example, the existence of weak solutions is known for $p > \frac{8}{5}$ (cf. [8], [9], [11], [18]). However, regularity properties of solutions and their uniqueness are known only for $p > \frac{20}{9}$ (cf. [11], [1]), and the results are much weaker than the corresponding ones in the space-periodic case (cf. Proposition 1.1). Moreover, we do not wish to burden the already complicated analysis with further technical difficulties. In fact all the difficulties which appear in the investigation of the continuous problem will also appear in the numerical analysis. Finally, we restrict ourselves to the three-dimensional case.

Now we will state precisely the problem in which we are interested. Let $\Omega = (0, L)^3$, $L \in (0, \infty)$, be a cube in $\mathbb{R}^3$. Let us denote $\Gamma_j = \partial\Omega \cap \{x_j = 0\}$ and $\Gamma_{j+3} = \partial\Omega \cap \{x_j = L\}$ for $j = 1, 2, 3$. For $T \in (0, \infty)$, we denote by Q_T the time-space cylinder $I \times \Omega$, where $I = (0, T)$ is the time interval. Assume that $\mathbf{S}$ satisfies assumptions (1.1), (1.2). For a given external body force $\mathbf{f} : Q_T \to \mathbb{R}^3$ and a given initial velocity $\mathbf{u}_0 : \Omega \to \mathbb{R}^3$ we seek a velocity field $\mathbf{u} = (u_1, u_2, u_3)^\top : Q_T \to \mathbb{R}^3$ and a pressure function $\pi \colon \Omega \to \mathbb{R}$ solving the system

$$\begin{aligned} \partial_t \mathbf{u} - \operatorname{div} \mathbf{S}(\mathbf{D}\mathbf{u}) + [\nabla \mathbf{u}]\mathbf{u} + \nabla \pi &= \mathbf{f} && \text{in } Q_T, \\ \operatorname{div} \mathbf{u} &= 0 && \text{in } Q_T, \\ \mathbf{u}(0) &= \mathbf{u}_0 && \text{on } \Omega, \end{aligned} \tag{1.3}$$

and satisfying the space-periodicity requirements

$$\mathbf{u}\big|_{\Gamma_j} = \mathbf{u}\big|_{\Gamma_{j+3}}, \qquad \nabla \mathbf{u}\big|_{\Gamma_j} = \nabla \mathbf{u}\big|_{\Gamma_{j+3}}, \qquad \pi\big|_{\Gamma_j} = \pi\big|_{\Gamma_{j+3}} \tag{1.4}$$

for $j = 1, 2, 3$. The term $\mathbf{D}\mathbf{u} := \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^\top)$ denotes the symmetric part of the velocity gradient $\nabla \mathbf{u}$. We refer to (1.3), (1.4) as problem (NS_p).

This paper studies the following time discretization of (NS_p).

ALGORITHM 1. Given a time step size $k > 0$ and the corresponding net $I_k = \{t_m\}_{m=0}^{M}$, for $m \geq 1$ and $\mathbf{u}^{m-1}$ given from the previous step, compute an iterate $\mathbf{u}^m$ that solves

$$\begin{aligned} d_t \mathbf{u}^m - \operatorname{div} \mathbf{S}(\mathbf{D}\mathbf{u}^m) + [\nabla \mathbf{u}^m]\mathbf{u}^{m-1} + \nabla \pi^m &= \mathbf{f}(t_m), \\ \operatorname{div} \mathbf{u}^m &= 0, \\ \mathbf{u}^0 &= \mathbf{u}_0 \end{aligned} \tag{NS_p^k}$$

endowed with space-periodic boundary conditions (1.4), where we denote by $d_t\mathbf{u}^m := k^{-1}(\mathbf{u}^m - \mathbf{u}^{m-1})$ the divided difference in time. We assume that $\mathbf{S}$ has p-structure, i.e., it satisfies (1.1) and (1.2).

To our knowledge, numerical analysis for the full model (NS_p) starts with [13, 4], where error estimates for a first order fully implicit space-time discretization of (NS_p) are derived in the context of locally existing strong solutions; new tools had to be developed to efficiently control errors—albeit surprisingly only valid for restricted values of p. This observation motivated nonlinear stabilization strategies in [3, 4] which significantly extended the range of admissible p's. Proper strategies here are to add a q–Laplacian-type operator ($q \geq 2$) to the problem (scaled with the discretization parameter), or substitute p-growth of the underlying functional $\Phi : \mathbb{R} \to \mathbb{R}$ at a numerical threshold by a quadratic one; the motivation for these strategies is to "strengthen" the dissipative mechanism in the scheme for small values of p in relation to the convective term.

The goal of this paper is to show a similar effect for the scheme (NS_p^k) that is due to its semi-implicit character—with no need for additional stabilizing terms. The key step in our analysis is a new lemma of Gronwall-type (see Lemma 3.2). The verified rate of convergence with respect to the step size $k > 0$ will be the same as for the (more expensive) fully implicit stabilized schemes. The range of admissible p's for the derivation of the error estimates and for the existence of strong solutions will even be extended to $p > \frac{3}{2}$.

The existence of *strong* solutions locally in time for large data of the problem (NS_p) is ensured by the following assertion.[1]

PROPOSITION 1.1. *Let* $\mathbf{f} \in L^\infty(I, W^{1,2}(\Omega))$, $\partial_t\mathbf{f} \in L^2(I, L^2(\Omega))$, $\mathbf{u}_0 \in W^{2,2}_{\mathrm{div}}(\Omega)$, *and* $\frac{7}{5} < p \leq 2$. *Then there exists a time interval* $I = (0, T)$, $T > 0$, *and a unique solution* $\mathbf{u}$, π *which satisfies for all* $1 \leq r < 6(p-1)$

$$\begin{aligned} &\mathbf{u} \in C(I, W^{1,r}(\Omega)), \\ &\partial_t\mathbf{u} \in L^\infty(I, L^2(\Omega)) \cap L^p(I, L^{3p}(\Omega)), \\ &\partial_t^2\mathbf{u} \in L^2(I, (W^{1,2}_{\mathrm{div}}(\Omega))^*). \end{aligned} \tag{1.5}$$

The proof of this proposition for $p \in (7/5, 2]$ can be found in [3] and [5] under the additional assumption that $\mathbf{S}$ is given by a potential. However, this fact is never used in these papers, and the result also extends to the situation without a potential (cf. [16] where the case $p \in (3/2, 2]$ is treated under the assumptions (1.1) and (1.2) only). The case $p > 5/3$ is already covered in [10]. In [13] the existence of a weak solution $\mathbf{u}^m \in l^\infty(I_k; L^2(\Omega)) \cap l^p(I_k; W^{1,p}_{\mathrm{div}})$ of the fully implicit time discretization of (NS_p) is proved for $p > 3/2$, but the analysis applies to (NS_p^k) as well with only minor changes. In order to analyze the above algorithm for all $p > 3/2$ we first derive suboptimal convergence rates for the error between the strong solution $\mathbf{u}$ from Proposition 1.1 and the unique weak solution $\mathbf{u}^m$ of the discrete problem (NS_p^k) from Lemma 3.1. In fact, we show with the help of a new Gronwall-type lemma that the following proposition holds.

PROPOSITION 1.2. *Suppose that* $\mathbf{u}_0 \in W^{2,2}(\Omega) \cap W^{1,p}_{\mathrm{div}}$, $\mathbf{f} \in C(I; W^{1,2}(\Omega))$, $\partial_t\mathbf{f} \in C(I; L^2(\Omega))$ *are given. Let* $\mathbf{u}$ *be the strong solution of problem* (NS_p) *for* $p \in (\frac{3}{2}, 2]$ *satisfying* (1.5) *and let* $\mathbf{u}^m$ *be the unique weak solution of problem* (NS_p^k) *satisfying* (3.1) *and* $t_M \leq T$. *Then the following error estimate is valid provided that the time*

[1] We use standard notation for Lebesgue, Sobolev, and Bochner spaces, which will be defined precisely in section 2.

step size k is chosen sufficiently small, i.e., $k \le k_1(p,T)$:

$$\max_{0\le m\le M} \|\mathbf{u}(t_m)-\mathbf{u}^m\|_2^2 + k\sum_{m=0}^{M} \|\nabla(\mathbf{u}(t_m)-\mathbf{u}^m)\|_p^2 \le c_1\, k^{2\beta}\,, \qquad \max_{0\le m\le M} \|\nabla(\mathbf{u}(t_m)-\mathbf{u}^m)\|_p \le 1, \tag{1.6}$$

with $c_1 = c_1(p,T,\mathbf{u}_0,\mathbf{f})$ and

$$\beta := \frac{5p-6}{2p}\,.$$

Then we use this error estimate to show by an induction argument that the weak solution of problem (NS_p^k) is actually a strong one. Namely, we show that the following theorem holds.

THEOREM 1.3. *Let $\mathbf{u}_0$, $\mathbf{f}$, p, $\mathbf{u}$, $\mathbf{u}^m$, T, and t_M be as in Proposition* 1.2. *Then*

$$\max_{1\le m\le M} \|d_t\mathbf{u}^m\|_2^2 + k\sum_{m=1}^{M} \Big(\mathcal{I}_p(\mathbf{u}^m)^{\frac{5p-6}{2-p}} + \mathcal{K}_p(\mathbf{u}^m)\Big) \le c(\mathbf{f},\mathbf{u}_0), \tag{1.7}$$

where $\mathcal{I}_p$ and $\mathcal{K}_p$ are defined below by (2.1) *and* (2.2). *In particular, due to Lemma* 2.3 *for all $1<r<6(p-1)$ we have*

$$\begin{aligned} \mathbf{u}^m &\in l^{\frac{5p-6}{2-p}}(I_k; W^{2,\frac{3p}{p+1}}(\Omega)) \cap l^\infty(I_k; V_r),\\ d_t\mathbf{u}^m &\in l^{\frac{p(5p-6)}{(3p-2)(p-1)}}(I_k; W^{1,\frac{3p}{p+1}}(\Omega)) \cap l^\infty(I_k; L^2(\Omega)). \end{aligned} \tag{1.8}$$

Now using this regularity we can improve the convergence rate from Proposition 1.2 and show that the following theorem holds.

THEOREM 1.4. *Let $\mathbf{u}_0$, $\mathbf{f}$, p, $\mathbf{u}$, $\mathbf{u}^m$, T, and t_M be as in Proposition* 1.2. *Then for all*

$$\alpha < \alpha_0(p) := \frac{5p-6}{4(p-1)},$$

there exists a constant $c_2 = c_2(p,T,\mathbf{u}_0,\mathbf{f},\alpha)$, such that the following error estimate is valid for k chosen sufficiently small, i.e., $k \le k_2(p,T,\alpha)$:

$$\max_{0\le m\le M} \|\mathbf{u}(t_m)-\mathbf{u}^m\|_2^2 + k\sum_{m=0}^{M} \|\nabla(\mathbf{u}(t_m)-\mathbf{u}^m)\|_p^2 \le c_2\, k^{2\alpha}\,. \tag{1.9}$$

REMARK 1.5. *Theorem* 1.4 *improves Theorem* 1.10 *in* [4] *considerably, both with respect to the range of admissible p's and the regularity of the solution $\mathbf{u}^m$. In* [13] *it is proved that for $p \in (\frac{11+\sqrt{21}}{10}, 2] \approx (1.5583, 2]$ estimate* (1.9) *holds. However, note that the discrete solution $\mathbf{u}^m$ in* [4] *is only a weak solution; i.e., only* (3.1) *holds. The regularity of $\mathbf{u}^m$ ensured by Theorem* 1.3 *is proved in* [4] *only for stabilized schemes. For a subsequent analysis of a spatial discretization it is shown in* [13] *that the existence and characterization of strong solutions to (NS_p^k) are essential. For example, uniform a priori bounds in* [13] *in a comparable situation are obtained only for restricted values $p \in (\frac{9}{5}, 2]$. Using the results in this paper one can carry out the*

analysis of the fully discrete system based on (NS_p^k) *along the lines of* [13]. *Within the framework of the DFG–Forschergruppe "Nonlinear Partial Differential Equations: Theoretical and Numerical Analysis" we currently develop a robust and efficient solver for problems with p-structure. Part of the research is a comparison of semi-implicit and fully implicit schemes. However, at the moment results are not yet available.*

The remainder of this paper is organized as follows. Section 2 provides the mathematical setup to study (NS_p^k) and collects the consequences of our assumptions (1.1), (1.2). In section 3 first the existence of weak solutions is verified and then the proof of Proposition 1.2 is given, where Lemma 3.2 is the main tool. Section 4 presents proofs of Theorems 1.3 and 1.4.

2. Notation and technical preliminaries. In this section we fix the notation and collect some useful consequences of the assumptions (1.1), (1.2). Recall that $\Omega = (0,L)^3$, $L \in (0,\infty)$, is a cube in $\mathbb{R}^3$. By $\mathcal{D}(\Omega)$ we denote the space of smooth periodic functions with mean value zero. Let further $p, q > 1$ and $k > 0$. Then $(L^p(\Omega), \|\cdot\|_p)$ (respectively, $(W^{k,p}(\Omega), \|\cdot\|_{k,p})$) is used for the usual Lebesgue (respectively, Sobolev) spaces of periodic functions with mean value zero. We will further make frequent use of spaces of divergence-free functions defined by

$$\mathcal{V} := \big\{\boldsymbol{\psi} \in \mathcal{D}(\Omega) : \operatorname{div} \boldsymbol{\psi} = 0\big\},$$
$$W^{1,p}_{\mathrm{div}} := \text{the closure of } \mathcal{V} \text{ with respect to the } \|\nabla\cdot\|_p\text{-norm}.$$

By $\langle g, h\rangle$ we denote the scalar product $\int_\Omega g(x)\,h(x)\,dx$. For two Banach spaces X_0, X_1 and $\theta \in (0,1)$ the complex interpolation space is $[X_0, X_1]_{[\theta]}$. Moreover, we denote by $L^q(I;X)$ Bochner spaces which are equipped with the norm $\big(\int_I \|\cdot\|_X^q\,ds\big)^{1/q}$. We refer the reader to [7] for more details. We make frequent use of the discrete counterparts of these spaces. Let $I_k = \{t_m\}_{m=0}^M$ be a given net in an interval $I = [0, t_M]$ with a constant time step size $k := t_m - t_{m-1}$. We denote by $d_t\mathbf{u}^m := k^{-1}(\mathbf{u}^m - \mathbf{u}^{m-1})$ the divided difference in time. By $l^p(I_k;X)$ we denote the space of functions $\{\varphi^m\}_{m=0}^M$ with finite norm $\big(k\sum_{m=0}^M \|\varphi^m\|_X^p\big)^{1/p}$. In the case $p=\infty$, functions $\{\varphi^m\}_{m=0}^M$ need to satisfy the bound $\max_{0\le m\le M}\|\varphi^m\|_X < \infty$.

Let us introduce some notation for terms which arise from $\mathbf{S}$ when we test (NS_p) with $-\Delta\mathbf{u}$ or with $\partial_t^2\mathbf{u}$. Namely, for $p > 1$ we set

$$\begin{aligned}\mathcal{I}_p(\mathbf{u}) &= \int_\Omega \big(1+|\mathbf{D}\mathbf{u}|^2\big)^{\frac{p-2}{2}} |\mathbf{D}(\nabla\mathbf{u})|^2\,dx,\\ \mathcal{J}_p(\mathbf{u}) &= \int_\Omega \big(1+|\mathbf{D}\mathbf{u}|^2\big)^{\frac{p-2}{2}} |\mathbf{D}(\partial_t\mathbf{u})|^2\,dx.\end{aligned} \tag{2.1}$$

The discrete analogue for $\mathcal{J}_p(\mathbf{u})$ for a function defined on a net I_k reads as follows:

$$\mathcal{K}_p(\mathbf{u}^m) = \int_\Omega \left(1+\frac{1}{2}|\mathbf{D}\mathbf{u}^m|^2 + \frac{1}{2}|\mathbf{D}\mathbf{u}^{m-1}|^2\right)^{\frac{p-2}{2}} |\mathbf{D}(d_t\mathbf{u}^m)|^2\,dx. \tag{2.2}$$

Let us now summarize some important estimates for $\mathbf{S}$, $\mathcal{I}_p$, $\mathcal{J}_p$, and $\mathcal{K}_p$. The proofs can be found, for example, in [10, Lemmas 5.1.19 and 5.1.35] and [13, Lemma 2.8].

LEMMA 2.1. *Suppose that* Φ *and* $\mathbf{S}$ *satisfy* (1.1), (1.2) *for some* $p > 1$. *Then there are constants* $c_3 = c_3(p)$ *and* $c_4 = c_4(p)$ *such that for all* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$

$$\mathbf{S}(\mathbf{A})\cdot\mathbf{A} \ge c_3\big(1+|\mathbf{A}|\big)^{p-2}|\mathbf{A}|^2, \tag{2.3}$$

$$|\mathbf{S}(\mathbf{A})| \le c_4\big(1+|\mathbf{A}|\big)^{p-1}. \tag{2.4}$$

Additionally, we have

$$\big(\mathbf{S}(\mathbf{A})-\mathbf{S}(\mathbf{B})\big)\cdot\big(\mathbf{A}-\mathbf{B}\big)\geq c_3\,\big|\mathbf{A}-\mathbf{B}\big|^2\big(1+|\mathbf{B}|+|\mathbf{A}-\mathbf{B}|\big)^{p-2}, \tag{2.5}$$

$$\big|\mathbf{S}(\mathbf{A})-\mathbf{S}(\mathbf{B})\big|\leq c_4\,|\mathbf{A}-\mathbf{B}|\big(1+|\mathbf{B}|+|\mathbf{A}-\mathbf{B}|\big)^{p-2}. \tag{2.6}$$

REMARK 2.2. *From* (2.5) *it follows that for* $r\in[1,\infty)$ *and* $p\in(1,2]$

$$\begin{aligned}&\int_\Omega(\mathbf{S}(\mathbf{Du})-\mathbf{S}(\mathbf{Dv}))\cdot\mathbf{D}(\mathbf{u}-\mathbf{v})\,dx\\&\quad\geq c_5\,\|\mathbf{Du}-\mathbf{Dv}\|^2_{\frac{2r}{2-p+r}}(1+\|\mathbf{Du}\|_r+\|\mathbf{Du}-\mathbf{Dv}\|_r)^{p-2}.\end{aligned} \tag{2.7}$$

The following lemmas are proved in [4].

LEMMA 2.3. *Let* $\mathbf{u}\in C^1(I;\mathbf{C}^2(\Omega))$ *be a space periodic function with mean value zero and* $p\in(1,2]$. *Then there exists a constant* c *depending only on* Ω *and* p *such that for* $s\in[1,\infty)$

$$\|\nabla\mathbf{u}\|^2_{\frac{6s}{6-3p+s}}+\|\nabla^2\mathbf{u}\|^2_{\frac{2s}{2-p+s}}\leq c\,\mathcal{I}_p(\mathbf{u})\big(1+\|\nabla\mathbf{u}\|_s\big)^{2-p}, \tag{2.8}$$

$$\|\nabla\mathbf{u}\|^p_{3p}+\|\nabla^2\mathbf{u}\|^p_{\frac{3p}{p+1}}\leq c\,(1+\mathcal{I}_p(\mathbf{u})), \tag{2.9}$$

$$\|\partial_t\mathbf{u}\|^2_{\frac{6s}{6-3p+s}}+\|\nabla\partial_t\mathbf{u}\|^2_{\frac{2s}{2-p+s}}\leq c\,\mathcal{J}_p(\mathbf{u})(1+\|\nabla\mathbf{u}\|_s)^{2-p}, \tag{2.10}$$

$$\begin{aligned}\|\partial_t\mathbf{u}\|^p_{3p}+\|\nabla\partial_t\mathbf{u}\|^p_{\frac{3p}{p+1}}&\leq c\,(1+\mathcal{I}_p(\mathbf{u}))^{\frac{2-p}{2}}\mathcal{J}_p(\mathbf{u})^{\frac{p}{2}}\\&\leq c\,(1+\mathcal{I}_p(\mathbf{u})+\mathcal{J}_p(\mathbf{u})).\end{aligned} \tag{2.11}$$

Moreover, for $1\leq r<6(p-1)$ *we have*

$$\sup_{t\in I}\|\nabla\mathbf{u}\|^p_r\leq c\bigg(1+\int_I\mathcal{I}_p(\mathbf{u})^{\frac{5p-6}{2-p}}+\mathcal{J}_p(\mathbf{u})\,dt\bigg). \tag{2.12}$$

LEMMA 2.4. *Let* $\mathbf{u}\in l^\infty(I_k;\mathbf{C}^2(\Omega))$ *be a space-periodic function with mean value zero, and let* $p\in(1,2]$. *Then there exists a constant* c *depending only on* Ω *and* p *such that for* $s\in[1,\infty)$

$$\begin{aligned}&\|d_t\mathbf{u}^m\|^2_{\frac{6s}{6-3p+s}}+\|d_t\nabla\mathbf{u}^m\|^2_{\frac{2s}{2-p+s}}\\&\qquad\leq c\,\mathcal{K}_p(\mathbf{u}^m)(1+\|\nabla\mathbf{u}^m\|_s+\|\nabla\mathbf{u}^{m-1}\|_s)^{2-p},\end{aligned} \tag{2.13}$$

$$\begin{aligned}&\|d_t\mathbf{u}^m\|^p_{3p}+\|d_t\nabla\mathbf{u}^m\|^p_{\frac{3p}{p+1}}\\&\qquad\leq c\,(1+\mathcal{I}_p(\mathbf{u}^m)+\mathcal{I}_p(\mathbf{u}^{m-1}))^{\frac{2-p}{2}}\mathcal{K}_p(\mathbf{u}^m)^{\frac{p}{2}}\end{aligned} \tag{2.14}$$

$$\leq c\,(1+\mathcal{I}_p(\mathbf{u}^m)+\mathcal{I}_p(\mathbf{u}^{m-1})+\mathcal{K}_p(\mathbf{u}^m)). \tag{2.15}$$

Moreover, for $1\leq r<6(p-1)$

$$\max_{1\leq m\leq M}\|\nabla\mathbf{u}^m\|^p_r\leq c(r)\left(1+k\sum_{m=1}^M(\,\mathcal{I}_p(\mathbf{u}^m)^{\frac{5p-6}{2-p}}+\mathcal{I}_p(\mathbf{u}^{m-1})^{\frac{5p-6}{2-p}}+\mathcal{K}_p(\mathbf{u}^m)\,)\right). \tag{2.16}$$

3. Proof of Proposition 1.2. In this section we will present a proof of Proposition 1.2; i.e., we will derive preliminary estimates for the error $\mathbf{u}(t_m) - \mathbf{u}^m$. Before we do so, let us offer a few words on the existence of the solution $\mathbf{u}^m$ to the system (NS_p^k). The strategy employed in the proof of Proposition 1.1 to ensure the existence of strong solutions is not applicable to the discrete system (NS_p^k). However, the existence of *weak* solutions to the fully discrete analogue of problem (NS_p^k) is ensured, e.g., by [13, Lemma 4.1] which we recall here.

LEMMA 3.1. *Let* $\mathbf{u}_0$ *and* $\mathbf{f}$ *be as in Proposition* 1.1*, and let* $p > 3/2$*. Then there exists a unique, weak solution* $\mathbf{u}^m$ *of the problem* (NS_p^k) *satisfying*

$$\max_{0\le m\le M} \|\mathbf{u}^m\|_2^2 + k\sum_{m=0}^{M} \|\mathbf{D}\mathbf{u}^m\|_p^p \le c(\mathbf{f},\mathbf{u}_0), \tag{3.1}$$

whenever $p > 3/2$.

The proof of Lemma 3.1 is based on the fact that each $\mathbf{u}^m$ is just the solution to a stationary Stokes-like problem which is then solved by the techniques developed in [6] and [15]. Lemma 3.1 was originally developed for the fully implicit time discretization, i.e., with convective term $[\nabla\mathbf{u}^m]\mathbf{u}^m$, without the statement for uniqueness. In our case of a semi-implicit time discretization, i.e., with convective term $[\nabla\mathbf{u}^m]\mathbf{u}^{m-1}$, the proof could even be simplified, since the convective term is linear in $\mathbf{u}^m$. The same reason ensures that the solution $\mathbf{u}^m$ for each time step is unique.

Proof of Proposition 1.2. Let us introduce some notation. We set $\mathbf{e}^0 = \mathbf{0}$ and we define for $1 \le m \le M$

$$\mathbf{e}^m := \mathbf{u}(t_m) - \mathbf{u}^m, \qquad \eta^m := \pi(t_m) - \pi^m, \qquad \mathbf{R}^m := d_t\mathbf{u}(t_m) - \partial_t\mathbf{u}(t_m).$$

It has been shown in [13] and [4] under the assumptions (1.5) that $\mathbf{R}^m$ is well defined and satisfies

$$\begin{aligned} \|\mathbf{R}^m\|_{l^2(I_k;(W^{1,2}_{\mathrm{div}}(\Omega))^*)} &\le c\,k\,\|\partial_t^2\mathbf{u}\|_{L^2(I;(W^{1,2}_{\mathrm{div}}(\Omega))^*)},\\ \|\mathbf{R}^m\|_{l^\infty(I_k;L^2(\Omega))} &\le c\,\|\partial_t\mathbf{u}\|_{L^\infty(I;L^2(\Omega))}. \end{aligned}$$

Especially, by (1.5) we have

$$\|\mathbf{R}^m\|_{l^2(I_k;(W^{1,2}_{\mathrm{div}}(\Omega))^*)} \le c\,k, \qquad \|\mathbf{R}^m\|_{l^\infty(I_k;L^2(\Omega))} \le c. \tag{3.2}$$

With this new notation, system (NS_p) reads at time $t_m > 0$ as follows:

$$\begin{aligned} d_t\mathbf{u}(t_m) - \operatorname{div}\mathbf{S}(\mathbf{D}\mathbf{u}(t_m)) + [\nabla\mathbf{u}(t_m)]\mathbf{u}(t_m) + \nabla\pi(t_m) &= \mathbf{f}(t_m) + \mathbf{R}^m,\\ \operatorname{div}\mathbf{u}(t_m) &= 0,\\ \mathbf{u}(0) &= \mathbf{u}_0, \end{aligned}$$

with $1 \le m \le M$. This and (NS_p^k) imply for $1 \le m \le M$ that

$$\begin{aligned} &d_t\mathbf{e}^m - \operatorname{div}(\mathbf{S}(\mathbf{D}\mathbf{u}(t_m)) - \mathbf{S}(\mathbf{D}\mathbf{u}^m)) + \nabla\eta^m\\ &\quad = -k\,[\nabla\mathbf{u}(t_m)]d_t\mathbf{u}(t_m) - [\nabla\mathbf{u}(t_m)]\mathbf{e}^{m-1} - [\nabla\mathbf{e}^m]\mathbf{u}^{m-1} + \mathbf{R}^m,\\ &\operatorname{div}\mathbf{e}^m = 0,\\ &\mathbf{e}^0 = 0. \end{aligned} \tag{3.3}$$

We use the test function $\mathbf{e}^m$ for this system and get

$$\begin{aligned}&\big\langle d_t\mathbf{e}^m,\mathbf{e}^m\big\rangle+\big\langle\mathbf{S}(\mathbf{Du}(t_m))-\mathbf{S}(\mathbf{Du}^m),\mathbf{De}^m\big\rangle\\&\quad\le k\,\big|\big\langle[\nabla\mathbf{u}(t_m)]d_t\mathbf{u}(t_m),\mathbf{e}^m\big\rangle\big|+\big|\big\langle[\nabla\mathbf{u}(t_m)]\mathbf{e}^{m-1},\mathbf{e}^m\big\rangle\big|+\big|\langle\mathbf{R}^m,\mathbf{e}^m\rangle\big|\\&\quad=:K_1^m+K_2^m+K_3^m,\end{aligned}\tag{3.4}$$

where we have used that $\langle\nabla\eta^m,\mathbf{e}^m\rangle=0$ and $\langle[\nabla\mathbf{e}^m]\mathbf{u}^{m-1},\mathbf{e}^m\rangle=0$ since $\operatorname{div}(\mathbf{u}^{m-1})=\operatorname{div}(\mathbf{u}^m)=0$. With $\langle d_t\mathbf{e}^m,\mathbf{e}^m\rangle=\frac12 d_t\|\mathbf{e}^m\|_2^2+\frac{k}{2}\|d_t\mathbf{e}^m\|_2^2$, Remark 2.2, with $r=p$, and Korn's inequality we get

$$\tfrac12 d_t\|\mathbf{e}^m\|_2^2+\frac{c_5\,\|\nabla\mathbf{e}^m\|_p^2}{(1+\|\nabla\mathbf{u}(t_m)\|_p+\|\nabla\mathbf{e}^m\|_p)^{2-p}}\le c\,\big(K_1^m+K_2^m+K_3^m\big).$$

Since $\mathbf{u}\in C(I;W^{1,p}(\Omega))$ by (1.5), this implies the existence of a constant $c_6>0$ such that

$$\tfrac12 d_t\|\mathbf{e}^m\|_2^2+\frac{c_6\,\|\nabla\mathbf{e}^m\|_p^2}{(1+\|\nabla\mathbf{e}^m\|_p)^{2-p}}\le c\,\big(K_1^m+K_2^m+K_3^m\big).\tag{3.5}$$

We will now estimate K_1^m, K_2^m, and K_3^m in such a way that we can apply Lemma 3.2 below with

$$\begin{aligned}a_m&:=\|\mathbf{e}^m\|_2,\\b_m&:=\|\nabla\mathbf{e}^m\|_p,\\r_m&:=c\,\big(\|\mathbf{R}^m\|_{-1,p'}+k\,\|d_t\mathbf{u}(t_m)\|_{\frac{12}{8-3p}}\big),\\\beta&:=\frac{5p-6}{2p}.\end{aligned}\tag{3.6}$$

Let us first show that r_m satisfies (3.14). From the embedding ($p\le2$)

$$\begin{aligned}[l^2(I_k;W^{-1,2}(\Omega)),l^\infty(I_k;L^2(\Omega))]_{[\frac{5p-6}{2p}]}&\hookrightarrow[l^2(I_k;W^{-1,2}(\Omega)),l^\infty(I_k;W^{-1,6}(\Omega))]_{[\frac{5p-6}{2p}]}\\&\hookrightarrow l^{\frac{4p}{5p-6}}(I_k;W^{-1,p'}(\Omega))\\&\hookrightarrow l^2(I_k;W^{-1,p'}(\Omega))\end{aligned}$$

and (3.2) it follows that

$$\begin{aligned}\|\mathbf{R}^m\|_{l^2(I_k;W^{-1,p'}(\Omega))}&\le c\,\|\mathbf{R}^m\|_{[l^2(I_k;W^{-1,2}(\Omega)),l^\infty(I_k,L^2(\Omega))]_{[\frac{5p-6}{2p}]}}\\&\le c\,\|\mathbf{R}^m\|_{l^\infty(I_k;L^2(\Omega))}^{\frac{6-3p}{2p}}\|\mathbf{R}^m\|_{l^2(I_k;W^{1,2}(\Omega)^*)}^{\frac{5p-6}{2p}}\\&\le c\,k^{\frac{5p-6}{2p}}.\end{aligned}\tag{3.7}$$

From $\partial_t\mathbf{u}\in L^\infty(I;L^2(\Omega))$ and $\partial_t\mathbf{u}\in L^p(I;L^{3p}(\Omega))$ it follows by complex interpolation that

$$\partial_t\mathbf{u}\in L^2(I;L^{\frac{12}{8-3p}}(\Omega))=[L^\infty(I;L^2(\Omega)),L^p(I;L^{3p}(\Omega))]_{[\frac{p}{2}]}.$$

We estimate

$$k\sum_{m=1}^M\|d_t\mathbf{u}(t_m)\|_{\frac{12}{8-3p}}^2=k^{-1}\sum_{m=1}^M\left\|\int_{t_{m-1}}^{t_m}\partial_t\mathbf{u}(t)\,dt\right\|_{\frac{12}{8-3p}}^2\le\int_0^{t_M}\|\partial_t\mathbf{u}(t)\|_{\frac{12}{8-3p}}^2\,dt.$$

In particular,

$$\|d_t\mathbf{u}(t_m)\|_{l^2(I_k;L^{\frac{12}{8-3p}}(\Omega))} \le \|\partial_t\mathbf{u}\|_{L^2(I;L^{\frac{12}{8-3p}}(\Omega))} \le c. \tag{3.8}$$

From (3.7) and (3.8) follows

$$\begin{aligned} k\sum_{m=1}^{M} r_m^2 &\le c\left(\|\mathbf{R}^m\|^2_{l^2(I_k;W^{-1,p'}(\Omega))} + k^2\,\|d_t\mathbf{u}(t_m)\|^2_{l^2(I_k;L^{\frac{12}{8-3p}}(\Omega))}\right)\\ &\le c\,k^{\frac{5p-6}{p}} + c\,k^2 \le c\,k^{2\beta}. \end{aligned}$$

This proves (3.14). Let us return to the estimation of K_1^m, K_2^m, and K_3^m. First,

$$K_3^m = \left|\langle \mathbf{R}^m, \mathbf{e}^m\rangle\right| \le \|\mathbf{R}^m\|_{-1,p'}\|\mathbf{e}^m\|_{1,p} \le c\,\|\mathbf{R}^m\|_{-1,p'}\|\nabla\mathbf{e}^m\|_p. \tag{3.9}$$

Second,

$$K_1^m \le k\,\|\nabla\mathbf{u}(t_m)\|_{\frac{12p}{3p^2+8p-12}}\|d_t\mathbf{u}(t_m)\|_{\frac{12}{8-3p}}\|\mathbf{e}^m\|_{\frac{3p}{3-p}}.$$

We use $6(p-1) > \frac{12p}{3p^2+8p-12}$ for all $p \in [\frac{3}{2}, 2]$ and (1.5). Then

$$K_1^m \le ck\,\|d_t\mathbf{u}(t_m)\|_{\frac{12}{8-3p}}\|\nabla\mathbf{e}^m\|_p. \tag{3.10}$$

From (3.9) and (3.10) we deduce

$$K_1^m + K_3^m \le r_m b_m. \tag{3.11}$$

Third,

$$\begin{aligned} K_2^m &= \left|\left\langle [\nabla\mathbf{u}(t_m)]\mathbf{e}^{m-1}, \mathbf{e}^m\right\rangle\right|\\ &\le \|\nabla\mathbf{u}(t_m)\|_3\|\mathbf{e}^m\|_{\frac{p}{p-1}}\|\mathbf{e}^{m-1}\|_{\frac{3p}{3-p}}\\ &\le c\,\|\mathbf{e}^m\|_{\frac{p}{p-1}}\|\nabla\mathbf{e}^{m-1}\|_p, \end{aligned}$$

where we have used that $3 < 6(p-1)$ for all $p \in (\frac{3}{2}, 2]$, and (1.5).[2] Since $p \in (\frac{3}{2}, 2]$ there exists $\theta \in (0,1]$ with $\|\mathbf{e}^m\|_{\frac{p}{p-1}} \le c\,\|\nabla\mathbf{e}^m\|_p^{1-\theta}\|\mathbf{e}^m\|_2^{\theta}$. Thus

$$K_2^m \le c\,\|\nabla\mathbf{e}^{m-1}\|_p\|\nabla\mathbf{e}^m\|_p^{1-\theta}\|\mathbf{e}^m\|_2^{\theta} = c\,b_{m-1}b_m^{1-\theta}a_m^{\theta}. \tag{3.12}$$

Due to the embedding $W_0^{1,p}(\Omega) \hookrightarrow L^2(\Omega)$ it holds that $a_m \le c\,b_m$ which immediately implies

$$K_2^m \le c\,b_{m-1}b_m. \tag{3.13}$$

We combine (3.5), (3.11), (3.12), and (3.13) to obtain

$$\begin{aligned} d_t a_m^2 + c_6(1+b_m)^{p-2}b_m^2 &\le b_m r_m + c\,b_{m-1}b_m,\\ d_t a_m^2 + c_6(1+b_m)^{p-2}b_m^2 &\le b_m r_m + c\,b_{m-1}b_m^{1-\theta}a_m^{\theta}, \end{aligned}$$

[2]Note that this is the crucial estimate which limits the analysis to $p > 3/2$. Since the extra stress tensor $\mathbf{S}$ in problem (NS_p) depends on the symmetric part of the velocity gradient $\mathbf{D}$, the regularity stated in (1.5) is at the present time optimal. Thus the method presented here cannot be extended to smaller values of p.

which proves the validity of (3.15) and (3.16). Overall, we have shown that we can apply Lemma 3.2 below. This ensures the existence of k_3 such that if $0 < k < k_3$, then the error $\mathbf{e}^m$ satisfies

$$\max_{0\le m\le M} \|\nabla \mathbf{e}^m\|_p \le 1,$$

$$\max_{0\le m\le M} \|\mathbf{e}_m\|_2^2 + c_6\, k \sum_{m=0}^{M} \|\nabla \mathbf{e}^m\|_p^2 \le c\, k^{\frac{5p-6}{p}} \exp(c\,T).$$

This proves Proposition 1.2. □

We now state and prove the crucial Lemma 3.2 which is of Gronwall-type.

LEMMA 3.2. *Let $1 < p \le 2$. Let $0 \le a_m, b_m, r_m < \infty$, with $a_0 = b_0 = 0$, and let*

$$k \sum_{m=1}^{M} r_m^2 \le c_7 k^{2\beta}, \tag{3.14}$$

with $\frac{1}{2} < \beta$. Further, let

$$d_t a_m^2 + c_8\,(c_9 + b_m)^{p-2} b_m^2 \le b_m r_m + c_{10}\, b_{m-1} b_m, \tag{3.15}$$

$$d_t a_m^2 + c_8\,(c_9 + b_m)^{p-2} b_m^2 \le b_m r_m + c_{11}\, b_{m-1} b_m^{1-\theta} a_m^{\theta}, \tag{3.16}$$

with $0 < c_8, c_9 \le 1$, $c_{10}, c_{11} \ge 1$, and some $0 < \theta \le 1$. Then there exists $k_4 > 0$, such that if (3.15) and (3.16) hold for $0 < k < k_4 = \min(k_5, k_6)$, then

$$\max_{0\le m\le M} b_m \le 1, \tag{3.17}$$

$$\max_{0\le m\le M} a_m^2 + c_8\, k \sum_{m=0}^{M} b_m^2 \le c_{12} k^{2\beta} \exp(c_{13} k M), \tag{3.18}$$

where

$$c_{12} := 8\, c_7\, c_8^{-1}, \qquad c_{13} := 2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}}\, c_{11}^{\frac{2}{\theta}},$$

$$k_5 := \left(8\, c_8^{-2} \left(c_7 + c_{10}^2\, c_{12} \exp(c_{13} T)\right)\right)^{\frac{-1}{2\beta-1}}, \qquad k_6 := \frac{1}{2}\big(2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}}\, c_{11}^{\frac{2}{\theta}}\big)^{-1}.$$

Proof. We prove (3.17) and (3.18) by induction over M.

Case $M = 0$. Obvious.

Case $(M-1) \mapsto M$. We start with the proof of (3.17). There is nothing to show for $b_M \le 1$ so assume $b_M > 1$. Especially, we have $0 \le b_1, \dots, b_{M-1} \le 1 < b_M$. Summation of (3.15) over $\{1, \dots, M\}$ implies (with $a_0 = 0$)

$$\begin{aligned} a_M^2 + c_8\, k \sum_{m=0}^{M} (c_9 + b_m)^{p-2}\, b_m^2 &\le k \sum_{m=0}^{M} b_m \big(r_m + c_{10} b_{m-1}\big) \\ &\le c_8\, \frac{k}{2} \sum_{m=0}^{M} (c_9 + b_m)^{p-2} b_m^2 + \frac{1}{c_8}\, k \sum_{m=0}^{M} (c_9 + b_m^2)^{2-p} \big(r_m^2 + c_{10}^2 b_{m-1}^2\big). \end{aligned}$$

We absorb the first term of the right-hand side and then neglect all summands on the left-hand side, except for $m = M$:

$$\begin{aligned} c_8\, k\, (c_9 + b_M)^{p-2}\, b_M^2 &\le \frac{2}{c_8} k \sum_{m=0}^{M} (c_9 + b_m^2)^{2-p} \big(r_m^2 + c_{10}^2 b_{m-1}^2\big) \\ &\le \frac{2}{c_8} (c_9 + b_M)^{2-p} k \sum_{m=0}^{M} \big(r_m^2 + c_{10}^2 b_{m-1}^2\big) \\ &\le \frac{2}{c_8} (c_9 + b_M)^{2-p} k^{2\beta} \big(c_7 + c_{10}^2\, c_{12}\, \exp(c_{13}\, kM)\big), \end{aligned}$$

where we have used (3.18) for $0 \le m \le M-1$. With $0 \le c_9 \le 1 < b_M$

$$\begin{aligned} b_M^2 &\le 2 c_8^{-2} (c_9 + b_M)^{2(2-p)} k^{2\beta-1} \big(c_7 + c_{10}^2\, c_{12}\, \exp(c_{13}\, kM)\big) \\ &\le 2 c_8^{-2} b_M^{2(2-p)} 2^{2(2-p)} k^{2\beta-1} \big(c_7 + c_{10}^2\, c_{12}\, \exp(c_{13}\, kM)\big). \end{aligned}$$

In particular, with $1 < p \le 2$

$$1 < b_M^{2(p-1)} \le 8\, k^{2\beta-1}\, c_8^{-2} \big(c_7 + c_{10}^2\, c_{12}\, \exp(c_{13}\, T)\big).$$

If $0 < k < k_5$ with

$$k_5 := \left(8\, c_8^{-2} \left(c_7 + c_{10}^2\, c_{12} \exp(c_{13} T)\right)\right)^{\frac{-1}{2\beta-1}},$$

we get the desired contradiction $1 < 1$. This proves $0 \le b_M \le 1$, i.e., (3.17).

We continue with the proof of (3.18). From (3.16) and Young's inequality we deduce

$$\begin{aligned} d_t a_m^2 + c_8\, (c_9 + b_m)^{p-2} b_m^2 &\le r_m\, b_m + c_{11}\, b_m^{1-\theta} a_m^\theta b_{m-1} \\ &\le c_{11} \big(b_m^{2-\theta} + b_{m-1}^{2-\theta}\big) a_m^\theta + 2\, c_8^{-1}\, r_m^2 + \frac{c_8}{8}\, b_m^2. \end{aligned}$$

Now $b_1, \dots, b_M \le 1$, $0 < c_9 \le 1$, $0 < \theta \le 1$, $c_{11} \ge 1$, and Young's inequality imply

$$\begin{aligned} d_t a_m^2 + \frac{c_8}{2} b_m^2 &\le c_{11} \big(b_m^{2-\theta} + b_{m-1}^{2-\theta}\big) a_m^\theta + 2\, c_8^{-1} r_m^2 + \frac{c_8}{8} b_m^2 \\ &\le \frac{c_8}{8} \big(b_m^2 + b_{m-1}^2\big) + 2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}}\, c_{11}^{\frac{2}{\theta}} a_m^2 + 2\, c_8^{-1} r_m^2. \end{aligned}$$

Taking the sum $m = 1, \dots, M$ with $a_0 = 0$ implies

$$a_M^2 + \frac{c_8}{4} k \sum_{k=1}^{M} b_m^2 \le 2\, c_8^{-1}\, k \sum_{k=1}^{M} r_m^2 + 2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}}\, k \sum_{k=1}^{M} a_m^2\,.$$

Let $k_6 := \frac{1}{2}\big(2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}}\big)^{-1}$. Then for $0 < k < k_6$ we absorb $k\, 2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}} a_M^2 < \frac{1}{2} a_M^2$ on the left-hand side. Thus

$$a_M^2 + \frac{c_8}{2} k \sum_{k=1}^{M} b_m^2 \le \frac{4}{c_8}\, k \sum_{k=1}^{M} r_m^2 + 2 \cdot 8^{\frac{2-\theta}{\theta}}\, c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}}\, k \sum_{k=1}^{M-1} a_m^2.$$

An application of Gronwall's lemma implies

$$a_M^2 + \frac{c_8}{2} k \sum_{m=0}^{M} b_m \le \left(\frac{4}{c_8} k \sum_{m=0}^{M} r_m^2 \right) \exp\left(2 \cdot 8^{\frac{2-\theta}{\theta}} c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}} kM \right)$$
$$\le \frac{4\, c_7}{c_8}\, k^{2\beta} \exp\left(2 \cdot 8^{\frac{2-\theta}{\theta}} c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}}\, kM \right).$$

Let $c_{12} := 8\, c_7\, c_8^{-1}$ and $c_{13} := 2 \cdot 8^{\frac{2-\theta}{\theta}} c_8^{\frac{\theta-2}{\theta}} c_{11}^{\frac{2}{\theta}}$; then

$$a_M^2 + c_8\, k \sum_{m=0}^{M} b_m \le c_{12} k^{2\beta} \exp(c_{13} kM).$$

This proves (3.15). The choice $k_4 := \min(k_5, k_6)$ concludes the lemma. ∎

4. Proof of Theorems 1.3 and 1.4. In this section we will show that the solution $\mathbf{u}^m$ of system (NS_p^k) is not only a weak solution but a strong solution. In particular, we will derive some a priori estimates of second derivatives of $\mathbf{u}^m$ which will be independent of the time step size k (as long as $k \le k_1$). The results of this are summarized in Theorem 1.3.

Before we start with the derivation of the (global in time) a priori estimates of $\mathbf{u}^m$ we will show that each $\mathbf{u}^m$ has some higher regularity in space. This regularity in terms of norms crucially depends on the time step size k and the time step m. Nevertheless, we will need this in order to justify all the calculations later, in particular that all terms involved are finite.

LEMMA 4.1. *Let p, $\mathbf{f}$, and $\mathbf{u}_0$ be as in Theorem* 1.3. *Let $\mathbf{u}^m$ be the weak solution of system (NS_p^k) as in Lemma* 3.1. *Then for all $m \in \{-1, 0, \dots, M\}$ it holds that*

$$k^{-1} \|\nabla \mathbf{u}^m\|_2^2 + \mathcal{I}_p(\mathbf{u}^m) \le c(m, k^{-1}) < \infty. \tag{4.1}$$

Proof. We proceed by induction of m. For each time index m the function $\mathbf{u}^m$ is just the solution of a stationary quasi–Stokes-like system equipped with periodic boundary conditions

$$\begin{aligned} \frac{1}{k}\mathbf{u}^m - \operatorname{div} \mathbf{S}(\mathbf{D}\mathbf{u}^m) + [\nabla \mathbf{u}^m]\mathbf{u}^{m-1} &= \mathbf{f}^m + \frac{1}{k}\mathbf{u}^{m-1}, \\ \operatorname{div} \mathbf{u}^m &= 0. \end{aligned} \tag{4.2}$$

Either by means of the difference quotient method or by a Galerkin approach with eigenfunctions of the Stokes operator, it is possible to justify the formal use of $-\Delta \mathbf{u}^m$ as a test function of (4.2). As in [13] this implies

$$\tfrac{1}{2} d_t \|\nabla \mathbf{u}^m\|_2^2 + C_1\, \mathcal{I}_p(\mathbf{u}^m) \le c(\mathbf{f}) + \int_\Omega |\nabla \mathbf{u}^m|^2\, |\nabla \mathbf{u}^{m-1}|\, dx. \tag{4.3}$$

Here we have used

$$\begin{aligned}\langle -\operatorname{div}\mathbf{S}(\mathbf{D}\mathbf{u}^m), -\Delta\mathbf{u}^m\rangle &= \sum_{i,j,k}\langle \partial_j S_{ij}(\mathbf{D}\mathbf{u}^m), \partial_k^2 u_i^m\rangle\\ &= \sum_{i,j,k}\langle \partial_k S_{ij}(\mathbf{D}\mathbf{u}^m), \partial_j\partial_k u_i^m\rangle\\ &= \sum_{i,j,k,l,m}\big\langle \partial_{ij}\partial_{lm}\Phi(|\mathbf{D}\mathbf{u}^m|)D_{ij}(\partial_k\mathbf{u}^m), D_{ij}(\partial_k\mathbf{u}^m)\big\rangle\\ &\geq C_1\int_\Omega (1+|\mathbf{D}\mathbf{u}^m|^2)^{\frac{p-2}{2}}|\mathbf{D}(\nabla\mathbf{u}^m)|^2\,dx \qquad \text{(by (1.1))}\\ &= C_1\,\mathcal{I}_p(\mathbf{u}^m)\end{aligned}$$

and

$$\begin{aligned}\big\langle [\nabla\mathbf{u}^m]\mathbf{u}^{m-1}, -\Delta\mathbf{u}^m\big\rangle &= \sum_{j,k,l}\big\langle \partial_j u_k^m u_k^{m-1}, \partial_l^2 u_k^m\big\rangle\\ &= \sum_{j,k,l}\big\langle \partial_j u_k^m \partial_l u_j^{m-1}, \partial_l u_k^m\big\rangle + \sum_{j,k,l}\int_\Omega \tfrac{1}{2}\partial_j\big((\partial_j u_k^m)^2\big)u_j^{m-1}\,dx\\ &= \sum_{j,k.l}\big\langle \partial_j u_k^m \partial_l u_j^{m-1}, \partial_l u_k^m\big\rangle \qquad \text{(by } \operatorname{div}\mathbf{u}^{m-1}=0)\\ &\leq \int_\Omega |\nabla\mathbf{u}^m|^2|\nabla\mathbf{u}^{m-1}|\,dx,\end{aligned}$$

where we have frequently used the periodicity. Now, Young's inequality and (2.8) imply

$$\begin{aligned}\int_\Omega |\nabla\mathbf{u}^m|^2|\nabla\mathbf{u}^{m-1}|\,dx &\leq \|\nabla\mathbf{u}^{m-1}\|_{3p}\|\nabla\mathbf{u}^m\|^2_{\frac{6p}{3p-1}}\\ &\leq c\,\|\nabla\mathbf{u}^{m-1}\|_{3p}\|\nabla\mathbf{u}^m\|_p^{\frac{5p-7}{p}}\|\nabla\mathbf{u}^m\|^{\frac{7-3p}{p}}_{\frac{3p}{3-p}}\\ &\leq c_\delta\,\|\nabla\mathbf{u}^{m-1}\|_{3p}^{\frac{2p}{5p-7}}\|\nabla\mathbf{u}^m\|_p^2 + \delta\|\nabla\mathbf{u}^m\|^2_{\frac{3p}{3-p}}\\ &\leq c_\delta\,\|\nabla\mathbf{u}^{m-1}\|_{3p}^{\frac{2p}{5p-7}}\|\nabla\mathbf{u}^m\|_p^2 + \delta\,c\,\|\nabla^2\mathbf{u}^m\|_p^2\\ &\leq c_\delta\,\|\nabla\mathbf{u}^{m-1}\|_{3p}^{\frac{2p}{5p-7}}\|\nabla\mathbf{u}^m\|_p^2 + \delta\,c\,\mathcal{I}_p(\mathbf{u}^m).\end{aligned}$$

We absorb the last term on the left-hand side of (4.3) and get

$$\frac{1}{2}d_t\|\nabla\mathbf{u}^m\|_2^2 + C_1\,\mathcal{I}_p(\mathbf{u}^m) \leq c(\mathbf{f}) + c\,\|\nabla\mathbf{u}^{m-1}\|_{3p}^{\frac{2p}{5p-7}}\|\nabla\mathbf{u}^m\|_p^2.$$

From $\mathcal{I}_p(\mathbf{u}^{m-1})<\infty$, $\|\nabla\mathbf{u}^m\|_2<\infty$, and (2.9) we know that the right-hand side is finite. This proves the lemma. □

Note that estimate (4.1) depends on k^{-1} and m. Nevertheless, it will justify all of the following calculations. We will now get to the proof of Theorem 1.3.

Proof of Theorem 1.3. Let $\mathbf{u}^m$ be the weak solution of Lemma 3.1. We will show that $\mathbf{u}^m$ satisfies (1.7). Unfortunately, the proof of this cannot be reduced to a simple Gronwall argument but is rather subtle. Let q, r be such that $3<q<r<6(p-1)$.

By assumption (1.5) on $\mathbf{u}$ there exists $c_{14} = c_{14}(\mathbf{f}, \mathbf{u}_0) \geq 1$ such that for all M with $kM \leq T$ it holds that

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}(t_m)\|_p + \max_{0\leq m\leq M} \|\nabla \mathbf{u}(t_m)\|_r \leq c_{14}. \tag{4.4}$$

From Proposition 1.2, it follows that, for $k \leq k_1$, $\max_{0\leq m\leq M} \|\nabla \mathbf{e}^m\|_p \leq 1$ holds. This, together with $c_{14} \geq 1$ and (4.4), implies

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}^m\|_p \leq 2\, c_{14}. \tag{4.5}$$

We proceed by induction over M with step $M-1 \mapsto M$. Instead of (1.7) we will show step by step that $\mathbf{u}^m$ satisfies

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}^m\|_2^2 + k \sum_{m=0}^{M} \mathcal{I}_p(\mathbf{u}^m) \leq c_{15} = c_{15}(c_{14}, p), \tag{4.6a}$$

$$\max_{1\leq m\leq M} \|d_t \mathbf{u}^m\|_2^2 + k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m) \leq c_{16} = c_{16}(c_{14}, p), \tag{4.6b}$$

$$k \sum_{m=0}^{M} \big(\mathcal{I}_p(\mathbf{u}^m)\big)^{\frac{5p-6}{2-p}} \leq c_{17} = c_{17}(c_{14}, p), \tag{4.6c}$$

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}^m\|_r \leq c_{18} = c_{18}(c_{14}, p), \tag{4.6d}$$

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}(t_m) - \nabla \mathbf{u}^m\|_q \leq c_{14}, \tag{4.6e}$$

$$\max_{0\leq m\leq M} \|\nabla \mathbf{u}^m\|_q \leq 2\, c_{14}. \tag{4.6f}$$

Obviously, by $\mathbf{u}(0) = \mathbf{u}^0 = \mathbf{u}_0$ and the assumptions on $\mathbf{u}_0$ and $p \leq 2$, inequalities (4.6) are valid for $M = 0$. Note that due to (2.16) the inequality (4.6f) seems to contain less information than (4.6b) and (4.6c). The point here is that c_{14} is much smaller than c_{17} and is given in advance by (1.5). In the proof of (4.6a), (4.6b), and (4.6e) we will need some smallness of the step size k, which is dependent on c_{14} and c_1. Therefore, we assume $0 < k \leq k_7(c_{14}, c_1, k_1)$, where c_1 and k_1 are from Proposition 1.2. The exact dependence of k_5 on c_{14}, c_1, and k_1 is given later. The proof of (4.6e) will further rely on the error estimates of Proposition 1.2. We will prove (4.6a)–(4.6f) in the same order as stated. Assume in the following that (4.6) holds for $M-1$. From Lemma 4.1 we further know that $\mathbf{u}^m$ has enough regularity to justify the calculations below.

Proof of (4.6a). Using the test function $-\Delta \mathbf{u}^m$ for the system (NS_p^k) we conclude as in the proof of (4.3) in Lemma 4.1 that

$$\begin{aligned} \tfrac{1}{2} d_t \|\nabla \mathbf{u}^m\|_2^2 + C_1\, \mathcal{I}_p(\mathbf{u}^m) &\leq c(\mathbf{f}) + \int_\Omega |\nabla \mathbf{u}^m|^2 |\nabla \mathbf{u}^{m-1}|\, dx \\ &\leq c(\mathbf{f}) + \|\nabla \mathbf{u}^m\|_3^2 \|\nabla \mathbf{u}^{m-1}\|_3. \end{aligned} \tag{4.7}$$

Since $2 < 3 < \frac{3p}{3-p}$, there exists $\theta = \theta(p) \in (0,1)$ with $L^3(\Omega) = [L^2(\Omega), L^{\frac{3p}{3-p}}(\Omega)]_{[\theta]}$.

This and (4.6f) for $M-1$ imply

$$\begin{aligned}\frac{1}{2}d_t\|\nabla\mathbf{u}^m\|_2^2+C_1\,\mathcal{I}_p(\mathbf{u}^m)&\le c(\mathbf{f})+c_{14}\,\|\nabla\mathbf{u}^m\|_3^2\\&\le c(\mathbf{f})+c_{14}\,\|\nabla\mathbf{u}^m\|_2^{2(1-\theta)}\|\nabla\mathbf{u}^m\|_{\frac{3p}{3-p}}^{2\theta}\\&\le c(\mathbf{f})+c(c_{14},p,\varepsilon)\,\|\nabla\mathbf{u}^m\|_2^2+\varepsilon\|\nabla\mathbf{u}^m\|_{\frac{3p}{3-p}}^2,\end{aligned}\tag{4.8}$$

with $\varepsilon>0$. From (2.8) with $s=p$ and (4.5) follows

$$\frac{1}{2}d_t\|\nabla\mathbf{u}^m\|_2^2+C_1\,\mathcal{I}_p(\mathbf{u}^m)\le c(\mathbf{f})+c(c_{14},p,\varepsilon)\,\|\nabla\mathbf{u}^m\|_2^2+\varepsilon\,\mathcal{I}_p(\mathbf{u}^m)\,(1+2\,c_{14})^{2-p}.$$

For small ε we obtain

$$d_t\|\nabla\mathbf{u}^m\|_2^2+C_1\,\mathcal{I}_p(\mathbf{u}^m)\le c(\mathbf{f})+c(c_{14},p)\|\nabla\mathbf{u}^m\|_2^2.\tag{4.9}$$

Now, Gronwall's inequality provides the existence of $k_8=k_8(c_{14},p)$ and $c_{15}=c_{15}(c_{14},p)$ such that (4.6a) holds, provided that $k_7\le k_8$.

Proof of (4.6b). We want to use the test function $d_t\mathbf{u}^M$. In order to give $d_t\mathbf{u}^0$ a meaning we introduce $\mathbf{u}^{-1}$. For that we set for all $\boldsymbol{\varphi}\in\mathcal{V}$

$$\frac{1}{k}\langle\mathbf{u}^0-\mathbf{u}^{-1},\boldsymbol{\varphi}\rangle+\big\langle\mathbf{S}(\mathbf{D}(\mathbf{u}^0)),\mathbf{D}(\boldsymbol{\varphi})\big\rangle+\langle[\nabla\mathbf{u}^0]\mathbf{u}^0,\boldsymbol{\varphi}\rangle=\langle\mathbf{f}(0),\boldsymbol{\varphi}\rangle.$$

Using $\mathbf{u}^0=\mathbf{u}_0$, $p\le 2$, and the assumption on $\mathbf{u}_0$, we obtain

$$\|d_t\mathbf{u}^0\|_2^2\le\|\mathbf{f}(0)\|_2^2+\|[\nabla\mathbf{u}_0]\mathbf{u}_0\|_2^2+\big\|\operatorname{div}\mathbf{S}(\mathbf{D}(\mathbf{u}_0))\big\|_2^2\le c(\mathbf{f},\mathbf{u}_0).\tag{4.10}$$

Now we can take the discrete time derivative of the weak formulation (NS_p^k), use $d_t\mathbf{u}^m$ as a test function, and sum up to obtain

$$\begin{aligned}&\|d_t\mathbf{u}^M\|_2^2+k^{-1}\sum_{m=1}^M\big\langle\mathbf{S}(\mathbf{D}\mathbf{u}^m)-\mathbf{S}(\mathbf{D}\mathbf{u}^{m-1}),\mathbf{D}\mathbf{u}^m-\mathbf{D}\mathbf{u}^{m-1}\big\rangle\\&\quad\le c(\mathbf{f},\mathbf{u}_0)+c\,k\sum_{m=1}^M\big|\big\langle d_t\big([\nabla\mathbf{u}^m]\mathbf{u}^{m-1}\big),d_t\mathbf{u}^m\big\rangle\big|\\&\quad\le c(\mathbf{f},\mathbf{u}_0)+c\,k\sum_{m=1}^M\big|\big\langle[\nabla\mathbf{u}^{m-1}]d_t\mathbf{u}^{m-1},d_t\mathbf{u}^m\big\rangle\big|,\end{aligned}$$

where we used (4.10), $d_t\big([\nabla\mathbf{u}^m]\mathbf{u}^{m-1}\big)=[d_t\nabla\mathbf{u}^m]\mathbf{u}^{m-1}+[\nabla\mathbf{u}^{m-1}]d_t\mathbf{u}^{m-1}$, and $\langle[d_t\nabla\mathbf{u}^m]\mathbf{u}^{m-1},d_t\mathbf{u}^m\rangle=0$ as $\operatorname{div}\mathbf{u}^{m-1}=0$. From (2.5) and the definition of $\mathcal{K}_p$ we deduce

$$\big\langle\mathbf{S}(\mathbf{D}\mathbf{u}^m)-\mathbf{S}(\mathbf{D}\mathbf{u}^{m-1}),\mathbf{D}\mathbf{u}^m-\mathbf{D}\mathbf{u}^{m-1}\big\rangle\ge c\,k^2\,\mathcal{K}_p(\mathbf{u}^m).$$

Overall,

$$\begin{aligned}\|d_t\mathbf{u}^M\|_2^2+k\sum_{m=1}^M\mathcal{K}_p(\mathbf{u}^m)&\le c(\mathbf{f},\mathbf{u}_0)+c\,k\sum_{m=1}^M\big|\big\langle[\nabla\mathbf{u}^{m-1}]d_t\mathbf{u}^{m-1},d_t\mathbf{u}^m\big\rangle\big|\\&\le c(\mathbf{f},\mathbf{u}_0)+c\,k\sum_{m=1}^M\|\nabla\mathbf{u}^{m-1}\|_3\|d_t\mathbf{u}^{m-1}\|_3\|d_t\mathbf{u}^m\|_3.\end{aligned}$$

Thus, by $q > 3$ and (4.6f) for $M-1$ follows

$$\|d_t \mathbf{u}^M\|_2^2 + k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m) \le c(\mathbf{f}, \mathbf{u}_0) + c\, c_{14}\, k \sum_{m=1}^{M} \|d_t \mathbf{u}^{m-1}\|_3 \|d_t \mathbf{u}^m\|_3.$$

Analogously to the step from (4.7) to (4.8) we get

$$\|d_t \mathbf{u}^M\|_2^2 + k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m) \le c(\mathbf{f}, \mathbf{u}_0) + c(c_{14}, p, \varepsilon)\, k \sum_{m=1}^{M} \big(\|d_t \mathbf{u}^{m-1}\|_2^2 + \|d_t \mathbf{u}^m\|_2^2\big) + \varepsilon\, k \sum_{m=1}^{M} \|d_t \mathbf{u}^m\|_{\frac{3p}{3-p}}^2.$$

From (2.13) with $s = p$ and (4.5) follows

$$\|d_t \mathbf{u}^M\|_2^2 + k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m) \le c(\mathbf{f}, \mathbf{u}_0) + c(c_{14}, p, \varepsilon)\, k \sum_{m=1}^{M} \big(\|d_t \mathbf{u}^{m-1}\|_2^2 + \|d_t \mathbf{u}^m\|_2^2\big) + \varepsilon\, k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m)\, (1 + 2\, c_{14})^{2-p}.$$

For small ε we obtain

(4.11)
$$\|d_t \mathbf{u}^M\|_2^2 + k \sum_{m=1}^{M} \mathcal{K}_p(\mathbf{u}^m) \le c(\mathbf{f}, \mathbf{u}_0) + c(c_{14}, p, \varepsilon)\, k \sum_{m=1}^{M} \big(\|d_t \mathbf{u}^{m-1}\|_2^2 + \|d_t \mathbf{u}^m\|_2^2\big).$$

Now, Gronwall's inequality provides the existence of $k_9 = k_9(c_{14}, p)$ and $c_{16} = c_{16}(c_{14}, p)$ such that (4.6b) holds, provided that $k_7 \le k_9$.

Proof of (4.6c). As in the proof of (4.6a) we use $-\Delta \mathbf{u}^m$ as a test function. But, instead of retrieving information from the term $\langle d_t \mathbf{u}^m, \Delta \mathbf{u}^m \rangle$ as in the proof of (4.6a), will we estimate it on the right-hand side. In particular, instead of (4.9) we get

$$\begin{aligned} C_1\, \mathcal{I}_p(\mathbf{u}^m) &\le c(\mathbf{f}) + c(c_{14}, p) \|\nabla \mathbf{u}^m\|_2^2 + \big|\langle d_t \mathbf{u}^m, \Delta \mathbf{u}^m \rangle\big| \\ &\le c(\mathbf{f}) + c(c_{14}, p) + \big|\langle d_t \mathbf{u}^m, \Delta \mathbf{u}^m \rangle\big|, \end{aligned} \tag{4.12}$$

where we have used (4.6a). We will proceed as in [4], [16]:

$$\big|\langle d_t \mathbf{u}^m, \Delta \mathbf{u}^m \rangle\big| \le \|d_t \mathbf{u}^m\|_{\frac{3p}{2p-1}} \|\nabla^2 \mathbf{u}^m\|_{\frac{3p}{p+1}} \le c\, \|d_t \mathbf{u}^m\|_{\frac{3p}{2p-1}} \big(1 + \mathcal{I}_p(\mathbf{u}^m)\big)^{\frac{1}{p}},$$

where we used (2.9). With (4.12) we get

$$\big(1 + \mathcal{I}_p(\mathbf{u}^m)\big)^{\frac{p-1}{p}} \le c(c_{14}, p) \big(1 + \|d_t \mathbf{u}^m\|_{\frac{3p}{2p-1}}\big).$$

Now, we interpolate $L^{\frac{3p}{2p-1}}(\Omega)$ between $L^2(\Omega)$ and $L^{3p}(\Omega)$ and use (4.6b) and (2.14) to arrive at

$$\big(1 + \mathcal{I}_p(\mathbf{u}^m)\big)^{\frac{p-1}{p}} \le c(c_{14}, p) \Big(1 + \mathcal{K}_p(\mathbf{u}^m)^{\frac{\lambda}{2}} \big(1 + \mathcal{I}_p(\mathbf{u}^m) + \mathcal{I}_p(\mathbf{u}^{m-1})\big)^{\lambda\, \frac{2-p}{2p}}\Big),$$

with $\lambda = \frac{2-p}{3p-2}$. We raise this inequality to the power γ and apply Young's inequality to get

$$\begin{aligned}
&(1+\mathcal{I}_p(\mathbf{u}^m)\,)^{\gamma\,\frac{p-1}{p}}\\
&\quad\le c(c_{14},p)\left(1+\mathcal{K}_p(\mathbf{u}^m)^{\gamma\,\frac{\lambda}{2}}(1+\mathcal{I}_p(\mathbf{u}^m)+\mathcal{I}_p(\mathbf{u}^{m-1})\,)^{\gamma\,\lambda\,\frac{2-p}{2p}}\right)\\
&\quad\le c(c_{14},p)\left(1+c_\varepsilon\mathcal{K}_p(\mathbf{u}^m)+\varepsilon\big(1+\mathcal{I}_p(\mathbf{u}^m)+\mathcal{I}_p(\mathbf{u}^{m-1})\,\big)^{\frac{2\gamma\,\lambda}{2-\gamma\lambda}\,\frac{2-p}{2p}}\right). \qquad (4.13)
\end{aligned}$$

We now require

$$\gamma\frac{p-1}{p}=\frac{2\gamma\,\lambda}{2-\gamma\lambda}\frac{2-p}{2p},$$

which gives $\gamma=\frac{p}{p-1}\frac{5p-6}{2-p}$. With this γ and with ε sufficiently small we can absorb the last term in (4.13) into the left-hand side after summation over all time steps. Thus, we have derived

$$k\sum_{m=1}^{M}\mathcal{I}_p(\mathbf{u}^m)^{\frac{5p-6}{2-p}}\le c(c_{14},p)\left(1+k\sum_{m=1}^{M}\mathcal{K}_p(\mathbf{u}^m)\right)\le c(c_{14},p),$$

where we have used (4.6b). This proves (4.6c).

Proof of (4.6d). This is a direct consequence of (2.16), (4.6b), and (4.6c).

Proof of (4.6e). From (4.4) and (4.6d) we deduce

$$\max_{0\le m\le M}\|\nabla\mathbf{u}(t_m)-\nabla\mathbf{u}^m\|_r\le c(c_{14},p). \qquad (4.14)$$

On the other hand from (1.6) we know that

$$\max_{0\le m\le M}\|\nabla\mathbf{u}(t_m)-\nabla\mathbf{u}^m\|_p\le c_1\,k^{\frac{4p-6}{2p}}\le c_1\,k^{\frac{2p-3}{p}} \qquad (4.15)$$

for $p\in(\frac{3}{2},2]$. Since $p<q<r$, there exists by interpolation of (4.14) and (4.15) some $k_{10}=k_{10}(p,c_{14})>0$ such that

$$\max_{0\le m\le M}\|\nabla\mathbf{u}(t_m)-\nabla\mathbf{u}^m\|_q\le c_{14}$$

as long as $k\le k_{10}$. This proves (4.6e).

Proof of (4.6f). By (4.4) and $p<q<r$ there follows by interpolation

$$\max_{0\le m\le M}\|\nabla\mathbf{u}(t_m)\|_q\le c_{14}.$$

This and (4.6e) immediately imply (4.6f).

The proof of Theorem 1.3 is complete. □

Based on Theorem 1.3 we can now improve the convergence rate from Proposition 1.2.

Proof of Theorem 1.4. We will proceed as in the proof of Proposition 1.2. However, due to the better regularity properties of $\mathbf{u}^m$ we can extract more information from the second term on the left-hand side in (3.4). Namely, from Remark 2.2, with $r\in(p,6(p-1))$, (1.5), and (1.8) we deduce instead of (3.5)

$$\frac{1}{2}d_t\|\mathbf{e}^m\|_2^2+c_{19}\left(\|\nabla\mathbf{e}^m\|_p^2+\|\nabla\mathbf{e}^m\|^2_{\frac{2r}{2-p+r}}\right)\le c\left(K_1^m+K_2^m+K_3^m\right). \qquad (4.16)$$

Now we set instead of (3.6)

$$
\begin{aligned}
a_m &:= \|\mathbf{e}^m\|_2,\\
\widetilde{b}_m &:= \|\nabla \mathbf{e}^m\|_p + \|\nabla \mathbf{e}^m\|_{\frac{2r}{2-p+r}},\\
\widetilde{r}_m &:= c\left(\|\mathbf{R}^m\|_{-1,(\frac{2r}{2-p+r})'} + k\,\|d_t\mathbf{u}(t_m)\|_{\frac{12}{8-3p}}\right),\\
\beta = \beta(p,r) &:= \frac{2r+3p-6}{2r},
\end{aligned}
\tag{4.17}
$$

and want to use Lemma 3.2 with b_m replaced by $\widetilde{b}_m$ and r_m replaced by $\widetilde{r}_m$. Note that

$$b_m \le \widetilde{b}_m$$

since $r > p$. Thus we can replace in all estimates in the proof of Proposition 1.2 the term b_m by $\widetilde{b}_m$. In order to show that $\widetilde{r}_m$ satisfies (3.14) we use the embedding

$$
\begin{aligned}
&[l^2(I_k;W^{-1,2}(\Omega)), l^\infty(I_k;L^2(\Omega))]_{[\frac{2r+3p-6}{2r}]}\\
&\quad\hookrightarrow [l^2(I_k;W^{-1,2}(\Omega)), l^\infty(I_k;W^{-1,6}(\Omega))]_{[\frac{2r+3p-6}{2r}]}\\
&\quad\hookrightarrow l^{\frac{4r}{2r+3p-6}}(I_k;W^{-1,(\frac{2r}{2-p+r})'}(\Omega))\\
&\quad\hookrightarrow l^2(I_k;W^{-1,(\frac{2r}{2-p+r})'}(\Omega))
\end{aligned}
$$

and (3.2) to show that

$$
\begin{aligned}
\|\mathbf{R}^m\|_{l^2(I_k;W^{-1,p'}(\Omega))} &\le c\,\|\mathbf{R}^m\|_{[l^2(I_k;W^{-1,2}(\Omega)),l^\infty(I_k,L^2(\Omega))]_{[\frac{2r+3p-6}{2r}]}}\\
&\le c\,\|\mathbf{R}^m\|_{l^\infty(I_k;L^2(\Omega))}^{\frac{6-3p}{2r}}\|\mathbf{R}^m\|_{l^2(I_k;W^{1,2}(\Omega)^*)}^{\frac{2r+3p-6}{2r}}\\
&\le c\,k^{\frac{2r+3p-6}{2r}} = c\,k^{\beta(p,r)}.
\end{aligned}
\tag{4.18}
$$

From this and (3.8) follows

$$
\begin{aligned}
k\sum_{m=1}^M \widetilde{r}_m^2 &\le c\left(\|\mathbf{R}^m\|^2_{l^2(I_k;W^{-1,(\frac{2r}{2-p+r})'}(\Omega))} + k^2\,\|d_t\mathbf{u}(t_m)\|^2_{l^2(I_k;L^{\frac{12}{8-3p}}(\Omega))}\right)\\
&\le c\,k^{\frac{2r+3p-6}{r}} + c\,k^2 \le c\,k^{2\beta(p,r)}.
\end{aligned}
$$

This proves (3.14). Thus Lemma 3.2 in particular implies

$$\max_{0\le m\le M}\|\nabla\mathbf{e}^m\|_p \le 1,$$

$$\max_{0\le m\le M}\|\mathbf{e}_m\|_2^2 + c_{19}\,k\sum_{m=0}^M\|\nabla\mathbf{e}^m\|_p^2 \le c\,k^{2\beta(p,r)}\exp(c\,T).$$

Since

$$\lim_{r\to 6(p-1)}\beta(p,r) = \alpha_0(p) = \frac{5p-6}{4(p-1)},$$

the proof of Theorem 1.4 is complete. □

Acknowledgment. The authors would like to thank the referees for their comments and remarks.

REFERENCES

[1] H. Beirão da Veiga, *On the Regularity of Flows with Ladyzhenskaya Shear-Dependent Viscosity and Slip or Nonslip Boundary Conditions. Part* II: *The Evolution Problem*, preprint, University of Pisa, Pisa, Italy, 2004.

[2] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, 2nd ed., John Wiley, New York, 1987.

[3] L. Diening, *Theoretical and Numerical Results for Electrorheological Fluids*, Ph.D. thesis, University of Freiburg, Freiburg, Germany, 2002.

[4] L. Diening, A. Prohl, and M. Růžička, *On time-discretizations for generalized Newtonian fluids*, in Nonlinear Problems in Mathematical Physics and Related Topics, II, Int. Math. Ser. (N.Y.) 2, Kluwer/Plenum, New York, 2002, pp. 89–118.

[5] L. Diening and M. Růžička, *Strong solutions for generalized Newtonian fluids*, J. Math. Fluid Mech., 7 (2005), pp. 413–450.

[6] J. Frehse, J. Málek, and M. Steinhauer, *An existence result for fluids with shear dependent viscosity—steady flows*, Nonlinear Anal., 30 (1997), pp. 3041–3049.

[7] A. Kufner, O. John, and S. Fučík, *Function Spaces*, Academia, Prague, 1977.

[8] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, London, Paris, 1969.

[9] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[10] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman & Hall, London, 1996.

[11] J. Málek, J. Nečas, and M. Růžička, *On weak solutions to a class of non-Newtonian incompressible fluids in bounded three-dimensional domains. The case* $p \geq 2$, Adv. Differential Equations, 6 (2001), pp. 257–302.

[12] J. Málek, K. R. Rajagopal, and M. Růžička, *Existence and regularity of solutions and the stability of the rest state for fluids with shear dependent viscosity*, Math. Models Methods Appl. Sci., 5 (1995), pp. 789–812.

[13] A. Prohl and M. Růžička, *On fully implicit space-time discretization for motions of incompressible fluids with shear-dependent viscosities: The case* $p \leq 2$, SIAM J. Numer. Anal., 39 (2001), pp. 214–249.

[14] K. R. Rajagopal, *Mechanics of non-Newtonian fluids*, in Recent Developments in Theoretical Fluid Mechanics, Pitman Res. Notes Math. Ser. 291, G. P. Galdi and J. Nečas, eds., Longman, Harlow, UK, 1993, pp. 129–162.

[15] M. Růžička, *A note on steady flow of fluids with shear dependent viscosity*, Nonlinear Anal., 30 (1997), pp. 3029–3039.

[16] M. Růžička, *Modeling, mathematical and numerical analysis of electrorheological fluids*, Appl. Math., 49 (2004), pp. 565–609.

[17] G. Thäter, *Natural Convection, Dissipation & Power-Law Rheology: Mathematical Models & Results*, Habilitation Thesis, University of Hannover, Hannover, Germany, 2003.

[18] J. Wolf, *Existence of weak solutions to the equations of nonstationary motion of non-Newtonian fluids with shear-dependent viscosity*, J. Math. Fluid Mech., accepted.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1191–1217

PERFECTLY MATCHED LAYERS FOR TIME-HARMONIC ACOUSTICS IN THE PRESENCE OF A UNIFORM FLOW*

E. BÉCACHE†, A.-S. BONNET-BEN DHIA‡, AND G. LEGENDRE‡§

Abstract. This paper is devoted to the resolution of the time-harmonic linearized Galbrun equation, which models, via a mixed Lagrangian–Eulerian representation, the propagation of acoustic and hydrodynamic perturbations in a given flow of a compressible fluid. We consider here the case of a uniform subsonic flow in an infinite, two-dimensional duct. Using a limiting absorption process, we characterize the outgoing solution radiated by a compactly supported source. Then we propose a Fredholm formulation with perfectly matched absorbing layers for approximating this outgoing solution. The convergence of the approximated solution to the exact one is proved, and error estimates with respect to the parameters of the absorbing layers are derived. Several significant numerical examples are included.

Key words. aeroacoustics, Galbrun's equation, limiting absorption principle, perfectly matched layers, acoustic waveguide, modal decomposition

AMS subject classifications. 65N12, 76Q05

DOI. 10.1137/040617741

1. Introduction.

Several industrial applications are concerned with the propagation of acoustic waves in a moving fluid. Aeronautics, for instance, requires accurate simulations of acoustic radiation in the presence of a flow in order to design efficient devices for noise reduction. In this context, most of the numerical simulations consist of solving in the time domain the hyperbolic system of linearized Euler equations using finite difference schemes or, more recently, discontinuous Galerkin methods. The computational domain being necessarily finite, artificial boundary conditions are needed, and the perfectly matched layers (PMLs), introduced by Bérenger [3] in computational electromagnetics, have already been used to this end, raising some specific difficulties related to instabilities [14, 13, 19, 1, 15].

The present work differs from the previous ones as it considers the time-harmonic regime and aims at developing a finite element approach.

We use a model introduced by Galbrun [8, 18], which assumes small perturbations of an isentropic flow of a perfect fluid and whose unknown is the Lagrangian displacement perturbation, expressed in terms of Eulerian variables with respect to the mean flow. It can be viewed as an alternative to the use of the linearized Euler equations, as the perturbations of density, velocity, and pressure can be retrieved from the knowledge of both the the Lagrangian displacement perturbation and the mean flow quantities [18]. The so-called Galbrun equation is a linear partial differential equation of second order in time and space and is well suited to variational approaches. However, its numerical solution by standard (i.e., nodal) finite element methods is subject to difficulties similar to those observed for Maxwell's equations

*Received by the editors October 26, 2004; accepted for publication (in revised form) December 19, 2005; published electronically June 30, 2006.
http://www.siam.org/journals/sinum/44-3/61774.html
†Laboratoire POEMS, UMR 2706 CNRS/ENSTA/INRIA, INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le Chesnay cedex, France (eliane.becache@inria.fr).
‡Laboratoire POEMS, UMR 2706 CNRS/ENSTA/INRIA, ENSTA, 32, boulevard Victor, 75739 Paris cedex 15, France (anne-sophie.bonnet-bendhia@ensta.fr, guillaume.legendre@ensta.fr).
§Laboratoire de mathématiques appliquées, UMR CNRS 7641, Université de Versailles Saint-Quentin-en-Yvelines, 45, avenue des États-Unis, 78035 Versailles, France.

in electromagnetism. We previously proposed a regularized formulation of the time-harmonic Galbrun equation [5, 16] that allowed the use of nodal finite elements for the discretization of the problem.

The simple case of a uniform subsonic flow in an infinite two-dimensional duct is considered here, but the method should, in its principle, be extended to problems involving arbitrary flows and geometries. The two main difficulties we are confronted with when solving this problem in the time-harmonic regime are the characterization of its outgoing solution and its reduction to a bounded domain. We settle the first difficulty by a classical limiting absorption process and use a PML technique for the latter.

In a previous paper [2], we dealt with PMLs for the convected Helmholtz equation in a waveguide, and two different models of PML were analyzed: the "classical" Bérenger model and a modified model, designed to avoid a possible exponential spatial growth of the solution in the downstream layer (responsible for the instabilities in the time domain). For both models, we proved the well-posedness and convergence of the method. Yet, this last problem being scalar, acoustic waves were the only ones taken into account, whereas one of the difficulties when applying the PML technique in aeroacoustics lies in the appropriate treatment in the layers of the vorticity waves, which are convected downstream of the mean flow. In the present work, we will see that the regularization of Galbrun's equation leads to the addition of a noncompactly supported source term, accounting for the vortical effects of the flow, which will necessitate a proper, nonstandard treatment in the layers based on the modified PML model studied in [2].

The outline is the following. The problem to be solved is introduced in section 2. In section 3, a limiting absorption result is established. The problem with PML is posed and analyzed in section 4, and its convergence is subsequently proved in section 5 via the combined use of vector potentials and scalar modal analysis. Finally, numerical applications are presented in section 6.

2. The physical problem posed in an infinite waveguide. We consider Galbrun's formulation for the propagation of acoustic waves in an infinite, rigid, two-dimensional duct in the presence of a uniform mean flow of subsonic speed v_0. A time-harmonic dependence of the form $\exp(-\mathrm{i}\omega t)$, $\omega > 0$ being the pulsation, is assumed throughout the paper. The displacement perturbation then satisfies the equation and boundary condition

$$\mathrm{D}^2\boldsymbol{u} - \boldsymbol{\nabla}\,(\operatorname{div}\boldsymbol{u}) = \boldsymbol{f} \quad \text{in } \Omega, \tag{2.1}$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{2.2}$$

where Ω and $\partial\Omega$ denote, respectively, the infinite duct of height l and its rigid walls (i.e., $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 \ ; \ 0 < x_2 < l\}$) and $\boldsymbol{n}$ is a unit outward normal to $\partial\Omega$. In (2.1), the letter D stands for the time-harmonic material derivative in the subsonic uniform flow, given by $\mathrm{D}\boldsymbol{u} = -\mathrm{i}k\,\boldsymbol{u} + M\,\partial_{x_1}\boldsymbol{u}$, where the scalar $k = \frac{\omega}{c_0}$ is the wave number and the scalar $M = \frac{v_0}{c_0}$ is the Mach number ($0 < M < 1$), c_0 being the sound velocity in the mean flow. Thus, in extended form, (2.1) reads

$$-k^2\,\boldsymbol{u} - 2\mathrm{i}kM\,\partial_{x_1}\boldsymbol{u} + M^2\,\partial_{x_1}^2\boldsymbol{u} - \boldsymbol{\nabla}\,(\operatorname{div}\boldsymbol{u}) = \boldsymbol{f} \quad \text{in } \Omega.$$

Note that the equation resulting from (2.1) when the fluid is at rest (i.e., when $M = 0$) arises in several fluid-structure interaction problems (see, for instance, [12]).

An additional hypothesis is made on the compactly supported source $\boldsymbol{f}$, which is assumed to admit the Helmholtz decomposition $\boldsymbol{f} = \boldsymbol{\nabla} g_a + \mathbf{curl}\, g_h$, where g_a and g_h are also compactly supported. From a physical point of view, the source term $\boldsymbol{f}$ is meant to contain an "acoustic" part g_a, which generates irrotational perturbations (i.e., pressure fluctuations), and a vortical part g_h, which creates hydrodynamic perturbations. Note here that $\mathbf{curl}\, f = \partial_{x_2} f\, \boldsymbol{e_1} - \partial_{x_1} f\, \boldsymbol{e_2}$, where $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ are the vectors of the canonical basis of $\mathbb{R}^2$, is the vectorial form of the curl operator when applied to scalar functions. We denote by $\operatorname{curl} \boldsymbol{v} = \partial_{x_1} v_2 - \partial_{x_2} v_1$ the dual form of this operator when applied to vector fields.

The source $\boldsymbol{f}$ is also assumed to belong to the space $H(\operatorname{curl}; \Omega)$, which implies some regularity on g_a and g_h ($g_a \in H^1(\Omega)$ and $g_h \in H^2(\Omega)$, for instance).

The problem (2.1)–(2.2) admits an infinite number of solutions as long as an additional condition at infinity is not given. We are interested in the unique solution associated with the time-harmonic regime. In the next section, we characterize this solution through the study of a dissipative problem and the use of the limiting absorption principle [7].

3. Well-posedness—The limiting absorption principle.

3.1. The dissipative problem. A dissipative problem associated with (2.1)–(2.2) is readily obtained by replacing the real wave number k by a complex number $k_\varepsilon = k + \mathrm{i}\varepsilon$, where ε is a positive real number. The physical case then becomes the limiting case in which ε is equal to zero. In what follows, we prove that the unique solution of "finite energy" (that is, which belongs to the space $H^1(\Omega)^2$) of the dissipative problem converges, as ε tends to zero, in the $H^1_{\mathrm{loc}}(\Omega)^2$ sense to a limit, which will be called the "outgoing" solution of (2.1)–(2.2).

3.2. Study of the dissipative problem. We seek a function $\boldsymbol{u}^\varepsilon$ in $H^1(\Omega)^2$ satisfying

$$\mathrm{D}_\varepsilon^2 \boldsymbol{u}^\varepsilon - \boldsymbol{\nabla}\,(\operatorname{div} \boldsymbol{u}^\varepsilon) = \boldsymbol{f} \quad \text{in } \Omega, \tag{3.1}$$

$$\boldsymbol{u}^\varepsilon \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{3.2}$$

where $\mathrm{D}_\varepsilon = -\mathrm{i}k_\varepsilon + M\,\partial_{x_1}$. To be able to prove the well-posedness of this problem, it must be regularized, as proposed in [5]. To this end, we introduce the function $\psi^\varepsilon = \operatorname{curl} \boldsymbol{u}^\varepsilon$, belonging to $L^2(\Omega)$, which is a solution to the following ordinary differential equation with constant coefficients (obtained by taking the curl of (3.1)):

$$\mathrm{D}_\varepsilon^2 \psi^\varepsilon = \operatorname{curl} \boldsymbol{f} \quad \text{in } \Omega. \tag{3.3}$$

We first state a preliminary result.

LEMMA 3.1. *Equation* (3.3) *has a unique solution* ψ^ε *in* $L^2(\Omega)$. *This solution vanishes upstream of the support of the source* $\boldsymbol{f}$.

Proof. Introducing the causal Green's function of the differential operator D_ε^2,

$$G_\varepsilon(x_1) = \frac{x_1}{M^2} H(x_1)\, e^{\mathrm{i}\frac{k_\varepsilon}{M} x_1} \quad \forall x_1 \in \mathbb{R},$$

where H denotes the Heaviside function, one can derive the following particular solution to (3.3):

$$\psi^\varepsilon(x_1, x_2) = G_\varepsilon * \operatorname{curl} \boldsymbol{f}(\cdot, x_2)(x_1) = \frac{1}{M^2} \int_{-\infty}^{x_1} (x_1 - z)\, e^{\mathrm{i}\frac{k_\varepsilon}{M}(x_1 - z)} \operatorname{curl} \boldsymbol{f}(z, x_2)\, \mathrm{d}z.$$

We easily verify that this function vanishes upstream of the source and belongs to $L^2(\Omega)$ (see the appendix). Uniqueness of the solution follows from the fact that the solutions to the homogeneous equation $\mathrm{D}_\varepsilon^2\psi = 0$ are of the form

$$(c(x_2) + x_1\, d(x_2))\; e^{\mathrm{i}\frac{k_\varepsilon}{M}\, x_1} \qquad \forall (x_1, x_2) \in \Omega$$

and therefore do not belong to $L^2(\Omega)$, except for the trivial solution $c = d \equiv 0$. ∎

Now, if $\boldsymbol{u}^\varepsilon$ is a solution to (3.1)–(3.2), it clearly satisfies the so-called *regularized* or *augmented* problem

$$\begin{aligned} &\mathrm{D}_\varepsilon^2\boldsymbol{u}^\varepsilon - \boldsymbol{\nabla}\,(\operatorname{div}\boldsymbol{u}^\varepsilon) + \mathbf{curl}\;(\operatorname{curl}\boldsymbol{u}^\varepsilon - \psi^\varepsilon) = \boldsymbol{f} \quad \text{in } \Omega,\\ &\boldsymbol{u}^\varepsilon \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega,\\ &\operatorname{curl}\boldsymbol{u}^\varepsilon = \psi^\varepsilon \quad \text{on } \partial\Omega. \end{aligned} \tag{3.4}$$

Setting $V(\Omega) = \big\{\boldsymbol{v} \in H^1(\Omega)^2 \mid \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\big\}$, a variational formulation of this last problem reads as follows: *find* $\boldsymbol{u}^\varepsilon \in V(\Omega)$ *such that*

$$a_\Omega(k_\varepsilon; \boldsymbol{u}^\varepsilon, \boldsymbol{v}) = \int_\Omega (\boldsymbol{f}\cdot\overline{\boldsymbol{v}} + \psi^\varepsilon(\operatorname{curl}\overline{\boldsymbol{v}}))\;\mathrm{d}\boldsymbol{x} \qquad \forall \boldsymbol{v} \in V(\Omega), \tag{3.5}$$

where the sesquilinear form $a_\Omega(k_\varepsilon;\cdot,\cdot)$ is defined by

$$\begin{aligned} a_\Omega(k_\varepsilon; \boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \big((\operatorname{div}\boldsymbol{u})(\operatorname{div}\overline{\boldsymbol{v}}) + (\operatorname{curl}\boldsymbol{u})(\operatorname{curl}\overline{\boldsymbol{v}}) - M^2\,\partial_{x_1}\boldsymbol{u}\cdot\partial_{x_1}\overline{\boldsymbol{v}}\big)\;\mathrm{d}\boldsymbol{x}\\ + \int_\Omega \big(-{k_\varepsilon}^2\,\boldsymbol{u}\cdot\overline{\boldsymbol{v}} - 2\mathrm{i}k_\varepsilon M\,\partial_{x_1}\boldsymbol{u}\cdot\overline{\boldsymbol{v}}\big)\;\mathrm{d}\boldsymbol{x}. \end{aligned}$$

THEOREM 3.2. *The variational problem* (3.5) *is well-posed.*

Proof. Integrating by parts gives

$$\int_\Omega \partial_{x_1}\boldsymbol{u}\cdot\overline{\boldsymbol{u}}\,\mathrm{d}\boldsymbol{x} = -\int_\Omega \boldsymbol{u}\cdot\partial_{x_1}\overline{\boldsymbol{u}}\,\mathrm{d}\boldsymbol{x} = -\overline{\int_\Omega \partial_{x_1}\boldsymbol{u}\cdot\overline{\boldsymbol{u}}\,\mathrm{d}\boldsymbol{x}} \qquad \forall \boldsymbol{u} \in H^1(\Omega)^2;$$

hence $\int_\Omega \partial_{x_1}\boldsymbol{u}\cdot\overline{\boldsymbol{u}}\,\mathrm{d}\boldsymbol{x} \in \mathrm{i}\mathbb{R}$. We then have

$$\operatorname{Im}\left(-\frac{1}{k_\varepsilon}\,a_\Omega(k_\varepsilon;\boldsymbol{u},\boldsymbol{u})\right) = \int_\Omega \operatorname{Im}(k_\varepsilon)\left(|\boldsymbol{u}|^2 + \frac{|\operatorname{div}\boldsymbol{u}|^2 + |\operatorname{curl}\boldsymbol{u}|^2 - M^2\;|\partial_{x_1}\boldsymbol{u}|^2}{|k_\varepsilon|^2}\right)\mathrm{d}\boldsymbol{x}.$$

Since $M^2 < 1$ and $\operatorname{Im}(k_\varepsilon) > 0$, the sesquilinear form $a_\Omega(k_\varepsilon;\cdot,\cdot)$ is coercive on $H^1(\Omega)^2$, due to Theorem 4.1 of [6]. It is also clear that this form is continuous on the same space. Moreover, estimate (6.2) (see the appendix) allows one to establish the continuity of the antilinear form by simply using the Cauchy–Schwarz inequality. The well-posedness of problem (3.5) is then a consequence of the Lax–Milgram lemma. ∎

By construction, every solution to (3.1)–(3.2) belonging to $H^1(\Omega)$ is a solution to (3.5). The converse statement results from the following theorem.

THEOREM 3.3. *The solution* $\boldsymbol{u}^\varepsilon$ *to problem* (3.5) *is such that* $\operatorname{curl}\boldsymbol{u}^\varepsilon = \psi^\varepsilon$.

Proof. Taking as test function $\boldsymbol{v} = \mathbf{curl}\,\varphi$ with $\varphi \in \big\{\phi \in H^3(\Omega) \mid \phi_{|\partial\Omega} = 0\big\}$, we obtain, after some integration by parts and the use of boundary conditions of problem (3.5), the following orthogonality relation:

$$\int_\Omega (\operatorname{curl}\boldsymbol{u}^\varepsilon - \psi^\varepsilon)\,(\mathcal{H}_{k_\varepsilon,M}\overline{\varphi})\;\mathrm{d}\boldsymbol{x} = 0.$$

Here $\mathcal{H}_{k_\varepsilon,M}$ denotes the operator $\mathrm{D}_\varepsilon^2 - \Delta$. Owing to a density result (Theorem 1.6.2 of [11]), this relation holds for any function φ of $D(\mathcal{H}_{k_\varepsilon,M}) = H^2(\Omega) \cap H_0^1(\Omega)$. To conclude that $\operatorname{curl} \boldsymbol{u}^\varepsilon = \psi^\varepsilon$ in $L^2(\Omega)$, it suffices to show that $\mathcal{H}_{k_\varepsilon,M}$ is surjective from $D(\mathcal{H}_{k_\varepsilon,M})$ to $L^2(\Omega)$. For all φ in $D(\mathcal{H}_{k_\varepsilon,M})$, we have

$$(\mathcal{H}_{k_\varepsilon,M}\varphi,\varphi)_{L^2(\Omega)} = \int_\Omega \left(-{k_\varepsilon}^2\,|\varphi|^2 + 2\mathrm{i}k_\varepsilon M\,\partial_{x_1}\varphi\overline{\varphi} + |\boldsymbol{\nabla}\varphi|^2 - M^2\,\left|\partial_{x_1}\varphi\right|^2\right)\,\mathrm{d}\boldsymbol{x}.$$

Again, we have $\int_\Omega \partial_{x_1}\varphi\overline{\varphi}\,\mathrm{d}\boldsymbol{x} \in \mathrm{i}\mathbb{R}$, and we deduce that

$$\operatorname{Im}\left(-\frac{1}{k_\varepsilon}\,(\mathcal{H}_{k_\varepsilon,M}\varphi,\varphi)_{L^2(\Omega)}\right) = \int_\Omega \operatorname{Im}(k_\varepsilon)\left(|\varphi|^2 + \frac{|\boldsymbol{\nabla}\varphi|^2 - M^2\,\left|\partial_{x_1}\varphi\right|^2}{|k_\varepsilon|^2}\right)\,\mathrm{d}\boldsymbol{x}.$$

The surjectivity of the operator is then a consequence of the Lax–Milgram lemma. ∎

COROLLARY 3.4. *Problem* (3.1)–(3.2) *has a unique solution in* $H^1(\Omega)^2$ *which is the solution to problem* (3.5).

Proof. We choose $\boldsymbol{v} \in \mathscr{D}(\Omega)^2 \subset V(\Omega)$ in the variational formulation (3.5). Integration by parts and Theorem 3.3 imply that the unique solution $\boldsymbol{u}^\varepsilon$ to (3.5) verifies (3.1) in the distributional sense. The boundary condition (3.2) is also satisfied since $\boldsymbol{u}^\varepsilon \in V(\Omega)$. ∎

3.3. Convergence of the dissipative problem. We will prove in this subsection that, if k is not a cut-off wave number for acoustic modes, the solution $\boldsymbol{u}_\varepsilon$ to problem (3.5) converges to a limit $\boldsymbol{u}$ in $H^1_{\mathrm{loc}}(\Omega)^2$ as ε tends to zero. This limit is clearly a solution to (2.1)–(2.2) and, contrary to $\boldsymbol{u}_\varepsilon$, does not belong to $H^1(\Omega)^2$, since it does not decrease at infinity. The proof of convergence is based on a Helmholtz decomposition of the field $\boldsymbol{u}_\varepsilon$ and the use of convergence results for scalar problems.

3.3.1. Use of potentials. Let us consider the following problems: *find* $\varphi_a^\varepsilon \in H^1(\Omega)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2\varphi_a^\varepsilon - \Delta\varphi_a^\varepsilon = g_a \quad \text{in } \Omega,\\ &\partial_{\boldsymbol{n}}\varphi_a^\varepsilon = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.6}$$

and *find* $\varphi_h^\varepsilon \in L^2(\Omega)$ *such that*

$$\mathrm{D}_\varepsilon^2\varphi_h^\varepsilon = g_h \quad \text{in } \Omega. \tag{3.7}$$

Both problems are well-posed. Indeed, problem (3.6) has been studied in [4, Theorem 1], and the regularity of g_a and of the domain Ω imply that its solution φ_a^ε belongs to the space $H^2(\Omega)$. Problem (3.7) was dealt with in Lemma 3.1, the regularity of g_h implying that φ_h^ε is in $H^2(\Omega)$ (see the appendix). It then clearly ensues that the function $\boldsymbol{\nabla}\varphi_a^\varepsilon + \mathbf{curl}\,\varphi_h^\varepsilon$ is a solution to (3.1)–(3.2) (or equivalently to problem (3.5)), since $\boldsymbol{\nabla}\left(\mathrm{D}_\varepsilon^2\varphi_a^\varepsilon - \Delta\varphi_a^\varepsilon\right) + \mathbf{curl}\,\left(\mathrm{D}_\varepsilon^2\varphi_h^\varepsilon\right) = \boldsymbol{\nabla}g_a + \mathbf{curl}\,g_h = \boldsymbol{f}$ in Ω and $\partial_{\boldsymbol{n}}\varphi_a^\varepsilon + \mathbf{curl}\,\varphi_h^\varepsilon \cdot \boldsymbol{n} = 0$ on $\partial\Omega$ (the function g_h being compactly supported, φ_h^ε vanishes on the boundary $\partial\Omega$, which implies that $\mathbf{curl}\,\varphi_h^\varepsilon \cdot \boldsymbol{n} = 0$ on $\partial\Omega$). Hence, the uniqueness of the solution to (3.5) implies that

$$\boldsymbol{u}^\varepsilon = \boldsymbol{\nabla}\varphi_a^\varepsilon + \mathbf{curl}\,\varphi_h^\varepsilon. \tag{3.8}$$

We now prove the convergence of the respective solutions to problems (3.6) and (3.7) as ε tends to zero.

3.3.2. Limit and convergence of the acoustic problem. In order to obtain the limit in $H^2_{\rm loc}(\Omega)$ of φ_a^ε as ε tends to zero, we use some theoretical results previously established in [4] for scalar problems of the same type. First, problem (3.6) is equivalently set in a bounded domain Ω_b, which contains the supports of g_a and g_h and is situated in between the two vertical boundaries $\Sigma_\pm$, respectively located at $x_1 = x_\pm$. To this end, we make use of the Dirichlet-to-Neumann (DtN) operators $T_\pm^{N,\varepsilon} : H^{1/2}(\Sigma_\pm) \to H^{-1/2}(\Sigma_\pm)$ (the superscript N refers here to the Neumann boundary condition in problem (3.11)), defined by $T_\pm^{N,\varepsilon}\phi = \mp \mathrm{i} \sum_{n=0}^{+\infty} \beta_n^{\varepsilon\pm} \, (\phi, \mathrm{C}_n)_{L^2(\Sigma_\pm)} \, \mathrm{C}_n(x_2)$, where

$$\beta_n^{\varepsilon\pm} = \frac{-k_\varepsilon M \pm \sqrt{{k_\varepsilon}^2 - \frac{n^2\pi^2}{l^2}(1-M^2)}}{1-M^2}, \tag{3.9}$$

the square root being defined by $\sqrt{z} = \sqrt{|z|}\, e^{\mathrm{i}\frac{\arg(z)}{2}}$, $0 \le \arg(z) < 2\pi$, and where

$$\mathrm{C}_0(x_2) = \sqrt{\frac{1}{l}} \quad \text{and} \quad \mathrm{C}_n(x_2) = \sqrt{\frac{2}{l}} \cos\Big(\frac{n\pi}{l} x_2\Big) \ \ \forall n \in \mathbb{N}^*. \tag{3.10}$$

An equivalent formulation of problem (3.6) is then as follows: *find* $\varphi_a^\varepsilon \in H^1(\Omega_b)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2 \varphi_a^\varepsilon - \Delta \varphi_a^\varepsilon = g_a \quad \textit{in } \Omega_b, \\ &\partial_{\boldsymbol{n}} \varphi_a^\varepsilon = 0 \quad \textit{on } \partial\Omega \cap \partial\Omega_b, \\ &\partial_{\boldsymbol{n}} \varphi_a^\varepsilon = -T_\pm^{N,\varepsilon} \varphi_a^\varepsilon \quad \textit{on } \Sigma_\pm. \end{aligned} \tag{3.11}$$

We now can formally derive a limiting problem for (3.11). Observe that, because of the definition of the complex square root, one has

$$\lim_{\varepsilon\to 0} \sqrt{{k_\varepsilon}^2 - \frac{n^2\pi^2}{l^2}(1-M^2)} = \begin{cases} \sqrt{k^2 - \dfrac{n^2\pi^2}{l^2}(1-M^2)} \in \mathbb{R}_+ & \text{if } k \ge \dfrac{n\pi}{l}\sqrt{1-M^2}, \\ \mathrm{i}\sqrt{\dfrac{n^2\pi^2}{l^2}(1-M^2) - k^2} \in \mathrm{i}\mathbb{R}_+ & \text{if } k < \dfrac{n\pi}{l}\sqrt{1-M^2}. \end{cases}$$

Hence, the respective limits $\beta_n^\pm \ \forall n \in \mathbb{N}$ of axial wave numbers $\beta_n^{\varepsilon\pm}$ are defined by

$$\beta_n^\pm = \begin{cases} \dfrac{-kM \pm \sqrt{k^2 - \frac{n^2\pi^2}{l^2}(1-M^2)}}{1-M^2} & \text{if } k \ge \dfrac{n\pi}{l}\sqrt{1-M^2}, \\ \dfrac{-kM \pm \mathrm{i}\sqrt{\frac{n^2\pi^2}{l^2}(1-M^2) - k^2}}{1-M^2} & \text{if } k < \dfrac{n\pi}{l}\sqrt{1-M^2}. \end{cases} \tag{3.12}$$

The limiting problem to be considered is then as follows: *find* $\varphi_a \in H^1(\Omega_b)$ *such that*

$$\begin{aligned} &\mathrm{D}^2 \varphi_a - \Delta \varphi_a = g_a \quad \textit{in } \Omega_b, \\ &\partial_{\boldsymbol{n}} \varphi_a = 0 \quad \textit{on } \partial\Omega \cap \partial\Omega_b, \\ &\partial_{\boldsymbol{n}} \varphi_a = -T_\pm^N \varphi_a \quad \textit{on } \Sigma_\pm, \end{aligned} \tag{3.13}$$

with the following obvious definition for the DtN operators $T_\pm^N : H^{1/2}(\Sigma_\pm) \to H^{-1/2}(\Sigma_\pm)$, $T_\pm^N \phi = \mp \mathrm{i} \sum_{n=0}^{+\infty} \beta_n^\pm \, (\phi, \mathrm{C}_n)_{L^2(\Sigma_\pm)} \, \mathrm{C}_n(x_2)$.

THEOREM 3.5. *Problem* (3.13) *is well-posed, except if* $k = k_n$ *for* $n \in \mathbb{N}$, *with* $k_n = \sqrt{1-M^2}\,\frac{n\pi}{l}$.

A proof of this theorem is available in [2, Theorem 2.2]. The scalars k_n, $n \in \mathbb{N}$, are called the cut-off wave numbers of the modes. We are now in a position to prove a convergence result for problem (3.6).

THEOREM 3.6. *If $k \neq k_n$ $\forall n \in \mathbb{N}$, the solution φ_a^ε to problem* (3.6) *tends in $H^2(\Omega_b)$ to φ_a as ε tends to zero, φ_a being the solution to* (3.13).

Proof. Theorem 4 of [4] gives the convergence of φ_a^ε to φ_a in $H^1(\Omega_b)$. We then have $(1-M^2)\,\partial_{x_1}^2\varphi_a^\varepsilon + \partial_{x_2}^2\varphi_a^\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} (1-M^2)\,\partial_{x_1}^2\varphi_a + \partial_{x_2}^2\varphi_a$ in $L^2(\Omega_b)$. The domain Ω_b being convex, we deduce $\varphi_a^\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} \varphi_a$ in $H^2(\Omega_b)$ (see [10]). ∎

3.3.3. Limit and convergence of the hydrodynamic problem. The solution to problem (3.7) is explicitly given by the convolution product $\varphi_h^\varepsilon(x_1,x_2) = G_\varepsilon * g_h(.,x_2)(x_1)\ \forall (x_1,x_2) \in \Omega$, where the kernel G_ε denotes the causal Green's function of the differential operator D_ε^2. Introducing $G(x_1) = \frac{x_1}{M^2} H(x_1)\, e^{\mathrm{i}\frac{k}{M}x_1}\ \forall x_1 \in \mathbb{R}$ as the formal limit of G_ε as ε tends to zero, one can show that G_ε converges to G in $L^2_{\mathrm{loc}}(\mathbb{R})$. Consider now $\varphi_h(x_1,x_2) = G * g_h(.,x_2)(x_1)$, which is a solution to the limiting problem *find $\varphi_h \in L^2_{loc}(\Omega)$ such that*

$$\mathrm{D}^2\varphi_h = g_h \quad \text{in } \Omega. \tag{3.14}$$

One has $|\varphi_h^\varepsilon - \varphi_h| = |(G_\varepsilon - G) * g_h(.,x_2)|\ \forall x_2 \in [0,l]$, and, using the Cauchy–Schwarz inequality,

$$\|\varphi_h^\varepsilon - \varphi_h\|_{L^2(\Omega_b)} \leq \left(\int_{x_-}^{x_+} |G_\varepsilon(z) - G(z)|^2 \,\mathrm{d}z \right)^{1/2} \|g_h\|_{L^2(\Omega)}.$$

The convergence of G_ε to G in $L^2_{\mathrm{loc}}(\Omega)$ implies that φ_h^ε converges to φ_h in $L^2(\Omega_b)$ as ε tends to zero. Since $g_h \in H^2(\Omega)$ and using classical properties of the convolution of distributions, the above estimate is also true in the $H^2(\Omega)$ norm and we deduce the following theorem.

THEOREM 3.7. *The solution φ_h^ε to problem* (3.7) *tends to φ_h in $H^2(\Omega_b)$ as ε tends to zero.*

3.3.4. Conclusion. We finally infer from Theorems 3.6 and 3.7 the following result.

THEOREM 3.8. *If k is not a cut-off wave number, the solution $\boldsymbol{u}^\varepsilon$ to problem* (3.1)–(3.2) *tends to $\boldsymbol{u} = \boldsymbol{\nabla}\varphi_a + \mathbf{curl}\,\varphi_h$ in $H^1(\Omega_b)^2$ as ε tends to zero, where φ_a is the unique solution to* (3.13) *and $\varphi_h(x_1,x_2) = G * g_h(.,x_2)(x_1)$.*

The potential φ_a can be extended via a modal expansion to the whole domain Ω. The field $\boldsymbol{u}$ is therefore also defined in the whole duct Ω, where it obviously satisfies (2.1) and (2.2). From now on, this field will be referred to as the outgoing solution to problem (2.1)–(2.2).

COROLLARY 3.9. *The function $\psi^\varepsilon = \operatorname{curl}\boldsymbol{u}^\varepsilon$ tends to $\operatorname{curl}\boldsymbol{u}$ in $L^2(\Omega_b)$. We set $\psi = \operatorname{curl}\boldsymbol{u}$.*

In the remainder of the paper, we assume that k is not a cut-off wave number:

$$k \neq k_n \quad \forall n \in \mathbb{N}. \tag{3.15}$$

3.3.5. Another characterization of $\boldsymbol{\varphi_h}$. We observe that using the Helmholtz decomposition (3.8) of $\boldsymbol{u}^\varepsilon$ in the regularized problem (3.4) would lead, as $\operatorname{curl}(\mathbf{curl}) = -\Delta$ for a scalar field, to the following problem for the hydrodynamic potential:

find $\varphi_h^\varepsilon \in H^1(\Omega)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2 \varphi_h^\varepsilon - \Delta \varphi_h^\varepsilon = g_h + \psi^\varepsilon \quad \text{in } \Omega, \\ &\varphi_h^\varepsilon = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.16}$$

in place of (3.7). Nevertheless, problems (3.7) and (3.16) are equivalent, since $-\Delta\varphi_h^\varepsilon = \psi^\varepsilon$, but using the latter to find the limit of φ_h^ε as ε tends to zero is less natural and more delicate than what has been proposed in subsection 3.3.3. Even still, we detail this alternative approach here, since it provides another characterization of the potential φ_h, which will prove to be useful in what follows.

Notice that problem (3.16) is very similar to (3.6), except for the homogeneous Dirichlet boundary condition, which replaces the homogeneous Neumann boundary condition of (3.6), and for the right-hand side term, which does not have compact support. To prove the convergence of problem (3.16), we define a similar problem with compactly supported data, which then fits into the previous framework. This leads to the following statement.

LEMMA 3.10. *The limit* φ_h *of the potential* φ_h^ε *has the following two equivalent characterizations:*

1. φ_h *is the unique solution to* (3.14) *which vanishes upstream of the source* g_h.
2. $\varphi_h = \widetilde{\varphi}_h + \chi\zeta$, *where* $\widetilde{\varphi}_h$ *is the unique (outgoing) solution to problem* (3.21) *and* χ *and* ζ *are defined below.*

Proof. We first introduce the function $\psi^{\varepsilon,\infty}$, which coincides with ψ^ε downstream of the support of curl $\boldsymbol{f}$ and is the sum of two functions with separated variables. More precisely, if $\mathrm{supp}(\mathrm{curl}\,\boldsymbol{f}) \subset \{(x_1,x_2) \in \Omega \mid d_- < x_1 < d_+\}$, we have by definition that $\psi^{\varepsilon,\infty}(x_1,x_2) = \psi^\varepsilon(x_1,x_2)$ for $x_1 > d_+$ and, from (6.3), that

$$\psi^{\varepsilon,\infty}(x_1,x_2) = (a_\varepsilon(x_2) + x_1\, b_\varepsilon(x_2))\, e^{\mathrm{i}\frac{k_\varepsilon}{M}x_1} \quad \forall (x_1,x_2) \in \Omega.$$

Notice that $\psi^{\varepsilon,\infty}$ does not generally vanish upstream of the support of the source, contrary to ψ^ε. Taking advantage of the particular form of $\psi^{\varepsilon,\infty}$, one can explicitly determine a function ζ^ε satisfying the problem *find* $\zeta^\varepsilon \in H^1(\Omega)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2 \zeta^\varepsilon - \Delta \zeta^\varepsilon = \psi^{\varepsilon,\infty} \quad \text{in } \Omega, \\ &\zeta^\varepsilon = 0 \quad \text{on } \partial\Omega, \end{aligned}$$

which is sought in the form $\zeta^\varepsilon(x_1,x_2) = (A_\varepsilon(x_2) + x_1\, B_\varepsilon(x_2))\, e^{\mathrm{i}\frac{k_\varepsilon}{M}x_1}\ \forall (x_1,x_2) \in \Omega$. This leads to the solution of the following two problems:

$$\begin{aligned} &-B_\varepsilon''(x_2) + \frac{{k_\varepsilon}^2}{M^2} B_\varepsilon(x_2) = b_\varepsilon(x_2), \\ &B_\varepsilon(0) = B_\varepsilon(l) = 0, \end{aligned} \tag{3.17}$$

$$\begin{aligned} &-A_\varepsilon''(x_2) + \frac{{k_\varepsilon}^2}{M^2} A_\varepsilon(x_2) = 2\mathrm{i}\frac{k_\varepsilon}{M}\, B_\varepsilon(x_2) + a_\varepsilon(x_2), \\ &A_\varepsilon(0) = A_\varepsilon(l) = 0. \end{aligned} \tag{3.18}$$

One can easily prove that problem (3.17) is well-posed in the space $H_0^1([0,l])$ and consequently compute the function B_ε, which allows us in turn to determine A_ε (using the same argument) and finally find the function ζ^ε.

Now using their explicit formulas (see the appendix), one can show that a_ε and b_ε, respectively, tend, as ε tends to zero, to some functions a and b which are related

to $\psi = \operatorname{curl} \boldsymbol{u}$ through the relation $\psi(x_1, x_2) = (a(x_2) + x_1\, b(x_2))\, e^{\mathrm{i}\frac{k}{M}x_1}\ \forall (x_1, x_2) \in]d_+, +\infty[\times[0, l]$. We are then able to compute the respective limits of A_ε and B_ε, denoted A and B, by simply solving similar problems. We have consequently defined a function $\zeta(x_1, x_2) = (A(x_2) + x_1\, B(x_2))\, e^{\mathrm{i}\frac{k}{M}x_1}\ \forall (x_1, x_2) \in \Omega$, which satisfies $\mathrm{D}^2\zeta - \Delta\zeta = \psi\ \forall (x_1, x_2) \in]d_+, +\infty[\times[0, l]$ and which is the limit of ζ^ε as ε tends to zero.

Now, consider a cut-off function $\chi \in C^\infty(\mathbb{R})$ such that $\chi(x_1) = 1$ if $x_1 > d_+$ and vanishing for $x_1 < d_-$. It is easily seen that $\widetilde{\varphi}_h^\varepsilon = \varphi_h^\varepsilon - \chi\zeta^\varepsilon$ satisfies the following problem: *find* $\widetilde{\varphi}_h^\varepsilon \in H^1(\Omega)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2\widetilde{\varphi}_h^\varepsilon - \Delta\widetilde{\varphi}_h^\varepsilon = \widetilde{g}_h^\varepsilon \quad \textit{in } \Omega,\\ &\widetilde{\varphi}_h^\varepsilon = 0 \quad \textit{on } \partial\Omega, \end{aligned}$$

where $\widetilde{g}_h^\varepsilon = g_h + \psi^\varepsilon - \big(\mathrm{D}_\varepsilon^2(\chi\zeta^\varepsilon) - \Delta(\chi\zeta^\varepsilon)\big)$. By construction, $\psi^\varepsilon - \big(\mathrm{D}_\varepsilon^2(\chi\zeta^\varepsilon) - \Delta(\chi\zeta^\varepsilon)\big)$ has a compact support contained in Ω_b, and therefore $\widetilde{g}_h^\varepsilon$ is compactly supported in Ω_b. Using DtN operators, this last problem may equivalently be rewritten as a problem set in Ω_b: *find* $\widetilde{\varphi}_h^\varepsilon \in H^1(\Omega_b)$ *such that*

$$\begin{aligned} &\mathrm{D}_\varepsilon^2\widetilde{\varphi}_h^\varepsilon - \Delta\widetilde{\varphi}_h^\varepsilon = \widetilde{g}_h^\varepsilon \quad \textit{in } \Omega_b,\\ &\widetilde{\varphi}_h^\varepsilon = 0 \quad \textit{on } \partial\Omega \cap \partial\Omega_b,\\ &\partial_{\boldsymbol{n}}\widetilde{\varphi}_h^\varepsilon = -T_\pm^{D,\varepsilon}\widetilde{\varphi}_h^\varepsilon \quad \textit{on } \Sigma_\pm, \end{aligned} \tag{3.19}$$

where the operators $T_\pm^{D,\varepsilon} : H^{1/2}(\Sigma_\pm) \to H^{-1/2}(\Sigma_\pm)$ (the superscript D refers to the Dirichlet boundary condition in (3.19)) are defined by

$$T_\pm^{D,\varepsilon}\phi = \mp\mathrm{i}\sum_{n=1}^{+\infty} \beta_n^{\varepsilon\pm}\ (\phi, \mathrm{S}_n)_{L^2(\Sigma_\pm)}\ \mathrm{S}_n(x_2),$$

the numbers $\beta_n^{\varepsilon\pm}$, $n \in \mathbb{N}^*$, being defined in (3.9) and with

$$\mathrm{S}_n(x_2) = \sqrt{\frac{2}{l}}\ \sin\Big(\frac{n\pi}{l}x_2\Big)\ \ \forall n \in \mathbb{N}^*. \tag{3.20}$$

One can prove that problem (3.19) is well-posed, due to hypothesis (3.15). It is now possible to pass to the limit as ε tends to zero in the same way as previously done for the acoustic potential φ_a^ε and finally show that $\lim_{\varepsilon\to 0}\widetilde{\varphi}_h^\varepsilon = \widetilde{\varphi}_h$ in $H^1(\Omega_b)$, where $\widetilde{\varphi}_h$ is the solution to the following: *find* $\widetilde{\varphi}_h \in H^1(\Omega_b)$ *such that*

$$\begin{aligned} &\mathrm{D}^2\widetilde{\varphi}_h - \Delta\widetilde{\varphi}_h = \widetilde{g}_h \quad \textit{in } \Omega_b,\\ &\widetilde{\varphi}_h = 0 \quad \textit{on } \partial\Omega \cap \partial\Omega_b,\\ &\partial_{\boldsymbol{n}}\widetilde{\varphi}_h = -T_\pm^D\widetilde{\varphi}_h \quad \textit{on } \Sigma_\pm, \end{aligned} \tag{3.21}$$

where $\widetilde{g}_h = g_h + \psi - \big(\mathrm{D}^2(\chi\zeta) - \Delta(\chi\zeta)\big)$, the DtN operators $T_\pm^D : H^{1/2}(\Sigma_\pm) \to H^{-1/2}(\Sigma_\pm)$ being defined by $T_\pm^D\phi = \mp\mathrm{i}\sum_{n=0}^{+\infty}\beta_n^\pm\ (\phi, \mathrm{S}_n)_{L^2(\Sigma_\pm)}\ \mathrm{S}_n(x_2)$.

Using that $\lim_{\varepsilon\to 0}\zeta^\varepsilon = \zeta$, one has $\lim_{\varepsilon\to 0}\varphi_h^\varepsilon = \lim_{\varepsilon\to 0}(\widetilde{\varphi}_h^\varepsilon + \chi\zeta^\varepsilon) = \widetilde{\varphi}_h + \chi\zeta$. By uniqueness of the limit, we conclude that $\varphi_h = \widetilde{\varphi}_h + \chi\zeta$. ∎

4. Setting of the problem with perfectly matched layers. Our goal in this section is to develop a finite element method to compute an approximation of the outgoing solution $\boldsymbol{u}$ to problem (2.1)–(2.2). To do so, we must address two main difficulties. First, this problem is set in an unbounded domain. Second, the operator in Galbrun's equation is not coercive.

As already seen during the study of the dissipative problem, the coerciveness can be restored by applying a regularization technique. On the other hand, we use PMLs (see, for example, [2] and references therein for a presentation of this methodology) in order to truncate the computational domain. A posteriori, the regularization will prove to be necessary not only for the finite element method, but also for the PML method (see subsection 5.5).

In a previous paper [2], we proved the convergence of the solutions to PML formulations of the scalar problem (3.13). Two different models of PMLs were studied: a "classical" one, derived directly from Bérenger's original model, and a modified one, designed to avoid a possible growing of the solution in the downstream layer (due to the presence of the so-called inverse upstream modes). This last property will be useful for the vectorial problem at hand, and the modified PML model (to be recalled in the next subsection) will be the only one considered here.

4.1. The PML formulation. We introduce the bounded domain Ω^L, composed of domain Ω_b and surrounding layers $\Omega^L_\pm$, which are respectively defined by $\Omega^L_- = \{(x_1,x_2)\in\Omega\,|\,x_- - L < x_1 < x_-\}$ and $\Omega^L_+ = \{(x_1,x_2)\in\Omega\,|\,x_+ < x_1 < x_+ + L\}$, the external vertical boundaries of the layers $\Omega^L_\pm$ being respectively denoted by $\Sigma^L_\pm$.

The so-called modified PML model consists of the formal transformation of the differential operator

$$\partial_{x_1} \longrightarrow \alpha(x_1)\,\partial_{x_1} + \mathrm{i}\lambda(x_1) \tag{4.1}$$

in the governing equations of the problem. The complex function α is assumed to be unity in Ω_b and, in order to simplify the study, constant and equal to the complex scalar α^*, satisfying the hypotheses

$$\mathrm{Re}(\alpha^*) > 0, \quad \mathrm{Im}(\alpha^*) < 0 \tag{4.2}$$

in $\Omega\backslash\Omega_b$ (see [2] for a justification), but it can generally depend on the coordinate x_1 in the layers. The function λ is assumed to be zero in Ω_b and to be constant and equal to

$$\lambda^* = -\frac{kM}{1-M^2} \tag{4.3}$$

in $\Omega\backslash\Omega_b$. In Bérenger's model, one has $\lambda\equiv 0$. Note that the results obtained subsequently can be extended to the case of a varying coefficient α^*, as done in [2].

As a consequence of the transformation (4.1), the various modified operators will now be indexed by α and λ.

As seen in section 3, Galbrun's equation must be regularized in order to be numerically solved in a stable fashion by a nodal (i.e., H^1-conforming) finite element method. The PML formulation of the problem is no exception to this rule. Consequently, we introduce a function $\psi_{\alpha,\lambda}$, defined as the unique solution to $\mathrm{D}^2_{\alpha,\lambda}\psi_{\alpha,\lambda} = \mathrm{curl}\,\boldsymbol{f}$ in Ω, which vanishes upstream of the source. One can easily verify that $\psi_{\alpha,\lambda} = \psi$ in Ω_b and that, $\forall (x_1,x_2)\in\Omega_+$,

$$\psi_{\alpha,\lambda}(x_1,x_2) = e^{\mathrm{i}\left(\frac{k}{M}\,x_+ + \left(\frac{k}{M}-\lambda^*\right)\frac{x_1-x_+}{\alpha^*}\right)}\left(a(x_2) + \left(x_+ + \frac{(x_1-x_+)}{\alpha^*}\right)b(x_2)\right).$$

We then set the problem for the approximated displacement field: *find* $\boldsymbol{u}^L\in H^1(\Omega^L)^2$

such that

$$
\begin{aligned}
&\mathrm{D}^2_{\alpha,\lambda}\boldsymbol{u}^L - \boldsymbol{\nabla}_{\alpha,\lambda}\left(\mathrm{div}_{\alpha,\lambda}\,\boldsymbol{u}^L\right) + \mathbf{curl}_{\,\alpha,\lambda}\left(\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L - \psi_{\alpha,\lambda}\right) = \boldsymbol{f} \quad \text{in } \Omega^L,\\
&\boldsymbol{u}^L\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L,\\
&\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L = \psi_{\alpha,\lambda} \quad \text{on } \partial\Omega^L.
\end{aligned}
\tag{4.4}
$$

4.2. Well-posedness. We first establish a variational formulation of problem (4.4): *find* $\boldsymbol{u}^L \in V(\Omega^L) = \left\{\boldsymbol{v}\in H^1(\Omega^L)^2 \mid \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega^L\right\}$ *such that*

$$
a_{\Omega^L}(\boldsymbol{u}^L,\boldsymbol{v}) + b_{\Omega^L}(\boldsymbol{u}^L,\boldsymbol{v}) = l_{\Omega^L}(\boldsymbol{v})\ \forall \boldsymbol{v}\in V(\Omega^L),
\tag{4.5}
$$

where

$$
\begin{aligned}
a_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}) &= \int_{\Omega^L}\left(\boldsymbol{u}\cdot\overline{\boldsymbol{v}} + \frac{1}{\alpha}\left(\mathrm{div}_{\alpha,\lambda}\,\boldsymbol{u}\ \mathrm{div}_{\alpha,-\lambda}\,\overline{\boldsymbol{v}} + \mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}\,\mathrm{curl}_{\,\alpha,-\lambda}\overline{\boldsymbol{v}}\right)\right.\\
&\qquad\left. - \alpha M^2\,\partial_{x_1}\boldsymbol{u}\cdot\partial_{x_1}\overline{\boldsymbol{v}}\right)\mathrm{d}\boldsymbol{x},\\
b_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}) &= \int_{\Omega^L}\left(\frac{1}{\alpha}\left(-k^2 + 2kM\lambda - M^2\lambda^2 - \alpha\right)\,\boldsymbol{u}\cdot\overline{\boldsymbol{v}} + \left(\mathrm{i}\lambda M^2 - 2\mathrm{i}kM\right)\,\partial_{x_1}\boldsymbol{u}\cdot\overline{\boldsymbol{v}}\right.\\
&\qquad\left. - \mathrm{i}\lambda M^2\,\boldsymbol{u}\cdot\,\partial_{x_1}\overline{\boldsymbol{v}}\right)\mathrm{d}\boldsymbol{x},\\
l_{\Omega^L}(\boldsymbol{v}) &= \int_{\Omega^L}\frac{1}{\alpha}\left(\boldsymbol{f}\cdot\overline{\boldsymbol{v}} + \psi_{\alpha,\lambda}\,\mathrm{curl}_{\,\alpha,-\lambda}\overline{\boldsymbol{v}}\right)\,\mathrm{d}\boldsymbol{x} - \int_{\Sigma^L_{\pm}} M^2\psi_{\alpha,\lambda}\overline{v_2}(\boldsymbol{n}\cdot\boldsymbol{e_1})\,\mathrm{d}\sigma.
\end{aligned}
$$

THEOREM 4.1. *If the assumption* (4.2) *is satisfied, the variational problem* (4.5) *is of Fredholm type.*

Proof. We prove that the sesquilinear form $a_{\Omega^L}(\cdot\,,\cdot)$ defines, via the Riesz representation theorem, an operator which is the sum of an isomorphism and a compact operator on $V(\Omega^L)$. Let us write $a_{\Omega^L}(\cdot\,,\cdot)$ as the sum

$$
a_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}) = a^0_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}) + \lambda\, a^1_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}) + \lambda^2\, a^2_{\Omega^L}(\boldsymbol{u},\boldsymbol{v}),
$$

where the sesquilinear forms $a^i_{\Omega^L}(\cdot\,,\cdot)$, $i = 0,1,2$, are independent of λ. The sesquilinear form $a^0_{\Omega^L}(\cdot\,,\cdot)$ is coercive on $V(\Omega^L)$. Indeed, $\forall\boldsymbol{u}$, $\boldsymbol{v}$ in $V(\Omega^L)$, we have

$$
\int_{\Omega^L}\left(\mathrm{curl}_{\,\alpha,0}\boldsymbol{u}\,\mathrm{curl}_{\,\alpha,0}\overline{\boldsymbol{v}} + \mathrm{div}_{\alpha,0}\,\boldsymbol{u}\ \mathrm{div}_{\alpha,0}\,\overline{\boldsymbol{v}}\right)\,\mathrm{d}\boldsymbol{x} = \int_{\Omega^L}\boldsymbol{\nabla}_{\alpha,0}\boldsymbol{u} : \boldsymbol{\nabla}_{\alpha,0}\overline{\boldsymbol{v}}\,\mathrm{d}\boldsymbol{x},
$$

and due to assumption (4.2), we have $\forall\boldsymbol{u}$ of $V(\Omega^L)$

$$
\begin{aligned}
\mathrm{Re}\left(a^0_{\Omega^L}(\boldsymbol{u},\boldsymbol{u})\right) &= \int_{\Omega^L}\left(|\boldsymbol{u}|^2 + \mathrm{Re}(\alpha)(1-M^2)\,|\partial_{x_1}\boldsymbol{u}|^2 + \mathrm{Re}\left(\frac{1}{\alpha}\right)|\partial_{x_2}\boldsymbol{u}|^2\right)\,\mathrm{d}\boldsymbol{x}\\
&\geq C_\alpha\,\|\boldsymbol{u}\|_{H^1(\Omega^L)^2},
\end{aligned}
$$

with $C_\alpha = \min\left(1-M^2, \mathrm{Re}(\alpha^*)(1-M^2), \mathrm{Re}\left(\frac{1}{\alpha^*}\right)\right)$.

On the other hand, the forms $a^1_{\Omega^L}(\cdot\,,\cdot)$ and $a^2_{\Omega^L}(\cdot\,,\cdot)$ both define a compact operator on $V(\Omega^L)$, due to the compact embedding of $H^1(\Omega^L)$ into $L^2(\Omega^L)$. The same argument is used to show that the bounded operator defined on $H^1(\Omega^L)^2$ by the sesquilinear form $b_{\Omega^L}(\cdot\,,\cdot)$ is compact, which ends the proof. ∎

We now prove the equivalence between the variational formulation (4.5) and the following (strong) problem: *find* $\boldsymbol{u}^L \in H^1(\Omega^L)^2$ *such that*

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\boldsymbol{u}^L - \boldsymbol{\nabla}_{\alpha,\lambda}(\mathrm{div}_{\alpha,\lambda}\,\boldsymbol{u}^L) = \boldsymbol{f} \quad \text{in } \Omega^L,\\ &\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L = \psi_{\alpha,\lambda} \quad \text{in } \Omega^L,\\ &\boldsymbol{u}^L\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L. \end{aligned}$$

LEMMA 4.2. *There exists a strictly positive constant* δ*, depending on* k*,* M*, and* $\theta = \arg(\alpha^*)$*, such that if* $L/|\alpha^*| \geq \delta$*, any solution to variational problem* (4.5) *is such that* $\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L = \psi_{\alpha,\lambda}$ *in* Ω^L*.*

Proof. Considering a test function of the form $\boldsymbol{v} = \mathbf{curl}_{\,\overline{\alpha},-\lambda}\phi$ with $\phi \in H^3(\Omega^L)\cap H^1_0(\Omega^L)$ in (4.5), we obtain, after integrating by parts,

$$\begin{aligned}\int_{\Omega^L}\frac{1}{\alpha}\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L\left(\mathrm{D}^2_{\alpha,-\lambda}\overline{\phi} - \Delta_{\alpha,-\lambda}\overline{\phi}\right)\mathrm{d}\boldsymbol{x} &= \int_{\Omega^L}\frac{1}{\alpha}\left(\overline{\phi}\,\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{f} - \psi_{\alpha,\lambda}\Delta\overline{\varphi}\right)\mathrm{d}\boldsymbol{x}\\ &\quad + \int_{\Sigma^L_-\cup\Sigma^L_+}\alpha M^2\psi_{\alpha,\lambda}\partial_{x_1}\overline{\phi}(\boldsymbol{n}\cdot\boldsymbol{e_1})\,\mathrm{d}\sigma.\end{aligned}$$

Knowing that the function $\psi_{\alpha,\lambda}$ satisfies $\mathrm{D}^2_{\alpha,\lambda}\psi_{\alpha,\lambda} = \mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{f}$ in Ω^L, we finally have

$$\int_{\Omega^L}\frac{1}{\alpha}\left(\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L - \psi_{\alpha,\lambda}\right)\left(\mathrm{D}^2_{\alpha,-\lambda}\overline{\phi} - \Delta_{\alpha,-\lambda}\overline{\phi}\right)\mathrm{d}\boldsymbol{x} = 0.$$

Owing to a density argument, this equality holds for any function ϕ in $H^2(\Omega^L)\cap H^1_0(\Omega^L)$. Then, if $L/|\alpha^*|$ is large enough, the operator $\mathrm{D}^2_{\alpha,-\lambda}-\Delta_{\alpha,-\lambda}$ is surjective from $H^2(\Omega^L)\cap H^1_0(\Omega^L)$ onto $L^2(\Omega^L)$ (see subsection 5.1), and we deduce that $\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L = \psi_{\alpha,\lambda}$ almost everywhere in Ω^L. ∎

We finish this study by an existence and uniqueness result. In addition to hypothesis (3.15) on the wave number k, we now assume that

$$k \neq \frac{n\pi}{l} \quad \forall n \in \mathbb{N}. \tag{4.6}$$

The need for this second assumption will be made clear in the next section.

THEOREM 4.3. *Assume that hypotheses* (4.2) *are satisfied and choice* (4.3) *is made. Then, there exists a strictly positive constant* δ *such that problem* (4.5) *admits a unique solution if* $L/|\alpha^*| \geq \delta$*.*

Proof. The problem being of Fredholm type, proving the existence of its solution amounts to proving uniqueness. Let $\boldsymbol{w} \in H^1(\Omega^L)^2$ denote the difference between two solutions to problem (4.5). It verifies that

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\boldsymbol{w} - \boldsymbol{\nabla}_{\alpha,\lambda}(\mathrm{div}_{\alpha,\lambda}\,\boldsymbol{w}) = \boldsymbol{0} \quad \text{in } \Omega^L,\\ &\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{w} = 0 \quad \text{in } \Omega^L,\\ &\boldsymbol{w}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L. \end{aligned}$$

We then consider the function $\check{\boldsymbol{w}}(x_1,x_2) = \boldsymbol{w}(x_1,x_2)\,e^{\mathrm{i}\lambda x_1}$, so that $\mathrm{curl}_{\,\alpha,0}\check{\boldsymbol{w}} = \mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{w}\,e^{\mathrm{i}\lambda x_1}$, and we use the following result.

THEOREM 4.4. *A function* $\boldsymbol{v}$ *in* $L^2(\Omega^L)^2$ *satisfies* $\mathrm{curl}_{\,\alpha,0}\boldsymbol{v} = 0$ *in* Ω^L *if and only if there exists a function* ϕ *in* $H^1(\Omega^L)$ *such that* $\boldsymbol{v} = \boldsymbol{\nabla}_{\alpha,0}\phi$*. This function is unique up to an additive constant.*

A proof of the above theorem can be obtained by slightly modifying the proof of Theorem 2.9 of [9, Chapter 1].

The field $\check{\boldsymbol{w}}$ then derives from a scalar potential ϕ, which is a solution to

$$\begin{aligned} &\boldsymbol{\nabla}_{\alpha,0}\left(\mathrm{D}^2_{\alpha,0}\phi - \Delta_{\alpha,0}\phi\right) = 0 \quad \text{in } \Omega^L,\\ &\boldsymbol{\nabla}_{\alpha,0}\phi\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L; \end{aligned}$$

hence

$$\begin{aligned} &\mathrm{D}^2_{\alpha,0}\phi - \Delta_{\alpha,0}\phi = C \quad \text{in } \Omega^L,\\ &\boldsymbol{\nabla}_{\alpha,0}\phi\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L, \end{aligned} \tag{4.7}$$

where C is a complex constant. This last problem is well-posed if $L/|\alpha^*|$ is large enough (this will be proved later; see subsection 5.1). Besides, one can easily verify that $\phi \equiv -\frac{C}{k^2}$ is a solution to system (4.7) and, thus, the unique solution of the problem. As a consequence, one has $\boldsymbol{w} = \boldsymbol{0}$ in Ω^L. ∎

5. Convergence results of PMLs for Galbrun's equation. Our objective is to show that $\boldsymbol{u}^L$, the solution to (4.5), tends to $\boldsymbol{u} = \boldsymbol{\nabla}\varphi_a + \mathbf{curl}\,\varphi_h$ in Ω_b when the ratio $L/|\alpha^*|$ tends to infinity. A natural idea is to introduce two approximate potentials φ_a^L and φ_h^L, which converge, respectively, to φ_a and φ_h and are such that $\boldsymbol{u}^L = \boldsymbol{\nabla}\varphi_a^L + \mathbf{curl}\,\varphi_h^L$.

We now briefly present a sketch of the convergence proof and point out several difficulties that need to be addressed in the analysis as well. Indeed, it appears there is not a unique Helmholtz decomposition for $\boldsymbol{u}^L$, which leaves us with the delicate task of choosing the adequate approximate potentials. Second, the "natural candidates" for these potentials satisfy scalar problems with boundary conditions that are, as we shall see, a priori unusual for PML problems.

Let us characterize the potentials φ_h^L and φ_a^L as the respective solutions to the following scalar problems: *find* $\varphi_h^L \in H^1(\Omega^L)$ *such that*

$$\mathrm{D}^2_{\alpha,\lambda}\varphi_h^L - \Delta_{\alpha,\lambda}\varphi_h^L = g_h + \psi_{\alpha,\lambda} \quad \textit{in } \Omega^L, \tag{5.1a}$$

$$\varphi_h^L = 0 \quad \textit{on } \partial\Omega^L \cap \partial\Omega, \tag{5.1b}$$

$$-\Delta_{\alpha,\lambda}\varphi_h^L = \psi_{\alpha,\lambda} \quad \textit{on } \Sigma_\pm^L, \tag{5.1c}$$

and *find* $\varphi_a^L \in H^1(\Omega^L)$ *such that*

$$\mathrm{D}^2_{\alpha,\lambda}\varphi_a^L - \Delta_{\alpha,\lambda}\varphi_a^L = g_a \quad \textit{in } \Omega^L, \tag{5.2a}$$

$$\boldsymbol{\nabla}_{\alpha,\lambda}\varphi_a^L\cdot\boldsymbol{n} = 0 \quad \textit{on } \partial\Omega^L \cap \partial\Omega, \tag{5.2b}$$

$$\boldsymbol{\nabla}_{\alpha,\lambda}\varphi_a^L\cdot\boldsymbol{n} = -\mathbf{curl}_{\,\alpha,\lambda}\varphi_h^L\cdot\boldsymbol{n} \quad \textit{on } \Sigma_\pm^L. \tag{5.2c}$$

The first step of the proof consists of showing that these two problems are well-posed. The field $\boldsymbol{\nabla}\varphi_a^L + \mathbf{curl}\,\varphi_h^L$ is then clearly a solution to problem (4.4), which establishes that $\boldsymbol{u}^L = \boldsymbol{\nabla}\varphi_a^L + \mathbf{curl}\,\varphi_h^L$. It remains to prove the convergence of the potentials φ_h^L and φ_a^L to their counterparts φ_h and φ_a, using the convergence analysis previously done in [2]. However, problems (5.1) and (5.2) do not exactly enter the framework considered in [2] for three main reasons. First, the boundary condition (5.1c) is nonstandard (from a functional point of view) and nonhomogeneous. Second, the right-hand side term in (5.1a) has a part of its support contained in the PMLs (we nevertheless observe that both this datum and the one in (5.1c) are exponentially decreasing in the layers). Third, the boundary condition (5.2c) is not homogeneous. Note that this last condition links the potential φ_a^L to the tangential trace of $\mathbf{curl}\,\varphi_h^L$

on the boundaries $\Sigma_\pm^L$. Since φ_a satisfies an analogous problem with homogeneous boundary conditions, we expect this trace to tend to zero in order to be able to prove the convergence of the method. As a consequence, we will first investigate the convergence of φ_h^L and then that of φ_a^L.

Before doing so, we first recall and give convergence results for scalar problems of the type of (5.1) and (5.2), but with compactly supported data and homogeneous boundary conditions.

5.1. Convergence results of PMLs for scalar problems. Some straightforward extensions of the results of [2] will be needed in what follows. We do not detail their proofs here, as they are only slight modifications from the ones in the cited reference.

We study the following model problem. Suppose $g \in L^2(\Omega_b)$ is a source with compact support and consider the scalar field φ in $H^1_{\rm loc}(\Omega)$, which satisfies

$$\begin{aligned} &\mathrm{D}^2\varphi - \Delta\varphi = g \quad \text{in } \Omega, \\ &\varphi = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.3}$$

and an additional radiation condition at infinity, which selects the outgoing solution of the problem. As already seen in subsection 3.3.5, this (nonlocal) condition may be expressed through the DtN operators $T_\pm^D$ on $\Sigma_\pm$.

In the following subsections, we give three PML formulations of this model problem, each one of them using a different boundary condition at the end of the layers. One should note that the results obtained here are still valid if a homogeneous Neumann boundary condition is applied on $\partial\Omega$ and, for the sake of brevity, we do not duplicate the statements. Indeed, such a change induces only a modification of the modal basis that appears in the proofs—the sine functions $(\mathrm{S}_n)_{n\in\mathbb{N}^*}$ introduced in (3.20) being replaced by the cosine functions $(\mathrm{C}_n)_{n\in\mathbb{N}}$ defined in (3.10).

5.1.1. Problem A. We first consider a PML formulation of problem (5.3) with a homogeneous Dirichlet boundary condition on $\Sigma_\pm^L$: *find* $\varphi^L \in H^1(\Omega^L)$ *such that*

$$\mathrm{D}^2_{\alpha,\lambda}\varphi^L - \Delta_{\alpha,\lambda}\varphi^L = g \quad \textit{in } \Omega^L, \tag{5.4a}$$

$$\varphi^L = 0 \quad \textit{on } \partial\Omega^L. \tag{5.4b}$$

Equation (5.4a) has to be understood in the distributional sense, so that it implies the following transmission conditions at the interfaces between Ω_b and $\Omega_\pm^L$:

$$\left[\varphi^L\right]_{\Sigma_\pm} = 0 \quad \text{and} \quad \left[\alpha\,\partial_{x_1}\varphi^L + \mathrm{i}\lambda\,\varphi^L\right]_{\Sigma_\pm} = 0. \tag{5.5}$$

In the domain Ω_b, the function φ^L, solution to (5.4), is meant to be an approximation of φ, solution to (5.3).

Adapting the proofs in [2], one can easily show the following lemma.

LEMMA 5.1. *Assume that problem* (5.4) *has a solution. Then, this solution can be written as*

$$\varphi^L(x_1,x_2) = \sum_{n=1}^{+\infty} \left(\varphi^L(x_\pm,.),\mathrm{S}_n\right)_{L^2(\Sigma_\pm)} \Big(A_n^+(\pm L)\, e^{\mathrm{i}\gamma_n^+(x_1-x_\pm)} + A_n^-(\pm L)\, e^{\mathrm{i}\gamma_n^-(x_1-x_\pm)} \Big)\, \mathrm{S}_n(x_2)$$

in the layers $\Omega_\pm^L$, *where* $\gamma_n^\pm = \frac{\beta_n^\pm - \lambda^*}{\alpha^*}$ *and* $A_n^\pm(L) = \mp \frac{e^{\mathrm{i}\beta_n^\mp L/\alpha^*}}{e^{\mathrm{i}\beta_n^+ L/\alpha^*} - e^{\mathrm{i}\beta_n^- L/\alpha^*}}$.

We easily check that the denominators of $(A_n^{\pm}(\pm L))_{n\in\mathbb{N}}$ do not vanish due to assumptions (3.15) and (4.2). We are then able to write exact boundary conditions satisfied by ${\varphi^L}_{|_{\Omega_\pm^L}} = \varphi_\pm^L$ on $\Sigma_\pm$, that is,

$$\partial_{x_1}\varphi_{\pm|_{\Sigma_\pm}}^L = \mathrm{i}\sum_{n=1}^{+\infty}\left(\varphi_\pm^L(x_\pm,.),\mathrm{S}_n\right)_{L^2(\Sigma_\pm)}\left(A_n^+(\pm L)\gamma_n^+ + A_n^-(\pm L)\gamma_n^-\right)\mathrm{S}_n,$$

which in turn yield, using the transmission conditions (5.5),

$$\partial_{x_1}{\varphi_b^L}_{|_{\Sigma_\pm}} = \mathrm{i}\sum_{n=1}^{+\infty}\left(\varphi_b^L(x_\pm,.),\mathrm{S}_n\right)_{L^2(\Sigma_\pm)}\nu_n(\pm L)\,\mathrm{S}_n,$$

where we have set $\nu_n(L) = A_n^+(L)\beta_n^+ + A_n^-(L)\beta_n^- = \beta_n^+ + \frac{\beta_n^- - \beta_n^+}{1-e^{\mathrm{i}(\beta_n^- - \beta_n^+)L/\alpha^*}}$ $\forall n\in\mathbb{N}$. Observe here that these quantities do not depend on the value of λ^*.

Having in mind the comparison between φ^L and φ in Ω_b, we reformulate (5.4) as an equivalent problem posed solely in this domain: *find* $\varphi^L\in H^1(\Omega_b)$ *such that*

$$\begin{aligned}
&\mathrm{D}^2\varphi^L - \Delta\varphi^L = g \quad \text{in } \Omega_b,\\
&\varphi^L = 0 \quad \text{on } \partial\Omega_b\cap\partial\Omega,\\
&\partial_{\boldsymbol{n}}\varphi^L = -T_\pm^L\varphi^L \quad \text{on } \Sigma_\pm,
\end{aligned} \tag{5.6}$$

where the DtN operators $T_\pm^L : H^{1/2}(\Sigma_\pm)\to H^{-1/2}(\Sigma_\pm)$ are defined by $T_\pm^L\phi = \mp\mathrm{i}\sum_{n=1}^{+\infty}\nu_n(\pm L)\ (\phi,\mathrm{S}_n)_{L^2(\Sigma_\pm)}\ \mathrm{S}_n(x_2)$.

On the other hand, problem (5.3) has the following equivalent formulation: *find* $\varphi\in H^1(\Omega_b)$ *such that*

$$\begin{aligned}
&\mathrm{D}^2\varphi - \Delta\varphi = g \quad \text{in } \Omega_b,\\
&\varphi = 0 \quad \text{on } \partial\Omega_b\cap\partial\Omega,\\
&\partial_{\boldsymbol{n}}\varphi = -T_\pm^D\varphi \quad \text{on } \Sigma_\pm.
\end{aligned}$$

Since the scalars $\nu_n(\pm L)$ tend to $\beta_n^\pm$ as $L/|\alpha^*|$ tends to infinity for any integer n, we have the following convergence result (see [2] for a proof).

THEOREM 5.2. *Suppose that assumptions* (3.15) *and* (4.2) *are verified and let* g *in* $L^2(\Omega^L)$ *be compactly supported in* Ω_b. *Then, there exists a strictly positive constant* δ, *depending on* k, M, *and* $\theta = \arg(\alpha^*)$, *such that if* $L/|\alpha^*|\geq\delta$, *problem* (5.4) *is well-posed. Furthermore, the restriction to* Ω_b *of the solution* φ^L *to problem* (5.4) *converges to the restriction of the solution* φ *to problem* (5.3) *as* $L/|\alpha^*|$ *tends to infinity. There also exists a constant* C, *depending on* M *and* k, *such that*

$$\|\varphi-\varphi^L\|_{H^2(\Omega_b)} \leq C\,e^{-\eta\frac{L}{|\alpha^*|}}\|\varphi\|_{H^2(\Omega_b)},$$

with

$$\eta = \frac{2k}{1-M^2}\min\left(-\sin(\theta)\sqrt{1-\frac{{N_0}^2}{{K_0}^2}},\cos(\theta)\sqrt{\frac{(N_0+1)^2}{{K_0}^2}-1}\right), \tag{5.7}$$

where $K_0 = \frac{kl}{\pi\sqrt{1-M^2}}$, N_0 *is the floor of* K_0, *and* $\theta = \arg(\alpha^*)$.

5.1.2. Problem B. We now consider the homogeneous Neumann boundary condition

$$\boldsymbol{\nabla}_{\alpha,\lambda}\varphi^L \cdot \boldsymbol{n} = 0 \quad \text{on } \Sigma_{\pm}^L, \tag{5.8}$$

instead of the previous Dirichlet boundary condition. In this case, the claim of Theorem 5.2 remains true if we furthermore assume that $k \neq \frac{n\pi}{l}$ $\forall n \in \mathbb{N}$. Indeed, the sketch of the proof for the problem at hand is nearly identical to the preceding one, with the following values: $A_n^{\pm}(L) = \mp \frac{\beta_n^{\mp} e^{\mathrm{i}\beta_n^{\mp} L/\alpha^*}}{\beta_n^+ e^{\mathrm{i}\beta_n^+ L/\alpha^*} - \beta_n^- e^{\mathrm{i}\beta_n^- L/\alpha^*}}$ and $\nu_n(L) = \beta_n^+ + \frac{\beta_n^+(\beta_n^- - \beta_n^+)}{\beta_n^+ - \beta_n^- e^{\mathrm{i}(\beta_n^- - \beta_n^+)L/\alpha^*}}$.

Again, one can verify that the scalars $(A_n^{\pm}(L))_{n\in\mathbb{N}}$ are always defined if the assumptions (3.15) and (4.2) are satisfied. Obviously, the scalar $\nu_n(L)$ tends to β_n^+ as $L/|\alpha^*|$ tends to $+\infty$ for any integer n. On the other hand, if there exists an integer j such that $k = \frac{j\pi}{l}$, then the corresponding axial wave number β_j^+ vanishes and the scalar $\nu_j(-L) = \beta_j^- + \frac{\beta_j^-(\beta_j^+ - \beta_j^-)}{\beta_j^- - \beta_j^+ e^{\mathrm{i}(\beta_j^- - \beta_j^+)L/\alpha^*}} = 0$ cannot converge to $\beta_j^- \neq 0$ when $L/|\alpha^*|$ tends to infinity; hence we have the supplementary assumption.

Additionally, the well-posedness of problem (5.4a)–(5.8) in the more general case of a source g with noncompact support and/or if the boundary condition (5.8) is not homogeneous can be proved by means of the Fredholm alternative.

5.1.3. Problem C. We finally consider the following homogeneous condition:

$$\Delta_{\alpha,\lambda}\varphi^L = 0 \quad \text{on } \Sigma_{\pm}^L. \tag{5.9}$$

Theorem 5.2 is still valid. This time, in the proof, we have

$$A_n^{\pm}(L) = \mp \frac{(M\beta_n^{\mp} - k)^2\, e^{\mathrm{i}\beta_n^{\mp} L/\alpha^*}}{(M\beta_n^+ - k)^2\, e^{\mathrm{i}\beta_n^+ L/\alpha^*} - (M\beta_n^- - k)^2\, e^{\mathrm{i}\beta_n^- L/\alpha^*}} \tag{5.10}$$

and $\nu_n(L) = \beta_n^+ + \frac{(M\beta_n^+ - k)^2(\beta_n^- - \beta_n^+)}{(M\beta_n^+ - k)^2 - (M\beta_n^- - k)^2\, e^{\mathrm{i}(\beta_n^- - \beta_n^+)L/\alpha^*}}$.

Contrary to both previous problems, the well-posedness of problem (5.4a)–(5.9) cannot be extended to the case of an arbitrary source term g, since the boundary condition (5.9) does not allow us to write a variational formulation of the problem.

5.2. Well-posedness and convergence analysis for $\boldsymbol{\varphi_h^L}$. Once again, the idea is to deal with an equivalent problem, whose source term is compactly supported and whose boundary conditions are homogeneous, and the approach previously used in subsection 3.3.5 is followed closely. This new problem then permits the construction of the solution via a modal decomposition.

We first introduce the functions $\psi_{\alpha,\lambda}^{\infty}$ such that, $\forall (x_1, x_2) \in \Omega_+^L$,

$$\psi_{\alpha,\lambda}^{\infty}(x_1,x_2) = \begin{cases} (a(x_2) + b(x_2)x_1)\; e^{\mathrm{i}\frac{k}{M}x_1} \;\; \forall (x_1,x_2) \in]d_+, x_+[\times]0,l[, \\ \left(a(x_2) + b(x_2)\left(x_+ + \dfrac{x_1 - x_+}{\alpha^*}\right)\right)\; e^{\mathrm{i}\left(\frac{k}{M}x_+ + \left(\frac{k}{M} - \lambda^*\right)\frac{x_1 - x_+}{\alpha^*}\right)}, \end{cases}$$

and $\zeta_{\alpha,\lambda}$ such that, $\forall (x_1, x_2) \in \Omega_+^L$,

$$\zeta_{\alpha,\lambda}(x_1,x_2) = \begin{cases} (A(x_2) + B(x_2)x_1)\; e^{\mathrm{i}\frac{k}{M}x_1} \;\; \forall (x_1,x_2) \in]d_+, x_+[\times]0,l[, \\ \left(A(x_2) + B(x_2)\left(x_+ + \dfrac{x_1 - x_+}{\alpha^*}\right)\right)\; e^{\mathrm{i}\left(\frac{k}{M}x_+ + \left(\frac{k}{M} - \lambda^*\right)\frac{x_1 - x_+}{\alpha^*}\right)}, \end{cases}$$

the functions a, b, A, and B having been characterized in the subsection 3.3.5. We then set $\varphi_h^L = \widetilde{\varphi}_h^L + \chi\zeta_{\alpha,\lambda}$, where the function $\widetilde{\varphi}_h^L$ satisfies

$$\begin{array}{ll} \mathrm{D}^2_{\alpha,\lambda}\widetilde{\varphi}_h^L - \Delta_{\alpha,\lambda}\widetilde{\varphi}_h^L = \widetilde{g}_h & \text{in } \Omega^L, \\ \widetilde{\varphi}_h^L = 0 \quad \text{on } \partial\Omega^L \cap \partial\Omega, & \\ -\Delta_{\alpha,\lambda}\widetilde{\varphi}_h^L = 0 \quad \text{on } \Sigma_\pm^L. & \end{array}$$

Indeed, the quantity $g_h + \psi_{\alpha,\lambda} - \mathrm{D}^2_{\alpha,\lambda}(\chi\zeta_{\alpha,\lambda}) - \Delta_{\alpha,\lambda}(\chi\zeta_{\alpha,\lambda})$ coincides with $\widetilde{g}_h$, since $\psi_{\alpha,\lambda}$ and $\zeta_{\alpha,\lambda}$, respectively, coincide with ψ and ζ in Ω_b. This last problem is well-posed, owing to the results of subsection 5.1, and we easily deduce the following result.

THEOREM 5.3. *If the ratio $L/|\alpha^*|$ is large enough, problem* (5.1) *has a unique solution which is $\varphi_h^L = \widetilde{\varphi}_h^L + \chi\zeta_{\alpha,\lambda}$. Moreover, the function φ_h^L converges to $\widetilde{\varphi}_h + \chi\zeta = \varphi_h$ in $H^2(\Omega_b)$ as $L/|\alpha^*|$ tends to infinity, and one has the estimate*

$$\left\|\varphi_h - \varphi_h^L\right\|_{H^2(\Omega_b)} \leq C\, e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}} \left\|\varphi_h\right\|_{H^2(\Omega_b)},$$

where the constant C depends on k and M and η is defined in (5.7).

Proof. In the domain Ω_b, one has $\varphi_h^L - \varphi_h = \widetilde{\varphi}_h^L + \chi\zeta_{\alpha,\lambda} - \widetilde{\varphi}_h - \chi\zeta = \widetilde{\varphi}_h^L - \widetilde{\varphi}_h$. The convergence result then directly follows from subsection 5.1. ∎

COROLLARY 5.4. *Suppose that assumptions* (4.2) *and* (4.3) *hold. Then, the traces $(\mathbf{curl}\,_{\alpha,\lambda}\varphi_h^L \cdot \boldsymbol{n})_{|_{\Sigma_\pm^L}}$ tend to zero in $H^{1/2}(\Sigma_\pm^L)$ as $L/|\alpha^*|$ tends to infinity. More precisely, for $L/|\alpha^*|$ large enough, one has the estimate*

$$\left\|\mathbf{curl}\,_{\alpha,\lambda}\varphi_h^L \cdot \boldsymbol{n}\right\|_{H^{1/2}(\Sigma_\pm^L)} \leq C\left(e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}}\left\|\varphi_h\right\|_{H^2(\Omega_b)} + e^{\left(\frac{k}{M}-\lambda^*\right)\sin(\theta)\frac{L}{|\alpha^*|}}\left\|g_h\right\|_{L^2(\Omega_b)}\right),$$

where the constant C depends on k and M, η is defined in (5.7), *and θ denotes the argument of α^*.*

Proof. Using a modal decomposition for $\widetilde{\varphi}_h^L$ in the layers $\Omega_\pm^L$, similar to the ones used in the subsection 5.1, one has, on Σ_+^L, for instance, $\forall x_2 \in [0,l]$,

$$\partial_{x_2}\widetilde{\varphi}_h^L(x_+ + L, x_2) = \sum_{n=1}^{+\infty}\left(\widetilde{\varphi}_h^L(x_+,\cdot),\mathrm{S}_n\right)_{L^2([0,l])}\left(A_n^+(L)e^{\mathrm{i}\gamma_n^+ L} + A_n^-(L)e^{\mathrm{i}\gamma_n^- L}\right)\frac{n\pi}{l}\mathrm{C}_n(x_2),$$

the scalars $(A_n^\pm(L))_{n\in\mathbb{N}}$ being defined in (5.10). Setting

$$\tau_n = A_n^+(L)\,e^{\mathrm{i}\gamma_n^+ L} + A_n^-(L)\,e^{\mathrm{i}\gamma_n^- L} = \frac{M(\beta_n^- - \beta_n^+)\,(2k + M(\beta_n^+ + \beta_n^-))\;e^{\mathrm{i}\gamma_n^+ L}}{(M\beta_n^+ - k)^2\,e^{\mathrm{i}(\beta_n^+ - \beta_n^-)\frac{L}{\alpha^*}} - (M\beta_n^- - k)^2} \quad \forall n \in \mathbb{N},$$

we obtain, if $L/|\alpha^*|$ is large enough and after some majorizations,

$$|\tau_n| \leq C\, e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}} \quad \forall n \in \mathbb{N},$$

the constant η being defined in (5.7). Thus, we have

$$\begin{aligned} \left\|\partial_{x_2}\widetilde{\varphi}_h^L\right\|^2_{H^{1/2}(\Sigma_+^L)} &\leq \sum_{n=1}^{+\infty}\left(1 + \frac{n^2\pi^2}{l^2}\right)^{3/2}|\tau_n|^2\left|\left(\widetilde{\varphi}_h^L(x_+,\cdot),\mathrm{S}_n\right)_{L^2([0,l])}\right|^2 \\ &\leq C\, e^{-\eta\frac{L}{|\alpha^*|}}\left\|\widetilde{\varphi}_h^L\right\|^2_{H^{3/2}(\Sigma_+)}. \end{aligned}$$

We then deduce an estimate of this quantity using the convergence of $\widetilde{\varphi}_h^L$ to $\widetilde{\varphi}_h$ in $H^2(\Omega_b)$ and a trace theorem. On the other hand, one has

$$\|\partial_{x_2}\zeta_{\alpha,\lambda}\|^2_{H^{1/2}(\Sigma_+^L)} \leq \left(\|A\|^2_{H^2([0,l])} + \|B\|^2_{H^2([0,l])}\left|x_+ + \frac{L}{\alpha^*}\right|^2\right) e^{2\left(\frac{k}{M}-\lambda^*\right)\sin(\theta)\frac{L}{|\alpha^*|}}$$
$$\leq C\left(1+\frac{L}{|\alpha^*|}\right)^2 e^{2\left(\frac{k}{M}-\lambda^*\right)\sin(\theta)\frac{L}{|\alpha^*|}}\,\|g_h\|^2_{L^2(\Omega)}.$$

One obviously has similar estimates on Σ_-^L, and thus the announced result is obtained. ∎

We emphasize here that this corollary is valid only for the PML model corresponding to the transformation (4.1) with the choice (4.3) for λ^*. Indeed, every propagative mode, and, more particularly, every inverse upstream mode, has to be damped in the layers in order to prove the claim. This appears to be different from the convergence results previously given in [2] and subsequently extended in subsection 5.1, which are also true for the original Bérenger model (that is, when $\lambda \equiv 0$).

5.3. Well-posedness and convergence analysis for $\boldsymbol{\varphi_a^L}$. Since the analysis in section 5.1 was concerned with PML problems with homogeneous boundary conditions on $\Sigma_\pm^L$, the only difficulty in proving the convergence of problem (5.2) comes from the nonhomogeneous boundary condition (5.2c). Consequently, the scalar field φ_a^L is split in the following manner: $\varphi_a^L = \varphi_a^{L,1} + \varphi_a^{L,2}$, with

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\varphi_a^{L,1} - \Delta_{\alpha,\lambda}\varphi_a^{L,1} = g_a \quad \text{in } \Omega^L,\\ &\boldsymbol{\nabla}_{\alpha,\lambda}\varphi_a^{L,1}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L, \end{aligned} \tag{5.11}$$

and

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\varphi_a^{L,2} - \Delta_{\alpha,\lambda}\varphi_a^{L,2} = 0 \quad \text{in } \Omega^L,\\ &\boldsymbol{\nabla}_{\alpha,\lambda}\varphi_a^{L,2}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L\cap\partial\Omega,\\ &\boldsymbol{\nabla}_{\alpha,\lambda}\varphi_a^{L,2}\cdot\boldsymbol{n} = -\mathbf{curl}_{\,\alpha,\lambda}\varphi_h^L\cdot\boldsymbol{n} \quad \text{on } \Sigma_\pm^L. \end{aligned} \tag{5.12}$$

As seen in subsection 5.1, both of these problems are well-posed. Moreover, the results of subsection 5.1.2 readily give the convergence of the solution $\varphi_a^{L,1}$ to problem (5.11) to φ_a in $H^2(\Omega_b)$ as the ratio $L/|\alpha^*|$ tends to infinity.

We next prove the following lemma.

LEMMA 5.5. *If the ratio $L/|\alpha^*|$ is large enough, the solution $\varphi_a^{L,2}$ to problem* (5.12) *satisfies the estimate*

$$\left\|\varphi_a^{L,2}\right\|_{H^2(\Omega_b)} \leq C\,e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}}\left\|\mathbf{curl}_{\,\alpha,\lambda}\varphi_h^L\cdot\boldsymbol{n}\right\|_{H^{1/2}\left(\Sigma_\pm^L\right)},$$

where the constant C depends on k and M and the constant η is defined in (5.7).

Proof. Let us set $q^\pm = -(\mathbf{curl}_{\,\alpha,\lambda}\varphi_h^L\cdot\boldsymbol{n})_{|_{\Sigma_\pm^L}}$. The main tool of the proof is again a modal decomposition. Since the functions $q^\pm$, respectively, belong to $H^{1/2}(\Sigma_\pm^L)$, they can be written as $q^\pm(x_2) = \sum_{n=0}^{+\infty} q_n^\pm\,\mathrm{C}_n(x_2)$ on $\Sigma_\pm^L$, and we have $\|q^\pm\|^2_{H^{1/2}(\Sigma_\pm^L)} = \sum_{n=0}^{+\infty}(1+\frac{n^2\pi^2}{l^2})^{1/2}\,|q_n^\pm|^2$.

We look for a solution to problem (5.12) of the same form,

$$\varphi_a^{L,2}(x_1,x_2) = \sum_{n=0}^{+\infty}\phi_n(x_1)\,\mathrm{C}_n(x_2) \quad \text{in } \Omega^L, \tag{5.13}$$

which yields the following ODEs: $\mathrm{D}^2_{\alpha,\lambda}\phi_n - (\alpha\,\mathrm{d}_{x_1} + \mathrm{i}\lambda)^2\,\phi_n + \frac{n^2\pi^2}{l^2}\phi_n = 0$, $n \in \mathbb{N}$, with boundary conditions $\alpha\,\phi_n{}' + \mathrm{i}\lambda\,\phi_n = \pm q_n^{\pm}$ on $\Sigma_{\pm}^L$ and transmission conditions, between Ω_b and the PMLs, $[\phi_n]_{\Sigma_\pm} = 0$ and $\left[\alpha\,\phi_n{}' + \mathrm{i}\lambda\,\phi_n\right]_{\Sigma_\pm} = 0$.

There are three different zones:

$$\phi_n(x_1) = \begin{cases} A_n^-\,e^{\mathrm{i}\gamma_n^- x_1} + A_n^+\,e^{\mathrm{i}\gamma_n^+ x_1} & \text{if } x_1 \le x_-, \\ B_n^-\,e^{\mathrm{i}\beta_n^-(x_1-x_+)} + B_n^+\,e^{\mathrm{i}\beta_n^+(x_1-x_-)} & \text{if } x_- \le x_1 \le x_+, \\ C_n^-\,e^{\mathrm{i}\gamma_n^- x_1} + C_n^+\,e^{\mathrm{i}\gamma_n^+ x_1} & \text{if } x_1 \ge x_+. \end{cases}$$

Expressing the boundary and transmission conditions gives a 6-by-6 linear system to solve in order to obtain the coefficients $(A_n^{\pm}, B_n^{\pm}, C_n^{\pm})$. After some manipulations, we finally obtain $B_n^- = -\mathrm{i}\frac{q_n^-}{\beta_n^-}\,e^{-\mathrm{i}\gamma_n^- L}\frac{1+z_n^+}{1-z_n^+ z_n^-}$ and $B_n^+ = \mathrm{i}\frac{q_n^+}{\beta_n^+}\,e^{\mathrm{i}\gamma_n^+ L}\frac{1+z_n^-}{1-z_n^+ z_n^-}$, where $z_n^+ = e^{\mathrm{i}\beta_n^+(x_+-x_-)}e^{2\mathrm{i}\gamma_n^+ L}$ and $z_n^- = e^{-\mathrm{i}\beta_n^-(x_+-x_-)}e^{-2\mathrm{i}\gamma_n^- L}$.

Note that, due to assumption (4.6), the scalars $B_n^{\pm}$ are well defined. Using classical properties of propagative and evanescent modes, it is easy to show that, for $L/|\alpha^*|$ sufficiently large, we have $|B_n^{\pm}| \le C\left|\frac{q_n^{\pm}}{\beta_n^{\pm}}\right| e^{\mp \mathrm{Im}(\gamma_n^{\pm})L}$, which yield $|B_n^{\pm}| \le C\left|\frac{q_n^{\pm}}{\beta_n^{\pm}}\right| e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}}$, the constant η being the one defined in (5.7).

Then, successively integrating (5.13) with respect to x_2 and x_1 and using the estimates above, we conclude that

$$\left\|\varphi_a^{L,2}\right\|_{H^2(\Omega_b)} \le C\,e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}}\left\|q^{\pm}\right\|_{H^{1/2}\left(\Sigma_{\pm}^L\right)},$$

which ends the proof. ∎

Using Corollary 5.4, we finally deduce the following corollary.

COROLLARY 5.6. *The function $\varphi_a^{L,2}$ converges to zero as $L/|\alpha^*|$ tends to infinity. More precisely, for $L/|\alpha^*|$ large enough, one has the estimate*

$$\left\|\varphi_a^{L,2}\right\|_{H^2(\Omega_b)} \le C\left(e^{-\eta\frac{L}{|\alpha^*|}}\left\|\varphi_h\right\|_{H^2(\Omega_b)} + e^{\left(\left(\frac{k}{M}-\lambda^*\right)\sin(\theta)-\frac{\eta}{2}\right)\frac{L}{|\alpha^*|}}\left\|g_h\right\|_{L^2(\Omega)}\right),$$

where the constant C depends on k and M, η is defined in (5.7)*, and θ denotes the argument of α^*.*

5.4. Conclusion. The gathering of the preceding results yields the inequality

$$\begin{aligned}\|\boldsymbol{u}^L - \boldsymbol{u}\|_{H^1(\Omega_b)^2} \le C\,e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}}\Bigg(&\|\boldsymbol{\nabla}\varphi_a\|_{H^1(\Omega_b)^2} + \|\mathbf{curl}\,\varphi_h\|_{H^1(\Omega_b)^2} \\ &+ \left(1+\frac{L}{|\alpha^*|}\right)e^{\left(\frac{k}{M}-\lambda^*\right)\sin(\theta)\frac{L}{|\alpha^*|}}\left\|g_h\right\|_{L^2(\Omega)}\Bigg)\end{aligned}$$

and consequently allows us to state a final theorem relative to the convergence of the solution to problem (4.4) when the ratio $L/|\alpha^*|$ tends to infinity.

THEOREM 5.7. *Suppose that assumptions* (4.2) *and* (4.3) *hold. Then, the field $\boldsymbol{u}^L$ tends to $\boldsymbol{u}$ in $H^1(\Omega_b)^2$ as $L/|\alpha^*|$ tends to infinity. Furthermore, for $L/|\alpha^*|$ large enough, one has the estimate*

$$\left\|\boldsymbol{u}^L - \boldsymbol{u}\right\|_{H^1(\Omega_b)^2} \le C\,e^{-\frac{\eta}{2}\frac{L}{|\alpha^*|}},$$

where the constant C depends on k, M, and the solution $\boldsymbol{u}$, the constant η being defined in (5.7).

5.5. Remark on the use of PMLs without regularization. Let us finally tackle the claim made in the introduction of section 4, namely, that PMLs do not work without regularization. Indeed, assume that the approximated displacement field $\boldsymbol{u}^L$ is computed as the solution to the problem

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\boldsymbol{u}^L - \boldsymbol{\nabla}_{\alpha,\lambda}\left(\mathrm{div}_{\alpha,\lambda}\,\boldsymbol{u}^L\right) = \boldsymbol{f} \quad \text{in } \Omega^L, \\ &\boldsymbol{u}^L\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega^L, \end{aligned} \tag{5.14}$$

completed by an additional boundary condition at the end of the layers—for instance, $\mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L = 0$ on $\partial\Omega^L$.

Then, the function $\psi^L = \mathrm{curl}_{\,\alpha,\lambda}\boldsymbol{u}^L$ is a solution to the following problem:

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\psi^L = \mathrm{curl}\,\boldsymbol{f} \quad \text{in } \Omega^L, \\ &\psi^L = 0 \quad \text{on } \Sigma^L_{\pm}. \end{aligned} \tag{5.15}$$

For the PML method to work, the field $\boldsymbol{u}^L$ must converge to $\boldsymbol{u}$ in Ω_b as the ratio $L/|\alpha^*|$ tends to infinity, and, consequently, ψ^L must converge to $\psi = \mathrm{curl}\,\boldsymbol{u}$ in Ω_b. We now show that this convergence does not hold. Indeed, the solution ψ^L to problem (5.15) can be sought as the sum $\psi^L = \psi_{\alpha,\lambda} + \widetilde{\psi}^L$, where $\widetilde{\psi}^L$ solves the following problem:

$$\begin{aligned} &\mathrm{D}^2_{\alpha,\lambda}\widetilde{\psi}^L = 0 \quad \text{in } \Omega^L, \\ &\widetilde{\psi}^L = 0 \quad \text{on } \Sigma^L_-, \\ &\widetilde{\psi}^L = -\psi_{\alpha,\lambda} \quad \text{on } \Sigma^L_+. \end{aligned} \tag{5.16}$$

Using the expression of $\psi_{\alpha,\lambda}$ derived previously, we find that

$$\psi_{\alpha,\lambda}(x_+ + L, x_2) = \left(a(x_2) + b(x_2)\left(x_+ + \frac{L}{\alpha^*}\right)\right) e^{\mathrm{i}\left(\frac{k}{M}x_+ + \left(\frac{k}{M}-\lambda^*\right)\frac{L}{\alpha^*}\right)} \quad \forall x_2 \in]0, l[.$$

Then, the expression of $\widetilde{\psi}^L$ can be easily derived by seeking a solution of the form

$$\widetilde{\psi}^L(x_1, x_2) = \begin{cases} \widetilde{a}(x_2)\left(1 + t\,\dfrac{x_1 - x_-}{\alpha^*}\right) e^{\mathrm{i}\left(\frac{k}{M}\,x_- + \left(\frac{k}{M}-\lambda^*\right)\frac{x_1-x_-}{\alpha^*}\right)} & \text{if } x_1 \le x_-, \\ \widetilde{a}(x_2)\left(1 + t\,(x_1 - x_-)\right) e^{\mathrm{i}\frac{k}{M}x_1} & \text{if } x_- \le x_1 \le x_+, \\ \widetilde{a}(x_2)\left(1 + t\left(x_+ - x_- + \dfrac{x_1 - x_+}{\alpha^*}\right)\right) e^{\mathrm{i}\left(\frac{k}{M}\,x_+ + \left(\frac{k}{M}-\lambda^*\right)\frac{x_1-x_+}{\alpha^*}\right)} & \\ \qquad\qquad \text{if } x_1 \ge x_+, & \end{cases}$$

where the function $\widetilde{a}$ and the real number t need to be determined. By construction, the transmission conditions on $\Sigma_\pm$ are automatically satisfied, and we simply use the boundary conditions on $\Sigma^L_\pm$, which give us the two equalities

$$\begin{aligned} &1 - t\frac{L}{\alpha^*} = 0, \\ &\widetilde{a}(x_2)\left(1 + t\left(x_+ - x_- + \frac{L}{\alpha^*}\right)\right) = -\left(a(x_2) + b(x_2)\left(x_+ + \frac{L}{\alpha^*}\right)\right), \end{aligned}$$

so that, finally, $t = \alpha^*/L$ and $\widetilde{a}(x_2) = -\frac{a(x_2)+b(x_2)\left(x_+ + \frac{L}{\alpha^*}\right)}{2 + \frac{\alpha^*}{L}(x_+ - x_-)}$. Clearly, we see that $\widetilde{\psi}^L$ does not converge to zero as $L/|\alpha^*|$ tends to infinity.

6. Numerical applications. Due to its Fredholm type, variational formulation (4.5) is well suited for a finite element approximation, and, for given values of the PML parameters α^* and L, usual rates of convergence (with respect to the mesh stepsize) are ensured. All computations are done with the finite element library MÉLINA [17]. We use P_2 Lagrange finite elements on a nonstructured mesh, and the length of the PML is equal to 10% of the length of the domain $\Omega_b = [0,2] \times [0,1]$. As in the theoretical analysis, the function α is constant in the layer, and the argument of the complex number α^* is fixed and equal[1] to $-\frac{\pi}{4}$, its modulus $|\alpha^*|$ being a parameter in the simulations.

We are interested here in simulating the radiation of a compactly supported source situated in a two-dimensional rigid duct; this is a problem for which no explicit reference solution is available. Nonetheless, we consider as a preliminary study the propagation of acoustic and vortical modes in order to validate the method.

6.1. Mode propagation in a rigid duct.

6.1.1. Acoustic and vortical modes: Some definitions. The so-called modes are solutions with separated variables to the homogeneous, nonregularized Galbrun equation, with the rigid wall boundary condition (2.2). These solutions are of two distinct kinds, called acoustic and vortical modes.

The acoustic modes are of the form

$$\boldsymbol{u}(x_1,x_2) = \begin{cases} C\, e^{\mathrm{i}\beta_n^{\pm} x_1}\, \boldsymbol{e_1} & \text{if } n = 0, \\ C\, e^{\mathrm{i}\beta_n^{\pm} x_1} \left(-\dfrac{\mathrm{i}\beta_n^{\pm} l}{n\pi} \cos\left(\dfrac{n\pi}{l} x_2\right) \boldsymbol{e_1} + \sin\left(\dfrac{n\pi}{l} x_2\right) \boldsymbol{e_2} \right) & \text{if } n \in \mathbb{N}^*, \end{cases}$$

where C is a complex constant and the axial wave numbers $\beta_n^{\pm}$ are given by (3.12). One can see that these fields are irrotational—hence their name. The acoustic modes associated to real valued axial wave numbers are called propagative and evanescent otherwise. Propagative modes with positive (resp., negative) group velocity $\frac{\partial \omega}{\partial \beta}$ are called downstream (resp., upstream) modes, since their energy propagates downstream (resp., upstream) of the mean flow.

The second "family" of solutions to Galbrun's equation consists of a continuum of fields such that

$$\boldsymbol{u}(x_1,x_2) = C\, e^{\mathrm{i}\frac{k}{M} x_1} \left(\frac{\mathrm{i}k}{M}\, \varphi'(x_2)\, \boldsymbol{e_1} + \varphi(x_2)\, \boldsymbol{e_2} \right),$$

where C is a complex constant and φ denotes a scalar function belonging to $H_0^1([0,l])$. These solutions propagate only downstream and are called vortical modes, since they are divergence-free.

6.1.2. Description of the simulations. The following numerical simulations consist of solving problems similar to (4.4), with only one absorbing layer downstream (denoted by Ω_+^L). Each problem was designed in such a way that one of the previously introduced modes is its exact solution. For the propagation of an acoustic mode, we have $\boldsymbol{f} \equiv \boldsymbol{0}$ and $\psi_{\alpha,\lambda} \equiv 0$, the mode being imposed via a nonhomogeneous condition on the boundary Σ_- for the normal displacement $\boldsymbol{u}\cdot\boldsymbol{n}$. In the case of a vortical mode, we still have $\boldsymbol{f} \equiv \boldsymbol{0}$ and a nonhomogeneous boundary condition on Σ_-, but the field $\psi_{\alpha,\lambda}$ has to be computed a priori as the curl of the considered mode.

[1] While apparently arbitrary, this choice simply makes the quantities $-\sin(\arg(\alpha^*))$ and $\cos(\arg(\alpha^*))$, which appear in the definition (5.7) of the coefficient η, to be equal.

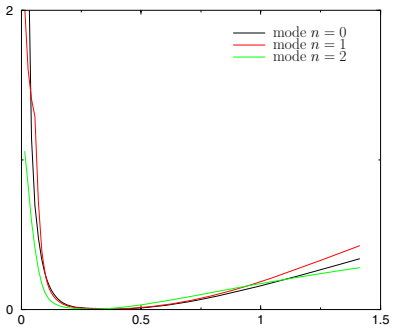


FIG. 1. *Relative error in the* $H^1(\Omega_b)^2$ *norm as a function of* $|\alpha^*|$ *for the computated propagative downstream acoustic modes* ($k = 8$ *and* $M = 0.4$).

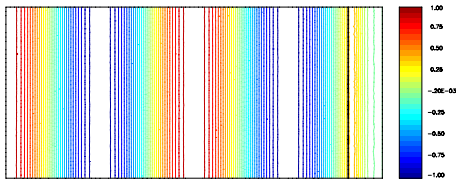


FIG. 2. *Contours of the real part of the component* u_1 *of the computed displacement field for the propagative downstream mode* ($n = 0$, $k = 8$, $M = 0.4$, $\alpha^* = 0.25(1 - \mathrm{i})$).

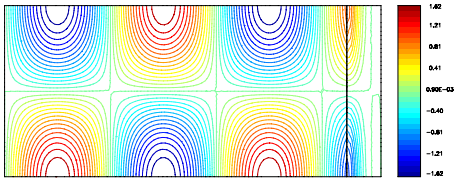


FIG. 3. *Contours of the real part of the component* u_1 *of the computed displacement field for the propagative downstream mode* ($n = 1$, $k = 8$, $M = 0.4$, $\alpha^* = 0.25(1 - \mathrm{i})$).

6.1.3. Numerical results for acoustic modes. In the chosen configuration, characterized by the values $l = 1$, $k = 8$, and $M = 0.4$, six (i.e., three upstream and three downstream) acoustic modes are propagative. The curves plotting the relative error in the $H^1(\Omega_b)$ norm for the computed displacement versus the modulus of α^* for the propagative downstream modes are shown in Figure 1. We observe that each curve contains a minimum plateau where the relative error is below a few percent. For large values of $|\alpha^*|$, the error increases due to the reflection at the end of the layer and behaves as theoretically predicted. For small values of $|\alpha^*|$, the method diverges, the mesh resolution being too coarse to adequately represent the modes in the PML medium, thus producing spurious numerical errors. Similar results were obtained for the propagative upstream modes.

We show in Figures 2 to 4 the contours of the components of the computed displacement for a value of $|\alpha^*|$ such that the error of the method is below one percent.

6.1.4. Numerical results for vortical modes. For the study of the method on vortical perturbations, the transverse dependence of the modes is arbitrarily chosen as $\varphi(x_2) = \sin\left(\frac{m\pi}{l}\, x_2\right)$, where m is a given nonzero integer. In Figure 5 we show

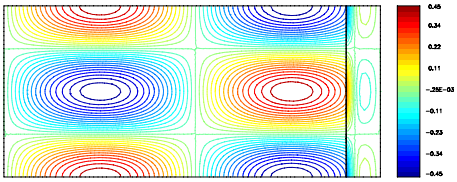


FIG. 4. *Contours of the real part of the component u_1 of the computed displacement field for the propagative downstream mode ($n = 2$, $k = 8$, $M = 0.4$, $\alpha^* = 0.25(1 - \mathrm{i})$).*

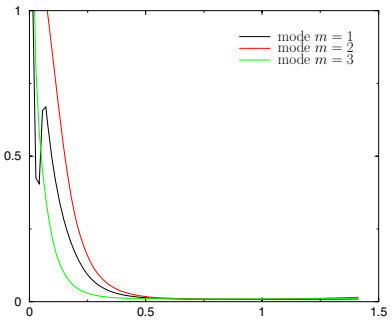


FIG. 5. *Relative error in the $H^1(\Omega_b)^2$ norm as a function of $|\alpha^*|$ for the computed vortical modes ($k = 8$, $M = 0.4$).*

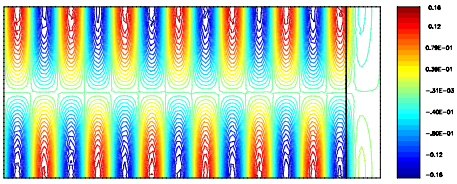


FIG. 6. *Contours of the real part of the component u_1 of the computed displacement field for a vortical mode ($m = 1$, $k = 8$, $M = 0.4$, $\alpha^* = 0.65(1 - \mathrm{i})$).*

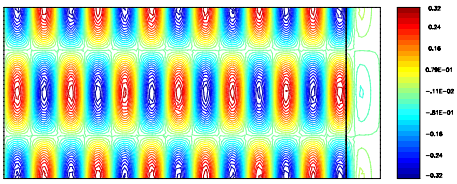


FIG. 7. *Contours of the real part of the component u_1 of the computed displacement field for a vortical mode ($m = 2$, $k = 8$, $M = 0.4$, $\alpha^* = 0.65(1 - \mathrm{i})$).*

the relative error of the method, plotted versus the modulus of the coefficient α^* for values of the integer m equal to 1, 2, and 3. The contours of the components of the corresponding computed solutions when the relative error is below one percent are presented in Figures 6 to 8.

The convergence of the PML method is also obtained in this case for appropriate values of $|\alpha^*|$. However, by comparing Figures 1 and 5, one can already point out a potential difficulty that may be encountered in practice when both irrotational and

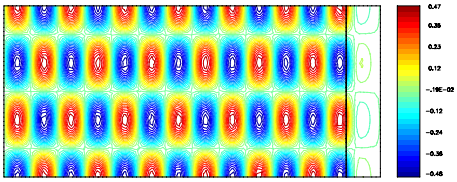


FIG. 8. *Contours of the real part of the component u_1 of the computed displacement field for a vortical mode ($m = 3$, $k = 8$, $M = 0.4$, $\alpha^* = 0.65(1 - \mathrm{i})$).*

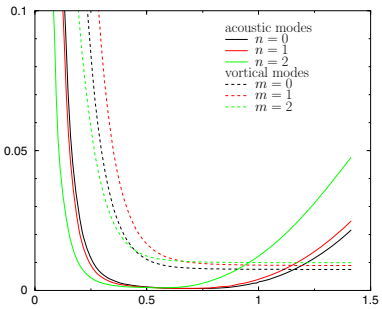


FIG. 9. *Relative error in the $H^1(\Omega_b)^2$ norm as a function of $|\alpha^*|$ for several computed modes ($k = 8$, $M = 0.4$).*

vortical perturbations are present, as the values of $|\alpha^*|$ that allow a good agreement between the exact and the computed solutions in the two cases are quite different. Indeed, the propagation constant of the vortical modes, which is equal to $\frac{k}{M}$, leads to more significant damping in the layer for these modes than for the acoustic ones. On the other hand, the finite element error for vortical modes is higher, since their wavelength is generally much shorter than that of their acoustic counterparts. This fact may cause some discretization issues as $|\alpha^*|$ becomes small.[2]

However, a compromise can be found by using a thicker layer. For instance, we have repeated the previous simulations with a layer of length equal to 25% of the length of the domain Ω_b. The obtained results are presented in Figure 9, and one can now observe a partial match between the respective ranges of values of $|\alpha^*|$ for which the relative errors are below a few percent.

6.2. Radiation of compactly supported sources.

6.2.1. Acoustic source. We next simulate the radiation of an irrotational, compactly supported source which is placed in the duct, defined in polar coordinates by $\boldsymbol{f}(r, \theta) = |r - r_C|\, \boldsymbol{e_r}$ in the ball of center $C = (1, 0.5)$ and with radius equal to 0.15, and vanishing elsewhere. This case happens to be more complex than the previous one because of the absence of any reference solution, which would permit the measuring of the precision of the method and the choosing of an adequate value for the parameter $|\alpha^*|$. The real parts of the components of the computed displacement field are shown in Figure 10. The acoustic mode mainly radiated by the source is the one of index $n = 2$. The convective effect of the uniform flow can clearly be seen, as the wavelength of the computed solution is shorter upstream of the source than downstream.

[2]Note that these issues, related to the mesh stepsize, appear as well without PMLs, when the speed of the flow is slow (that is, when the Mach number M is close to zero) or at high frequencies (that is, when the wave number k is large).

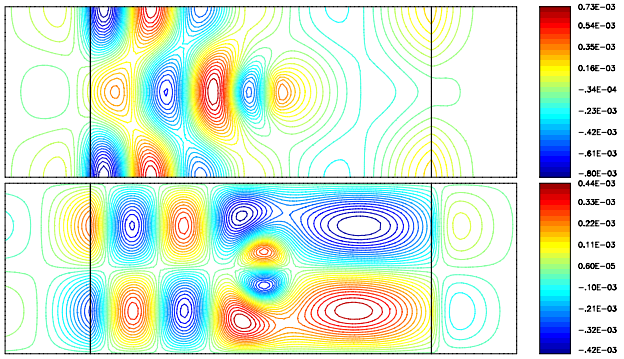


FIG. 10. *Contours of the real part of components u_1 (top) and u_2 (bottom) of the computed displacement field for the radiation of an acoustic source ($k = 8$, $M = 0.4$, $\alpha^* = 0.5(1 - \mathrm{i})$).*

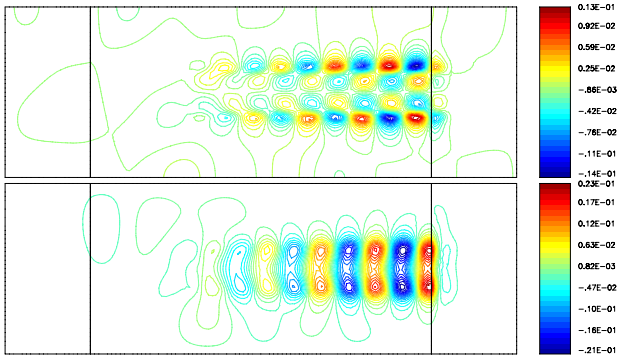


FIG. 11. *Contours of the real part of components u_1 (top) and u_2 (bottom) of the computed displacement field for the radiation of a vortical source ($k = 8$, $M = 0.4$, $\alpha^* = 0.5(1 - \mathrm{i})$).*

6.2.2. Rotational source. The simulation of the radiation of a compactly supported source whose curl is nonzero, which is defined in polar coordinates by $\boldsymbol{f}(r, \theta) = |r - r_C| \arctan(\frac{r \sin\theta - r_C \sin\theta_C}{r \cos\theta - r_C \cos\theta_C})\, \boldsymbol{e_r}$ in the ball of center $C = (0.7, 0.5)$ and with radius equal to 0.15, and vanishing elsewhere, is presented in Figure 11. It is located slightly upstream of the center of the domain Ω_b. One can observe the hydrodynamic wake generated by the source and convected by the flow, whose amplitude increases linearly with respect to the coordinate x_1. The acoustic perturbations, where present, have a negligible amplitude compared to the vortical ones.

Appendix. Study of a second order transport equation. This appendix is devoted to the second order transport equation (3.3), which is written here as

$$\mathrm{D}_\varepsilon^2 \psi^\varepsilon = g, \tag{6.1}$$

with g a compactly supported source belonging to $L^2(\Omega)$. We know from Lemma 3.1 that (6.1) has a unique solution in $L^2(\Omega)$. We also have the following result.

LEMMA 6.1. *The solution in $L^2(\Omega)$ to* (6.1) *vanishes upstream of the support of the source g and satisfies the estimate*

$$\|\psi^\varepsilon\|_{L^2(\Omega)} \leq C_\varepsilon \|g\|_{L^2(\Omega)}, \tag{6.2}$$

where C_ε denotes a positive constant depending on the parameter ε.

Proof. From the proof of Lemma 3.1, we know that

$$\psi^\varepsilon(x_1,x_2) = G_\varepsilon * g(\cdot,x_2)(x_1) = \frac{1}{M^2}\int_{-\infty}^{x_1} (x_1-z)\, e^{\mathrm{i}\frac{k_\varepsilon}{M}(x_1-z)} g(z,x_2)\,\mathrm{d}z,$$

where the kernel G_ε denotes the causal Green's function of the differential operator D^2_ε. Setting the bounds $d_- = \min_{x_2\in[0,h]}\{x_1\in\mathbb{R} \mid (x_1,x_2)\in \operatorname{supp} g\}$ and $d_+ = \max_{x_2\in[0,h]}\{x_1\in\mathbb{R} \mid (x_1,x_2)\in \operatorname{supp} g\}$, we consider the following cases.

If $x_1 < d_-$, one has $]-\infty,x_1]\cap[d_-,d_+] = \varnothing$; thus $\psi^\varepsilon(x_1,x_2)\equiv 0$ for any $x_2\in[0,l]$. The solution then vanishes upstream of the support of the source.

If $d_- \le x_1 \le d_+$, one has $\psi^\varepsilon(x_1,x_2) = \frac{1}{M^2}\int_{d_-}^{x_1}(x_1-z)\, e^{\mathrm{i}\frac{k_\varepsilon}{M}(x_1-z)} g(z,x_2)\,\mathrm{d}z$ for any $x_2\in[0,l]$. The Cauchy–Schwarz inequality then yields

$$|\psi^\varepsilon(x_1,x_2)|^2 \le \frac{1}{M^4}\left(\int_{d_-}^{x_1}(x_1-z)^2\, e^{-\frac{2\varepsilon}{M}(x_1-z)}\,\mathrm{d}z\right)\left(\int_{d_-}^{x_1}|g(z,x_2)|^2\,\mathrm{d}z\right),$$

and one obtains

$$\int_{d_-}^{d_+}\int_0^l |\psi^\varepsilon(x_1,x_2)|^2\,\mathrm{d}x_1\,\mathrm{d}x_2 \le C_{1\varepsilon}\,\|g\|^2_{L^2(\Omega)},$$

with $C_{1\varepsilon} = \frac{(d_+-d_-)^4}{M^4} e^{-\frac{2\varepsilon}{M}(d_+-d_-)}$.

Finally, if $x_1 > d_+$, we have

$$\psi^\varepsilon(x_1,x_2) = \left(-\frac{1}{M^2}\int_{d_-}^{d_+} z\, e^{-\mathrm{i}\frac{k_\varepsilon}{M}z} g(z,x_2)\,\mathrm{d}z + \frac{x_1}{M^2}\int_{d_-}^{d_+} e^{-\mathrm{i}\frac{k_\varepsilon}{M}z} g(z,x_2)\,\mathrm{d}z\right) e^{\mathrm{i}\frac{k_\varepsilon}{M}x_1}.$$

Observing that the variables can be separated in the expression above so that

$$\psi^\varepsilon(x_1,x_2) = (a_\varepsilon(x_2) + x_1\, b_\varepsilon(x_2))\; e^{\mathrm{i}\frac{k_\varepsilon}{M}x_1}, \tag{6.3}$$

one arrives at

$$\int_{d_+}^{+\infty}\int_0^l |\psi^\varepsilon(x_1,x_2)|^2\,\mathrm{d}x_1\,\mathrm{d}x_2 \le \|a_\varepsilon\|_{L^2([0,l])}\int_{d_+}^{+\infty} e^{-\frac{2\varepsilon}{M}x_1}\,\mathrm{d}x_1 + \|b_\varepsilon\|_{L^2([0,l])}\int_{d_+}^{+\infty}|x_1|\, e^{-\frac{2\varepsilon}{M}x_1}\,\mathrm{d}x_1.$$

Additionally, we have that $|a_\varepsilon(x_2)|^2 \le \frac{1}{M^4}\int_{d_-}^{d_+} z^2\, e^{-\frac{2\varepsilon}{M}z}\,\mathrm{d}z\int_{d_-}^{d_+}|g(z,x_2)|^2\,\mathrm{d}z$ and $|b_\varepsilon(x_2)|^2 \le \frac{1}{M^4}\int_{d_-}^{d_+} e^{-\frac{2\varepsilon}{M}z}\,\mathrm{d}z\int_{d_-}^{d_+}|g(z,x_2)|^2\,\mathrm{d}z$; hence

$$\int_{d_+}^{+\infty}\int_0^l |\psi^\varepsilon(x_1,x_2)|^2\,\mathrm{d}x_1\,\mathrm{d}x_2 \le C_{2\varepsilon}\,\|g\|^2_{L^2(\Omega)},$$

with $C_{2\varepsilon} = \frac{1}{M^4}\max\left(\int_{d_-}^{d_+} z^2 e^{-\frac{2\varepsilon}{M}z}\,\mathrm{d}z\int_{d_+}^{+\infty} e^{-\frac{2\varepsilon}{M}x_1}\,\mathrm{d}x_1, \int_{d_-}^{d_+} e^{-\frac{2\varepsilon}{M}z}\,\mathrm{d}z\int_{d_+}^{+\infty}|x_1|\, e^{-\frac{2\varepsilon}{M}x_1}\,\mathrm{d}x_1\right)$.

We finally set $C_\varepsilon = \sqrt{\max(C_{1\varepsilon}, C_{2\varepsilon})}$. ∎

REFERENCES

[1] S. Abarbanel, D. Gottlieb, and J. S. Hesthaven, *Well-posed perfectly matched layers for advective acoustics*, J. Comput. Phys., 154 (1999), pp. 266–283.

[2] E. Bécache, A.-S. Bonnet-Ben Dhia, and G. Legendre, *Perfectly matched layers for the convected Helmholtz equation*, SIAM J. Numer. Anal., 42 (2004), pp. 409–433.

[3] J.-P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[4] A.-S. Bonnet-Ben Dhia, L. Dahi, E. Lunéville, and V. Pagneux, *Acoustic diffraction by a plate in a uniform flow*, Math. Models Methods Appl. Sci., 12 (2002), pp. 625–647.

[5] A.-S. Bonnet-Ben Dhia, G. Legendre, and E. Lunéville, *Analyse mathématique de l'équation de Galbrun en écoulement uniforme*, C. R. Acad. Sci. Paris Sér. IIb Méc., 329 (2001), pp. 601–606.

[6] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[7] D. M. Eidus, *The principle of limiting absorption*, Amer. Math. Soc. Transl., 47 (1965), pp. 157–191.

[8] H. Galbrun, *Propagation d'une onde sonore dans l'atmosphère terrestre et théorie des zones de silence*, Gauthier-Villars, Paris, France, 1931.

[9] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.

[10] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, 1985.

[11] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, 1992.

[12] M. A. Hamdi, Y. Ousset, and G. Verchery, *A displacement method for the analysis of vibrations of coupled fluid-structure systems*, Internat. J. Numer. Methods Engrg., 13 (1978), pp. 139–150.

[13] J. S. Hesthaven, *On the analysis and construction of perfectly matched layers for the linearized Euler equations*, J. Comput. Phys., 142 (1998), pp. 129–147.

[14] F. Q. Hu, *On absorbing boundary conditions for linearized Euler equations by a perfectly matched layer*, J. Comput. Phys., 129 (1996), pp. 201–219.

[15] F. Q. Hu, *A stable, perfectly matched layer for linearized Euler equations in unsplit physical variables*, J. Comput. Phys., 173 (2001), pp. 455–480.

[16] G. Legendre, *Rayonnement acoustique dans un fluide en écoulement : analyse mathématique et numérique de l'équation de Galbrun*, Ph.D. thesis, Université Paris VI, Paris, France, 2003.

[17] D. Martin, *On-line documentation of* mélina, http://perso.univ-rennes1.fr/daniel.martin/melina/www/homepage.html.

[18] B. Poirée, *Les équations de l'acoustique linéaire et non-linéaire dans un écoulement de fluide parfait*, Acustica, 57 (1985), pp. 5–25.

[19] C. K. W. Tam, L. Auriault, and F. Cambuli, *Perfectly matched layer as an absorbing boundary condition for the linearized Euler equations in open and ducted domains*, J. Comput. Phys., 144 (1998), pp. 213–234.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1218–1247

FINITE ELEMENT APPROXIMATION OF SOLUBLE SURFACTANT SPREADING ON A THIN FILM*

JOHN W. BARRETT†, ROBERT NÜRNBERG†, AND MARK R. E. WARNER†

Abstract. We consider a fully practical finite element approximation of the following system of nonlinear degenerate parabolic equations:

$$\frac{\partial u}{\partial t} + \frac{1}{2}\,\nabla\cdot(u^2\,\nabla[\sigma(v)]) - \frac{1}{3}\,\nabla\cdot(u^3\,\nabla w) = 0,$$
$$w = -c\,\Delta u - \delta\,u^{-\nu} + a\,u^{-3},$$
$$\frac{\partial v}{\partial t} + \nabla\cdot(u\,v\,\nabla[\sigma(v)]) - \frac{1}{2}\,\nabla\cdot(u^2\,v\,\nabla w) - \rho_s\,\Delta v - K\,(\psi - v) = 0,$$
$$\frac{\partial \psi}{\partial t} + \frac{1}{2}\,u\,\nabla[\sigma(v)]\cdot\nabla\psi - \frac{1}{3}\,u^2\,\nabla w\cdot\nabla\psi - \rho_b\,u^{-1}\nabla\cdot(u\nabla\psi) + \beta\,K\,u^{-1}(\psi - v) = 0.$$

The above equations model a Marangoni driven thin film laden with a soluble surfactant, in which the bulk surfactant concentration has been vertically averaged. The model accounts for the presence of both attractive, $a \geq 0$, and repulsive, $\delta > 0$ with $\nu > 3$, van der Waals forces. Here u is the height of the film, v is the concentration of the interfacial surfactant species, ψ is the concentration of the surfactant species within the bulk phase, and $\sigma(v) := 1 - v$ is the typical surface tension. Moreover, $\rho_s \geq 0$, $\rho_b > 0$, and $c > 0$ are the inverses of the surface Peclet number, the bulk Peclet number, and the modified capillary number, respectively; finally, $\beta > 0$ and $K > 0$ are parameters that characterize the solubility and the rate of interfacial adsorption. In addition to showing stability bounds for our approximation, we prove convergence and hence existence of a solution to this nonlinear degenerate parabolic system (i) in one space dimension when $\rho_s > 0$ and, moreover, (ii) in two space dimensions if, in addition, $\nu \geq 7$. Furthermore, iterative schemes for solving the resulting nonlinear discrete system are discussed. Finally, some numerical experiments are presented.

Key words. thin film flow, surfactant, fourth order degenerate parabolic system, finite elements, convergence analysis

AMS subject classifications. 65M60, 65M12, 35K55, 35K65, 35K35, 76A20, 76D08

DOI. 10.1137/040618400

1. Introduction.

In the recent papers [5, 6], abbreviated to BGN and BN throughout this paper, the authors proposed and analyzed a fully practical finite element approximation of a system of nonlinear degenerate parabolic equations describing an *insoluble* surfactant driven monolayer. Here, we generalize that system to a model in which the chemical may demonstrate varying degrees of solubility allowing for adsorption and desorption between the bulk phase and an interfacial concentration. This extended model is given by

$$\frac{\partial u}{\partial t} + \frac{1}{2}\,\nabla\cdot(u^2\,\nabla[\sigma(v)]) - \frac{1}{3}\,\nabla\cdot(u^3\,\nabla w) = 0, \tag{1.1a}$$

$$w = -c\,\Delta u + \phi(u), \tag{1.1b}$$

$$\frac{\partial v}{\partial t} + \nabla\cdot(u\,v\,\nabla[\sigma(v)]) - \frac{1}{2}\,\nabla\cdot(u^2\,v\,\nabla w) - \rho_s\,\Delta v - K\,(\psi - v) = 0, \tag{1.1c}$$

*Received by the editors November 5, 2004; accepted for publication (in revised form) January 13, 2006; published electronically June 30, 2006.
http://www.siam.org/journals/sinum/44-3/61840.html

†Department of Mathematics, Imperial College, London, SW7 2AZ, UK (j.barrett@imperial.ac.uk, rn@imperial.ac.uk, m.r.warner@imperial.ac.uk). The third author was supported by EPSRC grant GR/S35660/01.

$$\frac{\partial \psi}{\partial t} + \frac{1}{2}\, u\, \nabla[\sigma(v)] \cdot \nabla \psi - \frac{1}{3}\, u^2\, \nabla w \cdot \nabla \psi - \rho_b\, u^{-1} \nabla \cdot\, (u\, \nabla \psi)$$

$$+ \beta\, K\, u^{-1}\, (\psi - v) = 0 \tag{1.1d}$$

in Ω_T, where $\Omega_T := \Omega \times (0, T]$ and Ω is a bounded domain in $\mathbb{R}^d$, $d = 1$ or 2. The above model, derived using lubrication theory and a cross-sectional averaging technique that removes the vertical dependence of the bulk species [12], models the flow of a surface tension gradient driven surfactant (chemical) laden thin film. Here u denotes the film height, v the concentration of the interfacial surfactant species, ψ the cross-sectionally averaged chemical concentration per unit height within the fluid layer, and w the pressure (reduced if van der Waals forces are present, that is, $\phi \not\equiv 0$). In addition, $\sigma \in C^1(\mathbb{R}_{\geq 0})$ with

$$\sigma(s) \geq 0, \quad \sigma'(s) < 0 \quad \forall\ s \in \mathbb{R}_{\geq 0} \tag{1.2}$$

is the constitutive equation of state relating the surface tension σ to the interfacial concentration v. We note that σ is a strictly monotonically decreasing function of v, which is natural to assume as the surfactant lowers surface tension. An empirical model proposed by Sheludko [15], often used in the engineering literature and that maps $\sigma : [0, 1] \to [0, 1]$, is

$$\sigma(s) := (\alpha + 1)\, [1 + \theta(\alpha)\, s]^{-3} - \alpha, \quad \text{where } \theta(\alpha) := (1 + \alpha^{-1})^{\frac{1}{3}} - 1, \tag{1.3}$$

in which $\alpha \in \mathbb{R}_{>0}$ relates to the activity of the surfactant; cf. [11, p. 262]. Of course the above model assumes that $v(\cdot, \cdot) \in [0, 1]$, which is a physically reasonable assumption. In modeling studies it is often further assumed that the surfactant concentration is dilute, in which case the limit $\alpha \to \infty$ is taken, and the equation of state (1.3) simplifies to $\sigma(s) := 1 - s$. For the van der Waals forces in (1.1b), we take the form suggested in [13]; that is,

$$\phi(u) = \phi^+(u) + \phi^-(u), \quad \phi^+(u) := -\delta\, u^{-\nu}, \quad \nu > 3, \quad \phi^-(u) := a\, u^{-3}, \tag{1.4}$$

where $a \in \mathbb{R}_{\geq 0}$ is the scaled dimensionless Hamaker constant and $\delta \in \mathbb{R}_{\geq 0}$ represents the effect of repulsive van der Waals forces. In (1.1a)–(1.1d), $\rho_s \in \mathbb{R}_{\geq 0}$, $\rho_b \in \mathbb{R}_{>0}$, and $c \in \mathbb{R}_{>0}$ are a nondimensional surface diffusivity (inverse of the surface Peclet number), a bulk diffusivity (inverse of the bulk Peclet number), and the modified capillary number, respectively. In order to permit the cross-sectional averaging process one has assumed that vertical diffusion is sufficiently fast for the bulk concentration to become independent of y, the vertical variable, at leading order. A necessary condition of this process is that the product of ρ_b and the dimensional film aspect ratio squared must be negligible at leading order in the lubrication approximation, and hence in practical applications ρ_b is not "too large." The parameter $\beta \in \mathbb{R}_{>0}$ indicates the degree of solubility of the chemical and emerges from the lubrication scaling as a ratio of the rate of adsorption to the rate of desorption of the chemical at the interface, $y = u$. We note that in the insoluble limit, $\beta \to \infty$, whereby the chemical accumulates preferentially at the interface, (1.1d) collapses to $\psi = v$, and the system (1.1a)–(1.1d) reduces to the insoluble surfactant system, (1.1a)–(1.1c) with $\psi \equiv v$, considered in BGN and BN. Finally, $K \in \mathbb{R}_{>0}$ is a parameter that describes the ratio of the time scale of the flow to the time scale of desorption. Applications of the system (1.1a)–(1.1d) range from the medical treatment of premature infants to industrial coating and drying processes; see BGN for further details and references.

As u and v can take on zero values, (1.1a)–(1.1d) is a degenerate parabolic system, which is fourth order in u. This degeneracy makes the analysis/numerical analysis of the system particularly difficult. As there is no maximum principle for parabolic equations of fourth order, a naive discretization does not guarantee the nonnegativity of the approximation to u. If $\delta = 0$, following [2], BGN imposed the nonnegativity of the discrete approximation to u as a constraint; whereas if $\delta > 0$, the positivity of the approximation to u can be guaranteed for an appropriate discretization through the singularity in ϕ^+. In both cases, BGN proposed a finite element approximation of the insoluble surfactant system, (1.1a)–(1.1c) with $\psi \equiv v$, and were able to derive stability bounds in space dimensions $d = 1$ and 2. However, their main convergence result was restricted to $\rho_s > 0$ and one space dimension. The latter was due to the fact that the a priori bounds they derived guarantee only in one space dimension that the discrete approximation to u is uniformly bounded and equicontinuous, which was necessary to be able to pass to the limit in the discrete problem. For similar reasons, the results on related degenerate parabolic equations of fourth order in [1, 2, 3, 4, 8, 10] were restricted to one space dimension. However, recently in [9], Grün proved convergence in two space dimensions of a finite element approximation to the thin film equation in the absence of surfactant/chemical, (1.1a)–(1.1b) with $v \equiv 0$ and $u(\cdot, 0) > 0$. In BN the techniques in BGN and [9] were adapted to propose a finite element approximation to the insoluble surfactant system, (1.1a)–(1.1c) with $\psi \equiv v$, and prove convergence in one space dimension if $\rho_s, a, \delta, u(\cdot, 0) > 0$ and, moreover, in two space dimensions if in addition $\nu \geq 7$. It is the aim of this paper to adapt the techniques in BN in order to prove convergence of a finite element approximation to (1.1a)–(1.1d). To this end, we will identify and exploit an underlying Lyapunov structure and extend the usage of two entropy-type estimates that were introduced in BGN and BN, respectively.

Throughout this paper, as in BGN and BN, we restrict ourselves to the linearized form of the constitutive equation of state

$$\sigma(v) := 1 - v, \tag{1.5}$$

the $\alpha \to \infty$ limit of (1.3). However, the techniques in this paper do apply to a general σ satisfying (1.2); see Remark 2.2 below. As remarked previously, the physically relevant values of v lie in the interval $[0, 1]$. Noting this, it is convenient for the analysis in this paper, as well as in BGN and BN, to replace the terms $u^i\, v$, $i = 1 \to 2$, in (1.1c) by $u^i\, \lambda(v)$, and similarly replace ψ in the first three terms of (1.1d) by $\lambda(\psi)$, where $\lambda : \mathbb{R} \to (-\infty, 1]$ is defined as

$$\lambda(s) := \min\{s, \lambda_M\}, \quad \text{with } \lambda_M := 1. \tag{1.6}$$

We will return to this point later in this section.

Altogether, in this paper we consider the following initial boundary value problem.

(P) Find functions $u, w, v, \psi : \Omega \times [0, T] \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \frac{1}{2}\,\nabla \cdot (u^2\, \nabla[\sigma(v)]) - \frac{1}{3}\,\nabla \cdot (u^3\, \nabla\, w) = 0 \qquad \text{in } \Omega_T, \tag{1.7a}$$

$$w = -c\,\Delta u + \phi(u) \qquad \text{in } \Omega_T, \tag{1.7b}$$

$$\frac{\partial v}{\partial t} + \nabla \cdot (u\, \lambda(v)\, \nabla[\sigma(v)]) - \frac{1}{2}\,\nabla \cdot (u^2\, \lambda(v)\, \nabla w) - \rho_s\,\Delta v - K(\psi - v) = 0 \qquad \text{in } \Omega_T, \tag{1.7c}$$

$$\frac{\partial (u\, \lambda(\psi))}{\partial t} + \frac{1}{2}\,\nabla \cdot \left(u^2\, \lambda(\psi)\, \nabla[\sigma(v)]\right)$$

$$-\frac{1}{3}\,\nabla\cdot\left(u^3\,\lambda(\psi)\,\nabla w\right)-\rho_b\,\nabla\cdot(u\,\nabla\psi)+\beta\,K\,(\psi-v)=0\quad\text{in }\Omega_T, \tag{1.7d}$$

$$u(x,0)=u^0(x)>0,\quad v(x,0)=v^0(x)\geq 0,\quad \psi(x,0)=\psi^0(x)\geq 0\quad\forall\, x\in\Omega, \tag{1.7e}$$

$$\begin{aligned}&\frac{1}{2}\,u^2\,\frac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}}-\frac{1}{3}\,u^3\,\frac{\partial w}{\partial\nu_{\partial\Omega}}=\frac{\partial u}{\partial\nu_{\partial\Omega}}=u\,\lambda(v)\,\frac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}}-\frac{1}{2}\,u^2\,\lambda(v)\,\frac{\partial w}{\partial\nu_{\partial\Omega}}-\rho_s\,\frac{\partial v}{\partial\nu_{\partial\Omega}}\\&=\frac{1}{2}\,u^2\,\lambda(\psi)\,\frac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}}-\frac{1}{3}\,u^3\,\lambda(\psi)\,\frac{\partial w}{\partial\nu_{\partial\Omega}}-\rho_b\,u\,\frac{\partial\psi}{\partial\nu_{\partial\Omega}}=0\quad\text{on }\partial\Omega\times(0,T),\end{aligned} \tag{1.7f}$$

where $\nu_{\partial\Omega}$ is normal to $\partial\Omega$, the Lipschitz boundary of Ω, and $T>0$ is a fixed positive time. In the above $c,\rho_b,K,\beta\in\mathbb{R}_{>0}$, and $\rho_s\in\mathbb{R}_{\geq 0}$ are given constants, while $\sigma\in C^1(\mathbb{R}_{\geq 0})$ and $\phi:\mathbb{R}_{>0}\to\mathbb{R}$ are given by (1.2) and (1.4), with $a\geq 0$, $\delta>0$, and u^0, v^0, and ψ^0 are given initial profiles.

Note that (1.7d) is just a combination of (1.1a) and the modified (1.1d), obtained by multiplying (1.1a) with $\lambda(\psi)$ and the modified (1.1d) with u. This is crucial for the analysis in this paper, as it allows us to exploit a Lyapunov structure that was not available before. The other main ingredients of our approach are two energy estimates for the surfactant driven flow combined with a regularization procedure. In particular, for any given $\varepsilon\in(0,\lambda_M)$, we introduce the regularized function

$$\lambda_\varepsilon(s):=\max\{\lambda(s),\varepsilon\}, \tag{1.8}$$

which yields the regularized system (P_ε), i.e., (P) with $\{u,w,v,\psi,\lambda\}$ replaced by $\{u_\varepsilon,w_\varepsilon,v_\varepsilon,\psi_\varepsilon,\lambda_\varepsilon\}$. On defining the horizontal velocity field $\vec{\mathcal{V}}_\varepsilon(y)$, similarly to BGN and BN, as

$$\vec{\mathcal{V}}_\varepsilon(y)=y\,\nabla[\sigma(v_\varepsilon)]+\left(\frac{1}{2}\,y^2-y\,u_\varepsilon\right)\nabla w_\varepsilon, \tag{1.9}$$

where the modified pressure $w_\varepsilon=-c\,\Delta u_\varepsilon+\phi(u_\varepsilon)$, we can recast the system (P_ε) in terms of this velocity field as follows:

$$\frac{\partial u_\varepsilon}{\partial t}+\nabla\cdot\left(\int_0^{u_\varepsilon}\vec{\mathcal{V}}_\varepsilon(y)\;\mathrm{d}y\right)=0, \tag{1.10a}$$

$$\frac{\partial v_\varepsilon}{\partial t}+\nabla\cdot(\lambda_\varepsilon(v_\varepsilon)\,\vec{\mathcal{V}}_\varepsilon(u_\varepsilon))=\rho_s\,\Delta v_\varepsilon+K\,(\psi_\varepsilon-v_\varepsilon)\,, \tag{1.10b}$$

$$\begin{aligned}&\frac{\partial(u_\varepsilon\,\lambda_\varepsilon(\psi_\varepsilon))}{\partial t}+\nabla\cdot\left(\lambda_\varepsilon(\psi_\varepsilon)\int_0^{u_\varepsilon}\vec{\mathcal{V}}_\varepsilon(y)\;\mathrm{d}y\right)=\rho_b\,\nabla\cdot\,(u_\varepsilon\,\nabla\psi_\varepsilon)-\beta\,K(\psi_\varepsilon-v_\varepsilon),\\&\nu_{\partial\Omega}\cdot\int_0^{u_\varepsilon}\vec{\mathcal{V}}_\varepsilon(y)\;\mathrm{d}y=\nu_{\partial\Omega}\cdot(\lambda_\varepsilon(v_\varepsilon)\,\vec{\mathcal{V}}_\varepsilon(u_\varepsilon)-\rho_s\,\nabla v_\varepsilon)\\&=\nu_{\partial\Omega}\cdot\left(\lambda_\varepsilon(\psi_\varepsilon)\int_0^{u_\varepsilon}\vec{\mathcal{V}}_\varepsilon(y)\;\mathrm{d}y-\rho_b\,u_\varepsilon\,\nabla\psi_\varepsilon\right)=0.\end{aligned} \tag{1.10c, 1.10d}$$

We see from (1.9) that (P) is derived on assuming a no-slip condition at $y=0$.

In order to derive the crucial energy estimates, we introduce the regularized function F_ε such that

$$F_\varepsilon''(s) = [\lambda_\varepsilon(s)]^{-1} \quad \text{and} \quad F_\varepsilon(1) = F_\varepsilon'(1) = 0, \tag{1.11}$$

which implies that

$$F_\varepsilon(s) := \begin{cases} \frac{s^2-\varepsilon^2}{2\,\varepsilon} + (\ln\varepsilon - 1)\,s + 1, & s \le \varepsilon, \\ s\,(\ln s - 1) + 1, & \varepsilon \le s \le 1, \\ \frac{1}{2}\,(s-1)^2, & 1 \le s. \end{cases} \tag{1.12}$$

Hence $F_\varepsilon \in C^{2,1}(\mathbb{R})$ and, for later purposes, we note that

$$F_\varepsilon(s) \ge \frac{s^2}{4} - \frac{1}{2} \quad \forall\ s \ge 0 \qquad \text{and} \qquad F_\varepsilon(s) \ge \frac{s^2}{2\,\varepsilon} \quad \forall\ s \le 0\,; \tag{1.13}$$

see (2.4) in BGN. In addition, it is easily deduced that

$$\frac{[F_\varepsilon'(s)]^2}{F_\varepsilon''(s)} \le \begin{cases} 2\,\varepsilon^{-1}\,[s]_-^2 + 8\,\exp(-1), & s \le \varepsilon, \\ 4\,\exp(-2), & \varepsilon \le s \le 1, \end{cases} \quad \text{where } [s]_\pm := \pm\max\{\pm s, 0\}. \tag{1.14}$$

We also introduce

$$\widehat{F}_\varepsilon(s) := F_\varepsilon(\lambda_\varepsilon(s)) \equiv \lambda_\varepsilon(s)\,F_\varepsilon'(s) - s + 1. \tag{1.15}$$

As F_ε is convex, it follows that

$$[\lambda_\varepsilon(s) - \lambda_\varepsilon(r)]F_\varepsilon'(s) \ge \widehat{F}_\varepsilon(s) - \widehat{F}_\varepsilon(r) \quad \forall\ r, s \in \mathbb{R}. \tag{1.16}$$

We will now derive several formal estimates for $\{u_\varepsilon, w_\varepsilon, v_\varepsilon, \psi_\varepsilon\}$. Testing equation (1.10a) with w_ε and combining and noting (1.9) and (1.10d) yields that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \Big[\frac{c}{2}\,|\nabla u_\varepsilon|^2 + \Phi(u_\varepsilon)\Big]\,\mathrm{d}x + \int_\Omega \left(\int_0^{u_\varepsilon} |\partial_y \vec{\mathcal{V}}_\varepsilon(y)|^2\,\mathrm{d}y\right)\mathrm{d}x = -\int_\Omega \vec{\mathcal{V}}_\varepsilon(u_\varepsilon)\cdot\nabla v_\varepsilon\ \mathrm{d}x, \tag{1.17}$$

where Φ is an antiderivative of ϕ, i.e., $\Phi' \equiv \phi$. Testing (1.10b) with $F_\varepsilon'(v_\varepsilon)$ and noting (1.10d) and (1.11) yields that

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega F_\varepsilon(v_\varepsilon)\ \mathrm{d}x + \rho_s \int_\Omega F_\varepsilon''(v_\varepsilon)\,|\nabla v_\varepsilon|^2\,\mathrm{d}x \\ &\qquad = \int_\Omega \vec{\mathcal{V}}_\varepsilon(u_\varepsilon)\cdot\nabla v_\varepsilon\ \mathrm{d}x + K\int_\Omega (\psi_\varepsilon - v_\varepsilon)\,F_\varepsilon'(v_\varepsilon)\ \mathrm{d}x. \end{aligned} \tag{1.18}$$

Moreover, it follows from testing (1.10c) with $F_\varepsilon'(\psi_\varepsilon)$ and testing (1.10a) with $\psi_\varepsilon - 1$, on noting (1.15), $F_\varepsilon'(\lambda_\varepsilon(\psi_\varepsilon))\,\frac{\partial[\lambda_\varepsilon(\psi_\varepsilon)]}{\partial t} = F_\varepsilon'(\psi_\varepsilon)\,\frac{\partial[\lambda_\varepsilon(\psi_\varepsilon)]}{\partial t}$, (1.10d), and (1.11), that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u_\varepsilon\,\widehat{F}_\varepsilon(\psi_\varepsilon)\ \mathrm{d}x &= \int_\Omega \frac{\partial(u_\varepsilon\,\lambda_\varepsilon(\psi_\varepsilon))}{\partial t}\,F_\varepsilon'(\psi_\varepsilon)\ \mathrm{d}x - \int_\Omega \frac{\partial u_\varepsilon}{\partial t}\,(\psi_\varepsilon - 1)\ \mathrm{d}x \\ &= -\rho_b\int_\Omega u_\varepsilon\,F_\varepsilon''(\psi_\varepsilon)\,|\nabla\psi_\varepsilon|^2\,\mathrm{d}x - \beta\,K\int_\Omega (\psi_\varepsilon - v_\varepsilon)\,F_\varepsilon'(\psi_\varepsilon)\ \mathrm{d}x. \end{aligned} \tag{1.19}$$

Combining (1.17), (1.18), and (1.19) yields that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \left[\frac{c}{2}\,|\nabla u_\varepsilon|^2 + \Phi(u_\varepsilon) + F_\varepsilon(v_\varepsilon) + \frac{1}{\beta}\,u_\varepsilon\,\widehat{F}_\varepsilon(\psi_\varepsilon)\right]\mathrm{d}x + \rho_s\int_\Omega F_\varepsilon''(v_\varepsilon)\,|\nabla v_\varepsilon|^2\,\mathrm{d}x\\ + \frac{\rho_b}{\beta}\int_\Omega u_\varepsilon\,F_\varepsilon''(\psi_\varepsilon)\,|\nabla\psi_\varepsilon|^2\,\mathrm{d}x + \int_\Omega\left(\int_0^{u_\varepsilon}|\partial_y\vec{\mathcal{V}}_\varepsilon(y)|^2\,\mathrm{d}y\right)\mathrm{d}x\\ + K\int_\Omega (F_\varepsilon'(\psi_\varepsilon) - F_\varepsilon'(v_\varepsilon))\,(\psi_\varepsilon - v_\varepsilon)\,\mathrm{d}x = 0.\end{aligned} \tag{1.20}$$

Due to the singularity in Φ at the origin, it immediately follows from (1.20) that $u_\varepsilon(\cdot,t) > 0$ for all $t \in (0,T)$ if $u_\varepsilon(\cdot,0) > 0$. In addition, it follows from (1.9), on applying a Young inequality

$$|r\,s| \le \frac{\gamma}{2}\,r^2 + \frac{1}{2\gamma}\,s^2 \quad \forall\ r,\,s \in \mathbb{R},\quad \gamma \in \mathbb{R}_{>0}, \tag{1.21}$$

that

$$\int_0^{u_\varepsilon}|\partial_y\vec{\mathcal{V}}_\varepsilon(y)|^2\,\mathrm{d}y \ge \frac{1}{8}\,u_\varepsilon\,|\nabla[\sigma(v_\varepsilon)]|^2 + \frac{1}{21}\,u_\varepsilon^3\,|\nabla w_\varepsilon|^2\,; \tag{1.22}$$

see (1.7) in BGN for details.

From (1.20), (1.11), (1.8), and (1.6), one can deduce uniform bounds on ∇v_ε and $u_\varepsilon\,\nabla\psi_\varepsilon$ in $L^2(\Omega_T)$. We note the crucial role that the cut-off λ_M in (1.6) plays in these estimates. Of course the cut-off λ_M can be chosen arbitrarily large, and it played no real role in our finite element approximation of (P_ε), as our computed approximations to both v_ε and ψ_ε were always strictly less than λ_M, which we set to be one. However, as it does not appear possible to obtain a priori $L^\infty(\Omega_T)$ bounds on v_ε and ψ_ε, one requires some (arbitrarily large) cut-off in (1.6) and hence in certain coefficients in (1.7a)–(1.7f), as the Lyapunov structure above is based on the relationship (1.11).

In order to obtain the second estimate, we also define a function $G \in C^\infty(\mathbb{R}_{>0})$ such that $u^3\,\nabla[G'(u)] = \nabla u$; that is,

$$G''(s) = s^{-3} \Rightarrow G'(s) = -\frac{1}{2}\,s^{-2} \Rightarrow G(s) = \frac{1}{2}\,s^{-1}, \tag{1.23}$$

where the constants of integration have been chosen to be zero. Testing (1.10a) with $G'(u_\varepsilon)$ and testing (1.1b) with $-\Delta u_\varepsilon$ yields, on noting (1.10d) and applying (1.21) (see (1.10)–(1.12) in BN for details), that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega G(u_\varepsilon)\,\mathrm{d}x + \frac{c}{3}\int_\Omega |\Delta u_\varepsilon|^2\,\mathrm{d}x + \frac{1}{4}\int_\Omega (\phi^+)'(u_\varepsilon)\,|\nabla u_\varepsilon|^2\,\mathrm{d}x\\ \le C\left[\int_\Omega u_\varepsilon\,|\nabla[\sigma(v_\varepsilon)]|^2\,\mathrm{d}x + \int_\Omega |\nabla u_\varepsilon|^2\,\mathrm{d}x\right].\end{aligned} \tag{1.24}$$

From (1.24), (1.20), and (1.22) one can show that u_ε is uniformly bounded in $L^2(0,T;H^2(\Omega))$ if $u_\varepsilon(\cdot,0) > 0$. Furthermore, the bound (1.13) together with (1.20) yields that $\int_{\Omega_T}[v_\varepsilon]_-^2\,\mathrm{d}x\,\mathrm{d}t \le C\,\varepsilon$. One can use this, together with the last bound in (1.20), to deduce that $\int_{\Omega_T} F_\varepsilon(\psi_\varepsilon)\,\mathrm{d}x\,\mathrm{d}t \le C$ and hence that $\int_{\Omega_T}[\psi_\varepsilon]_-^2\,\mathrm{d}x\,\mathrm{d}t \le C\,\varepsilon$.

It is the goal of this paper to derive a finite element method that is consistent with the formal energy estimates (1.20) and (1.24). We stress that it is only the bound (1.24) that requires the presence of the repulsive van der Waals forces, $\delta > 0$, to control the surfactant term in (1.10a). In the absence of a surfactant/chemical, (1.24) holds with $(a, \delta > 0)$ and without $(a = \delta = 0)$ van der Waals forces.

This paper is organized as follows. In section 2 we formulate a fully practical finite element approximation of the degenerate system (P) and derive discrete analogues of the energy estimates (1.20) and (1.24). In doing so, we adapt a technique introduced in [17] and [10] for deriving a discrete entropy bound for the thin film equation. In section 3 we prove convergence, and hence existence, of a solution to the system (P) in one space dimension if $\rho_s, \delta > 0$ and in two space dimensions if, in addition, $\nu \geq 7$. In section 4 we state an iterative scheme for solving the nonlinear discrete system at each time level and present some numerical computations in both one and two space dimensions.

Finally we note that there is very little work in the PDE literature on surfactant-type problems. To our knowledge, there is no work on the degenerate soluble system (P). A numerical study of (P) can be found in [16]. We stress that this paper is a nontrivial extension of the insoluble surfactant system, (1.7a)–(1.7c) with $\psi = v$, studied in BN. First, one has to identify the Lyapunov structure for (P), which we have outlined in this introduction. Second, proving convergence of our finite element approximation to (P) and hence proving existence of a solution to (P) is far more difficult in this case. As stated earlier, we will establish a finite element approximation, which satisfies discrete analogues of (1.20) and (1.24). For the insoluble surfactant one has control on the discrete analogue of $\int_{\Omega_T} |\nabla v_\varepsilon|^2 \,\mathrm{d}x\,\mathrm{d}t$ if $\rho_s > 0$, whereas for the chemical we have control only on the discrete analogue of $\int_{\Omega_T} u_\varepsilon \,|\nabla \psi_\varepsilon|^2 \,\mathrm{d}x\,\mathrm{d}t$. This degeneracy, as we have no a priori positive lower bound on u_ε, causes a number of new difficulties in the convergence analysis.

Notation and auxiliary results. Let $D \subset \mathbb{R}^d$, $d = 1$ or 2, with a Lipschitz boundary ∂D if $d = 2$. We adopt the standard notation for Sobolev spaces, denoting the norm of $W^{m,q}(D)$ $(m \in \mathbb{N}, q \in [1, \infty])$ by $\|\cdot\|_{m,q,D}$ and the seminorm by $|\cdot|_{m,q,D}$. We extend these norms and seminorms in the natural way to the corresponding spaces of vector and matrix valued functions. For $q = 2$, $W^{m,2}(D)$ will be denoted by $H^m(D)$ with the associated norm and seminorm written as, respectively, $\|\cdot\|_{m,D}$ and $|\cdot|_{m,D}$. For notational convenience, we drop the domain subscript on the above norms and seminorms in the case $D \equiv \Omega$. Throughout $(\cdot,\cdot)$ denotes the standard L^2 inner product over Ω, while q' denotes for any $q \in [1, \infty]$ the "dual exponent" such that $\frac{1}{q} + \frac{1}{q'} = 1$. In addition we define $\fint \eta := (\eta, 1)/\underline{m}(\Omega)$ for all $\eta \in L^1(\Omega)$, where $\underline{m}(D)$ denotes the measure of D.

It is convenient to introduce the operator $\mathcal{G} : (W^{1,q'}(\Omega))' \to W^{1,q}(\Omega)$ such that

$$(\nabla \mathcal{G} z, \nabla \eta) + (\mathcal{G} z, \eta) = \langle z, \eta \rangle_{q'} \quad \forall\ \eta \in W^{1,q'}(\Omega), \tag{1.25}$$

where here and throughout $\langle \cdot, \cdot \rangle_{q'}$ denotes the duality pairing between $(W^{1,q'}(\Omega))'$ and $W^{1,q'}(\Omega)$ for any $q \in (1, 2]$.

Throughout C denotes a generic constant independent of h, τ, and ε, the mesh and temporal discretization parameters and the regularization parameter. In addition, $C(a_1, \ldots, a_I)$ denotes a constant depending on the arguments $\{a_i\}_{i=1}^I$. Furthermore, $.^{(\star)}$ denotes an expression with or without the superscript $\star$.

2. Finite element approximation. We consider the finite element approximation of (P) under the following assumptions on the mesh.

(A) Let Ω be a convex polygonal domain if $d = 2$. Let $\{\mathcal{T}^h\}_{h>0}$ be a quasi-uniform family of partitionings of Ω into disjoint open simplices κ with $h_\kappa := \operatorname{diam}(\kappa)$ and $h := \max_{\kappa\in\mathcal{T}^h} h_\kappa$, so that $\overline{\Omega} = \cup_{\kappa\in\mathcal{T}^h}\overline{\kappa}$. In addition, it is assumed for $d = 2$ that all simplices $\kappa \in \mathcal{T}^h$ are right-angled.

Associated with $\mathcal{T}^h$ is the finite element space $S^h := \{\chi \in C(\overline{\Omega}) : \chi|_\kappa \text{ is linear for all } \kappa \in \mathcal{T}^h\} \subset H^1(\Omega)$. We also introduce $S^h_{\geq 0} := \{\chi \in S^h : \chi \geq 0 \text{ in } \Omega\} \subset H^1_{\geq 0}(\Omega) := \{\eta \in H^1(\Omega) : \eta \geq 0 \text{ a.e. in } \Omega\}$ and similarly $S^h_{>0}$ and $H^1_{>0}(\Omega)$. Let J be the set of nodes of $\mathcal{T}^h$ and $\{p_j\}_{j\in J}$ the coordinates of these nodes. Let $\{\chi_j\}_{j\in J}$ be the standard basis functions for S^h; that is, $\chi_j \in S^h_{\geq 0}$ and $\chi_j(p_i) = \delta_{ij}$ for all $i, j \in J$. We introduce $\pi^h : C(\overline{\Omega}) \to S^h$, the interpolation operator, such that $(\pi^h\eta)(p_j) = \eta(p_j)$ for all $j \in J$. A discrete semi-inner product on $C(\overline{\Omega})$ is then defined by

$$(\eta_1, \eta_2)^h := \int_\Omega \pi^h(\eta_1(x)\,\eta_2(x))\,\mathrm{d}x = \sum_{j\in J} m_j\,\eta_1(p_j)\,\eta_2(p_j), \tag{2.1}$$

where $m_j := (1, \chi_j) > 0$. The induced discrete seminorm is then $|\eta|_h := [\,(\eta,\eta)^h\,]^{\frac{1}{2}}$, where $\eta \in C(\overline{\Omega})$. We introduce also the L^2 projection $Q^h : L^2(\Omega) \to S^h$ defined by $(Q^h\eta, \chi)^h = (\eta, \chi)$ for all $\chi \in S^h$.

Similarly to the approach in [10, 17], we introduce matrices $\Lambda_\varepsilon : S^h \to [L^\infty(\Omega)]^{d\times d}$ and $\Xi : S^h_{>0} \to [L^\infty(\Omega)]^{d\times d}$ such that for all $z^h \in S^h$, $\chi \in S^h_{>0}$ and a.e. in Ω

$$\Lambda_\varepsilon(z^h),\ \Xi(\chi) \text{ are symmetric and positive semidefinite,} \tag{2.2a}$$

$$\Lambda_\varepsilon(z^h)\,\nabla\pi^h[F'_\varepsilon(z^h)] = \nabla z^h, \quad [\Xi(\chi)]^3\,\nabla\pi^h[G'(\chi)] = \nabla\chi. \tag{2.2b}$$

The construction of these matrices can be found in BN. Throughout we make use of the fact that the matrices $\Xi(\chi)$ and $\Lambda_\varepsilon(z^h)$ commute for any $\chi \in S^h_{>0}$ and $z^h \in S^h$.

As in BN, it is convenient to split Φ (recall (1.20)) into its convex and concave parts. We have for given $a \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{>0}$, and $\nu > 3$ that for all $s \in \mathbb{R}_{>0}$

$$\Phi(s) = \Phi^+(s) + \Phi^-(s), \quad \text{where } \Phi^+(s) := \frac{\delta}{\nu-1}\,s^{1-\nu}, \quad \Phi^-(s) := -\frac{a}{2}\,s^{-2}. \tag{2.3}$$

It holds, on recalling (1.4), that $\phi^+ \equiv (\Phi^+)'$ and $\phi^- \equiv (\Phi^-)'$. For future reference, we note that the following hold for all $r, s \in \mathbb{R}_{>0}$:

$$\begin{aligned} \Phi(s) &\geq -\frac{a\,(\nu-3)}{2\,(\nu-1)}\left(\frac{a}{\delta}\right)^{\frac{2}{\nu-3}} \quad \text{and} \\ -\Phi^-(s) &\leq \frac{a\,(\nu-3)}{2\,(\nu-1)}\left(\frac{2\,a}{\delta}\right)^{\frac{2}{\nu-3}} + \frac{1}{2}\,\Phi^+(s). \end{aligned} \tag{2.4}$$

In addition to $\mathcal{T}^h$, let $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ be a partitioning of $[0, T]$ into possibly variable time steps $\tau_n := t_n - t_{n-1}$, $n = 1 \to N$. We set $\tau := \max_{n=1\to N}\tau_n$. For any given $\varepsilon \in (0, 1)$, we then consider the following fully practical finite element approximation of (P) with $\sigma(v)$ given by (1.5) and $\phi(u)$ given by (1.4):

($\mathrm{P}_\varepsilon^{h,\tau}$) For $n \geq 1$ find $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\} \in [S^h]^4$ such that for all $\chi \in S^h$

$$\left(\frac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, \chi\right)^h + \frac{1}{3}\left([\Xi(U_\varepsilon^n)]^3 \nabla W_\varepsilon^n, \nabla\chi\right)$$

$$= -\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac{3}{2}} [\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}} \nabla V_\varepsilon^{n-1}, \nabla\chi\right), \tag{2.5a}$$

$$c\,(\nabla U_\varepsilon^n, \nabla\chi) + (\phi^+(U_\varepsilon^n) + \phi^-(U_\varepsilon^{n-1}), \chi)^h = (W_\varepsilon^n, \chi)^h, \tag{2.5b}$$

$$\left(\frac{V_\varepsilon^n - V_\varepsilon^{n-1}}{\tau_n}, \chi\right)^h + \rho_s\,(\nabla V_\varepsilon^n, \nabla\chi) + \left(\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla\chi\right)$$

$$- K\,(\Psi_\varepsilon^n - V_\varepsilon^n, \chi)^h = -\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^2\,\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla W_\varepsilon^n, \nabla\chi\right), \tag{2.5c}$$

$$\left(\frac{U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1})}{\tau_n}, \chi\right)^h + \rho_b\,(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n, \nabla\chi)$$

$$+ \frac{1}{3}\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi_\varepsilon^n)\,\nabla W_\varepsilon^n, \nabla\chi\right) + \beta\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, \chi)^h$$

$$= -\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac{3}{2}} [\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\Lambda_\varepsilon(\Psi_\varepsilon^n)\,\nabla V_\varepsilon^{n-1}, \nabla\chi\right), \tag{2.5d}$$

where $U_\varepsilon^0 \in S^h_{>0}$, $V_\varepsilon^0 \in S^h$, and $\Psi_\varepsilon^0 \in S^h$ are approximations of u^0, v^0, and ψ^0, respectively, e.g., $U_\varepsilon^0 \equiv \pi^h u^0$ or $Q^h u^0$ and similarly for V_ε^0 and Ψ_ε^0.

Remark 2.1. ($\mathrm{P}_\varepsilon^{h,\tau}$) is the natural extension of the approximation of the insoluble surfactant system studied in BN. In particular, on setting $\Psi_\varepsilon^n \equiv V_\varepsilon^n$, $n = 1 \to N$, equations (2.5a)–(2.5c) collapse to the approximation in BN. Note that we approximate u^2 by $[\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}$ in (2.5d) in order for our discrete stability bounds, the analogues of (1.20) and (1.24), to hold. Furthermore, as $U_\varepsilon^0 > 0$, one can ensure that $\Xi(U_\varepsilon^{n-1})$ and $\phi^-(U_\varepsilon^{n-1})$ are well defined for $n \geq 1$; see Theorem 2.3 below.

Remark 2.2. The restriction of σ to the linear case (1.5) is not crucial for the analysis in this paper. However, this choice simplifies our considerations and is also more practical. Different choices of σ, e.g., (1.3), can be incorporated; see Remark 2.2 in BN for details.

Below we recall some well-known results concerning S^h for any $\kappa \in \mathcal{T}^h$, $\chi, z^h \in S^h$, $m \in \{0,1\}$, $p \in [1,\infty]$, $s \in [2,\infty]$ if $d = 1$, and $s \in (2,\infty]$ if $d = 2$:

$$|\chi|_{m,r,\kappa} \leq C\,h_\kappa^{-d\,(\frac{1}{p}-\frac{1}{r})}\,|\chi|_{m,p,\kappa} \quad \text{for any } r \in [p,\infty]\,, \tag{2.6}$$

$$\lim_{h\to 0} \|(I-\pi^h)\eta\|_{1,s} = 0 \quad \forall\ \eta \in W^{1,s}(\Omega)\,, \tag{2.7}$$

$$|(I-\pi^h)\eta|_{m,s,\kappa} \leq C\,h_\kappa^{1-m}\,|\eta|_{1,s,\kappa} \quad \forall\ \eta \in W^{1,s}(\kappa)\,, \tag{2.8}$$

$$\|\pi^h[\chi\,z^h]\|_{1,p} \leq C\,\left[|\chi\,z^h|_{0,p} + |\chi\,\nabla z^h|_{0,p} + |z^h\,\nabla\chi|_{0,p}\right], \tag{2.9}$$

$$\int_\kappa \chi^2\,\mathrm{d}x \leq \int_\kappa \pi^h[\chi^2]\,\mathrm{d}x \leq (d+2)\int_\kappa \chi^2\,\mathrm{d}x\,, \tag{2.10}$$

$$|(\chi, z^h) - (\chi, z^h)^h| \leq |(I-\pi^h)(\chi\,z^h)|_{0,1} \leq C\,h^{1+m}\,|\chi|_{m,p}\,|z^h|_{1,p'}. \tag{2.11}$$

On recalling (2.1), we see that the operator Q^h satisfies

$$(Q^h\eta)(p_j) = m_j^{-1}(\eta,\chi_j) \quad \forall\, j\in J \Rightarrow |Q^h\eta|_{0,\infty} \le |\eta|_{0,\infty}, \quad \forall\, \eta\in L^\infty(\Omega), \tag{2.12}$$

and, in addition, it holds for $m\in\{0,1\}$ that

$$|(I-Q^h)\eta|_{m,r} \le C\,h^{1-m}\,|\eta|_{1,r} \quad \forall\,\eta\in W^{1,r}(\Omega), \quad \text{for any } r\in[2,\infty]. \tag{2.13}$$

We note that assumption (A) and (1.11) yield that

$$\int_\kappa \nabla z^h\cdot\nabla\pi^h[F_\varepsilon'(z^h)]\,\mathrm{d}x \ge |z^h|^2_{1,\kappa} \qquad \forall\, z^h\in S^h, \quad \forall\,\kappa\in\mathcal{T}^h; \tag{2.14}$$

see (2.13) in BN for details. It is also easily established that

$$|z^h|_{0,q} \le C\,h^{-1}\,\|\mathcal{G}z^h\|_{1,q} \quad \forall\, z^h\in S^h, \quad \text{for any } q\in(1,2]. \tag{2.15}$$

We note that the results (2.13) and (2.15) above exploit the fact that we have a quasi-uniform family of partitionings $\{\mathcal{T}^h\}_{h>0}$. Finally, we introduce the "discrete Laplacian" operator $\Delta^h : S^h\to S^h$ such that $(\Delta^h z^h,\chi)^h = -(\nabla z^h,\nabla\chi)$ for all $\chi\in S^h$.

THEOREM 2.3. *Let $\phi(\cdot)$ satisfy* (1.4) *with $\delta>0$. Let the assumptions* (A) *hold and $\{U_\varepsilon^{n-1},V_\varepsilon^{n-1},\Psi_\varepsilon^{n-1}\}\in S^h_{>0}\times[S^h]^2$. Then for all $\varepsilon\in(0,1)$ and for all h, $\tau_n>0$ there exists a solution $\{U_\varepsilon^n,W_\varepsilon^n,V_\varepsilon^n,\Psi_\varepsilon^n\}\in S^h_{>0}\times[S^h]^3$ to the nth step of* ($\mathrm{P}_\varepsilon^{h,\tau}$) *with $\fint U_\varepsilon^n = \fint U_\varepsilon^{n-1}$ and $\fint(V_\varepsilon^n+\frac{1}{\beta}\,\pi^h[U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)]) = \fint(V_\varepsilon^{n-1}+\frac{1}{\beta}\,\pi^h[U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1})])$.*

Proof. Existence of a solution $\{U_\varepsilon^n,W_\varepsilon^n\}\in S^h_{>0}\times S^h$ to (2.5a)–(2.5b) follows from Lemma 2.1 in BN. To prove the existence of $\{V_\varepsilon^n,\Psi_\varepsilon^n\}$ to (2.5c)–(2.5d) we will make use of the *Brouwer fixed point theorem* (see, e.g., [14, Theorem 9.36, p. 357]). This is a nontrivial extension of the existence proof for V_ε^n in Theorem 2.1 of BGN. Let $\mathcal{J}:=\#J$ and let $g:\mathbb{R}^{2\mathcal{J}}\to\mathbb{R}^{2\mathcal{J}}$ be defined by

$$\begin{aligned}
g_j(\underline{V},\underline{\Psi}) &:= \frac{1}{\tau_n}\,(V,\chi_j)^h + \rho_s\,(\nabla V,\nabla\chi_j) + (\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V)\,\nabla V,\nabla\chi_j)\\
&\qquad - K\,(\Psi-V,\chi_j)^h + \frac{1}{2}\,([\Xi(U_\varepsilon^n)]^2\,\Lambda_\varepsilon(V)\,\nabla W_\varepsilon^n,\nabla\chi_j),\\
g_{j+\mathcal{J}}(\underline{V},\underline{\Psi}) &:= \frac{1}{\beta}\left[\frac{1}{\tau_n}\,(U_\varepsilon^n\,\lambda_\varepsilon(\Psi),\chi_j)^h + \rho_b\,(U_\varepsilon^n\,\nabla\Psi,\nabla\chi_j)\right.\\
&\qquad + \frac{1}{3}\,([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi)\,\nabla W_\varepsilon^n,\nabla\chi_j) + \beta\,K\,(\Psi-V,\chi_j)^h\\
&\qquad \left. + \frac{1}{2}\,([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\Lambda_\varepsilon(\Psi)\,\nabla V_\varepsilon^{n-1},\nabla\chi_j)\right] \quad \forall\, j\in J,
\end{aligned}$$

where $V\equiv\sum_{j\in J}V_j\,\chi_j$, $\Psi\equiv\sum_{j\in J}\Psi_j\,\chi_j$, and $\{\underline{V},\underline{\Psi}\}:=(V_1,\ldots,V_{\mathcal{J}},\Psi_1,\ldots,\Psi_{\mathcal{J}})^T\in\mathbb{R}^{2\mathcal{J}}$. Hence a solution $\{V_\varepsilon^n,\Psi_\varepsilon^n\}$ of (2.5c)–(2.5d) is such that for $j=1\to\mathcal{J}$

$$g_j(\underline{V}_\varepsilon^n,\underline{\Psi}_\varepsilon^n) = \tfrac{1}{\tau_n}(V_\varepsilon^{n-1},\chi_j)^h, \qquad g_{j+\mathcal{J}}(\underline{V}_\varepsilon^n,\underline{\Psi}_\varepsilon^n) = \tfrac{1}{\beta\,\tau_n}(U_\varepsilon^{n-1}\lambda_\varepsilon(\Psi_\varepsilon^{n-1}),\chi_j)^h.$$

On noting Lemma 2.1 in BGN we have that g is continuous, and hence it is sufficient to show that g is coercive. We have that for all $\{V,\Psi\}\in[S^h]^2$

$$\begin{aligned}\sum_{j\in J}\left(g_j(\underline{V},\underline{\Psi})\,V_j+g_{j+\mathcal{J}}(\underline{V},\underline{\Psi})\,\Psi_j\right)&=\frac{1}{\tau_n}\,|V|_h^2+\rho_s\,|V|_1^2\\&+\left(\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V)\,\nabla V,\nabla V\right)+K\,|\Psi-V|_h^2+\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^2\,\Lambda_\varepsilon(V)\,\nabla W_\varepsilon^n,\nabla V\right)\\&+\frac{1}{\beta}\left[\frac{1}{\tau_n}\left(U_\varepsilon^n\,\lambda_\varepsilon(\Psi),\Psi\right)^h+\rho_b\,|(U_\varepsilon^n)^{\frac12}\,\nabla\Psi|_0^2+\frac{1}{3}\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi)\,\nabla W_\varepsilon^n,\nabla\Psi\right)\right.\\&\left.+\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac32}\,[\Xi(U_\varepsilon^{n-1})]^{\frac12}\,\Lambda_\varepsilon(\Psi)\,\nabla V_\varepsilon^{n-1},\nabla\Psi\right)\right].\end{aligned}\tag{2.16}$$

From (1.21) and Lemma 2.1 in BGN we have that

$$\begin{aligned}&\frac{1}{2}\left|\left([\Xi(U_\varepsilon^n)]^2\,\Lambda_\varepsilon(V)\,\nabla W_\varepsilon^n,\nabla V\right)\right|\\&\quad\le\frac{1}{2}\left(\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V)\,\nabla V,\nabla V\right)+\frac{1}{8}\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(V)\,\nabla W_\varepsilon^n,\nabla W_\varepsilon^n\right)\\&\quad\le\frac{1}{2}\left(\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V)\,\nabla V,\nabla V\right)+C(U_\varepsilon^n,\,W_\varepsilon^n),\end{aligned}\tag{2.17a}$$

and similarly, on additionally noting Lemma 2.2 in BN and $U_\varepsilon^n\in S^h_{>0}$, that

$$\frac{1}{3\beta}\left|\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi)\,\nabla W_\varepsilon^n,\nabla\Psi\right)\right|\le\frac{\rho_b}{4\beta}\left(U_\varepsilon^n\,\nabla\Psi,\nabla\Psi\right)+C(\beta,\rho_b,U_\varepsilon^n,\,W_\varepsilon^n),\tag{2.17b}$$

$$\begin{aligned}&\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac32}\,[\Xi(U_\varepsilon^{n-1})]^{\frac12}\,\Lambda_\varepsilon(\Psi)\,\nabla V_\varepsilon^{n-1},\nabla\Psi\right)\\&\quad\le\frac{\rho_b}{4\beta}\left(U_\varepsilon^n\,\nabla\Psi,\nabla\Psi\right)+C(\beta,\rho_b,U_\varepsilon^{n-1},U_\varepsilon^n,\,V_\varepsilon^{n-1}).\end{aligned}\tag{2.17c}$$

Moreover, it follows from (1.8) that

$$\frac{1}{\tau_n\beta}\left|\left(U_\varepsilon^n\,\lambda_\varepsilon(\Psi),\Psi\right)^h\right|\le\gamma\,|\Psi|_h^2+C(\gamma,\beta,\tau_n,U_\varepsilon^n)\qquad\text{for any fixed }\gamma>0.\tag{2.17d}$$

Combining (2.16) and (2.17)–(2.17d) yields that

$$\begin{aligned}\sum_{j\in J}\left(g_j(\underline{V},\underline{\Psi})\,V_j+g_{j+\mathcal{J}}(\underline{V},\underline{\Psi})\,\Psi_j\right)&\ge\frac{1}{\tau_n}\,|V|_h^2+K\,|\Psi-V|_h^2-\gamma\,|\Psi|_h^2-C\\&\ge\left(\frac{1}{\tau_n}-K_0\right)|V|_h^2+\left(\frac{K_0}{2}-\gamma\right)|\Psi|_h^2-C\\&\ge\frac{K_0}{4}\left[\,|V|_h^2+|\Psi|_h^2\,\right]-C\qquad\forall\;\{V,\Psi\}\in[S^h]^2,\end{aligned}\tag{2.18}$$

where $K_0\in(0,K]$ and $\gamma\in\mathbb{R}_{>0}$ are chosen sufficiently small. Hence the coerciveness of g follows from (2.18) and (2.1). Therefore, on noting the aforementioned theorem, we have existence of $\{V_\varepsilon^n,\Psi_\varepsilon^n\}$ to (2.5c)–(2.5d) and hence existence of a solution

$\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\}$ to $(\mathrm{P}_\varepsilon^{h,\tau})$. The integral relations follow immediately from choosing $\chi \equiv 1$ in (2.5a), (2.5c), and (2.5d). ∎

LEMMA 2.4. *Let the assumptions of Theorem* 2.3 *hold. Then for all* $\varepsilon \in (0,1)$ *and for all* h, $\tau_n > 0$ *a solution* $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\}$ *to the* n*th step of* $(\mathrm{P}_\varepsilon^{h,\tau})$ *is such that*

$$\begin{aligned}
&\mathcal{E}(U_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n) + \frac{c}{2}\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + \frac{1}{2}\,|V_\varepsilon^n - V_\varepsilon^{n-1}|_h^2 \\
&\qquad + \rho_s\,\tau_n\,(\nabla V_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(V_\varepsilon^n)]) + \frac{1}{\beta}\,\rho_b\,\tau_n\,\big(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n, \nabla\pi^h[F_\varepsilon'(\Psi_\varepsilon^n)]\big) \\
&\qquad + \tau_n\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n) - F_\varepsilon'(V_\varepsilon^n))^h + \frac{\tau_n}{24}\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n) \\
&\qquad + \frac{5}{8}\,\tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n) \\
&\quad \le \mathcal{E}(U_\varepsilon^{n-1}, V_\varepsilon^{n-1}, \Psi_\varepsilon^{n-1}) + \frac{\tau_n}{2}\,(\Xi(U_\varepsilon^{n-1})\,\nabla V_\varepsilon^{n-1}, \nabla V_\varepsilon^{n-1}),
\end{aligned} \tag{2.19}$$

where $\mathcal{E}(U_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n) := \frac{c}{2}\,|U_\varepsilon^n|_1^2 + (\Phi(U_\varepsilon^n) + F_\varepsilon(V_\varepsilon^n), 1)^h + \frac{1}{\beta}\,(U_\varepsilon^n, \widehat{F}_\varepsilon(\Psi_\varepsilon^n))^h$ *and* $\widehat{F}_\varepsilon$ *is as defined in* (1.15). *Furthermore, if* $\phi(\cdot)$ *satisfies* (1.4) *with* $\nu \ge 7$, *then*

$$\begin{aligned}
&(G(U_\varepsilon^n), 1)^h + \frac{\tau_n}{4}\,(\nabla\pi^h[\phi^+(U_\varepsilon^n)], \nabla U_\varepsilon^n) + \frac{c}{3}\,\tau_n\,|\Delta^h U_\varepsilon^n|_h^2 \\
&\qquad \le (G(U_\varepsilon^{n-1}), 1)^h + \frac{\tau_n}{8}(\nabla\pi^h[\phi^+(U_\varepsilon^{n-1})], \nabla U_\varepsilon^{n-1}) \\
&\qquad\quad + C\,\tau_n\,\big[\,|U_\varepsilon^n|_1^2 + |U_\varepsilon^{n-1}|_1^2\,\big] + \frac{\tau_n}{4}\,(\Xi(U_\varepsilon^{n-1})\,\nabla V_\varepsilon^{n-1}, \nabla V_\varepsilon^{n-1}).
\end{aligned} \tag{2.20}$$

Proof. First, upon substitution of the trial function $\chi \equiv W_\varepsilon^n$ into the height equation (2.5a), we obtain

$$\begin{aligned}
&(U_\varepsilon^n - U_\varepsilon^{n-1}, W_\varepsilon^n)^h + \frac{\tau_n}{3}([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n) \\
&\qquad = -\frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla V_\varepsilon^{n-1}, \nabla W_\varepsilon^n).
\end{aligned} \tag{2.21}$$

Substituting $\chi \equiv U_\varepsilon^n - U_\varepsilon^{n-1}$ into (2.5b) and noting the identity $2\,r\,(r-s) = (r^2 - s^2) + (r-s)^2$, the convexity of Φ^+, and the concavity of Φ^- gives

$$\begin{aligned}
&\frac{c}{2}\,|U_\varepsilon^n|_1^2 - \frac{c}{2}\,|U_\varepsilon^{n-1}|_1^2 + \frac{c}{2}\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + (\Phi(U_\varepsilon^n) - \Phi(U_\varepsilon^{n-1}), 1)^h \\
&\qquad \le (W_\varepsilon^n, U_\varepsilon^n - U_\varepsilon^{n-1})^h.
\end{aligned} \tag{2.22}$$

Combining (2.21) and (2.22) yields

$$\begin{aligned}
&\frac{c}{2}\,|U_\varepsilon^n|_1^2 + (\Phi(U_\varepsilon^n), 1)^h + \frac{\tau_n}{3}\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n) + \frac{c}{2}\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 \\
&\quad \le \frac{c}{2}\,|U_\varepsilon^{n-1}|_1^2 + (\Phi(U_\varepsilon^{n-1}), 1)^h - \frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla V_\varepsilon^{n-1}, \nabla W_\varepsilon^n).
\end{aligned} \tag{2.23}$$

Furthermore, choosing $\chi \equiv \pi^h[F_\varepsilon'(V_\varepsilon^n)]$ in the interfacial equation (2.5c) and noting the properties (2.2a)–(2.2b) of Λ_ε yields that

$$\begin{aligned}&(V_\varepsilon^n - V_\varepsilon^{n-1}, F_\varepsilon'(V_\varepsilon^n))^h + \rho_s\,\tau_n\,(\nabla V_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(V_\varepsilon^n)])\\ &\quad + \tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n) - \tau_n\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, \pi^h[F_\varepsilon'(V_\varepsilon^n)])^h\\ &\quad = -\frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^2\,\nabla W_\varepsilon^n, \nabla V_\varepsilon^n).\end{aligned} \tag{2.24}$$

Now $F_\varepsilon'' \geq 1$ implies that

$$(V_\varepsilon^n - V_\varepsilon^{n-1}, F_\varepsilon'(V_\varepsilon^n))^h \geq (F_\varepsilon(V_\varepsilon^n) - F_\varepsilon(V_\varepsilon^{n-1}), 1)^h + \frac{1}{2}\,|V_\varepsilon^{n-1} - V_\varepsilon^n|_h^2\,. \tag{2.25}$$

Combining (2.24) and (2.25) gives

$$\begin{aligned}&(F_\varepsilon(V_\varepsilon^n), 1)^h + \frac{1}{2}\,|V_\varepsilon^{n-1} - V_\varepsilon^n|_h^2 + \rho_s\,\tau_n\,(\nabla V_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(V_\varepsilon^n)])\\ &\quad + \tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n) - \tau_n\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(V_\varepsilon^n))^h\\ &\quad \leq (F_\varepsilon(V_\varepsilon^{n-1}), 1)^h - \frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^2\,\nabla W_\varepsilon^n, \nabla V_\varepsilon^n).\end{aligned} \tag{2.26}$$

Similarly, choosing $\chi \equiv \pi^h[F_\varepsilon'(\Psi_\varepsilon^n)]$ in the bulk chemical equation (2.5d) and using the properties (2.2a)–(2.2b) gives

$$\begin{aligned}&(U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1}), F_\varepsilon'(\Psi_\varepsilon^n))^h + \tau_n\,\rho_b\,(U_\varepsilon^n\,\nabla \Psi_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(\Psi_\varepsilon^n)])\\ &\qquad + \frac{\tau_n}{3}\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla \Psi_\varepsilon^n) + \tau_n\,\beta\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n))^h\\ &\quad = -\frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla V_\varepsilon^{n-1}, \nabla \Psi_\varepsilon^n).\end{aligned}$$

Combining this with the height equation (2.5a) for $\chi \equiv \Psi_\varepsilon^n$ gives

$$\begin{aligned}&(U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1}), F_\varepsilon'(\Psi_\varepsilon^n))^h - (U_\varepsilon^n - U_\varepsilon^{n-1}, \Psi_\varepsilon^n)^h\\ &\quad = -\tau_n\,\rho_b\,(U_\varepsilon^n\,\nabla \Psi_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(\Psi_\varepsilon^n)]) - \tau_n\,\beta\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n))^h.\end{aligned} \tag{2.27}$$

We recall (1.15) and (1.16) and also note that $(U_\varepsilon^n - U_\varepsilon^{n-1}, 1)^h = 0$, which follows from inserting $\chi \equiv 1$ into (2.5a). Then the left-hand side of (2.27) may be rewritten as

$$\begin{aligned}&(U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1}), F_\varepsilon'(\Psi_\varepsilon^n))^h - (U_\varepsilon^n - U_\varepsilon^{n-1}, \Psi_\varepsilon^n)^h\\ &\quad = (U_\varepsilon^n - U_\varepsilon^{n-1}, \lambda_\varepsilon(\Psi_\varepsilon^n)\,F_\varepsilon'(\Psi_\varepsilon^n) - \Psi_\varepsilon^n)^h + (\lambda_\varepsilon(\Psi_\varepsilon^n) - \lambda_\varepsilon(\Psi_\varepsilon^{n-1}), U_\varepsilon^{n-1}\,F_\varepsilon'(\Psi_\varepsilon^n))^h\\ &\quad \geq (U_\varepsilon^n - U_\varepsilon^{n-1}, \widehat{F}_\varepsilon(\Psi_\varepsilon^n) - 1)^h + (\widehat{F}_\varepsilon(\Psi_\varepsilon^n) - \widehat{F}_\varepsilon(\Psi_\varepsilon^{n-1}), U_\varepsilon^{n-1})^h\\ &\quad = (U_\varepsilon^n\,\widehat{F}_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\widehat{F}_\varepsilon(\Psi_\varepsilon^{n-1}), 1)^h,\end{aligned}$$

and thus

$$\begin{aligned}&(U_\varepsilon^n\,\widehat{F}_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\widehat{F}_\varepsilon(\Psi_\varepsilon^{n-1}), 1)^h\\ &\quad \leq -\tau_n\,\rho_b\,(U_\varepsilon^n\,\nabla \Psi_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(\Psi_\varepsilon^n)]) - \tau_n\,\beta\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n))^h.\end{aligned} \tag{2.28}$$

Combining (2.23), (2.26), and (2.28) and noting Young's inequality (1.21) yields that

$$\begin{aligned}
&\mathcal{E}(U_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n) + \frac{c}{2}\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + \frac{1}{2}\,|V_\varepsilon^n - V_\varepsilon^{n-1}|_h^2 + \rho_s\,\tau_n\,(\nabla V_\varepsilon^n, \nabla \pi^h[F_\varepsilon'(V_\varepsilon^n)])\\
&\qquad + \frac{1}{\beta}\,\tau_n\,\rho_b\,(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n, \nabla\pi^h[F_\varepsilon'(\Psi_\varepsilon^n)]) + \tau_n\,K\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n) - F_\varepsilon'(V_\varepsilon^n))^h\\
&\qquad + \frac{\tau_n}{3}\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n) + \tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n)\\
&\le \mathcal{E}(U_\varepsilon^{n-1}, V_\varepsilon^{n-1}, \Psi_\varepsilon^{n-1}) - \frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^2\,\nabla W_\varepsilon^n, \nabla V_\varepsilon^n)\\
&\qquad - \frac{\tau_n}{2}\,([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla W_\varepsilon^n, \nabla V_\varepsilon^{n-1})\\
&\le \mathcal{E}(U_\varepsilon^{n-1}, V_\varepsilon^{n-1}, \Psi_\varepsilon^{n-1}) + \frac{\zeta+\gamma}{4}\,\tau_n\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n)\\
&\qquad + \frac{\tau_n}{4\zeta}\,(\Xi(U_\varepsilon^{n-1})\,\nabla V_\varepsilon^{n-1}, \nabla V_\varepsilon^{n-1}) + \frac{\tau_n}{4\gamma}\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n)
\end{aligned}$$

for arbitrary choices of ζ, $\gamma > 0$. Choosing $\zeta = \frac{1}{2}$ and $\gamma = \frac{2}{3}$ leads to the desired result for the discrete energy structure (2.19).

The desired result (2.20) was derived in Lemma 2.4 in BN. ∎

Remark 2.5. We note that (2.19) and (2.20) are the discrete analogues of the formal energy estimates (1.20) (on noting (1.22)) and (1.24), respectively.

THEOREM 2.6. *Let $\phi(\cdot)$ satisfy* (1.4) *with $\delta > 0$. Let the assumptions* (A) *hold and $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \in S_{>0}^h \times [S^h]^2$. Then for all $\varepsilon \in (0,1)$, $h > 0$ and for all time partitions $\{\tau_n\}_{n=1}^N$ a solution $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\}_{n=1}^N$ to $(\mathrm{P}_\varepsilon^{h,\tau})$ is such that $\fint U_\varepsilon^n = \fint U_\varepsilon^0$ and $\fint (V_\varepsilon^n + \frac{1}{\beta}\,\pi^h[U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)]) = \fint (V_\varepsilon^0 + \frac{1}{\beta}\,\pi^h[U_\varepsilon^0\,\lambda_\varepsilon(\Psi_\varepsilon^0)])$, $n = 1 \to N$, and if $\tau_n \le \frac{5}{4}\,\omega\,\tau_{n-1}$, $n = 2 \to N$, for an $\omega \in (0,1)$, then*

$$\begin{aligned}
&c\max_{n=1\to N}\|U_\varepsilon^n\|_1^2 + \max_{n=1\to N}(\Phi(U_\varepsilon^n),1)^h + \max_{n=1\to N}(F_\varepsilon(V_\varepsilon^n),1)^h\\
&\qquad + c\sum_{n=1}^N \|U_\varepsilon^n - U_\varepsilon^{n-1}\|_1^2 + \sum_{n=1}^N |V_\varepsilon^n - V_\varepsilon^{n-1}|_0^2\\
&\qquad + \rho_s\sum_{n=1}^N \tau_n\,(\nabla V_\varepsilon^n, \nabla\pi^h[F_\varepsilon'(V_\varepsilon^n)]) + \sum_{n=1}^N \tau_n\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n, \nabla W_\varepsilon^n)\\
&\qquad + (1-\omega)\sum_{n=1}^N \tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n, \nabla V_\varepsilon^n) + \rho_b\sum_{n=1}^N \tau_n\,(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n, \nabla\Psi_\varepsilon^n)\\
&\qquad + K\sum_{n=1}^N \tau_n\,(\Psi_\varepsilon^n - V_\varepsilon^n, F_\varepsilon'(\Psi_\varepsilon^n) - F_\varepsilon'(V_\varepsilon^n))^h \le C\,\mathcal{C}^0,
\end{aligned} \tag{2.29a}$$

where

$$\mathcal{C}^0 := 1 + \|U_\varepsilon^0\|_1^2 + (\Phi(U_\varepsilon^0) + F_\varepsilon(V_\varepsilon^0),1)^h + (U_\varepsilon^0, \widehat{F}_\varepsilon(\Psi_\varepsilon^0))^h + (\Xi(U_\varepsilon^0)\,\nabla V_\varepsilon^0, \nabla V_\varepsilon^0). \tag{2.29b}$$

In addition,

$$\max_{n=1\to N}|V_\varepsilon^n|_0^2+\varepsilon^{-1}\max_{n=1\to N}|\pi^h[V_\varepsilon^n]_-|_0^2+\rho_s\sum_{n=1}^N\tau_n\,\|V_\varepsilon^n\|_1^2+K\sum_{n=1}^N\tau_n\,|\Psi_\varepsilon^n-V_\varepsilon^n|_0^2$$

$$+\sum_{n=1}^N\tau_n\,\big[\,(F_\varepsilon(\Psi_\varepsilon^n),1)^h+|\Psi_\varepsilon^n|_0^2+\varepsilon^{-1}\,|\pi^h[\Psi_\varepsilon^n]_-|_0^2\,\big]\le C\,\mathcal{C}^0, \tag{2.30}$$

and, on letting $B_\varepsilon^n:=\pi^h[U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)]$, $n=0\to N$, *we have that*

$$\sum_{n=1}^N\tau_n\left[\left\|\mathcal{G}\left[\frac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}\right]\right\|_{1,q}^2+\left\|\mathcal{G}\left[\frac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}\right]\right\|_{1,q}^2+\left\|\mathcal{G}\left[\frac{B_\varepsilon^n-B_\varepsilon^{n-1}}{\tau_n}\right]\right\|_{1,q}^2\right]$$

$$+\sum_{n=1}^N\tau_n\,\|B_\varepsilon^n\|_{1,q}^2\le C\left(\max_{n=0\to N}\{|[\Xi(U_\varepsilon^n)]^3|_{0,r},|(U_\varepsilon^n)^{\frac12}|_{0,r}\}\right)\mathcal{C}^0, \tag{2.31a}$$

where $q=2$ *and* $r=\infty$ *if* $d=1$, $q\in(1,2)$; *and* $r=\frac{2q}{2-q}$ *if* $d=2$; *and*

$$|[\Xi(U_\varepsilon^n)]^\alpha|_{0,s}\le C\,\|U_\varepsilon^n\|_1^\alpha\qquad\forall\ \alpha\in(0,\infty),\quad\forall\ s\in\begin{cases}[1,\infty] & \text{if } d=1,\\ [1,\infty) & \text{if } d=2.\end{cases} \tag{2.31b}$$

Furthermore, if $\phi(\cdot)$ *satisfies* (1.4) *with* $\nu\ge7$, *then*

$$\max_{n=1\to N}(G(U_\varepsilon^n),1)^h+c\sum_{n=1}^N\tau_n\,|\Delta^hU_\varepsilon^n|_h^2+\sum_{n=1}^N\tau_n\,(\nabla\pi^h[\phi^+(U_\varepsilon^n)],\nabla U_\varepsilon^n)$$

$$\le C\,\big[\,\mathcal{C}^0+(G(U_\varepsilon^0),1)^h+(\nabla\pi^h[\phi^+(U_\varepsilon^0)],\nabla U_\varepsilon^0)\,\big]. \tag{2.32}$$

Proof. Summing the discrete energy estimate (2.19) from $n=1\to k$ and using $\tau_n\le\frac54\,\omega\,\tau_{n-1}$, $n=2\to k$, yields for any $k\le N$ that

$$\begin{aligned}&\mathcal{E}(U_\varepsilon^k,V_\varepsilon^k,\Psi_\varepsilon^k)+\frac12\sum_{n=1}^k\big[c\,|U_\varepsilon^n-U_\varepsilon^{n-1}|_1^2+|V_\varepsilon^n-V_\varepsilon^{n-1}|_h^2\big]\\&\quad+\rho_s\sum_{n=1}^k\tau_n\,(\nabla V_\varepsilon^n,\nabla\pi^h[F_\varepsilon'(V_\varepsilon^n)])+\frac{1}{\beta}\,\rho_b\sum_{n=1}^k\tau_n\,(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n,\nabla\pi^h[F_\varepsilon'(\Psi_\varepsilon^n)])\\&\quad+\sum_{n=1}^k\tau_n\left[K\,(\Psi_\varepsilon^n-V_\varepsilon^n,F_\varepsilon'(\Psi_\varepsilon^n)-F_\varepsilon'(V_\varepsilon^n))^h+\frac{1}{24}\,([\Xi(U_\varepsilon^n)]^3\,\nabla W_\varepsilon^n,\nabla W_\varepsilon^n)\right]\\&\quad+\frac58\,(1-\omega)\sum_{n=1}^k\tau_n\,(\Xi(U_\varepsilon^n)\,\nabla V_\varepsilon^n,\nabla V_\varepsilon^n)\\&\quad\le\mathcal{E}(U_\varepsilon^0,V_\varepsilon^0,\Psi_\varepsilon^0)+\frac{\tau_1}{2}\,(\Xi(U_\varepsilon^0)\,\nabla V_\varepsilon^0,\nabla V_\varepsilon^0).\end{aligned} \tag{2.33}$$

Similarly to (2.14), it holds that

$$(U_\varepsilon^n\,\nabla z^h,\nabla\pi^h[F_\varepsilon'(z^h)])\ge|(U_\varepsilon^n)^{\frac12}\,\nabla z^h|_0^2\qquad\forall\ z^h\in S^h. \tag{2.34}$$

Therefore, on noting (1.13), (2.4), (2.1), (2.10), (2.34), and a Poincaré inequality, the bounds (2.29a) follow directly from (2.33).

Combining the third bound in (2.29a) and (1.13) yields the first two bounds in (2.30). These, together with the sixth bound in (2.29a), yield, on noting (2.14), the third bound in (2.30). Moreover, the fourth bound in (2.30) follows from (1.11) and the last bound in (2.29a). We will now prove the final three bounds in (2.30). First, by the convexity of F_ε, we have for all r, $s \in \mathbb{R}$ that

(2.35)
$$F_\varepsilon(r) \leq F_\varepsilon(s) + (r-s)\,F_\varepsilon'(r) = F_\varepsilon(s) + (r-s)\,(F_\varepsilon'(r) - F_\varepsilon'(s)) + (r-s)\,F_\varepsilon'(s).$$

The last term on the right-hand side of (2.35) is only nonnegative if either $r \leq s \leq 1$ or $r \geq s \geq 1$, in which case we have that

(2.36)
$$2\,(r-s)\,F_\varepsilon'(s) \leq F_\varepsilon''(s)\,(r-s)^2 + \frac{[F_\varepsilon'(s)]^2}{F_\varepsilon''(s)} \leq (r-s)\,(F_\varepsilon'(r) - F_\varepsilon'(s)) + \frac{[F_\varepsilon'(s)]^2}{F_\varepsilon''(s)}.$$

The fifth bound in (2.30) then follows from (2.35), (2.36), (2.29a), (1.14), and the first two bounds in (2.30). This, together with (1.13), then yields the last two bounds in (2.30).

From (1.25), (2.5c), Lemma 2.1 in BGN, and (2.13) we obtain that

$$\begin{aligned}
&\left(\nabla\mathcal{G}\left[\frac{V_\varepsilon^n - V_\varepsilon^{n-1}}{\tau_n}\right], \nabla\eta\right) + \left(\mathcal{G}\left[\frac{V_\varepsilon^n - V_\varepsilon^{n-1}}{\tau_n}\right], \eta\right) = \left(\frac{V_\varepsilon^n - V_\varepsilon^{n-1}}{\tau_n}, Q^h\eta\right)^h \\
&\quad = -\rho_s\,(\nabla V_\varepsilon^n, \nabla Q^h\eta) - \Bigg(\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla V_\varepsilon^n \\
&\qquad + \frac{1}{2}\,[\Xi(U_\varepsilon^n)]^2\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla W_\varepsilon^n, \nabla Q^h\eta\Bigg) + K\,(\Psi_\varepsilon^n - V_\varepsilon^n, Q^h\eta)^h \qquad (2.37)\\
&\quad \leq C[\,\rho_s\,|\nabla V_\varepsilon^n|_0 + |[\Xi(U_\varepsilon^n)]^{\frac12}|_{0,r}(|[\Xi(U_\varepsilon^n)]^{\frac12}\,\nabla V_\varepsilon^n|_0 \\
&\qquad + |[\Xi(U_\varepsilon^n)]^{\frac32}\,\nabla W_\varepsilon^n|_0)\,]|\eta|_{1,q'} + C\,K\,|\Psi_\varepsilon^n - V_\varepsilon^n|_0\,|\eta|_{0,q'}.
\end{aligned}$$

In a similar fashion, it follows from (2.5d) that

(2.38)
$$\begin{aligned}
&\left(\nabla\mathcal{G}\left[\frac{B_\varepsilon^n - B_\varepsilon^{n-1}}{\tau_n}\right], \nabla\eta\right) + \left(\mathcal{G}\left[\frac{B_\varepsilon^n - B_\varepsilon^{n-1}}{\tau_n}\right], \eta\right) \\
&\quad = \left(\frac{U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n) - U_\varepsilon^{n-1}\,\lambda_\varepsilon(\Psi_\varepsilon^{n-1})}{\tau_n}, Q^h\eta\right)^h \\
&\quad = -\rho_b\left(U_\varepsilon^n\,\nabla\Psi_\varepsilon^n, \nabla Q^h\eta\right) - \frac{1}{3}\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi_\varepsilon^n)\,\nabla W_\varepsilon^n, \nabla Q^h\eta\right) \\
&\qquad - \beta\,K\left(\Psi_\varepsilon^n - V_\varepsilon^n, Q^h\eta\right)^h - \frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac32}\,[\Xi(U_\varepsilon^{n-1})]^{\frac12}\,\Lambda_\varepsilon(\Psi_\varepsilon^n)\,\nabla V_\varepsilon^{n-1}, \nabla Q^h\eta\right) \\
&\quad \leq C[\,|[\Xi(U_\varepsilon^n)]^{\frac32}|_{0,r}(|[\Xi(U_\varepsilon^{n-1})]^{\frac12}\,\nabla V_\varepsilon^{n-1}|_0 + |[\Xi(U_\varepsilon^n)]^{\frac32}\,\nabla W_\varepsilon^n|_0) \\
&\qquad + \rho_b\,|(U_\varepsilon^n)^{\frac12}|_{0,r}\,|(U_\varepsilon^n)^{\frac12}\,\nabla\Psi_\varepsilon^n|_0\,]|\eta|_{1,q'} + C\,\beta\,K\,|\Psi_\varepsilon^n - V_\varepsilon^n|_0\,|\eta|_{0,q'}.
\end{aligned}$$

Moreover, it follows from (2.9) and (1.8) that

$$\begin{aligned}\|B_\varepsilon^n\|_{1,q} &\leq C\,\left[|U_\varepsilon^n\,\pi^h[\lambda_\varepsilon(\Psi_\varepsilon^n)]|_{0,q} + |U_\varepsilon^n\,\nabla\pi^h[\lambda_\varepsilon(\Psi_\varepsilon^n)]|_{0,q} + |\pi^h[\lambda_\varepsilon(\Psi_\varepsilon^n)]\,\nabla U_\varepsilon^n|_{0,q}\right]\\ &\leq C\,\left[\|U_\varepsilon^n\|_1 + |(U_\varepsilon^n)^{\frac12}|_{0,r}\,|(U_\varepsilon^n)^{\frac12}\,\nabla\Psi_\varepsilon^n|_0\right].\end{aligned} \tag{2.39}$$

Combining (2.37), a similar bound for the discrete U_ε time derivative (see [6, (2.73)]), (2.38), (2.39), the assumptions on τ_n, and the bounds (2.29a) and (2.30) yields the bounds (2.31a).

The desired result (2.31b) was proved in Theorem 2.2 in BN.

Finally, summing (2.20) from $n = 1 \to k$, observing that $\tau_n \leq \frac{5}{4}\,\omega\,\tau_{n-1}$, $n = 2 \to k$, and noting the first and eighth bounds in (2.29a) yields (2.32). ∎

Remark 2.7. We note that all the results in this section hold also in the absence of attractive van der Waals forces, $a = 0$. The same holds true for the results quoted from BN, even though it was not explicitly stated there.

LEMMA 2.8. *Let u^0, v^0, $\psi^0 \in H^1_{\geq 0}(\Omega)$, with $u^0 \in L^\infty(\Omega)$ and $u^0(x) \geq \zeta > 0$ for a.e. $x \in \Omega$, and let the assumptions* (A) *hold. On choosing either $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \equiv \{Q^h u^0, Q^h v^0, Q^h \psi^0\}$ or $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \equiv \{\pi^h u^0, \pi^h v^0, \pi^h \psi^0\}$ if either $d = 1$ or $\{u^0, v^0, \psi^0\} \in [W^{1,e}(\Omega)]^3$ with $e > 2$, it follows that $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \in [S^h_{\geq 0}]^3$ with $U_\varepsilon^0 \geq \zeta$ are such that for all $h > 0$*

$$\mathcal{C}^0 + (G(U_\varepsilon^0), 1)^h + (\nabla\pi^h[\phi^+(U_\varepsilon^0)], \nabla U_\varepsilon^0) \leq C. \tag{2.40}$$

Proof. The desired result (2.40) follows immediately from (2.29b), (2.12), (2.8), (2.13), (2.3), (1.12), (1.23), (1.4), and Lemma 2.2 in BN. ∎

3. Convergence. Let

$$U_\varepsilon(t) := \frac{t - t_{n-1}}{\tau_n}\,U_\varepsilon^n + \frac{t_n - t}{\tau_n}\,U_\varepsilon^{n-1}, t \in [t_{n-1}, t_n], \quad n \geq 1, \tag{3.1a}$$

and

$$U_\varepsilon^+(t) := U_\varepsilon^n, \qquad U_\varepsilon^-(t) := U_\varepsilon^{n-1}, t \in (t_{n-1}, t_n], \quad n \geq 1. \tag{3.1b}$$

We note for future reference that

$$U_\varepsilon - U_\varepsilon^\pm = (t - t_n^\pm)\,\frac{\partial U_\varepsilon}{\partial t}, \quad t \in (t_{n-1}, t_n), \quad n \geq 1, \tag{3.2}$$

where $t_n^+ := t_n$ and $t_n^- := t_{n-1}$. We introduce also $\bar\tau(t) := \tau_n$ for $t \in (t_{n-1}, t_n]$, $n \geq 1$.

Using the above notation, and introducing analogous notation for $W_\varepsilon, V_\varepsilon, \Psi_\varepsilon$, and B_ε (recall (2.31a)), ($\mathrm{P}_\varepsilon^{h,\tau}$) can be restated as follows.

Find $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\} \in C([0,T]; S^h) \times L^\infty(0,T; S^h) \times [C([0,T]; S^h)]^2$ such that

for all $\chi \in L^2(0,T;S^h)$

$$\int_0^T \left[\left(\frac{\partial U_\varepsilon}{\partial t}, \chi \right)^h + \frac{1}{3} \left([\Xi(U_\varepsilon^+)]^3 \, \nabla W_\varepsilon^+, \nabla \chi \right) \right] \mathrm{d}t$$

$$= -\frac{1}{2} \int_0^T \left([\Xi(U_\varepsilon^+)]^{\frac{3}{2}} \, [\Xi(U_\varepsilon^-)]^{\frac{1}{2}} \, \nabla V_\varepsilon^-, \nabla \chi \right) \mathrm{d}t, \tag{3.3a}$$

$$\int_0^T \left[\left(\frac{\partial V_\varepsilon}{\partial t}, \chi \right)^h + \rho_s \left(\nabla V_\varepsilon^+, \nabla \chi \right) + \left(\Xi(U_\varepsilon^+) \, \Lambda_\varepsilon(V_\varepsilon^+) \, \nabla V_\varepsilon^+, \nabla \chi \right) \right.$$

$$\left. - K \left(\Psi_\varepsilon^+ - V_\varepsilon^+, \chi \right)^h \right] \mathrm{d}t = -\frac{1}{2} \int_0^T \left([\Xi(U_\varepsilon^+)]^2 \, \Lambda_\varepsilon(V_\varepsilon^+) \, \nabla W_\varepsilon^+, \nabla \chi \right) \mathrm{d}t, \tag{3.3b}$$

$$\int_0^T \left[\left(\frac{\partial B_\varepsilon}{\partial t}, \chi \right)^h + \rho_b \left(U_\varepsilon^+ \, \nabla \Psi_\varepsilon^+, \nabla \chi \right) + \frac{1}{3} \left([\Xi(U_\varepsilon^+)]^3 \, \Lambda_\varepsilon(\Psi_\varepsilon^+) \, \nabla W_\varepsilon^+, \nabla \chi \right) \right.$$

$$\left. + \beta \, K \left(\Psi_\varepsilon^+ - V_\varepsilon^+, \chi \right)^h \right] \mathrm{d}t$$

$$= -\frac{1}{2} \int_0^T \left([\Xi(U_\varepsilon^+)]^{\frac{3}{2}} [\Xi(U_\varepsilon^-)]^{\frac{1}{2}} \, \Lambda_\varepsilon(\Psi_\varepsilon^+) \, \nabla V_\varepsilon^-, \nabla \chi \right) \mathrm{d}t, \tag{3.3c}$$

where for a.a. $t \in (0,T)$ and for all $z^h \in S^h$

$$\left(W_\varepsilon^+(\cdot,t), z^h \right)^h = c \left(\nabla U_\varepsilon^+(\cdot,t), \nabla z^h \right) + \left(\phi^+(U_\varepsilon^+(\cdot,t)) + \phi^-(U_\varepsilon^-(\cdot,t)), z^h \right)^h ; \tag{3.3d}$$

that is, $W_\varepsilon^+ \equiv -c \, \Delta^h U_\varepsilon^+ + \pi^h[\phi^+(U_\varepsilon^+) + \phi^-(U_\varepsilon^-)]$.

LEMMA 3.1. *Let $\rho_s > 0$, $\phi(\cdot)$ satisfy (1.4) with $\delta > 0$, and $u^0 \in H^1(\Omega) \cap L^\infty(\Omega)$ with $u^0 \geq \zeta > 0$ a.e. and $v^0, \psi^0 \in H^1_{\geq 0}(\Omega)$. Let $\{\mathcal{T}^h, U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0, \{\tau_n\}_{n=1}^N, \varepsilon\}_{h>0}$ be such that*

(i) *either $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \equiv \{Q^h u^0, Q^h v^0, Q^h \psi^0\}$ or $\{U_\varepsilon^0, V_\varepsilon^0, \Psi_\varepsilon^0\} \equiv \{\pi^h u^0, \pi^h v^0, \pi^h \psi^0\}$ if either $d = 1$ or $\{u^0, v^0, \psi^0\} \in [W^{1,e}(\Omega)]^3$ with $e > 2$;*

(ii) *Ω and $\{\mathcal{T}^h\}_{h>0}$ fulfill assumption (A), $\varepsilon \in (0,1)$, and $\tau_n \leq \frac{5}{4} \, \omega \, \tau_{n-1}$, $n = 2 \to N$, for an $\omega \in (0,1)$;*

(iii) *$\tau \, h^{-d\left(1-\frac{2}{p}\right)} \to 0$ and $\varepsilon \, h^{-d\left(\frac{1}{2}-\frac{1}{p}\right)} \to 0$ as $h \to 0$, where $p = 2$ if $d = 1$ and $p > 2$ if $d = 2$.*

Then there exist a subsequence of $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}_h$, where $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}$ solve ($\mathrm{P}_\varepsilon^{h,\tau}$), and functions

$$u \in L^\infty(0,T;H^1_{\geq 0}(\Omega)) \cap H^1(0,T;(W^{1,q'}(\Omega))'), \tag{3.4a}$$

$$v \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1_{\geq 0}(\Omega)) \cap H^1(0,T;(W^{1,q'}(\Omega))'), \tag{3.4b}$$

$$\psi \in L^2(0,T;L^2_{\geq 0}(\Omega)), \text{ such that} \tag{3.4c}$$

$$\lambda(v), \, \lambda(\psi) \in L^\infty(\Omega_T), \tag{3.4d}$$

with $u(\cdot,0) = u^0(\cdot)$ *in* Y_1, $v(\cdot,0) = v^0(\cdot)$ *in* Y_2, *where* $H^1(\Omega) \overset{c}{\hookrightarrow} Y_1$, $L^2(\Omega) \overset{c}{\hookrightarrow} Y_2$, *and* $\fint u(\cdot,t) = \fint u^0 > 0$, $\fint [v(\cdot,t) + \frac{1}{\beta}\, u(\cdot,t)\, \lambda(\psi(\cdot,t))] = \fint [v^0 + \frac{1}{\beta}\, u^0\, \lambda(\psi^0)]$ *for a.a.* $t \in (0,T)$ *such that as* $h \to 0$

$$\text{(3.5a)}\quad U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{weak-}* \text{ in } L^\infty(0,T;H^1(\Omega)),$$

$$\text{(3.5b)}\quad V_\varepsilon,\ V_\varepsilon^\pm \to v \qquad \text{weak-}* \text{ in } L^\infty(0,T;L^2(\Omega)),\ \text{weakly in } L^2(0,T;H^1(\Omega)),$$

$$\text{(3.5c)}\quad \Psi_\varepsilon^+ \to \psi \qquad \text{weakly in } L^2(\Omega_T),$$

$$\text{(3.5d)}\quad \mathcal{G}\frac{\partial U_\varepsilon}{\partial t} \to \mathcal{G}\frac{\partial u}{\partial t} \quad \text{and} \quad \mathcal{G}\frac{\partial V_\varepsilon}{\partial t} \to \mathcal{G}\frac{\partial v}{\partial t} \qquad \text{weakly in } L^2(0,T;W^{1,q}(\Omega)),$$

$$\text{(3.6a)}\quad (U_\varepsilon)^\alpha, (U_\varepsilon^\pm)^\alpha \to u^\alpha \quad \text{for any } \alpha \in (0,\infty), \qquad \text{strongly in } L^2(0,T;L^s(\Omega)),$$

$$\text{(3.6b)}\quad V_\varepsilon,\ V_\varepsilon^\pm \to v \qquad \text{strongly in } L^2(0,T;L^p(\Omega)),$$

$$\text{(3.7a)}\quad [\Xi(U_\varepsilon^\pm)]^\alpha \to u^\alpha\, \mathcal{I} \quad \text{for any } \alpha \in (0,\infty), \qquad \text{strongly in } L^2(0,T;L^s(\Omega)),$$

$$\text{(3.7b)}\quad \Lambda_\varepsilon(V_\varepsilon^+) \to \lambda(v)\, \mathcal{I} \qquad \text{strongly in } L^2(0,T;L^p(\Omega)),$$

where $s \in [2,\infty]$ *and* $q = 2$ *if* $d = 1$, $s \in [2,\infty)$, *and* $q \in (1,2)$ *if* $d = 2$.

Furthermore, if $d = 1$, *or* $d = 2$ *and* $\nu \geq 7$ *in* (1.4), *then* u *in addition to* (3.4a) *satisfies*

$$u \in L^2(0,T;H^2(\Omega)), \tag{3.8}$$

and there exists a subsequence of $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}_h$ *satisfying* (3.5a)–(3.5d), (3.6a)–(3.6b), (3.7a)–(3.7b), *and as* $h \to 0$

$$\text{(3.9a)}\quad \Delta^h U_\varepsilon^+ \to \Delta u \qquad \text{weakly in } L^2(\Omega_T),$$

$$\text{(3.9b)}\quad U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{weakly in } L^2(0,T;W^{1,p}(\Omega)),$$

$$\text{(3.9c)}\quad U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{strongly in } L^2(0,T;C^{0,\gamma}(\overline{\Omega})) \quad \text{for any } \gamma \in \left(0, 1 - \tfrac{d}{p}\right),$$

and for a.a. $t \in (0,T)$

$$\text{(3.9d)}\quad u(\cdot,t) \in C^{0,\gamma}(\overline{\Omega}) \quad \text{with} \quad u(x,t) \geq \zeta(t) > 0 \qquad \forall\ x \in \overline{\Omega}.$$

On extracting a further subsequence, it also holds as $h \to 0$ *that for a.a.* $t \in (0,T)$

$$\text{(3.10a)}\quad \pi^h[\phi^\pm(U_\varepsilon^\pm)](\cdot,t) \to \phi^\pm(u(\cdot,t)) \qquad \text{strongly in } C(\overline{\Omega}),$$

$$\text{(3.10b)}\quad W_\varepsilon^+(\cdot,t) \to w(\cdot,t) \equiv -c\,\Delta u(\cdot,t) + \phi(u(\cdot,t)) \qquad \text{weakly in } H^1(\Omega),$$

$$\text{(3.10c)}\quad [\Xi(U_\varepsilon^+)]^{\frac{3}{2}}\, \nabla W_\varepsilon^+ \to u^{\frac{3}{2}}\, \nabla w \qquad \text{weakly in } L^2(\Omega_T).$$

Moreover, we have that

$$U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{strongly in } L^2(0,T;H^1(\Omega)). \tag{3.11}$$

Proof. Noting the definitions (3.1a)–(3.1b) and [6, (1.19)] , the bounds in (2.29a)–

(2.31b) together with (2.40) and our assumption (i) imply that

(3.12)

$$\begin{aligned}
&\|U_\varepsilon^{(\pm)}\|^2_{L^\infty(0,T;H^1(\Omega))} + \|V_\varepsilon^{(\pm)}\|^2_{L^\infty(0,T;L^2(\Omega))} + \rho_s \,\|V_\varepsilon^{(\pm)}\|^2_{L^2(0,T;H^1(\Omega))}\\
&\quad + \varepsilon^{-1}\|\pi^h[V_\varepsilon^+]_-\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\Psi_\varepsilon^+\|^2_{L^2(\Omega_T)} + \varepsilon^{-1}\|\pi^h[\Psi_\varepsilon^+]_-\|^2_{L^2(\Omega_T)}\\
&\quad + \left\|\bar\tau^{\frac12}\frac{\partial U_\varepsilon}{\partial t}\right\|^2_{L^2(0,T;H^1(\Omega))} + \left\|\bar\tau^{\frac12}\frac{\partial V_\varepsilon}{\partial t}\right\|^2_{L^2(\Omega_T)} + \|\,[\Xi(U_\varepsilon^+)]^{\frac32}\,\nabla W_\varepsilon^+\|^2_{L^2(\Omega_T)}\\
&\quad + \|(U_\varepsilon^+)^{\frac12}\,\nabla\Psi_\varepsilon^+\|^2_{L^2(\Omega_T)} + \left\|\mathcal{G}\frac{\partial U_\varepsilon}{\partial t}\right\|_{L^2(0,T;W^{1,q}(\Omega))} + \left\|\mathcal{G}\frac{\partial V_\varepsilon}{\partial t}\right\|_{L^2(0,T;W^{1,q}(\Omega))}\\
&\quad \le C.
\end{aligned}$$

Furthermore, we deduce from (3.2), (3.12), and (2.6) that

$$\|U_\varepsilon - U_\varepsilon^\pm\|^2_{L^2(0,T;H^1(\Omega))} \le \left\|\bar\tau\,\frac{\partial U_\varepsilon}{\partial t}\right\|^2_{L^2(0,T;H^1(\Omega))} \le C\,\tau, \tag{3.13a}$$

$$\|V_\varepsilon - V_\varepsilon^\pm\|^2_{L^2(0,T;L^p(\Omega))} \le C\,h^{-d\,(1-\frac2p)}\,\|V_\varepsilon - V_\varepsilon^\pm\|^2_{L^2(\Omega_T)} \le C\,h^{-d\,(1-\frac2p)}\,\tau. \tag{3.13b}$$

Hence on noting (3.12), (3.13a)–(3.13b), $U_\varepsilon > 0$, (1.6), assumption (iii), and a standard compact embedding result (see, e.g., [6, (1.20a)]), we can choose a subsequence $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}_h$ such that the convergence results (3.4a)–(3.4c), at first without the nonnegativity constraints on v and ψ, (3.5a)–(3.5d), and (3.6a)–(3.6b) for $\alpha = 1$ hold. Then (3.4a)–(3.4c) and Theorem 2.3 yield, on noting a standard compact embedding result (see, e.g., [6, (1.20b)]), assumption (i), (2.8), and (2.13), that the subsequence satisfies the additional initial and integral conditions.

The proof of the results (3.6a) for $\alpha \in (0,\infty)$, (3.7a)–(3.7b) can be found in the proof of Lemma 3.1 in BN. Furthermore, we note that Lemma 2.1 in BGN and (3.7b) imply that $\lambda(v) \ge 0$ a.e. $\Rightarrow v \ge 0$ a.e., and hence $H^1_{\ge0}(\Omega)$ in (3.4b). Moreover, on noting that $\|[\Psi_\varepsilon^+]_-\|_{L^2(\Omega_T)} \le \|\pi^h[\Psi_\varepsilon^+]_-\|_{L^2(\Omega_T)}$, the sixth bound in (3.12) shows that $[\Psi_\varepsilon^+]_- \to 0$ weakly in $L^2(\Omega_T)$, which implies that $[\Psi_\varepsilon^+]_+ \to \psi$ weakly in $L^2(\Omega_T)$. Hence $L^2_{\ge0}(\Omega)$ in (3.4c), and the results (3.4d) hold on noting (1.6).

The proof, using the key entropy estimate (2.32) in the case $d = 2$, of the results (3.8)–(3.9b), the result (3.9c) if $d = 1$, and the result on U_ε in (3.9c) if $d = 2$ can be found in the proof of Lemma 3.1 in BN. To prove the result on $U_\varepsilon^\pm$ in (3.9c) for the case $d = 2$, we note the following. For any $\gamma \in (0, 1-\frac2p)$ and any $\bar p \in (\frac{2}{1-\gamma}, p)$ it holds on noting the compact embedding $W^{1,\bar p}(\Omega) \overset{c}{\hookrightarrow} C^{0,\gamma}(\overline\Omega)$, (3.13a), and (3.9b) that

$$\begin{aligned}
\|U_\varepsilon - U_\varepsilon^\pm\|_{L^2(0,T;C^{0,\gamma}(\overline\Omega))} &\le \|U_\varepsilon - U_\varepsilon^\pm\|_{L^2(0,T;W^{1,\bar p}(\Omega))}\\
&\le \|U_\varepsilon - U_\varepsilon^\pm\|^{\mu}_{L^2(0,T;H^1(\Omega))}\,\|U_\varepsilon - U_\varepsilon^\pm\|^{1-\mu}_{L^2(0,T;W^{1,p}(\Omega))} \le C\,\tau^{\frac{\mu}{2}},
\end{aligned} \tag{3.14}$$

where $\mu = \frac{2\,(p-\bar p)}{(p-2)\,\bar p} \in (0,1)$. Combining (3.14), assumption (iii), and the established result on U_ε in (3.9c) yields the desired result (3.9c). Then the strong convergence result (3.9c) yields the remaining results (3.9d)–(3.10c); see the proof of Lemma 3.2 in BN for details.

Finally, we have that

$$\|\nabla(U_\varepsilon^+ - u)\|^2_{L^2(\Omega_T)} \le \left|\int_{\Omega_T} \nabla(U_\varepsilon^+ - u)\cdot\nabla u \,\mathrm{d}x\,\mathrm{d}t\right| + \left|\int_{\Omega_T} \nabla(U_\varepsilon^+ - \pi^h u)\cdot\nabla U_\varepsilon^+ \,\mathrm{d}x\,\mathrm{d}t\right| + \left|\int_{\Omega_T} \nabla(\pi^h u - u)\cdot\nabla U_\varepsilon^+ \,\mathrm{d}x\,\mathrm{d}t\right|, \tag{3.15a}$$

where, on noting (2.10),

$$\begin{aligned}\left|\int_{\Omega_T} \nabla(U_\varepsilon^+ - \pi^h u).\nabla U_\varepsilon^+ \,\mathrm{d}x\,\mathrm{d}t\right| &= \left| - \int_0^T (\Delta^h U_\varepsilon^+, U_\varepsilon^+ - \pi^h u)^h \,\mathrm{d}t\right| \\ &\le C\,\|\Delta^h U_\varepsilon^+\|_{L^2(\Omega_T)}\,\|U_\varepsilon^+ - \pi^h u\|_{L^2(\Omega_T)}.\end{aligned} \tag{3.15b}$$

Combining (3.15a)–(3.15b), (3.5a), (2.32), (2.7), (3.4a), (3.6a), and (3.13a) yields (3.11). ∎

Remark 3.2. We remark that in the case $d = 1$ one can prove stronger versions of (3.9c)–(3.10b); see the proof of Lemma 3.1 in BN for details. We note that in BN a further time step assumption was introduced for $d = 2$, in order to prove the results (3.9c)–(3.10c). However, the proof given here shows that this assumption is not necessary; see (3.14).

Remark 3.3. One can adapt the approximation ($\mathrm{P}_\varepsilon^{h,\tau}$) when there are no repulsive van der Waals forces ($\delta = 0$), by replacing Ξ with Ξ_ε (see Remark 3.2 in BN for details) and here in addition by replacing $U_\varepsilon^n\,\nabla\Psi_\varepsilon^n$ with $\pi^h[U_\varepsilon^n]_+\,\nabla\Psi_\varepsilon^n$ in (2.5d). Similarly, to the insoluble surfactant system studied in BN one can now no longer guarantee the nonnegativity of U_ε. However, in contrast to the system studied in BN, it is not clear that one can prove convergence in the case $d = 1$ by adapting the techniques in BGN.

LEMMA 3.4. *Let all the assumptions of Lemma* 3.1 *hold, and in addition assume that if* $d = 2$, *then* $p \in (2,6)$, $q \in [\frac{4p}{3p-2}, 2)$, *and* $\tau\, h^{-3d(\frac12-\frac1p)} \to 0$ *as* $h \to 0$. *Then there exists a subsequence of* $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}_h$ *such that as* $h \to 0$

$$\mathcal{G}\frac{\partial B_\varepsilon}{\partial t} \to \mathcal{G}\frac{\partial(u\,\lambda(\psi))}{\partial t} \quad \text{weakly in } L^2(0,T;W^{1,q}(\Omega)), \tag{3.16a}$$

$$\Xi[U_\varepsilon^+]\,\Lambda_\varepsilon(\Psi_\varepsilon^+) \to u\,\lambda(\psi)\,\mathcal{I} \quad \text{strongly in } L^2(0,T;L^p(\Omega)), \tag{3.16b}$$

$$(U_\varepsilon^+)^{\frac12}\,\nabla\Psi_\varepsilon^+ \to u^{\frac12}\,\nabla\psi \quad \text{weakly in } L^2(\Omega_T). \tag{3.16c}$$

Proof. Noting the definitions (3.1a)–(3.1b) and the bounds in (2.29a)–(2.31b) together with (2.40), (2.39) for $n = 0$, our assumption (i), (2.12), (2.13), and (2.8) imply that

$$\|B_\varepsilon^{(\pm)}\|^2_{L^2(0,T;W^{1,q}(\Omega))} + \left\|\mathcal{G}\frac{\partial B_\varepsilon}{\partial t}\right\|_{L^2(0,T;W^{1,q}(\Omega))} \le C \tag{3.17}$$

and hence, on noting a standard compact embedding result (see, e.g., [6, (1.20a)]), that

$$\begin{aligned} B_\varepsilon &\to b \quad \text{strongly in } L^2(0,T;L^s(\Omega)) \quad \text{as } h \to 0, \\ \mathcal{G}\frac{\partial B_\varepsilon}{\partial t} &\to \mathcal{G}\frac{\partial b}{\partial t} \quad \text{weakly in } L^2(0,T;W^{1,q}(\Omega)) \quad \text{as } h \to 0, \end{aligned} \tag{3.18}$$

where $b \in L^2(0,T;L^s(\Omega))$ still needs to be identified. Moreover, on noting (2.6), (1.25), (3.2), and (3.17), it holds for $q = p = 2$ if $d = 1$ and $q \in [\frac{4p}{3p-2}, 2)$ if $d = 2$ that

$$\begin{aligned}
\|B_\varepsilon - B_\varepsilon^\pm\|^2_{L^2(0,T;L^p(\Omega))} &\leq C\,h^{-2d(\frac12-\frac1p)}\,\|B_\varepsilon - B_\varepsilon^\pm\|^2_{L^2(0,T;L^2(\Omega))}\\
&\leq C\,\tau\,h^{-2d(\frac12-\frac1p)}\left\|\mathcal{G}\frac{\partial B_\varepsilon}{\partial t}\right\|_{L^2(0,T;W^{1,q}(\Omega))}\|B_\varepsilon^\pm\|_{L^2(0,T;W^{1,q'}(\Omega))}\\
&\leq C\,\tau\,h^{-2d(\frac12-\frac1p)}\,h^{-d(\frac1q-\frac17)}\|B_\varepsilon^\pm\|_{L^2(0,T;W^{1,q}(\Omega))}\\
&\leq C\,\tau\,h^{-2d(\frac12-\frac1p)}\,h^{-d(\frac{2-q}{q})} \leq C\,\tau\,h^{-3d(\frac12-\frac1p)} \to 0 \quad \text{as } h\to 0. \qquad (3.19)
\end{aligned}$$

In addition, on noting (2.8), (1.8), and (2.6), it holds for $n = 0 \to N$ that

$$\begin{aligned}
|B_\varepsilon^n - U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)|^2_{0,p} &\leq \sum_{\kappa\in\mathcal{T}^h}[\underline{m}(\kappa)]^{\frac2p}\,|(I-\pi^h)[U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)]|^2_{0,\infty,\kappa}\\
&\leq C\sum_{\kappa\in\mathcal{T}^h}[\underline{m}(\kappa)]^{\frac2p}h_\kappa^2\,|U_\varepsilon^n\,\Psi_\varepsilon^n|^2_{1,\infty,\kappa} \leq C\sum_{\kappa\in\mathcal{T}^h}h_\kappa^2\,|U_\varepsilon^n\,\Psi_\varepsilon^n|^2_{1,p,\kappa}\\
&\leq C\sum_{\kappa\in\mathcal{T}^h}h_\kappa^2\,h_\kappa^{-2d(\frac1q-\frac1p)}|U_\varepsilon^n\,\Psi_\varepsilon^n|^2_{1,q,\kappa} \leq C\,h^{2-2d(\frac1q-\frac1p)}\,|U_\varepsilon^n\,\Psi_\varepsilon^n|^2_{1,q}\\
&\leq C\,h^{2-2d(\frac1q-\frac1p)}\,\big[|U_\varepsilon^n|_1^2 + |(U_\varepsilon^n)^{\frac12}|^2_{0,r}\,|(U_\varepsilon^n)^{\frac12}\,\nabla\Psi_\varepsilon^n|_0^2\big], \qquad (3.20)
\end{aligned}$$

where $r = \infty$ if $d = 1$ and $r = \frac{2q}{2-q}$ if $d = 2$. Therefore it follows from (3.20), (3.12), and our assumptions on p and q that $\|B_\varepsilon^+ - U_\varepsilon^+\,\lambda_\varepsilon(\Psi_\varepsilon^+)\|_{L^2(0,T;L^p(\Omega))} \to 0$ and hence, on noting (3.18) and (3.19), that

$$U_\varepsilon^+\,\lambda_\varepsilon(\Psi_\varepsilon^+) \to b \qquad \text{strongly in } L^2(0,T;L^p(\Omega)) \quad \text{as } h\to 0. \qquad (3.21)$$

Combining (3.21), (3.6a), (3.9d), (3.12), and (1.8) yields, on possibly extracting a further subsequence, that $\lambda_\varepsilon(\Psi_\varepsilon^+) \to u^{-1}b$ strongly in $L^2(0,T;L^p(\Omega))$ as $h\to 0$ and in particular that $u^{-1}b \in [0,1]$ a.e. in Ω_T. Moreover, it follows from (1.6), (1.8), (3.12), and assumption (iii) that

$$\|\lambda(\Psi_\varepsilon^+) - \lambda_\varepsilon(\Psi_\varepsilon^+)\|_{L^2(0,T;L^2(\Omega))} \leq C\,\varepsilon + \|[\Psi_\varepsilon^+]_-\|_{L^2(0,T;L^2(\Omega))} \leq C\,\varepsilon^{\frac12} \to 0 \quad \text{as } h\to 0.$$

Hence

$$\lambda(\Psi_\varepsilon^+) \to u^{-1}b \qquad \text{strongly in } L^2(0,T;L^2(\Omega)) \quad \text{as } h\to 0. \qquad (3.22)$$

Combining (3.22), (1.6), and (3.5c) yields that $u^{-1}b = \psi = \lambda(\psi)$ a.e. where $u^{-1}b < 1$. It remains to identify b, where $u^{-1}b = 1$. Let $\mathcal{A} := \{(x,t) \in \Omega_T : (u^{-1}b)(x,t) = 1,\ \psi(x,t) < 1\}$. On assuming that $\underline{m}(\mathcal{A}) > 0$, it follows from (3.5c), (1.6), and (3.22) that

$$\begin{aligned}
\underline{m}(\mathcal{A}) &= \int_{\Omega_T}\mathcal{H}_\mathcal{A}\,\mathrm{d}x\,\mathrm{d}t > \int_{\Omega_T}\psi\,\mathcal{H}_\mathcal{A}\,\mathrm{d}x\,\mathrm{d}t \leftarrow \int_{\Omega_T}\Psi_\varepsilon^+\,\mathcal{H}_\mathcal{A}\,\mathrm{d}x\,\mathrm{d}t\\
&\geq \int_{\Omega_T}\lambda(\Psi_\varepsilon^+)\,\mathcal{H}_\mathcal{A}\,\mathrm{d}x\,\mathrm{d}t \to \int_{\Omega_T}1\,\mathcal{H}_\mathcal{A}\,\mathrm{d}x\,\mathrm{d}t = \underline{m}(\mathcal{A}),
\end{aligned}$$

where $\mathcal{H}_\mathcal{A}$ is the characteristic function of $\mathcal{A}$. This is a contradiction and hence $\underline{m}(\mathcal{A}) = 0$. This means that $\psi \geq 1$ a.e. where $u^{-1}b = 1$, i.e., $u^{-1}b = 1 = \lambda(\psi)$ a.e.

where $u^{-1}b = 1$. Combining this with the earlier result on b yields that $b = u\,\lambda(\psi)$ a.e. in Ω_T. This proves, on recalling (3.18), that $B_\varepsilon \to u\,\lambda(\psi)$ strongly in $L^2(0,T;L^s(\Omega))$ as $h \to 0$, and that (3.16a) holds.

Similarly to (3.20), we have on noting Lemma 2.1 in BGN, Lemma 2.3 in BN, and (3.7a) that

$$\begin{aligned}
&\|U_\varepsilon^+\,\lambda_\varepsilon(\Psi_\varepsilon^+)\,\mathcal{I} - \Xi(U_\varepsilon^+)\,\Lambda_\varepsilon(\Psi_\varepsilon^+)\|^2_{L^2(0,T;L^p(\Omega))}\\
&\quad\le \|U_\varepsilon^+\,[\lambda_\varepsilon(\Psi_\varepsilon^+)\,\mathcal{I} - \Lambda_\varepsilon(\Psi_\varepsilon^+)]\|^2_{L^2(0,T;L^p(\Omega))} + \|\,[U_\varepsilon^+\,\mathcal{I} - \Xi(U_\varepsilon^+)]\,\Lambda_\varepsilon(\Psi_\varepsilon^+)\|^2_{L^2(0,T;L^p(\Omega))}\\
&\quad\le \int_0^T \sum_{\kappa\in\mathcal{T}^h} h_\kappa^2\,|U_\varepsilon^+\,\nabla\Psi_\varepsilon^+|^2_{0,p,\kappa}\,\mathrm{d}t + \|U_\varepsilon^+\,\mathcal{I} - \Xi(U_\varepsilon^+)\|^2_{L^2(0,T;L^p(\Omega))} \to 0 \quad \text{as } h\to 0.
\end{aligned}$$

Combining this and (3.21) yields the desired result (3.16b).

Finally, it follows from (3.12) that $(U_\varepsilon^+)^{\frac12}\,\nabla\Psi_\varepsilon^+ \to z$ weakly in $L^2(\Omega_T)$, where $z \in L^2(\Omega_T)$. But for any $\eta \in C_0^\infty(\overline{\Omega}_T)$, which is dense in $L^2(\Omega_T)$, we have, on recalling (3.6a), (3.5c), and (3.11), that

$$\begin{aligned}
\int_0^T (u^{\frac12}\,z,\eta)\,\mathrm{d}t &\leftarrow \int_0^T (U_\varepsilon^+\,\nabla\Psi_\varepsilon^+,\eta)\,\mathrm{d}t = -\int_0^T \left[(\Psi_\varepsilon^+\,\nabla U_\varepsilon^+,\eta) + (\Psi_\varepsilon^+\,U_\varepsilon^+,\nabla\cdot\eta)\right]\,\mathrm{d}t\\
&\to -\int_0^T (\psi\,\nabla u,\eta)\,\mathrm{d}t - \int_0^T (\psi\,u,\nabla\cdot\eta)\,\mathrm{d}t = \int_0^T (u\,\nabla\psi,\eta)\,\mathrm{d}t. \qquad (3.23)
\end{aligned}$$

Hence $z = u^{\frac12}\,\nabla\psi$ in $L^2(\Omega_T)$ and (3.16c) holds. ∎

THEOREM 3.5. *Let all the assumptions of Lemma* 3.4 *hold. Then there exist a subsequence of* $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}_h$, *where* $\{U_\varepsilon, W_\varepsilon^+, V_\varepsilon, \Psi_\varepsilon\}$ *solve* $(\mathrm{P}_\varepsilon^{h,\tau})$, *and functions* $\{u,w,v,\psi\}$ *satisfying* (3.4a)–(3.4d), (3.8), *and* (3.9d). *In addition, as* $h \to 0$ *the following hold:* (3.5a)–(3.5d), (3.6a)–(3.6b), (3.7a)–(3.7b), (3.9a)–(3.9c), (3.10a)–(3.10b) *for a.a.* $t \in (0,T)$, (3.10c), (3.11), *and* (3.16a)–(3.16c). *Moreover, we have that* u *and* v *fulfill* $u(\cdot,0) = u^0(\cdot)$ *in* Y_1, $v(\cdot,0) = v^0(\cdot)$ *in* Y_2, *where* $H^1(\Omega) \overset{c}{\hookrightarrow} Y_1$, $L^2(\Omega) \overset{c}{\hookrightarrow} Y_2$. *Furthermore,* $\{u,w,v,\psi\}$ *satisfy for all* $\eta \in L^2(0,T;W^{1,q'}(\Omega))$, *with* $q' = 2$ *if* $d = 1$ *and* $q' \in (2, \frac{4p}{p+2}]$, *where* $p \in (2,6)$, *if* $d = 2$,

$$\int_0^T \left\langle \frac{\partial u}{\partial t},\eta\right\rangle_{q'}\mathrm{d}t + \int_{\Omega_T}\left[\frac13\,u^3\,\nabla w\cdot\nabla\eta + \frac12\,u^2\,\nabla v\cdot\nabla\eta\right]\mathrm{d}x\,\mathrm{d}t = 0, \qquad (3.24a)$$

$$\begin{aligned}
&\int_0^T \left\langle \frac{\partial v}{\partial t},\eta\right\rangle_{q'}\mathrm{d}t + \int_{\Omega_T}[\,\rho_s\,\nabla v\cdot\nabla\eta + u\,\lambda(v)\,\nabla v\cdot\nabla\eta\,]\,\mathrm{d}x\,\mathrm{d}t\\
&\qquad + \int_{\Omega_T}\left[\frac12\,u^2\,\lambda(v)\,\nabla w\cdot\nabla\eta - K\,(\psi - v)\,\eta\right]\mathrm{d}x\,\mathrm{d}t = 0, \qquad (3.24b)
\end{aligned}$$

$$\begin{aligned}
&\int_0^T \left\langle \frac{\partial(u\,\lambda(\psi))}{\partial t},\eta\right\rangle_{q'}\mathrm{d}t + \int_{\Omega_T}\left[\rho_b\,u\,\nabla\psi\cdot\nabla\eta + \frac13\,u^3\,\lambda(\psi)\,\nabla w\cdot\nabla\eta\right]\mathrm{d}x\,\mathrm{d}t\\
&\qquad + \int_{\Omega_T}\left[\frac12\,u^2\,\lambda(\psi)\,\nabla v\cdot\nabla\eta + \beta\,K\,(\psi - v)\,\eta\right]\mathrm{d}x\,\mathrm{d}t = 0, \qquad (3.24c)
\end{aligned}$$

where for a.a. $t \in (0,T)$

$$\int_\Omega [\,w(\cdot,t)\,\xi - c\,\nabla u(\cdot,t).\,\nabla\xi - \phi(u(\cdot,t))\,\xi\,]\,\mathrm{d}x = 0 \qquad \forall\ \xi \in H^1(\Omega). \qquad (3.24d)$$

Proof. On choosing $z^h \equiv \pi^h \tilde{\xi}$, where $\tilde{\xi} \in W^{1,p}(\Omega)$, in (3.3d); it follows from (2.1), (2.10), (2.7), (3.5a), and (3.10a)–(3.10b) that (3.24d) holds for $\xi \equiv \tilde{\xi}$. The desired result (3.24d) then holds for any $\xi \in H^1(\Omega)$ via a density argument.

For any $\eta \in L^2(0,T;W^{1,q'}(\Omega))$ and $\widetilde{\eta} \in H^1(0,T;W^{1,\infty}(\Omega))$, we choose $\chi \equiv \pi^h \eta$ in (3.3a)–(3.3c) and then analyze the subsequent terms. The results (3.24a) and (3.24b), for the case $K = 0$, are then derived from (3.3a) and (3.3b), with $K = 0$; their proof can be found in the proof of Theorem 3.1 in BN. Hence it is sufficient to prove (3.24c), as the convergence of the term involving K in (3.24b) then follows from the convergence of the corresponding term in (3.24c).

First, (2.11), (2.15), an interpolation estimate in time (see [6, (1.19)]), (3.12), (3.17), and (2.8), on noting that $\|B_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq \|U_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))}$, yield that

(3.25)
$$
\begin{aligned}
&\left| \int_0^T \left[\left(\frac{\partial B_\varepsilon}{\partial t}, \pi^h \eta \right)^h - \left(\frac{\partial B_\varepsilon}{\partial t}, \pi^h \eta \right) \right] \mathrm{d}t \right| \\
&\leq \left| \int_0^T \left[\left(\frac{\partial B_\varepsilon}{\partial t}, \pi^h [\eta - \widetilde{\eta}] \right)^h - \left(\frac{\partial B_\varepsilon}{\partial t}, \pi^h [\eta - \widetilde{\eta}] \right) \right] \mathrm{d}t \right| \\
&\quad + \left| - \int_0^T \left(B_\varepsilon, \frac{\partial (\pi^h \widetilde{\eta})}{\partial t} \right)^h \mathrm{d}t + (B_\varepsilon(\cdot,T), \pi^h \widetilde{\eta}(\cdot,T))^h - (B_\varepsilon(\cdot,0), \pi^h \widetilde{\eta}(\cdot,0))^h \right. \\
&\qquad \left. + \int_0^T \left(B_\varepsilon, \frac{\partial (\pi^h \widetilde{\eta})}{\partial t} \right) \mathrm{d}t - (B_\varepsilon(\cdot,T), \pi^h \widetilde{\eta}(\cdot,T)) + (B_\varepsilon(\cdot,0), \pi^h \widetilde{\eta}(\cdot,0)) \right| \\
&\leq C \left\| \mathcal{G} \frac{\partial B_\varepsilon}{\partial t} \right\|_{L^2(0,T;W^{1,q}(\Omega))} \|\pi^h [\eta - \widetilde{\eta}]\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\qquad + C\, h\, \|B_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))}\, \|\pi^h \widetilde{\eta}\|_{H^1(0,T;H^1(\Omega))} \\
&\leq C\, \|\eta - \widetilde{\eta}\|_{L^2(0,T;W^{1,q'}(\Omega))} + C\, h\, \|\widetilde{\eta}\|_{H^1(0,T;W^{1,q'}(\Omega))}.
\end{aligned}
$$

Furthermore, it follows from (1.25) and (3.17) that

$$
\begin{aligned}
\left| \int_0^T \left(\frac{\partial B_\varepsilon}{\partial t}, (I - \pi^h)\eta \right) \mathrm{d}t \right| &\leq C \left\| \mathcal{G} \frac{\partial B_\varepsilon}{\partial t} \right\|_{L^2(0,T;W^{1,q}(\Omega))} \|(I - \pi^h)\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\leq C\, \|(I - \pi^h)\eta\|_{L^2(0,T;W^{1,q'}(\Omega))}. \qquad (3.26)
\end{aligned}
$$

Combining (3.25), the denseness of $H^1(0,T;W^{1,\infty}(\Omega))$ in $L^2(0,T;W^{1,q'}(\Omega))$, (3.26), (2.7), (1.25), and (3.16a) yields for all $\eta \in L^2(0,T;W^{1,q'}(\Omega))$ that

$$
\int_0^T \left(\frac{\partial B_\varepsilon}{\partial t}, \pi^h \eta \right)^h \mathrm{d}t \to \int_0^T \left\langle \frac{\partial (u\, \lambda(\psi))}{\partial t}, \eta \right\rangle_{q'} \mathrm{d}t \quad \text{as } h \to 0. \qquad (3.27)
$$

Similarly to the above, it follows from (2.1), (2.11), (2.8), (3.5b), and (3.5c) that for all $\eta \in L^2(0,T;W^{1,q'}(\Omega))$

$$
\int_0^T \left(\Psi_\varepsilon^+ - V_\varepsilon^+, \pi^h \eta \right)^h \mathrm{d}t \to \int_0^T (\psi - v, \eta)\, \mathrm{d}t \quad \text{as } h \to 0. \qquad (3.28)
$$

In view of (3.12), (3.4a), [6, (1.19)], and (3.16c), we deduce with r as defined in

(2.31a) that for all $\eta \in L^2(0,T;W^{1,q'}(\Omega))$ and for all $\widetilde{\eta} \in H^1(0,T;W^{1,\infty}(\Omega))$

$$\begin{aligned}
&\left|\int_0^T (U_\varepsilon^+ \nabla \Psi_\varepsilon^+, \nabla[\widetilde{\eta} - \pi^h \eta]\,)\,\mathrm{d}t\right| + \left|\int_0^T (u\,\nabla\psi, \nabla[\widetilde{\eta} - \eta]\,)\,\mathrm{d}t\right| \\
&\quad \le \|(U_\varepsilon^+)^{\frac12}\|_{L^\infty(0,T;L^r(\Omega))}\, \|(U_\varepsilon^+)^{\frac12}\,\nabla \Psi_\varepsilon^+\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \pi^h\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\qquad + \|u^{\frac12}\|_{L^\infty(0,T;L^r(\Omega))}\, \|u^{\frac12}\,\nabla\psi\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\quad \le C\,\big[\|(I-\pi^h)\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} + \|\eta - \widetilde{\eta}\|_{L^2(0,T;W^{1,q'}(\Omega))}\big]. \qquad (3.29)
\end{aligned}$$

Similarly to the above, on noting in addition (2.31b), Lemma 2.1 in BGN, (3.10c), (3.4d), and (3.4b), we deduce for all $\eta \in L^2(0,T;W^{1,q'}(\Omega))$ and for all $\widetilde{\eta} \in H^1(0,T;W^{1,\infty}(\Omega))$

$$\begin{aligned}
&\left|\int_0^T ([\Xi(U_\varepsilon^+)]^3\,\Lambda_\varepsilon(\Psi_\varepsilon^+)\,\nabla W_\varepsilon^+, \nabla[\widetilde{\eta} - \pi^h\eta]\,)\,\mathrm{d}t\right| + \left|\int_0^T (u^3\,\lambda(\psi)\,\nabla w, \nabla[\widetilde{\eta} - \eta]\,)\,\mathrm{d}t\right| \\
&\quad \le \|[\Xi(U_\varepsilon^+)]^{\frac32}\|_{L^\infty(0,T;L^r(\Omega))}\, \|[\Xi(U_\varepsilon^+)]^{\frac32}\,\nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \pi^h\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\qquad + \|\lambda(\psi)\|_{L^\infty(\Omega_T)}\, \|u^{\frac32}\|_{L^\infty(0,T;L^r(\Omega))}\, \|u^{\frac32}\,\nabla w\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\quad \le C\,\big[\|(I-\pi^h)\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} + \|\eta - \widetilde{\eta}\|_{L^2(0,T;W^{1,q'}(\Omega))}\big]
\end{aligned}$$

and

$$\begin{aligned}
&\left|\int_0^T ([\Xi(U_\varepsilon^+)]^{\frac32}\,[\Xi(U_\varepsilon^-)]^{\frac12}\,\Lambda_\varepsilon(\Psi_\varepsilon^+)\,\nabla V_\varepsilon^-, \nabla[\widetilde{\eta} - \pi^h\eta]\,)\,\mathrm{d}t\right| \\
&\qquad + \left|\int_0^T (u^2\,\lambda(\psi)\,\nabla v, \nabla[\widetilde{\eta} - \eta]\,)\,\mathrm{d}t\right| \\
&\quad \le \|[\Xi(U_\varepsilon^+)]^{\frac32}\,[\Xi(U_\varepsilon^-)]^{\frac12}\|_{L^\infty(0,T;L^r(\Omega))}\, \|\nabla V_\varepsilon^-\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \pi^h\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\qquad + \|\lambda(\psi)\|_{L^\infty(\Omega_T))}\, \|u^2\|_{L^\infty(0,T;L^r(\Omega))}\, \|\nabla v\|_{L^2(\Omega_T)}\, \|\widetilde{\eta} - \eta\|_{L^2(0,T;W^{1,q'}(\Omega))} \\
&\quad \le C\,\big[\|(I-\pi^h)\eta\|_{L^2(0,T;W^{1,q'}(\Omega))} + \|\eta - \widetilde{\eta}\|_{L^2(0,T;W^{1,q'}(\Omega))}\big]. \qquad (3.30)
\end{aligned}$$

For all $\widetilde{\eta} \in H^1(0,T;W^{1,\infty}(\Omega))$, it follows from (3.6a), (3.16c), (3.7a), (3.16b), (3.10c), (3.4a), (3.4d), [6, (1.19)], (2.31b), (3.1b), (3.12), and (3.5b) that as $h \to 0$

$$\int_{\Omega_T} (U_\varepsilon^+)^{\frac12}\,[(U_\varepsilon^+)^{\frac12}\,\nabla\Psi_\varepsilon^+]\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t \to \int_{\Omega_T} u^{\frac12}\,[u^{\frac12}\,\nabla\psi]\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t, \qquad (3.31a)$$

$$\begin{aligned}
&\int_{\Omega_T} [\Xi(U_\varepsilon^+)]^{\frac12}\,[\Xi(U_\varepsilon^+)\,\Lambda_\varepsilon(\Psi_\varepsilon^+)]\,[\,[\Xi(U_\varepsilon^+)]^{\frac32}\,\nabla W_\varepsilon^+]\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t \\
&\qquad \to \int_{\Omega_T} u^{\frac12}\,[u\,\lambda(\psi)]\,[u^{\frac32}\,\nabla w]\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t, \qquad (3.31b)
\end{aligned}$$

$$\begin{aligned}
&\int_{\Omega} [\Xi(U_\varepsilon^+)]^{\frac12}\,[\Xi(U_\varepsilon^-)]^{\frac12}\,[\,[\Xi(U_\varepsilon^+)]\,\Lambda_\varepsilon(\Psi_\varepsilon^+)]\,\nabla V_\varepsilon^-\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t \\
&\qquad \to \int_{\Omega_T} u\,[u\,\lambda(\psi)]\,\nabla v\,\nabla\widetilde{\eta}\,\mathrm{d}x\,\mathrm{d}t. \qquad (3.31c)
\end{aligned}$$

Noting that $H^1(0,T;W^{1,\infty}(\Omega))$ is dense in $L^2(0,T;W^{1,q'}(\Omega))$ and (2.7), we obtain the desired result (3.24c) on combining (3.3c), (3.27)–(3.30), and (3.31a)–(3.31c). Hence $\{u,\, w,\, v,\, \psi\}$ satisfy (3.24a)–(3.24d) as well as the stated results of Lemmas 3.1 and 3.4. ∎

4. Numerical results. Before presenting some numerical results in both one and two space dimensions, we briefly state algorithms for solving the resulting system of algebraic equations for $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\}$ arising at each time level from the approximation ($\mathrm{P}_\varepsilon^{h,\tau}$). As (2.5a)–(2.5b) in ($\mathrm{P}_\varepsilon^{h,\tau}$) are independent of $\{V_\varepsilon^n, \Psi_\varepsilon^n\}$, we first solve these to obtain $\{U_\varepsilon^n, W_\varepsilon^n\}$; then we solve (2.5c)–(2.5d) for $\{V_\varepsilon^n, \Psi_\varepsilon^n\}$. We use the following iterative approach to solve (2.5a)–(2.5b) for $\{U_\varepsilon^n, W_\varepsilon^n\}$: Given $U_\varepsilon^{n,0} \in S^h_{>0}$, for $k \geq 1$ find $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}\} \in [S^h]^2$ such that for all $\chi \in S^h$

$$\left(\frac{U_\varepsilon^{n,k} - U_\varepsilon^{n-1}}{\tau_n}, \chi\right)^h + \frac{1}{3}\left([\Xi(U_\varepsilon^{n,k-1})]^3\, \nabla W_\varepsilon^{n,k}, \nabla\chi\right)$$

$$= -\frac{1}{2}\left([\Xi(U_\varepsilon^{n,k-1})]^{\frac{3}{2}}\, [\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\, \nabla V_\varepsilon^{n-1}, \nabla\chi\right), \tag{4.1a}$$

$$c\,(\nabla U_\varepsilon^{n,k}, \nabla\chi) + (\phi^+(U_\varepsilon^{n,k}) + \phi^-(U_\varepsilon^{n-1}), \chi)^h = (W_\varepsilon^{n,k}, \chi)^h. \tag{4.1b}$$

Then, having obtained $\{U_\varepsilon^n, W_\varepsilon^n\}$, we solve (2.5c)–(2.5d) for $\{V_\varepsilon^n,\ \Psi_\varepsilon^n\}$ using the following iterative approach: Given $\{V_\varepsilon^{n,0}, \Psi_\varepsilon^{n,0}\} \in [S^h]^2$, for $k \geq 1$ find $\{V_\varepsilon^{n,k}, \Psi_\varepsilon^{n,k}\} \in [S^h]^2$ such that for all $\chi \in S^h$

$$\left(\frac{V_\varepsilon^{n,k} - V_\varepsilon^{n-1}}{\tau_n}, \chi\right)^h + \rho_s\,(\nabla V_\varepsilon^{n,k}, \nabla\chi) + (\Xi(U_\varepsilon^n)\,\Lambda_\varepsilon(V_\varepsilon^{n,k-1})\,\nabla V_\varepsilon^{n,k}, \nabla\chi)$$

$$- K\left(\Psi_\varepsilon^{n,k-1} - V_\varepsilon^{n,k}, \chi\right)^h = -\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^2\,\Lambda_\varepsilon(V_\varepsilon^{n,k-1})\,\nabla W_\varepsilon^n, \nabla\chi\right), \tag{4.2a}$$

$$\left(\frac{U_\varepsilon^n \lambda_\varepsilon(\Psi_\varepsilon^{n,k}) - U_\varepsilon^{n-1}\lambda_\varepsilon(\Psi_\varepsilon^{n-1})}{\tau_n}, \chi\right)^h + \rho_b\left(U_\varepsilon^n\,\nabla\Psi_\varepsilon^{n,k}, \nabla\chi\right)$$

$$+ \frac{1}{3}\left([\Xi(U_\varepsilon^n)]^3\,\Lambda_\varepsilon(\Psi_\varepsilon^{n,k-1})\,\nabla W_\varepsilon^n, \nabla\chi\right) + \beta K\left(\Psi_\varepsilon^{n,k} - V_\varepsilon^{n,k}, \chi\right)^h$$

$$= -\frac{1}{2}\left([\Xi(U_\varepsilon^n)]^{\frac{3}{2}}\,[\Xi(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\Lambda_\varepsilon(\Psi_\varepsilon^{n,k-1})\,\nabla V_\varepsilon^{n-1}, \nabla\chi\right). \tag{4.2b}$$

Equations (4.1a)–(4.1b) and (4.2a)–(4.2b) are natural extensions of the iterative procedure proposed in [10] for solving a finite element approximation of the thin film equation. Note that we have chosen the iterative method such that (4.2a) and (4.2b) decouple. As $U_\varepsilon^{n,k-1} > 0$, it is easily established on noting Lemma 2.2 in BN that there exists a unique solution $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}\} \in S^h_{>0} \times S^h$ to (4.1a)–(4.1b). As (4.2a) is linear, existence of $V_\varepsilon^{n,k}$ follows from uniqueness; this is easily established on noting (2.2a) and $\rho_s \geq 0$. Existence and uniqueness of $\Psi_\varepsilon^{n,k}$ follow from the monotonicity of λ_ε and the positivity of $U_\varepsilon^n > 0$. Hence the iterations (4.1a)–(4.1b) and (4.2a)–(4.2b) are well defined.

For the iterative algorithms (4.1a)–(4.1b) and (4.2a)–(4.2b) we set, for $n \geq 1$, $\{U_\varepsilon^{n,0}, V_\varepsilon^{n,0}, \Psi_\varepsilon^{n,0}\} \equiv \{U_\varepsilon^{n-1}, V_\varepsilon^{n-1}, \Psi_\varepsilon^{n-1}\}$ and adopted the stopping criteria

$$|U_\varepsilon^{n,k} - U_\varepsilon^{n,k-1}|_{0,\infty} < tol,\ |V_\varepsilon^{n,k} - V_\varepsilon^{n,k-1}|_{0,\infty} < tol, \quad \text{and} \quad |\Psi_\varepsilon^{n,k} - \Psi_\varepsilon^{n,k-1}|_{0,\infty} < tol,$$

respectively, with $tol = 10^{-8}$. Furthermore, we then set $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n, \Psi_\varepsilon^n\} \equiv \{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}, V_\varepsilon^{n,k}, \Psi_\varepsilon^{n,k}\}$ for (4.1a)–(4.1b) and (4.2a)–(4.2b).

Remark 4.1. The nonlinear system (4.1a)–(4.1b) can be solved using an inexact Newton's method, applying a BiCGSTAB algorithm at each Newton iteration. The linear system (4.2a), on the other hand, can be solved efficiently using a conjugate

gradient algorithm. The nonlinear system (4.2b) can be solved with a nonlinear SOR method, similar to the one employed in [7]. In particular, in each step and at each vertex a scalar nonlinear equation of the form $\alpha_1\,\lambda_\varepsilon(s) + \alpha_2\,s = \alpha_3$, $s \in \mathbb{R}$, where α_1, $\alpha_2 > 0$, has to be solved, which is straightforward as λ_ε is monotone.

Although we are unable to show convergence of the iterations (4.1a)–(4.1b) and (4.2a)–(4.2b) for $\{U_\varepsilon^n, W_\varepsilon^n\}$ and $\{V_\varepsilon^n, \Psi_\varepsilon^n\}$, respectively, we observed good convergence properties in practice.

4.1. Numerical results for $d = 1$. First, we present numerical experiments in one space dimension. Throughout we choose a uniform partitioning of $\Omega = (-L, L)$, where $L \geq 1$, with mesh points $p_j = -L + (j-1)h$, $j = 1 \to 2^{10}+1$, where $h = 2^{-9}\,L$. In addition we choose uniform time steps $\tau_n = \tau = 10^{-3}$ and throughout set the regularization parameter $\varepsilon = 10^{-5}$. For the initial profiles u^0, v^0, and ψ^0, we set

$$u^0(x) = 1 \quad \text{and} \quad v^0(x) = \frac{1}{2}\left[1 - \tanh(10\,|x| - 5)\right], \quad \psi^0(x) = 0, \tag{4.3}$$

which resembles a uniform liquid film of unit height with surfactant on top of it, and the film is uncontaminated by the chemical. In the absence of both surfactant and chemical, a uniform film is a steady state. We choose $U_\varepsilon^0 \equiv \pi^h u^0$, $V_\varepsilon^0 \equiv \pi^h v^0$, and $\Psi_\varepsilon^0 \equiv \pi^h \psi^0$ as the discrete initial data on noting that u^0, v^0, $\psi^0 \in W^{1,\infty}(\Omega)$.

We now report on the evolutions of U_ε, V_ε, and Ψ_ε for similar parameters as in some of the experiments in [16, Fig. 5]. We set $L = 10$, $c = 10^{-3}$, $\rho_s = 10^{-5}$, $\rho_b = 10^{-2}$, $K = 1$, $a = 0$, $\delta = 10^{-5}$, and $\nu = 4$ and used the initial data (4.3). We then varied the solubility parameter, by choosing $\beta = 0.01, 1$, or 100. The different evolution results can be seen in Figure 1, where we plot U_ε and V_ε both at time $t = 5$ and at different final times T. Note that for brevity only V_ε is displayed since in all cases after a short time Ψ_ε is graphically indistinguishable from it. This is to be expected from the Lyapunov structure (2.29a) (see also (1.20)) on noting (1.11) and (1.8).

One can clearly see the effect of the parameter β on the evolution. On the one hand, the larger the value of β the faster Ψ_ε attains the profile of V_ε. On the other hand, since the quantity $\fint(V_\varepsilon^n + \frac{1}{\beta}\,\pi^h[U_\varepsilon^n\,\lambda_\varepsilon(\Psi_\varepsilon^n)]) = \fint(V_\varepsilon^0 + \frac{\varepsilon}{\beta})$ is preserved, the value of β dictates how much surfactant material V_ε remains on the film surface. In particular, after a sufficiently long time it holds that $\fint V_\varepsilon^n \approx \fint \Psi_\varepsilon^n$ and $\fint(V_\varepsilon^n + \frac{1}{\beta}\,\pi^h[U_\varepsilon^n\,V_\varepsilon^n]) \approx \fint V_\varepsilon^0$. Hence if β is such that $\varepsilon \ll \beta \ll 1$, the original drop of surfactant almost completely disappears, leading to a comparatively small change in the liquid film height that quickly smooths out. In the case of a very large β, recall that $\beta \to \infty$ models insoluble surfactant spreading; the initial amount of surfactant is almost completely preserved, leading to a fast propagating wave front.

Finally, note that the presence of repulsive van der Waals forces ($\delta > 0$) has no effect on the evolution in this case, as the film height is always bounded well away from zero.

When attractive van der Waals forces are included, however, this has a marked effect on the film evolution. We repeated the above experiments for a value of $a = 5 \times 10^{-4}$, and the results can be seen in Figure 2. Note that for $\beta = 0.01$ and 1 the film has thinned considerably in some areas due to the presence of attractive van der Waals forces, although it can never actually rupture ($U_\varepsilon = 0$) due to the repulsive van der Waals forces, ϕ^+. In fact, it holds that $\min_{x\in\overline{\Omega}} U_\varepsilon(x,T) = \arg\min_{s\in\mathbb{R}_{>0}} \Phi(s) = (\frac{\delta}{a})^{\frac{1}{\nu-3}} = 0.02$. For $\beta = 100$, on the other hand, the film has not yet completely thinned at the displayed time.

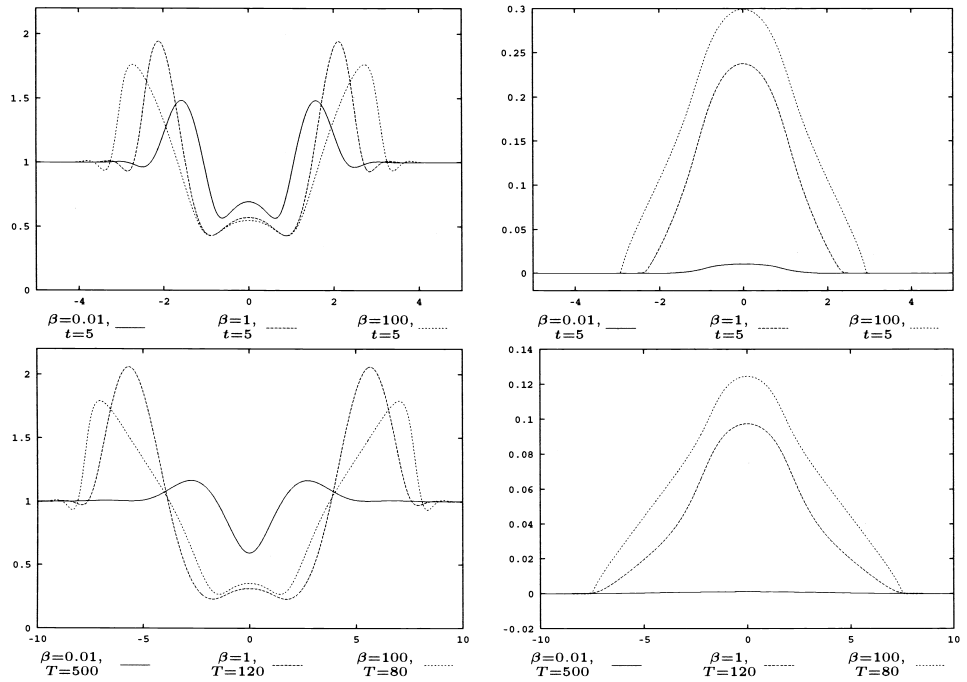


FIG. 1. *$U_\varepsilon(x,t)$ and $V_\varepsilon(x,t)$ for $\beta \in \{0.01, 1, 100\}$ and time $t = 5$ (above), and for different final times $t = T$ (below).*

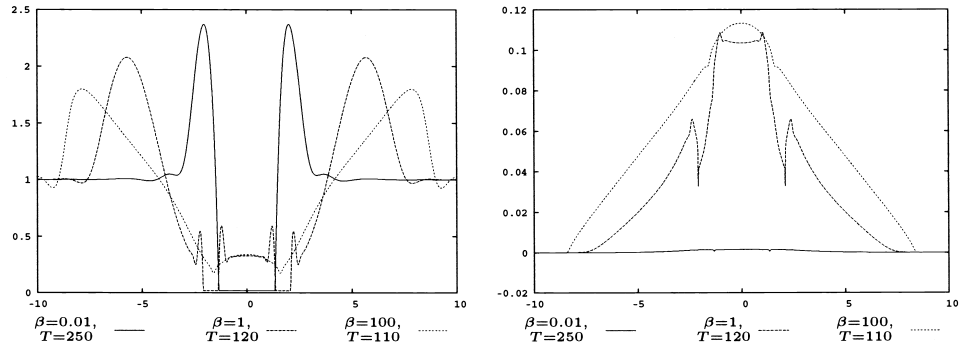


FIG. 2. *$U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ for $\beta \in \{0.01, 1, 100\}$ for different final times T.*

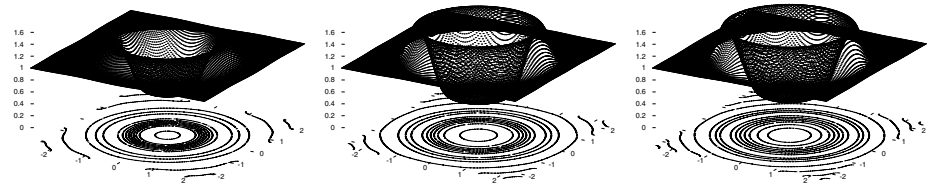


FIG. 3. *$U_\varepsilon(x,T)$ for $\beta = 0.01$, $T = 5$ (left), for $\beta = 1$, $T = 3$ (middle), and for $\beta = 100$, $T = 2$ (right).*

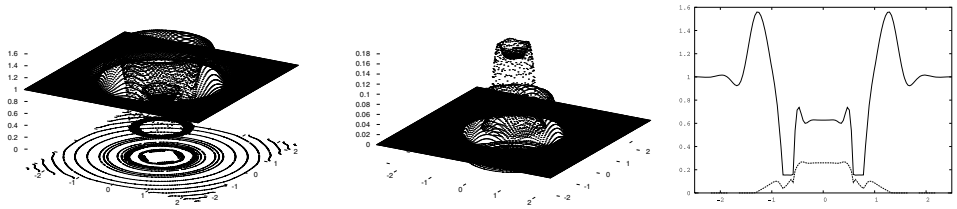


FIG. 4. $U_\varepsilon(x,T)$ *and* $V_\varepsilon(x,T)$ *for* $\beta = 0.01$ *and* $T = 1$. *On the right a plot of* $U_\varepsilon(x,T)\,|_{x_2=0}$ *and* $V_\varepsilon(x,T)\,|_{x_2=0}$.

4.2. Numerical results for $d = 2$. Finally, we present a numerical experiment in two space dimensions with $\Omega = (-L,L) \times (-L,L)$. We took a uniform mesh of squares of length $h = \frac{2L}{128}$, each of which was divided into two triangles by its northeast diagonal. We chose the following parameters for $(\mathrm{P}_\varepsilon^{h,\tau})$: $L = 2.5$, $c = 10^{-3}$, $\rho_s = 10^{-5}$, $\rho_b = 10^{-2}$, $K = 1$, $a = 0$, $\delta = 10^{-5}$, $\nu = 7$, $\tau_n = \tau = 10^{-3}$, and $\varepsilon = 10^{-5}$. For the initial profiles we chose (4.3). We set $U_\varepsilon^0 \equiv \pi^h u^0$, $V_\varepsilon^0 \equiv \pi^h v^0$, and $\Psi_\varepsilon \equiv \pi^h \psi^0$. In Figure 3 we plot $U_\varepsilon(x,T)$ for $\beta = 0.01, 1$, and 100 at different final times T.

Though on a slower time scale, the results are qualitatively similar to the experiments in one space dimension. The same holds true when including attractive van der Waals forces in the simulation. See Figure 4, where we plot U_ε and V_ε for $\beta = 0.01$ and the same parameters as above except $a = 0.02$. Again for brevity only V_ε is displayed since Ψ_ε is graphically indistinguishable from it. Note that here $\min_{x\in\overline{\Omega}} U_\varepsilon(x,T) \approx \arg\min_{s\in\mathbb{R}_{>0}} \Phi(s) = (\frac{\delta}{a})^{\frac{1}{\nu-3}} \approx 0.15$.

REFERENCES

[1] J. W. BARRETT AND J. F. BLOWEY, *Finite element approximation of a degenerate Allen–Cahn/Cahn–Hilliard system*, SIAM J. Numer. Anal., 39 (2001), pp. 1598–1624.

[2] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *Finite element approximation of a fourth order nonlinear degenerate parabolic equation*, Numer. Math., 80 (1998), pp. 525–556.

[3] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *Finite element approximation of the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Numer. Anal., 37 (1999), pp. 286–318.

[4] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *On fully practical finite element approximations of degenerate Cahn–Hilliard systems*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 713–748.

[5] J. W. BARRETT, H. GARCKE, AND R. NÜRNBERG, *Finite element approximation of surfactant spreading on a thin film*, SIAM J. Numer. Anal., 41 (2003), pp. 1427–1464.

[6] J. W. BARRETT AND R. NÜRNBERG, *Convergence of a finite element approximation of surfactant spreading on a thin film in the presence of van der Waals forces*, IMA J. Numer. Anal., 24 (2004), pp. 323–363.

[7] J. W. BARRETT AND R. NÜRNBERG, *Finite element approximation of a Stefan problem with degenerate Joule heating*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 633–652.

[8] F. BERNIS AND A. FRIEDMAN, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[9] G. GRÜN, *On the convergence of entropy consistent schemes for lubrication type equations in multiple space dimensions*, Math. Comp., 72 (2003), pp. 1251–1279.

[10] G. GRÜN AND M. RUMPF, *Nonnegativity preserving numerical schemes for the thin film equation*, Numer. Math., 87 (2000), pp. 113–152.

[11] O. E. JENSEN AND J. B. GROTBERG, *Insoluble surfactant spreading on a thin viscous film: Shock evolution and film rupture*, J. Fluid Mech., 240 (1992), pp. 259–288.

[12] O. E. Jensen and J. B. Grotberg, *The spreading of heat or soluble surfactant along a thin film*, Phys. Fluids A, 5 (1993), pp. 58–68.
[13] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.
[14] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1992.
[15] A. Sheludko, *Thin liquid films*, Adv. Colloid Interface Sci., 1 (1967), pp. 391–464.
[16] M. R. E. Warner, R. V. Craster, and O. K. Matar, *Fingering phenomena created by a soluble surfactant deposition on a thin liquid film*, Phys. Fluids, 16 (2004), pp. 2933–2951.
[17] L. Zhornitskaya and A. L. Bertozzi, *Positivity-preserving numerical schemes for lubrication-type equations*, SIAM J. Numer. Anal., 37 (2000), pp. 523–555.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1248–1274

CONTINUOUS INTERIOR PENALTY FINITE ELEMENT METHOD FOR OSEEN'S EQUATIONS*

ERIK BURMAN†, MIGUEL A. FERNÁNDEZ‡, AND PETER HANSBO§

Abstract. In this paper we present an extension of the continuous interior penalty method of Douglas and Dupont [*Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216] to Oseen's equations. The method consists of a stabilized Galerkin formulation using equal order interpolation for pressure and velocity. To counter instabilities due to the pressure/velocity coupling, or due to a high local Reynolds number, we add a stabilization term giving L^2-control of the jump of the gradient over element faces (edges in two dimensions) to the standard Galerkin formulation. Boundary conditions are imposed in a weak sense using a consistent penalty formulation due to Nitsche. We prove energy-type a priori error estimates independent of the local Reynolds number and give some numerical examples recovering the theoretical results.

Key words. finite element methods, stabilized methods, continuous interior penalty, Oseen's equations

AMS subject classifications. 65N12, 65N30, 76M10, 76D07

DOI. 10.1137/040617686

1. Introduction.

The construction of finite element methods for the incompressible Navier–Stokes equations that are robust and accurate for a wide range of Reynolds numbers remains a challenging problem. The standard Galerkin method requires the fulfillment of the inf-sup or Babuska–Brezzi condition, which leads to the need for formulations using mixed interpolations (see [7, 27]). From the computational point of view it is, however, more practical to use equal order interpolation for the velocity and pressure spaces, which requires that stability is imposed in some other fashion. One possibility is to construct *stabilized* finite element methods where some terms are added to the standard Galerkin formulation in order to enhance the stability properties of the method. To be useful the method must also be stable with respect to the convective terms and give sufficient control of the incompressibility condition.

A favored approach has been to stabilize both the velocities and the pressure using the streamline upwind Petrov–Galerkin (SUPG) method originally proposed by Brooks and Hughes in [9]. This method was first analyzed for the Navier–Stokes equations in a velocity vorticity formulation by Johnson and Saranen in [32], and then in a pressure velocity formulation by Hansbo and Szepessy in [29], by Franca and Frey

*Received by the editors October 16, 2004; accepted for publication (in revised form) February 7, 2006; published electronically June 30, 2006. A preliminary version of this paper summarizing the main results was published in the Eccomas 2004 proceedings [12]. This work was partially supported by the Research Training Network "Mathematical Modelling of the Cardiovascular System (HaeMOdel)," contract HPRN-CT-2002-00270 of the European Community.

http://www.siam.org/journals/sinum/44-3/61768.html

†Institut d'Analyse et Calcul Scientifique (CMCS/IACS), Ecole Polytechnique Federale de Lausanne, Switzerland (Erik.Burman@epfl.ch). This author was supported by INRIA during his visit to REO team at Rocquencourt.

‡INRIA projet REO, Rocquencourt BP 105, F-78153 Le Chesnay Cedex, France (miguel.fernandez@inria.fr).

§Department of Applied Mechanics, Chalmers University of Technology, SE-41296 Göteborg, Sweden (hansbo@am.chalmers.se). This author was supported by the Semester Program "Mathematical Modelling of the Cardiovascular System" of the Bernoulli Center at the EPFL.

in [24], and by Tobiska and Verfürth in [38]. The SUPG method owes its success to the unified treatment of velocities and pressures. It allows for a priori error estimates that are independent of the Reynolds number and has been used extensively in practice with good results. Nevertheless, the SUPG method has some undesirable features:

- artificial boundary conditions on velocities and pressure are introduced;
- artificial nonsymmetric terms are introduced;
- the least squares term introduces nonphysical pressure/velocity couplings;
- the least squares term makes mass lumping impossible and the choice of time-stepping methods limited; most clear-cut from a theoretical point of view is a space-time finite element approach using discontinuous approximation in time;
- it is not yet fully understood how to use mixed finite elements in combination with the SUPG method (for recent advances see [26]).

To overcome these disadvantages, alternative stabilization techniques have been developed such as the projection method proposed by Codina [17] and Codina and Blasco [18], the subgrid viscosity method or local projection method proposed by Guermond [28] and Becker and Braack [1], the polynomial pressure projection method by Dohrmann and Bochev reported in [19], and the pressure-Poisson stabilization of the Stokes equations proposed by Bochev and Gunzburger in [2].

Recently, the continuous interior penalty method of Douglas and Dupont [20] was revived as an alternative. The idea is to add a least squares penalization on the gradient jump between neighboring elements as a unified treatment of all the above-mentioned instabilities. It was shown in [13, 15] that the method stabilizes both instabilities due to dominating convection and instabilities due to the velocity/pressure coupling. Moreover, it was shown in [10] how this method provides a natural link between conforming and nonconforming stabilized finite element methods. It was used in [14] to provide a Reynolds number independent stabilized formulation for the classical nonconforming P_1 Crouzeix–Raviart approximation for the velocities combined with elementwise constant pressures.

In this paper we extend the face oriented stabilization method to Oseen's equations, using equal order interpolation for velocities and pressure. For the case of similar stabilization strategies for element pairs satisfying the inf-sup condition we refer the reader to [11]. We follow the framework proposed in [10] using weakly imposed boundary conditions as introduced by Nitsche (see [36, 25]). Although the constants of our analysis inevitably depend on the parameters of the problem (since the solution depends on the physical parameters), the stabilization terms allow us to trade the need of coercivity in the H^1-norm for coercivity in the weaker L^2-norm plus the stabilization term, which vanishes at optimal rate under refinement. To exploit this in the analysis, we add a zero order term to Oseen's equations. With this additional term we obtain estimates that do not explode as the viscosity goes to zero, provided the exact solution is sufficiently regular.

With the proposed method, all the above-mentioned inconveniences of the SUPG method are alleviated. The formulation allows for general unstructured meshes and variable polynomial degree. The main inconveniences of the present method, however, are as follows:

- Added couplings in the Jacobian matrix: the bandwidth of the system matrix doubles in two space dimensions and triples in three space dimensions. This may increase the computational cost of the linear system solution, for instance, if an incomplete LU factorization is used as preconditioner.

- The method requires stabilization terms to be evaluated on the faces of the elements and hence a table of nearest neighbors for computation of the jumps. Such features are present typically when using adaptivity with a posteriori error estimation or for discontinuous Galerkin (DG) methods.

However, all stabilization terms have the same structure, allowing for one computation of one global stabilization matrix based on the gradient jumps of one component. The parameter values that change from time-step to time-step may then be updated using locally averaged weights. When time-stepping the Navier–Stokes equations this means that the stabilization matrix has to be constructed only once and for one component. This is in stark contrast to the SUPG method, where the stabilization terms have to be reconstructed at every time-step for consistency.

The outline of the paper is as follows: In the next section we introduce our model problem, Oseen's equations, and formulate the interior penalty finite element method. In section 3 we discuss the question of stability, we prove a lemma of fundamental importance for the stability of the method, and we show that the discrete problem has a unique solution. We then proceed and prove (quasi-) optimal a priori error estimates in section 4 with special focus on how to make the estimates independent of the local Reynolds number. Finally, in section 5, we study the performance of the numerical scheme on some linear model cases in three space dimensions. We make some concluding remarks in section 6 and some outlooks to future developments, with special emphasis on the relation between the present method and variational multiscale methods (VMS) in large eddy simulations (LES).

2. A finite element method for Oseen's equations. Let Ω be a Lipschitz-continuous domain in $\mathbb{R}^d$ ($d = 2$ or 3) with a polyhedral boundary $\partial\Omega$ and outward pointing normal $\mathbf{n}$. We will consider the Sobolev spaces $W^{m,q}(\Omega)$, with norm $\|\cdot\|_{m,q,\Omega}$, $m \geq 0$, and $q \geq 1$. In particular, we have $L^q(\Omega) = W^{0,q}(\Omega)$. We use the standard notation $H^m(\Omega) \stackrel{\text{def}}{=} W^{m,2}(\Omega)$. The norm of $H^m(\Omega)$ is denoted by $\|\cdot\|_{m,\Omega}$ and its seminorm by $|\cdot|_{m,\Omega}$. The space of $L^2(\Omega)$ divergence free functions is denoted by $H_0(\mathrm{div};\Omega)$. The scalar product in $L^2(\Omega)$ is denoted by $(\cdot,\cdot)$ and its norm by $\|\cdot\|_{0,\Omega}$. The closed subspaces $H_0^1(\Omega)$, consisting of functions in $H^1(\Omega)$ with zero trace on $\partial\Omega$, and $L_0^2(\Omega)$, consisting of functions in $L^2(\Omega)$ with zero mean in Ω, will also be used. Oseen's equations take the form

$$\left\{\begin{array}{rl} \sigma\mathbf{u} + \boldsymbol{\beta}\cdot\nabla\mathbf{u} - 2\nabla\cdot(\nu\boldsymbol{\epsilon}(\mathbf{u})) + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = 0 & \text{on } \partial\Omega, \end{array}\right. \tag{2.1}$$

where $\mathbf{u} \in [H_0^1(\Omega)]^d \cap H_0(\mathrm{div};\Omega)$, $\boldsymbol{\beta} \in [W^{1,\infty}(\Omega)]^d \cap H_0(\mathrm{div};\Omega)$, $p \in L_0^2(\Omega)$, $\boldsymbol{\epsilon}(\mathbf{u})$ stands for the strain rate tensor

$$\boldsymbol{\epsilon}(\mathbf{u}) \stackrel{\text{def}}{=} \frac{1}{2}\Big(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\Big),$$

$\mathbf{f} \in [L^2(\Omega)]^d$ is a given source term, and σ, ν are positive constants.

The weak formulation of problem (2.1) reads as follows: find $(\mathbf{u}, p) \in [H_0^1(\Omega)]^d \times L_0^2(\Omega)$ such that

$$\left\{\begin{array}{c} a(\mathbf{u},\mathbf{v}) + b(p,\mathbf{v}) = (\mathbf{f},\mathbf{v}), \\ b(q,\mathbf{u}) = 0 \end{array}\right. \quad \forall (\mathbf{v},q) \in [H_0^1(\Omega)]^d \times L_0^2(\Omega), \tag{2.2}$$

where

$$
\begin{aligned}
a(\mathbf{u},\mathbf{v}) &\stackrel{\text{def}}{=} (\sigma\,\mathbf{u},\,\mathbf{v}) + (\boldsymbol{\beta}\cdot\nabla\mathbf{u},\,\mathbf{v}) + 2(\nu\boldsymbol{\epsilon}(\mathbf{u}),\boldsymbol{\epsilon}(\mathbf{v})),\\
b(p,\mathbf{v}) &\stackrel{\text{def}}{=} -(p,\nabla\cdot\mathbf{v}).
\end{aligned}
\tag{2.3}
$$

The well-posedness of this problem follows by the Lax–Milgram lemma applied in the space $[H_0^1(\Omega)]^d \cap H_0(\mathrm{div};\Omega)$ (see, for instance, [27]).

Let $\{\mathcal{T}_h\}_{0<h\le 1}$ denote a family of triangulations of the domain Ω without hanging nodes. For each triangulation $\mathcal{T}_h$, the subscript $h \in (0,1]$ refers to the level of refinement of the triangulation, which is defined by

$$
h \stackrel{\text{def}}{=} \max_{K\in\mathcal{T}_h} h_K,
$$

with h_K the diameter of K. We also define the elementwise constant function $\tilde{h}|_K = h_K$. The interior of a triangle K will be denoted by $\overset{\circ}{K}$, and $\mathcal{N}(K)$ will stand for the set of elements sharing at least one node with the element K. Moreover, we will assume that the family $\{\mathcal{T}_h\}_{0<h\le 1}$ has the following regularity properties:

1. Local shape regularity: for all $K \in \mathcal{T}_h$ with $h \in (0,1]$ there holds

$$
\frac{h_K}{\rho_K} < c_0,
\tag{2.4}
$$

where ρ_K stands for the diameter of the largest ball contained in K, and c_0 is a fixed positive constant.

2. Local quasi-uniformity: for all $K \in \mathcal{T}_h$ with $h \in (0,1]$ there holds

$$
\frac{1}{\rho} h_{K'} \le h_K \le \rho h_{K'} \quad \forall K' \in \mathcal{N}(K),
\tag{2.5}
$$

where $\rho > 1$ is a given parameter depending on the local uniformity of $\{\mathcal{T}_h\}_{0<h\le 1}$.

We will also assume that the data are sufficiently well resolved in the sense that there exists $\rho_{\boldsymbol{\beta}} > 1$ such that

$$
\frac{1}{\rho_{\boldsymbol{\beta}}}\|\boldsymbol{\beta}\|_{0,\infty,K'} \le \|\boldsymbol{\beta}\|_{0,\infty,K} \le \rho_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_{0,\infty,K'} \quad \forall K' \in \mathcal{N}(K).
\tag{2.6}
$$

Note that this is a hypothesis on the mesh and not on the data. Under the assumption that $\boldsymbol{\beta} \in W^{1,\infty}(\mathcal{N}(K))$ this can be ensured by

$$
|\boldsymbol{\beta}|_{1,\infty,\mathcal{N}(K)} \le c_{\boldsymbol{\beta}} h_K^{-1} \|\boldsymbol{\beta}\|_{0,\infty,\mathcal{N}(K)}
\tag{2.7}
$$

for some constant $c_{\boldsymbol{\beta}} > 0$ small enough.

For the error analysis, we shall use the trace inequality

$$
\|v\|_{0,\partial K}^2 \le C_{\mathrm{T}}\Big(h_K^{-1}\|v\|_{0,K}^2 + h_K\|v\|_{1,K}^2\Big) \quad \forall v \in H^1(K),
\tag{2.8}
$$

where C_{T} is a generic constant independent of h_K (for a proof, see [37, p. 26]).

For a given piecewise continuous function φ, the jump $[\![\varphi]\!]$ over a face f is defined by

$$
[\![\varphi]\!](\mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} \lim\limits_{t\to 0^+} (\varphi(\mathbf{x}-t\mathbf{n}_f) - \varphi(\mathbf{x}+t\mathbf{n}_f)) & \text{if} \quad f \not\subset \partial\Omega,\\ 0 & \text{if} \quad f \subset \partial\Omega, \end{cases}
$$

where $\mathbf{n}_f$ is a normal unit vector on f and $\mathbf{x} \in f$.

In this paper we let V_h^k denote the standard space of continuous functions of piecewise polynomial order $k \geq 1$,

$$V_h^k \stackrel{\text{def}}{=} \left\{ v \in H^1(\Omega) \,:\, v_{|K} \in P_k(K) \quad \forall K \in \mathcal{T}_h \right\},$$

and $H^2(\mathcal{T}_h)$ the space of piecewise H^2 functions

$$H^2(\mathcal{T}_h) \stackrel{\text{def}}{=} \left\{ v : \Omega \longrightarrow \mathbb{R} \,:\, v_{|K} \in H^2(K) \quad \forall K \in \mathcal{T}_h \right\}.$$

For the velocities we will use the space $[V_h^k]^d$ and for the pressure we will use $Q_h^k = V_h^k \cap L_0^2(\Omega)$. In what follows, we let $\pi_{h,k}$, $i_{h,k}$, and $\mathcal{C}_{h,k}$ denote (respectively) the L^2-projection operator, the nodal interpolation operator, and the Clément interpolant onto the finite element spaces, and we make no notational difference between the projection onto the velocity and pressure spaces. We also introduce a piecewise linear approximated velocity $\boldsymbol{\beta}_h \in [V_h^1]^d$ such that

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_{0,\infty,K} \leq C h_K |\boldsymbol{\beta}|_{1,\infty,K} \quad \forall K \in \mathcal{T}_h. \tag{2.9}$$

Here and in the following C denotes a constant independent of the problem parameters and the local mesh size, but not necessarily of the local mesh geometry. Denoting the product space $W_h^k \stackrel{\text{def}}{=} [V_h^k]^d \times Q_h^k$ our finite element method reads as follows: find $(\mathbf{u}_h, p_h) \in W_h^k$ such that

$$a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(p_h, \mathbf{v}_h) - b_h(q_h, \mathbf{u}_h) + j_{\mathbf{u}}(\mathbf{u}_h, \mathbf{v}_h) + j_p(p_h, q_h) = (\mathbf{f}, \mathbf{v}_h) \tag{2.10}$$

for all $(\mathbf{v}_h, q_h) \in W_h^k$, and with

$$\begin{aligned} a_h(\mathbf{u}_h, \mathbf{v}_h) \stackrel{\text{def}}{=}\ & a(\mathbf{u}_h, \mathbf{v}_h) - \langle 2\nu\boldsymbol{\epsilon}(\mathbf{u}_h)\mathbf{n}, \mathbf{v}_h\rangle_{\partial\Omega} - \langle \mathbf{u}_h, 2\nu\boldsymbol{\epsilon}(\mathbf{v}_h)\mathbf{n}\rangle_{\partial\Omega} \\ & - \langle \boldsymbol{\beta}\cdot\mathbf{n}\mathbf{u}_h, \mathbf{v}_h\rangle_{\partial\Omega_{\text{in}}} + \langle \gamma(\nu/\tilde{h})\mathbf{u}_h, \mathbf{v}_h\rangle_{\partial\Omega} \\ & + \langle \gamma \max\{|\boldsymbol{\beta}|, \nu/\tilde{h}\}\mathbf{u}_h\cdot\mathbf{n}, \mathbf{v}_h\cdot\mathbf{n}\rangle_{\partial\Omega}, \end{aligned} \tag{2.11}$$

$$b_h(p_h, \mathbf{v}_h) \stackrel{\text{def}}{=} b(p_h, \mathbf{v}_h) + \langle p_h, \mathbf{v}_h\cdot\mathbf{n}\rangle_{\partial\Omega}, \tag{2.12}$$

$$\begin{aligned} j_{\mathbf{u}}(\mathbf{u}_h, \mathbf{v}_h) \stackrel{\text{def}}{=}\ & \sum_{K\in\mathcal{T}_h} \gamma\xi(\mathrm{Re_K}) h_K^2 \int_{\partial K} \|\boldsymbol{\beta}\cdot\mathbf{n}\|_{0,\infty,\partial K} [\![\mathbf{n}\cdot\nabla\mathbf{u}_h]\!]\cdot[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]\,\mathrm{d}s \\ & + \sum_{K\in\mathcal{T}_h} \gamma\xi(\mathrm{Re_K})\|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2 \int_{\partial K} [\![\nabla\cdot\mathbf{u}_h]\!][\![\nabla\cdot\mathbf{v}_h]\!]\,\mathrm{d}s, \end{aligned} \tag{2.13}$$

$$j_p(p_h, q_h) \stackrel{\text{def}}{=} \sum_{K\in\mathcal{T}_h} \gamma\xi(\mathrm{Re_K}) \frac{h_K^2}{\|\boldsymbol{\beta}\|_{0,\infty,K}} \int_{\partial K} [\![\nabla p_h]\!]\cdot[\![\nabla q_h]\!]\,\mathrm{d}s, \tag{2.14}$$

$\mathbf{n}$ the outward pointing normal to $\partial\Omega$, and using the notation

$$\mathrm{Re}_K \stackrel{\text{def}}{=} \frac{\|\boldsymbol{\beta}\|_{0,\infty,K} h_K}{\nu}, \quad \xi(\lambda) \stackrel{\text{def}}{=} \min\{1,\lambda\}, \quad \partial\Omega_{\text{in}} \stackrel{\text{def}}{=} \{\mathbf{x}\in\partial\Omega \,:\, (\boldsymbol{\beta}\cdot\mathbf{n})(\mathbf{x}) < 0\},$$

$$\langle x, y\rangle_{\partial\Omega} \stackrel{\text{def}}{=} \sum_{\substack{K\,\in\,\mathcal{T}_h \\ K\,\cap\,\partial\Omega\,\neq\,\emptyset}} \int_{\partial K\cap\partial\Omega} xy\,\mathrm{d}s, \quad \tilde{h}\in H^2(\mathcal{T}_h) \quad \text{with} \quad \tilde{h}_{|K} \stackrel{\text{def}}{=} h_K.$$

To keep down notation we have used a canonical stabilization parameter γ for all terms. In practice this term, however, can be chosen distinctly for different terms. The gradient jump terms serve three purposes:

1. stabilization of the convective terms (the first sum in (2.13));
2. giving additional control of the incompressibility condition (the second sum in (2.13)); and
3. making the discretization inf-sup stable (the sum in (2.14)).

We will see in the analysis that these three objectives are all obtained in the same fashion and that essentially the gradient jump operator can stabilize any instability provoked by a first order term.

Assuming sufficient regularity of the exact solution the above formulation is strongly consistent. More generally we have the following result.

LEMMA 2.1 (modified Galerkin orthogonality). *Assume that* $(\mathbf{u}, p)$*, the solution of* (2.1)*, belongs to the space* $[H^{3/2+\varepsilon}(\Omega)]^d \times L^2_0(\Omega)$*, with* $\varepsilon > 0$*, and let* $(\mathbf{u}_h, p_h) \in W_h^k$ *be the solution of* (2.10)*. Then*

$$a_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + b_h(p - p_h, \mathbf{v}_h) - b_h(q_h, \mathbf{u} - \mathbf{u}_h) + j_{\mathbf{u}}(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) - j_p(p_h, q_h) = 0$$

for all $(\mathbf{v}_h, q_h) \in W_h^k$*.*

Proof. This is an immediate consequence of the consistency of the standard Galerkin method and the fact that, under the regularity assumptions, $j_{\mathbf{u}}(\mathbf{u}, \mathbf{v}_h) = 0$ since $[\![\nabla \mathbf{u}]\!]_f = 0$ for all interior faces f. ∎

3. Stability of the method. Stability in the face oriented stabilization method is based on the following lemma, which was proved for piecewise linear continuous approximation in [10] (for a similar result with applications to DG methods see [33]). Here we extend this result to arbitrary polynomial degree. Note that we give a lower bound as well. This is not needed for the analysis but shows that in some sense the stabilizing terms are optimal.

LEMMA 3.1. *There exist an interpolation operator* $\pi^*_{h,k} : [H^2(\mathcal{T}_h)]^d \longrightarrow [V_h^k]^d$ *and constants* γ*,* γ_{low} *depending on the local mesh geometry and the polynomial degree, but not on the local mesh size, such that*

$$\gamma_{low} j_{\boldsymbol{\beta}}(\mathbf{v}_h, \mathbf{v}_h) \le \| h^{\frac{1}{2}} \big(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h - \pi^*_{h,k}(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h) \big) \|^2_{0,\Omega} \le j_{\boldsymbol{\beta}}(\mathbf{v}_h, \mathbf{v}_h)$$

for all $\mathbf{v}_h \in [V_h^k]^d$*, where*

$$j_{\boldsymbol{\beta}}(\mathbf{v}_h, \mathbf{v}_h) = \gamma \sum_{K \in \mathcal{T}_h} \int_{\partial K} h_K^2 |\boldsymbol{\beta}_h \cdot \mathbf{n}|^2 |[\![\nabla \mathbf{v}_h \mathbf{n}]\!]|^2 \, \mathrm{d}s.$$

Proof. First note that, as pointed out in [10], $[\![\boldsymbol{\beta}_h \cdot \nabla \mathbf{u}_h]\!] = \boldsymbol{\beta}_h \cdot \mathbf{n} [\![\mathbf{n} \cdot \nabla \mathbf{u}_h]\!]$. For each node x_i, let n_i be the number of elements containing x_i as a node. Then we define a quasi-interpolant $\pi^*_{h,k}$ of degree k by

$$\pi^*_{h,k} \mathbf{v}(x_i) \stackrel{\text{def}}{=} \frac{1}{n_i} \sum_{\{K \,:\, x_i \in K\}} \mathbf{v}_{|K}(x_i) \quad \forall \mathbf{v} \in [H^2(\mathcal{T}_h)]^d.$$

For each element $K \in \mathcal{T}_h$ consider the function

$$\delta_K \stackrel{\text{def}}{=} h_K^{\frac{1}{2}} \big(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_{h|K} - \pi^*_{h,k}(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h)_{|K} \big).$$

Clearly, $\delta_K(x_j) = 0$ for each interior node $x_j \in \overset{\circ}{K}$, whereas on the element faces, i.e., for all nodes $x_j \in \partial K$, we have

$$\begin{aligned}\delta_K(x_j) &= h_K^{\frac{1}{2}} \frac{1}{n_j} \sum_{\{K' : x_j \in K'\}} \boldsymbol{\beta}_h \cdot \big(\nabla \mathbf{v}_{h|K}(x_j) - \nabla \mathbf{v}_{h|K'}(x_j)\big) \\ &= h_K^{\frac{1}{2}} \frac{1}{n_j} \sum_{\{K' : x_j \in K'\}} \sum_{f \in P(K,K')} \boldsymbol{\beta}_h(x_j) \cdot [\![\nabla \mathbf{v}_h]\!]_f(x_j), \end{aligned} \tag{3.1}$$

where $P(K, K')$ stands for the set of faces between K and K' (the shortest path). We now introduce the reference element $\hat{K}$ and, for each $K \in \mathcal{T}_h$, the affine mapping

$$F_K(\hat{\mathbf{x}}) = B_K \hat{\mathbf{x}} + \mathbf{b}_K \quad \forall \hat{\mathbf{x}} \in \hat{K},$$

such that $F_K(\hat{K}) = K$. Finally, let φ_j^K for $j = 1, \dots, k$ be the basis functions on K. Since $\delta_K(x_j) = 0$ for each interior node $x_j \in \overset{\circ}{K}$, $\|\delta_K \circ F_K\|_{0,\partial\hat{K}}^2 = 0$ implies that $\delta_K \circ F_K = 0$ in $\hat{K}$. Therefore, by equivalence of norms on discrete spaces, using a standard scaling argument (see [27, p. 96]) and (3.1), it follows that

$$\begin{aligned}\|\delta_K\|_{0,K}^2 &= \det B_K \|\delta_K \circ F_K\|_{0,\hat{K}}^2 \\ &\le C \det B_K \|\delta_K \circ F_K\|_{0,\partial\hat{K}}^2 \\ &= \int_{\partial\hat{K}} \frac{1}{|B_K^{-\mathrm{T}} \hat{\mathbf{n}}|} |\delta_K \circ F_K|^2 \underbrace{\det B_K |B_K^{-\mathrm{T}} \hat{\mathbf{n}}| \, \mathrm{d}\hat{s}}_{\mathrm{d}s} \\ &\le C |B_K^{\mathrm{T}}| \int_{\partial K} |\delta_K|^2 \, \mathrm{d}s \\ &\le C h_K \int_{\partial K} |\delta_K|^2 \, \mathrm{d}s \\ &\le C h_K \int_{\partial K} \sum_{j=1}^{k} |\delta_K(x_j)|^2 (\varphi_j^K)^2 \, \mathrm{d}s \\ &\le C h_K^2 \int_{\partial K} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{\{K' : x_j \in K'\}} \sum_{e \in P(K,K')} |\boldsymbol{\beta}_h(x_j) \cdot [\![\nabla \mathbf{v}_h]\!]_e(x_j)|^2 (\varphi_j^K)^2 \, \mathrm{d}s \\ &\le C h_K^2 \int_{\partial K} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{f \in E(K)} |\boldsymbol{\beta}_h(x_j) \cdot [\![\nabla \mathbf{v}_h]\!]_f(x_j)|^2 (\varphi_j^K)^2 \, \mathrm{d}s \\ &\le C h_K^2 \sum_{f \in E(K)} \int_f |\boldsymbol{\beta}_h \cdot [\![\nabla \mathbf{v}_h]\!]_f|^2 \, \mathrm{d}s, \end{aligned}$$

where, in the two last inequalities, $E(K)$ denotes the set of faces containing some node of K. On the other hand, the local quasi-regularity of $\mathcal{T}_h$ implies that the maximum number of occurrences of a face in all the sets $E(K)$ is bounded by a fixed constant independent of h_K. Then, by summation on K, we get the upper bound

$$\begin{aligned}\|h^{\frac{1}{2}} \big(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h - \pi_{h,k}^*(\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h)\big)\|_{0,\Omega}^2 &\le C \sum_{K \in \mathcal{T}_h} h_K^2 \sum_{f \in E(K)} \int_f |\boldsymbol{\beta}_h \cdot [\![\nabla \mathbf{v}_h]\!]_e|^2 \, \mathrm{d}s, \\ &\le C \sum_{K \in \mathcal{T}_h} \int_{\partial K} h_K^2 |\boldsymbol{\beta}_h \cdot [\![\nabla \mathbf{v}_h]\!]|^2 \, \mathrm{d}s. \end{aligned}$$

The lower bound follows by considering the L^2-norm of the discontinuous function δ over the reference patch $\hat{G}$ consisting of the reference element $\hat{K}$ and its nearest neighbors. Clearly if $\|\delta\|_{\hat{G}} = 0$, then $\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h = \pi^*_{h,k} \boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h$ in $\hat{G}$. This means that $\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h$ is continuous in $\hat{G}$ and hence $\sum_{f \in E(K)} \int_f h_K [\![\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h]\!]^2 \, \mathrm{d}s = 0$. Hence by norm equivalence on discrete spaces we have

$$\sum_{f \in E(K)} \int_f [\![\boldsymbol{\beta}_h \cdot \nabla \mathbf{v}_h]\!]^2 \, \mathrm{d}s \leq \|\delta_G\|^2_{0,G}.$$

The claim then follows in the same fashion as the first part of the proof by scaling and extension to all of $\mathcal{T}_h$. ∎

Using the same technique we immediately have the following corollary where for simplicity the lower bounds are omitted.

COROLLARY 3.2. *Under the same assumptions as Lemma* 3.1 *we have, with* $\alpha > 0$ *and* ϕ *some function that is constant per element,*

$$\begin{aligned} \|\phi^{\frac{1}{2}} \big(\nabla \cdot \mathbf{v}_h - \pi^*_{h,k}(\nabla \cdot \mathbf{v}_h)\big)\|^2_{0,\Omega} &\leq \gamma \sum_{K \in \mathcal{T}_h} \int_{\partial K} \phi h_K [\![\nabla \cdot \mathbf{v}_h]\!]^2 \, \mathrm{d}s, \\ \|\phi^{\frac{1}{2}} \big(\nabla q_h - \pi^*_{h,k}(\nabla q_h)\big)\|^2_{0,\Omega} &\leq \gamma \sum_{K \in \mathcal{T}_h} \int_{\partial K} \phi h_K |[\![\nabla q_h]\!]|^2 \, \mathrm{d}s \end{aligned} \tag{3.2}$$

for all $(\mathbf{v}_h, q_h) \in W^k_h$ *and with* $\gamma > 0$ *constants independent of* h.

We now introduce the following mesh-dependent norm for the velocity:

$$\begin{aligned} |\!|\!| \mathbf{v}_h |\!|\!|^2 \stackrel{\text{def}}{=} \|\sigma^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\Omega} + \|\nu^{\frac{1}{2}} \nabla \mathbf{v}_h\|^2_{0,\Omega} + j_{\mathbf{u}}(\mathbf{v}_h, \mathbf{v}_h) + \||\boldsymbol{\beta} \cdot \mathbf{n}|^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\partial\Omega} \\ + \|\gamma^{\frac{1}{2}} (\nu/\tilde{h})^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\partial\Omega} + \|\gamma^{\frac{1}{2}} \max\{|\boldsymbol{\beta}|, \nu/\tilde{h}\}^{\frac{1}{2}} \mathbf{v}_h \cdot \mathbf{n}\|^2_{0,\partial\Omega} \end{aligned} \tag{3.3}$$

for all $\mathbf{v}_h \in [V^k_h]^d$.

The following lemma gives the coercivity of our discrete operator with respect to this mesh-dependent norm.

LEMMA 3.3 (coercivity). *There exists a constant* $C > 0$, *depending only on* Ω *and* γ, *such that*

$$a_h(\mathbf{v}_h, \mathbf{v}_h) + j_{\mathbf{u}}(\mathbf{v}_h, \mathbf{v}_h) \geq C |\!|\!| \mathbf{v}_h |\!|\!|^2$$

for all $\mathbf{v}_h \in [V^k_h]^d$.

Proof. From (2.10) we get

$$\begin{aligned} a_h(\mathbf{v}_h, \mathbf{v}_h) + j_{\mathbf{u}}(\mathbf{v}_h, \mathbf{v}_h) \geq \|\sigma^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\Omega} + 2\|\nu^{\frac{1}{2}} \boldsymbol{\epsilon}(\mathbf{v}_h)\|^2_{0,\Omega} + j_{\mathbf{u}}(\mathbf{v}_h, \mathbf{v}_h) \\ + \frac{1}{2} \||\boldsymbol{\beta} \cdot \mathbf{n}|^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\partial\Omega} + \|\gamma^{\frac{1}{2}} (\nu/\tilde{h})^{\frac{1}{2}} \mathbf{v}_h\|^2_{0,\partial\Omega} \\ + \|\gamma^{\frac{1}{2}} \max\{|\boldsymbol{\beta}|, \nu/\tilde{h}\}^{\frac{1}{2}} \mathbf{v}_h \cdot \mathbf{n}\|^2_{0,\partial\Omega} \\ - \langle 4\nu \boldsymbol{\epsilon}(\mathbf{v}_h) \mathbf{n}, \mathbf{v}_h \rangle_{\partial\Omega}, \end{aligned} \tag{3.4}$$

where we used the fact that, after integration by parts and since $\nabla \cdot \boldsymbol{\beta} = 0$,

$$(\boldsymbol{\beta} \cdot \nabla \mathbf{v}_h, \mathbf{v}_h) = \frac{1}{2} \langle \boldsymbol{\beta} \cdot \mathbf{n} \mathbf{v}_h, \mathbf{v}_h \rangle_{\partial\Omega}.$$

The last term in (3.4) can be bounded using the Cauchy–Schwarz inequality followed by a trace inequality, to obtain

$$|\langle 4\nu\boldsymbol{\epsilon}(\mathbf{v}_h)\mathbf{n},\mathbf{v}_h\rangle_{\partial\Omega}| \leq 8\frac{C_{\mathrm{T}}}{\gamma}\|\nu^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v}_h)\|^2_{0,\Omega} + \frac{1}{2}\|\gamma^{\frac{1}{2}}(\nu/\tilde{h})^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega}.$$

In what follows we will assume that

$$\gamma > 4C_{\mathrm{T}} > 0, \tag{3.5}$$

and therefore

$$\lambda(\gamma) \stackrel{\text{def}}{=} 2 - 8\frac{C_{\mathrm{T}}}{\gamma} > 0.$$

From (3.4), we then get

$$\begin{aligned} a_h(\mathbf{v}_h,\mathbf{v}_h) + j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h) \geq{}& \|\sigma^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\Omega} + \lambda(\gamma)\|\nu^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v}_h)\|^2_{0,\Omega} + j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h) \\ &+ \frac{1}{2}\||\boldsymbol{\beta}\cdot\mathbf{n}|^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega} + \frac{1}{2}\|\gamma^{\frac{1}{2}}(\nu/\tilde{h})^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega} \\ &+ \|\gamma^{\frac{1}{2}}\max\{|\boldsymbol{\beta}|,\nu/\tilde{h}\}^{\frac{1}{2}}\mathbf{v}_h\cdot\mathbf{n}\|^2_{0,\partial\Omega}, \end{aligned}$$

and consequently

$$\begin{aligned} a_h(\mathbf{v}_h,\mathbf{v}_h) + j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h) \geq{}& \|\sigma^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\Omega} + j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h) \\ &+ \min_{\substack{K\in\mathcal{T}_h \\ K\cap\partial\Omega\neq\emptyset}} \left\{\lambda(\gamma), \frac{\gamma}{4h_K}\right\}\left(\|\nu^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v}_h)\|^2_{0,\Omega} + \|\nu^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega}\right) \\ &+ \frac{1}{2}\||\boldsymbol{\beta}\cdot\mathbf{n}|^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega} + \frac{1}{4}\|\gamma^{\frac{1}{2}}(\nu/\tilde{h})^{\frac{1}{2}}\mathbf{v}_h\|^2_{0,\partial\Omega} \\ &+ \|\gamma^{\frac{1}{2}}\max\{|\boldsymbol{\beta}|,\nu/\tilde{h}\}^{\frac{1}{2}}\mathbf{v}_h\cdot\mathbf{n}\|^2_{0,\partial\Omega}. \end{aligned}$$

In particular, since $0 < h \leq 1$ and by choosing (accordingly with (3.5))

$$\gamma \stackrel{\text{def}}{=} \varepsilon + 4C_{\mathrm{T}},$$

with $\varepsilon > 0$ sufficiently small, one obtains

$$\lambda(\gamma) < \frac{\gamma}{4h_K} \quad \forall K\in\mathcal{T}_h, \quad K\cap\partial\Omega\neq\emptyset.$$

We then conclude the proof using Korn's inequality (see [6]). ∎

In what follows, we shall make use of the following discrete pressure and velocity subspaces:

$$\begin{aligned} C^1_{h,k} &\stackrel{\text{def}}{=} \left\{q_h\in Q^k_h \,:\, j_p(q_h,q_h) = 0\right\}, \\ V^{\mathrm{div}}_{h,k} &\stackrel{\text{def}}{=} \left\{\mathbf{v}_h\in[V^k_h]^d \,:\, b_h(q_h,\mathbf{v}_h) = 0 \quad \forall q_h\in C^1_{h,k}\right\}. \end{aligned}$$

In addition, $Q^k_h\backslash C^1_{h,k}$ will stand for the supplementary of $C^1_{h,k}$ in Q^k_h, i.e.,

$$Q^k_h = (Q^k_h\backslash C^1_{h,k}) \oplus C^1_{h,k}.$$

The following lemma ensures, in particular, that $V_{h,k}^{\mathrm{div}}$ is not trivial.

LEMMA 3.4. *There exists a constant $\boldsymbol{\beta} > 0$, independent of h, such that*

$$\inf_{q_h \in C_{h,k}^1} \sup_{\mathbf{v}_h \in [V_h^k]^d} \frac{|b_h(q_h, \mathbf{v}_h)|}{\|q_h\|_{0,\Omega} \|\mathbf{v}_h\|_{1,\Omega}} \geq \beta.$$

Proof. Let $q_h \in C_{h,k}^1$. From [27, Corollary 2.4], there exists $\mathbf{v}_q \in [H_0^1(\Omega)]^d$ such that

$$\nabla \cdot \mathbf{v}_q = q_h, \quad \|\mathbf{v}_q\|_{1,\Omega} \leq C \|q_h\|_{0,\Omega}. \tag{3.6}$$

Thus, using integration by parts and (2.12), we have

$$\begin{aligned} \|q_h\|_{0,\Omega}^2 &= (q_h, \nabla \cdot \mathbf{v}_q) \\ &= (q_h, \nabla \cdot \mathbf{v}_q - \nabla \cdot \pi_{h,k} \mathbf{v}_q) + (q_h, \nabla \cdot \pi_{h,k} \mathbf{v}_q) \\ &= (\nabla q_h, \mathbf{v}_q - \pi_{h,k} \mathbf{v}_q) - \langle q_h, (\pi_{h,k} \mathbf{v}_q) \cdot \mathbf{n} \rangle_{\partial\Omega} \\ &\quad + (q_h, \nabla \cdot \pi_{h,k} \mathbf{v}_q) \\ &= (\nabla q_h, \mathbf{v}_q - \pi_{h,k} \mathbf{v}_q) - b_h(q_h, \pi_{h,k} \mathbf{v}_q). \end{aligned} \tag{3.7}$$

Since $q_h \in C_{h,k}^1$, it follows that $\nabla q_h \in [V_h^k]^d$. Thus, using the orthogonality of the L^2-projection, we have

$$(\nabla q_h, \mathbf{v}_q - \pi_{h,k} \mathbf{v}_q) = 0.$$

Thus, from (3.7), it follows that

$$|b_h(q_h, \pi_{h,k} \mathbf{v}_q)| = \|q_h\|_{0,\Omega}^2.$$

In addition, using H^1-stability of the L^2-projection (see [5]) and (3.6), we have

$$\begin{aligned} \|\pi_{h,k} \mathbf{v}_q\|_{1,\Omega} &\leq C \|\mathbf{v}_q\|_{1,\Omega} \\ &\leq C \|q_h\|_{0,\Omega}, \end{aligned}$$

which completes the proof. ∎

We now state the well-posedness of the discrete problem.

THEOREM 3.5. *The discrete problem* (2.10) *has a unique solution.*

Proof. Problem (2.10) can be written, in operator form, as

$$\begin{aligned} A\mathbf{u}_h + B^T p_h &= M\mathbf{f} \quad \text{in} \quad \left([V_h^k]^d\right)', \\ B\mathbf{u}_h &= Jp_h \quad \text{in} \quad [Q_h^k]', \end{aligned} \tag{3.8}$$

with $A \in \mathcal{L}([V_h^k]^d, ([V_h^k]^d)')$, $M \in \mathcal{L}([L^2(\Omega)]^d, ([V_h^k]^d)')$, $B \in \mathcal{L}\left([V_h^k]^d, (Q_h^k)'\right)$, and $J \in \mathcal{L}\left(Q_h^k, (Q_h^k)'\right)$ defined by

$$\begin{aligned} \langle A\mathbf{u}_h, \mathbf{v}_h \rangle &\stackrel{\text{def}}{=} a_h(\mathbf{u}_h, \mathbf{v}_h) + j_{\mathbf{u}}(\mathbf{u}_h, \mathbf{v}_h), \\ \langle M\mathbf{f}, \mathbf{v}_h \rangle &\stackrel{\text{def}}{=} (\mathbf{f}, \mathbf{v}_h), \\ \langle B\mathbf{v}_h, q_h \rangle &\stackrel{\text{def}}{=} b_h(q_h, \mathbf{v}_h), \\ \langle Jp_h, q_h \rangle &\stackrel{\text{def}}{=} j(p_h, q_h). \end{aligned}$$

We also introduce the operator $B^1 \in \mathcal{L}([V_h^k]^d, (C_{h,k}^1)')$ defined by

$$\langle B^1 \mathbf{v}_h, q_h \rangle \stackrel{\text{def}}{=} b_h(q_n, \mathbf{v}_h) \quad \forall (\mathbf{v}_h, q_h) \in [V_h^k]^d \times C_{h,k}^1;$$

in other words,

$$B^1 \mathbf{v}_h \stackrel{\text{def}}{=} (B\mathbf{v}_h)_{|C_{h,k}^1} \quad \forall \mathbf{v}_h \in [V_h^k]^d.$$

From Lemma 3.4, it follows that B^1 is surjective and $(B^1)^{\mathrm{T}}$ is injective (see [27, p. 58]). We then deduce that $V_{h,k}^{\mathrm{div}} \stackrel{\text{def}}{=} \mathrm{Ker}(B^1) \neq \{0\}$.

Let us consider the following reduced formulation (derived from (2.10) with $(\mathbf{v}_h, q_h) \in V_{h,k}^{\mathrm{div}} \times (Q_h^k \backslash C_{h,k}^1)$): find $(\mathbf{u}_h, \tilde{p}_h) \in V_{h,k}^{\mathrm{div}} \times (Q_h^k \backslash C_{h,k}^1)$ such that

$$\begin{aligned} A\mathbf{u}_h + B^T \tilde{p}_h &= M\mathbf{f} \quad \text{in} \quad \left(V_{h,k}^{\mathrm{div}}\right)', \\ B\mathbf{u}_h &= J\tilde{p}_h \quad \text{in} \quad \left(Q_h^k \backslash C_{h,k}^1\right)'. \end{aligned} \tag{3.9}$$

Since, by construction, $C_{h,k}^1 = \mathrm{Ker}(J)$, we conclude that J is invertible in $Q_h^k \backslash C_{h,k}^1$. Hence, from (3.9), we have

$$\tilde{p}_h = J^{-1}_{|Q_h^k \backslash C_{h,k}^1} B\mathbf{u}_h. \tag{3.10}$$

By plugging this expression into the first equation of (3.9), we obtain that $\mathbf{u}_h \in V_{h,k}^{\mathrm{div}}$ solves

$$\left(A + B^T J^{-1}_{|Q_h^k \backslash C_{h,k}^1} B\right) \mathbf{u}_h = M\mathbf{f} \quad \text{in} \quad \left(V_{h,k}^{\mathrm{div}}\right)'.$$

Existence and uniqueness of $\mathbf{u}_h$ follow by the positivity of A (Lemma 3.3) and the nonnegativity of $B^T J^{-1}_{|Q_h^k \backslash C_{h,k}^1} B$. We may then recover $\tilde{p}_h$ uniquely from (3.10). Therefore, the reduced problem (3.9) has a unique solution. On the other hand, from the first equation of (3.9), it follows that

$$A\mathbf{u}_h + B^T \tilde{p}_h - M\mathbf{f} \in \left(\mathrm{Ker}(B^1)\right)^0,$$

with $\left(\mathrm{Ker}(B^1)\right)^0$ standing for the *polar set* of $\mathrm{Ker}(B^1)$. From Lemma 3.4, it follows that B^1 is an isomorphism from $C_{h,k}^1$ onto $\left(\mathrm{Ker}(B^1)\right)^0$ (see [27, p. 58]). Thus, there exists a unique $p^1 \in C_{h,k}^1$ such that

$$A\mathbf{u}_h + B^T \tilde{p}_h - M\mathbf{f} = (B^1)^{\mathrm{T}} p^1 \quad \text{in} \quad \left([V_h^k]^d\right)'. \tag{3.11}$$

Therefore, from (3.11) and (3.9), and by noticing that $(B^1)^{\mathrm{T}} p^1 = B^{\mathrm{T}} p^1$ and $Jp^1 = 0$, it follows that problem (2.10) has a unique solution, given by $(\mathbf{u}_h, p_h \stackrel{\text{def}}{=} \tilde{p}_h - p^1)$. ∎

4. Convergence of the method. The parameter for the pressure stabilization scales as $h_K^2/\|\boldsymbol{\beta}\|_{0,\infty,K}$ when the local Reynolds number Re_K is big, and as h_K^3/ν when Re_K is small. The stabilizing terms acting on the velocity scale as $\|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2$ at a high local Reynolds number and as $\mathrm{Re}_K \|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2$ for a low Reynolds number. The factor $\mathrm{Re}_K \|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2$ in the velocity stabilization may be omitted in the low Reynolds regime without perturbing the convergence. We will now show that this scaling gives optimal a priori error estimates in the high (local) Reynolds number

regime when the solution is smooth, $(\mathbf{u}, p) \in [H^{k+1}(\Omega)]^{d+1}$, and in the low (local) Reynolds number regime under standard regularity assumptions. We then prove, using the Aubin–Nitsche duality technique (see, e.g., [21]), that the velocities have optimal convergence order also in the L^2-norm, when the local Reynolds number is low, without any modification of the stabilization.

First, we summarize some stability properties of the L^2-projection with weighted norms and show an approximability result for the triple norm (3.3).

REMARK 4.1. *In the remainder of this section, $C > 0$ stands for a generic constant independent of h and the physical parameters.*

To prove approximability for the L^2-projection on locally quasi-uniform meshes we need some additional stability for the L^2-projection from [3] that we state here without proof.

LEMMA 4.2. *For $\rho, \eta > 0$ sufficiently small and for all $\phi \in V_h^1$ satisfying*

$$\phi > 0, \quad |\nabla \phi(\mathbf{x})| \leq \eta h_K^{-1} \phi(\mathbf{x}) \quad \forall \mathbf{x} \in K, \quad \forall K \in \mathcal{T}_h,$$

there holds

$$\|\phi \pi_{h,k} u\|_{0,\Omega} \leq C \|\phi u\|_{0,\Omega} \quad \forall u \in L^2(\Omega),$$
$$\|\phi \nabla \pi_{h,k} u\|_{0,\Omega} \leq C \|\phi \nabla u\|_{0,\Omega} \quad \forall u \in H^1(\Omega).$$

A direct consequence of this result is stated in the following corollary.

COROLLARY 4.3. *Under the assumptions of the previous lemma, we have*

$$\left(\sum_{|\boldsymbol{\alpha}| \leq l} \|\phi \partial^{\boldsymbol{\alpha}} (u - \pi_{h,k} u)\|_{0,\Omega}^2 \right)^{\frac{1}{2}} \leq C \left(\sum_{K \in \mathcal{T}_h} \|\phi\|_{0,\infty,K}^2 h^{2(r_u - l)} \|u\|_{r_u,\Omega}^2 \right)^{\frac{1}{2}}$$

for all $u \in H^r(\Omega)$, with $r \geq 1$, $r_u \stackrel{\text{def}}{=} \min\{r, k+1\}$, $0 \leq l \leq r_u$, $\boldsymbol{\alpha} \in \mathbb{N}^d$, and ∂^α the standard multi-index notation for high order derivatives.

In order to obtain localized estimates we now show that the weights appearing in our stabilization allow for L^2-stability.

LEMMA 4.4. *Let $\phi_i \in H^2(\mathcal{T}_h)$, $i = 1, \ldots, 5$, be piecewise constant functions defined by*

$$\phi_{1|K} \stackrel{\text{def}}{=} \nu^{-\frac{1}{2}} \min\left\{ \mathrm{Re}_K^{-\frac{1}{2}}, 1 \right\}, \quad \phi_{2|K} \stackrel{\text{def}}{=} \|\boldsymbol{\beta}\|_{0,\infty,K}^{\frac{1}{2}} h_K^{-\frac{1}{2}},$$
$$\phi_{3|K} \stackrel{\text{def}}{=} h_K^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,K}^{\frac{1}{2}} \xi(\mathrm{Re}_K)^{-\frac{1}{2}}, \quad \phi_{4|K} \stackrel{\text{def}}{=} \phi_{3|K}^{-1}, \quad \phi_{5|K} \stackrel{\text{def}}{=} h_K^{-r},$$

with $r \geq 1$, for all $K \in \mathcal{T}_h$, and let $\phi_i^* \stackrel{\text{def}}{=} \pi_{h,1}^* \phi_i$. Then, there holds

$$\left.\begin{aligned} \phi_i (\rho_{\boldsymbol{\beta}} \rho)^{-\frac{1}{2}} \leq \phi_i^* \leq (\rho_{\boldsymbol{\beta}} \rho)^{\frac{1}{2}} \phi_i \\ |\nabla \phi_i^*| \leq c_0 (\rho_{\boldsymbol{\beta}} \rho - 1) h_K^{-1} \phi_i^* \end{aligned}\right\} \quad \text{for} \quad i = 1, 2,$$

$$\left.\begin{aligned} \phi_i \rho_{\boldsymbol{\beta}}^{-\frac{1}{2}} \rho^{-1} \leq \phi_i^* \leq \rho_{\boldsymbol{\beta}}^{\frac{1}{2}} \rho \phi_i \\ |\nabla \phi_i^*| \leq c_0 (\rho_{\boldsymbol{\beta}} \rho^2 - 1) h_K^{-1} \phi_i^* \end{aligned}\right\} \quad \text{for} \quad i = 3, 4,$$

$$\left.\begin{aligned} \phi_5 \rho^{-r} \leq \phi_5^* \leq \rho^r \phi_5, \\ |\nabla \phi_5^*| \leq c_0 (\rho^{2r} - 1) h_K^{-1} \phi_5^* \end{aligned}\right.$$

in K, for all $K \in \mathcal{T}_h$, with $c_0 > 0$ the constant in (2.4).

Proof. We give the proof only for ϕ_1; the argument for the rest is similar. First, note that for all $K \in \mathcal{T}_h$,

$$\begin{aligned}\max_{x\in K}\phi_1^* &\le \max_{K'\in\mathcal{N}(K)} \nu^{-\frac{1}{2}} \min\{\mathrm{Re}_{K'}^{-\frac{1}{2}}, 1\} \\ &= \max_{K'\in\mathcal{N}(K)} \min\{\|\boldsymbol{\beta}\|_{0,\infty,K'}^{-\frac{1}{2}} h_{K'}^{-\frac{1}{2}}, \nu^{-\frac{1}{2}}\}.\end{aligned} \tag{4.1}$$

We now distinguish two cases. On one hand, if $\mathrm{Re}_K \le 1$, we have $\phi_{1|K} = \nu^{-\frac{1}{2}}$. Thus, from (4.1) and since $\rho_{\boldsymbol{\beta}}\rho > 1$, it follows that

$$\begin{aligned}\max_{x\in K}\phi_1^* &\le \nu^{-\frac{1}{2}} \\ &\le (\rho\rho_\beta)^{\frac{1}{2}}\phi_{1|K}.\end{aligned}$$

On the other hand, if $\mathrm{Re}_K > 1$, we get $\phi_{1|K} = \|\boldsymbol{\beta}\|_{0,\infty,K}^{-\frac{1}{2}} h_K^{-\frac{1}{2}}$. Therefore, from (4.1) and the assumptions on the mesh (2.5) and (2.6), we have

$$\begin{aligned}\max_{x\in K}\phi_1^* &\le \max_{K'\in\mathcal{N}(K)} \{\|\boldsymbol{\beta}\|_{0,\infty,K'}^{-\frac{1}{2}} h_{K'}^{-\frac{1}{2}}\} \\ &\le (\rho\rho_\beta)^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,K}^{-\frac{1}{2}} h_K^{-\frac{1}{2}} \\ &= (\rho\rho_\beta)^{\frac{1}{2}}\phi_{1|K}.\end{aligned}$$

The lower bound follows in a similar fashion.

Finally, for the derivative, using the bounds on ϕ_1^* and the regularity of the mesh (2.4), and since $\rho_{\boldsymbol{\beta}}\rho > 1$, we obtain

$$\begin{aligned}|\nabla\phi_{1|K}^*| &\le \frac{\max_{x\in K}\phi_1^* - \min_{x\in K}\phi_1^*}{\rho_K} \\ &\le \frac{(\rho_{\boldsymbol{\beta}}\rho)^{\frac{1}{2}} - (\rho_{\boldsymbol{\beta}}\rho)^{-\frac{1}{2}}}{\rho_K}\phi_{1|K} \\ &\le c_0(\rho_{\boldsymbol{\beta}}\rho - 1)h_K^{-1}\phi_1^*,\end{aligned}$$

which completes the proof. ∎

REMARK 4.5. *It follows from Lemma* 4.4 *that for the weight functions* ϕ_i^*, $1 \le i \le 5$, *the stability estimate of Lemma* 4.2 *holds, provided* $\rho_{\boldsymbol{\beta}}$ *and* ρ *are sufficiently close to* 1. *From now on we assume that this is the case.*

The following lemma states the approximation properties of the L^2-projection in the triple norm $|\!|\!| \cdot |\!|\!|$.

LEMMA 4.6 (velocity approximability). *Assume that* $\rho_{\boldsymbol{\beta}}$ *and* ρ *are sufficiently close to* 1. *Then, there holds*

$$|\!|\!|\mathbf{u} - \pi_{h,k}\mathbf{u}|\!|\!|^2 \le C \sum_{K\in\mathcal{T}_h} \left(\sigma h_K^{2r_{\mathbf{u}}} + \max\{\nu, \|\boldsymbol{\beta}\|_{0,\infty,K} h_K\}\, h_K^{2(r_{\mathbf{u}}-1)}\right) \|\mathbf{u}\|_{r_{\mathbf{u}},K}^2$$

for all $\mathbf{u} \in [H^r(\Omega)]^d$, *with* $r \ge 2$ *and* $r_{\mathbf{u}} = \min\{k+1, r\}$.

Proof. First note that

$$|\!|\!|\mathbf{u} - \pi_{h,k}\mathbf{u}|\!|\!|^2 \le |\!|\!|i_{h,k}\mathbf{u} - \pi_{h,k}\mathbf{u}|\!|\!|^2 + |\!|\!|i_{h,k}\mathbf{u} - \mathbf{u}|\!|\!|^2.$$

We give the proof for the first term only. The argument for the second term is similar. By the stability estimate for the L^2-projection Lemma 4.2 we have

$$\begin{aligned}\|\sigma^{\frac{1}{2}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,\Omega}^2 &\le C\|\sigma^{\frac{1}{2}}(\mathbf{u}-i_{h,k}\mathbf{u})\|_{0,\Omega}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\sigma h_K^{2r_{\mathbf{u}}}\|\mathbf{u}\|_{r_{\mathbf{u}},K}^2.\end{aligned}$$

Using now the H^1-stability of the L^2-projection on locally quasi-uniform meshes (see [5]) we get

$$\begin{aligned}\|\nu^{\frac{1}{2}}\nabla(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,\Omega}^2 &\le C\|\nu^{\frac{1}{2}}\nabla(i_{h,k}\mathbf{u}-\mathbf{u})\|_{0,\Omega}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\nu h^{2(r_{\mathbf{u}}-1)}\|\mathbf{u}\|_{r_{\mathbf{u}},\Omega}.\end{aligned}$$

We treat the boundary terms using the trace inequality (2.8) in combination with Lemma 4.2 and approximation, which yields

$$\begin{aligned}(4.2)\quad &\|\max\{|\boldsymbol{\beta}|,\nu/\tilde{h}\}^{\frac{1}{2}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,\partial\Omega}^2\\ &\le C\sum_{\substack{K\in\mathcal{T}_h\\ K\cap\partial\Omega\neq\emptyset}}\|\max\{\|\boldsymbol{\beta}\|_{0,\infty,K},\nu h_K^{-1}\}^{\frac{1}{2}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,K\cap\partial\Omega}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\|h_K^{-\frac{1}{2}}\max\{\|\boldsymbol{\beta}\|_{0,\infty,K},\nu h_K^{-1}\}^{\frac{1}{2}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,K}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\|\phi_3^*(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,K}^2\\ &= C\sum_{K\in\mathcal{T}_h}\|\phi_3^*\pi_{h,k}(i_{h,k}\mathbf{u}-\mathbf{u})\|_{0,K}^2\\ &\le C\sum_K\max\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K,\nu\}h_K^{2(r_{\mathbf{u}}-1)}\|\mathbf{u}\|_{r_{\mathbf{u}},K}^2.\end{aligned}$$

The interior penalty terms are treated in the same fashion as the boundary terms. We have

$$\begin{aligned}(4.3)\quad j_{\mathbf{u}}(\mathbf{u}-\pi_{h,k}\mathbf{u},\mathbf{u}-\pi_{h,k}\mathbf{u}) &\le j_{\mathbf{u}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u},i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\\ &\quad + j_{\mathbf{u}}(\mathbf{u}-i_{h,k}\mathbf{u},\mathbf{u}-i_{h,k}\mathbf{u}).\end{aligned}$$

The first term in this inequality can be estimated using that $\xi(\mathrm{Re}_K)\le 1$, the trace inequality (2.8), an inverse inequality, and the H^1-stability of the L^2-projection (see [5]), which yields

(4.4)

$$\begin{aligned}j_{\mathbf{u}}(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u},i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u}) &\le C\sum_{K\in\mathcal{T}_h}\|\boldsymbol{\beta}\|_{0,\infty,K}\xi(\mathrm{Re}_K)h_K^2\|\nabla(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,\partial K}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\|\boldsymbol{\beta}\|_{0,\infty,K}h_K^{-1}\|i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u}\|_{0,K}^2\\ &\le C\|\phi_2^*(i_{h,k}\mathbf{u}-\pi_{h,k}\mathbf{u})\|_{0,\Omega}^2\\ &\le C\sum_K\|\boldsymbol{\beta}\|_{0,\infty,K}h_K^{2r_{\mathbf{u}}-1}\|\mathbf{u}\|_{r_{\mathbf{u}},K}^2,\end{aligned}$$

and so the proof is finished. ∎

For the pressure, we have the following result.

LEMMA 4.7 (pressure approximability). *Under the assumptions of Lemma* 4.6*, there holds*

$$\|\tilde{h}^{\frac{1}{2}}\phi_1^*(p-\pi_{h,k}p)\|_{0,\partial\Omega}^2+\|\phi_1^*(p-\pi_{h,k}p)\|_{0,\Omega}^2+j(\pi_{h,k}p,\pi_{h,k}p)$$
$$\leq C\sum_{K\in\mathcal{T}_h}\min\{\|\boldsymbol{\beta}\|_{0,\infty,K}^{-1},h_K/\nu\}h_K^{2s_p-1}\|p\|_{s_p,K}^2$$

for all $p\in H^s(\Omega)$ *with* $s\geq 1$ *and* $s_p\overset{\text{def}}{=}\min\{k+1,s\}$.

Proof. As p may be only $H^1(\Omega)$, we must replace the nodal interpolant by the Clément interpolant in the analysis. The proof for the first term is similar to (4.2), by replacing $i_{h,k}$ by $\mathcal{C}_{h,k}$. The estimate for the second term follows from Corollary 4.3. Finally, for the interior penalty term, since $[\![\mathcal{C}_{h,k}\nabla p]\!]=0$ and using a trace inequality followed by an inverse inequality, we have

$$\begin{aligned}
j(\pi_{h,k}p,\pi_{h,k}p)&=\sum_{K\in\mathcal{T}_h}\xi(\mathrm{Re_K})\frac{h_K^2}{\|\boldsymbol{\beta}\|_{0,\infty,K}}|[\![\nabla\pi_{h,k}p-\mathcal{C}_{h,k}\nabla p]\!]|^2\\
&\leq C\sum_{K\in\mathcal{T}_h}\xi(\mathrm{Re_K})\frac{h_K}{\|\boldsymbol{\beta}\|_{0,\infty,K}}\|\nabla\pi_{h,k}p-\mathcal{C}_{h,k}\nabla p\|_{0,K}^2\\
&\leq C(\|\phi_4^*\nabla\pi_{h,k}(p-\mathcal{C}_{h,k}p)\|_{0,K}^2+\|\phi_4^*\nabla(p-\mathcal{C}_{h,k}p)\|_{0,K}^2\\
&\quad+\|\phi_4^*(\nabla p-\mathcal{C}_{h,k}\nabla p)\|_{0,K}^2)\\
&\leq C\sum_{K\in\mathcal{T}_h}\min\{\|\boldsymbol{\beta}\|_{0,\infty,K}^{-1}h_K,h_K^2/\nu\}h_K^{2(s_p-1)}\|p\|_{s_p,K}^2,
\end{aligned}$$

where we concluded using the stability lemma, Lemma 4.2, with weight function ϕ_4^*, and the optimal approximation properties of the Clément interpolant (see [16, 21]). ∎

4.1. Energy norm error estimate. In this section we prove convergence in the triple norm. These results are optimal independently of the local Reynolds number when the exact solution is sufficiently smooth.

We start by proving a technical lemma.

LEMMA 4.8. *For all* $\mathbf{v}_h\in[V_h^k]^d$*, there holds*

$$\sum_{K\in\mathcal{T}_h}h_K^2\int_{\partial K}\|\boldsymbol{\beta}_h\cdot\mathbf{n}\|_{0,\infty,\partial K}|[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]|^2\,\mathrm{d}s\leq C\left(j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h)+\|\nu^{\frac{1}{2}}\nabla\mathbf{v}_h\|_{0,\Omega}^2\right),$$

$$\sum_{K\in\mathcal{T}_h}\phi_{1|K}^{-1}h_K\int_{\partial K}[\![\nabla\cdot\mathbf{v}_h]\!]^2\,\mathrm{d}s\leq C\left(j_{\mathbf{u}}(\mathbf{v}_h,\mathbf{v}_h)+\|\nu^{\frac{1}{2}}\nabla\mathbf{v}_h\|_{0,\Omega}^2\right).$$

Proof. Let A_1 denote the set of elements $K\in\mathcal{T}_h$ such that $\xi(\mathrm{Re}_K)\geq 1$, and A_2 the set of elements such that $\xi(\mathrm{Re}_K)<1$. It then follows that $|\beta|_{\infty,0,K}h_K<\nu$ for $K\in A_2$, and we may write

$$\sum_{K\in\mathcal{T}_h} h_K^2 \int_{\partial K} \|\boldsymbol{\beta}_h\cdot\mathbf{n}\|_{0,\infty,\partial K} |[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]|^2\,\mathrm{d}s$$
$$\le \sum_{K\in A_1} h_K^2 \int_{\partial K} \|\boldsymbol{\beta}_h\cdot\mathbf{n}\|_{0,\infty,\partial K} |[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]|^2\,\mathrm{d}s + \sum_{K\in A_2} h_K\nu \int_{\partial K} |[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]|^2\,\mathrm{d}s$$
$$\le \sum_{K\in\mathcal{T}_h} \int_{\partial K} h_K^2 \xi(\mathrm{Re}_K) \|\boldsymbol{\beta}_h\cdot\mathbf{n}\|_{0,\infty,\partial K} |[\![\mathbf{n}\cdot\nabla\mathbf{v}_h]\!]|^2\,\mathrm{d}s + C\|\nu^{\frac{1}{2}}\nabla\mathbf{v}_h\|_{0,\Omega}^2,$$

where the last inequality follows by a trace inequality, an inverse inequality in the second term, and extending the sums over all $\mathcal{T}_h$.

The second inequality follows in a similar fashion, noting that

$$\sum_{K\in\mathcal{T}_h} \phi_{1|K}^{-1} h_K \int_{\partial K} [\![\nabla\cdot\mathbf{v}_h]\!]^2\,\mathrm{d}s \le C \sum_{K\in\mathcal{T}_h} h_K \max\{\|\boldsymbol{\beta}\|_{0,\infty,K} h_K, \nu\} \int_{\partial K} [\![\nabla\cdot\mathbf{v}_h]\!]^2\,\mathrm{d}s$$
$$\le \sum_{K\in A_1} h_K^2 \|\boldsymbol{\beta}\|_{0,\infty,K} \xi(\mathrm{Re}_K) \int_{\partial K} [\![\nabla\cdot\mathbf{v}_h]\!]^2\,\mathrm{d}s + \|\nu^{\frac{1}{2}}\nabla\mathbf{v}_h\|_{0,A_2}^2,$$

and so the proof is completed. ∎

The main result of this paragraph is stated in the following theorem.

THEOREM 4.9. *Assume* $(\mathbf{u},p)\in[H^r(\Omega)]^d\times H^s(\Omega)$, *with* $r\ge 2$ *and* $s\ge 1$, *is the solution of* (2.1) *and* $(\mathbf{u}_h,p_h)\in W_h^k$ *is the solution of* (2.10). *Then, under the assumptions of Lemma* 4.6, *there holds*

$$|\!|\!|\mathbf{u}-\mathbf{u}_h|\!|\!| \le C\left[\sum_{K\in\mathcal{K}} \left(\sigma h_K^{2r_{\mathbf{u}}} + \max\{\|\boldsymbol{\beta}\|_{0,\infty,K} h_K,\nu\} h_K^{2(r_{\mathbf{u}}-1)}\right) \|\mathbf{u}\|_{r_{\mathbf{u}},K}^2\right]^{\frac{1}{2}}$$
$$+C\max_{K\in\mathcal{T}_h}\left\{\sigma^{-\frac{1}{2}}|\boldsymbol{\beta}|_{1,\infty,K} h_K^{r_{\mathbf{u}}}\right\}\|\mathbf{u}\|_{r_{\mathbf{u}},\Omega} + C\left(\sum_{K\in\mathcal{T}_h} \min\{\|\boldsymbol{\beta}\|_{0,\infty,K}^{-1}, h_K/\nu\} h_K^{2s_p-1} \|p\|_{s_p,K}^2\right)^{\frac{1}{2}},$$

with $r_{\mathbf{u}} = \min\{k+1,r\}$ *and* $s_p = \min\{k+1,s\}$.

Proof. Let us decompose the error $\mathbf{u}-\mathbf{u}_h$ in two parts:

$$\mathbf{u}-\mathbf{u}_h = \underbrace{\mathbf{u}-\pi_{h,k}\mathbf{u}}_{\mathbf{e}^\pi} + \underbrace{\pi_{h,k}\mathbf{u}-\mathbf{u}_h}_{-\mathbf{e}_h} = \mathbf{e}^\pi - \mathbf{e}_h.$$

We also consider the discrete pressure error

$$y_h \overset{\mathrm{def}}{=} p_h - \pi_{h,k}p. \tag{4.5}$$

It follows then that

$$|\!|\!|\mathbf{u}-\mathbf{u}_h|\!|\!| \le |\!|\!|\mathbf{e}^\pi|\!|\!| + |\!|\!|\mathbf{e}_h|\!|\!|.$$

Lemma 4.6 gives an estimate for $|\!|\!|\mathbf{e}^\pi|\!|\!|$. Hence, it suffices to estimate $|\!|\!|\mathbf{e}_h|\!|\!|$.

Using coercivity and orthogonality, namely, Lemmas 3.3 and 2.1, we get

$$\begin{aligned} C|\!|\!|\mathbf{e}_h|\!|\!|^2 + j_p(y_h,y_h) &\le a_h(\mathbf{e}_h,\mathbf{e}_h) + j_{\mathbf{u}}(\mathbf{e}_h,\mathbf{e}_h) + j_p(y_h,y_h) \\ &= a_h(\mathbf{e}^\pi,\mathbf{e}_h) + b_h(y^\pi,\mathbf{e}_h) - b_h(y_h,\mathbf{e}^\pi) \\ &\quad + j_{\mathbf{u}}(\mathbf{e}^\pi,\mathbf{e}_h) - j_p(\pi_{h,k}p,y_h). \end{aligned} \tag{4.6}$$

By an application of the Cauchy–Schwarz inequality in the symmetric part of the discrete elliptic operator and integrating by parts in the convective term, we obtain

$$a_h(\mathbf{e}^\pi,\mathbf{e}_h) \le |\!|\!|\mathbf{e}^\pi|\!|\!|\,|\!|\!|\mathbf{e}_h|\!|\!| + |(\mathbf{e}^\pi,\boldsymbol{\beta}\cdot\nabla\mathbf{e}_h)| \\ - \langle 2\nu\boldsymbol{\epsilon}(\mathbf{e}^\pi)\mathbf{n},\mathbf{e}_h\rangle_{\partial\Omega} - \langle \mathbf{e}^\pi, 2\nu\boldsymbol{\epsilon}(\mathbf{e}_h)\mathbf{n}\rangle_{\partial\Omega},$$

where, for simplicity, the boundary term from the integration by parts has been included in the first term on the right-hand side. We note that in the same way we have, using the Cauchy–Schwarz inequality, a trace inequality and a local inverse inequality,

$$\langle 2\nu\boldsymbol{\epsilon}(\mathbf{e}_h)\mathbf{n},\mathbf{e}^\pi\rangle_{\partial\Omega} \le C|\!|\!|\mathbf{e}_h|\!|\!|\,|\!|\!|\mathbf{e}^\pi|\!|\!|. \tag{4.7}$$

For the second boundary term we use the Cauchy–Schwarz inequality followed by a trace inequality and an approximation argument, similar to (4.3)–(4.4), to obtain

$$\langle 2\nu\boldsymbol{\epsilon}(\mathbf{e}^\pi)\mathbf{n},\mathbf{e}_h\rangle_{\partial\Omega} \le C\left(\sum_{K\in\mathcal{T}_h}\nu h_K^{2(r_\mathbf{u}-1)}\|\mathbf{u}\|^2_{r_\mathbf{u},K}\right)^{\frac12}|\!|\!|\mathbf{e}_h|\!|\!|. \tag{4.8}$$

The convective term is controlled using a local inverse inequality, Lemma 4.4, Corollary 4.3, and the orthogonality of the L^2-projection, after having replaced the continuous velocity field $\boldsymbol{\beta}$ by its piecewise linear interpolant $\boldsymbol{\beta}_h$,

$$\begin{aligned}
|(\mathbf{e}^\pi,\boldsymbol{\beta}\cdot\nabla\mathbf{e}_h)| &\le |(\mathbf{e}^\pi,(\boldsymbol{\beta}-\boldsymbol{\beta}_h)\cdot\nabla\mathbf{e}_h)| + |(\mathbf{e}^\pi,\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h)| \\
&\le C\sum_{K\in\mathcal{T}_h}|\boldsymbol{\beta}|_{1,\infty,K}\|\mathbf{e}^\pi\|_{0,K}h_K\|\nabla\mathbf{e}_h\|_{0,K} \\
&\quad + \left|\left(\mathbf{e}^\pi,\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h - \pi^*_{h,k}(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h)\right)\right| \\
&\le C\sum_{K\in\mathcal{T}_h}\sigma^{-\frac12}|\boldsymbol{\beta}|_{1,\infty,K}h_K^{r_\mathbf{u}}\|\phi_5^*\mathbf{e}^\pi\|_{0,K}\|\sigma^{\frac12}\mathbf{e}_h\|_{0,K} \\
&\quad + \left|\left(\mathbf{e}^\pi,\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h - \pi^*_{h,k}(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h)\right)\right| \\
&\le C\max_{K\in\mathcal{T}_h}\{\sigma^{-\frac12}|\boldsymbol{\beta}|_{1,\infty,K}h_K^{r_\mathbf{u}}\}\|\mathbf{u}\|_{r_\mathbf{u},\Omega}|\!|\!|\mathbf{e}_h|\!|\!| \\
&\quad + \|\phi_2\mathbf{e}^\pi\|_{0,\Omega}\left\|\phi_2^{-1}\left(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h - \pi^*_{h,k}(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h)\right)\right\|_{0,\Omega}.
\end{aligned}$$

Now we apply Lemma 3.1 to obtain

$$\begin{aligned}
&\|\phi_2\mathbf{e}^\pi\|_{0,\Omega}\left\|\phi_2^{-1}\left(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h - \pi^*_{h,k}(\boldsymbol{\beta}_h\cdot\nabla\mathbf{e}_h)\right)\right\|_{0,\Omega} \\
&\qquad \le C\|\phi_2^*\mathbf{e}^\pi\|_{0,\Omega}\left(\sum_{K\in\mathcal{T}_h}\int_{\partial K}h_K^2\|\boldsymbol{\beta}\cdot\mathbf{n}\|_{0,\infty,\partial K}|[\![\mathbf{n}\cdot\nabla\mathbf{u}_h]\!]|^2\,\mathrm{d}s\right)^{\frac12} \\
&\qquad\qquad \le C\left(\sum_{K\in\mathcal{T}_h}\|\boldsymbol{\beta}\|_{0,\infty,K}h_K^{2r_\mathbf{u}-1}\|\mathbf{u}\|^2_{r_\mathbf{u},K}\right)^{\frac12}|\!|\!|\mathbf{e}_h|\!|\!|,
\end{aligned}$$

where we used Corollary 4.3 and Lemma 4.8 in the last inequality.

Collecting terms we have

$$\begin{aligned}
a_h(\mathbf{e}^\pi,\mathbf{e}_h) &\le C|\!|\!|\mathbf{e}^\pi|\!|\!|\,|\!|\!|\mathbf{e}_h|\!|\!| + C\max_{K\in\mathcal{T}_h}\{\sigma^{-\frac12}|\boldsymbol{\beta}|_{1,\infty,K}h_K^{r_\mathbf{u}}\}|\!|\!|\mathbf{e}_h|\!|\!| \\
&\quad + C\left(\sum_{K\in\mathcal{T}_h}\max\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K,\nu\}h_K^{2(r_\mathbf{u}-1)}\|\mathbf{u}\|^2_{r_\mathbf{u},K}\right)^{\frac12}|\!|\!|\mathbf{e}_h|\!|\!|.
\end{aligned} \tag{4.9}$$

For the second term in (4.6), using the orthogonality of the L^2-projection, Lemmas 4.7 and 4.8, and replacing $\mathbf{u}$ with p in (4.2), we have

$$\begin{aligned} b_h(y^\pi, \mathbf{e}_h) &= -(y^\pi, \nabla\cdot\mathbf{e}_h - \pi^*_{h,k}(\nabla\cdot\mathbf{e}_h)) + \langle y^\pi, \mathbf{e}_h\cdot\mathbf{n}\rangle_{\partial\Omega} \\ &\le \|\phi_1^* y^\pi\|_{0,\Omega}\|\phi_1^{-1}(\nabla\cdot\mathbf{e}_h - \pi^*_{h,k}(\nabla\cdot\mathbf{e}_h))\|_{0,\Omega} + \|\phi_1^*\tilde{h}^{\frac12} y^\pi\|_{0,\partial\Omega}|\!|\!|\mathbf{e}_h|\!|\!| \\ &\le C\left(\sum_{K\in\mathcal{T}_h}\min\{\|\boldsymbol{\beta}\|^{-1}_{0,\infty,K}, h_K/\nu\}h_K^{2s_p-1}\|p\|^2_{s_p,K}\right)^{\frac12}|\!|\!|\mathbf{e}_h|\!|\!|. \end{aligned} \tag{4.10}$$

In a similar fashion, after integration by parts in the third term, one obtains

$$\begin{aligned} b_h(y_h, \mathbf{e}^\pi) &= -(y_h, \nabla\cdot\mathbf{e}^\pi) + \langle y_h, \mathbf{e}^\pi\cdot\mathbf{n}\rangle_{\partial\Omega} \\ &= (\nabla y_h, \mathbf{e}^\pi) \\ &= (\nabla y_h - \pi^*_{h,k}(\nabla y_h), \mathbf{e}^\pi) \\ &\le C\|\phi_3^{-1}\big(\nabla y_h - \pi^*_{h,k}(\nabla y_h)\big)\|_{0,\Omega}\|\phi_3^*\mathbf{e}^\pi\|_{0,\Omega} \\ &\le C j_p(y_h, y_h)^{\frac12}\left(\sum_{K\in\mathcal{T}_h}\max\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K, \nu\}h_K^{2(r_{\mathbf{u}}-1)}\|\mathbf{u}\|^2_{r_{\mathbf{u}},K}\right)^{\frac12}. \end{aligned} \tag{4.11}$$

Finally, using Lemma 4.7, for the interior penalty terms we have

$$\begin{aligned} j_{\mathbf{u}}(\mathbf{e}^\pi, \mathbf{e}_h) + j_p(\pi_{h,k}p, y_h) &\le C|\!|\!|\mathbf{e}^\pi|\!|\!||\!|\!|\mathbf{e}_h|\!|\!| + j_p(\pi_{h,k}p, \pi_{h,k}p)^{\frac12} j_p(y_h, y_h)^{\frac12} \\ &\le C|\!|\!|\mathbf{e}^\pi|\!|\!||\!|\!|\mathbf{e}_h|\!|\!| + C j_p(y_h, y_h)^{\frac12}\left(\sum_{K\in\mathcal{T}_h}\min\{\|\boldsymbol{\beta}\|^{-1}_{0,\infty,K}, h_K/\nu\}h_K^{2s_p-1}\|p\|^2_{s_p,K}\right)^{\frac12}. \end{aligned} \tag{4.12}$$

We conclude the proof by collecting the results of (4.9)–(4.12) in (4.6) and applying the approximation lemma, Lemma 4.6. ∎

The following corollary follows from (4.6) in combination with (4.5) and Lemma 4.7.

COROLLARY 4.10. *Under the assumptions of Theorem* 4.9*, there holds*

$$\begin{aligned} j_p(p_h, p_h) \le{}& C\max_{K\in\mathcal{T}_h}\left\{\sigma^{-\frac12}|\boldsymbol{\beta}|_{1,\infty,K}h_K^{r_{\mathbf{u}}}\right\}\|\mathbf{u}\|_{r_{\mathbf{u}},\Omega} \\ &+ C\left[\sum_{K\in\mathcal{K}}\left(\sigma h_K^{2r_{\mathbf{u}}} + \max\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K, \nu\}h^{2(r_{\mathbf{u}}-1)}\right)\|\mathbf{u}\|^2_{r_{\mathbf{u}},K}\right]^{\frac12} \\ &+ C\left(\sum_{K\in\mathcal{T}_h}\min\{\|\boldsymbol{\beta}\|^{-1}_{0,\infty,K}, h_K/\nu\}h_K^{2s_p-1}\|p\|^2_{s_p,K}\right)^{\frac12}. \end{aligned}$$

REMARK 4.11. *In a physical situation the velocity gradient on boundaries with no-slip conditions is known to scale as* $|\boldsymbol{\beta}|_{1,\infty,\partial\Omega} \sim \nu^{-\frac12}$*. If in the boundary layer* $h_K \sim \nu$*, that is, a low local Reynolds number on the boundary, then the estimate is dominated by the* $H^1(\Omega)$ *contribution from the boundary that converges at optimal rate since the layer is resolved. The condition* (2.7) *is satisfied with* $c_{\boldsymbol{\beta}} \sim \nu^{\frac12}$ *showing that the strongest constraint on the mesh is not that of* (2.7)*, but that of the* $\frac{h_K}{\nu}$ *contribution on the boundary. In laminar free-flow we can expect* $|\boldsymbol{\beta}|_{1,\infty,K} \le c\|\boldsymbol{\beta}\|_{0,\infty,K}$ *to hold, and hence the convergence in the* L^2*-norm in this regime is of the quasi-optimal rate* $h^{k+\frac12}$ *for a sufficiently regular solution.*

4.2. Recovering the pressure. In this section, we provide an estimate of the L^2-norm of the pressure error. This is the aim of the following theorem, which ensures that the pressure converges at the rate of the velocity.

THEOREM 4.12. *Assume* $(\mathbf{u},p)\in[H^r(\Omega)]^d\times H^s(\Omega)$, *with* $r\geq 2$ *and* $s\geq 1$, *is the solution of* (2.1) *and* $(\mathbf{u}_h,p_h)\in W_h^k$ *is the solution of* (2.10). *Then, under the assumptions of Lemma* 4.6, *there holds*

$$\|p-p_h\|_{0,\Omega}\leq C\left(C_{\mathrm{L}}\sigma^{\frac{1}{2}}+\max_{K\in\mathcal{T}_h}\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K,\nu\}^{\frac{1}{2}}+\sigma^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\right)C_{\mathbf{u}},$$

with $C_{\mathbf{u}}$ *the convergence rate of* $|||\mathbf{u}-\mathbf{u}_h|||$ *given by Theorem* 4.9, *and* C_L *a positive constant depending only on* Ω.

Proof. Following [27, Corollary 2.4], there exists $\mathbf{v}_p\in[H_0^1(\Omega)]^d$ such that

$$\nabla\cdot\mathbf{v}_p=p-p_h,\quad \|\mathbf{v}_p\|_{0,\Omega}\leq C_{\mathrm{L}}\|p-p_h\|_{0,\Omega},\quad |\mathbf{v}_p|_{1,\Omega}\leq C\|p-p_h\|_{0,\Omega},\tag{4.13}$$

with $C_{\mathrm{L}}>0$ a constant, depending on Ω, which scales as a distance. Thus, using the modified Galerkin orthogonality (Lemma 2.1), we readily obtain

$$\begin{aligned}\|p-p_h\|_{0,\Omega}^2&=(p-p_h,\nabla\cdot\mathbf{v}_p)\\&=(p-p_h,\nabla\cdot(\mathbf{v}_p-\pi_{h,k}\mathbf{v}_p))+\langle p-p_h,\pi_{h,k}\mathbf{v}_p\cdot\mathbf{n}\rangle_{\partial\Omega}\\&\quad+a_h(\mathbf{u}-\mathbf{u}_h,\pi_{h,k}\mathbf{v}_p)+j_{\mathbf{u}}(\mathbf{u}-\mathbf{u}_h,\pi_{h,k}\mathbf{v}_p).\end{aligned}$$

Thus, after integrating by parts, we get

$$\|p-p_h\|_{0,\Omega}^2=\underbrace{(\nabla(p-p_h),\mathbf{v}_p-\pi_{h,k}\mathbf{v}_p)}_{T_1}+\underbrace{a_h(\mathbf{u}-\mathbf{u}_h,\pi_{h,k}\mathbf{v}_p)+j_{\mathbf{u}}(\mathbf{u}-\mathbf{u}_h,\pi_{h,k}\mathbf{v}_p)}_{T_2}.\tag{4.14}$$

For the first term, using the orthogonality of the L^2-projection, the Cauchy–Schwarz inequality, Corollary 3.2, (4.13), and Corollary 4.3, we get

$$\begin{aligned}T_1&=(\nabla(p-p_h)-\pi_{h,k}\nabla p+\pi_{h,k}^*\nabla p_h,\mathbf{v}_p-\pi_{h,k}\mathbf{v}_p)\\&\leq\|\tilde{h}(\nabla p-\pi_{h,k}\nabla p)\|_{0,\Omega}\|\tilde{h}^{-1}(\mathbf{v}_p-\pi_{h,k}\mathbf{v}_p)\|_{0,\Omega}\\&\quad+C\|\phi_3^{-1}(\nabla p_h-\pi_{h,k}^*\nabla p_h)\|_{0,\Omega}\|\phi_3^*(\mathbf{v}_p-\pi_{h,k}\mathbf{v}_p)\|_{0,\Omega}\\&\leq C\Bigg[\Bigg(\sum_{K\in\mathcal{T}_h}h_K^{2s_p-1}\|p\|_{s_p,K}^2\Bigg)^{\frac{1}{2}}\\&\quad+\max_{K\in\mathcal{T}_h}\{\nu,\|\boldsymbol{\beta}\|_{0,\infty,K}h_K\}^{\frac{1}{2}}j_p(p_h,p_h)^{\frac{1}{2}}\Bigg]\|p-p_h\|_{0,\Omega}.\end{aligned}\tag{4.15}$$

Using the definition (2.11) of the bilinear form a_h, and after integration by parts in the convective term, we have

$$\begin{aligned}T_2&\leq|||\mathbf{u}-\mathbf{u}_h|||\,|||\pi_{h,k}\mathbf{v}_p|||+(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\beta}\cdot\nabla\pi_{h,k}\mathbf{v}_p)\\&\quad-\langle 2\nu\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\mathbf{n},\pi_{h,k}\mathbf{v}_p\rangle_{\partial\Omega}-\langle\mathbf{u}-\mathbf{u}_h,2\nu\boldsymbol{\epsilon}(\pi_{h,k}\mathbf{v}_p)\mathbf{n}\rangle_{\partial\Omega}.\end{aligned}\tag{4.16}$$

For the convective term we have, using the H^1-stability of the L^2-projection (see [5]) and (4.13),

$$\begin{aligned}(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\beta}\cdot\nabla\pi_{h,k}\mathbf{v}_p)&\leq\sigma^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\sigma^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)\|_{0,\Omega}\|\nabla\pi_{h,k}\mathbf{v}_p\|_{0,\Omega}\\&\leq C\sigma^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}|||\mathbf{u}-\mathbf{u}_h|||\,\|p-p_h\|_{0,\Omega}.\end{aligned}\tag{4.17}$$

The boundary terms are controlled in the following fashion:

$$\begin{aligned}(4.18)\quad &\langle 2\nu\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\mathbf{n}, \pi_{h,k}\mathbf{v}_p\rangle_{\partial\Omega} + \langle \mathbf{u}-\mathbf{u}_h, 2\nu\boldsymbol{\epsilon}(\pi_{h,k}\mathbf{v}_p)\mathbf{n}\rangle_{\partial\Omega}\\ &\quad\le C\|(\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\|_{0,\partial\Omega}|\!|\!|\pi_{h,k}\mathbf{v}_p|\!|\!| + C\|(\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\pi_{h,k}\mathbf{v}_p)\|_{0,\partial\Omega}|\!|\!|\mathbf{u}-\mathbf{u}_h|\!|\!|.\end{aligned}$$

In addition, as in (4.7) and (4.8), we have

$$\begin{aligned}(4.19)\quad \|(\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\|_{0,\partial\Omega} &\le \|(\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{e}^{\pi})\|_{0,\partial\Omega} + \|(\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{e}_h)\|_{0,\partial\Omega},\\ &\le C\left[\left(\sum_{K\in\mathcal{T}_h}\nu h_K^{2(r_{\mathbf{u}}-1)}\|\mathbf{u}\|^2_{r_{\mathbf{u}},K}\right)^{\frac{1}{2}} + |\!|\!|\mathbf{e}_h|\!|\!|\right].\end{aligned}$$

In the same fashion, we obtain

$$(4.20)\qquad \|(2\nu\tilde{h})^{\frac{1}{2}}\boldsymbol{\epsilon}(\pi_{h,k}\mathbf{v}_p)\|_{0,\partial\Omega} \le C|\!|\!|\pi_{h,k}\mathbf{v}_p|\!|\!|.$$

Finally, from (4.13), it follows that

$$(4.21)\qquad |\!|\!|\pi_{h,k}\mathbf{v}_p|\!|\!| \le C\left(C_L\sigma^{\frac{1}{2}} + \max_{K\in\mathcal{T}_h}\{\|\boldsymbol{\beta}\|_{0,\infty,K}h_K,\nu\}^{\frac{1}{2}}\right)\|p-p_h\|_{0,\Omega}.$$

We conclude the proof by collecting the estimations (4.15)–(4.20) in (4.14), using (4.21) and Theorem 4.9. ∎

REMARK 4.13. *Let us notice that the three terms appearing in the error estimate of the previous theorem scale with the right dimensions.*

4.3. Low Reynolds number optimality. The following theorem gives an optimal L^2-error estimate for velocity when the local Reynolds number is low.

THEOREM 4.14. *Assume that the solution* $(\mathbf{u},p)$ *of* (2.1) *belongs to* $[H^2(\Omega)]^d \times H^1(\Omega)$ *and let* $(\mathbf{u}_h,p_h)\in W_h^k$ *be the solution of* (2.10). *Assume also that*

$$(4.22)\qquad \|\boldsymbol{\beta}\|_{0,\infty,K}h_K \le \nu \quad \forall K\in\mathcal{T}_h,$$

and that the solution (φ,ψ) *of the adjoint problem*

$$(4.23)\qquad \begin{cases}\sigma\varphi - \boldsymbol{\beta}\cdot\nabla\varphi - 2\nu\nabla\cdot\boldsymbol{\epsilon}(\varphi) - \nabla\psi = \mathbf{u}-\mathbf{u}_h \quad in \quad \Omega,\\ \nabla\cdot\varphi = 0 \quad in \quad \Omega,\\ \varphi = 0 \quad on \quad \partial\Omega\end{cases}$$

belongs to $[H^2(\Omega)]^d\times[H^1(\Omega)]$ *and satisfies*

$$(4.24)\qquad \|\varphi\|_{2,\Omega} + \|\psi\|_{1,\Omega} \le C\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}.$$

Then, there holds

$$\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega} \le ch^2\big(\|\mathbf{u}\|_{2,\Omega} + \|p\|_{1,\Omega}\big),$$

with constant $c>0$ *independent of* h, *but depending on the physical parameters.*

Proof. Multiplying the first equation of (4.23) by $\mathbf{u}-\mathbf{u}_h$ and the second by $-(p-p_h)$, integrating by parts, and using the modified Galerkin orthogonality (Lemma 2.1), it follows that

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}^2 &= a_h(\mathbf{u}-\mathbf{u}_h,\varphi)+b_h(p-p_h,\varphi)-b_h(\psi,\mathbf{u}-\mathbf{u}_h)\\ &= \underbrace{a_h(\mathbf{u}-\mathbf{u}_h,\varphi-\pi_{h,k}\varphi)+b_h(p-p_h,\varphi-\pi_{h,k}\varphi)-b_h(\psi-\pi_{h,k}\psi,\mathbf{u}-\mathbf{u}_h)}_{T_1}\\ &\quad+\underbrace{j_{\mathbf{u}}(\mathbf{u}-\mathbf{u}_h,\varphi-\pi_{h,k}\varphi)}_{T_2}+\underbrace{j_p(p_h,\pi_{h,k}\psi)}_{T_3}.\end{aligned}$$

Following the argument of the proofs of Theorems 4.9 and 4.12, and using Lemma 4.6 and (4.22), we get

$$\begin{aligned}T_1 &\le |||\mathbf{u}-\mathbf{u}_h|||\,|||\varphi-\pi_{h,k}\varphi||| + |(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\beta}\cdot\nabla(\varphi-\pi_{h,k}\varphi))|\\ &\quad-\langle 2\nu\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\mathbf{n},\varphi-\pi_{h,k}\varphi\rangle_{\partial\Omega}-\langle\mathbf{u}-\mathbf{u}_h,2\nu\boldsymbol{\epsilon}(\varphi-\pi_{h,k}\varphi)\mathbf{n}\rangle_{\partial\Omega}\\ &\le Ch\left(|||\mathbf{u}-\mathbf{u}_h|||\,\|\varphi\|_{2,\Omega}+\|p-p_h\|_{0,\Omega}\|\varphi\|_{2,\Omega}+|||\mathbf{u}-\mathbf{u}_h|||\,\|\psi\|_{1,\Omega}\right).\end{aligned}$$

Using Cauchy–Schwarz, Lemma 4.6, and (4.22), one obtains

$$\begin{aligned}T_2 &\le j_{\mathbf{u}}(\mathbf{u}-\mathbf{u}_h,\mathbf{u}-\mathbf{u}_h)^{\frac12}\,j_{\mathbf{u}}(\varphi-\pi_{h,k}\varphi,\varphi-\pi_{h,k}\varphi)^{\frac12}\\ &\le Ch^{\frac32}|||\mathbf{u}-\mathbf{u}_h|||\,\|\varphi\|_{2,\Omega}.\end{aligned}$$

Finally, from Lemma 4.7 and (4.22), for the last term we have

$$\begin{aligned}T_3 &\le j_p(p_h,p_h)^{\frac12}\,j_p(\pi_{h,k}\psi,\pi_{h,k}\psi)^{\frac12}\\ &\le Chj_p(p_h,p_h)^{\frac12}\,\|\psi\|_{1,\Omega}.\end{aligned}$$

The proof concludes by combining the above estimations with Theorems 4.9 and 4.12, Corollary 4.10, (4.22), and the assumed regularizing behavior (4.24). ∎

Let us sum up the results provided by Theorems 4.9 and 4.12. When the local Reynolds number is high and the solution is regular, we enjoy an optimal $O(h^{k+\frac12})$ convergence order of the error in the L^2-norm for the velocity and the pressure. For less regular solutions, for instance, when the pressure is in $H^1(\Omega)$ and the velocity is in $[H^2(\Omega)]^d$, we get an optimal $O(h)$ estimate in the energy norm, when the local Reynolds number is low, but a suboptimal estimate of $O(h^{\frac12})$ when the local Reynolds number is high. This is due to the fact that the inconsistencies in the pressure stabilization pollute the energy norm estimate for the velocities.

REMARK 4.15. *Note that by adding the L^2-coercivity, we can use the stabilization term to control the convective term without using the H^1-coercivity; this leads to a quasi-optimal estimate in the weaker L^2-norm, with a ν-weighted H^1 contribution showing that the stabilization handles the numerical instability induced by treating nonsymmetric terms using the standard Galerkin method. In case $\sigma=0$ the H^1 estimate obtained by a standard energy argument will scale as $\nu^{-\frac12}$, reflecting the physical instability of the problem.*

5. Numerical results. In this section we report several numerical experiments that show the good convergence properties of our stabilized finite element method. In particular, we recover the convergence rates obtained in section 4.

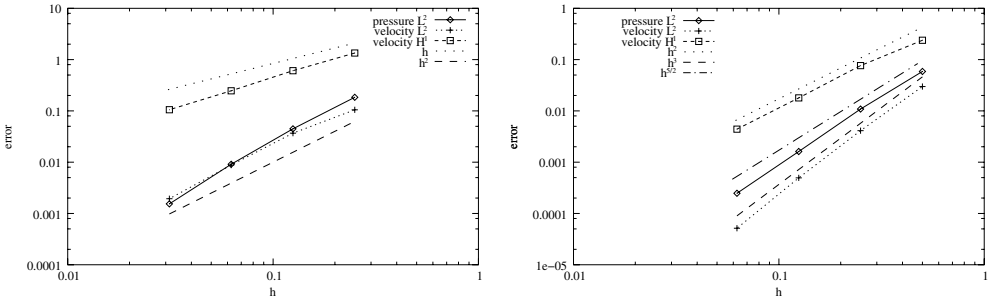


FIG. 5.1. *Convergence history: Linear elements* ($k = 1$) *(left) and quadratic elements* ($k = 2$) *(right).*

We consider problem (2.1) in three dimensions with nonhomogeneous boundary conditions. The right-hand side $\mathbf{f}$ and the boundary data are chosen in order to ensure that the exact solution of (2.1) is given by the following expression [22]:

$$
\begin{aligned}
u_1(x_1,x_2,x_3) &= be^{a(x_1-x_3)+b(x_2-x_3)} - ae^{a(x_3-x_2)+b(x_1-x_2)},\\
u_2(x_1,x_2,x_3) &= be^{a(x_2-x_1)+b(x_3-x_2)} - ae^{a(x_1-x_3)+b(x_2-x_3)},\\
u_3(x_1,x_2,x_3) &= be^{a(x_3-x_2)+b(x_1-x_2)} - ae^{a(x_2-x_1)+b(x_3-x_1)},\\
p(x_1,x_2,x_3) &= (a^2+b^2+ab)\Big[e^{a(x_1-x_2)+b(x_1-x_3)} + e^{a(x_2-x_3)+b(x_2-x_1)}\\
&\qquad + e^{a(x_3-x_1)+b(x_3-x_2)}\Big]
\end{aligned}
\tag{5.1}
$$

with $\boldsymbol{\beta} = \mathbf{u}$, $\sigma = 1$, $\nu = 10^{-4}$, $a = b = 0.75$, and $\Omega = (0,1)^3$ the unit cube.

The resulting continuous problem was solved approximately using the stabilized discrete formulation (2.10); however, the boundary conditions were strongly enforced. All numerical tests have been performed using conforming linear and quadratic finite elements for velocity and pressure, namely, P_1/P_1 and P_2/P_2 (implemented in a three-dimensional research code [23]). The stabilization parameter involved in the jumps terms (2.13) and (2.14) were chosen as

$$
\gamma = \begin{cases} \dfrac{1}{8} & \text{if} \quad k = 1, \\ \dfrac{1}{32} & \text{if} \quad k = 2. \end{cases}
$$

In Figure 5.1 we show, respectively, the velocity and pressure convergence histories for $k = 1$ and $k = 2$. Note that, in both cases, the numerical solution exhibits optimal convergence order and is hence in agreement with Theorems 4.9 and 4.12.

We show in Figure 5.2 the pressure contours in two different meshes (which are depicted in Figure 5.3) using linear elements. No spurious pressure oscillations are observed. We report in Figure 5.4 the contours of the second component of $\mathbf{u}_h$, u_{h2}, in the left plot with full stabilization and in the right plot setting the stabilization parameter for the term associated with the streamline derivative to zero, on the cutting plane $x = 0.5$. Although the exact solution is smooth, the plot of the unstabilized solution (right) exhibits spurious oscillations. Note that the spurious velocity oscillations (right) are completely controlled by the streamline-derivative jumps (left).

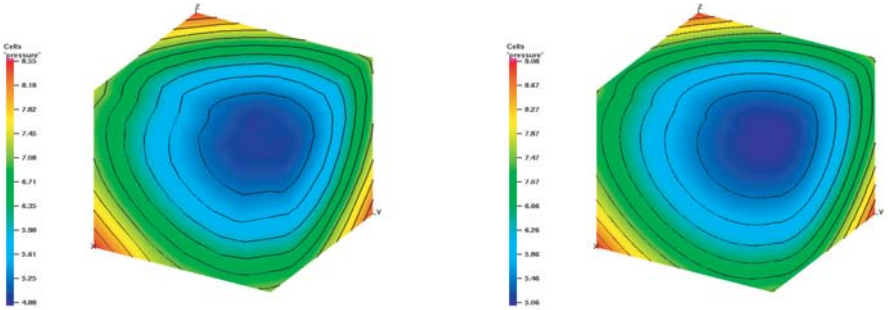


FIG. 5.2. *Pressure contours: Coarse mesh (left) and fine mesh (right).*

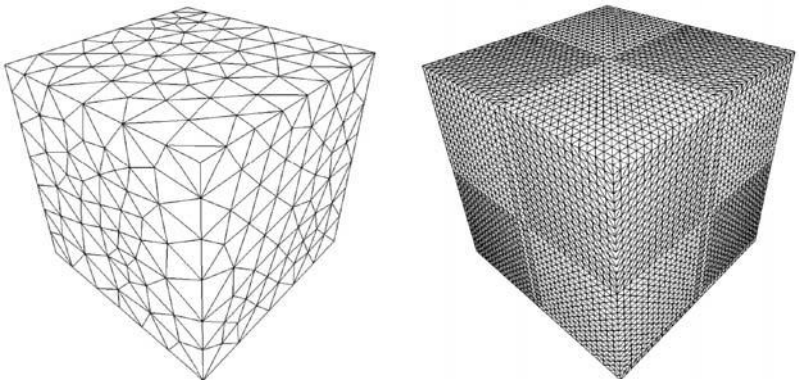


FIG. 5.3. *Coarse mesh (*2929 *tetrahedra) and fine mesh (*196608 *tetrahedra).*

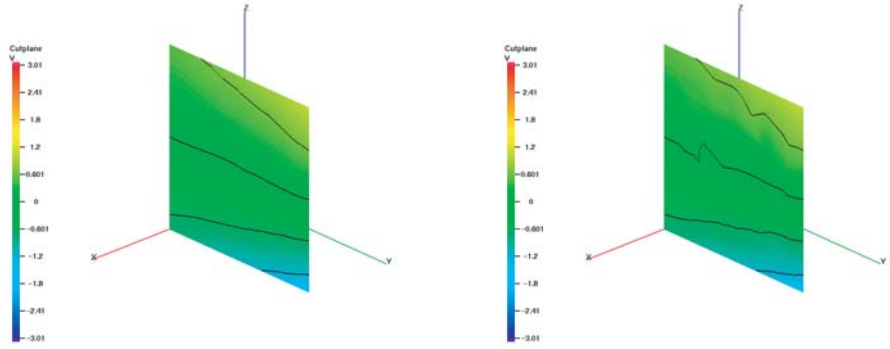


FIG. 5.4. *Velocity (u_{h2}) contours on a cutting plane: Stabilized (left), with $\gamma_\beta = 0$ (right).*

In what follows we will replace, in (5.1), the expression for the pressure by

$$p(x_1, x_2, x_3) = \begin{cases} 2x_2 & \text{if} \quad 0 \le x_2 \le \dfrac{1}{2}, \\ 2(1 - x_2) & \text{if} \quad \dfrac{1}{2} \le x_2 \le 1. \end{cases}$$

Clearly, this function satisfies $p \in H^1(\Omega)$ but does not belong to $H^2(\Omega)$. Figure 5.5

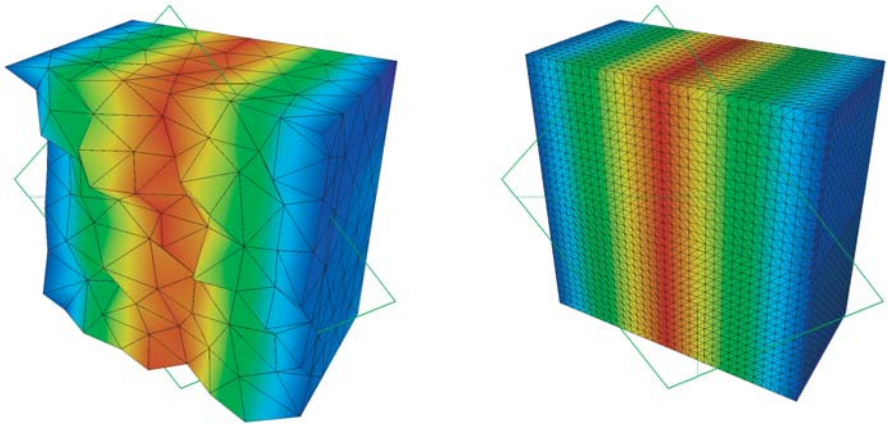


FIG. 5.5. *Cutting plane pressure: Coarse mesh and fine mesh.*

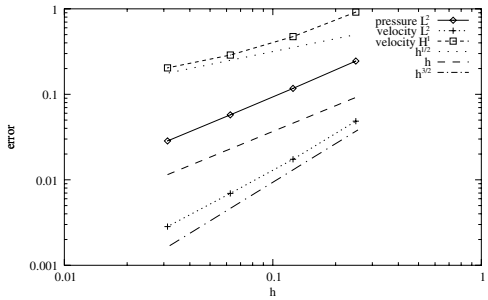


FIG. 5.6. *Convergence history: Linear elements, nonsmooth pressure, stabilization, parameters chosen as in* (2.13) *and* (2.14).

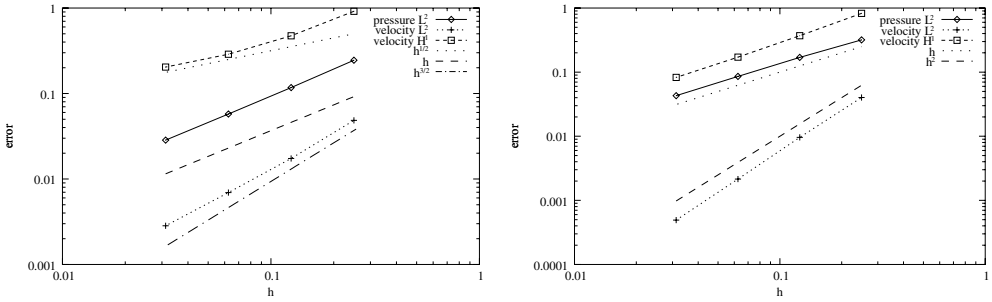


FIG. 5.7. *Convergence history: Linear elements, nonsmooth pressure, parameters chosen as in* (2.13) *and* (2.14). *Left: High Reynolds number. Right: Low Reynolds number.*

shows the pressure contours in a cut of a coarse and a fine mesh. Once more no spurious pressure oscillations are observed. Figure 5.6 shows the velocity and pressure convergence histories using linear elements. We get the suboptimal $O(h^{\frac{1}{2}})$ order for the velocity in the H^1-norm in the case of high local Reynolds numbers, in agreement with Theorem 4.9. The L^2-norm of the velocities, on the other hand, is still not far from the quasi-optimal $O(h^{\frac{3}{2}})$ convergence order. As expected, when the local Reynolds is low (for instance $\nu = 0.1$), we recover the optimal $O(h)$; see Figure 5.7,

left graphic. In addition, as predicted in Theorem 4.14, we notice that the convergence order for the velocity in the L^2-norm is $O(h^2)$.

6. Conclusion and outlook. In this paper, we have extended the results reported in [15, 13] to Oseen's equations using equal order interpolation and finite element spaces of arbitrary polynomial order. The stability properties of the method are based on an interior penalty term giving L^2-control of the jump of the gradient over interior element faces. We have shown that such a stabilization operator may be used to control all the nonsymmetric first order terms of Oseen's equations and that they give control only of the part of the operator that is not in the finite element space. In this sense the proposed method is a minimal stabilized method (see [8]).

The convergence analysis shows that the method has (quasi-) optimal convergence properties both in the L^2-norm and in the energy norm when the solution is sufficiently regular or the local Reynolds number is low. When physically realistic regularities are considered ($p \in H^1(\Omega)$) and the local Reynolds number is high, the convergence may become suboptimal $O(h^{\frac{1}{2}})$ due to the inconsistencies in the pressure stabilization. In some numerical examples we illustrated the theoretical results. The method shows very good performance in all regimes. In particular, we observe that in the high Reynolds number regime the scheme degenerates to the theoretical $O(h^{k+\frac{1}{2}})$ convergence in the L^2-norm predicted by the theory only in the case where the pressure is only H^1 and where the theoretical prediction is $O(h^{\frac{1}{2}})$.

The method presented here has some common features with VMS for LES as introduced in [30]. However, unlike the VMS, where two scales V_h and V_H defined by hierarchic meshes are considered (see, e.g., [35, 31, 4]), in our case the finite element space V_h represents the only resolved scale and the "turbulent" viscosity acts only on the gradient component that is not resolved on V_h. Recently, John and Kaya [31] proposed a VMS using a projection method framework which essentially takes the form of a standard Galerkin formulation for $\mathbf{u}_h$ supplemented with the turbulent viscosity acting only on the fine scales in the form of an additional term

$$(\nu_T(I - P_H)\boldsymbol{\epsilon}(\mathbf{u}_h), (I - P_H)\boldsymbol{\epsilon}(\mathbf{v}_h)), \tag{6.1}$$

where P_H is some map from fine scales to coarse scales. Comparing this now with the face oriented stabilization method, we would choose $H = h$ and thus make the turbulent viscosity act only on the scales that are not resolved on the space V_h. Applying Lemma 3.1 we immediately get an interior penalty interpretation of the term (6.1), with $P_H \overset{\text{def}}{=} \pi^*_{h,k}$,

$$\|\nu_T^{\frac{1}{2}}(I - \pi^*_{h,k})\boldsymbol{\epsilon}(\mathbf{u}_h)\|^2_\Omega \leq \sum_{K\in\mathcal{T}_h} \int_{\partial K} \nu_T h_K [\![\boldsymbol{\epsilon}(\mathbf{u}_h)]\!] : [\![\boldsymbol{\epsilon}(\mathbf{u}_h)]\!] \,\mathrm{d}s,$$

and we conclude that a possible subgrid modeling term would be

$$j_T(\mathbf{u}_h, \mathbf{v}_h) = \sum_{K\in\mathcal{T}_h} \int_{\partial K\setminus\partial\Omega} \nu_T h_K [\![\boldsymbol{\epsilon}(\mathbf{u}_h)]\!] : [\![\boldsymbol{\epsilon}(\mathbf{v}_h)]\!] \,\mathrm{d}s,$$

where the choice of ν_T now is a modeling issue. It should be noted that the choice $\nu_T = \gamma h_K$ gives us a term which is asymptotically equivalent to the face penalty operator using the whole gradient. However, other choices of ν_T based on modeling considerations are possible.

For sufficiently high polynomial degree there exists a C^1 subspace of V_h with approximation properties. It follows that the solution may be decomposed into one C^1 part which is untouched by the stabilizing terms and another C^0 part which is penalized. We conclude that the method enjoys the scale separation property characteristic for VMS as proposed in [30] by polynomial order rather than by hierarchic meshes. Future work will focus on the extension of the present method to the Navier–Stokes equations both from a numerical and a theoretical standpoint.

Finally, we remark that the Nitsche-type weak boundary conditions used in this paper, while nonstandard, have the benefit of acting as slip boundary conditions in the high Reynolds number regime and as no-slip conditions when the boundary layers are resolved; this may be favorable in LES (see Layton [34]).

Acknowledgment. The authors wish to thank the anonymous reviewers for their careful reading of the manuscript and their many constructive comments.

REFERENCES

[1] R. Becker and M. Braack, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 173–199.

[2] P. Bochev and M. Gunzburger, *An absolutely stable pressure-Poisson stabilized finite element method for the Stokes equations*, SIAM J. Numer. Anal., 42 (2004), pp. 1189–1207.

[3] M. Boman, *Estimates for the L_2-projection onto continuous finite element spaces in a weighted L_p-norm*, BIT, 46 (2006), to appear.

[4] M. Braack and E. Burman, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method*, SIAM J. Numer. Anal, 43 (2006), pp. 2544–2566.

[5] J. Bramble, J. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 71 (2002), pp. 147–156.

[6] P. G. Ciarlet, *Mathematical Elasticity, Volume I: Three-Dimensional Elasticity*, North-Holland, Amsterdam, 1988.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[8] F. Brezzi and M. Fortin, *A minimal stabilisation procedure for mixed finite element methods*, Numer. Math., 89 (2001), pp. 457–491.

[9] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[10] E. Burman, *A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty*, SIAM J. Numer. Anal., 43 (2005), pp. 2012–2033.

[11] E. Burman and M. A. Fernández, *Stabilized finite element schemes for incompressible flow using velocity/pressure spaces satisfying the LBB-condition*, in Proceedings of the 6th World Congress in Computational Mechanics (WCCM VI), Beijing, China, 2004.

[12] E. Burman, M. A. Fernández, and P. Hansbo, *Edge stabilization: An interior penalty method for the incompressible Navier-Stokes equation*, in Proceedings of the Fourth European Congress on Computational Methods in Applied Sciences and Engineering, Vol. I, Jyväskylä, Finland, 2004.

[13] E. Burman and P. Hansbo, *Edge stabilization for the generalized Stokes problem: A continuous interior penalty method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2393–2410.

[14] E. Burman and P. Hansbo, *A stabilized nonconforming finite element method for incompressible flow*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2881–2899.

[15] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453.

[16] P. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 9 (1975), pp. 77–84.

[17] R. Codina, *Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales*, Publication CIMNE 289, International Center for Numerical Methods

in Engineering (CIME), Barcelona, Spain, 2006.
[18] R. Codina and J. Blasco, *Analysis of a pressure-stabilized finite element approximation of the stationary Navier-Stokes equations*, Numer. Math., 87 (2000), pp. 59–81.
[19] C. R. Dohrmann and P. B. Bochev, *A stabilized finite element method for the Stokes problem based on polynomial pressure projections*, Internat. J. Numer. Methods Fluids, 46 (2004), pp. 183–201.
[20] J. Douglas and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences (Second International Symposium, Versailles, 1975), Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216.
[21] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.
[22] C. Ethier and D. Steinman, *Exact fully* 3*-D Navier-Stokes solutions for benchmarking*, Internat. J. Numer. Methods Fluids, 19 (1994), pp. 369–375.
[23] M. A. Fernández, L. Formaggia, A. Gauthier, J. Gerbeau, C. Prud'homme, and A. Veneziani, *The LifeV Project*, http://www.lifev.org.
[24] L. P. Franca and S. L. Frey, *Stabilized finite element methods.* II. *The incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.
[25] J. Freund and R. Stenberg, *On weakly imposed boundary conditions for second order problems*, in Proceedings of the Ninth International Conference on Finite Elements in Fluids, Venice, Italy, 1995, pp. 327–336.
[26] T. Gelhard, G. Lube, M. A. Olshanskii, and J.-H. Starcke, *Stabilized finite element schemes with LBB-stable elements for incompressible flows*, J. Comput. Appl. Math., 177 (2005), pp. 243–267.
[27] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Series in Comput. Math. 5, Springer-Verlag, Berlin, 1986.
[28] J. L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.
[29] P. Hansbo and A. Szepessy, *A velocity-pressure streamline diffusion finite element method for the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 84 (1990), pp. 175–192.
[30] T. Hughes, L. Mazzei, and K. Jansen, *Large eddy simulation and the variational multiscale method.*, Comput. Vis. Sci., 3 (2000), pp. 47–59.
[31] V. John and S. Kaya, *A finite element variational multiscale method for the Navier–Stokes equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1485–1503.
[32] C. Johnson and J. Saranen, *Streamline diffusion methods for the incompressible Euler and Navier-Stokes equations*, Math. Comp., 47 (1986), pp. 1–18.
[33] O. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.
[34] W. Layton, *Weak imposition of "no-slip" conditions in finite element methods*, Comput. Math. Appl., 38 (1999), pp. 129–142.
[35] W. Layton, *A connection between subgrid scale eddy viscosity and mixed methods*, Appl. Math. Comput., 133 (2002), pp. 147–157.
[36] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
[37] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Comput. Math. 25, Springer-Verlag, Berlin, 1997.
[38] L. Tobiska and R. Verfürth, *Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations*, SIAM J. Numer. Anal., 33 (1996), pp. 107–127.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1275–1296

PRECONDITIONERS FOR GENERALIZED SADDLE-POINT PROBLEMS*

CHRIS SIEFERT† AND ERIC DE STURLER‡

Abstract. We propose and examine block-diagonal preconditioners and variants of indefinite preconditioners for block two-by-two generalized saddle-point problems. That is, we consider the nonsymmetric, nonsingular case where the (2,2) block is small in norm, and we are particularly concerned with the case where the (1,2) block is different from the transposed (2,1) block. We provide theoretical and experimental analyses of the convergence and eigenvalue distributions of the preconditioned matrices. We also extend the results of [de Sturler and Liesen, *SIAM J. Sci. Comput.*, 26 (2005), pp. 1598–1619] to matrices with nonzero (2,2) block and to the use of approximate Schur complements. To demonstrate the effectiveness of these preconditioners we show convergence results, spectra, and eigenvalue bounds for two model Navier–Stokes problems.

Key words. saddle-point problems, generalized saddle-point problems, iterative methods, preconditioning, Krylov subspace methods, eigenvalue bounds

AMS subject classification. 65F10

DOI. 10.1137/040610908

1. Introduction.

We examine preconditioners for real systems of the form

$$\mathcal{A}\begin{bmatrix} x \\ y \end{bmatrix} \equiv \begin{bmatrix} A & B^T \\ C & D \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, and $n > m$. For many relevant problems, $D = 0$ and $B \neq C$, and such problems are referred to as generalized saddle-point problems [24]. For other problems we consider $D \neq 0$, but $\|D\|_2$ is small enough that the problem retains the characteristics of a generalized saddle-point problem. In many such problems, the nonzero (2,2) block arises from a stabilization term. However, this is not always the case. In a problem involving metal deformation [35], for example, it derives from very slight compressibility. In addition, we note that certain approaches to stabilization lead to systems where $B \neq C$ [3, 24], [27, sections 7.5 and 9.4] although many other problems have $B = C$. Finally, our preconditioners allow A to be singular. We consider all of these cases, which arise in many applications, ranging from stabilized formulations of the Navier–Stokes equations [4, 11, 27] to metal deformation [35] and interior point methods [13].

Problems of this type have been of recent interest [1, 8, 9, 18, 20, 23], as have their symmetric counterparts [7, 10, 14, 26, 31, 34] and the case where $D = 0$ [2, 5, 6, 8, 15, 19, 21, 23, 32]. However, preconditioners for the case where $B \neq C$ have not received as much attention. Though they are considered in [8, 18, 23], these papers do not provide numerical experiments for such problems. We will do this in the present paper. In [8], a detailed analysis is provided for two classes of preconditioners for

*Received by the editors July 1, 2004; accepted for publication (in revised form) December 2, 2005; published electronically July 7, 2006. This work was supported in part by the U.S. Department of Energy under grant DOE LLNL B341494 through the Center for the Simulation of Advanced Rockets.
http://www.siam.org/journals/sinum/44-3/61090.html

†Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (siefert@uiuc.edu).

‡Department of Mathematics, 460 McBryde, Virginia Tech, Blacksburg, VA 24061-0123 (sturler @vt.edu).

the case where $B \neq C$ and $D = 0$. Here, we extend these preconditioners to the case where $D \neq 0$ and to allow for approximations to the Schur complement matrix that arises in the preconditioner. Our preconditioners for (1.1) derive from a matrix splitting, $A = F - E$. Our purpose is to derive preconditioners that result in tightly clustered eigenvalues. In general, this leads to fast convergence for Krylov subspace methods, although in the nonsymmetric case the eigenvectors may play a role as well.

In this paper we assume that the matrix is nonsingular or that the singularity can be easily removed, such as the constant pressure mode in the Oseen problem [12]. For the splitting, we assume that F and $(D - CF^{-1}B^T)$ are nonsingular. In section 2, we propose a block-diagonal preconditioner that is a generalization of the preconditioners discussed in [18] and [8]. In section 3, we use this preconditioner to derive a second preconditioned system, which is a generalization of the *related system* presented in [8]. For the $D = 0$ case, the related system corresponds to an efficient implementation of a constraint preconditioner; see also [5, 6, 14, 26]. In section 4, we extend both types of preconditioner to the use of approximate Schur complements. Our analysis focuses on the $D \neq 0$ case, but we provide specializations to the $D = 0$ case as well. While the block-diagonally preconditioned system may be very effective or more convenient in certain situations, the related system is generally the better preconditioner, offering much faster convergence for a modest increase in the computational cost per iteration. Therefore, in section 5 on numerical experiments we focus on the related system.

We propose preconditioners with exact (sections 2 and 3) and with approximate (section 4) Schur complements, and we discuss the convergence for the preconditioned systems and the clustering of the eigenvalues. We explore two model problems in section 5. The first, which arises from a finite element discretization of the Navier–Stokes equations, has $D \neq 0$ and $A \neq A^T$. The second, which arises from a spectral collocation approach for an incompressible Stokes problem, has $B \neq C$ and $D = 0$. We use eigenvalue bounds and numerical experiments to illustrate that reasonable choices for splittings and approximate Schur complements yield good convergence. Our analysis also illustrates the issues involved in choosing splittings and approximate Schur complements to achieve effective preconditioning. Although eigenvalue bounds are often wide, they nevertheless indicate good eigenvalue clustering for reasonable choices for splittings and approximate Schur complements.

2. Block-diagonal preconditioners (exact Schur complement). We consider a splitting of the (1,1) block, $A = F - E$, where F is easy to solve with and $(D - CF^{-1}B^T)^{-1}$ exists. Note that $-(D - CF^{-1}B^T)$ is the Schur complement of the matrix

$$\begin{bmatrix} F & B^T \\ C & D \end{bmatrix}, \tag{2.1}$$

and we will use the phrase *exact Schur complement* to refer to $-(D - CF^{-1}B^T)$. Next, we introduce the following block-diagonal preconditioner as a straightforward generalization of preconditioners in [8, 18]:

$$\mathcal{P}(F) = \begin{bmatrix} F^{-1} & 0 \\ 0 & -(D - CF^{-1}B^T)^{-1} \end{bmatrix}. \tag{2.2}$$

Preconditioning from the left or the right with $\mathcal{P}$ yields a system of the form

$$\mathcal{B}(F) \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \tag{2.3}$$

where $\mathcal{B}(F)$ is either $\mathcal{PA}$ or $\mathcal{AP}$. For example, the matrix from the left-preconditioned system is

$$\mathcal{P}(F)\mathcal{A} = \begin{bmatrix} I - F^{-1}E & F^{-1}B^T \\ -(D - CF^{-1}B^T)^{-1}C & -(D - CF^{-1}B^T)^{-1}D \end{bmatrix},$$

implicitly defining S, N, M, and Q in (2.3) for the left-preconditioned case. Apart from the preconditioned (2,2) block Q, this resembles the system arising from the zero (2,2) block case. For the rest of this paper, we assume that Q is diagonalizable. While $MN = I$ for the $D = 0$ case [23, 8], for $D \neq 0$ we have

$$\begin{aligned} MN &= -(D - CF^{-1}B^T)^{-1}CF^{-1}B^T = -(D - CF^{-1}B^T)^{-1}(-D + CF^{-1}B^T + D) \\ &= I + Q. \end{aligned} \tag{2.4}$$

This is true for both the left- and right-preconditioned cases. In the $D = 0$ case, NM is a projector [8]. For the $D \neq 0$ case, it is not, as $(NM)^2 = NM + NQM$.

In section 2.1 we derive the eigendecomposition of the matrix

$$\mathcal{B}_0 = \begin{bmatrix} I & N \\ M & Q \end{bmatrix}, \tag{2.5}$$

when $I+Q$ (and thus B^T and C) have full rank. We use this in section 2.2 to develop bounds for the eigenvalues of $\mathcal{B}(F)$ using perturbation theory. Finally, in section 2.3, we discuss the case when $I + Q$ is rank-deficient.

2.1. Eigenvalues and eigenvectors of $\mathcal{B}_0$. Assume that $I + Q$ (and thus B^T and C) have full rank. We wish to find λ, u, and v such that

$$u + Nv = \lambda u, \tag{2.6}$$

$$Mu + Qv = \lambda v. \tag{2.7}$$

First, we assume $\lambda = 1$. Substituting this into (2.6) and using $Q = MN - I$ in (2.7) yields

$$Nv = 0 \quad \text{and} \quad Mu = 2v. \tag{2.8}$$

Since B^T has full column rank by assumption, this implies that $v = 0$ and that $\mathcal{B}_0$ has only eigenpairs of the form

$$\left(1, \begin{bmatrix} u \\ 0 \end{bmatrix}\right), \qquad \text{where } u \in \text{null}\,(M)\,. \tag{2.9}$$

Since C has full row rank, so does M, and $\mathcal{B}_0$ has precisely $n-m$ distinct eigenpairs of this type. Next, we consider the case where $\lambda \neq 1$. Solving (2.6) for u and substituting into (2.7) yields

$$\lambda Q v_j = (\lambda^2 - \lambda - 1)v_j. \tag{2.10}$$

Hence, the v_j must be eigenvectors of Q. We have assumed that Q has a full set of eigenpairs, $Qv_j = \delta_j v_j$, for $j = 1, \ldots, m$. We then solve (2.10) for λ to yield

$$\lambda_j^{\pm} = \frac{(1 + \delta_j) \pm \sqrt{4 + (1 + \delta_j)^2}}{2}; \tag{2.11}$$

cf. [11]. Using (2.6) with the eigenvectors of Q for v yields the vectors u. We finally rescale the eigenvector by $(\lambda_j^{\pm} - 1)$ to yield eigenpairs of the form

$$\left(\lambda_j^{\pm}, \begin{bmatrix} Nv_j \\ (\lambda_j^{\pm} - 1)v_j \end{bmatrix}\right). \tag{2.12}$$

Note that $\lambda_j^- \neq 1$ regardless of the choice of δ_j, and $\lambda_j^+ = 1$ only if $\delta_j = -1$. However, the latter would contradict the assumption that $I + Q$ has full rank. Thus, $\mathcal{B}_0$ has $2m$ eigenpairs corresponding to $\lambda \neq 1$. This completes a full set of eigenpairs for $\mathcal{B}_0$. Let U_1 be a matrix whose columns form an orthonormal basis for null (M) (cf. (2.9)), and let U_2 be the matrix with normalized columns $u_j = Nv_j$, where $Qv_j = \delta_j v_j$; cf. (2.12). Furthermore, let $\Lambda^+ = \operatorname{diag}(\lambda_j^+)$ and $\Lambda^- = \operatorname{diag}(\lambda_j^-)$, where $\operatorname{diag}(\cdot)$ denotes the diagonal matrix with the given arguments. Then, the following matrix, $\mathcal{Y}$, is an eigenvector matrix of $\mathcal{B}_0$:

$$\mathcal{Y} \equiv \left[\begin{array}{c|c} Y_{11} & Y_{12} \\ \hline Y_{21} & Y_{22} \end{array}\right] = \left[\begin{array}{cc|c} U_1 & U_2 & U_2 \\ 0 & V(\Lambda^+ - I) & V(\Lambda^- - I) \end{array}\right]. \tag{2.13}$$

For our perturbation results we also need

$$\mathcal{Z} = \mathcal{Y}^{-1} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{2.14}$$

Using the block-inversion formula in [17, section 0.7.3] we obtain [29, 30]

$$Z_{11} = \begin{bmatrix} I_{n-m} & 0 \\ 0 & \Upsilon^+ \end{bmatrix} Y_{11}^{-1} = \hat{I}_n Y_{11}^{-1}, \tag{2.15}$$

$$Z_{21} = -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1}, \tag{2.16}$$

$$Z_{12} = -\begin{bmatrix} 0 \\ (\Lambda^- - \Lambda^+)^{-1} V^{-1} \end{bmatrix}, \tag{2.17}$$

$$Z_{22} = \left(V(\Lambda^- - \Lambda^+)\right)^{-1}, \tag{2.18}$$

with $\Upsilon^+ = \operatorname{diag}((\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+))$ and $\Upsilon^- = \operatorname{diag}((\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+))$. For $Q = 0$ (because $D = 0$), the eigendecomposition of $\mathcal{B}_0$ reduces to the case discussed in [8].

2.2. Perturbation bounds on the eigenvalues of $\mathcal{B}(F)$. We are now ready to consider the eigenvalues of $\mathcal{B}(F)$ and derive bounds on the spectrum. Throughout this paper $\|\cdot\|$ indicates the 2-norm.

THEOREM 2.1. *Consider matrices $\mathcal{B}(F)$ of the form* (2.3). *Let $\mathcal{Y}$ be the eigenvector matrix of $\mathcal{B}_0$, as given by* (2.13). *Then for each eigenvalue $\lambda_{\mathcal{B}}$ of $\mathcal{B}(F)$ there exists an eigenvalue λ of $\mathcal{B}_0$ such that*

$$|\lambda_{\mathcal{B}} - \lambda| \leq \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| \tag{2.19}$$

$$\leq 2 \max\left(1, \|\Upsilon^+\|, \|\Upsilon^-\|\right) \|Y_{11}^{-1} S Y_{11}\|. \tag{2.20}$$

Proof. Since $\mathcal{B}_0$ is diagonalizable, (2.19) follows from a classic result in perturbation theory [33, Theorem IV.1.12]. We expand the right-hand side of (2.19) using

(2.13)–(2.17) to get (see also [8])

$$
\begin{aligned}
|\lambda_{\mathcal{B}} - \lambda| &\le \left\| \begin{bmatrix} \hat{I}_n Y_{11}^{-1} S Y_{11} & \hat{I}_n Y_{11}^{-1} S Y_{12} \\ -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1} S Y_{11} & -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1} S Y_{12} \end{bmatrix} \right\| \\
&\le \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \\
&\quad \cdot \left\| \begin{bmatrix} Y_{11}^{-1} S U_1 & Y_{11}^{-1} S U_2 & Y_{11}^{-1} S U_2 \\ -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1} S U_1 & -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1} S U_2 & -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1} S U_2 \end{bmatrix} \right\|.
\end{aligned}
$$

Using the consistency of the 2-norm, we can simplify this to (see also [8])

$$
\begin{aligned}
|\lambda_{\mathcal{B}} - \lambda| &\le \sqrt{2} \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \left\| \begin{bmatrix} Y_{11}^{-1} S Y_{11} \\ -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1} S Y_{11} \end{bmatrix} \right\| \\
&\le 2 \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \left\| Y_{11}^{-1} S Y_{11} \right\|. \qquad \square
\end{aligned}
$$

The $\Upsilon^{\pm}$ terms can be large only if $\delta_j \approx -1 \pm 2i$. For the problems discussed in section 5, the δ_j's are well separated from this value, because $\|D\|$ is small and the problem and preconditioner are relatively well conditioned. The following lemma provides bounds on the $\|\Upsilon^{\pm}\|$. We explicitly consider the special case where the δ_j's are real (and thus bounded away from $-1 \pm 2i$). This occurs in the important case that D is symmetric and the Schur complement is definite. For the following proof and subsequent discussions, we define the function $p(z) = 4 + (1+z)^2$.

LEMMA 2.2. *Let Υ^+ and Υ^- be defined as above.*

1. *If $\delta_j \in \mathbb{R}$, for all j, then*

$$\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \le \frac{1+\sqrt{2}}{2}.$$

Moreover, if $\delta_j \ge -1$, for all j, then $\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) = 1$.

2. *If $\delta_j \in \mathbb{C}$ and $\exists \alpha : |\delta_j| \le \alpha < \sqrt{5}$ for $j = 1, \ldots, m$, then*

$$\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \le \max\left(1, \frac{1}{2} + \frac{1+\alpha}{2\sqrt{2\left(\sqrt{5}-\alpha\right)}}\right).$$

Proof. Substituting $\lambda_j^{\pm}$ from (2.11) into $\Upsilon^+ = \operatorname{diag}(\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+)$ and $\Upsilon^- = \operatorname{diag}(\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+)$ gives

$$\Upsilon^{\pm} = \operatorname{diag}\left(\frac{1-\delta_j}{2\sqrt{4+(1+\delta_j)^2}} \pm \frac{1}{2}\right) = \operatorname{diag}\left(\frac{1-\delta_j}{2\sqrt{p(\delta_j)}} \pm \frac{1}{2}\right). \tag{2.21}$$

The proof for the real case now follows from basic calculus.

For the complex case, note that $p(\delta) = (\delta + 1 + 2i)(\delta + 1 - 2i)$. Any δ must be at least distance 2 from one of the roots of $p(\delta)$. We assume without loss of generality that δ is near $-1+2i$. The value $\delta_* = (-1+2i)\alpha/\sqrt{5}$ minimizes $|\delta + 1 - 2i|$ subject to $|\delta| \le \alpha$, and we have $|\delta_* + 1 - 2i| = \sqrt{5} - \alpha$. So, we have $|p(\delta)| \ge 2\left(\sqrt{5} - \alpha\right)$. Using this inequality for $|p(\delta)|$ after taking norms in (2.21) completes the proof. $\square$

In practice, the bound for the complex case is quite modest. For example, if $|\delta_j| \le 1$ for all j, then our bound on $\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|)$ is about 1.136. Likewise, if $|\delta_j| \le 2$ for all j, the bound is about 1.470.

We derive a bound on $\left\|Y_{11}^{-1}SY_{11}\right\|$ following the approach in [8]. Recall that $Y_{11} = [U_1\ U_2]$, where $U_1^T U_1 = I$, and $U_2 = NV$ with unit columns. Let $U_2 = V_2\Theta$, where $V_2^T V_2 = I$. Furthermore, let $\omega_1 = \|U_1^T V_2\|$, which is the cosine of the smallest principal angle between range $(U_1) =$ null (NM) and range $(U_2) =$ range (NM).

LEMMA 2.3. *Define Y_{11}, S, U_1, U_2, V_2, Θ, and ω_1 as above, and let $\kappa(.)$ denote the 2-norm condition number. Then*

$$\left\|Y_{11}^{-1}SY_{11}\right\| \leq \kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\|. \tag{2.22}$$

Proof. We have $\|Y_{11}^{-1}SY_{11}\| \leq \kappa(Y_{11})\|S\|$, where

$$Y_{11} = \begin{bmatrix} U_1 & V_2 \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & \Theta \end{bmatrix},$$

since U_2 has unit columns $\|\Theta\| \geq 1$ and $\|\Theta^{-1}\| \geq 1$. So, our bound simplifies to

$$\|Y_{11}^{-1}SY_{11}\| \leq \kappa(\Theta)\ \kappa\left(\begin{bmatrix} U_1 & V_2 \end{bmatrix}\right)\|S\| \leq \kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\|, \tag{2.23}$$

where the second inequality follows from the bound on $\kappa([U_1\ V_2])$ from Lemma 3.6 in [8]. ∎

COROLLARY 2.4. *Let Θ and ω_1 be defined as above.*

1. *If $\delta_j \in \mathbb{R}$ for all j, then*

$$|\lambda_{\mathcal{B}} - \lambda| \leq (1+\sqrt{2})\kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\|. \tag{2.24}$$

2. *If $\delta_j \in \mathbb{C}$ and $\exists\alpha : |\delta_j| \leq \alpha < \sqrt{5}$ for $j = 1, \ldots, m$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \leq 2\max\left(1, \frac{1}{2} + \frac{1+\alpha}{2\sqrt{2\left(\sqrt{5}-\alpha\right)}}\right)\kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\|. \tag{2.25}$$

Proof. Use Lemmas 2.2 and 2.3 in Theorem 2.1. ∎

We see that the clustering of the eigenvalues depends mainly on $\|S\|$ and the size of the δ_j, unless $\omega_1 \approx 1$ or $\kappa(\Theta)$ large. This implies that the block-diagonally preconditioned system can have as many as $2m+1$ eigenvalue clusters, one for $\lambda = 1$ and one for each $\lambda_j^{\pm}$. Hence, the convergence of Krylov methods may not be very good for the block-diagonally preconditioned system, even if $\|S\|$ is small. Examples in section 5 will illustrate this. However, when the δ_j and $\|S\|$ are small, the block-diagonal preconditioner will give good convergence. This typically happens for small mesh width when D and Q are h-dependent; see Table 5.1. In addition, the block-diagonal preconditioner provides an intermediate step to a better preconditioner described in section 3.

2.3. Rank-deficiency of $I + Q$. In section 2.1, we made the assumption that $I + Q$ has full rank (for $D = 0$, this is always true). We now briefly discuss the rank-deficient case.

There are three sources of potential rank-deficiency in $I + Q$. The first two are rank-deficiency in C and B^T. The third is when there are vectors v such that $Nv \neq 0$

and $Nv \in \text{null}\,(M)$. This implies that $MNv = (I+Q)v = 0$ and v is an eigenvector of Q. This case occurs when F^{-1} (for left preconditioning) or $-(D - CF^{-1}B^T)^{-1}$ (for right preconditioning) maps a nontrivial vector from range $\left(B^T\right)$ into null (C).

Assume that $I+Q$, C, and B^T are rank deficient by k, l_c, and l_b, respectively. Note that $k \geq \max(l_b, l_c)$, since $I+Q = -(D-CF^{-1}B^T)^{-1}CF^{-1}B^T$ and the product of matrices cannot be of higher rank than any of its factors.

Our previous analysis remains valid for the $2(m-k)$ eigenpairs (2.11) that correspond to $\delta_j \neq -1$. It is also valid for the k eigenpairs where $\delta_j = -1$ that correspond to λ_j^-. Since the Schur complement is invertible, M must also be rank deficient by l_c. Thus, the number of eigenpairs of the form (2.9) equals $\dim(\text{null}\,(M)) = n - m + l_c$. This gives a total of $n + m - k + l_c$ eigenpairs, leaving us to find $k - l_c$ eigenpairs.

From (2.8), we have that all eigenvectors corresponding to $\lambda = 1$ must satisfy $Nv = 0$ and $Mu = 2v$. Since $\dim(\text{null}\,(N)) = l_b$, there are l_b independent vectors v that satisfy $Nv = 0$. Unfortunately, there may be as many as l_c independent vectors v where $Mu = 2v$ has no solution. If we do not have $k - l_c$ independent vectors v such that $Mu = 2v$ has a solution, then $\mathcal{B}_0$ is defective. The analysis of section 2.1 does not permit any other eigenvectors.

For the missing eigenpairs we have that $\lambda_j^+ \to 1$ as $\delta_j \to -1$. Therefore, we look for principal vectors of grade two (see [16]) for $\lambda = 1$. These vectors satisfy the equations

$$Nv = \tilde{u} \quad \text{and} \quad Mu = 2v, \tag{2.26}$$

where $\tilde{u} \neq 0$ and $\tilde{u} \in \text{null}\,(M)$. We note that there are k independent vectors v such that $(I+Q)v = 0$. Since there are precisely l_b independent vectors v such that $Nv = 0$, there must be $k - l_b$ such vectors v that satisfy $Nv = \tilde{u}$ with $\tilde{u} \neq 0$ and $M\tilde{u} = 0$. This gives k independent vectors v that satisfy the first equation of either (2.8) or (2.26).

There exists a space of dimension l_c such that $Mu = 2v$ has no solution. However, since we have k independent v's to propose, we are guaranteed to find $k - l_c$ independent vectors v's that satisfy this equation. This gives us either our remaining eigenvectors or principal vectors of grade two. This also guarantees us that we have Jordan blocks of size at most two.

In the special case when $k = l_b = l_c$, we have $k - l_c = 0$, so we have a full set of eigenvectors. We can apply the analysis described in the full rank case with k additional eigenpairs $(1, [\tilde{u}_{n-m+j}^T, 0^T]^T)$ for $j = 1, \ldots, k$, replacing the corresponding eigenpairs $(\lambda_j^+, [(Nv_j)^T, (\lambda_j^+ - 1)v_j^T]^T)$ for which $\delta_j = -1$. Let U_1 be such that $U_1^T U_1 = I_{n-m+l_c}$ and range $(U_1) = \text{null}\,(M)$. Let $\widetilde{V}$ be such that $\widetilde{V}^T\widetilde{V} = I_{l_c}$ and range$(\widetilde{V}) = \text{null}\,(I+Q)$. Further, let the columns of $\widehat{V}$ be the eigenvectors of Q corresponding to the eigenvalues $\delta_j \neq -1$, scaled such that $U_2 = N\widehat{V}$ has unit columns. Finally, let the diagonal matrices $\widehat{\Lambda}^+$ and $\widehat{\Lambda}^-$ contain the eigenvalues λ_j^+ and λ_j^- corresponding to the eigenvalues $\delta_j \neq -1$ ordered consistently with the columns of $\widehat{V}$. Then the eigenvector matrix of $\mathcal{B}_0$ is given by

$$\mathcal{Y} = \left[\begin{array}{cc|cc} U_1^{(n-m+l_c)} & U_2^{(m-l_c)} & N\tilde{V}^{(l_c)} & U_2^{(m-l_c)} \\ \hline 0 & \widehat{V}(\widehat{\Lambda}^+ - I) & -2\widetilde{V} & \widehat{V}(\widehat{\Lambda}^- - I) \end{array}\right], \tag{2.27}$$

where superscripts in the top row indicate the number of columns. The corresponding eigenvalues are those from (2.9) and (2.11). We can then use the eigenvector matrix

of $\mathcal{B}_0$ given in (2.27) to derive bounds on the eigenvalues, as for the full rank case. The reduction in the number of columns of U_2 may in fact reduce the factor $\kappa(\Theta)$ in Corollary 2.4. An important example of this case is the stabilized Navier–Stokes (Oseen) problem [11], where $C = B$ and F is positive definite.

3. Fixed-point method and its related system (exact Schur complement). We now consider an alternative solution method that leads to faster convergence in general; cf. [8]. In the $D = 0$ case this approach leads to an efficient implementation of so-called constraint preconditioners; cf. [6, 5, 14, 26]. We can derive the following splitting from (2.3):

$$\mathcal{B}(F)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} I-S & N \\ M & Q \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \left(\mathcal{B}_0 - \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\right)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}. \tag{3.1}$$

Note that

$$\mathcal{B}_0^{-1} = \begin{bmatrix} I-NM & N \\ M & -I \end{bmatrix}. \tag{3.2}$$

We left-multiply (3.1) by $\mathcal{B}_0^{-1}$ to yield the fixed-point iteration,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} (I-NM)S & 0 \\ MS & 0 \end{bmatrix}\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{3.3}$$

Note that this iteration is formally the same as for the $D = 0$ case in [5, 8]. Since x_{k+1} and y_{k+1} depend only on x_k, we need to iterate only on the x_k variables; see also [4, pp. 214–215] and [8]. The x-component of the fixed-point of (3.3) satisfies the so-called *related system* for the fixed-point iteration [16],

$$(I-(I-NM)S)x = \hat{f}.\ ^{1} \tag{3.4}$$

The full-size related system (that is, with the y component) and $D \neq 0$ has been examined elsewhere for special cases. In [26], A is symmetric positive definite and spectrally equivalent to the identity, and so the splitting $F = I$ is used. In [14], F is symmetric positive definite. In both of these cases $B = C$.

3.1. Eigenvalue bounds for fixed-point matrix and related system. In this section we assume $n-m \geq m$, but equivalent results are obtained for $m > n-m$. Let U_1 and U_2 be defined as in (2.13), $\Delta = \operatorname{diag}(\delta_j)$, and let $U_2 = V_2\Theta$ with $V_2^T V_2 = I$. Then, we have $NMU_1 = 0$, $NMU_2 = NMNV = NV(I+\Delta)$, and therefore

$$(I-NM) = \begin{bmatrix} U_1 & V_2 \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & -\Theta\Delta\Theta^{-1} \end{bmatrix}\begin{bmatrix} U_1 & V_2 \end{bmatrix}^{-1}. \tag{3.5}$$

In the rank-deficient case, we can use (2.27). So, for this approach rank-deficiency has a potential advantage in terms of the conditioning of Θ. To analyze $\|I-NM\|$ we need the following singular value decomposition (SVD):

$$U_1^T V_2 = \Phi\Omega\Psi^T, \quad \text{where } 1 > \omega_1 \geq \omega_2 \geq \cdots \geq \omega_m. \tag{3.6}$$

[1] The full-size related system derives from using (2.1) as a left-preconditioner; see also [8].

Following [8], we define W by $W\Sigma = V_2\Psi - U_1\Phi\Omega$, where the diagonal matrix $\Sigma = \mathrm{diag}((1-\omega_j^2)^{1/2})$ contains the sines of the principal angles between range (U_1) and range (V_2). Then, $[U_1\ W]$ is orthogonal, and we can decompose V_2 as follows:

$$V_2 = U_1\Phi\Omega\Psi^T + W\Sigma\Psi^T. \tag{3.7}$$

THEOREM 3.1. *Let U_1,V_2, and ω_1 be defined as above. Let λ_R be an eigenvalue of the related system matrix in* (3.4). *Then,*

$$\left.\begin{matrix}\rho((I-NM)S)\\ |1-\lambda_R|\end{matrix}\right\} \le (1-\omega_1^2)^{-1/2}(1+\|\Theta\Delta\Theta^{-1}\|)\|S\|,$$

where $\rho(\cdot)$ designates the spectral radius.

Proof. The proof of this theorem largely follows [8]. Note that the result for $\rho((I-NM)S)$ immediately implies the result for $|1-\lambda_R|$. We have $\rho((I-NM)S) \le \|I-NM\|\|S\|$. Let $Z = -\Theta\Delta\Theta^{-1}$. Then,

$$\|I-NM\| = \left\|[U_1\ V_2]\begin{bmatrix} I & 0\\ 0 & Z\end{bmatrix}[U_1\ V_2]^{-1}\right\| \tag{3.8}$$

$$\le \left\|[U_1\ V_2]\begin{bmatrix} I & 0\\ 0 & 0\end{bmatrix}[U_1\ V_2]^{-1}\right\| + \left\|[U_1\ V_2]\begin{bmatrix} 0 & 0\\ 0 & Z\end{bmatrix}[U_1\ V_2]^{-1}\right\| \tag{3.9}$$

$$\le (1-\omega_1^2)^{-1/2} + (1-\omega_1^2)^{-1/2}\|Z\| = (1-\omega_1^2)^{-1/2}(1+\|Z\|). \tag{3.10}$$

The first term in (3.9) is the norm of an oblique projection. Given the SVD in (3.6), this norm equals $(1-\omega_1^2)^{-1/2}$ [22, section 5.15]. We establish the bound for the second term as follows:

$$\left\|[U_1\ V_2]\begin{bmatrix} 0 & 0\\ 0 & Z\end{bmatrix}[U_1\ V_2]^{-1}\right\| = \max_{U_1a+V_2b\neq 0}\frac{\|V_2Zb\|}{\|U_1a+V_2b\|}. \tag{3.11}$$

Without loss of generality we may assume $\|b\| = 1$, so that $\|V_2Zb\| \le \|Z\|$. From (3.7) we see that $\|U_1a+V_2b\| = \|U_1a+U_1\Phi\Omega\Psi^Tb+W\Sigma\Psi^Tb\|$, which for any given b is minimized by $a = -\Phi\Omega\Psi^Tb$. This gives $\|U_1a+V_2b\| = \|W\Sigma\Psi^Tb\|$, which in turn is minimized for $b = \psi_1$. Hence, we have

$$\left\|[U_1\ V_2]\begin{bmatrix} 0 & 0\\ 0 & Z\end{bmatrix}[U_1\ V_2]^{-1}\right\| = \max_{U_1a+V_2b\neq 0}\frac{\|V_2Zb\|}{\|U_1a+V_2b\|} \le (1-\omega_1^2)^{-1/2}\|Z\|. \tag{3.12}$$

Therefore, by using (3.8)–(3.12) we have

$$\rho((I-NM)S) \le (1-\omega_1^2)^{-1/2}(1+\|\Theta\Delta\Theta^{-1}\|)\|S\|. \qquad \square$$

If the δ_j are clustered, the influence of $\kappa(\Theta)$ is small.

COROLLARY 3.2. *Let $\hat\delta = \arg\min_{z\in\mathbb{C}}\max_j|z-\delta_j|$ and $\tilde\delta_j = \delta_j - \hat\delta$. Then*

$$\left.\begin{matrix}\rho((I-NM)S)\\ |1-\lambda_R|\end{matrix}\right\} \le (1-\omega_1^2)^{-1/2}(1+\hat\delta+\kappa(\Theta)\max|\tilde\delta_j|)\|S\|.$$

Proof. Note that $\Delta = \hat\delta I + \mathrm{diag}(\tilde\delta_j)$, so $\Theta\Delta\Theta^{-1} = \hat\delta I + \Theta\ \mathrm{diag}(\tilde\delta_j)\,\Theta^{-1}$. $\square$

So, the eigenvalues of the related system cluster around 1, and the tightness of the clustering is controlled through $\|S\|$. Note that the factor containing ω_1 in Corollary 3.2 is no larger than the corresponding factor for the block-diagonally preconditioned system in Corollary 2.4. In addition, the influence of the $\kappa(\Theta)$ term is smaller for the related system if the δ_j are clustered. This generally leads to better clustering and tighter bounds for the related system than for the block-diagonally preconditioned system. Because of these advantages, the related system will generally have faster convergence than the block-diagonally preconditioned system.

3.2. Satisfying "constraints". In the $D = 0$ case, the second block of equations in (1.1) often represents a set of constraints. For the $D \neq 0$ case, this may or may not be the case. So-called constraint preconditioners in the $D = 0$ case have the advantage that each iterate of a Krylov subspace method for the preconditioned system satisfies the constraints if the initial guess is chosen appropriately. Fixed-point methods such as (3.3) often satisfy the constraints after a single step. This is the case for the fixed-point method proposed in [8] for $D = 0$. It turns out that we can prove an analogous property for the $D \neq 0$ case.

LEMMA 3.3. *For any initial guess* $[x_0^T, y_0^T]^T$, *the iterates,* $[x_k^T, y_k^T]^T$, *for* $k = 1, 2, \ldots,$ *of* (3.3) *satisfy* $Mx_k + Qy_k = \tilde{g}$ *in* (2.3) *and* $Cx_k + Dy_k = g$ *in* (1.1).

The proof can be found in [29, 30].

COROLLARY 3.4. *After the first iteration of* (3.3), *all fixed-point updates are in the null space of* $[M\ Q]$.

This follows trivially from Lemma 3.3.

We can also show that the iterates of a Krylov subspace method will satisfy the constraints if the initial guess satisfies the constraints (cf. [8]). We first give a general result and then specialize it to our problem. For the remainder of this section, A and C are arbitrary matrices, not the matrices referred to in (1.1). We return to the nomenclature of (1.1) in the next section.

THEOREM 3.5. *Let* $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $C \in \mathbb{R}^{m\times n}$, *and* $d \in \mathbb{R}^m$, *and define the iteration* $x_{k+1} = Ax_k + b$. *Further, let the iterates* x_k *satisfy* $Cx_k = d$ *for* $k \geq 1$ *and any starting vector* x_0. *Then, the iterates* $x^{(m)}$, $m \geq 0$, *of a Krylov method applied to the (related) system,* $(I - A)x = b$, *will satisfy* $Cx^{(m)} = d$ *if* $Cx^{(0)} = d$.

The proof can be found in [29, 30].

COROLLARY 3.6. *The iterates,* $[x^{(m)^T}, y^{(m)^T}]^T$, *of any Krylov method applied to the full* $n+m$ *related system for* (3.3) *satisfy* $Mx^{(m)} + Qy^{(m)} = \tilde{g}$ *and* $Cx^{(m)} + Dy^{(m)} = g$ *if the initial guess is the result of at least one step of fixed-point iteration* (3.3).

Proof. Use Theorem 3.5, with A as fixed-point iteration matrix in (3.3), $b = [\hat{f}^T\ \hat{g}^T]^T$, $C = [M\ Q]$, and $d = \hat{g}$. ∎

4. Approximate Schur complement. It may be expensive to compute the Schur complement matrix $(D - CF^{-1}B^T)$ or to compute and apply its inverse (or factors). So, we would like to use a cheap approximation to the inverse of the Schur complement. We now consider the effect of such an approximation on the eigenvalue clustering of the preconditioned matrices and on the resulting convergence. Let $S_1 = -(D - CF^{-1}B^T)$ denote the actual Schur complement and S_2^{-1} denote our approximation to its inverse. As we only need to apply S_2^{-1}, no explicit representation of S_2 is needed. Finally, let $S_2^{-1}S_1 = I + \mathcal{E}$.

4.1. Eigenvalue analysis of the block-diagonally preconditioned system. Now, the block-diagonal preconditioner is as follows:

$$\mathcal{P}(F, S_2) = \begin{bmatrix} F^{-1} & 0 \\ 0 & S_2^{-1} \end{bmatrix}.$$

We multiply (1.1) from the left by $\mathcal{P}(F, S_2)$. We refer to the resulting preconditioned matrix as $\mathcal{B}(F, S_2)$. The system of equations with $\mathcal{B}(F, S_2)$ is as follows:

$$\begin{bmatrix} I - S & N \\ M_2 & Q_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \left(\begin{bmatrix} I & N \\ M & Q \end{bmatrix} - \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \tag{4.1}$$

where M, N, and Q are defined as in section 2, $M_2 = S_2^{-1}C$, and $Q_2 = S_2^{-1}D$. Note also that $M_2 = S_2^{-1}S_1S_1^{-1}C = (I+\mathcal{E})M$ and analogously $Q_2 = (I+\mathcal{E})Q$. Using (4.1), we can bound the eigenvalues of $\mathcal{B}(F,S_2)$ by considering the perturbation of the eigenvalues of $\mathcal{B}_0$ analogously to our bounds in section 2.2.

THEOREM 4.1. *Let $\lambda_{\mathcal{B}}$ be an eigenvalue of $\mathcal{B}(F,S_2)$, λ be an eigenvalue of $\mathcal{B}_0$, and $Qv_j = \delta_j v_j$.*

1. *If $\delta_j \in \mathbb{R}$ for $j = 1,\dots,m$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \le (1+\sqrt{2})\kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\| + \max_j\left\{|1+\delta_j\lambda_j^+|, |1+\delta_j\lambda_j^-|\right\}\kappa(V)\|\mathcal{E}\|.$$

2. *If $\delta_j \in \mathbb{C}$ and $\exists\alpha > 0$ s.t. $|\delta_j| \le \alpha < \sqrt{5}$ for $j = 1,\dots,m$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \le 2\max\left(1, \frac{1}{2} + \frac{1+\alpha}{2\sqrt{2\left(\sqrt{5}-\alpha\right)}}\right)\kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\| + \frac{2+(1+\sqrt{5})\alpha+2\alpha^2}{\sqrt{2\left(\sqrt{5}-\alpha\right)}}\kappa(V)\|\mathcal{E}\|.$$

3. *If $D = 0$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \le 2\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\| + \frac{2\sqrt{5}}{5}\|\mathcal{E}\|.$$

Proof. In section 2.1 we have already derived the eigendecomposition of $\mathcal{B}_0$. From this decomposition we get the following perturbation bound (see [33, Theorem IV.1.12]):

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\le \left\|\mathcal{Y}^{-1}\begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix}\mathcal{Y}\right\| \\ &\le \left\|\mathcal{Y}^{-1}\begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\mathcal{Y}\right\| + \left\|\mathcal{Y}^{-1}\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y}\right\|. \end{aligned} \tag{4.2}$$

Corollary 2.4 gives bounds for the first term in (4.2). So, we need bounds only for the second term.

Define $\mathcal{X}$ such that

$$\mathcal{X} = \mathcal{Y}^{-1}\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y}.$$

We have

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y} = \begin{bmatrix} 0 & 0 \\ -\mathcal{E}(MY_{11}+QY_{21}) & -\mathcal{E}(MY_{12}+QY_{22}) \end{bmatrix},$$

where $MU_1 = 0$ and $MU_2 = MNV = (I+Q)V = V(I+\Delta)$. This gives $MY_{12} = MU_2 = V(I+\Delta)$, $MY_{11} = [0\ V(I+\Delta)]$, $QY_{22} = V\Delta(\Lambda^- - I)$, and $QY_{21} = [0\ V\Delta(\Lambda^+ - I)]$. So, the previous equation reduces to

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y} = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & -\mathcal{E}V(I+\Delta\Lambda^+) & -\mathcal{E}V(I+\Delta\Lambda^-) \end{array}\right]. \tag{4.3}$$

We then multiply (4.3) from the left by $\mathcal{Y}^{-1}$ (see (2.14)–(2.17)) and refactor to yield

$$\mathcal{X} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & (\Lambda^- - \Lambda^+)^{-1} & 0 \\ 0 & 0 & -(\Lambda^- - \Lambda^+)^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \end{bmatrix} \mathcal{W},$$

where

$$\mathcal{W} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I + \Delta\Lambda^+ & 0 \\ 0 & 0 & I + \Delta\Lambda^- \end{bmatrix}.$$

Using the consistency of the 2-norm, we have the following bound on $\|\mathcal{X}\|$:

$$\|\mathcal{X}\| \leq 2\|(\Lambda^- - \Lambda^+)^{-1}\| \max_j \left\{|1 + \delta_j\lambda_j^+|, |1 + \delta_j\lambda_j^-|\right\} \kappa(V)\|\mathcal{E}\|. \tag{4.4}$$

The remainder of the proof concerns the bounds on the right-hand side of (4.4) for each particular case.

For the first part of the theorem, assume $\delta_j \in \mathbb{R}$ for $j = 1, \ldots, m$. We have

$$\begin{aligned} \lambda_j^- - \lambda_j^+ &= \frac{1 + \delta_j - \sqrt{4 + (1+\delta)^2}}{2} - \frac{1 + \delta_j + \sqrt{4 + (1+\delta)^2}}{2} = -\sqrt{4 + (1+\delta_j)^2} \\ &= -\sqrt{p(\delta)}. \end{aligned}$$

Clearly, $|1/(\lambda_j^- - \lambda_j^+)|$ obtains its maximum at $\delta_j = -1$. This yields $|1/(\lambda_j^- - \lambda_j^+)| \leq 1/2$. We can use this in (4.4) to complete the proof of the first bound.

For the second part of the theorem, we assume $\exists \alpha > 0$ s.t. $|\delta_j| \leq \alpha < \sqrt{5}$ for $j = 1, \ldots, m$. First we derive a bound for $\|(\Lambda^- - \Lambda^+)^{-1}\|$. Recall the lower bound on $p(\delta)$ in the proof of Lemma 2.2 and note that $|1/(\lambda_j^- - \lambda_j^+)| = 2/\sqrt{|p(\delta_j)|}$. So, we have $\|(\Lambda^- - \Lambda^+)^{-1}\| \leq (2(\sqrt{5} - \alpha))^{-1/2}$. Furthermore, we have

$$|1 + \delta_j\lambda_j^\pm| = \left|1 + \delta_j \frac{1 + \delta_j \pm \sqrt{4 + (1+\delta_j)^2}}{2}\right| \leq 1 + \frac{|\delta_j||1 + \delta_j| + |\delta_j|\sqrt{|4 + (1+\delta_j)^2|}}{2}.$$

We can bound $|\delta + 1 - 2i|$ and $|\delta + 1 + 2i|$ from above by $\sqrt{5} + \alpha$; so, $\sqrt{|4 + (1+\delta_j)^2|} \leq \sqrt{5} + \alpha$. Thus, we have

$$|1 + \delta_j\lambda_j^\pm| \leq 1 + \frac{\alpha(1+\alpha) + \alpha\left(\sqrt{5} + \alpha\right)}{2} = 1 + \frac{1 + \sqrt{5}}{2}\alpha + \alpha^2.$$

Substituting these bounds into (4.4) yields

$$\|\mathcal{X}\| \leq \frac{2 + (1 + \sqrt{5})\alpha + 2\alpha^2}{\sqrt{2\left(\sqrt{5} - \alpha\right)}} \kappa(V)\|\mathcal{E}\|. \tag{4.5}$$

We can then substitute this result into (4.2) to prove the second part of the theorem.

For the third part of the theorem, we assume $D = 0$. We bound the first term in (4.2) using Theorem 2.1, Lemma 2.2 for $\delta \geq -1$, and Lemma 2.3 where $\kappa(\Theta) = 1$. This follows from the fact that U_2 can be chosen to be orthogonal (see [8]).

For the second term in (4.2), since $Q = 0$, $\delta_j = 0$, so $\lambda_j^- - \lambda_j^+ = -\sqrt{5}$, and we can choose $V = I$. We then substitute this into (4.4). ∎

In practice, in the complex case the term involving α will generally be modest. For example, if $\alpha = 1$, it is about 4.6022, and for $\alpha = 2$, it is about 23.9727.

If we compare the bounds from Theorem 4.1 with those from Corollary 2.4 for the block-diagonal preconditioner with the exact Schur complement, $(D - CF^{-1}B^T)$, we see that the deterioration of the bounds is $O(\|\mathcal{E}\|)$. Note that the factors that multiply the $\|\mathcal{E}\|$ are all constants with respect to the choice of the approximate Schur complement, S_2^{-1}. This is about as good as we can hope for. The bounds also demonstrate that there is no point in investing in a really good splitting when a poor approximation to the Schur complement is used or vice versa. Rather, we should be equally attentive to both if we want good eigenvalue clustering.

4.2. Eigenvalue analysis of the related system. If we follow the approach in section 3 to generate the related system for this problem, we would generate precisely the related system derived from (3.3), with S_1^{-1} instead of S_2^{-1} [8]. Therefore, we use an alternative splitting of $\mathcal{B}(F)$,

$$\mathcal{B}(F) = \begin{bmatrix} I & N \\ M_2 & Q_2 + \mathcal{E} \end{bmatrix} - \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix},$$

and derive the related system for this splitting. Due to the $\mathcal{E}$ term in the splitting, however, we cannot reduce the size of our system. Instead, we get

$$\begin{bmatrix} I - (I - NM_2)S & -N\mathcal{E} \\ -M_2 S & I + \mathcal{E} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{4.6}$$

For a special problem in magnetostatics, a linear system similar to (4.6) was derived in [26]. If we use the choices for the splitting and approximations from [26], we obtain basically the same system to be solved. In [26], the authors only outline the qualitative behavior of the eigenvalues in the case that $\mathcal{E}$ is sufficiently small.

THEOREM 4.2. *For any eigenvalue, λ_R, of the related system matrix* (4.6),

$$|1 - \lambda_R| \le \sqrt{1 + \|N\|^2}\sqrt{1 + \|M_2\|^2}\max\left(\|S\|, \|\mathcal{E}\|\right).$$

Proof. Note that the matrix in (4.6) can be split as follows:

$$\begin{aligned} \begin{bmatrix} I - (I - NM_2)S & -N\mathcal{E} \\ -M_2 S & I + \mathcal{E} \end{bmatrix} &= I - \begin{bmatrix} I - NM_2 & N \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix} \\ &= I - \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix}. \end{aligned}$$

Expressing our matrix as a perturbation of the identity and using a classic perturbation bound (see [33]) yields

$$|1 - \lambda_R| \le \left\| \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix} \right\|.$$

Noting that

$$\left\| \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \right\| \le \sqrt{1 + \|N\|^2} \qquad \text{and} \qquad \left\| \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \right\| \le \sqrt{1 + \|M_2\|^2},$$

we obtain

$$|1 - \lambda_R| \le \sqrt{1 + \|N\|^2}\sqrt{1 + \|M_2\|^2}\max\left(\|S\|, \|\mathcal{E}\|\right). \qquad \square$$

The terms $\|N\|$ and $\|M_2\|$ in the bound from Theorem 4.2 are fairly benign. They are bounded by the norms of the off-diagonal blocks of the unpreconditioned matrix (1.1) and the norms of the inverses of the splitting and approximate Schur complement. Note that the latter two are chosen by the user. Moreover, if we use a good preconditioner for this problem and therefore both our splitting and approximate Schur complement are reasonably accurate, the norms of their inverses will not be large relative to the norm of (1.1), unless (1.1) is itself poorly conditioned.

Just as for the block-diagonally preconditioned system, the eigenvalue perturbation of the related system depends on both $\|S\|$ and $\|\mathcal{E}\|$. Again, there is no advantage in making one significantly smaller than the other. Thus, we should be equally attentive to both $\|S\|$ and $\|\mathcal{E}\|$ in order to achieve tight clustering and fast convergence.

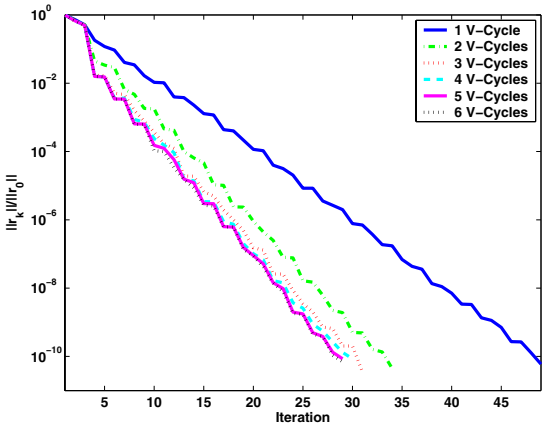
5. Numerical experiments. We present numerical experiments for two model problems, both arising from the Navier–Stokes equations.

The first model problem involves a stabilized finite element discretization of the Navier–Stokes equations. We use the software toolkit for a two-dimensional leaky lid-driven cavity problem developed for the Winter School in Scientific Computing and Iterative Methods hosted by the Chinese University of Hong Kong in December 1995 and made available by David Silvester [11]. Using this toolkit, we can easily apply the preconditioners and analysis from this paper to the stabilized Navier–Stokes problem (Oseen case). This problem is nonsymmetric but has $B = C$. Excellent work has been done by others on preconditioners for this specific problem [11, 31, 34], which we do not intend to supplant. Rather, our goal is to illustrate the effect of the preconditioners proposed in this paper on the convergence behavior and the eigenvalue distributions for a problem which is well understood and easily accessible to the community.

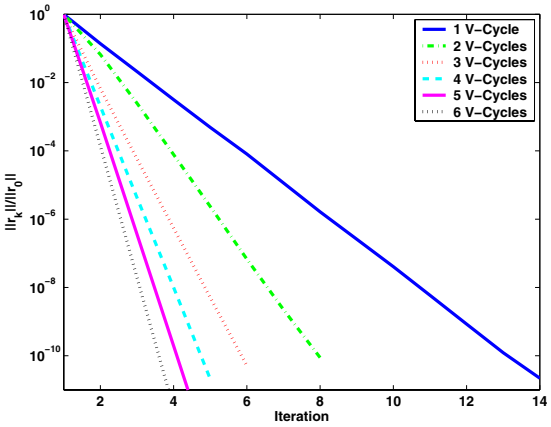
In particular, we show what happens to the convergence of GMRES, the eigenvalues, and our eigenvalue bounds as we improve the splitting ($\|S\| \to 0$) and the approximate Schur complement ($\|\mathcal{E}\| \to 0$). We also succinctly compare the block-diagonally preconditioned systems (2.3) and (4.1) with the related systems (3.4) and (4.6), in terms of both eigenvalues and convergence. We also illustrate the importance of *balancing* the quality of the splitting and the Schur complement to avoid wasted effort. Finally, we study the influence of the mesh width on the convergence of the related system.

The second model problem involves a spectral collocation discretization for the incompressible Stokes equations on a square [3, 27]. This application has $B \neq C$ and $D = 0$, and this particular formulation uses the Chebyshev nodes for the collocation sites to allow the rapid computation of Gauss–Lobatto quadrature. To our knowledge, this is the first presentation of convergence and eigenvalue results in the literature for preconditioners for generalized saddle-point problems with $B \neq C$. For this application, we present GMRES convergence results as well as the locations of the eigenvalues of the preconditioned system.

5.1. Navier–Stokes with finite elements. For our first experiments, we choose a 16×16 grid, viscosity parameter $\nu = 0.1$, and stabilization parameter $\beta = 0.25$. After removing the constant pressure mode, the system has 705 unknowns. Since multigrid cycles are actually matrix splittings, we use a number of multigrid V-cycles to define the splitting of the (1,1) block. For each V-cycle we use three SOR-Jacobi pre- and post-smoothing steps with relaxation parameter $\omega = 0.25$. As a purely algebraic alternative, we also employ an ILUT factorization of the (1,1) block and vary the drop tolerance to change the accuracy of our splitting [28].



(a) Block-Diagonal Preconditioner (2.3).

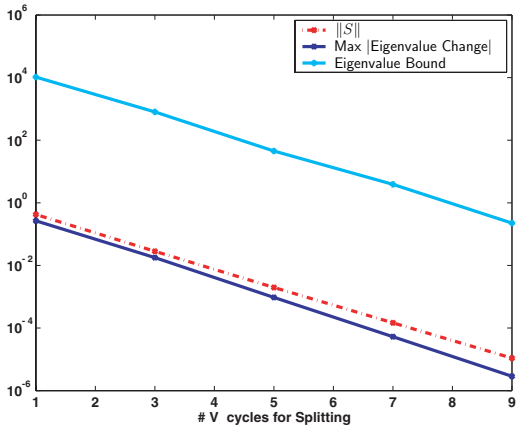


(b) Related System (3.4).

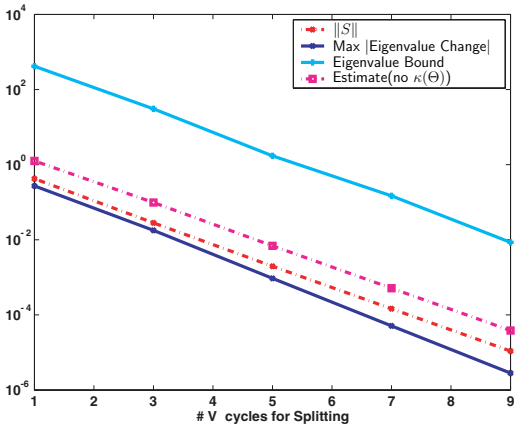
FIG. 5.1. *Convergence of GMRES for both types of preconditioners, using the exact Schur complement and varying the number of V-cycles for the splitting.*

We start with the exact Schur complement, varying the number of V-cycles for the splitting from one to six. Figures 5.1(a) and 5.1(b) show the convergence history for preconditioned GMRES for the block-diagonally preconditioned system and the related system, respectively. Note that the related system converges in significantly fewer iterations, for any choice of the number of V-cycles, demonstrating the performance difference between the two preconditioned systems.

We have also computed the eigenvalue perturbation and the eigenvalue bounds for both preconditioned systems, using up to nine V-cycles for the splitting, with the exact Schur complement. Figure 5.2(a) shows the maximum absolute eigenvalue perturbation from $\lambda \in \{1, \lambda_j^{\pm}\}$ for the block-diagonally preconditioned system (2.3), and Figure 5.2(b) shows the maximum absolute eigenvalue perturbation from 1 for the related system (3.4).



(a) Block-Diagonal Preconditioner (2.3).

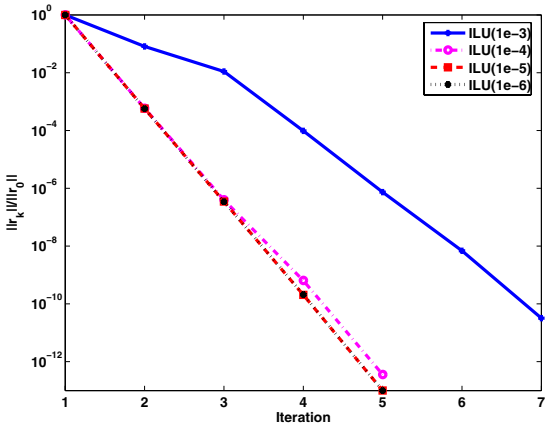


(b) Related System (3.4).

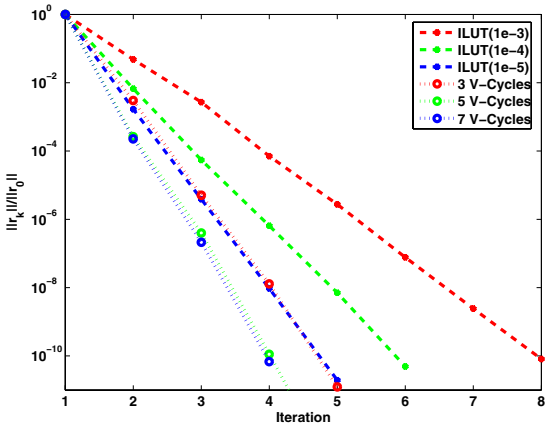
FIG. 5.2. *Maximum absolute eigenvalue perturbation and perturbation bounds, for both types of preconditioners, using the exact Schur complement and varying the number of V-cycles for the splitting.*

As we use a better splitting for A (more V-cycles), we see that the eigenvalue bound decreases with approximately the same rate as the corresponding eigenvalue perturbations, although the bound is pessimistic. This pessimism is mostly due to the $\kappa(\Theta)$ factor. Figure 5.2(b) includes an *estimate* of the perturbation for the related system, which consists of the bound in Corollary 3.2 with $\kappa(\Theta)$ replaced by one. Both the bound and our estimate follow the trend in the actual eigenvalue perturbation well as the number of V-cycles increases. The figure shows that the bounds and the estimate give good qualitative, respectively quantitative, descriptions of the eigenvalue perturbation as the splitting improves.

The eigenvalue perturbation bound for the related system (3.4) is much smaller than for the block-diagonally preconditioned system (2.3). However, the actual max-



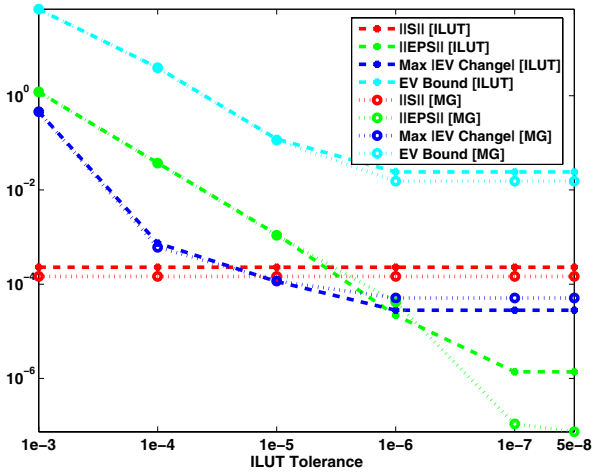
(a) Using five V-cycles for the splitting of the (1,1) block and varying the approximate Schur complement.



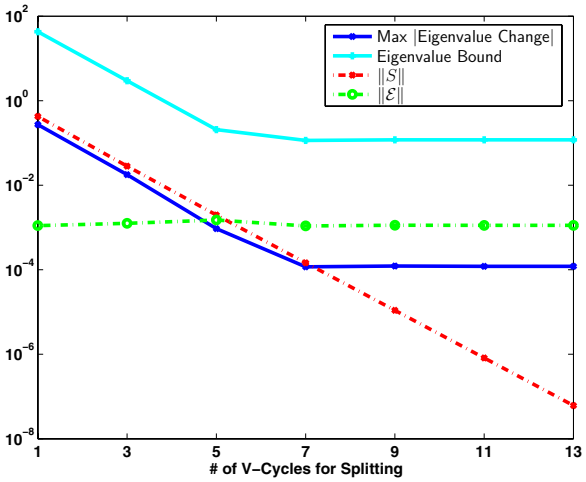
(b) Using the approximate Schur complement with ILUT($1e-5$) and varying the splitting of the (1,1) block (# V-cycles and ILUT tolerance).

FIG. 5.3. *Convergence results for the related system using an approximate Schur complement.*

imum eigenvalue perturbation for both systems is about equal. For the related system, this represents a single eigenvalue cluster around 1, which means that the bound proves fast convergence for about 6 V-cycles or more, and the actual (max) perturbation indicates good convergence already for 1 V-cycle. On the other hand, for the block-diagonally preconditioned system, this represents $2m+1$ (potentially) distinct clusters around 1 and $\lambda_j^{\pm}$ for $j = 1, \ldots, m$. The existence of multiple clusters in this case, compared with the single cluster for the related system, explains the difference in their convergence behavior. These multiple clusters also explain the diminishing returns of improving the splitting for the block-diagonal preconditioner shown in



(a) Using seven V-cycles or ILUT(1e − 5) for the splitting and varying the approximate Schur complement.



(b) Using the approximate Schur complement with ILUT(1e − 5) and varying the number of V-cycles for the splitting.

FIG. 5.4. *The effects of* $\|S\|$ *and* $\|\mathcal{E}\|$ *on the related system using the approximate Schur complement.*

Figure 5.1(a). As we see similar differences between the preconditioners for the other test cases, we show results only for the related system for the remainder of this section.

We illustrate the convergence behavior for the preconditioner with an approximate Schur complement as a function of the accuracy of the approximation by using an ILUT decomposition [28]. While this may not be a practical choice, it serves our purposes for this paper because it allows us to progressively increase the accuracy of

TABLE 5.1

Effect of the number of grid points per dimension (n) on $\max_j |\delta_j|$ and the number of GMRES iterations for the related system (4.6) using a splitting of 5 V-cycles and various approximate Schur complements.

		Number of GMRES iterations			
n	$\max \lvert\delta_j\rvert$	ILUT(1e-3)	ILUT(1e-4)	ILUT(1e-5)	ILUT(1e-6)
4	1.72e+00	5	5	5	5
8	5.92e-01	5	4	4	4
16	1.60e-01	7	5	5	5
32	4.07e-02	13	6	5	5

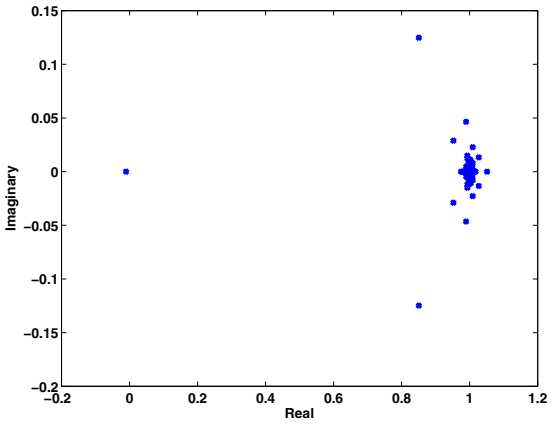
the approximation to the inverse of the Schur complement. We use drop tolerances ranging from $1e-3$ to $5e-8$.

Figures 5.3(a) and 5.3(b) show the effects of improving the splitting (for multigrid and ILUT) and the approximation to the Schur complement on the convergence of GMRES for the related system (4.6). First, in Figure 5.3(a), we vary the drop tolerance for the approximate Schur complement and fix the number of V-cycles for the splitting at five. Then, in Figure 5.3(b), we demonstrate a number of splittings using V-cycles and ILUT, and fix the drop tolerance at $1e-5$ for the approximate Schur complement. The convergence results are quite good, regardless of the choice of splitting.

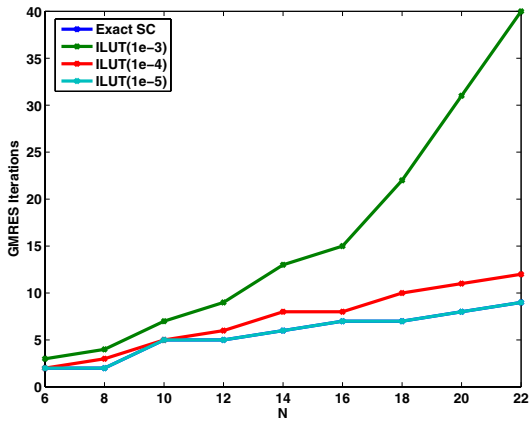
The convergence rates in Figures 5.3(a) and 5.3(b) hit a point of diminishing returns, past which improving either the splitting or the approximate Schur complement while leaving the other unchanged does not improve convergence. To explain this, we show the eigenvalue perturbations from 1 and the perturbation bound for the same example in Figure 5.4. In both plots, the eigenvalue perturbation (and bound) cease to decrease shortly after $\|S\|$ is less than $\|\mathcal{E}\|$ or vice versa. This demonstrates that the eigenvalue bound from Theorem 4.2 is indicative of the actual eigenvalue perturbation and the resulting convergence behavior, and that using a significantly more accurate splitting than approximate Schur complement, or vice versa, yields little additional benefit. Finally, note that for reasonable choices of splitting and approximation to the Schur complement the bounds are less than 1, indicating that the eigenvalues are clustered away from the origin. This should lead to rapid convergence for Krylov methods.

Varying the number of grid points per dimension, $n = 1/h$, gives some insight into how the convergence of the related system (4.6) depends on h. Table 5.1 summarizes these results. First, note that $|\delta_j|$ decreases with h. This leads to significant reductions of the factors involving δ_j in the theorems of sections 2, 3, and 4. In particular, with respect to Corollary 3.2 for the related system and Corollary 2.4 and Theorem 4.1 for the block-diagonal preconditioner, note that for small h the δ_j are nearly real. Moreover, note that the convergence of GMRES for the related system (4.6) depends only mildly on h. A good splitting and a reasonably accurate approximate Schur complement seem to lead to h-independent convergence.

5.2. Incompressible Stokes with spectral collocation. We will build discretizations with polynomials of degree up to 22 for this problem. The largest system will be of size 1241. We use an odd-even ordering for the velocity unknowns to exploit the orthogonality properties of Chebyshev polynomials and put the (1,1) block in block-diagonal form. We use ILUT with a drop tolerance of $1e-4$ for the splitting of the (1,1) block, and for the approximate Schur complement we use ILUT with a drop tolerance between $1e-3$ and $1e-5$. Figure 5.5(a) shows the eigenvalues of the



(a) Eigenvalues of related system for polynomial degree $N = 22$ using an approximate Schur complement with ILUT(1e − 4).



(b) GMRES iteration count versus maximum polynomial degree (N) for various approximate Schur complements. The iteration counts for the exact Schur complement coincide with those for the approximate Schur complement with ILUT(1e-5).

FIG. 5.5. *Eigenvalues and iteration counts for the related system* (4.6) *from spectral discretization of the incompressible Stokes equations with an ILUT*(1e−4) *splitting and an approximate Schur complement.*

related system for the largest problem, $N = 22$. Except for a single eigenvalue of $O(1e-2)$, the eigenvalues are tightly clustered around one. As expected, this leads to rapid convergence, as shown in Figure 5.5(b). Moreover, the GMRES iteration count for the related system with an approximate Schur complement (with the exception of ILUT(1e − 3)) shows only modest dependence on the maximum polynomial degree

N. Hence, even for fully asymmetric problems, our preconditioners are effective and show the potential of scaling well to larger problems.

6. Conclusions and future work. We have proposed and analyzed variants of indefinite preconditioners (the related system) and block-diagonal preconditioners for the $D \neq 0$ case, including the use of approximate Schur complements. We have illustrated their performance in terms of convergence, eigenvalue perturbations, and eigenvalue bounds using well-known model problems. Further analysis should help tighten the eigenvalue bounds, in particular using the consistency property of matrix norms less. We also aim to specialize our methods to particular problems. We are currently exploring applications from metal deformation, porous media flow, optimization, and electromagnetics.

Acknowledgments. We gratefully acknowledge the use of the software toolkit for a two-dimensional leaky lid-driven cavity problem developed by David Silvester in collaboration with Howard Elman, Bernd Fischer, Alison Ramage, and Andy Wathen.

We also gratefully acknowledge the reviewers for many suggestions that helped improve this paper.

REFERENCES

[1] M. Benzi and G. H. Golub, *A preconditioner for generalized saddle point problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 20–41.

[2] M. Benzi, M. J. Gander, and G. H. Golub, *Optimization of the Hermitian and skew-Hermitian splitting iteration for saddle-point problems*, BIT, 43 (2003), pp. 881–900.

[3] C. Bernardi, C. Canuto, and Y. Maday, *Generalized inf-sup conditions for Chebyshev spectral approximation of the Stokes problem*, SIAM J. Numer. Anal., 25 (1988), pp. 1237–1271.

[4] D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[5] D. Braess and R. Sarazin, *An efficient smoother for the Stokes problem*, Appl. Numer. Math., 23 (1997), pp. 3–19.

[6] D. Braess, P. Deuflhard, and K. Lipnikov, *A subspace cascadic multigrid method for mortar elements*, Computing, 69 (2002), pp. 205–225.

[7] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.

[8] E. de Sturler and J. Liesen, *Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems,* I: *Theory*, SIAM J. Sci. Comput., 26 (2005), pp. 1598–1619.

[9] H. C. Elman, *Preconditioning for the steady-state Navier–Stokes equations with low viscosity*, SIAM J. Sci. Comput., 20 (1999), pp. 1299–1316.

[10] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[11] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Iterative methods for problems in computational fluid dynamics*, in Winter School on Iterative Methods in Scientific Computing and Applications, Chinese University of Hong Kong, Hong Kong, 1996.

[12] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers*, Oxford University Press, New York, 2005.

[13] A. Forsgren, P. E. Gill, and M. H. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.

[14] G. H. Golub and A. J. Wathen, *An iteration for indefinite systems and its application to the Navier–Stokes equations*, SIAM J. Sci. Comput., 19 (1998), pp. 530–539.

[15] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.

[16] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[18] I .C. F. Ipsen, *A note on preconditioning nonsymmetric matrices*, SIAM J. Sci. Comput., 23 (2001), pp. 1050–1051.
[19] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.
[20] P. Krzyżanowski, *On block preconditioners for nonsymmetric saddle point problems*, SIAM J. Sci. Comput., 23 (2001), pp. 157–169.
[21] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.
[22] C. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[23] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput., 21 (2000), pp. 1969–1972.
[24] R. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems*, SIAM J. Numer. Anal., 19 (1982), pp. 349–357.
[25] M. Parks, E. de Sturler, G. Mackey, D. Johnson, and S. Maiti, *Recycling Krylov Subspaces for Sequences of Linear Systems*, Technical report UIUCDCS-R-2004-2421, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 2004.
[26] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.
[27] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, 2nd ed., Springer-Verlag, New York, 1997.
[28] Y. Saad, *ILUT: A dual threshold incomplete ILU factorization*, Numer. Linear Algebra Appl., 1 (1994), pp. 387–402.
[29] C. Siefert, *Preconditioners for Generalized Saddle-Point Problems*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 2005.
[30] C. Siefert and E. de Sturler, *Preconditioners for Generalized Saddle-Point Problems*, Technical report UIUCDCS-R-2004-2448, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 2004.
[31] D. Silvester and A. Wathen, *Fast iterative solution of stabilized Stokes systems,* II: *Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.
[32] D. Silvester, H. Elman, D. Kay, and A. Wathen, *Efficient preconditioning of the linearized Navier-Stokes equations for incompressible flow*, J. Comput. Appl. Math., 128 (2001), pp. 261–279.
[33] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[34] A. Wathen and D. Silvester, *Fast iterative solution of stabilized Stokes systems,* I: *Using simple diagonal preconditioners*, SIAM J. Numer. Anal, 30 (1993), pp. 630–649.
[35] L. Zhu, A. J. Beaudoin, and S. R. MacEwan, *A study of kinetics in stress relaxation of* AA 5182, in Proceedings of TMS Fall 2001: Microstructural Modeling and Prediction During Thermomechanical Processing, Indianapolis, IN, 2001, pp. 189–199.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1297–1313

STABILITY PROPERTIES OF THE PERONA–MALIK SCHEME*

SELIM ESEDOĞLU†

Abstract. The Perona–Malik scheme is a numerical technique for denoising digital images without blurring object boundaries (edges). In general, solutions generated by this scheme do not satisfy a comparison principle. We identify conditions under which two solutions initially ordered remain ordered, and we state (restricted) comparison principles. These allow us to study stability properties of the scheme. We also explore what these results say in the limit as the discretization size goes to 0.

Key words. computer vision, nonlinear diffusion, Perona–Malik equation

AMS subject classifications. 65M12, 68U10

DOI. 10.1137/S0036142903424817

1. Introduction.

Denoising is a fundamental procedure in digital image processing and an essential tool in many computer vision applications such as edge detection and segmentation. Its goal is to estimate a clean image from a given corrupted one. Mathematically, a two-dimensional grayscale image can be represented as a function $f(x)$ mapping a domain D in the plane (usually a rectangle, representing the computer screen) to the unit interval $[0, 1]$. The value of the function f at a given point $x \in D$ then represents the grayscale intensity (brightness) of the pixel found at that location: for instance, 0 can represent black, and 1 white. Image denoising tries to reduce, usually by some averaging operation, the rapid oscillations in f that are due to the presence of noise.

Partial differential equation (PDE) based image denoising models have enjoyed a great deal of success and have become very popular in the field. In this approach, the given noisy image f is taken to be the initial condition for some parabolic PDE, which is solved for a length of time chosen by the user. The solution at this later time is then taken to be the denoised version of the image. One of the first and most elementary denoising techniques, namely, convolution of the image by a Gaussian kernel, can be interpreted this way: it is the solution of the standard heat equation with the original noisy image taken as the initial condition, where the variance of the Gaussian kernel used is related to the length of time for which the PDE is solved. Although this is an effective denoising technique, it has the important disadvantage of blurring boundaries of objects (edges) in the image, where the function f has large gradients or a discontinuity. From an applications point of view, what is desired is a denoising method that can preserve sharp object boundaries.

In their seminal papers [6, 7], Perona and Malik proposed a numerical scheme for edge preserving image denoising that appears to be the finite difference discretization of a nonlinear diffusion equation which can become backward parabolic. The continuous in time version of their scheme has the following form:

$$\dot{u}_{i,j}(t) = D_1^- \left(R_\kappa \left(D_1^+ u_{i,j} \right) \right) + D_2^- \left(R_\kappa \left(D_2^+ u_{i,j} \right) \right), \tag{1}$$

*Received by the editors July 31, 2003; accepted for publication (in revised form) December 23, 2005; published electronically July 7, 2006.
http://www.siam.org/journals/sinum/44-3/42481.html
†Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (esedoglu@umich.edu).

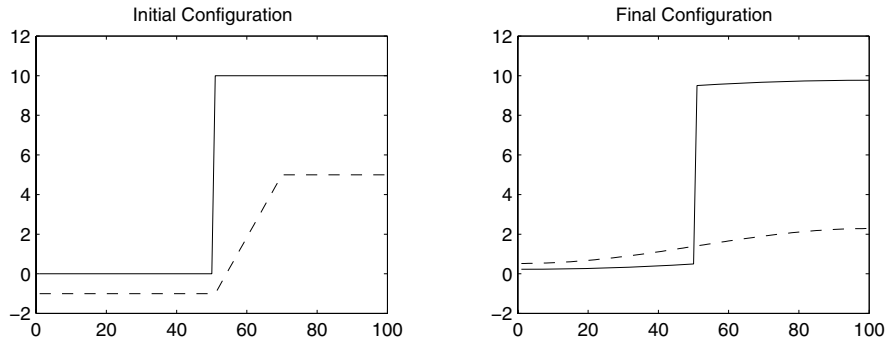


FIG. 1. *Violation of comparison principle. The solution represented by the dashed line remains in the well-posed (parabolic) regime at all times. Nevertheless, order is lost. It is very easy to understand why this happens: The dashed line has a small enough slope to be completely in the parabolic regime of scheme* (1)*, so that it is evolved essentially by the heat equation toward a constant state (its average). On the other hand, the solid line hardly evolves at all because its large gradient in the middle is far into the "edge preserving" regime of the scheme, where such "discontinuities" are maintained for large times by the scheme, which was designed precisely to behave as such.*

where D_m^+ and D_m^- are the standard forward and backward difference operators in the mth coordinate direction, and the function R_κ satisfies important properties which we explain in section 2. The indices i and j run over the pixels arranged in a two-dimensional uniform grid. In image processing literature, it is common to use either periodic or homogeneous Neumann boundary conditions so as to keep constant the total intensity of the image being processed.

The motivation of Perona and Malik in proposing their method (1) was to replace the standard heat equation with a nonlinear parabolic PDE which is designed to suppress diffusion at regions of large gradient, since such regions are likely to contain edges. Indeed, scheme (1) seems to be a natural discretization for the PDE

$$u_t = (R_\kappa(u_x))_x + (R_\kappa(u_y))_y. \tag{2}$$

The essence of Perona and Malik's technique is contained in the choice of the function $R_\kappa(\xi)$; the success of the scheme in preserving sharp edges (until they disappear) is due to this choice. But for the functions R_κ that Perona and Malik advocate in their papers, PDE (2) becomes backward parabolic in regions where the gradient of the solution is larger than some threshold that depends on the parameter κ. As such, there is no well-posedness theory for this PDE. That makes it interesting to investigate continuum limits (limits as $h \to 0$) for scheme (1). And understanding how the behavior of a denoising technique such as (1) depends on the discretization step size is important since it is very common to have the same image at various resolutions.

In this paper we state a *restricted* comparison principle for the semidiscrete Perona–Malik scheme (1). We note that a *restriction* of the kind we consider is necessary because the standard comparison result does not hold for scheme (1). We illustrate this easy-to-see fact by a simple numerical example (Figure 1). Subsequently, we apply the comparison principle to explore stability properties of the scheme at fixed discretization step size $h > 0$ and also in the limit $h \to 0^+$. We also expose some effects the precise shape of the function R_κ has on the behavior of the scheme.

In practice, scheme (1) exhibits much better stability properties than one would expect from backward diffusion equations [11], which are notoriously ill-posed. Moreover, it is extremely effective at its intended purpose. It has therefore become an intriguing issue to explain the better than expected stability properties of Perona and Malik's technique. Some aspects of this surprisingly tame behavior have been explained by previous authors; we believe with this paper we further our understanding of this problem.

Another interesting issue is what effects the precise shape of the nonlinear function R_κ has on the scheme. Perona and Malik, and subsequently many other authors, reported numerical experiments with a variety of choices (each of which conforms to the fundamental properties we listed in section 2), and on occasion mentioned differences in observed behavior [8]. Indeed, based on numerical experiments, even with functions R_κ that have identical parabolicity thresholds, the behavior can still be quite different. The results presented in this paper allow us to reveal and quantify some differences.

2. Perona–Malik scheme. As we remarked above, it is the choice of $R_\kappa(\xi)$ that distinguishes Perona and Malik's technique from previous techniques. In [6], they report numerical experiments using scheme (1) with

$$R_\kappa(\xi) = \frac{\xi}{1+\xi^2/\kappa} \text{ and } R_\kappa(\xi) = \xi \exp\left(\frac{-\xi^2}{\kappa}\right).$$

Other choices used in practice include

$$R_\kappa(\xi) = \xi\left(1+\frac{\xi^2}{\kappa}\right)^{(\beta-1)}, \text{ where } \beta \in \left(0, \frac{1}{2}\right).$$

These choices share the following essential characteristics:

1. $\xi R_\kappa(\xi) \geq 0$ for all ξ.
2. The parameter κ defines a positive critical value $z(\kappa)$ such that

$$R'_\kappa(\xi) \begin{cases} < 0 \text{ for } |\xi| > z(\kappa), \text{ and} \\ \geq 0 \text{ otherwise.} \end{cases} \tag{3}$$

3. $R_\kappa(\xi) \to 0$ as $|\xi| \to \infty$.

Figure 2 illustrates $R_\kappa(\xi)$ for such a choice. The desirable properties of the Perona–Malik scheme seem to hold in practice whenever the function $R_\kappa(\xi)$ in (1) satisfies the properties above. The results presented in this paper apply to (1) under all such choices of $R_\kappa(\xi)$.

In (2), R'_κ appears as the diffusion coefficient. Therefore, as we indicated in the previous section, the parameter κ constitutes a threshold value: when the gradient of grayscale intensity, $D_m^+ u_{i,j}$, is large compared to κ, (2) violates parabolicity.

Encouraged by favorable numerical results, some previous mathematical work on the Perona–Malik technique deals with understanding whether (2) can be given a well-posedness theory, so that the PDE (2) can be properly understood as the continuum limit of scheme (1). The paper [5] by Kichenassamy and the paper [4] by Kawohl and Kutev pursue this direction. The present work is drastically different from them in spirit: we do not try to understand (2) at all; instead, we deal directly with scheme (1) and study its properties. This is also the approach pursued in [2], where a continuum limit for (1) is established that differs from (2).

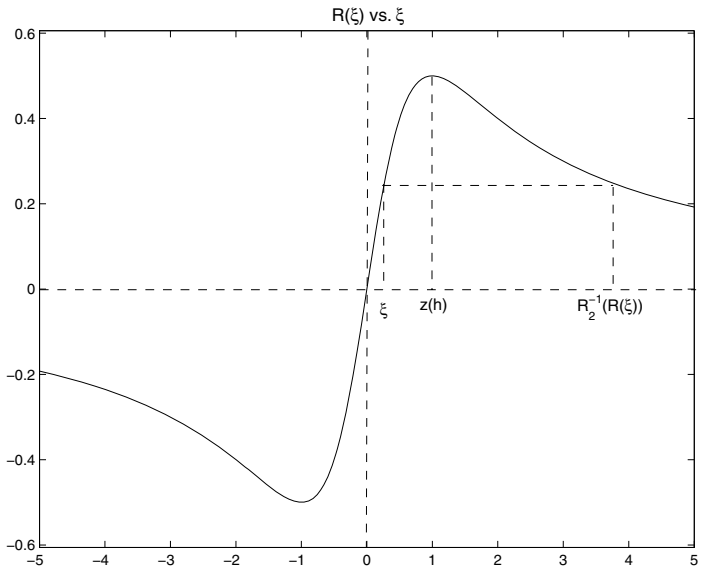


FIG. 2. *A typical choice for the function $R_h(\xi)$ that appears in the scheme. Here $R_h(\xi) = \xi/(1+\xi^2/h)$ with $h = 1$.*

Nevertheless, we make use of ideas from the work of these previous researchers; indeed, the motivation for this paper came from [4]. There, Kawohl and Kutev establish a restricted comparison principle for *continuum* solutions of (2). In this paper, we strive to find conditions under which two *discrete* solutions generated by the scheme (1) can be compared. In the end, however, we did not obtain direct discrete analogues of Kawohl and Kutev's results; the conditions and results in this paper are entirely different.

Let us note that computer vision is not the only context in which Perona–Malik-type equations and their associated issues come up. The recent paper [10] proposes a PDE model for granular flow that has much in common with the Perona–Malik equation and presents analysis directed at questions closely related to the ones raised in the image denoising literature.

Finally, we must mention that there are other very successful variational, edge preserving, image denoising techniques. One example is the total variation based image denoising model of Rudin, Osher, and Fatemi (ROF) [9]. Unlike the Perona–Malik scheme, the nonlinear PDE involved in this model has a well-understood theory, as it never becomes backward parabolic (although it can degenerate). Indeed, it is well known, for instance, that the nonlinear PDE that constitutes the essence of the ROF model (the total variation flow) is monotone (preserves the order of solutions) without the need for any restrictions such as the ones considered in this paper.

3. Comparison principle. We begin by introducing some notation. First, from now on the subscript κ in R_κ will be suppressed, but it will be understood that R comes with a parameter κ. Notice that for $y \notin \{-R(z(\kappa)), 0, R(z(\kappa))\}$, it holds that $R(\xi) = y$ for exactly two distinct values of ξ: one in $[-z(\kappa), z(\kappa)]$, and the other in $[-z(\kappa), z(\kappa)]^c$. We will denote by R_1 and R_2 the restriction of R to the domains

$[-z(\kappa), z(\kappa)]$ and $(-z(\kappa), z(\kappa))^c$, respectively:

$$R_1(\xi) := R(\xi)|_{[-z(\kappa),z(\kappa)]} \text{ and } R_2(\xi) := R(\xi)|_{(-z(\kappa),z(\kappa))^c}.$$

Then, R_1 and R_2 are one-to-one functions on their respective domains (the restriction is on the variable ξ); their inverses will be denoted R_1^{-1} and R_2^{-1}, respectively.

We will speak of the "jump set" $S(\phi)$ of a function $\{\phi_{i,j}\}$ defined on the grid; with that we mean the collection of indices defined by

$$S(\phi) := \left\{(i,j) : \max\left(\left|D_1^+\phi_{i,j}\right|, \left|D_2^+\phi_{i,j}\right|\right) \geq z(\kappa)\right\}.$$

Also, we adopt the terminology in [4] to say that ϕ is *supersonic* on $S(\phi)$ and *subsonic* elsewhere.

PROPOSITION 1. *Let $\{u_{i,j}(t)\}$ and $\{v_{i,j}(t)\}$ be solutions generated by the Perona–Malik scheme* (1)*, subject to Neumann boundary conditions. Assume that*

1. $|D_m^+ v_{i,j}(t)| < z(\kappa)$ *for all* (i,j), $t \in [0,T]$, *and* $m = 1, 2$, *and*
2. $|D_m^+ u_{i,j}(t)| \leq R_2^{-1}\left(R(|D_m^+ v_{i,j}(t)|)\right)$ *for all* (i,j), $t \in [0,T]$, *and* $m = 1, 2$.

Then if $\{u_{i,j}(t)\}$ and $\{v_{i,j}(t)\}$ are strictly ordered at $t = 0$, they remain ordered for all $t \in [0,T]$; i.e.,

1. *if* $u_{i,j}(0) > v_{i,j}(0)$, *then* $u_{i,j}(t) \geq v_{i,j}(t)$ *for all* $t \in [0,T]$, *and*
2. *if* $u_{i,j}(0) < v_{i,j}(0)$, *then* $u_{i,j}(t) \leq v_{i,j}(t)$ *for all* $t \in [0,T]$.

Proof. We treat only the first case $u_{i,j}(0) > v_{i,j}(0)$, the second case being completely analogous. Suppose the conclusion is false. Then there exists $t_0 \in (0,T]$ such that

$$\begin{aligned} &u_{i,j}(t) > v_{i,j}(t) \text{ for all } (i,j) \text{ and } t \in [0,t_0), \text{ and} \\ &u_{k,l}(t_0) = v_{k,l}(t_0) \text{ for some } (k,l). \end{aligned}$$

Consequently, $\dot{v}_{k,l}(t_0) - \dot{u}_{k,l}(t_0) \geq 0$, and hence

$$\begin{aligned} &D_1^- R\left(D_1^+ v_{k,l}(t_0)\right) - D_1^- R\left(D_1^+ u_{k,l}(t_0)\right) \\ &\quad + D_2^- R\left(D_2^+ v_{k,l}(t_0)\right) - D_2^- R\left(D_2^+ u_{k,l}(t_0)\right) \geq 0. \end{aligned}$$

Therefore,

$$\text{either } D_1^- R\left(D_1^+ v_{k,l}(t_0)\right) - D_1^- R\left(D_1^+ u_{k,l}(t_0)\right) \geq 0 \tag{4a}$$

$$\text{or } D_2^- R\left(D_2^+ v_{k,l}(t_0)\right) - D_2^- R\left(D_2^+ u_{k,l}(t_0)\right) \geq 0. \tag{4b}$$

Without loss of generality, assume that (4a) is true. That means that

$$R\left(D_1^+ v_{k,l}(t_0)\right) - R\left(D_1^+ u_{k,l}(t_0)\right) + R\left(D_1^+ u_{k-1,l}(t_0)\right) - R\left(D_1^+ v_{k-1,l}(t_0)\right) \geq 0. \tag{5}$$

But now, by hypotheses 1 and 2 of the proposition,

$$\left(R\left(D_1^+ v_{k,l}(t_0)\right) - R\left(D_1^+ u_{k,l}(t_0)\right)\right)\left(D_1^+ v_{k,l}(t_0) - D_1^+ u_{k,l}(t_0)\right) \geq 0, \tag{6a}$$

and

$$\begin{aligned} &\left(R\left(D_1^+ u_{k-1,l}(t_0)\right) - R\left(D_1^+ v_{k-1,l}(t_0)\right)\right) \\ &\quad \times \left(D_1^+ u_{k-1,l}(t_0) - D_1^+ v_{k-1,l}(t_0)\right) \geq 0. \end{aligned} \tag{6b}$$

Inequality (6a) can be verified by considering the three cases $R(D_1^+ u_{k,l}(t_0)) \in [-R_2^{-1}(R(|D_1^+ v_{k,l}(t_0)|)), -z(\kappa))$, and $R(D_1^+ u_{k,l}(t_0)) \in [-z(\kappa), z(\kappa)]$, and $R(D_1^+ u_{k,l}(t_0))$

$\in (z(\kappa), R_2^{-1}(R(|D_1^+ v_{k,l}(t_0))]$. Inequality (6b) can be verified by considering the analogous cases. By definition of t_0, we also have

$$\begin{aligned} D_1^+ v_{k,l}(t_0) - D_1^+ u_{k,l}(t_0) = D_1^+ \left(v_{k,l} - u_{k,l}\right)(t_0) \le 0, \text{ and} \\ D_1^+ u_{k-1,l}(t_0) - D_1^+ v_{k-1,l}(t_0) = D_1^+ \left(u_{k-1,l} - v_{k-1,l}\right)(t_0) \le 0. \end{aligned} \tag{7}$$

So we get, in particular,

$$\left(R\left(D_1^+ v_{k,l}(t_0)\right) - R\left(D_1^+ u_{k,l}(t_0)\right)\right) \times \left(R\left(D_1^+ u_{k-1,l}(t_0)\right) - R\left(D_1^+ v_{k-1,l}(t_0)\right)\right) \ge 0.$$

By (5),

$$\begin{aligned} R\left(D_1^+ v_{k,l}(t_0)\right) - R\left(D_1^+ u_{k,l}(t_0)\right) \ge 0, \\ R\left(D_1^+ u_{k-1,l}(t_0)\right) - R\left(D_1^+ v_{k-1,l}(t_0)\right) \ge 0. \end{aligned}$$

By (6a) and (6b) that means that

$$\begin{aligned} D_1^+ v_{k,l}(t_0) - D_1^+ u_{k,l}(t_0) \ge 0, \\ D_1^+ u_{k-1,l}(t_0) - D_1^+ v_{k-1,l}(t_0) \ge 0. \end{aligned} \tag{8}$$

Finally, (7) and (8) imply that

$$D_1^+ v_{j,l}(t_0) = D_1^+ u_{j,l}(t_0) \text{ for } j = k-1, k,$$

and thus

$$v_{j,l}(t_0) = u_{j,l}(t_0) \text{ for } j = k-1, k, k+1.$$

As a result, equality holds in (4a). Therefore, (4b) is also true. The same line of reasoning we used for (4a) now gives

$$v_{k,j}(t_0) = u_{k,j}(t_0) \text{ for } j = l-1, l, l+1.$$

Repetition of this argument (with k replaced by $k \pm 1$, l replaced by $l \pm 1$, and so on) gives

$$v_{i,j}(t_0) = u_{i,j}(t_0) \text{ for all } (i,j).$$

But then uniqueness of the solution to system (1) implies that

$$v_{i,j}(t) = u_{i,j}(t) \text{ for all } (i,j) \text{ and } t \ge t_0,$$

which proves the proposition. □

Remark. Proposition 1 has been stated and proved only on a two-dimensional grid for simplicity; the statement is true, and the proof works, with small modifications, for any space dimension.

Remark. It is not hard to check that the analogue of Proposition 1 holds for the discrete-in-time version of scheme (1), when forward Euler time steps are used with time step size satisfying $\delta t \le \frac{\min\{\delta x, \delta y\}^2}{4 \max\{1, \|R'\|_{L^\infty}\}}$, with δx and δy the uniform grid sizes in the two coordinate directions. Other statements in this paper can also be generalized to the fully discrete version of scheme (1).

Proposition 1 allows for comparison when one of the solutions is "smooth" (i.e., subsonic). In the case of one space dimension, we shall say a bit more: the next proposition allows for the comparison of more general one-dimensional signals. Its proof is a slight variation on that of Proposition 1. Its hypotheses will be justified in the next section, especially through Proposition 4. And eventually, it will find an application in section 4.3, where we will consider the behavior of scheme (1) as $h \to 0^+$.

PROPOSITION 2. *Let $\{u_j(t)\}$ and $\{v_j(t)\}$ be one-dimensional solutions generated by the Perona–Malik scheme* (1)*, subject to Neumann boundary conditions. Assume that*

1. $S(v(t)) = S(v(0)) := \{p_1, \dots, p_n\} \subseteq S(u(t))$ *for all* $t \in [0,T]$,
2. $|D^+u_j(t)| \le R_2^{-1}(R(|D^+v_j(t)|)$ *for all* $j \notin S(v(0))$ *and* $t \in [0,T]$, *and*
3. $sign(u_i(0) - v_i(0)) = sign(u_j(0) - v_j(0))(-1)^{k-k'} \neq 0$ *for* $i \in \{p_k + 1, \dots, p_{k+1}\}$ *and* $j \in \{p_{k'} + 1, \dots, p_{k'+1}\}$.

Then, for all $t \in [0,T]$ *we have*

$$\big(u_j(t) - v_j(t)\big)\big(u_j(0) - v_j(0)\big) \ge 0.$$

Proof. The conclusion is satisfied for some positive time by continuity; suppose that it fails for the first time at $t = t_0 < T$ and at index k. Without loss of generality, let us assume that $u_k(0) > v_k(0)$. Define $\alpha : \mathbf{Z} \to \{0,1\}$ as follows:

$$\alpha(j) := \begin{cases} 1 \text{ if } j \in S(v(0)), \\ 0 \text{ otherwise.} \end{cases}$$

By definitions of t_0 and k, and by hypothesis 3, we have

$$\begin{aligned} \big(D^+u_k(t_0) - D^+v_k(t_0)\big)(-1)^{\alpha(k)} &\ge 0, \\ \big(D^+u_{k-1}(t_0) - D^+v_{k-1}(t_0)\big)(-1)^{\alpha(k-1)} &\le 0. \end{aligned} \tag{9}$$

Hypothesis 2 implies, as in the proof of Proposition 1, that

$$\big(R(D^+u_j(t)) - R(D^+v_j(t))\big)\big(D^+u_j(t) - D^+v_j(t)\big) \ge 0 \text{ if } j \notin S(v(0)). \tag{10}$$

On the other hand, if $j \in S(v(0))$, then $|D^+u_j|, |D^+v_j| \ge z(\kappa)$; and since R is decreasing on $(-z(\kappa), z(\kappa))^c$ we get

$$\big(R(D^+u_j(t)) - R(D^+v_j(t))\big)\big(D^+u_j(t) - D^+v_j(t)\big) \le 0 \text{ if } j \in S(v(0)). \tag{11}$$

We can summarize (10) and (11) as

$$\big(R(D^+u_j(t)) - R(D^+v_j(t))\big)\big(D^+u_j(t) - D^+v_j(t)\big)(-1)^{\alpha(j)} \ge 0. \tag{12}$$

Furthermore, the definition of t_0 implies that

$$\begin{aligned} \dot{u}_k(t_0) - \dot{v}_k(t_0) &= R(D^+u_k(t_0)) - R(D^+v_k(t_0)) \\ &\quad - R(D^+u_{k-1}(t_0)) + R(D^+v_{k-1}(t_0)) \\ &\le 0. \end{aligned} \tag{13}$$

But now, (9) and (12) mean that

$$\big(R(D^+u_k(t_0)) - R(D^+v_k(t_0))\big) \times \big(R(D^+v_{k-1}(t_0)) - R(D^+u_{k-1}(t_0))\big) \ge 0. \tag{14}$$

Combined with (13), (14) implies that

$$\begin{aligned} R(D^+u_k(t_0)) - R(D^+v_k(t_0)) &\le 0, \\ R(D^+v_{k-1}(t_0)) - R(D^+u_{k-1}(t_0)) &\le 0. \end{aligned} \tag{15}$$

In light of (12), these inequalities lead to the following conclusion:

$$\begin{aligned} \left(D^+u_k(t_0) - D^+v_k(t_0)\right)(-1)^{\alpha(k)+1} &\ge 0, \\ \left(D^+u_{k-1}(t_0) - D^+v_{k-1}(t_0)\right)(-1)^{\alpha(k-1)} &\ge 0. \end{aligned} \tag{16}$$

But then, (16) and (9) give

$$D^+\left(u_j(t_0) - v_j(t_0)\right) = 0 \text{ for } j = k-1, k$$

so that

$$u_j(t_0) = v_j(t_0) \text{ for } j = k-1, k, k+1.$$

That, much like in the proof of Proposition 1, leads to the conclusion of the proposition. □

4. Applications. The comparison principles stated and proved in the previous section are simply tools; indeed, their hypotheses require knowledge of the solutions involved for all time. Here, in section 4.1, we use them to state some down-to-earth results, such as the stability property that is the content of Theorem 5. Then, in section 4.2, we give concrete examples of how those results can be applied in practice. Section 4.3 is devoted to exploring what these results say in the limit as $h \to 0^+$.

4.1. Stability results. We begin by recording a few simple but important properties of scheme (1) that will help us apply Propositions 1 and 2.

LEMMA 3. *Let $\{u_{i,j}(t)\}$ be the solution generated by scheme* (1) *from subsonic initial data. Then*

$$\sup_{i,j,t} |D_m^+ u_{i,j}(t)| \le \sup_{i,j} |D_m^+ u_{i,j}(0)| \text{ for each } m = 1, 2.$$

Proof. It is easy to see that in the subsonic regime, scheme (1) satisfies a maximum principle for difference quotients. This in turn prevents the solution from entering the supersonic regime if the initial data is subsonic. The conclusion of the lemma follows. □

In what follows, we will often specialize to the one-dimensional version of the Perona–Malik scheme (1), which then reduces to

$$\dot{u}_j(t) = D^-\left(R_\kappa\left(D^+u_j\right)\right). \tag{17}$$

Next, we recall an important property of scheme (17): supersonic regions shrink in time.

PROPOSITION 4. *Let $\{u_j(t)\}$ be a solution generated by scheme* (17). *Then $S(u(t_2)) \subseteq S(u(t_1))$ whenever $0 \le t_1 \le t_2$.*

Proof. See [3], where it first appeared, or [2]. □

Remark. The conclusion of Proposition 4 is false in two dimensions: supersonic regions can grow, as shown by the numerical experiment in Figure 3.

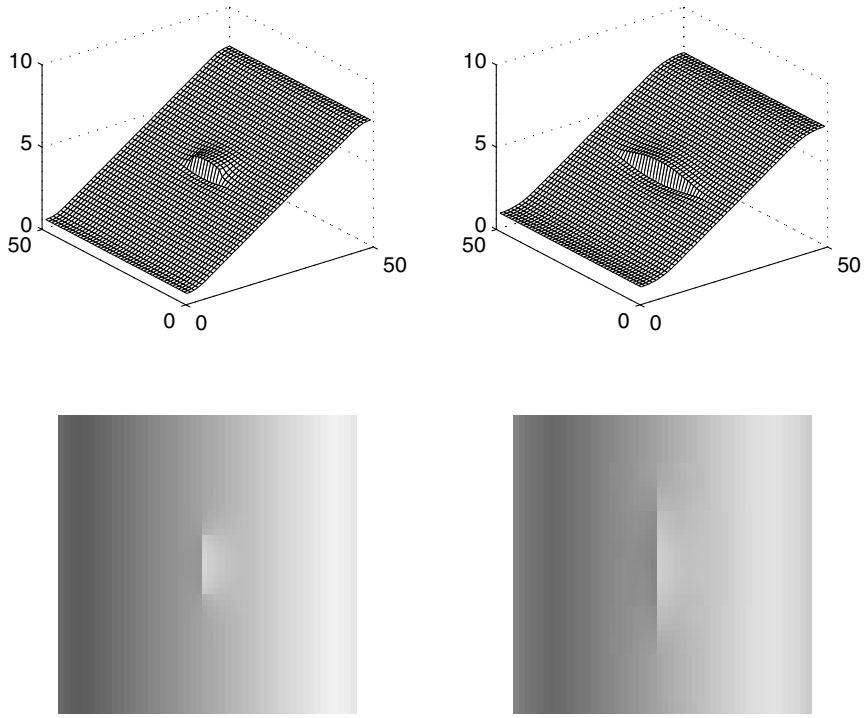


FIG. 3. *Jump set (or the supersonic regime) can grow in two dimensions. Here, a small "crack" in the initial data propagates.*

Our first application deals with subsonic data corrupted by low amplitude noise. We estimate the difference between the evolutions of corrupted and uncorrupted data in terms of the amplitude of the noise. The hypotheses of Proposition 1 involve all $t \geq 0$. We will use in our proof comparison functions that will bound, one from above and the other from below, the solution generated from the corrupted data. As we shall explain, these comparison functions will satisfy the hypotheses of Proposition 1 automatically for all time. Moreover, their definitions will involve only the uncorrupted initial data.

THEOREM 5. *Let $\{\phi_{i,j}\}$ be subsonic initial data, i.e.,*

$$M := \max_{i,j,m} |D_m^+ \phi_{i,j}| < z(\kappa).$$

Let $\{u_{i,j}(t)\}$ be the solution generated by scheme (1) *from $\{\phi_{i,j}\}$, and let $\{u_{i,j}^n(t)\}$ be the one generated from $\{(\phi + n)_{i,j}\}$. If*

$$\max_{i,j} |n_{i,j}| < \frac{h}{2} \left(R_2^{-1} \left(R(M) \right) - M \right),$$

then

$$\max_{i,j} \left| u_{i,j}^n(t) - u_{i,j}(t) \right| \leq \max_{i,j} |n_{i,j}| \ \textit{for all } t \geq 0.$$

Proof. Fix a $\delta > 0$ such that

$$\max_{i,j} |n_{i,j}| < \delta < \frac{h}{2}\left(R_2^{-1}\left(R(M)\right) - M\right).$$

The upper and lower comparison functions, which we shall denote by $v^-_{i,j}(t)$ and $v^+_{i,j}(t)$, respectively, will simply be

$$v^{\pm}_{i,j}(t) := u_{i,j}(t) \pm \delta. \tag{18}$$

Then $v^{\pm}_{i,j}(t)$ are clearly solutions of (1). Furthermore, $u^n_{i,j}(0) \in (v^-_{i,j}(0), v^+_{i,j}(0))$. Since the initial condition $\phi_{i,j}$ is subsonic, by Lemma 3, $u_{i,j}(t)$ and therefore also $v^{\pm}_{i,j}(t)$ are subsonic for all time. Thus, hypothesis 1 in Proposition 1 is satisfied. Moreover, again by virtue of Lemma 3, we have

$$\max_{i,j,m,t} |D^+_m u_{i,j}(t)| = \max_{i,j,m,t} |D^+_m v^{\pm}_{i,j}(t)| = M. \tag{19}$$

Also, the inequality in hypothesis 2 of Proposition 1 is *strictly* satisfied at $t = 0$ since

$$|D^+_m u^n_{i,j}(0)| \le |D^+_m u_{i,j}(0)| + \frac{2}{h}\max_{i,j}|n_{i,j}| < R_2^{-1}(R(M)) \le R_2^{-1}\left(R(|D^+_m u_{i,j}(0)|)\right)$$

by our assumption on the amplitude of the noise $n_{i,j}$. We will now show that in fact hypothesis 2 is strictly satisfied for all time. Suppose not; then there exists $t_0 > 0$ such that

$$\left|D^+_m u^n_{i,j}(t)\right| < R_2^{-1}\left(R\left(|D^+_m u_{i,j}(t)|\right)\right)$$

for all (i,j), $m \in \{1,2\}$, and $t \in [0, t_0)$, and

$$\left|D^+_m u^n_{k,l}(t_0)\right| = R_2^{-1}\left(R\left(|D^+_m u_{k,l}(t_0)|\right)\right)$$

for some (k,l) and some $m \in \{1,2\}$. By (19), that means that

$$\left|D^+_m u^n_{k,l}(t_0)\right| \ge R_2^{-1}\left(R(M)\right). \tag{20}$$

We also have

$$\begin{aligned} \left|D^+_m u^n_{k,l}(t_0)\right| &\le \left|D^+_m (u^n - u)_{k,l}(t_0)\right| + \left|D^+_m u_{k,l}(t_0)\right| \\ &\le \left|D^+_m (u^n - u)_{k,l}(t_0)\right| + M \end{aligned} \tag{21}$$

again by (19). Combining (20) and (21) we get

$$\left|D^+_m (u^n - u)_{k,l}(t_0)\right| \ge R_2^{-1}\left(R(M)\right) - M > \frac{2\delta}{h}.$$

That means we have

$$\left|u^n_{i,j}(t_0) - u_{i,j}(t_0)\right| > \delta \text{ for some } (i,j) \in \{k, k+1\} \times \{l, l+1\}. \tag{22}$$

On the other hand, since both hypotheses of Proposition 1 are satisfied on $t \in [0, t_0)$, we have

$$v^-_{i,j}(t) \le u^n_{i,j}(t) \le v^+_{i,j}(t) \text{ for all } t \in [0, t_0)$$

and, by continuity, also at $t = t_0$. Combined with (18) this means

$$\left|u^n_{i,j}(t_0) - u_{i,j}(t_0)\right| \le \delta \text{ for all } (i,j),$$

which contradicts (22). ∎

An immediate consequence of Theorem 5 is the following elementary corollary, which is, unlike the theorem, one-dimensional. It tells us that a smooth one-dimensional signal corrupted by low amplitude noise is rapidly denoised and provides an upper bound on the denoising time.

COROLLARY 6. *Let ϕ_j, n_j, u_j, u^n_j, and M be as in Theorem 5, and assume that n_j satisfies the hypothesis of that theorem. If we set*

$$T := \inf\{t_0 \ge 0 : S(u^n) \text{ is empty for all } t \ge t_0\},$$

then we have the estimate

$$T \le \frac{2\max_j |\phi_j + n_j|}{R\left(\frac{2}{h}\max_j |n_j| + M\right)}.$$

Proof. The interesting case is when $S(u^n(0))$ is nonempty; under that assumption, for any $\delta > \max_j |n_j|$ we have $2\delta/h + M > z(\kappa)$. Fix an $\varepsilon > 0$ small enough so that

$$R\left(z(\kappa) - \varepsilon\right) > R\left(\frac{2\delta}{h} + M\right).$$

For $k \in S(u^n(0))$, let

$$T^\varepsilon_k := \inf\left\{t \ge 0 : |D^+ u^n_k(t)| = z(\kappa) - \varepsilon\right\}.$$

In view of Proposition 4, we have

$$T < \max_k T^\varepsilon_k.$$

So fix a $k \in S(u^n(0))$; without loss of generality, we may assume that $D^+u^n_k(0) > 0$. By Theorem 5,

$$D^+ u^n_k(t) < \frac{2\delta}{h} + M \text{ for all } t \ge 0.$$

Furthermore, since R is a decreasing function on $[z(\kappa), \infty)$, for $t \in [0, T^\varepsilon_k]$ we have

$$z(\kappa) - \varepsilon \le D^+ u^n_k(t) < \frac{2\delta}{h} + M \Rightarrow R(D^+u^n_k(t)) > R\left(\frac{2\delta}{h} + M\right). \tag{23}$$

That gives

$$\frac{d}{dt}\sum_{j=1}^{k} h\left(\max_j |\phi_j + n_j| - u^n_j\right) = -R(D^+u^n_k) \le -R\left(\frac{2\delta}{h} + M\right),$$

where the inequality follows via (23) for as long as $t \in [0, T^\varepsilon_k]$. Also, since $|u^n_k(t)| \le \max_j |\phi_j + n_j|$ for all $t \ge 0$, we have

$$2\left(\max_j |\phi_j + n_j|\right) - T^\varepsilon_k R\left(\frac{2\delta}{h} + m\right) \ge 0.$$

Letting $\delta \to \max_j |n_j|$ from above leads to the desired inequality. ∎

4.2. Examples. In this section, we apply the results of the previous subsection in some practical situations.

Example 1. Take the function in Figure 4. It has maximum slope $M = 4$. The Perona–Malik scheme (17) is applied using the choice

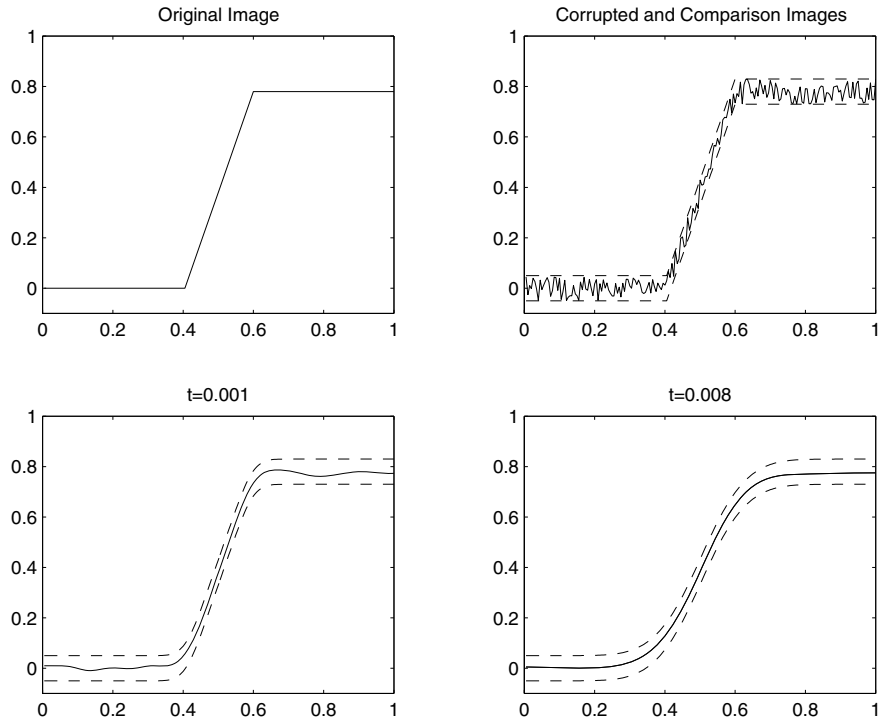


FIG. 4. *An application of the comparison principle. The corrupted initial data (which is not subsonic) quickly becomes subsonic, and no artificial "edges" are introduced. The noise amplitude here is* 0.05*, and the maximum amplitude allowed by Theorem* 5 *is* 0.0525*.*

$$R(\xi) = \frac{\xi}{1 + \frac{\xi^2}{100}} \tag{24}$$

for the nonlinear function appearing in the scheme. According to this choice, the threshold value of slope is $z(\kappa) = 10$. We calculate the maximum noise amplitude allowed by Theorem 5 to be 0.0525. The corrupted signal in the example of Figure 4, which is not subsonic, was obtained by adding noise of amplitude 0.05 to the original signal.

The evolution shown in Figure 5 is obtained by adding a specific perturbation of amplitude 0.06 to the original signal. We see how the comparison principle gets violated.

Example 2. We now compare the effects of the precise shape of the function $R(\xi)$ on the behavior of the scheme, by using the same original image as in our first example, but the different nonlinear function

$$R(\xi) = \xi \exp\left(-\frac{\xi^2}{200}\right), \tag{25}$$

which has the same threshold value of the slope as for (24) of the first example, namely, $z(\kappa) = 10$. The maximum amplitude of noise allowed by Theorem 5 this time (for $R(\xi)$ given by (25)) turns out to be between 0.034 and 0.03425—significantly smaller than that for (24).

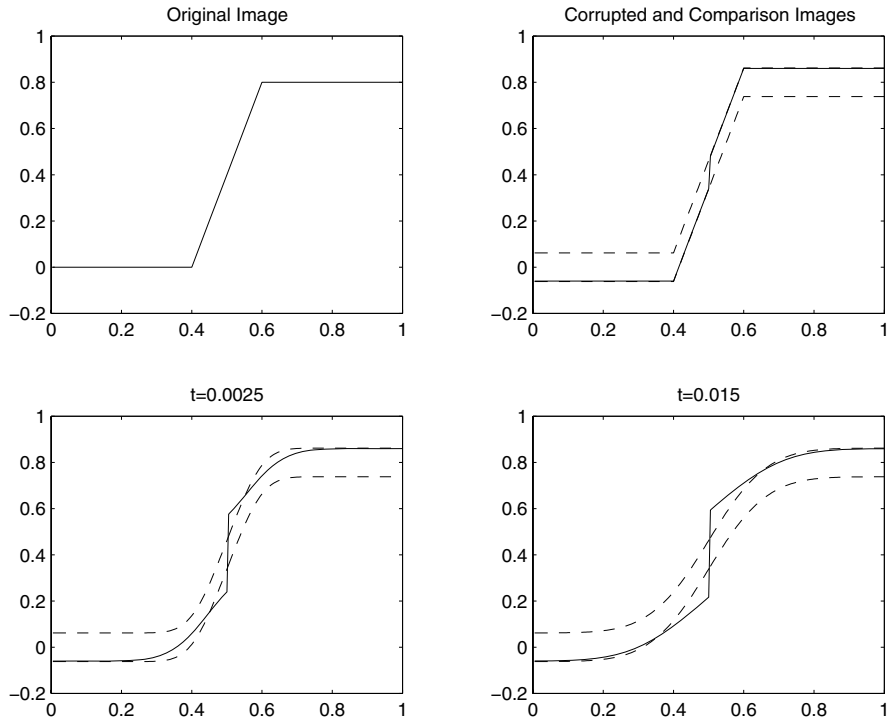


FIG. 5. *Effects of a perturbation with a (deliberately chosen) noise of amplitude that exceeds the maximum amplitude allowed. Here, the noise amplitude is* 0.06*, higher than what Theorem* 5 *allows, which is* 0.0525*. The result is an artificial "edge."*

Theorem 5 thus suggests a way to quantify the difference in stability properties of scheme (1) with respect to the two choices of $R_\kappa(\xi)$. It is easy to see that this difference is important; we illustrate it with a numerical example: Figure 6 shows the evolution of the original image perturbed by a (contrived) noise of amplitude 0.05 under scheme (17) using the two choices for $R(\xi)$ given in (24) and (25). We emphasize again that although the two choices of R are different, the value of κ in each case has been chosen so that the critical points of the two functions (that determine boundary of sub- and supersonic regions) are the same. However, the perturbation amplitude 0.05 is above the allowed limit for (25), and below it for (24). The evolutions can then be seen to be very different. The perturbation used in this and the previous example simply consisted of shifting the graph of the initial condition along $x \geq 0$ by a small amount, so as to introduce a "tear" in the middle of the graph where the gradient of the unperturbed initial condition is largest.

4.3. Limit as $h \to 0^+$. In [2], the continuum limit of scheme (17) is investigated with the function $R_\kappa(\xi)$ given by

$$R_\kappa(\xi) = \xi \left(1 + \frac{\xi^2}{\kappa}\right)^{(\beta-1)}, \text{ where } \beta \in \left[0, \frac{1}{2}\right). \tag{26}$$

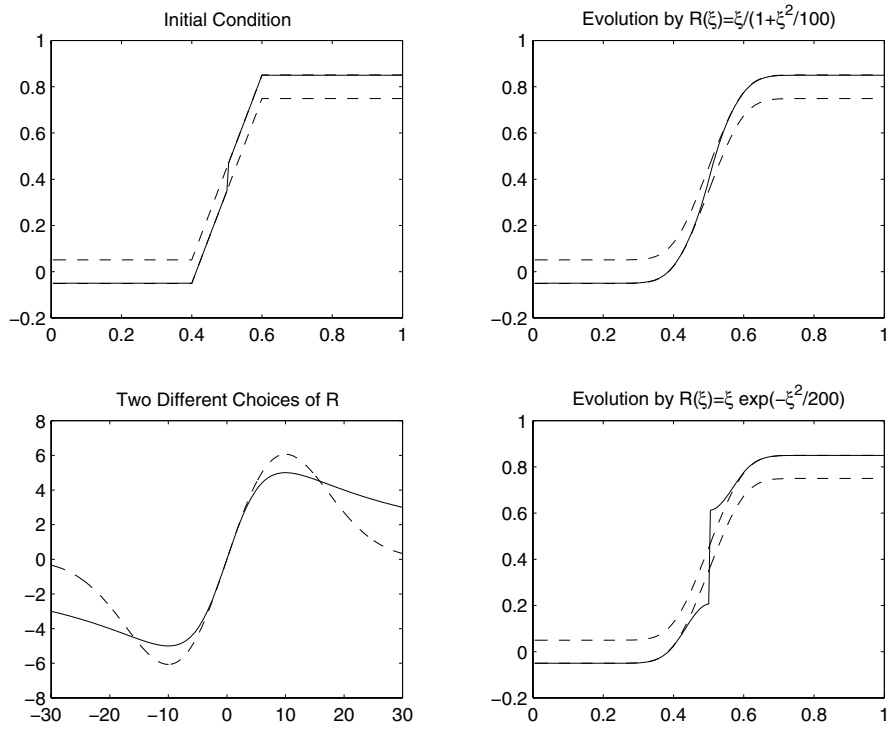


FIG. 6. *Evolution via different choices for the function R_κ. The upper left-hand figure illustrates the initial signal and comparison functions; the initial signal was obtained from that of Example* 1 (*Figure* 4) *after modifying by a deliberately chosen perturbation. The lower left-hand figure shows the two distinct choices for the function $R(\xi)$ used in this example: The dashed line is the graph of $R(\xi) = \xi \exp(-\xi^2/200)$, and the solid line is the graph of $R(\xi) = \xi/(1 + \xi^2/100)$. The second column shows the evolution of the initial image and comparison functions with $R(\xi) = \xi/(1+\xi^2/100)$ and $R(\xi) = \xi \exp(-\xi^2/200)$ (comparison functions are shown with dashed lines). For both of these functions, $z(\kappa) = 10$.*

When $\beta \in (0, \frac{1}{2})$ the evolution that (17) generates is the gradient descent for the discrete energy

$$\mathbf{E}_u^h(t) := \sum_j h\Phi_{\kappa,\beta}\left((D^+u_j(t))^2\right), \tag{27}$$

where

$$\Phi_{\kappa,\beta}(\xi) := \frac{\kappa}{\beta}\left(\left(1 + \frac{\xi}{\kappa}\right)^\beta - 1\right). \tag{28}$$

The continuum limit studied in [2] is obtained by scaling the parameter κ (and hence the threshold point $z(\kappa)$) with respect to the discretization step size h as follows:

$$\kappa(h) = h^{(2\beta-1)/(1-\beta)} \Rightarrow z(h) := \frac{1}{\sqrt{1-2\beta}} h^{(2\beta-1)/(2-2\beta)}. \tag{29}$$

Such scalings were studied previously in the stationary setting by Chambolle in [1] to obtain interesting continuum limits for discrete energies similar to (27); the approach

taken in [2] follows Chambolle's lead in adjusting the threshold z with respect to the grid size h but concerns the time dependent problem. The resulting evolution is defined for piecewise smooth one-dimensional signals; it takes place on a domain that changes at discrete times and is described as follows.

For a given piecewise smooth initial data $\phi : [0,1] \to [0,1]$ with jump discontinuities at $p_1, \dots, p_N \in (0,1)$, we solve the linear heat equation $u_t = u_{xx}$ on the domain $(0,1) - \{p_1, \dots, p_N\}$, subject to homogeneous Neumann boundary conditions at $x = 0$ and $x = 1$, and to the nonlinear boundary conditions

$$u_x(p_j^{\pm}, t) = \big(u(p_j^+, t) - u(p_j^-, t)\big)\,\big|u(p_j^+, t) - u(p_j^-, t)\big|^{2\beta-2} \tag{30}$$

at the discontinuity points. The condition (30) *becomes singular whenever one of the discontinuities, say, the one at p_j, heals (i.e., when $u(p_j^+, t) = u(p_j^-, t)$); at such special times, we merge the two intervals (p_{j-1}, p_j) and (p_j, p_{j+1}) into one longer interval (p_{j-1}, p_{j+1}) and continue the evolution according to the heat equation on the new (and smaller) collection of intervals.*

The claim is that the numerical solutions generated by scheme (17) converge, as $h \to 0$, to the continuum evolution described above, provided that the threshold $z(\kappa)$ is scaled with respect to h according to formula (29), and that the approximate (discrete) initial data ϕ^h converge to the continuum data ϕ in some suitable sense. The proof in [2] for this statement involves a number of technical hypotheses. One of the most restrictive among them requires the "jump sets" of $\phi^h$and ϕ to be *compatible*: it is assumed that $S(\phi^h)$ and $\{p_1, \dots, p_N\}$ are in one-to-one correspondence.

A discussion of the most general conditions under which convergence takes place would be very technical and out of place. But, as an application of Proposition 2, we will show that under suitable hypotheses, the jump sets $S(\phi^h)$ of the approximate initial data become compatible with $\{p_1, \dots, p_N\}$ after an arbitrarily small initial interval of time. To that end, let $\{h_n\}_{n=1}^{\infty}$ be a sequence of positive numbers such that $h_n \to 0$ as $n \to \infty$, and let $x_j^{h_n}$ denote the grid points for the uniform discretization size h_n. Assume that a sequence ϕ^{h_n} of discrete initial data satisfies

$$\lim_{n\to\infty} \max_j \left|\phi_j^{h_n} - \phi(x_j^{h_n})\right| = 0.$$

Then we have the following result.

THEOREM 7. *Let $\{u^{h_n}(t)\}_n$ be the discrete solutions generated from $\{\phi^{h_n}\}_n$ by scheme* (17)*, where R is given by* (26) *and κ is scaled as in* (29)*. Then, given any $\varepsilon > 0$, there exists $K \in \mathbf{N}$ such that for any $n > K$ the following property holds at some $t \in [0, \varepsilon)$:*

$$|D^+ u_j^{h_n}(t)| \geq z(\kappa) \text{ only if } \{p_1, \dots, p_N\} \cap [x_j^{h_n}, x_{j+1}^{h_n}] \text{ is nonempty.}$$

Proof. The jump sets of initial data ϕ^{h_n} can be much larger than that of ϕ; the idea is to construct comparison functions $\psi^{n,\pm}$ whose jump sets precisely match $\{p_1, \dots, p_N\}$, and then compare using Proposition 2. We first define

$$M := \max_i \max_{x \in (p_i, p_{i+1})} |\phi'(x)|, \qquad m := \min_i \left|\phi(p_i^+) - \phi(p_i^-)\right|,$$

$$\tilde{S}_n := \left\{ j \in \mathbf{N} : \left| D^+ \phi_j^{h_n} \right| \geq \frac{m}{2h_n} \right\}, \qquad \alpha(j) := \sharp \left\{ i \in \tilde{S}_n : i < j \right\}.$$

Convergence of the approximate initial data ϕ^{h_n} to ϕ uniformly on the grid implies

that for large enough n we have

$$\begin{aligned} &j \in \tilde{S}_n \Rightarrow p_i \in [x_j, x_{j+1}] \text{ for some } i, \text{ and} \\ &p_i \in [x_j, x_{j+1}] \Rightarrow \{j, j+1\} \cap \tilde{S}_n \neq \emptyset. \end{aligned} \tag{31}$$

For $\delta \in (0, \frac{m}{4})$, we construct the pair of comparison functions $\psi^{n,\pm}(t)$ via scheme (17) from the following initial data:

$$\psi_j^{n,\pm}(0) = \phi(x_j^{h_n}) \pm \delta(-1)^{\alpha(j)}.$$

Then, $\psi^{n,\pm}(0)$ satisfy the following properties:

$$S\left(\psi^{n,\pm}(0)\right) = \tilde{S}\left(\phi^{h_n}\right), \tag{32a}$$

$$\sharp S(\psi^{n,\pm}(0)) = N, \text{ and} \tag{32b}$$

$$\sup_n \mathbf{E}^{h_n}_{\psi^{n,\pm}}(0) < \infty. \tag{32c}$$

Furthermore,

$$\limsup_{n\to\infty} \max_{j \notin S(\psi^{n,\pm})} \left|D^+\psi^{n,\pm}\right| = M < \infty. \tag{32d}$$

We recall a few points from [2]: First, by virtue of property (32c), the evolutions $\{\psi^{n,\pm}(t)\}_n$ are Hölder continuous in time with values in L^∞ of space, *uniformly* in n. Moreover, the difference quotients of $\psi^{n,\pm}(t)$ satisfy the maximum principle on the complements of their jump sets, while the jump sets remain constant. In light of these comments and of Proposition 4, we can determine a $T > 0$ so that for all $t \in [0, T]$ the following hold:

$$S(\psi^{n,\pm}(t)) = S(\psi^{n,\pm}(0)), \tag{33a}$$

$$\max_{j \notin S(\psi^{n,\pm})} \left|D^+\psi^{n,\pm}(t)\right| \leq 2M. \tag{33b}$$

The dependence of κ on h_n, as prescribed in (29), implies that

$$C := \liminf_{n\to\infty} h_n R_2^{-1}\left(R(2M)\right) > 0 \tag{33c}$$

so that, by (33b) for large enough n,

$$R_2^{-1}\left(R(D^+\psi_j^{n,\pm}(t))\right) > \frac{C}{2h_n} \text{ for all } j \notin S\left(\psi_j^{n,\pm}(0)\right) \text{ and } t \in [0, T]. \tag{33d}$$

Choose $\delta < C/4$. For large enough n, we will certainly have

$$(-1)^{\alpha(j)}\psi_j^{n,-}(t) < u_j^{h_n}(t) < (-1)^{\alpha(j)}\psi_j^{n,+}(t) \tag{33e}$$

for some positive time. Then (32a), (33a), and (33e) verify hypotheses 1 and 3 of Proposition 2 for positive time. Meanwhile, (33d), (33e), and the choice of δ verify hypothesis 2 of Proposition 2 at $t = 0$. But as in the proof of Theorem 5, since (33d) holds for all $t \in [0, T]$, these are sufficient to ensure that the three hypotheses of Proposition 2 are satisfied for all $t \in [0, T]$ and lead to the conclusion that we want:

$$(-1)^{\alpha(j)}\psi_j^{n,-}(t) \leq u_j^{h_n}(t) \leq (-1)^{\alpha(j)}\psi_j^{n,+}(t) \text{ for all } t \in [0, T]. \tag{34}$$

At this stage, it is possible to repeat the proof of Corollary 6 to get an estimate of how quickly jumps of height less than 2δ vanish; the upper bound one obtains goes to 0 when we send first $n \to \infty$ and then $\delta \to 0$. □

Remark. In some sense, Theorem 7 says that given an initial piecewise smooth signal corrupted by noise (i.e., a signal with a few large and many small jumps), the Perona–Malik scheme denoises the signal by quickly removing the small jumps and maintaining the large ones. This stability property hinged on inequality (33c) in our proof, and is a result of the choice of the constitutive function (26) and scaling (29). It is in contrast to the different function and scaling considered in [3] that lead to a different continuum limit for which such a stability property is not to be expected.

Remark. As h gets smaller, the hypotheses of the discrete comparison principles discussed in this paper get more restrictive. In fact, if h is sent to 0 while κ is fixed, we arrive at the Perona–Malik PDE (2), to which our arguments do not extend. However, scaling (29) implies that $z(h) \to \infty$ as $h \to 0^+$ (albeit at a slower rate than $1/h$). This allows for some form of the restricted comparison principle to be maintained in the limit; it allowed us to state Theorem 2, which concerns a continuum evolution.

Acknowledgment. The author would like to thank his former advisor Robert V. Kohn for his continuing attention and encouragement.

REFERENCES

[1] A. Chambolle, *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations*, SIAM J. Appl. Math., 55 (1995), pp. 827–863.

[2] S. Esedoglu, *An analysis of the Perona–Malik scheme*, Comm. Pure Appl. Math., 54 (2001), pp. 1442–1487.

[3] M. Gobbino, *Gradient flow for the one-dimensional Mumford-Shah functional*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 27 (1998), pp. 145–193.

[4] B. Kawohl and N. Kutev, *Maximum and comparison principle for one-dimensional anisotropic diffusion*, Math. Ann., 311 (1998), pp. 107–123.

[5] S. Kichenassamy, *The Perona–Malik paradox*, SIAM J. Appl. Math., 57 (1997), pp. 1328–1342.

[6] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion*, in Proceedings of the IEEE Computer Society Workshop on Computer Vision, IEEE Computer Society, Los Alamitos, CA, 1987, pp. 16–22.

[7] P. Perona and J. Malik, *Scale Space and Edge Detection Using Anisotropic Diffusion*, Technical report, Department of EECS, University of California at Berkeley, Berkeley, CA, 1988.

[8] P. Perona, T. Shiota, and J. Malik, *Anisotropic diffusion*, in Geometry-Driven Diffusion in Computer Vision, B. ter Haar Romeny, ed., Comput. Imaging Vision 1, Kluwer Academic, Dordrecht, The Netherlands, 1994, pp. 73–92.

[9] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[10] T. P. Witelski, D. G. Schaeffer, and M. Shearer, *A discrete model for an ill-posed nonlinear parabolic PDE*, Phys. D, 160 (2001), pp. 189–221.

[11] Y. Yuo, W. Xu, A. Tannenbaum, and M. Kaveh, *Behavioral analysis of anisotropic diffusion in image processing*, IEEE Trans. Image Process., 5 (1996), pp. 1539–1553.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1314–1331

MATRIX-FREE INTERPOLATION ON THE SPHERE*

M. GANESH† AND H. N. MHASKAR‡

Abstract. We study a subspace of bivariate trigonometric polynomials for interpolating functions on the sphere. We give an explicit construction for a system of interpolation nodes, and the corresponding basis for this space, that allows a (discrete) fast Fourier transform-type formula for the interpolant. We prove that the uniform norm of our interpolation operator is of the order $(\log M)^2$, where M is the number of interpolation points. We also construct a minimal quadrature rule for our space (with a number of points equal to the dimension of the space), and describe an associated interpolation operator.

Key words. sphere, interpolation, discrete orthogonal projection, minimal quadrature

AMS subject classifications. 42A15, 65D32, 33C55

DOI. 10.1137/050624005

1. Introduction.

Approximation of differentiable functions on the sphere and analysis of the error in the uniform norm are important ingredients for solving partial differential equations on spherical geometries [4, 5, 6]. Advances in approximation theory on the sphere are required, for example, to measure earth's atmospheric flow and gravitational potential, to simulate sound waves scattered by spherical geometries, to identify the shape of the scattering objects, and for image reconstructions in cosmology.

One popular strategy for approximation of functions is interpolation. Construction of an interpolatory approximation to an unknown function f defined on a set Ω consists of (i) designing a set of nodes $x_j \in \Omega$ and a class of functions ϕ_j defined on Ω, $j = 1, \dots, M$, that forms a basis for a space V and (ii) computing an approximant $I_M f = \sum_{j=1}^{M} a_j(f)\phi_j \in V$ such that $I_M f(x_k) = f(x_k), k = 1, \dots, M$. The matrix $\mathbf{A} := [\phi_j(x_k)]_{j,k=1,\dots,M}$ is called the *interpolation matrix*. We say that $\{x_1, \dots, x_M\}$ is a *set of uniqueness* for V if $\mathbf{A}$ is invertible, or equivalently, for $v \in V$, the fact that $v(x_k) = 0$ for $k = 0, \dots, M$ implies that $v \equiv 0$. It is clear that the vector $\mathbf{a}$ of coefficients in $I_M f$ satisfies the matrix equation

$$\mathbf{A}\mathbf{a} = [f(x_1), \dots, f(x_M)]^T. \tag{1.1}$$

Thus, some of the important problems in this theory are an efficient evaluation of the coefficient vector $\mathbf{a}$ and an estimation of the *Lebesgue constant* (uniform norm) of the operator I_M, defined to be the infimum of all Λ_M for which

$$\sup_{x\in\Omega} |I_M f(x)| \le \Lambda_M \sup_{x\in\Omega} |f(x)|.$$

A classical example is the interpolation on the unit circle, $\{e^{ix} : x \in [0, 2\pi)\}$. We

*Received by the editors February 8, 2005; accepted for publication (in revised form) January 6, 2006; published electronically July 7, 2006.

http://www.siam.org/journals/sinum/44-3/62400.html

†Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (mganesh@mines.edu). The research of this author was supported in part by the Australian Research Council.

‡Department of Mathematics, California State University, Los Angeles, CA, 90032 (hmhaska@calstatela.edu). The research of this author was supported in part by grant DMS-0204704 from the National Science Foundation and grant W911NF-04-1-0339 from the U.S. Army Research Office.

define the nodes x_j and the basis ϕ_j by

$$x_j := \frac{2(j+N)\pi}{2N+1}, \quad \phi_j(x) := \frac{\exp(ijx)}{\sqrt{2N+1}}, \quad -N \le j \le N.$$

In this case (using a well-known identity for complex exponentials (see (4.1))), the inverse of the resulting interpolation matrix A_{trig} is $\overline{A_{trig}^T}$. Hence, for a function f defined on $[0, 2\pi]$, the solution of the interpolation problem can be written explicitly as follows:

$$I_{2N+1}f(x) = \sum_{j=-N}^{N} a_j(f)\phi_j(x) = \sum_{\ell=-N}^{N} f(x_\ell)K_N(x, x_\ell), \tag{1.2}$$

where

$$a_j(f) := \sum_{\ell=-N}^{N} f(x_\ell)\overline{\phi_j(x_\ell)}, \quad -N \le j \le N, \quad K_N(x,y) = \sum_{j=-N}^{N} \phi_j(x)\overline{\phi_j(y)}.$$

We note that each $a_j(f)$ is the discrete Fourier transform (DFT), and hence, for a fixed x, the $\mathcal{O}(N^2)$ summations in $I_{2N+1}f(x)$ can be computed with $\mathcal{O}(N \log N)$ operations, using the fast Fourier transform (FFT). The Lebesgue constant in this case is $\mathcal{O}(\log N)$.

Formulas similar to (1.2) also hold in the case of interpolation at the zeros of an arbitrary orthogonal polynomial system on a real interval [3, section I.4]. This fact is a simple consequence of the Gauss–Jacobi quadrature formula, which, although based on the zeros of a polynomial of degree n, is exact for integrating polynomials of degree $2n-1$. In Proposition 2.1 below, we will use the same arguments as in [3, section I.4] and [10, Lemma 3] to observe that the existence of suitable quadrature formulas implies the existence of interpolation formulas similar to (1.2). One common feature in all these examples is that one does not need to solve a matrix equation to solve the interpolation problem. We say that the interpolation is *matrix-free* when a system of interpolation nodes and basis functions is given explicitly, so that a formula analogous to the FFT-type representation (1.2) defines an interpolation operator.

If Ω is the unit sphere $\mathbb{S}^2 := \{(x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2 = 1\}$, one may think of a simple strategy for interpolation by considering the coordinate transformation

$$\hat{\mathbf{x}} = \boldsymbol{p}(\theta, \phi) := (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)^{\mathrm{T}}, \quad \hat{\mathbf{x}} \in \mathbb{S}^2. \tag{1.3}$$

We may then think of a function on $\mathbb{S}^2$ as a periodic function of θ and ϕ, and use bivariate trigonometric interpolation to fit the data. One problem with this straightforward approach is that not all the bivariate trigonometric polynomials are continuous functions of points on $\mathbb{S}^2$. Several authors have considered approximation by special classes of bivariate trigonometric polynomials and special bases for the same. We refer the reader to [2, sections 18.26, 18.27] and references therein for an interesting account.

In [5], a subspace of bivariate trigonometric polynomials was introduced for facilitating interpolation of functions on the sphere. However, a matrix-free interpolant construction was not discussed in [5]. In this paper, we introduce a different subspace spanned by a basis of the form $Q_n^m(\cos\theta)\exp(im\phi)$, where Q_n^m is the associated Chebyshev or Legendre functions of degree n and order at most 2. These two bases

lead to FFT-type matrix-free interpolation, corresponding to two systems of points on the sphere. This fact will be proved using a suitable quadrature formula as a consequence of Proposition 2.1.

We prove that the uniform norm of the interpolation operator, based on one of the systems of points, is $\mathcal{O}(\log^2 N)$. The other system of points allows us to construct quadrature formulas that are exact for integration of elements of a higher order space with respect to the area measure on $\mathbb{S}^2$. We demonstrate numerically that the two interpolation operators provide the same degree of approximation for the ten benchmark functions studied in [11] and references therein.

It might be interesting to compare interpolation from our spaces with that from the classical spherical polynomials. First, we observe that our space $\mathcal{X}_N$ defined after Theorem 2.1 below consists of all bivariate trigonometric polynomials of coordinate-wise degree at most N that are functions on the sphere and that do not contain terms with the frequencies $\sin N\theta$ (cf. (1.3)). Therefore, if $\mathbb{P}_n$ denotes the class of all spherical polynomials of degree at most n (i.e., the class of the restrictions to $\mathbb{S}^2$ of all trivariate algebraic polynomials of total degree at most n), then $\mathbb{P}_{N-1} \subset \mathcal{X}_N$. Moreover, the dimension of $\mathcal{X}_N$ is approximately twice that of $\mathbb{P}_{N-1}$.

The case of spherical polynomial interpolation is very different from our constructions. Sloan [10] has proved that it is not possible to construct a matrix–free interpolation operator onto $\mathbb{P}_N$ for $N \geq 3$ using a quadrature formula that is exact on $\mathbb{P}_{2N}$ (see [10, Lemma 3]). Several authors including Sündermann [13], Golitschek and Light [7], Xu [17, 18], and Laín Fernández [8] have constructed point systems with various symmetry properties for which a spherical polynomial interpolation matrix is guaranteed to be invertible. To the best of our knowledge, the norm of the resulting spherical polynomial interpolation operator as well as the computationally important condition number of the interpolation matrix have not been investigated.

The problem of finding points on the sphere for which a polynomial interpolation matrix is invertible and well conditioned has been studied computationally by Sloan and Womersley [11, 12] by imposing such additional constraints as to minimize the Lebesgue constant of the resulting operator or to maximize the determinant of the interpolatory matrix. It is conjectured that the Lebesgue constant of the best quality interpolation operator constructed in [11, 12] is $\mathcal{O}(N)$. In view of the smoothing properties of the surface integral operators arising in applications in elasticity theory, potential theory, and scattering of sound waves from three dimensional smooth spherical geometries, it is important for the projection operators to have Lebesgue constants that are $\mathcal{O}(N^\alpha)$ for some $\alpha < 1$ [4, 5, 6].

In the next section, we discuss the construction of the space and discuss the degree of approximation, interpolation, and quadrature in this space. Numerical experiments are presented in section 3, and the proofs of the results in section 2 are given in section 4.

2. Main results. In this paper, let C^* denote the class of all continuous functions on $\mathbb{R}^2$ that are 2π-periodic in each of their variables, equipped with the uniform norm $\|\circ\|_\infty^*$. The space of all continuous functions on $\mathbb{S}^2$, equipped with the uniform norm $\|\circ\|_\infty$, will be denoted by $C(\mathbb{S}^2)$. For $f \in C(\mathbb{S}^2)$, let $f^*(\theta,\phi) := f(\boldsymbol{p}(\theta,\phi))$, where $\boldsymbol{p}(\theta,\phi)$ are the polar coordinates defined in (1.3). It is clear that $f \in C(\mathbb{S}^2)$ if and only if $f^* \in C^*$, and f^* satisfies the following symmetry conditions:

$$f^*(-\theta,\phi+\pi) = f^*(\theta,\phi), \quad \theta,\phi \in \mathbb{R}, \tag{2.1}$$

and

$$f^*(0,\phi), \quad f^*(\pi,\phi) \text{ are independent of } \phi. \tag{2.2}$$

We will denote by C° the subspace of C^* comprising of functions satisfying the above two conditions. If $g \in C^\circ$, there exists a unique $f \in C(\mathbb{S}^2)$ such that $g = f^*$. We will write $f = g^\circ$. For integer $N \geq 2$, the space $\mathbb{H}_N$ denotes the class of all bivariate trigonometric polynomials of order at most N; i.e., the span of $\{e^{i\ell\theta}e^{im\phi}\}_{|\ell|,|m|\leq N}$. The space of all univariate algebraic polynomials of degree at most N will be denoted by Π_N.

Our first objective is to obtain a detailed description of $C^\circ \cap \mathbb{H}_N$.

THEOREM 2.1. *Let $N \geq 0$ be an integer. We have $T \in C^\circ \cap \mathbb{H}_N$ if and only if*

$$T(\theta,\phi) = S_0(\cos\theta) + \sin^2\theta \sum_{\substack{|\ell|\leq N,\ \ell\neq 0 \\ \ell \text{ even}}} Q_\ell(\cos\theta)\exp(i\ell\phi) + \sin\theta \sum_{\substack{|\ell|\leq N \\ \ell \text{ odd}}} R_\ell(\cos\theta)\exp(i\ell\phi) \tag{2.3}$$

$$= L(\cos\theta) + \sin^2\theta \sum_{\substack{|\ell|\leq N \\ \ell \text{ even}}} Q_\ell(\cos\theta)\exp(i\ell\phi) + \sin\theta \sum_{\substack{|\ell|\leq N \\ \ell \text{ odd}}} R_\ell(\cos\theta)\exp(i\ell\phi), \tag{2.4}$$

where $S_0 \in \Pi_N$, $L \in \Pi_1$, and for $|\ell| \leq N$, $Q_\ell \in \Pi_{N-2}$, $R_\ell \in \Pi_{N-1}$.

In this paper, we are interested in the spaces

$$\begin{aligned} \mathcal{X}_N^* &= \{T \in C^\circ \cap \mathbb{H}_N : T \text{ satisfies (2.3) with } R_\ell \in \Pi_{N-2}\}, \\ \mathcal{X}_N &= \{T^\circ : T \in \mathcal{X}_N^*\}. \end{aligned} \tag{2.5}$$

Thus, $\mathcal{X}_N$ comprises those T° for which the terms corresponding to $\sin N\theta$ are absent from the expansion of T. It is easy to see from (2.5) and (2.4) that the dimension of $\mathcal{X}_N$ is given by

$$d_N := 2 + (N-1)(2N+1) = 2N^2 - N + 1. \tag{2.6}$$

If $P \in \mathbb{P}_{N-1}$, then $P^* \in \mathbb{H}_{N-1} \cap C^\circ$. Therefore, $\mathbb{P}_{N-1} \subset \mathcal{X}_N$. The dimension of $\mathbb{P}_{N-1}$ is N^2, which is slightly more than half the dimension of $\mathcal{X}_N$. This is partly because $\mathcal{X}_N$ consists of polynomials with coordinatewise degree at most N, rather than total degree at most N, but clearly, $\mathcal{X}_N$ contains many elements which are not spherical polynomials of any degree.

Next, we describe the construction of a matrix-free interpolation scheme for the space $\mathcal{X}_N$. Before launching into the details of these constructions, we formulate in an abstract setting a proposition that shows a close connection between minimal quadrature formulas and interpolation. If Ω is a nonempty set, and V is a vector space of functions on Ω, we recall that a subset $\mathcal{C} \subseteq \Omega$ is a *set of uniqueness* for V if $P \in V$ and $P(x) = 0$ for all $x \in \mathcal{C}$ imply that $P \equiv 0$. The following proposition summarizes standard constructions in the theory of interpolation at the zeros of orthogonal polynomials on a real interval [3, section I.4] and appears in essence also as [10, Lemma 3].

PROPOSITION 2.1. *Let $d \geq 1$ be an integer, Ω be a set containing at least d elements, V be a d-dimensional vector space of functions on Ω, $\mathcal{C} = \{x_1, \ldots, x_d\} \subseteq \Omega$ be a set of uniqueness for V, and $w_1, \ldots, w_d > 0$. Let $\{\Phi_1, \ldots, \Phi_d\}$ be an orthonormal basis for V with respect to the inner product*

$$\langle P, Q\rangle = \sum_{k=1}^{d} w_k P(x_k)\overline{Q(x_k)},$$

and

$$K(x,y) := \sum_{k=1}^{d} \Phi_k(x)\overline{\Phi_k(y)}.$$

Then

$$w_k^{-1} = \sum_{j=1}^{d} |\Phi_j(x_k)|^2, \qquad k = 1, \ldots, d. \tag{2.7}$$

If $\mathcal{Y} := \{y_1, \ldots, y_d\} \subset \mathbb{C}$ *and*

$$g(\mathcal{Y}, x) := \sum_{k=1}^{d} w_k y_k K(x, x_k), \tag{2.8}$$

then $g(\mathcal{Y})$ *is the unique element of* V *satisfying* $g(\mathcal{Y}, x_j) = y_j$, $j = 1, \ldots, d$.

We note that $K(\circ, \circ)$ is the reproducing kernel of $(V, \langle \circ, \circ \rangle)$, and that for any $f \in V$, the interpolant $g(\{f(x_k)\}_{k=1,\ldots,d}, x)$ is just $\langle f, K(\circ, x)\rangle = f(x)$.

We now describe our matrix-free constructions for interpolation from the space $\mathcal{X}_N$. In what follows, we will assume that $N \geq 2$ and write

$$N_0 = N, \quad N_m = N - 2 \quad \text{for } m \neq 0. \tag{2.9}$$

First, we describe an orthonormal basis for $\mathcal{X}_N$. In the remainder of this section, let P_n denote the Legendre polynomial of degree n, normalized so that $P_n(1) = 1$. The associated Legendre function of degree n and order m is defined by

$$P_n^m(x) := (1 - x^2)^{m/2} \frac{d^m}{dx^m} P_n(x). \tag{2.10}$$

We recall that the classical spherical harmonics are defined by

$$Y_n^m(\boldsymbol{p}(\theta, \phi)) := \alpha_n^m P_n^{|m|}(\cos\theta) \exp(im\phi), \quad |m| \leq n, \tag{2.11}$$

where the *Condon–Shortley phase* α_n^m is given by

$$\alpha_n^m := \begin{cases} (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} & \text{if } m \geq 0, \\ (-1)^m \alpha_n^{|m|} & \text{if } m < 0. \end{cases} \tag{2.12}$$

The polynomials $\{Y_n^m \ : \ n = 0, \ldots, N, \ |m| \leq n\}$ form an orthonormal basis for the space $\mathbb{P}_N$ of all spherical polynomials of degree N, are eigenfunctions of the Laplace–Beltrami operator, and play a very important role in the theory of functions on the sphere. Following the representation (2.3), we define

$$G_n^m(\boldsymbol{p}(\theta, \phi)) := \begin{cases} Y_n^0(\boldsymbol{p}(\theta, 0)) & \text{if } m = 0, \\ Y_{n+1}^1(\boldsymbol{p}(\theta, 0)) \exp(im\phi) & \text{if } m \text{ is odd}, \\ Y_{n+2}^2(\boldsymbol{p}(\theta, 0)) \exp(im\phi) & \text{if } m \text{ is even}, m \neq 0. \end{cases} \tag{2.13}$$

Using (2.5) and (2.3), it is not difficult to check that $\{G_n^m \ : \ n = 0, \ldots, N_m, |m| \leq N\}$ is an orthonormal basis for $\mathcal{X}_N$:

$$\int_{\mathbb{S}^2} G_n^m(\boldsymbol{p}(\theta, \phi)) G_{n'}^{m'}(\boldsymbol{p}(\theta, \phi)) d(\boldsymbol{p}(\theta, \phi)) = \delta_{(n,m),(n',m')}, \tag{2.14}$$

where $d(\boldsymbol{p}(\theta,\phi))$ is the area element of $\mathbb{S}^2$, given by $\sin\theta d\theta d\phi$. They satisfy the following differential equations on $\mathbb{S}^2$ and in the unit ball in $\mathbb{R}^3$, where Δ_* denotes the Laplace–Beltrami operator, and Δ denotes the Laplacian in three variables.

$$(\Delta_* + n(n+1))\, G_n^m(\boldsymbol{p}(\theta,\phi)) = \begin{cases} 0 & \text{if } m=0, \\ \dfrac{1-m^2}{\sin^2(\theta)} G_n^m(\boldsymbol{p}(\theta,\phi)) & \text{if } m \text{ is odd}, \\ \dfrac{4-m^2}{\sin^2(\theta)} G_n^m(\boldsymbol{p}(\theta,\phi)) & \text{if } m \text{ is even}, m\neq 0, \end{cases} \tag{2.15}$$

$$\Delta\left(r^n G_n^m(\boldsymbol{p}(\theta,\phi))\right) = \begin{cases} 0 & \text{if } m=0, \\ \dfrac{1-m^2}{r^2\sin^2(\theta)} G_n^m(\boldsymbol{p}(\theta,\phi)) & \text{if } m \text{ is odd}, \\ \dfrac{4-m^2}{r^2\sin^2(\theta)} G_n^m(\boldsymbol{p}(\theta,\phi)) & \text{if } m \text{ is even}, m\neq 0. \end{cases} \tag{2.16}$$

In accordance with Proposition 2.1, we obtain a quadrature formula based on d_N points that is exact for integration of elements of $\mathcal{X}_{2N-1}$, and describe an interpolation operator for these nodes. This construction is based on the *Gauss–Lobatto quadrature rule*, given by the zeros of $P_N^1(\cos\theta)$. Thus, let $\widehat{\theta}_0 < \cdots < \widehat{\theta}_{N-2}$ be points on $(0,\pi)$, such that $P_N^{'}(\cos\widehat{\theta}_n) = 0$, $n = 0,\ldots,N-2$, $\widehat{\theta}_{N-1} = 0$, $\widehat{\theta}_N = \pi$, and

$$\widetilde{\phi}_m := \frac{2(m+N)\pi}{2N+1}, \quad -N \le m \le N. \tag{2.17}$$

Let

$$\mathcal{C}_N^q = \{\boldsymbol{p}(\widehat{\theta}_n, \widetilde{\phi}_m) \ : \ n = 0,\ldots,N-2,\ |m| \le N\} \cup \{\hat{\mathbf{n}}, \hat{\mathbf{s}}\}, \tag{2.18}$$

where $\hat{\mathbf{n}}$ and $\hat{\mathbf{s}}$ denote the north and south poles, respectively. We note that $\mathcal{C}_N^q$ contains exactly d_N elements. We define the corresponding discrete inner product by

$$\begin{aligned} \langle f, g\rangle_N^q = {} & \frac{4\pi}{N(N+1)(2N+1)} \sum_{|m|\le N} \sum_{n=0}^{N-2} \frac{f(\boldsymbol{p}(\widehat{\theta}_n,\widetilde{\phi}_m))\overline{g(\boldsymbol{p}(\widehat{\theta}_n,\widetilde{\phi}_m))}}{[P_N(\cos\widehat{\theta}_n)]^2} \\ & + \frac{4\pi}{N(N+1)}\left\{f(\hat{\mathbf{n}})\overline{g(\hat{\mathbf{n}})} + f(\hat{\mathbf{s}})\overline{g(\hat{\mathbf{s}})}\right\}. \end{aligned} \tag{2.19}$$

With

$$(g_n^m)^{-1} := \begin{cases} 2 + 1/N & \text{if } m=0 \text{ and } n=N, \\ 2 - 3/(N+2) & \text{if } 0\neq m \text{ even and } n = N-2, \\ 1 & \text{otherwise}, \end{cases} \tag{2.20}$$

we will show in the proof of Theorem 2.2 below that the functions $\{\sqrt{g_n^m} G_n^m\}$ form an orthonormal basis for $\mathcal{X}_N$, orthonormalized with respect to the inner product $\langle \circ, \circ\rangle_N^q$. In view of Proposition 2.1, a matrix–free interpolation operator can now be defined easily with the kernel

$$\mathcal{K}_N^q(\hat{\mathbf{x}},\hat{\mathbf{y}}) := \sum_{|m|\le N} \sum_{n=0}^{N_m} g_n^m G_n^m(\hat{\mathbf{x}}) G_n^m(\hat{\mathbf{y}}). \tag{2.21}$$

The following theorem summarizes some facts regarding our constructions so far.

THEOREM 2.2. *Let $N \geq 2$ be an integer. For $T \in \mathcal{X}_{2N-1}$, we have*

$$\int_{\mathbb{S}^2} T\left(\boldsymbol{p}(\theta,\phi)\right) d(\boldsymbol{p}(\theta,\phi)) = \frac{4\pi}{N(N+1)(2N+1)} \sum_{|m|\leq N} \sum_{n=0}^{N-2} \frac{T(\boldsymbol{p}(\widehat{\theta}_n, \widetilde{\phi}_m))}{[P_N(\cos\widehat{\theta}_n)]^2} + \frac{4\pi}{N(N+1)} \left\{T(\hat{\mathbf{n}}) + T(\hat{\mathbf{s}})\right\}. \tag{2.22}$$

For $f \in C(\mathbb{S}^2)$, let

$$\begin{aligned} \mathcal{G}_N f(\hat{\mathbf{x}}) := {} & \frac{4\pi}{N(N+1)(2N+1)} \sum_{|m|\leq N} \sum_{n=0}^{N-2} [P_N(\cos\widehat{\theta}_n)]^{-2} \\ & \times f(\boldsymbol{p}(\widehat{\theta}_n, \widetilde{\phi}_m)) \mathcal{K}_N^q(\hat{\mathbf{x}}, \boldsymbol{p}(\widehat{\theta}_n, \widetilde{\phi}_m)) \\ & + \frac{4\pi}{N(N+1)} \left\{f(\hat{\mathbf{n}}) \mathcal{K}_N^q(\hat{\mathbf{x}}, \hat{\mathbf{n}}) + f(\hat{\mathbf{s}}) \mathcal{K}_N^q(\hat{\mathbf{x}}, \hat{\mathbf{s}})\right\}. \end{aligned} \tag{2.23}$$

Then $\mathcal{G}_N f$ is the unique element of $\mathcal{X}_N$ that satisfies $\mathcal{G}_N f(\xi) = f(\xi)$ for $\xi \in \mathcal{C}_N^q$.

We believe that our proof of (2.31) below can be adapted to show that the Lebesgue constant of $\mathcal{G}_N$ is $\mathcal{O}(\sqrt{N})$, and we hope to report on this in the near future, along with similar constructions for spheres embedded in Euclidean spaces of dimensions higher than 3. On the other hand, we remark that a direct application of the representation (2.23) leads only to an estimate $\mathcal{O}(N)$, using standard techniques as in [14].

Next, we describe another construction for the nodes that leads to a matrix-free interpolation operator with uniform norm $\mathcal{O}((\log N)^2)$. Let

$$\widetilde{\theta}_n := \frac{(n+1)\pi}{N}, \quad n = 0, \ldots, N-2, \quad \widetilde{\theta}_{N-1} := 0,\ \widetilde{\theta}_N := \pi, \tag{2.24}$$

and $\widetilde{\phi}_m$ be defined by (2.17). For the points of interpolation, we choose the set

$$\mathcal{C}_N^i :== \{\boldsymbol{p}(\widetilde{\theta}_n, \widetilde{\phi}_m) : n = 0, \ldots, N-2,\ |m| \leq N\} \cup \{\hat{\mathbf{n}}, \hat{\mathbf{s}}\}.$$

To describe a basis for $\mathcal{X}_N$, we recall that the formula $T_n(\cos\theta) = \cos n\theta$ defines a unique polynomial T_n of degree n, called the Chebyshev polynomial (of first kind). Analogous to the associated Legendre functions, we define the associated Chebyshev functions by

$$C_n^m(x) = (1-x^2)^{m/2} \frac{d^m}{dx^m} T_n(x).$$

Our basis functions are defined by

$$Z_n^m(\boldsymbol{p}(\theta,\phi)) := \begin{cases} C_n^0(\cos\theta) & \text{if } m = 0, \\ C_{n+1}^1(\cos\theta) e^{im\phi} & \text{if } m \text{ is odd}, \\ C_{n+2}^2(\cos\theta) e^{im\phi} & \text{if } m \neq 0,\ m \text{ is even}. \end{cases} \tag{2.25}$$

These functions are not orthogonal with respect to the standard L^2 inner product on $\mathbb{S}^2$. However, we observe that an application of Proposition 2.1 requires only the

orthogonality of the functions with respect to a discrete inner product based at the points in question. Accordingly, we define

$$\langle f,g\rangle_N := \frac{2\pi^2}{N(2N+1)}\sum_{|m|\le N}\sum_{n=0}^{N-2} f(\boldsymbol{p}(\widetilde{\theta}_n,\widetilde{\phi}_m))\overline{g(\boldsymbol{p}(\widetilde{\theta}_n,\widetilde{\phi}_m))} + \frac{2\pi^2}{2N}\{f(\hat{\mathbf{n}})\overline{g(\hat{\mathbf{n}})}+f(\hat{\mathbf{s}})\overline{g(\hat{\mathbf{n}})}\}. \tag{2.26}$$

With the normalization factors

$$(z_{n,N}^m)^{-1} := \begin{cases} 2\pi^2 & \text{if } m=0,\ n=0,N,\\ \pi^2 & \text{if } m=0,\ n=1,\dots,N,\\ (n+1)^2\pi^2 & \text{if } m \text{ odd},\ n=0,\dots,N-2,\\ (n+2)^4\pi^2(1-(n+2)^{-2}) & \text{if } m \text{ even},\ m\neq 0,\ n=0,\dots,N-3,\\ 2(n+2)^4\pi^2(1-(n+2)^{-1}) & \text{if } m \text{ even},\ m\neq 0,\ n=N-2, \end{cases} \tag{2.27}$$

we will show in the proof of Theorem 2.3 below that the function $\{\sqrt{z_n^m}Z_n^m\}$ is an orthonormal basis for $\mathcal{X}_N$ with respect to this inner product. The interpolation operator can now be described using the kernel

$$\mathcal{K}_N(\hat{\mathbf{x}},\hat{\mathbf{y}}) := \sum_{|m|\le N}\sum_{n=0}^{N_m} z_{n,N}^m Z_n^m(\hat{\mathbf{x}})Z_n^m(\hat{\mathbf{y}}). \tag{2.28}$$

THEOREM 2.3. *Let $N\ge 2$ be an integer, $f\in C(\mathbb{S}^2)$, and*

$$\mathcal{I}_N f(\hat{\mathbf{x}}) := \frac{2\pi^2}{N(2N+1)}\sum_{|m|\le N}\sum_{n=0}^{N-2} f(\boldsymbol{p}(\widetilde{\theta}_n,\widetilde{\phi}_m))\mathcal{K}_N(\hat{\mathbf{x}},\boldsymbol{p}(\widetilde{\theta}_n,\widetilde{\phi}_m)) + \frac{\pi^2}{N}\left\{f(\hat{\mathbf{n}})\mathcal{K}_N(\hat{\mathbf{x}},\hat{\mathbf{n}})+f(\hat{\mathbf{s}})\mathcal{K}_N(\hat{\mathbf{x}},\hat{\mathbf{s}})\right\}. \tag{2.29}$$

Then $\mathcal{I}_N f$ is the unique element of $\mathcal{X}_N$ that satisfies $\mathcal{I}_N f(\xi)=f(\xi)$ for each $\xi\in\mathcal{C}_N^i$.

In the next theorem, we discuss the approximation properties of the operator $\mathcal{I}_N$. If $V\subset C^*$, we define

$$\text{dist }(f,V) := \inf_{P\in V}\|f-P\|_\infty^*, \quad f\in C^*, \tag{2.30}$$

with a similar definition for dist (f,V) when $f\in C(\mathbb{S}^2)$ and $V\subset C(\mathbb{S}^2)$. Throughout this paper, c denotes a generic constant, independent of N. Its value may be different at different occurrences, even within a single formula.

THEOREM 2.4. *For integer $N\ge 2$ and $f\in C(\mathbb{S}^2)$, we have*

$$\|\mathcal{I}_N f\|_\infty \le c(\log N)^2\|f\|_\infty \tag{2.31}$$

and

$$\|f-\mathcal{I}_N f\|_\infty \le c(\log N)^2 \text{ dist }(f^*,\mathbb{H}_{N-1}) \le c(\log N)^2 \text{ dist }(f,\mathbb{P}_{N-1}). \tag{2.32}$$

The connection between the smoothness of f^* and the superalgebraic convergence of dist $(f^*,\mathbb{H}_{N-1})$ to zero has been investigated in detail in classical approximation theory [16]. We note that since the space $\mathbb{P}_{N-1}^*$ corresponding to spherical polynomials of degree at most $N-1$ is contained in $\mathbb{H}_{N-1}$, dist $(f^*,\mathbb{H}_{N-1})\le$ dist $(f^*,\mathbb{P}_{N-1}^*)=$ dist$(f,\mathbb{P}_{N-1})$.

TABLE 1
Parameters for f_5.

i	$y_{i,1}$	$y_{i,2}$	$y_{i,3}$	α_i	β_i	γ_i
1	0	0	1	2	5	1
2	0.932039	0	0.362358	0.5	7	1
3	-0.362154	0.619228	0.696707	-2	6	2
4	0.904035	0.279651	-0.323290	-2	5	1
5	-0.0479317	-0.424684	-0.904072	0.2	2.1	1

3. Numerical experiments. We demonstrate the quality of our interpolatory operators by computing estimates of the uniform norm errors $\|\mathcal{I}_N f_i - f_i\|_\infty$ and $\|\mathcal{G}_N f_i - f_i\|_\infty,\ i = 1, \ldots, 10$, where $f_1, \ldots, f_{10}$ are the benchmark functions on the sphere (see [11] and references therein), defined for $\hat{\mathbf{x}} = (x_1, x_2, x_3) \in \mathbb{S}^2$, by

$$f_1(\hat{\mathbf{x}}) = x_1x_2x_3, \qquad f_2(\hat{\mathbf{x}}) = \exp(x_1), \qquad f_3(\hat{\mathbf{x}}) = \exp(x_1 + x_2 + x_3)/10,$$

$$f_4(\hat{\mathbf{x}}) = -5\sin(1 + 10x_3),\ f_5(\hat{\mathbf{x}}) = \sum_{i=1}^{5} \alpha_i \exp(-\beta_i \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathbf{y}}_i)^{2\gamma_i}),\ f_6(\hat{\mathbf{x}}) = \frac{1}{101 - 100x_3},$$

$$f_7(\hat{\mathbf{x}}) = |x_1| + |x_2| + |x_3|, \qquad f_8(\hat{\mathbf{x}}) = \frac{1}{f_7(\hat{\mathbf{x}})}, \qquad f_9(\hat{\mathbf{x}}) = \frac{\sin^2(1 + f_7(\hat{\mathbf{x}}))}{10},$$

and

$$f_{10}(\hat{\mathbf{x}}) = \begin{cases} \cos^2\left(\frac{3\pi}{2}\mathrm{dist}(\hat{\mathbf{x}}, \boldsymbol{p}(\pi/4, 5\pi/4))\right) & \text{if } \mathrm{dist}(\hat{\mathbf{x}}, \boldsymbol{p}(\pi/4, 5\pi/4)) < 1/3, \\ 0, & \text{if } \mathrm{dist}(\hat{\mathbf{x}}, \boldsymbol{p}(\pi/4, 5\pi/4)) \geq 1/3, \end{cases}$$

where the $\mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \cos^{-1}(\hat{\mathbf{x}} \cdot \hat{\mathbf{y}})$ is the geodesic distance between two points $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{S}^2$, and the parameters $\hat{\mathbf{y}}_i = (y_{i,1}, y_{i,2}, y_{i,3})$ and $\alpha_i, \beta_i, \gamma_i,\ i = 1, \ldots, 5$, in the test function f_5 are in Table 1. The functions $f_i,\ i = 1, \ldots, 6$, are analytic; f_7, f_8, and f_9 are continuous, but not continuously differentiable, and the locally supported cosine cap function f_{10} is once (but not twice) continuously differentiable function on $\mathbb{S}^2$.

We computed approximations of these functions in $\mathcal{X}_N$, using the interpolation operators $\mathcal{I}_N$ and $\mathcal{G}_N$. For each $i = 1, \ldots, 10$, the uniform norm errors $\|f_i - \mathcal{I}_N f_i\|_\infty$ and $\|f_i - \mathcal{G}_N f_i\|_\infty$ were estimated by taking the maximum of errors over $12,000$ points on the sphere.

The results in Tables 2–5 show that both $\mathcal{I}_N f_i$ and $\mathcal{G}_N f_i$ provide a similar quality of approximation of $f_i,\ i = 1, \ldots, 10$. The tables clearly demonstrate also that our operators yield a better reconstruction of these functions with their various smoothness properties than the (matrix-dependent) interpolatory and noninterpolatory polynomial approximations of the same functions discussed in [11, pp. 222–223]. (The functions denoted here by f_5 and f_6 are denoted in [11, pp. 222–223] by f_6 and f_5, respectively.) We note again that the construction of the interpolatory operators $\mathcal{I}_N f_i,\ \mathcal{G}_N f_i,\ i = 1, \ldots, 10$, does not require a numerical solution of any linear system of equations. Moreover, Theorem 2.4 shows that the Lebesgue constant of $\mathcal{I}_N$ is $\mathcal{O}\left((\log N)^2\right)$.

4. Proofs.

Proof of Theorem 2.1. It is clear that any expression of the form on the right-hand side of (2.4) is in $C^\circ \cap \mathbb{H}_N$. Let $T \in C^\circ \cap \mathbb{H}_N$, and

$$T(\theta, \phi) =: \sum_{|\ell|, |k| \leq N} a_{\ell,k} e^{ik\theta} e^{i\ell\phi}.$$

TABLE 2
Error in approximation of f_i by $\mathcal{I}_N f_i$, $i = 1, \ldots, 5$.

N	$\|f_1 - \mathcal{I}_N f_1\|_\infty$	$\|f_2 - \mathcal{I}_N f_2\|_\infty$	$\|f_3 - \mathcal{I}_N f_3\|_\infty$	$\|f_4 - \mathcal{I}_N f_4\|_\infty$	$\|f_5 - \mathcal{I}_N f_5\|_\infty$
4	1.5193e-15	1.0193e-03	5.4374e-02	1.1526e+01	7.0812e-01
8	1.5193e-15	2.1948e-08	1.9515e-05	6.7137e+00	1.3019e-01
16	1.5193e-15	8.2158e-15	2.2205e-14	7.1530e-03	2.6437e-03
32	1.5193e-15	1.5543e-14	2.2205e-14	1.0658e-13	5.9918e-07
64	1.5193e-15	4.4631e-14	2.0872e-14	2.3714e-13	2.1585e-11

TABLE 3
Error in approximation of f_i by $\mathcal{G}_N f_i$, $i = 1, \ldots, 5$.

N	$\|f_1 - \mathcal{G}_N f_1\|_\infty$	$\|f_2 - \mathcal{G}_N f_2\|_\infty$	$\|f_3 - \mathcal{G}_N f_3\|_\infty$	$\|f_4 - \mathcal{G}_N f_4\|_\infty$	$\|f_5 - \mathcal{G}_N f_5\|_\infty$
4	1.6653e-16	1.2257e-03	6.5224e-02	1.0562e+01	7.5962e-01
8	1.6653e-16	3.4587e-08	3.0874e-05	5.6223e+00	1.0930e-01
16	1.9429e-16	5.2180e-14	1.4522e-13	5.6956e-03	2.2566e-03
32	3.6082e-16	5.4622e-14	1.2346e-13	2.2027e-13	6.4830e-07
64	3.0531e-16	1.7064e-13	5.3824e-13	6.6702e-13	1.4792e-11

TABLE 4
Error in approximation of f_i by $\mathcal{I}_N f_i$, $i = 6, \ldots, 10$.

N	$\|f_6 - \mathcal{I}_N f_6\|_\infty$	$\|f_7 - \mathcal{I}_N f_7\|_\infty$	$\|f_8 - \mathcal{I}_N f_8\|_\infty$	$\|f_9 - \mathcal{I}_N f_9\|_\infty$	$\|f_{10} - \mathcal{I}_N f_{10}\|_\infty$
8	3.4509e-01	1.0125e-02	9.6995e-02	8.0456e-03	1.0205e-01
16	1.0003e-01	5.2213e-03	5.1501e-02	4.8926e-03	1.6087e-01
32	1.0757e-02	2.9341e-03	2.6055e-02	2.5565e-03	2.3648e-03
64	1.1523e-04	1.3091e-03	1.3079e-02	9.9133e-04	2.5106e-04
128	1.3881e-08	6.5481e-04	6.5467e-03	4.9564e-04	3.6030e-05

TABLE 5
Error in approximation of f_i by $\mathcal{G}_N f_i$, $i = 6, \ldots, 10$.

N	$\|f_6 - \mathcal{G}_N f_6\|_\infty$	$\|f_7 - \mathcal{G}_N f_7\|_\infty$	$\|f_8 - \mathcal{G}_N f_8\|_\infty$	$\|f_9 - \mathcal{G}_N f_9\|_\infty$	$\|f_{10} - \mathcal{G}_N f_{10}\|_\infty$
8	3.8245e-01	1.0690e-02	1.0331e-01	8.4990e-03	1.5496e-01
16	1.3193e-01	5.7501e-03	5.6980e-01	5.0567e-03	2.7559e-02
32	1.2847e-02	2.9679e-03	2.9636e-02	2.5629e-03	6.4724e-03
64	1.2910e-04	1.4640e-03	1.4645e-02	1.1080e-03	8.3897e-04
128	1.4840e-08	7.2705e-04	7.2712e-03	5.5022e-04	1.8309e-04

Then, recalling that for integers $k \geq 0$, $\cos k\theta$ and $\sin k\theta / \sin\theta$ are polynomials in $\cos\theta$ of degree k and $k-1$, respectively, we obtain

$$\begin{aligned} T(\theta,\phi) &= (1/2)(T(\theta,\phi) + T(-\theta,\phi+\pi)) \\ &= \sum_{|\ell|,|k|\leq N} a_{\ell,k} \frac{e^{ik\theta} + (-1)^\ell e^{-ik\theta}}{2} e^{i\ell\phi} \\ &= \sum_{\substack{|\ell|\leq N \\ \ell \text{ even}}} \sum_{k=0}^{N} a_{\ell,k} \cos k\theta e^{i\ell\phi} + i \sum_{\substack{|\ell|\leq N \\ \ell \text{ odd}}} \sum_{k=1}^{N} a_{\ell,k} \sin k\theta e^{i\ell\phi} \\ &= \sum_{\substack{|\ell|\leq N \\ \ell \text{ even}}} S_\ell(\cos\theta) e^{i\ell\phi} + \sin\theta \sum_{\substack{|\ell|\leq N \\ \ell \text{ odd}}} R_\ell(\cos\theta) e^{i\ell\phi}, \end{aligned}$$

where $S_\ell \in \Pi_N$ and $R_\ell \in \Pi_{N-1}$, $|\ell| \leq N$. There exist $Q_\ell \in \Pi_{N-2}$, $L_\ell \in \Pi_1$ such that $S_\ell(\cos\theta) = (1-\cos^2\theta)Q_\ell(\cos\theta) + L_\ell(\cos\theta)$. Since $T(0,\phi)$ and $T(\pi,\phi)$ are independent

of ϕ, we have $S_\ell(\pm 1) = 0$ if $\ell \neq 0$. Therefore, $L_\ell = 0$ for $|\ell| \leq N$, ℓ even, and $\ell \neq 0$. □

Proof of Proposition 2.1. Let $\mathbf{A} = [a_{j,k}]$ be the collocation matrix defined by $a_{j,k} = \Phi_k(x_j)$. Since $\mathcal{C}$ is a set of uniqueness, $\mathbf{A}$ is invertible. Also, the orthonormality of $\{\Phi_k \ : \ k = 1, \ldots, d\}$ is equivalent to the statement that $\overline{\mathbf{A}^T} D \mathbf{A} = I$, where $D = \text{diag}\,[w_1, \ldots, w_d]$. This leads to $\overline{\mathbf{A}^T} D = \mathbf{A}^{-1}$, and $D^{-1} = \mathbf{A}\overline{\mathbf{A}^T}$. This is (2.7). With $\mathbf{y} = [y_1, \ldots, y_d]^T$, and $\mathbf{b}(x) = [\Phi_1(x), \ldots, \Phi_d(x)]^T$, we have

$$g(\mathcal{Y}, x) = \mathbf{b}(x)^T \overline{\mathbf{A}^T} D \mathbf{y} = \mathbf{b}(x)^T \mathbf{A}^{-1} \mathbf{y}.$$

This completes the proof. □

The next lemma describes some sets of uniqueness for $\mathcal{X}_N$.

LEMMA 4.1. *Let $N \geq 2$ be an integer, $\theta_0, \ldots, \theta_{N-2}$ be distinct points in $(0, \pi)$, and $\theta_{N-1} = 0$, $\theta_N = \pi$, and ϕ_m, $|m| \leq N$, be distinct points on $[0, 2\pi)$. Then the set*

$$\mathcal{C} = \{\boldsymbol{p}(\theta_n, \phi_m) \ : \ n = 0, \ldots, N-2, \ |m| \leq N\} \cup \{\hat{\mathbf{n}}, \hat{\mathbf{s}}\}$$

consists of d_N distinct elements and is a set of uniqueness for $\mathcal{X}_N$.

Proof. Let $T \in \mathcal{X}_N^*$, and for $|\ell| \leq N$, $Q_\ell, R_\ell \in \Pi_{N-2}$, $Q_0 \in \Pi_N$ be found so that

$$T(\theta, \phi) = Q_0(\cos\theta) + \sin^2\theta \sum_{\substack{|\ell| \leq N, \ \ell \neq 0 \\ \ell \text{ even}}} Q_\ell(\cos\theta) \exp(i\ell\phi) + \sin\theta \sum_{\substack{|\ell| \leq N \\ \ell \text{ odd}}} R_\ell(\cos\theta) \exp(i\ell\phi),$$

and $T^\circ(\boldsymbol{p}(\theta_n, \phi_m)) = 0$, $n = 0, \ldots, N$, $|m| \leq N$. For any n, $T(\theta_n, \circ)$ is a trigonometric polynomial of degree at most N. Since this polynomial has $2N+1$ distinct zeros, $\{\phi_m\}_{|m| \leq N}$, it must be identically zero. This yields $Q_0(\cos\theta_n) = 0$ for $n = 0, \ldots, N$, and $Q_\ell(\cos\theta_n) = R_\ell(\cos\theta_n) = 0$, $n = 0, \ldots, N-2$, $\ell \neq 0$. Since $Q_0 \in \Pi_N$ and $Q_\ell, R_\ell \in \Pi_{N-2}$ for $\ell \neq 0$, this implies that each of these polynomials is identically equal to zero. Thus, $T \equiv 0$, and hence, $T^\circ \equiv 0$. □

We are now in a position to prove Theorems 2.2 and 2.3. We observe that for any integer $M \geq 1$, and integer k,

$$\frac{1}{M} \sum_{m=0}^{M-1} \exp(2\pi i k m/M) = \begin{cases} 1 & \text{if } k = 0 \bmod M, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

In particular,

$$\int_0^{2\pi} e^{ik\phi} d\phi = \frac{2\pi}{M} \sum_{m=0}^{M-1} \exp(2\pi i k m/M), \quad |k| \leq M-1. \tag{4.2}$$

Proof of Theorem 2.2. It is well known [1, (25.4.32), p. 888] that for $P \in \Pi_{2N-1}$, we have

$$\begin{aligned} &\int_0^\pi P(\cos\theta) \sin\theta d\theta \\ &\quad = \frac{2}{N(N+1)} \sum_{n=0}^{N-2} \frac{P(\cos\widehat{\theta}_n)}{\left[P_N(\cos\widehat{\theta}_n)\right]^2} + \frac{2}{N(N+1)} \left\{P(\cos\widehat{\theta}_0) + P(\cos\widehat{\theta}_N)\right\}. \end{aligned} \tag{4.3}$$

(The notation in [1] is somewhat different.) The equation (2.22) follows from the definition of the space $\mathcal{X}_N$ and the quadrature formulas (4.2) and (4.3).

In this proof only, we adopt the notation P_n^0 for the Legendre polynomial P_n and write

$$F_n^m(x) := \begin{cases} \alpha_n^0\ P_n^0(x) & \text{if } m = 0, \\ \alpha_{n+1}^1\ P_{n+1}^1(x) & \text{if } m \text{ is odd}, \\ \alpha_{n+2}^2\ P_{n+2}^2(x) & \text{if } m \text{ is even}, m \neq 0, \end{cases} \tag{4.4}$$

and we let $-N \le m, j \le N$, and $n = 0, \dots, N_m$, $l = 0, \dots, N_j$. Using (2.19), (2.13), (4.4), and (4.1) we get

$$\begin{aligned} \langle G_n^m, G_l^j \rangle_N^q = \delta_{m,j} \left[\frac{4\pi}{N(N+1)} \sum_{q=0}^{N-2} \frac{F_n^m(\cos\widehat{\theta}_q) F_l^j(\cos\widehat{\theta}_q)}{[P_N^0(\cos\widehat{\theta}_q)]^2} \right] \\ + \frac{4\pi}{N(N+1)} [F_n^m(-1)F_l^j(-1) + F_n^m(1)F_l^j(1)]. \end{aligned} \tag{4.5}$$

Let $j = m$. In view of (2.9) and (4.4), $F_n^m(x)F_l^m(x)$ is a polynomial of degree at most $2N-1$ on $[-1,1]$ for all $l = 0, \dots, N_m$, $0 \le n < N_m$, and also for $n = N_m$, if m is odd. Since the associated Legendre functions are orthonormal, (4.3) and (4.5) show that

$$\langle G_n^m, G_l^m \rangle_N^q = 0 \quad \text{if } n \neq l,$$

$$\langle G_n^m, G_n^m \rangle_N^q = 1 \quad \text{if } n \neq N_m, \quad \langle G_n^m, G_n^m \rangle_N^q = 1 \quad \text{if } n = N_m \quad \text{and} \quad m \text{ odd}.$$

Thus, we have shown that

$$\langle G_n^m, G_l^j \rangle_N^q = (g_n^m)^{-1} \delta_{n,l} \delta_{m,j} \tag{4.6}$$

for all n, m, l, j in question, except for the case when m is even, and $n = l = N_m$.

Next, let $m = 0$, and $n = l = N$. Using (4.5), (2.9), and (4.4), we have

$$\begin{aligned} \langle G_N^0, G_N^0 \rangle_N^q &= \frac{2N+1}{N(N+1)} \left(\sum_{q=0}^{N-2} \frac{[P_N^0(\cos\widehat{\theta}_q)]^2}{[P_N^0(\cos\widehat{\theta}_q)]^2} + [P_N^0(-1)]^2 + [P_N^0(1)]^2 \right) \\ &= \frac{2N+1}{N(N+1)} (N - 1 + 1 + 1) = 2 + 1/N. \end{aligned} \tag{4.7}$$

Finally, let m be even, $m \neq 0$, and $l = n = N_m = N-2$. In view of (4.4) and (2.12), we see that

$$[F_n^m(x)]^2 = \frac{2N+1}{4\pi} \frac{(N-2)!}{(N+2)!} [P_N^2(x)]^2.$$

Since $P_N^2(x) = (1-x^2)\frac{d^2}{dx^2} P_N^0(x)$, and P_N^0 is a solution of the Legendre differential equation [15, (4.2.1) with $\alpha = \beta = 0$], we have $P_N^2(x) = 2x\frac{d}{dx} P_N^0(x) - N(N+1)P_N^0(x)$. Since the Gauss–Lobatto quadrature points $x_q := \cos\widehat{\theta}_q \in (-1,1)$, $q = 0, \dots, N-2$, are zeros of the derivative of P_N^0, we have $P_N^2(x_q) = -N(N+1)P_N^0(x_q)$ for $q = 0, \dots, N-2$. We substitute this expression for $P_N^2(x_q)$ in (4.5), and recall that $P_N^2(\pm 1) = 0$ to obtain for even m

$$\langle G_{N-2}^m, G_{N-2}^m \rangle_N^q = \frac{2N+1}{(N-1)(N+2)} \left(\sum_{q=0}^{N-2} \frac{[P_N^0(x_q)]^2}{[P_N^0(x_q)]^2} \right) = \frac{2N+1}{(N+2)} = 2 - 3/(N+2). \tag{4.8}$$

The equations (4.6), (4.7), (4.8), and (2.20) show that $\{\sqrt{g_n^m}G_n^m\}$ is a basis for $\mathcal{X}_N$, orthonormalized with respect to the inner product $\langle \circ, \circ \rangle_N^q$. Lemma 4.1 shows that $\mathcal{C}_N^q$ is a set of uniqueness for $\mathcal{X}_N$. Therefore, the proof of Theorem 2.2 is complete in view of Proposition 2.1. ∎

The proof of Theorem 2.3 is very similar to that of Theorem 2.2, although the details are somewhat different. First, we obtain an analogue of (4.3).

LEMMA 4.2. *Let $N \geq 1$ be an integer. For $P \in \Pi_{2N-1}$, we have*

$$\int_0^\pi P(\cos\theta)d\theta = \frac{\pi}{N}\sum_{j=0}^{N-2} P(\cos\widetilde{\theta}_j) + \frac{\pi}{2N}\{P(\cos\widetilde{\theta}_{N-1}) + P(\cos\widetilde{\theta}_N)\}. \tag{4.9}$$

Proof. We observe that for $k = 0, \ldots, 2N-1$

$$\begin{aligned}
&\frac{1}{N}\sum_{j=0}^{N-2}\cos k\widetilde{\theta}_j + \frac{1}{2N}\{\cos k\widetilde{\theta}_{N-1} + \cos k\widetilde{\theta}_N\}\\
&\quad = \frac{1}{2N}\left\{\sum_{\ell=1}^{N-1}\exp(\pi ik\ell/N) + \exp(\pi ik(2N-\ell)/N)\right\} + \frac{1}{2N}\{\cos k(0) + \cos k\pi\}\\
&\quad = \frac{1}{2N}\sum_{\ell=0}^{2N-1}\exp(2\pi ik\ell/(2N)) = \delta_{k,0}.
\end{aligned}$$

This implies (4.9) when $P(\cos\theta) = \cos k\theta$, $k = 0, \ldots, 2N-1$. ∎

In the remainder of this section, we will write

$$\langle f, g\rangle_N^i := \frac{\pi}{N}\sum_{j=0}^{N-2} f(\cos\widetilde{\theta}_j)\overline{g(\cos\widetilde{\theta}_j)} + \frac{\pi}{2N}\left\{f(1)\overline{g(1)} + f(-1)\overline{g(-1)}\right\}. \tag{4.10}$$

The following lemma summarizes certain properties of the associated Chebyshev functions that we will need.

LEMMA 4.3. *Let $N \geq 2$ be an integer. Then $\langle C_n^m, C_{n'}^m\rangle_N^i = 0$ if $m = 0, 1, 2$, $n, n' = m, \ldots, N$, $n \neq n'$. Further,*

$$\langle C_n^m, C_n^m\rangle_N^i = \begin{cases} \pi & \text{if } m = 0,\ n = 0, N,\\ \pi/2 & \text{if } m = 0,\ n = 1, \ldots, N-1,\\ n^2\pi/2 & \text{if } m = 1,\ n = 1, \ldots, N-1,\\ (n^4\pi/2)(1-n^{-2}) & \text{if } m = 2,\ n = 2, \ldots, N-1,\\ n^4\pi(1-n^{-1}) & \text{if } m = 2,\ n = N. \end{cases} \tag{4.11}$$

Proof. In this proof only, we introduce the polynomials

$$U_n(\cos\theta) := \frac{\sin(n+1)\theta}{\sin\theta}, \qquad V_n(\cos\theta) := (1/2)U_{n+1}'(\cos\theta). \tag{4.12}$$

We note that

$$C_n^0 = T_n,\ C_n^1(\cos\theta) = n\sin n\theta = n\sin\theta U_{n-1}(\cos\theta),\ C_n^2(\cos\theta) = 2n\sin^2\theta V_{n-2}(\cos\theta).$$

Further, the polynomials U_n and V_n are the ultraspherical polynomials denoted in [15, p. 80] by $P_n^{(1)}$ and $P_n^{(2)}$, respectively. Therefore, using a straightforward computation

in the case of C_n^0 and C_n^1, and the orthogonality of $\{P_n^{(2)}\}$ in the case of C_n^2, we obtain

$$\int_0^\pi C_n^m(\cos\theta)C_{n'}^m(\cos\theta)d\theta = 0, \quad m = 0,1,2,\ n \neq n', \quad n = m, m+1, \ldots. \tag{4.13}$$

Similarly, using [15, (4.7.15), p. 81] in the case of C_n^2, we get

$$\int_0^\pi C_n^m(\cos\theta)^2 d\theta = \begin{cases} \pi & \text{if } n = m = 0, \\ \pi/2 & \text{if } m = 0, n = 1,2,\ldots, \\ n^2\pi/2 & \text{if } m = 1, n = 1,2,\ldots, \\ (n^4\pi/2)(1-n^{-2}) & \text{if } m = 2, n = 2,3,\ldots. \end{cases} \tag{4.14}$$

The quadrature formula (4.9) now shows that $\langle C_n^m, C_{n'}^m\rangle_N^i = 0$ if $m = 0,1,2$, $n, n' = m,\ldots,N$, $n \neq n'$, as well as all the equations in (4.11), except for the cases $m = 0$, $n = N$, and $m = 2$, $n = N$. The case $m = 0$, $n = N$ is clear from the definitions. Let $m = 2$, $n = N$. From the differential equation for Chebyshev polynomials, we see that with $x = \cos\theta$,

$$C_N^2(x) = (1-x^2)T_N''(x) = xT_N'(x) - N^2T_N(x) = N\cot\theta\sin(N\theta) - N^2\cos(N\theta).$$

Thus, $C_N^2(\pm 1) = 0$, and $C_N^2(\cos\widetilde{\theta}_j) = (-1)^j N^2$. The last equation in (4.11) is now easy to obtain from the definitions. ∎

Proof of Theorem 2.3. In this proof only, let

$$F_n^m(x) := \begin{cases} C_n^0(x) & \text{if } m = 0,\ n = 0,\ldots,N, \\ C_{n+1}^1(x) & \text{if } m \text{ is odd, } |m| \le N,\ n = 0,\ldots,N-2, \\ C_{n+2}^2(x) & \text{if } m \text{ is even, } |m| \le N,\ m \neq 0,\ n = 0,\ldots,N-2. \end{cases}$$

Using (4.1) with $M = 2N+1$ and the fact that $C_{n+m}^m(\pm 1) = 0$ if $m = 1,2$, we obtain as in the proof of Theorem 2.2 that for integers $|m|, |m'| \le N$, $n = 0,\ldots,N_m$, $n' = 0,\ldots,N_{m'}$,

$$\langle Z_{n'}^{m'}, Z_n^m\rangle_N = \begin{cases} 0, & m \neq m' \text{ or } n \neq n', \\ 2\pi\langle F_n^m, F_n^m\rangle_N^i & \text{if } n = n',\ m = m'. \end{cases} \tag{4.15}$$

Together with (2.27) and (4.11), this shows that $\{\sqrt{z_n^m}Z_n^m\}$ is an orthonormal basis for $\mathcal{X}_N$. Lemma 4.1 shows that $\mathcal{C}_N^i$ is a set of uniqueness for $\mathcal{X}_N$. Therefore, the proof of Theorem 2.2 is complete in view of Proposition 2.1. ∎

In order to prove Theorem 2.4, we need a representation for $\mathcal{I}_N f$ in (2.29) using the Dirichlet kernels. For integer $m \ge 1$, let

$$D_m^*(\theta) = \frac{1}{2} + \sum_{k=1}^{m-1}\cos k\theta + \frac{1}{2}\cos m\theta = \frac{\sin m\theta}{2\tan(\theta/2)},$$

$$D_m(\phi) = \sum_{|k|\le m}\exp(ik\phi) = \frac{\sin(m+1/2)\phi}{\sin(\phi/2)} = D_{m,e}(\phi) + D_{m,o}(\phi), \tag{4.16}$$

where

$$D_{m,e}(\phi) = \sum_{|2k|\le m}\exp(2ki\phi) = \frac{1}{2}\{D_m(\phi) + D_m(\phi+\pi)\},$$

$$D_{m,o}(\phi) = \sum_{|2k+1|\le m}\exp((2k+1)i\phi) = \frac{1}{2}\{D_m(\phi) - D_m(\phi+\pi)\}.$$

We note that

$$D_N^*(\widetilde{\theta}_j - \widetilde{\theta}_k) = N\delta_{j,k},\ D_N(\widetilde{\phi}_m - \widetilde{\phi}_\ell) = (2N+1)\delta_{\ell,m}. \tag{4.17}$$

LEMMA 4.4. *For $f \in C(\mathbb{S}^2)$, we have*

(4.18)

$$\begin{aligned}
&\mathcal{I}_N f(\boldsymbol{p}(\theta,\phi)) \\
&= \frac{1}{N(2N+1)} \sum_{j=0}^{N-2} \sum_{|m|\le N} f(\boldsymbol{p}(\widetilde{\theta}_j, \widetilde{\phi}_m))\{(D_N^*(\theta - \widetilde{\theta}_j) + D_N^*(\theta + \widetilde{\theta}_j))D_{N,e}(\phi - \widetilde{\phi}_m) \\
&\qquad + (D_N^*(\theta - \widetilde{\theta}_j) - D_N^*(\theta + \widetilde{\theta}_j))D_{N,o}(\phi - \widetilde{\phi}_m)\} \\
&\qquad + \frac{1}{N}\{f(\hat{\mathbf{n}})D_N^*(\theta) + f(\hat{\mathbf{s}})D_N^*(\theta - \pi)\}.
\end{aligned}$$

Proof. In this proof only, we denote the right-hand side of (4.18) by $T(\theta,\phi)$. Clearly, each of the summands on the right-hand side of (4.18), and hence T, is in $\mathbb{H}_N$. We observe that for all θ, ϕ, $j = 0,\dots,N-2$, $|m| \le N$,

$$\begin{aligned}
&(D_N^*(-\theta - \widetilde{\theta}_j) + D_N^*(-\theta + \widetilde{\theta}_j))D_{N,e}((\phi+\pi) - \widetilde{\phi}_m) \\
&= (D_N^*(\theta - \widetilde{\theta}_j) + D_N^*(\theta + \widetilde{\theta}_j))D_{N,e}(\phi - \widetilde{\phi}_m),
\end{aligned}$$

and

$$D_N^*(-\widetilde{\theta}_j) + D_N^*(\widetilde{\theta}_j) = D_N^*(\pi - \widetilde{\theta}_j) + D_N^*(\pi + \widetilde{\theta}_j) = 0.$$

Hence, for $j = 0,\dots,N-2$, $|m| \le N$,

$$(D_N^*(\theta - \widetilde{\theta}_j) + D_N^*(\theta + \widetilde{\theta}_j))D_{N,e}(\phi - \widetilde{\phi}_m) \in \mathcal{X}_N^*.$$

Similarly, for all θ, ϕ, $j = 0,\dots,N-2$, $|m| \le N$,

$$\begin{aligned}
&(D_N^*(-\theta - \widetilde{\theta}_j) - D_N^*(-\theta + \widetilde{\theta}_j))D_{N,o}((\phi+\pi) - \widetilde{\phi}_m) \\
&= (D_N^*(\theta - \widetilde{\theta}_j) - D_N^*(\theta + \widetilde{\theta}_j))D_{N,o}(\phi - \widetilde{\phi}_m),
\end{aligned}$$

and

$$D_N^*(-\widetilde{\theta}_j) - D_N^*(\widetilde{\theta}_j) = D_N^*(\pi - \widetilde{\theta}_j) - D_N^*(\pi + \widetilde{\theta}_j) = 0.$$

Moreover,

$$\cos N(\theta - \widetilde{\theta}_j) - \cos N(\theta + \widetilde{\theta}_j) = \cos(N\theta - j\pi - \pi) - \cos(N\theta + j\pi + \pi) = 0.$$

Therefore, none of the terms $D_N^*(\theta - \widetilde{\theta}_j) - D_N^*(\theta + \widetilde{\theta}_j)$ can contain a term involving $\sin N\theta$. Thus,

$$(D_N^*(\theta - \widetilde{\theta}_j) - D_N^*(\theta + \widetilde{\theta}_j))D_{N,o}(\phi - \widetilde{\phi}_m) \in \mathcal{X}_N^*.$$

It is clear that $D_N^*(\theta)$ and $D_N^*(\theta - \pi)$ are also in $\mathcal{X}_N^*$. Thus, each of the summands on the right-hand side in (4.18) may be viewed as functions on $\mathbb{S}^2$, and as such, are in $\mathcal{X}_N$. Thus, $T^\circ \in \mathcal{X}_N$.

Now, for $\ell = 0, \dots, N-2$, $|\nu| \le N$, we may use (4.17) and (4.1) to conclude that

$$\begin{aligned} &T(\widetilde{\theta}_\ell, \widetilde{\phi}_\nu) \\ &\quad = \frac{1}{2N+1} \sum_{|m|\le N} f(\boldsymbol{p}(\widetilde{\theta}_\ell, \widetilde{\phi}_m))\big\{D_{N,e}(\widetilde{\phi}_\nu - \widetilde{\phi}_m) + D_{N,o}(\widetilde{\phi}_\nu - \widetilde{\phi}_m)\big\} \\ &\quad = \frac{1}{2N+1} \sum_{|m|\le N} f(\boldsymbol{p}(\widetilde{\theta}_\ell, \widetilde{\phi}_m)) D_N(\widetilde{\phi}_\nu - \widetilde{\phi}_m) = f(\boldsymbol{p}(\widetilde{\theta}_\ell, \widetilde{\phi}_\nu)). \end{aligned}$$

The equation (4.17) also leads to $T^\circ(\hat{\mathbf{n}}) = f(\hat{\mathbf{n}})$ and $T^\circ(\hat{\mathbf{s}}) = f(\hat{\mathbf{s}})$. ∎

Our next lemma relates the degrees of approximation of a function $f \in C(\mathbb{S}^2)$ from $\mathcal{X}_N$ with that of $f^* \in C^\circ$ from $\mathbb{H}_N$.

LEMMA 4.5. *Let $N \ge 2$ be an integer, $f \in C(\mathbb{S}^2)$. Then*

$$\text{dist } (f^*, \mathbb{H}_N) \le \text{ dist } (f, \mathcal{X}_N) \le 5 \text{ dist } (f^*, \mathbb{H}_{N-1}). \tag{4.19}$$

Proof. The first inequality in (4.19) is obvious since $\mathcal{X}_N \subset \mathbb{H}_N$. Let $T \in \mathbb{H}_{N-1}$ be chosen so that $\|f^* - T\|_\infty^* = \text{ dist } (f^*, \mathbb{H}_{N-1})$. Then $U(\theta, \phi) := (1/2)[T(\theta, \phi) + T(-\theta, \phi + \pi)]$ satisfies (2.1), and

$$\|f^* - U\|_\infty^* = \text{dist } (f^*, \mathbb{H}_{N-1}). \tag{4.20}$$

Since U satisfies (2.1), there exist $S_\ell \in \Pi_{N-1}$, $R_\ell \in \Pi_{N-2}$, $|\ell| \le N-1$, such that

$$U(\theta, \phi) = \sum_{\substack{|\ell|\le N-1 \\ \ell \text{ even}}} S_\ell(\cos\theta) e^{i\ell\phi} + \sin\theta \sum_{\substack{|\ell|\le N-1 \\ \ell \text{ odd}}} R_\ell(\cos\theta) e^{i\ell\phi}.$$

For any $\phi \in \mathbb{R}$, we have

$$\max\left\{ \left| f(\hat{\mathbf{n}}) - \sum_{\substack{|\ell|\le N-1 \\ \ell \text{ even}}} S_\ell(1) e^{i\ell\phi} \right|, \ \left| f(\hat{\mathbf{s}}) - \sum_{\substack{|\ell|\le N-1 \\ \ell \text{ even}}} S_\ell(-1) e^{i\ell\phi} \right| \right\} \le \text{ dist } (f^*, \mathbb{H}_{N-1}). \tag{4.21}$$

Therefore,

$$|f(\hat{\mathbf{n}}) - S_0(1)| = \left| \frac{1}{2\pi} \int_0^{2\pi} \left\{ f(\hat{\mathbf{n}}) - \sum_{\substack{|\ell|\le N-1 \\ \ell \text{ even}}} S_\ell(1) e^{i\ell\phi} \right\} d\phi \right| \le \text{ dist } (f^*, \mathbb{H}_{N-1}). \tag{4.22}$$

Similarly,

$$|f(\hat{\mathbf{s}}) - S_0(-1)| \le \text{ dist } (f^*, \mathbb{H}_{N-1}). \tag{4.23}$$

The estimates (4.21), (4.22), (4.23) lead to

$$\left| \sum_{\substack{1\le|\ell|\le N-1 \\ \ell \text{ even}}} S_\ell(\pm 1) e^{i\ell\phi} \right| \le 2 \text{ dist } (f^*, \mathbb{H}_{N-1}). \tag{4.24}$$

Now, let

$$\widetilde{S}_\ell(x) = S_\ell(x) - S_\ell(1)(1+x)/2 - S_\ell(-1)(1-x)/2, \qquad 1 \le |\ell| \le N-1,\ \ell \text{ even},$$

and

$$\widetilde{U}(\theta,\phi) = S_0(\cos\theta) + \sum_{\substack{1\le|\ell|\le N-1\\ \ell \text{ even}}} \widetilde{S}_\ell(\cos\theta)e^{i\ell\phi} + \sin\theta \sum_{\substack{|\ell|\le N-1\\ \ell \text{ odd}}} R_\ell(\cos\theta)e^{i\ell\phi}.$$

In view of Theorem 2.1, $\widetilde{U} \in \mathcal{X}_N^*$. It is easy to verify using (4.20), (4.22), (4.23), and (4.24) that $\|f^* - \widetilde{U}\|_\infty^* \le 5 \text{ dist } (f^*, \mathbb{H}_{N-1})$. □

Finally, we are in a position to prove Theorem 2.4.

Proof of Theorem 2.4. In this proof only, let

$$\begin{aligned} L_N := \sup_{\theta,\phi\in\mathbb{R}} \Bigg[& \frac{1}{N(2N+1)} \sum_{j=0}^{N-2} \sum_{|m|\le N} \big\{ \big| D_N^*(\theta-\widetilde{\theta}_j) + D_N^*(\theta+\widetilde{\theta}_j)\big| |D_{N,e}(\phi-\widetilde{\phi}_m)| \\ & + \big| D_N^*(\theta-\widetilde{\theta}_j) - D_N^*(\theta+\widetilde{\theta}_j)\big| |D_{N,o}(\phi-\widetilde{\phi}_m)| \big\} \\ & + \frac{1}{N}\left\{ |D_N^*(\theta)| + |D_N^*(\theta-\pi)| \right\} \Bigg]. \end{aligned} \tag{4.25}$$

Using [9, (3.4), (3.6)], we estimate the discrete sums above by the integral norms of the Dirichlet kernels. Well-known bounds on Dirichlet kernels (see, for example, [5, 19]) now imply that

$$L_N \le c \int_{-\pi}^{\pi} |D_N^*(t)| dt \int_{-\pi}^{\pi} \{ |D_{N,e}(t)| + |D_{N,o}(t)| \}\, dt + c \le c(\log N)^2.$$

The estimate (2.31) is now clear from (4.18).

If $T \in \mathcal{X}_N$, then $\mathcal{I}_N(T) = T$. Therefore, using (2.31), we obtain that for any $T \in \mathcal{X}_N$,

$$\begin{aligned} \|f - \mathcal{I}_N f\|_\infty = \|f - T - \mathcal{I}_N(f-T)\|_\infty &\le \|f-T\|_\infty + c(\log N)^2 \|f-T\|_\infty \\ &\le c(\log N)^2 \|f-T\|_\infty. \end{aligned}$$

This implies (2.32). □

Acknowledgment. The authors thank the referees for their careful reading and many suggestions for improving the first draft of this paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, NY, 1972.
[2] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover Publications, Mineola, NY, 2001.
[3] G. Freud, *Orthogonal Polynomials*, Académiai Kiado, Budapest, 1971.
[4] M. Ganesh, I. G. Graham, and J. Sivaloganathan, *A pseudospectral three–dimensional boundary integral method applied to a nonlinear model problem from finite elasticity*, SIAM J. Numer. Anal., 31 (1994), pp. 1378–1414.
[5] M. Ganesh, I. G. Graham, and J. Sivaloganathan, *A new spectral boundary integral collocation method for three-dimensional potential problems*, SIAM J. Numer. Anal., 35 (1998), pp. 778–805.
[6] M. Ganesh and I.G. Graham, *A high-order algorithm for obstacle scattering in three dimensions*, J. Comput. Phys., 198 (2004), pp. 211–242.

[7] M. V. Golitschek and W. Light, *Interpolation by polynomials and radial basis functions on spheres*, Constr. Approx., 17 (2001), pp. 1–18.
[8] N. Laín Fernández, *Polynomial Bases on the Sphere*, Ph.D. thesis, Universität zu Lübeck, Logos Verlag, Berlin, 2003.
[9] H. N. Mhaskar and J. Prestin, *On the detection of singularities of a periodic function*, Adv. Comput. Math., 12 (2000), pp. 95–131.
[10] I.H. Sloan, *Polynomial interpolation and hyperinterpolation over general regions*, J. Approx. Theory, 83 (1995), pp. 238–254.
[11] I. H. Sloan and R. S. Womersley, *How good can polynomial interpolation on the sphere be?*, Adv. Comput. Math., 14 (2001) pp. 195–226.
[12] I. H. Sloan and R. S. Womersley, *Extremal systems of points and numerical integration on the sphere*, Adv. Comput. Math., 21 (2004), pp. 107–125.
[13] B. Sündermann, *Projektionen auf Polynomraumen in mehreren veranderlichen*, Ph.D. thesis, Universitat Dortmund, 1983.
[14] J. Szabados and P. Vértesi, *Interpolation of Functions*, World Scientific, Singapore, 1990.
[15] G. Szegő, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 22, AMS, Providence, RI, 1975.
[16] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Pergamon Press, 1963.
[17] Y. Xu, *Polynomial interpolation on the unit sphere*, SIAM J. Numer. Anal., 41 (2003), pp. 751–766.
[18] Y. Xu, *Polynomial interpolation on the unit ball and on the unit sphere*, Adv. Comput. Math., 20 (2004), pp. 247–260.
[19] A. Zygmund, *Trigonometric Series*,, Cambridge University Press, Cambridge, UK, 1977.

SIAM J. NUMER. ANAL.
Vol. 44, No. 3, pp. 1332–1350

A SPECTRAL ORDER METHOD FOR INVERTING SECTORIAL LAPLACE TRANSFORMS*

MARÍA LÓPEZ-FERNÁNDEZ†, CÉSAR PALENCIA†, AND ACHIM SCHÄDLE‡

Abstract. Laplace transforms which admit a holomorphic extension to some sector strictly containing the right half plane and exhibiting a potential behavior are considered. A spectral order, parallelizable method for their numerical inversion is proposed. The method takes into account the available information about the errors arising in the evaluations. Several numerical illustrations are provided.

Key words. Laplace transform, numerical inversion, parabolic, spectral order, parallelizable

AMS subject classifications. 65R10, 65J10

DOI. 10.1137/050629653

1. Introduction.

In a variety of situations, the problem arises of inverting numerically the Laplace transform $U(z)$ of a given mapping of interest $u(t)$. Roughly speaking, it turns out that the wider the set W where $U(z)$ can be computed, the easier the inversion is. For instance, if W is an interval (a, b) then the numerical inversion becomes an ill-posed problem [1, 5, 6]. On the other hand, if W is the complement of some bounded region, then the efficient Talbot method [16, 22] is at hand.

In the present paper we focus on the particular situation where W is a sector symmetric with respect to the real axis, strictly containing the right half plane, and we assume that $U(z)$ exhibits a potential behavior on W. We say then that $U(z)$ is *sectorial*. Precisely, there is a renewed interest in the numerical inversion of *sectorial* mappings [7, 8, 9, 10, 13, 14, 18] mainly due to its applicability to linear, non-homogeneous evolution equations of parabolic type (both in the context of abstract IVPs and Volterra equations), as well as their discretizations in space [3, 4, 11]. In this context, the inversion approach presents several computational advantages (a drastic reduction of the number of linear systems to be solved and two levels of parallelism), its main disadvantages being that the Laplace transform $F(z)$ of the source term must be *sectorial* (which in turn, as we comment below, demands the source term to be holomorphic on a sector containing the half axis $t > 0$) and that it requires evaluations of $F(z)$ at nodes with $\operatorname{Re} z < 0$. This issue is considered in [9]. Another way to overcome these difficulties is presented in [14], where the source term is locally approximated by holomorphic mappings with simple Laplace transforms. More recently, these restrictions are overcome in [12, 17], where the ideas in the present paper are adapted so as to provide accurate reconstructions of the traditional Runge–Kutta approximations to the solutions of such parabolic problems. These reconstructions require no regularity on the source term of the problem.

In the present paper we consider the issue of the numerical inversion of *sectorial* mappings by itself. By definition, a holomorphic mapping taking values in a complex

*Received by the editors April 20, 2005; accepted for publication January 27, 2006; published electronically July 7, 2006.

http://www.siam.org/journals/sinum/44-3/62965.html

†Departamento de Matemática Aplicada, Universidad de Valladolid, Valladolid, Spain (marial@mac.cie.uva.es, palencia@mac.cie.uva.es). The work of these authors was supported by DGI-MCYT under project MTM 2004-07194 cofinanced by FEDER funds.

‡ZIB Berlin, Takustr. 7, D-14195 Berlin, Germany (schaedle@zib.de). The work of this author was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

Banach space X,

$$U : W \subset \mathbb{C} \to X,$$

is said to be *sectorial* if W is the complement of some acute sector of the form

$$\Sigma_\delta = \{z \in \mathbb{C} : |\arg(-z)| \le \delta\}, \qquad 0 < \delta < \frac{\pi}{2}, \tag{1.1}$$

and if there exist constants $M > 0$ and $\mu \in \mathbb{R}$ such that

$$\|U(z)\| \le \frac{M}{|z|^\mu}, \qquad z \notin \Sigma_\delta. \tag{1.2}$$

It is known that a *sectorial* mapping is a Laplace transform (which for $\mu \le 0$ is understood in the Operational Calculus sense). In fact, let $U : W \to \mathbb{C}$ be a holomorphic mapping satisfying (1.1) and (1.2) for some $0 < \delta < \pi/2$, $M > 0$, and $\mu \in \mathbb{R}$. For $t > 0$, set

$$u(t) = \frac{1}{2\pi i} \int_\Gamma e^{tz} U(z)\, dz, \tag{1.3}$$

where Γ is a simple contour, lying in W, and parametrizable by a regular mapping $S : (-\infty, +\infty) \to \mathbb{C}$ such that

$$\lim_{x \to \pm\infty} \operatorname{Im} S(x) = \pm\infty \quad \text{and} \quad \lim_{x \to \pm\infty} \frac{\operatorname{Re} S(x)}{|x|} < 0.$$

Since the last condition implies

$$\operatorname{Re} z \le -b|z| \qquad \text{as } z \to \infty,\ z \in \Gamma,$$

for some $b > 0$, the integrand in (1.3) is absolutely convergent. Moreover, $u(t)$ is independent of the particular choice of Γ. Then, proceeding as in the proof of Theorem 2.6.1 in [2], it is easy to conclude that:

(i) u admits a holomorphic extension $u(\tau)$ to any sector of the form $|\arg(\tau)| \le \delta'$, with $0 < \delta' < \pi/2 - \delta$, and there $\|u(\tau)\| = \mathcal{O}(|\tau|^{\mu-1})$.

(ii) If $\mu > 0$, then U is the Laplace transform of u in the classical sense; i.e.,

$$U(z) = \int_0^{+\infty} e^{-zt} u(t)\, dt, \qquad \operatorname{Re} z > 0.$$

(iii) If $\mu \le 0$, then $u(t)$ might not be integrable in the neighborhood of the origin. However, after selecting an integer number $m \ge 1$, with $m + \mu \ge 1$, the previous comment shows that u is the derivative of order m of a mapping $v : (0, +\infty) \to X$, whose Laplace transform is $V(z) = U(z)/z^m$.

Thus, our goal is to numerically reconstruct $u(t)$ from the knowledge of a moderate number of evaluations of $U(z)$ at suitable nodes $z \notin \Sigma_\delta$. Let us point out that, from a practical point of view, it is essential to take into account that these evaluations are going to be affected by errors.

Notice that in case $U(z)$ satisfies a similar inequality,

$$\|U(z)\| \le \frac{M}{|z - \omega|^\mu}, \qquad z \notin \omega + \Sigma_\delta,$$

for some $\omega \in \mathbb{R}$, then, by using the shifting theorem, the inversion of $U(z)$ is reduced to that of a Laplace transform $\widetilde{U}(z)$ fulfilling (1.2). Since the respective originals $u(t)$ and $\widetilde{u}(t)$ are related by $u(t) = e^{\omega t}\widetilde{u}(t)$, we can just approximate $\widetilde{u}(t)$. This is why the analysis is restricted to the situation $\omega = 0$, i.e., to (1.2).

Now, as in [8, 9, 10, 13, 14], we choose Γ as the branch of a hyperbola and a parametrization $S : (-\infty, +\infty) \to \mathbb{C}$ of Γ which admits a holomorphic extension to a horizontal strip around the real axis. The numerical method we propose is simply the truncated trapezoidal rule, applied to the definite integral arising after parametrizing (1.3) by S, used with $2n + 1$ nodes $x_k = kh$, $-n \le k \le n$, and a suitable step size $h > 0$. The properties of S allow us to use the ideas and results in [20, 21], where the trapezoidal rule applied to holomorphic mappings on strips is considered. As already noted in [8, 9, 10, 13], the fast decay of our integrand yields an improvement of the more general estimates in [20, 21].

Very often, for instance, in the context of IVPs (see Illustration 3 in section 5), the main computational effort of the method is due to the evaluations of $U(z)$ at the nodes $z_k = S(x_k)$, $-n \le k \le n$. An important feature of the present approach is that the same evaluations can be used to approximate $u(t)$ at different $t > 0$ [7, 8, 9, 13, 16]. Accordingly, our goal is to obtain a uniform error estimate for the approximation of $u(t)$ on intervals of the form $[t_0, \Lambda t_0]$, with given $t_0 > 0$ and $\Lambda \ge 1$, rather than at a fixed $t > 0$. Essentially, this was the aim in [13], whose basic estimates we borrow. Notice also that the algorithm presents two levels of parallelism since, first, the evaluations of $U(z)$ at the involved nodes and, second, the evaluations of $u(t)$ at a selected finite set of values of $t \in [t_0, \Lambda t_0]$, can be carried out on different processors [7, 8, 9, 13, 14].

In the present paper, by considering a different choice of the geometrical and scale parameters from the one in [8, 9, 13], we improve the results there in two different ways:

(i) We get a better error bound, which now turns out to be a genuine spectral estimate of the form $\mathcal{O}(e^{-cn})$, instead of $\mathcal{O}(e^{-cn/\ln n})$.
(ii) We also get a weaker dependence of the exponential factor c on Λ than in [13], since now $c = \mathcal{O}(1/\ln \Lambda)$.

This means, in practice, that with a moderate number of evaluations of $U(z)$ we can accurately approximate $u(t)$ uniformly on intervals $[t_0, t_1]$ with $\Lambda = t_1/t_0 >> 1$, say, $\Lambda = 50$. Moreover, it is interesting to note that, for $\mu > 1$ in (1.2), with a different selection of parameters, we can achieve a uniform error estimate like $\mathcal{O}(e^{-c\sqrt{n}})$, for $0 \le t \le 1$, by using the same quadrature nodes. This can be shown by an argument similar to the one used in the proof of Proposition 2.7 in [9].

On the other hand, for the choice of parameters we propose, the precision ρ used in the evaluations of $U(z)$ at the required nodes plays a more relevant role than in [13]. In fact, ignoring that we always have $\rho > 0$ would result in large actual errors for $n >> 1$, as simple numerical experiments show (see Illustration 1 in section 5). This drawback is overcome by minimizing the estimate we get for the actual error (Theorem 2), which leads to a (ρ, n)-dependent choice of parameters. With this choice, the actual error finally behaves for moderate n like $\mathcal{O}(e^{-cn})$, with $c = \mathcal{O}(1/\ln \Lambda)$, and for large n like $\mathcal{O}(\rho)$. This optimal choice of parameters demands, of course, some information about the size of ρ. In the absence of it, we propose an n-dependent choice of parameters for which the actual error behaves like $\mathcal{O}(\rho + e^{-cn})$, with $c = \mathcal{O}(1/(\ln n + \ln \Lambda))$. All the above estimates are uniform on $t_0 \le t \le \Lambda t_0$, with fixed $t_0 > 0$ and $\Lambda > 1$. Moreover, the error constants are made explicit in the analysis and turn out to be reasonable.

The outline of the paper is as follows. In section 2 we describe the numerical method and show, in Theorem 1, how to achieve (i) and (ii). The propagation of errors is studied in section 3. The choice of parameters is considered in section 4, and four simple numerical illustrations of the theoretical results are provided in section 5.

2. The numerical method. Given δ in (1.1) and following the ideas in [13], we select α, $d > 0$ such that

$$0 < \alpha - d < \alpha + d < \frac{\pi}{2} - \delta. \tag{2.1}$$

Defining

$$T(w) = 1 - \sin(\alpha + iw),$$

this mapping transforms each horizontal straight line $\operatorname{Im} w = y$, $-d \le y \le d$, into the left branch of the hyperbola given by

$$\left(\frac{\operatorname{Re} z - 1}{\sin(\alpha - y)}\right)^2 - \left(\frac{\operatorname{Im} z}{\cos(\alpha - y)}\right)^2 = 1, \tag{2.2}$$

with center at $(1, 0)$ and foci at $(0, 0)$ and $(2, 0)$, whose asymptotes make angles $\pm[\pi/2 - (\alpha - y)]$ with the real axis. Therefore, T transforms the horizontal strip

$$D_d = \{z \in \mathbb{C} : |\operatorname{Im} z| \le d\}$$

into the region in the complex plane limited by the left branches corresponding to $y = \pm d$ in (2.2); cf. Figure 2.1.

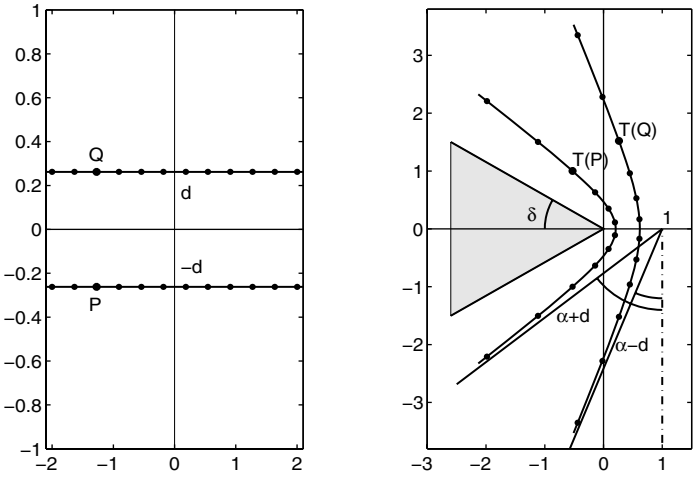


FIG. 2.1. *The horizontal strip (left) is transformed into the area limited by the left branches of two hyperbola (right). Additionally the sector Σ_δ is shown.*

Introducing a parameter $\lambda > 0$, the parametrization of Γ in (1.3) can be defined as

$$\Gamma = \{\lambda T(x) : x \in \mathbb{R}\};$$

i.e., Γ is the branch of a hyperbola corresponding to the image of the real axis under $S = \lambda T$. This results in

$$u(t) = \int_{-\infty}^{+\infty} G_t(x)\, dx\ , \qquad t > 0,$$

where $G_t : D_d \to X,\ t > 0$, is the mapping

$$G_t(w) = -\frac{\lambda}{2\pi i} \exp(\lambda t T(w)) U(\lambda T(w)) T'(w).$$

Once the parameters α, d, and λ have been fixed, we set $x_k = kh,\ k \in \mathbb{Z}$, and consider the approximation to $u(t)$ given by

$$u_n(t) = h \sum_{k=-n}^{n} G_t(x_k), \qquad t > 0. \tag{2.3}$$

The proof of the main result in [13, Theorem 2], shows that for $\mu = 1$ in (1.2)

(2.4)
$$\|u(t) - u_n(t)\| \le M \cdot \varphi(\alpha, d) \cdot L(\lambda t \sin(\alpha - d)) \cdot e^{\lambda t} \Big(\frac{1}{e^{2\pi d/h} - 1} + \frac{1}{e^{\lambda t \sin\alpha \cosh(nh)}} \Big),$$

where

$$\varphi(\alpha, d) = \frac{2}{\pi} \sqrt{\frac{1 + \sin(\alpha + d)}{1 - \sin(\alpha + d)}},$$

and $L(x)$, $x > 0$, is the function

$$L(x) = 1 + |\ln(1 - e^{-x})|.$$

Notice that $L(x)$ is decreasing in x, $L(x) \approx |\ln x|$, as $x \to 0^+$, and $L(x)$ tends to 1, as $x \to +\infty$.

As we commented in the introduction, in many applications the computational effort to obtain $u_n(t)$ is mainly due to the evaluations of $U(z)$ at $z = \lambda T(x_k),\ -n \le k \le n$, but these evaluations could be carried out in parallel. Another attractive feature of (2.3) is that the same evaluations of $U(z)$ can be used to compute $u_n(t)$ for different $t > 0$. In fact, as we see below, with the appropriate choice of parameters, we can use the same evaluations of $U(z)$ so as to have a spectral estimate

$$\|u(t) - u_n(t)\| = \mathcal{O}(e^{-cn})$$

uniform on intervals $t_0 \le t \le t_1$. The exponential factor c turns out to depend weakly on the ratio $\Lambda = t_1/t_0$, given that $c = \mathcal{O}(1/\ln\Lambda)$.

For simplicity the next theorem is restricted to the situation $\mu = 1$ in (1.2). The cases $\mu > 1$ and $\mu < 1$ are treated in subsequent remarks.

THEOREM 1. *Assume that U satisfies* (1.2) *with $\mu = 1$. Fixing α and d according to* (2.1)*, for $t_0 > 0$, $\Lambda \ge 1$, $0 < \theta < 1$, and $n \ge 1$, the choice of parameters*

$$h = \frac{1}{n} a(\theta), \qquad \lambda = \frac{2\pi d n (1 - \theta)}{t_0 \Lambda a(\theta)}, \tag{2.5}$$

with

$$a(\theta) = \operatorname{arccosh}\Big(\frac{\Lambda}{(1-\theta)\sin\alpha}\Big),$$

yields the uniform estimate on $t_0 \le t \le \Lambda t_0$,

$$\|u(t) - u_n(t)\| \le M \cdot \varphi(\alpha, d) \cdot L(\lambda t_0 \sin(\alpha - d)) \cdot \frac{2\epsilon_n(\theta)^\theta}{1 - \epsilon_n(\theta)}, \tag{2.6}$$

where

$$\epsilon_n(\theta) = \exp\Big(-\frac{2\pi d}{a(\theta)} n\Big).$$

The theorem shows, just by selecting any $0 < \theta < 1$, a genuine spectral order of convergence in n of the form $\mathcal{O}(e^{-cn})$, where $c = \mathcal{O}(1/\ln\Lambda)$ (cf. [8, 9, 10, 13]).

Proof. Set $\sigma = \lambda t_0$. For $t_0 \le t \le \Lambda t_0$, (2.4) implies the uniform bound

$$\|u(t) - u_n(t)\| \le M \cdot \varphi(\alpha, d) \cdot L(\sigma \sin(\alpha - d)) \cdot e^{\Lambda\sigma}\Big(\frac{1}{e^{2\pi d/h} - 1} + \frac{1}{e^{\sigma \sin\alpha \cosh(nh)}}\Big).$$

Our choice of h and λ is precisely the one guaranteeing that

$$\exp\Big(\frac{2\pi d}{h}\Big) = \exp(\sigma \sin\alpha \cosh(nh)) = \frac{1}{\epsilon_n(\theta)};$$

hence

$$\frac{1}{e^{2\pi d/h} - 1} + \frac{1}{e^{\sigma \sin\alpha \cosh(nh)}} \le \frac{2e^{-2\pi d/h}}{1 - e^{-2\pi d/h}} = \frac{2\epsilon_n(\theta)}{1 - \epsilon_n(\theta)}.$$

The proof ends after remarking that

$$e^{\Lambda\sigma}\epsilon_n(\theta) = \epsilon_n(\theta)^{\theta-1}\epsilon_n(\theta) = \epsilon_n(\theta)^\theta. \qquad \square$$

To end the section we comment, in the two following remarks, on the situation $\mu \ne 1$ in (1.2). We omit details in the proofs, which are completely analogous to the one of Theorem 1.

Remark 1. Assume that U satisfies (1.2) with $\mu > 1$. By Remark 1 in [13] we have

$$\|u(t) - u_n(t)\| \le M \cdot \varphi(\alpha, d, \mu) \cdot L(\lambda t \sin(\alpha - d)) \cdot \frac{e^{\lambda t}}{\lambda^{\mu-1}}\Big(\frac{1}{e^{2\pi d/h} - 1} + \frac{1}{e^{\lambda t \sin\alpha \cosh(nh)}}\Big),$$

where

$$\varphi(\alpha, d, \mu) = \frac{2}{\pi}\sqrt{\frac{1 + \sin(\alpha + d)}{(1 - \sin(\alpha + d))^{2\mu-1}}}.$$

Thus, for $0 < \theta < 1$, the same choice of values for h and λ as in Theorem 1 gives the bound

$$\|u(t) - u_n(t)\| \le M \cdot \varphi(\alpha, d, \mu) \cdot L(\lambda t_0 \sin(\alpha - d)) \cdot \lambda^{1-\mu} \cdot \frac{2\epsilon_n(\theta)^\theta}{1 - \epsilon_n(\theta)},$$

uniformly for $t_0 \le t \le \Lambda t_0$. This estimate is again spectral in n, since

$$\lambda^{1-\mu} = \mathcal{O}\Big(\Big(\frac{\Lambda t_0}{n}\Big)^{\mu-1}\Big).$$

Remark 2. Assume now that U satisfies (1.2) with $\mu < 1$. By Remark 1 in [13], for a fixed $s \in (0,1)$, there holds

$$\|u(t)-u_n(t)\| \le M\cdot\varphi_s(\alpha,d,\mu)\cdot L(s\lambda t\sin(\alpha-d))\cdot\frac{e^{\lambda t}}{t^{1-\mu}}\Big(\frac{1}{e^{2\pi d/h}-1}+\frac{1}{e^{s\lambda t\sin\alpha\cosh(nh)}}\Big),$$

where now

$$\varphi_s(\alpha,d,\mu) = \frac{2}{\pi}\sqrt{\frac{1+\sin(\alpha+d)}{1-\sin(\alpha+d)}}\Big(\frac{1-\mu}{(1-s)e\sin(\alpha-d)}\Big)^{1-\mu}.$$

In this situation, for $\theta \in (0,1)$ we choose

$$h = \frac{1}{n}a_s(\theta), \qquad \lambda = \frac{2\pi dn(1-\theta)}{t_0\Lambda a_s(\theta)},$$

where

$$a_s(\theta) = \operatorname{arccosh}\Big(\frac{\Lambda}{\mathrm{s}(1-\theta)\sin\alpha}\Big).$$

Setting

$$\epsilon_{s,n}(\theta) = \exp\Big(\frac{-2\pi dn}{a_s(\theta)}\Big),$$

we get the spectral estimate

$$\|u(t)-u_n(t)\| \le M\cdot\varphi_s(\alpha,d,\mu)\cdot L(s\lambda t_0\sin(\alpha-d))\cdot t_0^{\mu-1}\frac{2\epsilon_{s,n}(\theta)^\theta}{1-\epsilon_{s,n}(\theta)}$$

uniformly for $t_0 \le t \le \Lambda t_0$.

3. Error propagation. Numerical experiments (see section 5) show that for large values of n the estimate (2.6) is no longer true in practice. The explanation of this apparently contradictory behavior lies in the influence of the errors when evaluating U and the elementary functions involved. For the sake of simplicity, we consider first the case $\mu = 1$ in (1.2). The situations $\mu > 1$ and $\mu < 1$ are considered in subsequent remarks.

Let $z_k = \lambda T(x_k)$, $-n \le k \le n$, be the nodes used in (2.3). Clearly, in practice, as numerical approximation to $u(t)$ we actually obtain

$$\bar{u}_n(t) = \sum_{k=-n}^{n}\omega_k(t)U_k, \tag{3.1}$$

where, for $-n \le k \le n$, $\omega_k(t) \in \mathbb{C}$ and $U_k \in X$, are approximations to

$$-\frac{\lambda h}{2\pi i}\exp(\lambda t z_k)T'(x_k)$$

and $U(z_k)$, respectively.

To estimate the actual error $\|u(t) - \bar{u}_n(t)\|$ we need to make some assumptions on the approximations used. To this end, we are going to focus on two frequent possibilities, depending on whether we have information on absolute or relative errors due to the evaluations. To be precise, we are going to assume that there exists $\rho > 0$ such that, simultaneously for all $-n \le k \le n$, we have either

$$\|U(z_k) - U_k\| \le \rho \quad \text{and} \quad \omega_k(t) = -\frac{\lambda h}{2\pi i} \exp(\lambda t z_k) T'(x_k) \tag{3.2}$$

or

$$\left\| -\frac{\lambda h}{2\pi i} \exp(\lambda t z_k) T'(x_k) U(z_k) - \omega_k(t) U_k \right\| \le \rho \left\| -\frac{\lambda h}{2\pi i} \exp(\lambda t z_k) T'(x_k) U(z_k) \right\|. \tag{3.3}$$

Situation (3.2) arises, for instance, when $U_k \approx U(z_k)$ are provided by means of some auxiliary routine, say, by solving a linear system, with prescribed absolute accuracy ρ, and moreover the errors due to the evaluations of the elementary functions involved turn out to be negligible compared to ρ. Situation (3.3) is typical when $U(z)$ is an elementary function that can be evaluated with relative accuracy ρ.

The next theorem yields an estimate of the actual error for these situations. We maintain the notation introduced in Theorem 1.

THEOREM 2. *Assume that U satisfies* (1.2) *with $\mu = 1$. Fix α, d according to* (2.1). *For $t_0 > 0$, $\Lambda \ge 1$, $0 < \theta < 1$, and $n \ge 1$, select the parameters*

$$h = \frac{1}{n} a(\theta), \qquad \lambda = \frac{2\pi d n (1-\theta)}{t_0 \Lambda a(\theta)}.$$

Assume also that $\omega_k(t) \in \mathbb{C}$, $t_0 \le t \le t_1$, $U_k \in X$, $-n \le k \le n$, satisfy either (3.2) *or* (3.3). *Then, the actual error is estimated by*

$$\|u(t) - \bar{u}_n(t)\| \le M \cdot \Phi \cdot Q \cdot \left(\varepsilon \epsilon_n(\theta)^{\theta - 1} + \frac{\epsilon_n(\theta)^{\theta}}{1 - \epsilon_n(\theta)} \right), \tag{3.4}$$

uniformly on $t_0 \le t \le \Lambda t_0$, where either

(a) $\varepsilon = \rho/(M t_0)$,

$$\Phi = \max\left\{ \frac{2}{\pi} \sqrt{\frac{1 + \sin(\alpha + d)}{1 - \sin(\alpha + d)}}, \frac{1}{2\pi \sin \alpha} \right\},$$

and

$$Q = \max\{2L(\lambda t_0 \sin(\alpha - d)), 2 + (2 + \lambda t_0)h\}$$

in case (3.2) *holds, or*

(b) $\varepsilon = \rho$,

$$\Phi = \frac{2}{\pi} \sqrt{\frac{1 + \sin(\alpha + d)}{1 - \sin(\alpha + d)}},$$

and

$$Q = \max\{2L(\lambda t_0 \sin(\alpha - d)), 1/2(h + L(\lambda t_0 \sin \alpha))\},$$

in case (3.3) *holds.*

Notice that Q depends logarithmically on α, d, $1-\theta$, and Λ.

The estimate (3.4) given by the theorem, with a fixed $0<\theta<1$, shows again a spectral order of convergence $\mathcal{O}(e^{-cn})$, with $c=\mathcal{O}(1/\ln\Lambda)$, but only for moderate n, to be more precise, as long as $\epsilon_n(\theta)\geq\varepsilon$. On the other hand, for fixed θ, (3.4) goes to $+\infty$ exponentially as $n\to+\infty$. However, this apparent drawback is overcome by selecting θ in a suitable way, as we explain in section 4.

Proof. By writing

$$\|u(t)-\bar{u}_n(t)\|\leq\|u(t)-u_n(t)\|+\|u_n(t)-\bar{u}_n(t)\|,$$

and noticing that, for the corresponding Q, (2.6) implies

$$\|u(t)-u_n(t)\|\leq M\cdot\Phi\cdot Q\frac{\epsilon_n(\theta)^\theta}{1-\epsilon_n(\theta)},$$

the proof is reduced to show that

$$\|u_n(t)-\bar{u}_n(t)\|\leq M\cdot\Phi\cdot Q\varepsilon\epsilon_n(\theta)^{\theta-1}. \tag{3.5}$$

Assume first that (3.2) holds. This situation was already studied in section 5 in [13]. As it is shown there, for $t_0\leq t\leq\Lambda t_0$, we have

$$\|u_n(t)-\bar{u}_n(t)\|\leq\frac{\lambda e^{\lambda\Lambda t_0}\rho}{2\pi}h\sum_{k=-n}^{n}e^{-\gamma\cosh x_k}\cosh x_k,$$

where $\gamma=\lambda t_0\sin\alpha$. By noticing that the function $se^{-\gamma s}$, $s\geq0$, attains its maximum $1/(\gamma e)$ at the point $s_0=1/\gamma$ and is monotonic on the intervals $[0,s_0]$ and $[s_0,+\infty)$, it is easy to see that

$$\begin{aligned}h\sum_{k=-n}^{n}e^{-\gamma\cosh x_k}\cosh x_k&\leq h+2h\sum_{k=1}^{n}e^{-\gamma\cosh x_k}\cosh x_k\\&\leq h+\frac{4h}{\gamma e}+2\int_0^{+\infty}e^{-\gamma\cosh x}\cosh x\,dx\\&\leq\frac{(\gamma+2)h+2}{\gamma},\end{aligned}$$

whence, recalling that $\varepsilon=\rho/(t_0M)$, we get

$$\|u_n(t)-\bar{u}_n(t)\|\leq\frac{M}{2\pi\sin\alpha}[2+(2+\lambda t_0)h]e^{\lambda\Lambda t_0}\varepsilon. \tag{3.6}$$

Using now that

$$e^{\Lambda\lambda t_0}=\epsilon_n(\theta)^{\theta-1}, \tag{3.7}$$

we readily obtain (3.5).

Assume now that (3.3) holds. Proceeding as in the proof of Lemma 1 and Theorem 2 in [13], and denoting

$$\varphi(\alpha,0)=\frac{2}{\pi}\sqrt{\frac{1+\sin\alpha}{1-\sin\alpha}},$$

we get

$$\begin{aligned}\|u_n(t)-\bar{u}_n(t)\| &\le \frac{\rho M e^{\lambda t}}{2\pi} h \sum_{k=-n}^{n} e^{-\lambda t \sin\alpha \cosh x_k}\left|\frac{T'(x_k)}{T(x_k)}\right| \\ &\le \frac{M\varphi(\alpha,0)}{4}\rho e^{\lambda t} h \sum_{k=-n}^{n} e^{-\lambda t \sin\alpha \cosh x_k} \\ &\le \frac{M\varphi(\alpha,0)}{2}\rho e^{\lambda t}\left(h+\int_0^{+\infty} e^{-\lambda t \sin\alpha \cosh x}\,dx\right) \\ &\le \frac{M\varphi(\alpha,0)}{2}\rho e^{\lambda t}(h+L(\lambda t \sin\alpha)).\end{aligned}$$

Hence, using again (3.7) and the inequality $\varphi(\alpha,0)\le\varphi(\alpha,d)$, we deduce (3.5). ∎

The behavior for $\mu \neq 1$ in (1.2) is considered in the following remarks, whose proofs are a combination of Remarks 1 and 2 and the arguments used in [13, Theorem 2]. Notice that (3.6) is independent of μ.

Remark 3. Assume that $\mu > 1$ in (1.2) and fix $0<\theta<1$. Then, for the choice of parameters in Theorem 2 and uniformly on $t_0 \le t \le \Lambda t_0$, we have the following.

(a) In case (3.2) it holds that

$$\|u(t)-\bar{u}_n(t)\| \le M\cdot\Phi\cdot Q\cdot\left(\varepsilon\epsilon_n(\theta)^{\theta-1}+\lambda^{1-\mu}\frac{\epsilon_n(\theta)^{\theta}}{1-\epsilon_n(\theta)}\right),$$

with $\varepsilon=\rho/(Mt_0)$,

$$\Phi=\max\left\{\frac{2}{\pi}\sqrt{\frac{1+\sin(\alpha+d)}{(1-\sin(\alpha+d))^{2\mu-1}}},\frac{1}{2\pi\sin\alpha}\right\},$$

and

$$Q=\max\{2L(\lambda t_0\sin(\alpha-d)),2+(2+\lambda t_0)h\}.$$

(b) In case (3.3) it holds that

$$\|u(t)-\bar{u}_n(t)\| \le M\cdot\Phi\cdot Q\cdot\lambda^{1-\mu}\cdot\left(\varepsilon\epsilon_n(\theta)^{\theta-1}+\frac{\epsilon_n(\theta)^{\theta}}{1-\epsilon_n(\theta)}\right),$$

with $\varepsilon=\rho$,

$$\Phi=\frac{2}{\pi}\sqrt{\frac{1+\sin(\alpha+d)}{(1-\sin(\alpha+d))^{2\mu-1}}},$$

and

$$Q=\max\{2L(\lambda t_0\sin(\alpha-d)),1/2(h+L(\lambda t_0\sin\alpha))\}.$$

Remark 4. Assume that $\mu<1$ in (1.2) and fix $0<s,\theta<1$. Then, for the choice of parameters in Remark 2 and uniformly on $t_0\le t\le\Lambda t_0$, we have the following.

(a) In case (3.2) it holds that

$$\|u(t) - \bar{u}_n(t)\| \leq M \cdot \Phi \cdot Q \cdot \left(\varepsilon \epsilon_{s,n}(\theta)^{\theta-1} + t_0^{\mu-1} \frac{\epsilon_{s,n}(\theta)^{\theta}}{1 - \epsilon_{s,n}(\theta)} \right),$$

with $\varepsilon = \rho/(Mt_0)$,

$$\Phi = \max \left\{ \frac{2}{\pi} \sqrt{\frac{1+\sin(\alpha+d)}{1-\sin(\alpha+d)}} \Big(\frac{1-\mu}{(1-s)e\sin(\alpha-d)} \Big)^{1-\mu}, \frac{1}{2\pi \sin\alpha} \right\},$$

and

$$Q = \max\{2L(s\lambda t_0 \sin(\alpha - d)), 2 + (2 + \lambda t_0)h\}.$$

(b) In case (3.3) it holds that

$$\|u(t) - \bar{u}_n(t)\| \leq M \cdot \Phi \cdot Q \cdot \left(\lambda^{1-\mu} \varepsilon \epsilon_{s,n}(\theta)^{\theta-1} + t_0^{\mu-1} \frac{\epsilon_{s,n}(\theta)^{\theta}}{1 - \epsilon_{s,n}(\theta)} \right),$$

with $\varepsilon = \rho$,

$$\Phi = \frac{2}{\pi} \sqrt{\frac{1+\sin(\alpha+d)}{1-\sin(\alpha+d)}} \Big(\frac{1-\mu}{(1-s)e\sin(\alpha-d)} \Big)^{1-\mu},$$

and

$$Q = \max\{2L(s\lambda t_0 \sin(\alpha - d)), 1/2(h + L(s\lambda t_0 \sin\alpha))\}.$$

4. The choice of parameters. With Theorem 2 in mind, we now try to derive a strategy for the choice of parameters. First, (3.4) shows that it is of interest to select α away from zero and $\alpha + d$ away from $\pi/2$. The dependence of the actual error on $\alpha - d$ is less important, since it is logarithmic.

Suppose α and d have already been chosen; then for a given n we take h and λ as indicated in Theorem 2 and we fix $0 < \theta < 1$. Assume also that we have an estimation of ρ and set $\varepsilon = \rho/(Mt_0)$ or $\varepsilon = \rho$ as in Theorem 2. Then, since in practice we always have $\rho > 0$ and hence $\varepsilon > 0$, it turns out that $\varepsilon \epsilon_n(\theta)^{\theta-1} \to +\infty$ as $n \to +\infty$. Hence, it is clear that increasing the number of nodes might result in a worse estimate (3.4). In fact, increasing n may result in worse approximations, as Illustration 1 in section 5 shows.

To overcome this drawback we let θ be a free parameter for the moment. Given $\varepsilon > 0$ and n, after selecting α and d, neglecting the logarithmic factor Q, and taking into account that typically $\epsilon_n(\theta) << 1$, the best thing we can do is to choose $0 < \theta < 1$ so as to minimize the term

$$\varepsilon \epsilon_n(\theta)^{\theta-1} + \epsilon_n(\theta)^{\theta}; \tag{4.1}$$

i.e., we must tune θ depending on $\varepsilon > 0$ and n. By a direct calculation it can be proven that the first derivative of $\epsilon_n(\theta)^{\theta-1}$ with respect to θ is increasing in θ. The same is true for $\epsilon_n(\theta)^{\theta}$ (in this case the proof, though elementary, is more difficult). We conclude that the expression in (4.1) defines a strictly convex function of θ. Moreover, its derivative is < 0 at 0^+ and tends to $+\infty$ as $\theta \to 1^-$. Therefore, (4.1) attains its minimum exactly for one value $\theta_{\varepsilon,n} \in (0, 1)$, which is the one we propose to be used.

Though it is not easy to express the dependence of $\theta_{\varepsilon,n}$ on n and ε, this can be easily done numerically (see section 5).

Since, up to logarithmic factors, the choice $\theta = \theta_{\varepsilon,n}$ in (3.4) is optimal, it is clear that with this choice we get for the actual error:

(a) A spectral order of convergence $\mathcal{O}(e^{-cn})$, with $c = \mathcal{O}(1/\ln \Lambda)$, for moderate values of n. In fact, already for any fixed $0 < \theta < 1$, (3.4) shows that the error behaves like $\mathcal{O}(e^{-cn})$ as long as $\epsilon_n(\theta) \le \varepsilon$, i.e., for $n = \mathcal{O}(|\ln \varepsilon|)$.

(b) The errors are not propagated. In fact, already with the nonoptimal choice

$$\theta = 1 - \frac{1}{n},$$

(3.4) reads

$$\|u(t) - \bar{u}_n(t)\| = \mathcal{O}(\varepsilon + e^{-cn}) \tag{4.2}$$

uniformly on $t_0 \le t \le \Lambda t_0$, with $c = \mathcal{O}(1/(\ln \Lambda + \ln n))$. This remark tells us that, for large values of n, the actual error saturates at level ε, as observed in the numerical experiments (see section 5).

In the previous discussion it was essential to assume that we had some information about ε. Notice that, even in case we do not have such information, the choice $\theta = 1 - 1/n$, which led to (4.2), is always available. This bound is almost spectral in n, depends weakly on Λ, and prevents error amplification.

5. Numerical illustrations. In this section we give four numerical illustrations. The first two concern elementary Laplace transforms which are assumed to be computed with a relative error of order $\rho \approx$ eps, where eps stands for the machine precision (eps $= 10^{-16}$ in our computations). In the last two illustrations we do not assume any information about the errors due to the computations of the Laplace transforms.

Illustration 1. We first show, by means of a simple example, that for $n >> 1$ (2.6) fails in the presence of errors in the evaluations. To this end, we consider the mapping $u(t) = e^{-t}$, whose Laplace transform is $U(z) = 1/(1+z)$.

This function satisfies (1.2) for all $\delta > 0$ and $M = 1/\sin\delta$. We fix $\theta = 0.5$, $\alpha = 0.7$, and $d = 0.6$, and choose the parameters h, λ as stated in the theorem for all the values of n. In Figure 5.1 we plot in a semilogarithmic scale the absolute actual error, i.e.,

$$\ln \max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$$

versus n (recall that $\bar{u}_n(t)$ stands for the actual computed approximation to $u(t)$; see (3.1)). This is done for $\Lambda = 5, 50$ and $t_0 = 1$. This figure shows that the error decays exponentially for the first values of n, saturates near ε level, and then grows like $\mathcal{O}(e^{cn})$.

Next, to study the behavior of the error and its estimate with respect to θ, we repeat the previous experiment for $\Lambda = 5$, with $\theta = 0.5$ and $\theta = 0.99$. The actual error and the corresponding theoretical estimate are depicted in Figure 5.2. We observe that the value of n where the error starts growing is well predicted by Theorem 2. Also, we see that enlarging θ results in a slower error propagation.

Finally, we tune parameters as explained in section 4. For $\Lambda = 5,\ 50$, in Figure 5.3 (left) we plot the optimal values of θ against n. In Figure 5.3 (right) we plot

$$\ln \max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$$

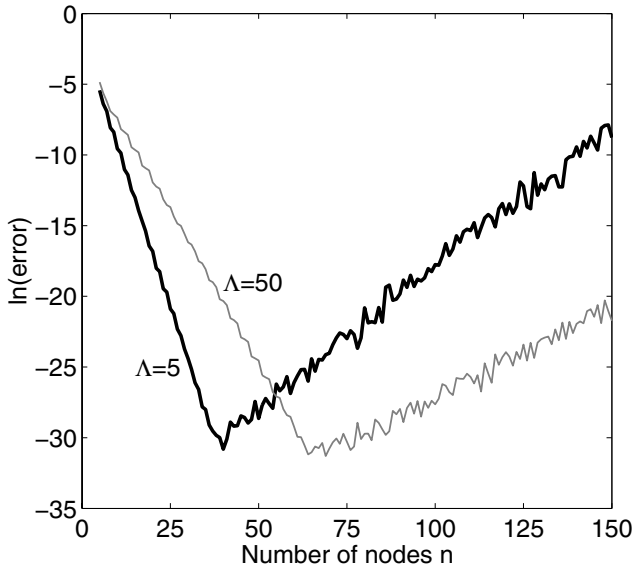


FIG. 5.1. $\ln \max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$ *versus* n *for* u *in Illustration* 1*, with* $\theta = 0.5$ *fixed,* $\alpha = 0.7$*,* $d = 0.6$*, and* $t_0 = 1$*. The gray line corresponds to* $\Lambda = 50$ *and the black to* $\Lambda = 5$*.*

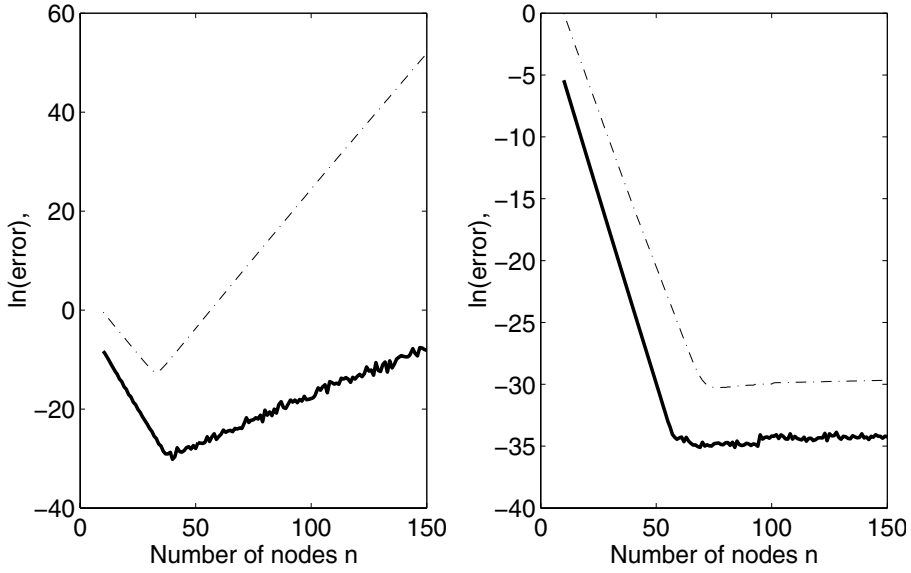


FIG. 5.2. *Observed error (continuous) and theoretical estimate (dashed) for* u *in Illustration* 1 *with fixed* θ *versus* n*, in semilogarithmic scale. Left: For* $\theta = 0.5$*. Right: For* $\theta = 0.99$*. In both cases,* $\Lambda = 5$*.*

(continuous line) and the logarithm of the corresponding values of the theoretical error estimate (dashed line) obtained in Theorem 2, versus n, once θ is optimal. We maintain $\alpha = 0.7$, $d = 0.6$, and $t_0 = 1$.

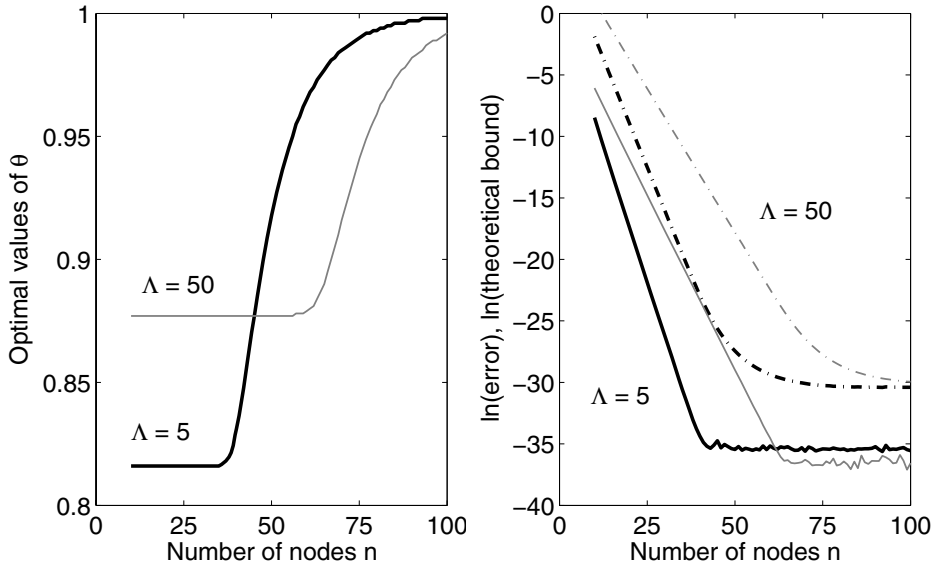


FIG. 5.3. *Left: Optimal θ versus n. Right: Natural logarithms of* $\max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$ *(continuous) and the theoretical estimate (dashed) versus n, for u in Illustration* 1. *The gray lines correspond to $\Lambda = 50$ and the black to $\Lambda = 5$.*

Illustration 2. Take $\beta = 1.5$ and set

$$U(z) = \frac{z^{\beta-1}}{z^\beta + 1};$$

i.e., $U(z)$ is the Laplace transform of

$$u(t) = M_\beta(-t^\beta),$$

where M_β stands for the Mittag–Leffler function of order β (see [15]). Notice that U satisfies (1.2) for any $\delta \in (\pi/3, \pi/2)$, with $\mu = 1$ and $M = 1/\sin(\beta(\pi - \delta))$. We consider here as an exact solution the one computed with 500 nodes and take $\alpha = \pi/12$, $d = 0.25$, and $t_0 = 1$.

This example was already considered in [13]. In order to compare the performance of the strategy proposed in [13] with the one proposed in the present paper, we first compute $\bar{u}_n(t)$ by selecting the parameters as in [13]. In Figure 5.4 (left) we plot in semilogarithmic scale the theoretical estimate and actual error for $\Lambda = 2, 5$, which are acceptable. In Figure 5.4 (right) we do the same for $\Lambda = 50$ and conclude that the approach in [13] is not at all useful for large values of Λ. However, the corresponding computation obtained by using the strategy in section 4 yields the plot in Figure 5.5, which shows a satisfactory spectral order of convergence even for $\Lambda = 50$.

Illustration 3. We consider the inhomogeneous heat equation on the unit square $\Omega = (0,1)^2$ with zero initial value and a convective heat flux at the boundary

$$\left\{ \begin{array}{lcl} u_t(t,x) & = & \Delta u(t,x) + f(x), \text{ for } x \in \Omega,\ t \geq 0, \\ \partial_\nu u(t,x) & = & -u(t,x), \text{ for } x \in \partial\Omega,\ t \geq 0, \\ u(0,x) & = & 0, \text{ for } x \in \Omega, \end{array} \right. \tag{5.1}$$

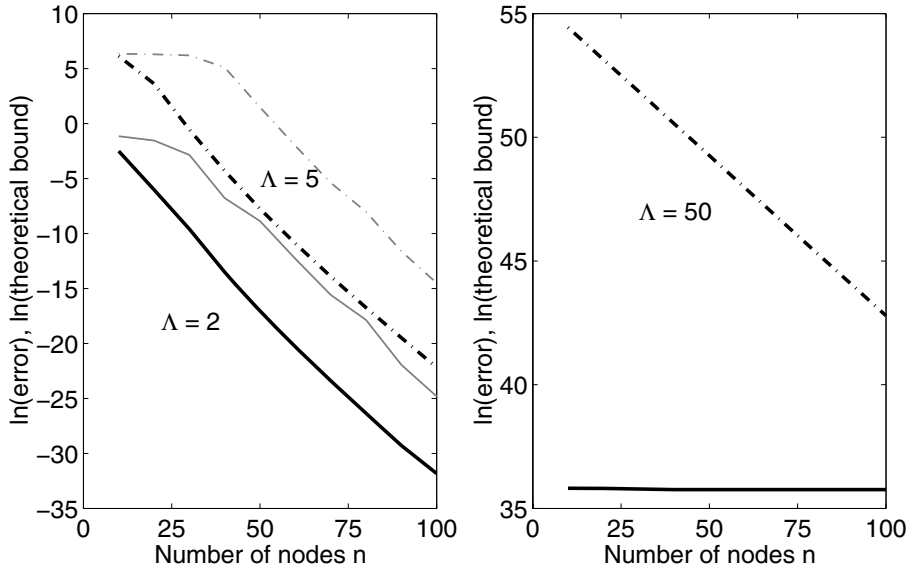


FIG. 5.4. *Natural logarithms of* $\max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$ *(continuous) and the theoretical estimate (dashed) versus* n*, for* u *in Illustration* 2 *proceeding as in* [13] *for* $\delta = \pi/3$*,* $t_0 = 1$*. The gray lines correspond to* $\Lambda = 50$ *and the black to* $\Lambda = 5$*.*

where f is the indicator function of the rectangle $R = [0.6, 0.8] \times [0.2, 0.8]$, i.e., $f = 1$ on R and $f = 0$ elsewhere.

Problem (5.1) is semidiscretized in space by using linear finite elements on a triangular grid. Denoting by $V_h \subset L^2(\Omega)$ the space of elements and by $U_h(z)$ the Laplace transform of the semidiscrete solution $u_h(t)$, we get

$$U_h(z) = \frac{1}{z}(z - \Delta_h)^{-1} P_h f,$$

with $\Delta_h : V_h \to V_h$ the discrete Laplacian and P_h the orthogonal projection of f onto V_h. Now, for fixed $h > 0$, we try to approximate $u_h(t)$ by inverting $U_h(z)$. Notice that, since Δ_h is definite negative, certainly $U_h(z)$ satisfies (1.2) for any $0 < \delta < \pi/2$ and $M = 1/\sin(\delta)$. Notice also that, working in coordinates relative to the standard basis of elements, $U_h(z)$ is represented by a vector valued mapping $\mathbf{U}_h(z)$ satisfying

$$zM_h\mathbf{U}_h(z) + S_h\mathbf{U}_h(z) = \frac{1}{z}\mathbf{f}_h,$$

where M_h and S_h stand for the mass and stiffness matrices and where $\mathbf{f}_h$ is the vector formed by the scalar products of f with the elements of the basis. Thus, one evaluation of $U(z_k)$ at a given node z_k requires the solution of one linear system of the above form.

In the experiment we generate a mesh, shown in the left of Figure 5.6, with 542 triangles by means of the mesh generator Triangle [19]. Linear systems are solved using MATLAB's sparse LU factorization UMFPACK. Since $u_h(t)$ is unknown, the errors are estimated in the $L^2(\Omega)$-norm with respect to a reference solution $\bar{u}_{h,500}(t)$ obtained with 500 nodes. In the absence of precise information about ρ, both for this reference solution and for the rest of the approximations $\bar{u}_{h,n}(t)$ to $u_h(t)$, we take

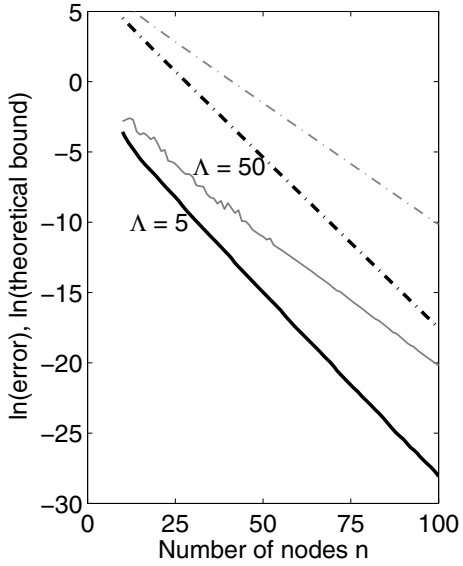


FIG. 5.5. *Natural logarithm of* $\max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$ *(continuous) and the theoretical estimate (dashed) versus* n*, for* u *in Illustration* 2*. The gray lines correspond to* $\Lambda = 50$ *and the black to* $\Lambda = 5$*.*

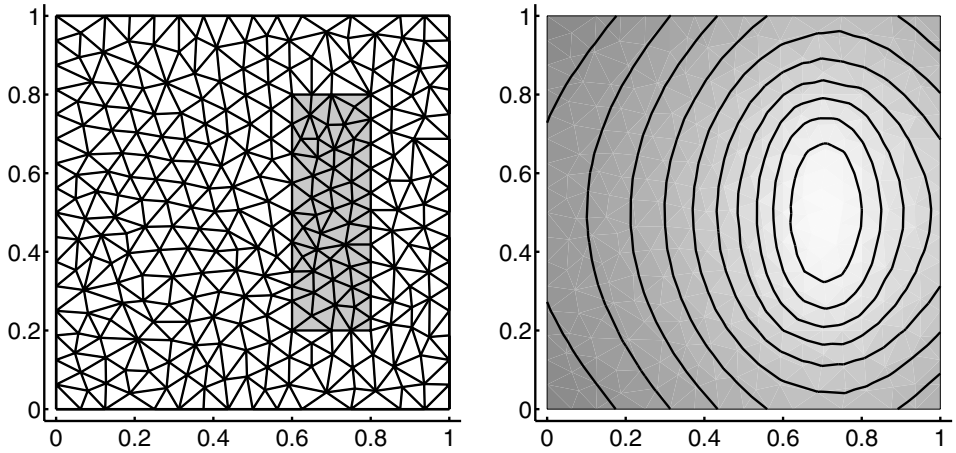


FIG. 5.6. *Left: Mesh of* Ω*, with the set* R *indicated in dark gray. Right: Temperature distribution at* $t = 0.5$ *in false-color representation (white corresponds to temperature* 1 *and black to* 0*).*

$\theta = 1 - 1/n$, as indicated in section 4. In Figure 5.7, for the parameters $\alpha = 0.7$, $d = 0.6$, $t_0 = 0.01$, and $\theta = 1 - 1/n$, we plot $\ln \max_{t\in[t_0,\Lambda t_0]} \|\bar{u}_{h,500}(t) - \bar{u}_{h,n}(t)\|$ against n for $\Lambda = 5$, 50. This plot shows the predicted behavior.

Illustration 4. We consider again the Laplace transform $U(z) = 1/(1+z)$ of the exponential function $u(t) = e^{-t}$ as in Illustration 1. The values of α, d, and t_0 are again 0.7, 0.6, and 1, respectively.

We add, on purpose, perturbations of maximum size 10^{-4} to the evaluations of

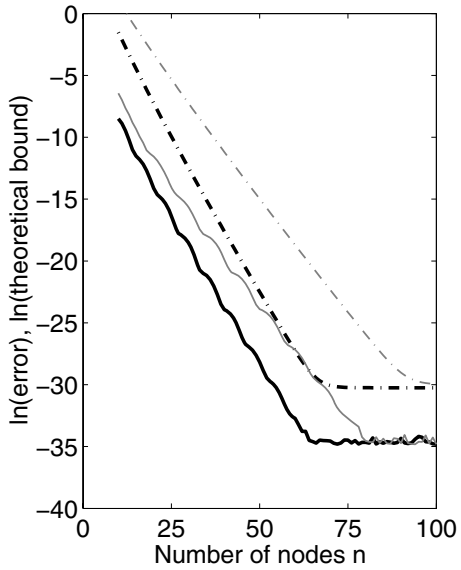


FIG. 5.7. *Left: Natural logarithm of* $\max_{t\in[t_0,\Lambda t_0]} \|u(t)-\bar{u}_n(t)\|$ *(continuous) and the theoretical estimate (dashed) versus* n*, for* u *in Illustration* 3*. The gray lines correspond to* $\Lambda = 50$ *and the black to* $\Lambda = 5$*.*

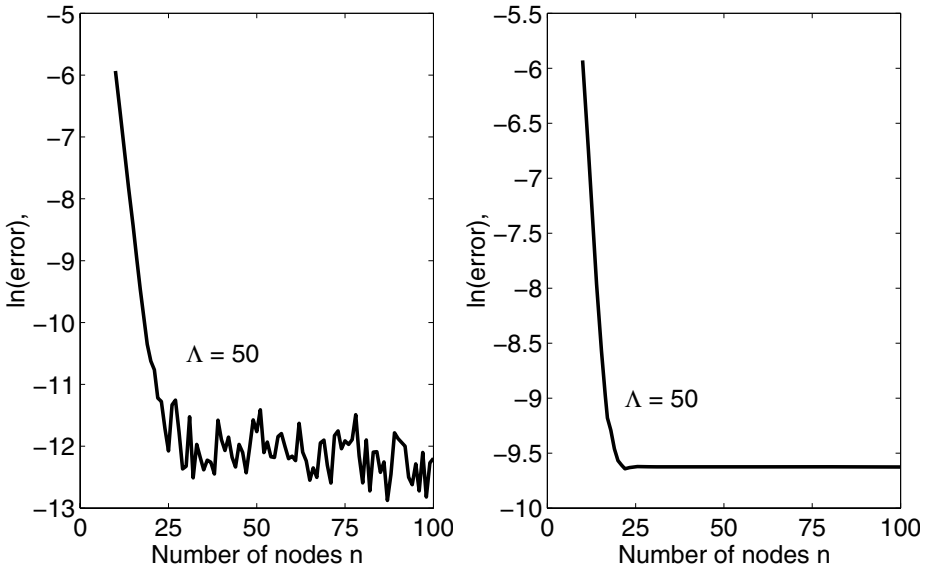


FIG. 5.8. $\ln \max_{t\in[t_0,\Lambda t_0]} \|u(t) - \bar{u}_n(t)\|$ *versus* n*, for* u *in Illustration* 4 *with* $\theta = 1 - 1/n$*,* $\alpha = 0.7$*,* $d = 0.6$*,* $t_0 = 1$*, and* $\Lambda = 50$*. Left: Random perturbation. Right: Worst case perturbation.*

U at the required nodes. Thus, we use (3.1) with

$$U_k = U(z_k) + \eta_k, \qquad -n \le k \le n,$$

with $|\eta_k| \le \rho = 10^{-4}$. Now we try to approximate $u(t)$ without using the available information about ρ. In this situation, as explained in section 4, we take $\theta = 1 - 1/n$.

In fact, we compare two types of perturbations.

We first generate complex, random, independent perturbations η_k in such a way that $|\eta_k|$ and $\arg(\eta_k)$ are uniformly distributed on $[0, 10^{-4}]$ and $[0, 2\pi]$, respectively. In Figure 5.8 (left), we show the resulting actual error, which behaves much better than predicted by (4.2). The explanation is that cancellations are likely compensating the effects of the independent random perturbations. A finer analysis of the observed behavior is beyond the scope of the present paper.

Second, for each $-n \le k \le n$, we consider the perturbation

$$\eta_k = 10^{-4} \exp(-i \arg(\omega_k(t_0))),$$

with $\omega_k(t_0)$ defined in (3.1). These perturbations correspond to the worst possible case in (3.2) for $t = t_0 = 1$. Now, the resulting actual error, plotted in Figure 5.8 (right), fits quite well with (4.2).

Acknowledgments. The authors wish to express their gratitude to Prof. T. Hohage for a suggestion concerning [13] which is used in section 3 of the present paper. The useful comments made by the referees during the revision process also helped to improve the quality of the work.

REFERENCES

[1] D. D. Ang, J. Lund, and F. Stenger, *Complex variable and regularization methods of inversion of the Laplace transform*, Math. Comp., 53 (1989), pp. 589–608.

[2] W. Arendt, C. J. K. Batty, M. Hieber, and F. Neubrander, *Vector-valued Laplace Transforms and Cauchy Problems*, Birkhäuser, Basel, 2001.

[3] A. Ashyralyev and P. Sobolevskii, *Well-Posedness of Parabolic Difference Equations*, Birkhäuser, Basel, 1994.

[4] N. Y. Bakaev, V. Thomée, and L. Wahlbin, *Maximum-norm estimates for resolvents of elliptic finite element operators*, Math. Comp., 72 (2002), pp. 1597–1610.

[5] C. Cunha and F. Viloche, *An iterative method for the numerical inversion of Laplace transforms*, Math. Comp., 64 (1995), pp. 1193–1198.

[6] C. Cunha and F. Viloche, *The Laguerre functions in the inversion of the Laplace transform*, Inverse Problems, 9 (1993), pp. 57–68.

[7] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *$\mathcal{H}$-matrix approximation for the operator exponential with applications*, Numer. Math., 92 (2002), pp. 83–111.

[8] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation to the operator-valued functions of elliptic operators*, Math. Comp., 73 (2004), pp. 1297–1324.

[9] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation to a class of operator-valued functions*, Math. Comp., 74 (2005), pp. 681–708.

[10] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent algorithms for the operator exponential with applications to inhomogeneous problems in Banach spaces*, SIAM J. Numer. Anal., 43 (2005), pp. 2144–2171.

[11] T. Hohage and F. J. Sayas, *Numerical solution of a heat diffusion problem by boundary element methods using the Laplace transform*, Numer. Math., 102 (2005), pp. 67–92.

[12] M. López-Fernández, Ch. Lubich, C. Palencia, and A. Schädle, *Fast Runge-Kutta approximation of inhomogeneous parabolic differential equations*, Numer. Math., 102 (2005), pp. 277–291.

[13] M. López-Fernández and C. Palencia, *On the numerical inversion of the Laplace transform of certain holomorphic mappings*, Appl. Numer. Math., 51 (2004), pp. 289–303.

[14] W. McLean and V. Thomée, *Time discretization of an evolution equation via Laplace transforms*, IMA J. Numer. Anal., 24 (2004), pp. 439–463.

[15] I. Polubny, *Fractional Differential Equations*, Math. Sci. Engrg. 198, Academic Press, San Diego, 1999.

[16] M. Rizzardi, *A modification of Talbot's method for the simultaneous approximation of several values of the inverse Laplace transform*, ACM Trans. Math. Software, 21 (1995), pp. 347–371.

[17] A. Schädle, M. López-Fernández, and C. Lubich, *Fast and oblivious convolution quadrature*, SIAM J. Sci. Comput., 28 (2006), pp. 421–438.

[18] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time discretization of parabolic equations based on Laplace transformation and quadrature*, Math. Comp., 69 (2000), pp. 177–195.

[19] J. R. Shewchuk, *Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator*, in Applied Computational Geometry: Towards Geometric Engineering, Lecture Notes in Comput. Sci. 1148, Springer-Verlag, New York, 1996, pp. 203–222.

[20] F. Stenger, *Approximations via Whittaker's cardinal function*, J. Approx. Theory, 17 (1976), pp. 222–240.

[21] F. Stenger, *Numerical methods based on Whittaker cardinal, or sinc functions*, SIAM Rev., 23 (1981), pp. 165–224.

[22] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1351–1374

PROJECTION METHODS AND CONDITION NUMBERS IN UNIFORM NORM FOR FREDHOLM AND CAUCHY SINGULAR INTEGRAL EQUATIONS*

M. C. DE BONIS† AND G. MASTROIANNI†

Dedicated to Professor Francesco Costabile on the occasion of his 60th birthday

Abstract. In this paper the authors propose a numerical method for the approximate solution of some classes of Fredholm and Cauchy integral equations including the "discrete collocation" and "collocation" methods.

Key words. Cauchy singular integral equation, Fredholm integral equations, projection method, Lagrange interpolation, Fourier sum

AMS subject classifications. 65R20, 45E05, 41A05, 41A10

DOI. 10.1137/050626934

1. Introduction. This paper deals with the numerical treatment of some classes of Fredholm and Cauchy integral equations. In the last few decades, several authors proposed numerical methods in order to obtain polynomial approximations of the solutions. Among them we mention [1, 4, 12, 14, 15, 16, 17, 18, 19, 32, 41, 45, 46, 47, 48], where the so-called *collocation* and *discrete collocation* methods are considered. Such procedures essentially consist of three steps: first, look for an approximate solution as a (weighted) polynomial; second, approximate the (Fredholm) integral by a quadrature formula; and third, project the equation onto a finite dimensional space of (weighted) polynomials by collocation. The stability and the convergence of the method are proved on a finite dimensional equation equivalent to the system. But, as a simple example shows, the uniform boundedness of the condition number of the discrete operator does not imply the well-conditioning of the system that is crucial in order to compute the approximate solution.

If the above-mentioned equations are considered in suitable L^p-spaces with $1 < p < +\infty$, then, using Marcinkievicz bases (see [20, 21]), one can associate a well-conditioned system of linear equations to a discrete and well-conditioned operator. The case $p = +\infty$ is still open and we are going to investigate it in this paper. To be more precise, we shall consider some classes of Fredholm and Cauchy singular integral equations in a subspace of continuous functions which are related to the kernels of the equations. Since in these spaces sequences of uniformly bounded polynomial projections do not exist, we shall use sequences of projections $\{\Gamma_m\}_m$ (essentially of Lagrange or Fourier type) such that $\|\Gamma_m\| \sim \log m$. The simple procedure showed in section 2 (see Remark 1), constructs uniquely solvable polynomial equations whose solutions converge as the best approximation (except for an extra log factor). Using a suitable basis, we derive a well-conditioned linear system equivalent to the finite dimensional equation, whose solution is the array of the coefficients of the polyno-

*Received by the editors March 16, 2005; accepted for publication (in revised form) December 21, 2005; published electronically July 7, 2006. This paper was sponsored by Istituto Nazionale di Alta Matematica, GNCS project "Trattamento numerico di equazioni integrali e connessi problemi di approssimazione e algebra lineare."

http://www.siam.org/journals/sinum/62693.html

†Dipartimento di Matematica, Università della Basilicata, Via dell'Ateneo Lucano 10, 85100 Potenza, Italy (mdebonis@unibas.it, mastroianni@unibas.it).

mial. The error estimates are sharp and cover the ones available in the literature (in the uniform norm). The proposed procedure includes the discrete collocation and collocation methods.

The case of nonsingular kernels is considered in subsection 2.3. In Theorem 2.2 we show that the approximation error does not change, while the linear system is strongly simplified.

Section 3 is devoted to the numerical treatment of the Cauchy singular integral equations with constant coefficients and compact perturbation, having index $\chi \in \{0, 1\}$. By regularization, we derive equivalent Fredholm equations having nonsingular kernels which are transforms of the original kernels by means of singular operators. Very recently it was shown in [9] that the problem is led back to the case of Fredholm equations with nonsingular kernel. In section 4 we give the proofs of the previous results.

2. Projection methods for Fredholm equations.

2.1. Function spaces. We are going to consider the integral equations in the space

$$C_v = \Big\{ f \in C^0((-1,1)) \ : \ \lim_{|x|\to 1} (fv)(x) = 0 \Big\},$$

where $C^0(A)$ is the collection of the continuous functions in $A \subset [-1,1]$ and $v(x) := v^{\gamma,\delta}(x) = (1-x)^\gamma (1+x)^\delta$ is a Jacobi weight. In the case $\gamma = 0$ (respectively, $\delta = 0$) C_v consists of all continuous functions on $(-1,1]$ (respectively, $[-1,1)$) such that

$$\lim_{x\to -1} (fv)(x) = 0 \quad \Big(\lim_{x\to 1} (fv)(x) = 0 \Big).$$

In the case $\gamma = \delta = 0$, we set $C_v = C^0([-1,1])$. The space C_v equipped with the norm

$$\|f\|_{C_v} := \max_{|x|\le 1} |(fv)(x)| =: \|fv\|$$

is complete. For brevity, we shall write $\|f\|_A := \max_{x\in A} |f(x)|, \ A \subseteq [-1,1]$.

In what follows we will also consider functions belonging to the Besov space $B_{r,q}(v)$. In order to define $B_{r,q}(v)$ we introduce the seminorms

$$\|f\|_{\gamma,\delta,r,q} = \begin{cases} \displaystyle\left(\int_0^1 \left[\frac{\Omega^k_\varphi(f,t)_{v^{\gamma,\delta}}}{t^r} \right]^q \frac{dt}{t} \right)^{\frac{1}{q}}, & 1 \le q < +\infty, \\ & k > r > 0, \\ \displaystyle\sup_{t>0} \frac{\Omega^k_\varphi(f,t)_{v^{\gamma,\delta}}}{t^r}, & q = +\infty, \end{cases} \tag{2.1}$$

where [10]

$$\Omega^k_\varphi(f,t)_{v^{\gamma,\delta}} := \sup_{0<h\le t} \|(\Delta^k_{h\varphi} f) v^{\gamma,\delta}\|_{I_{h,k}},$$

$k \in \mathbb{N}, I_{h,k} := [-1 + 4k^2h^2, 1 - 4k^2h^2], 0 < t < 1, \varphi(x) = \sqrt{1-x^2}$, and

$$\Delta^k_{h\varphi} f(x) := \sum_{i=0}^k (-1)^i \binom{k}{i} f\left(x + h\varphi(x) \left(\frac{k}{2} - i \right) \right).$$

We will set $\|f\|_{r,q} := \|f\|_{0,0,r,q}$.

Thus, the Besov spaces are [11]

$$B_{r,q}(v) = \{f \in C_v : \|f\|_{\gamma,\delta,r,q} < +\infty\}, \quad \gamma, \delta \geq 0, \quad r \in \mathbb{R}^+, \quad 1 \leq q \leq +\infty, \tag{2.2}$$

and they are equipped with the norm

$$\|f\|_{B_{r,q}(v)} = \|fv\| + \|f\|_{\gamma,\delta,r,q}. \tag{2.3}$$

As previously done we will set $B_{r,q} := B_{r,q}(v^{0,0})$. In the case $q = +\infty$, $B_{r,\infty}(v), r > 0$, are the well-known Zygmund spaces and we will set $Z_r(v) := B_{r,\infty}(v)$, $Z_r := Z_r(v^{0,0})$.

In the following $\mathcal{C}$ denotes a positive constant which may have different values in different formulas. We will write $\mathcal{C} \neq \mathcal{C}(a, b, \dots)$ to say that $\mathcal{C}$ is independent of the parameters $a, b, \dots$. If $A, B \geq 0$ are quantities depending on some parameters, we write $A \sim B$, if there exists a positive constant $\mathcal{C}$ independent of the parameters of A and B, such that

$$\frac{B}{\mathcal{C}} \leq A \leq \mathcal{C}B.$$

2.2. Projection methods. Now we consider the Fredholm integral equation of the second kind

$$f(y) + \lambda \int_{-1}^{1} h(x, y) f(x) v^{\alpha,\beta}(x) dx = g(y), \tag{2.4}$$

where $v^{\alpha,\beta}$ is a Jacobi weight, $\lambda \in \mathbb{R}$, and h, g are given functions. Letting

$$(Kf)(y) = \lambda \int_{-1}^{1} h(x, y) f(x) v^{\alpha,\beta}(x) dx,$$

we can rewrite (2.4) as

$$(I + K)f = g, \tag{2.5}$$

where I denotes the identity operator.

We will consider (2.5) in C_v with γ and δ according to

$$\begin{aligned} &\max\left\{0, \frac{\alpha}{2} + \frac{1}{4}\right\} \leq \gamma < \min\left\{\frac{\alpha}{2} + \frac{3}{4}, 1 + \alpha\right\}, \\ &\max\left\{0, \frac{\beta}{2} + \frac{1}{4}\right\} \leq \delta < \min\left\{\frac{\beta}{2} + \frac{3}{4}, 1 + \beta\right\}, \end{aligned} \tag{2.6}$$

and we state the following assumptions:

$$g \in Z_r(v), \tag{2.7}$$

$$\sup_{t>0} \frac{\Omega_\varphi^k(Kf, t)_v}{t^r} \leq \mathcal{C}\|fv\|, \quad k > r > 0, \quad f \in C_v. \tag{2.8}$$

Note that (2.8) can be true even if the kernel $h(x, y)$ is weakly singular. For example, if

$$h(x, y) = \frac{1}{|x - y|^\mu}, \quad 0 < \mu < 1,$$

then (2.8) is true with $r \leq 1-\mu$ (see [30, Lemma 4.1]). Obviously, if the kernel $h(x,y)$ belongs to the Zygmund space $Z_r(v)$ w.r.t. y, then (2.8) is automatically satisfied.

Under the assumptions (2.7)–(2.8), the solution f^* of (2.4) (if it exists) can be "well" approximated by polynomials. In order to show this we denote by $L_m^{\alpha,\beta}$ the Lagrange projection based on the zeros of the mth orthonormal Jacobi polynomial $p_m(v^{\alpha,\beta})$, i.e., with $F \in C_v$,

$$L_m^{\alpha,\beta}(F,x) = \sum_{i=1}^{m} l_i^{\alpha,\beta}(x) F(x_i), \quad l_i^{\alpha,\beta}(x) = \frac{p_m(v^{\alpha,\beta},x)}{p_m'(v^{\alpha,\beta},x_i)(x-x_i)},$$

$x_1 < x_2 < \cdots < x_m$, $x_i = x_{m,i}^{\alpha,\beta}$ being the zeros of $p_m(v^{\alpha,\beta})$.

By means of this projection we introduce the polynomial sequence $\{g_m\}_m$, with $g_m = L_m^{\alpha,\beta}(g)$, and the sequence of operators $\{K_m\}_m$, where $(K_m f)(y) = L_m^{\alpha,\beta}(Kf,y)$. Obviously, for every $f \in C_v$, we have $K_m f \in \mathbb{P}_{m-1}$, $\mathbb{P}_{m-1}$ being the set of all algebraic polynomials of degree at most $m-1$. So, we are going to solve the sequence of polynomial equations

$$(I+K_m)f_m = g_m, \quad m = 1,2,\ldots, \tag{2.9}$$

where $f_m \in \mathbb{P}_{m-1}$ is unknown. Denoting by $\lambda_k^{\alpha,\beta} = \lambda_k(v^{\alpha,\beta})$, $k = 1,\ldots,m$, the Christoffel numbers w.r.t. $v^{\alpha,\beta}$, the following theorem holds.

THEOREM 2.1. *Assuming that* $\operatorname{Ker}(I+K) = \{0\}$ *in* C_v, *we denote by* f^* *the unique solution of* (2.5) *for a given* g. *If* (2.6)–(2.8) *are satisfied, then, for* m *sufficiently large (say,* $m > m_0$*), the equation* $(I+K_m)f_m = g_m$ *has the unique solution* $f_m^* \in \mathbb{P}_{m-1}$ *satisfying the estimate*

$$\|(f^* - f_m^*)v\| \leq \mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v)}, \quad r \geq 1, \tag{2.10}$$

where $\mathcal{C} \neq \mathcal{C}(m, f^*)$.

If we expand f_m^* *in the basis* $\{\varphi_i\}_{i=1,\ldots,m}$, *with* $\varphi_i = \frac{l_i^{\alpha,\beta}}{v(x_i)}$, *i.e., we write*

$$f_m^*(y) = \sum_{i=1}^{m} a_i \varphi_i(y),$$

then the array $\mathbf{a} = (a_1,\ldots,a_m)$ *of the coefficients is the unique solution of the system of linear equations*

$$\sum_{k=1}^{m}\left[\delta_{i,k} + \lambda\ \lambda_k^{\alpha,\beta}\frac{v(x_i)}{v(x_k)}S_m^{\alpha,\beta}(h(\cdot,x_i),x_k)\right]a_k = (gv)(x_i), \quad i = 1,\ldots,m, \tag{2.11}$$

where $v(x) = v^{\gamma,\delta}(x)$ *and*

$$S_m^{\alpha,\beta}(F,x) = \sum_{\nu=0}^{m-1} c_\nu(y) p_\nu(v^{\alpha,\beta},x), \quad c_\nu(y) = \int_{-1}^{1} p_\nu(v^{\alpha,\beta},x)F(x)v^{\alpha,\beta}(x)dx,$$

is the Fourier sum of a function F.

Finally, denoting by A_m *the matrix of the system* (2.11) *and by* $\operatorname{cond}(A_m) = \|A_m\|\|A_m^{-1}\|$ *its condition number in uniform norm (the so-called row sum norm), if*

$$\sup_{|y|\leq 1}\int_{-1}^{1} v^{\alpha-\gamma,\beta-\delta}(x)|h(x,y)|\log\left(2 + v^{\alpha-\gamma,\beta-\delta}(x)|h(x,y)|\right)dx < +\infty, \tag{2.12}$$

we have

$$\sup_m \frac{\mathrm{cond}(A_m)}{\log m} < +\infty. \tag{2.13}$$

In conclusion, if the assumptions (2.7)–(2.8) and (2.12) are satisfied, then choose γ and δ according to (2.6), solve the system (2.11), which is well conditioned (except for an extra $\log m$ factor), and construct the approximate solution f_m^*. The degree of f_m^*, or, equivalently, the order of the linear system, is chosen according to the required error and using the estimate (2.10).

The following remark includes a short discussion on the assumptions.

Remark 1. The choice of the space $C_v, v = v^{\gamma,\delta}$, with γ, δ satisfying (2.6), is crucial. Indeed the norms in C_v of the projections $L_m^{\alpha,\beta}$ and $S_m^{\alpha,\beta}$ are the smallest (except for a constant) (see Lemmas 4.1 and 4.2). Moreover, since, by virtue of (2.6), we can always choose $\gamma, \delta > 0$, Theorem 2.1 covers some cases of kernels and known terms unbounded in ± 1.

The assumption (2.8) implies the compactness of the operator $K : C_v \to C_v$ (see, for example, [51, p. 93]). Finally, if the norms in (2.7)–(2.8) are replaced by the Besov norms, then Theorem 2.1 is still true (see its proof). We used the Zygmund norm only to simplify the notation in the proofs.

2.3. The case of nonsingular kernels. In the shown procedure the Fourier sum of $h(x, y)$ generates the main computational effort. On the other hand several procedures concerning the most frequently used kernels are available in the literature (among others we mention [42, 43, 34]). It appears necessary if the kernel is weakly singular. But, if the kernel is smooth (for example, it belongs to the Zygmund space $Z_r(v)$), then with $h_x(y) = h_y(x) = h(x, y)$, both the norms

$$\sup_{|y|\le 1} \|[S_m^{\alpha,\beta} h_y - h_y]v\| \quad \text{and} \quad \sup_{|y|\le 1} \|[L_m^{\alpha,\beta} h_y - h_y]v\|$$

are dominated by $\|h_y\|_{Z_r(v)} m^{-r} \log m$ and we can replace in the system (2.11) $S_m^{\alpha,\beta}(h(x_i, \cdot), x_k)$ by $L_m^{\alpha,\beta}(h(x_i, \cdot), x_k) = h(x_k, x_i)$. Obviously, the new system

$$\sum_{k=1}^{m} \left[\delta_{i,k} + \lambda\ \lambda_k^{\alpha,\beta} \frac{v(x_i)}{v(x_k)} h(x_k, x_i) \right] a_k = (gv)(x_i), \quad i = 1, \dots, m, \tag{2.14}$$

is much easier. Moreover, if A_m is the matrix of the system (2.14) and $\mathrm{cond}(A_m)$ is its condition number in uniform norm, we deduce the next theorem.

THEOREM 2.2. *Assume that in $C_{v^{\gamma,\delta}}$, with γ, δ satisfying* (2.6)*, we have* $\mathrm{Ker}(I + K) = \{0\}$*, and let f^* be the unique solution of* (2.5) *for a given g. If*

$$\|g\|_{Z_r(v)} < +\infty, \tag{2.15}$$

$$\sup_{|x|\le 1} \|h_x\|_{Z_r(v)} < +\infty, \tag{2.16}$$

$$\sup_{|y|\le 1} v(y) \|h_y\|_{Z_r} < +\infty, \tag{2.17}$$

then we get

$$\sup_m \frac{\mathrm{cond}(A_m)}{\log m} < +\infty. \tag{2.18}$$

Moreover, denoting by $(a_1, \dots, a_m)$ *the unique solution of* (2.14)*, the polynomial*

$$f_m^*(y) = \sum_{i=1}^{m} a_i \varphi_i(y), \quad \varphi_i = \frac{l_i^{\alpha,\beta}}{v(x_i)},$$

verifies the estimate

$$\|(f^* - f_m^*)v\| \le C \left(\frac{\log m}{m^r} \right), \quad r \ge 1, \tag{2.19}$$

where the constant C *is independent of* m *and* f^**.*

In the next section we will give an application of Theorem 2.2.

3. Cauchy singular integral equations with constant coefficients. We consider one class of Cauchy singular integral equations with constant coefficients a and b satisfying $a^2 + b^2 = 1$ and which can be defined with the help of only one parameter $\alpha \in (0, 1)$. This class of equations appears in several problems of applied sciences and a wide literature on this topic is available. In particular, we mention the fundamental books and papers [14, 15, 16, 17, 18, 26, 32, 35, 37, 41, 45] and the references therein.

In this section we will consider this class of equations in the space of continuous functions with uniform norm and, using the regularization method [32, 37, 41], we get Fredholm equations. This procedure seems to be more complicated than the direct methods [14, 15, 16, 17, 18, 45]. But the authors have recently proved in [9] precise results on the mapping properties of the singular operators that are the dominant part of the Cauchy singular integral equations. By virtue of such results and under suitable assumptions on the kernels and the known terms, the numerical treatment of such equations goes back to an application of Theorem 2.2.

3.1. Equations with index 0. Consider the equation

$$(\hat{A}f)(y) + \int_{-1}^{1} k(x, y) f(x) v^{\alpha,-\alpha}(x) dx = g(y), \tag{3.1}$$

where

$$(\hat{A}f)(y) = \cos \pi\alpha f(y) v^{\alpha,-\alpha}(y) - \frac{\sin \pi\alpha}{\pi} \int_{-1}^{1} \frac{f(x)}{x - y} v^{\alpha,-\alpha}(x) dx, \quad 0 < \alpha < 1. \tag{3.2}$$

Assume $g \in Z_r(v^{0,\alpha})$ and $k_x \in Z_r(v^{0,\alpha}), r \ge 1$, uniformly w.r.t. x. An equivalent Fredholm equation can be obtained multiplying (3.1) from the left by the operator A defined as

$$(Af)(y) = \cos \pi\alpha f(y) v^{-\alpha,\alpha}(y) + \frac{\sin \pi\alpha}{\pi} \int_{-1}^{1} \frac{f(x)}{x - y} v^{-\alpha,\alpha}(x) dx. \tag{3.3}$$

Since (see, e.g., [9, 30]) $A\hat{A}f = \hat{A}Af = f, f \in Z_r(v^{0,\alpha})$, under the assumptions on g and k_x we get

$$f(y) + \int_{-1}^{1} (Ak_x)(y) f(x) v^{\alpha,-\alpha}(x) dx = (Ag)(y). \tag{3.4}$$

The equations (3.1) and (3.4) are uniquely solvable if the respective homogeneous problems have only the trivial solutions. We consider (3.4) in $C_{v^{\alpha+\gamma,\delta}}$ with $(\alpha+\gamma)$ and δ satisfying (2.6) with $\beta=-\alpha$, i.e., with γ and δ such that

$$\max\left\{0,-\frac{\alpha}{2}+\frac{1}{4}\right\}\le\gamma,\quad \delta<\min\left\{-\frac{\alpha}{2}+\frac{3}{4},1-\alpha\right\}. \tag{3.5}$$

Moreover, setting

$$\Psi(x,y):=(Ak_x)(y),\quad G(y):=(Ag)(y),$$

we rewrite (3.4) as follows:

$$f(y)+\int_{-1}^{1}\Psi(x,y)f(x)v^{\alpha,-\alpha}(x)dx=G(y). \tag{3.6}$$

If we want to apply Theorem 2.2, it is sufficient that G and Ψ verify in $C_{v^{\alpha+\gamma,\delta}}$ the assumptions (2.15)–(2.17). Now, in [9] (see also [30]) the following equivalence has been proved:

$$\|g\|_{Z_r(v^{0,\alpha})}\sim\|Ag\|_{Z_r(v^{\alpha,0})}. \tag{3.7}$$

The last one is crucial to prove the following lemma.

LEMMA 3.1. *If $0<\alpha<1$ and $\gamma,\delta\ge 0$, then we have*

$$\|G\|_{Z_r(v^{\alpha+\gamma,\delta})}\le\mathcal{C}\|g\|_{Z_r(v^{0,\alpha})}, \tag{3.8}$$

$$\sup_{|x|\le1}\|\Psi_x\|_{Z_r(v^{\alpha+\gamma,\delta})}\le\mathcal{C}\sup_{|x|\le1}\|k_x\|_{Z_r(v^{0,\alpha})}, \tag{3.9}$$

$$\sup_{|y|\le1}v^{\alpha+\gamma,\delta}(y)\|\Psi_y\|_{Z_r}\le\mathcal{C}\left[\sup_{|y|\le1}v^{\gamma,\delta+\alpha}(y)\|k_y\|_{Z_r}+\sup_{|y|\le1}v^{\gamma,\delta+\alpha}(y)\left\|\frac{\partial}{\partial y}k_y\right\|_{Z_r}\right], \tag{3.10}$$

with $r>0$, $\Psi_x(y)=\Psi_y(x)=\Psi(x,y)$, and $\mathcal{C}\ne\mathcal{C}(G,\Psi,x,y)$.

As a consequence of Lemma 3.1, if the right-hand sides of (3.8)–(3.10) are finite, then the functions G and Ψ of (3.6) satisfy in $C_{v^{\alpha+\gamma,\delta}}$ the assumptions (2.15)–(2.17) of Theorem 2.2 and we can deduce the following proposition.

PROPOSITION 3.1. *Assume that the original equation* (3.1) *is uniquely solvable in $C_{v^{\alpha+\gamma,\delta}}$. If the kernel k and the known term g of* (3.1) *satisfy*

$$\|g\|_{Z_r(v^{0,\alpha})}<+\infty, \tag{3.11}$$

$$\sup_{|x|\le1}\|k_x\|_{Z_r(v^{0,\alpha})}<+\infty, \tag{3.12}$$

$$\sup_{|y|\le1}v^{\gamma,\delta+\alpha}(y)\|k_y\|_{Z_r}+\sup_{|y|\le1}v^{\gamma,\delta+\alpha}(y)\left\|\frac{\partial}{\partial y}k_y\right\|_{Z_r}<+\infty, \tag{3.13}$$

with γ,δ according to (3.5)*, then, for m sufficiently large (say, $m>m_0$), the polynomial*

$$f_m^*(y)=\sum_{i=1}^{m}a_i\varphi_i(y),\quad \varphi_i=\frac{l_i^{\alpha,-\alpha}}{v^{\alpha+\gamma,\delta}(x_i)},\quad p_m^{\alpha,-\alpha}(x_i)=0,$$

where $(a_1, \dots, a_m)$ *is the solution of the linear system*

(3.14)
$$\sum_{k=1}^{m} \left[\delta_{i,k} + \lambda_k^{\alpha,-\alpha} \frac{v^{\alpha+\gamma,\delta}(x_i)}{v^{\alpha+\gamma,\delta}(x_k)} \Psi(x_k, x_i) \right] a_k = G(x_i) v^{\alpha+\gamma,\delta}(x_i), \quad i = 1, 2, \dots, m,$$

converges to the exact solution f^* *and*

$$\| (f_m^* - f^*) v^{\alpha+\gamma,\delta} \| \leq \mathcal{C} \frac{\log m}{m^r} \| g \|_{Z_r(v^{0,\alpha})}, \tag{3.15}$$

where $\mathcal{C} \neq \mathcal{C}(m, f^*, g)$.

Of course, by virtue of Theorem 2.2, the matrix of the system (3.14) is well conditioned.

Now we want to give some numerical remarks on the computation of the quantities $\Psi(x_k, x_i)$ and $G(x_i), i, k = 1, \dots, m$, in the system (3.14). Looking at their expressions

$$\Psi(x_k, x_i) = \cos \pi\alpha k(x_k, x_i) v^{-\alpha,\alpha}(x_i) + \frac{\sin \pi\alpha}{\pi} \int_{-1}^{1} \frac{k(x_k, t)}{t - x_i} v^{-\alpha,\alpha}(t) dt \tag{3.16}$$

and

$$G(x_i) = \cos \pi\alpha g(x_i) v^{-\alpha,\alpha}(x_i) + \frac{\sin \pi\alpha}{\pi} \int_{-1}^{1} \frac{g(t)}{t - x_i} v^{-\alpha,\alpha}(t) dt, \tag{3.17}$$

we can see that the only difficulty (if the analytical expression is not available) consists in the computation of the Hilbert transforms. The last ones can be computed using one of the several methods available in literature which are based on Gaussian rules, on product rules [5, 6, 7, 8, 33], or on suitable transformation of the integrand [2, 13, 22, 36, 40, 50]. Here, for completeness, we propose substituting $G(x_i)$ and $\Psi(x_k, x_i), k, i = 1, \dots, m$, with

$$\Psi_m(x_k, x_i) = \frac{\sin \alpha\pi}{\pi} \sum_{j=1}^{m} \frac{k(x_k, t_j)}{t_j - x_i} \lambda_{m,j}^{-\alpha,\alpha} \tag{3.18}$$

and

$$G_m(x_i) = \frac{\sin \alpha\pi}{\pi} \sum_{j=1}^{m} \frac{g(t_j)}{t_j - x_i} \lambda_{m,j}^{-\alpha,\alpha}, \tag{3.19}$$

where t_j are the zeros of $p_m^{-\alpha,\alpha}$. Then, we solve the new system

$$\sum_{k=1}^{m} \left[\delta_{i,k} + \lambda_k^{\alpha,-\alpha} \frac{v^{\alpha+\gamma,\delta}(x_i)}{v^{\alpha+\gamma,\delta}(x_k)} \Psi_m(x_k, x_i) \right] \bar{a}_k = G_m(x_i) v^{\alpha+\gamma,\delta}(x_i), \quad i = 1, \dots, m, \tag{3.20}$$

and we construct the approximate solution $f_m^{**}(y) = \sum_{i=1}^{m} \bar{a}_i \varphi_i(y)$. Obviously, we have to compare the norm $\| (f^* - f_m^{**}) v^{\alpha+\gamma,\delta} \|$ with $\| (f^* - f_m^*) v^{\alpha+\gamma,\delta} \|$. To this end the following propositions hold.

PROPOSITION 3.2. *If we assume that $k(x,y)$ and g satisfy* (3.12) *and* (3.11)*, respectively, then we have*

$$\sup_{1\le i,k\le m} v^{\alpha,0}(x_i)|\Psi(x_k,x_i)-\Psi_m(x_k,x_i)| \le \mathcal{C}\frac{\log m}{m^r}\sup_{|x|\le 1}\|k_x\|_{Z_r(v^{0,\alpha})} \tag{3.21}$$

$$\sup_{1\le i\le m} v^{\alpha,0}(x_i)|G(x_i)-G_m(x_i)| \le \mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v^{0,\alpha})}, \tag{3.22}$$

where $\mathcal{C}\neq\mathcal{C}(m,k,g)$.

PROPOSITION 3.3. *If A_m and A_m^* denote the matrices of the systems* (3.14) *and* (3.20)*, respectively, then*

$$\lim_m \frac{\operatorname{cond}(A_m^*)}{\operatorname{cond}(A_m)} = 1 \tag{3.23}$$

and, moreover,

$$\|(f^*-f_m^{**})v^{\alpha+\gamma,\delta}\| \le \mathcal{C}\left(\frac{\log^2 m}{m^r}\right), \tag{3.24}$$

where the constant $\mathcal{C}$ is independent of m and f^*.

Therefore, the condition numbers of the systems (3.14) and (3.20) are comparable, and, if the system (3.20) replaces (3.14), then the estimate (3.15) is perturbed by a $\log m$ factor.

Finally, note that we could consider (3.1) in the space $C_{v^{\alpha,0}}$ as proposed in [30]. But such an approach implies considering two different linear systems in the cases $0<\alpha<\frac{1}{2}$ and $\alpha\ge\frac{1}{2}$. Anyway, the results on the condition numbers of the respective linear systems and the convergence of both methods are equivalent.

3.2. Equation with index 1. Concerning the equation

$$(Df)(y)+\int_{-1}^{1} k(x,y)f(x)v^{-\alpha,\alpha-1}(x)dx = g(y), \tag{3.25}$$

with

$$(Df)(y)=\cos\pi\alpha f(y)v^{-\alpha,\alpha-1}(y)+\frac{\sin\pi\alpha}{\pi}\int_{-1}^{1}\frac{f(x)}{x-y}v^{-\alpha,\alpha-1}(x)dx, \tag{3.26}$$

we assume $g\in Z_r(v^{\alpha,1-\alpha})$ and $k_x\in Z_r(v^{\alpha,1-\alpha})$, uniformly w.r.t. x. We multiply (3.25) from the left by the operator $\hat{D}$ with

$$(\hat{D}f)(y)=\cos\pi\alpha f(y)v^{\alpha,1-\alpha}(y)-\frac{\sin\pi\alpha}{\pi}\int_{-1}^{1}\frac{f(x)}{x-y}v^{\alpha,1-\alpha}(x)dx. \tag{3.27}$$

Since [9] $D\hat{D}f=f$ and

$$\hat{D}Df = f-\frac{\int_{-1}^{1}f(x)v^{-\alpha,\alpha-1}(x)dx}{\int_{-1}^{1}v^{-\alpha,\alpha-1}(x)dx},$$

with

$$\frac{\int_{-1}^{1}f(x)v^{-\alpha,\alpha-1}(x)dx}{\int_{-1}^{1}v^{-\alpha,\alpha-1}(x)dx}=A\in\mathbb{R}, \tag{3.28}$$

(3.25) becomes

$$f(y)+\int_{-1}^{1}(\hat{D}k_x)(y)f(x)v^{-\alpha,\alpha-1}(x)dx=(\hat{D}g)(y)+A. \tag{3.29}$$

Equation (3.25) cannot be uniquely solvable, since the index of the operator in the spaces under consideration is equal to 1. Consequently, one has to consider (3.25) together with the additional condition (3.28) for a given constant A. Then (3.25), (3.28) is equivalent to (3.29). Letting

$$\Gamma(x,y)=(\hat{D}k_x)(y) \quad \text{and} \quad G_1(y)=(\hat{D}g)(y),$$

and assuming $A=0$, (3.29) can be rewritten as

$$f(y)+\int_{-1}^{1}\Gamma(x,y)f(x)v^{-\alpha,\alpha-1}(x)dx=G_1(y). \tag{3.30}$$

We consider (3.30) in $C_{v^{\gamma,\delta}}$ and we choose γ,δ replacing α by $-\alpha$ and β by $\alpha-1$ in (2.6). Thus, we take

$$\begin{aligned}\max\left\{0,-\frac{\alpha}{2}+\frac{1}{4}\right\}\le\gamma<\min\left\{-\frac{\alpha}{2}+\frac{3}{4},1-\alpha\right\},\\ \max\left\{0,\frac{\alpha}{2}-\frac{1}{4}\right\}\le\delta<\min\left\{\frac{\alpha}{2}+\frac{1}{4},\alpha\right\}.\end{aligned} \tag{3.31}$$

Now the equivalence

$$\|g\|_{Z_r(v^{\alpha,1-\alpha})}\sim\|\hat{D}g\|_{Z_r} \tag{3.32}$$

was proved by the authors in [9, (3.19)]. Then, as in subsection 3.1, we can deduce the next proposition.

PROPOSITION 3.4. *Assume that* (3.30) *is uniquely solvable in* $C_{v^{\gamma,\delta}}$ *and* (3.31) *holds true. If*

$$\|g\|_{Z_r(v^{\alpha,1-\alpha})}<+\infty, \tag{3.33}$$

$$\sup_{|x|\le1}\|k_x\|_{Z_r(v^{\alpha,1-\alpha})}<+\infty, \tag{3.34}$$

$$\sup_{|y|\le1}v^{\alpha+\gamma,1-\alpha+\delta}(y)\|k_y\|_{Z_r}+\sup_{|y|\le1}v^{\alpha+\gamma,1-\alpha+\delta}(y)\left\|\frac{\partial}{\partial y}k_y\right\|_{Z_r}<+\infty, \tag{3.35}$$

then the polynomial

$$f_m^*(y)=\sum_{i=1}^{m}a_i\varphi_i(y),\quad \varphi_i=\frac{l_i^{-\alpha,\alpha-1}}{v^{\gamma,\delta}(x_i)},\quad p_m^{-\alpha,\alpha-1}(x_i)=0,$$

where $(a_1,\dots,a_m)$ *is the solution of the linear system*

$$\sum_{k=1}^{m}\left[\delta_{i,k}+\ \lambda_k^{-\alpha,\alpha-1}\frac{v^{\gamma,\delta}(x_i)}{v^{\gamma,\delta}(x_k)}\Gamma(x_k,x_i)\right]a_k=G_1(x_i)v^{\gamma,\delta}(x_i),\quad i=1,2,\dots,m, \tag{3.36}$$

converges to the unique solution f^* *of* (3.25) *with* $A = 0$ *and we have*

$$\|(f_m^* - f^*)v^{\gamma,\delta}\| \le \mathcal{C}\left(\frac{\log m}{m^r}\right), \tag{3.37}$$

where $\mathcal{C} \ne \mathcal{C}(m, f^*)$.

As in the previous case, the system (3.36) can be replaced by

$$\sum_{k=1}^{m}\left[\delta_{i,k} + \lambda_{m,k}^{-\alpha,\alpha-1}\frac{v^{\gamma,\delta}(x_i)}{v^{\gamma,\delta}(x_k)}\Gamma_{m-1}(x_k, x_i)\right]\bar{a}_k = G_{1,m-1}(x_i)v^{\gamma,\delta}(x_i), \quad i = 1, \dots, m, \tag{3.38}$$

where

$$\Gamma_{m-1}(x_k, x_i) = \frac{\sin\alpha\pi}{\pi}\sum_{j=1}^{m-1}\frac{k(x_k, t_j)}{x_i - t_j}\lambda_{m-1,j}^{\alpha,1-\alpha}, \tag{3.39}$$

$$G_{1,m-1}(x_i) = \frac{\sin\alpha\pi}{\pi}\sum_{j=1}^{m-1}\frac{g(t_j)}{x_i - t_j}\lambda_{m-1,j}^{\alpha,1-\alpha}, \tag{3.40}$$

and t_j are the zeros of the Jacobi polynomial $p_{m-1}^{\alpha,1-\alpha}$.

Propositions analogous to Propostions 3.2 and 3.3 hold true also in this case, but, for the sake of brevity, we omit the details.

4. Proofs. We need some notation and preliminary results.

We denote by

$$E_m(f)_{v^{\gamma,\delta}} = \inf_{P\in\mathbb{P}_m}\|(f - P)v^{\gamma,\delta}\|$$

the error of best approximation of a function f in $C_{v^{\gamma,\delta}}$. We set $E_m(f) := E_m(f)_{v^{0,0}}$.

For all functions $f \in C_{v^{\gamma,\delta}}$ we have [10, p. 94]

$$E_m(f)_{v^{\gamma,\delta}} \le \mathcal{C}\int_0^{\frac{1}{m}}\frac{\Omega_\varphi^k(f,t)_{v^{\gamma,\delta}}}{t}dt. \tag{4.1}$$

In particular, from (4.1) we deduce

$$E_m(f)_{v^{\gamma,\delta}} \le \frac{\mathcal{C}}{m^r}\|f\|_{Z_r(v^{\gamma,\delta})} \quad \forall f \in Z_r(v^{\gamma,\delta}), \tag{4.2}$$

$$E_m(f)_{v^{\gamma,\delta}} \le \frac{\mathcal{C}}{m^r}\|f\|_{B_{r,q}(v^{\gamma,\delta})} \quad \forall f \in B_{r,q}(v^{\gamma,\delta}). \tag{4.3}$$

The following lemmas will be useful in what follows.

LEMMA 4.1. *For* α, β, γ, *and* δ *satisfying* (2.6) *and for every* $f \in C_{v^{\gamma,\delta}}$, *we have*

$$\|(L_m^{\alpha,\beta}f)v^{\gamma,\delta}\| \le \mathcal{C}(\log m)\|fv^{\gamma,\delta}\| \tag{4.4}$$

or, equivalently,

$$\|(f - L_m^{\alpha,\beta}f)v^{\gamma,\delta}\| \le \mathcal{C}(\log m)E_{m-1}(f)_{v^{\gamma,\delta}}. \tag{4.5}$$

Moreover, for every function $f \in C^0([-1,1])$, we have

$$\int_{-1}^{1} |L_m^{\alpha,\beta}(f,x)| v^{\alpha-\gamma,\beta-\delta}(x)dx \le \mathcal{C}\|f\|, \tag{4.6}$$

$$\int_{-1}^{1} |f(x) - L_m^{\alpha,\beta}(f,x)| v^{\alpha-\gamma,\beta-\delta}(x)dx \le \mathcal{C}E_{m-1}(f). \tag{4.7}$$

Here the constant $\mathcal{C}$ is independent of m and f.

Proof. The bound (4.4) is Theorem 2.2 in [27], while (4.6) is a special case of Nevai's result in [38]. □

LEMMA 4.2. *For α, β, γ, and δ satisfying* (2.6) *and for every $g \in C_{v^{\gamma,\delta}}$, we have*

$$\|(S_m^{\alpha,\beta} g)v^{\gamma,\delta}\| \le \mathcal{C}(\log m)\|g v^{\gamma,\delta}\| \tag{4.8}$$

or, equivalently,

$$\|(g - S_m^{\alpha,\beta} g)v^{\gamma,\delta}\| \le \mathcal{C}(\log m)E_{m-1}(g)_{v^{\gamma,\delta}}. \tag{4.9}$$

Moreover, for every function g such that

$$A(g) := \int_{-1}^{1} v^{\alpha-\gamma,\beta-\delta}(x)|g(x)| \log\left(2 + v^{\alpha-\gamma,\beta-\delta}(x)|g(x)|\right)dx < +\infty, \tag{4.10}$$

we get

$$\int_{-1}^{1} |S_m^{\alpha,\beta}(g,x)| v^{\alpha-\gamma,\beta-\delta}(x)dx \le \mathcal{C}A(g). \tag{4.11}$$

Here the constant $\mathcal{C}$ is independent of m, f, and g.

Proof of Lemma 4.2. The bound (4.8) can be found in [23]. To prove (4.11), for the sake of simplicity of notation, we assume $\alpha = \beta$ and $\gamma = \delta$, and we set $v^l(x) = (1-x^2)^l$. Using the Pollard transformation (cf. [44]) we can write

$$\begin{aligned} S_m^{\alpha}(g,x) = {} & \alpha_m p_m(v^\alpha, x)c_m + \beta_m [p_m(v^\alpha, x)H(p_{m-1}(v^\alpha\varphi^2)\varphi^2 g v^\alpha, x) \\ & - p_{m-1}(v^\alpha, x)H(p_m(v^\alpha\varphi^2)\varphi^2 g v^\alpha, x)], \end{aligned}$$

where $\alpha_m \sim \beta_m \sim 1$, c_m is the mth Fourier coefficient, and H is the Hilbert transform. Denote by $\|F\|_1$ the L^1-norm of F. Since, by the Remez inequality (see [29] and [31]) we have

$$\|S_m^{\alpha}(g)v^{\alpha-\gamma}\|_1 \le \mathcal{C}\|S_m^{\alpha}(g)v^{\alpha-\gamma}\|_{L^1(I_m)},$$

where $I_m = \left[-1 + \frac{\mathcal{C}}{m^2}, 1 - \frac{\mathcal{C}}{m^2}\right]$, we get

$$\begin{aligned} \|S_m^{\alpha}(g)v^{\alpha-\gamma}\|_1 \le {} & \mathcal{C}\left(\|p_m(v^\alpha)v^{\alpha-\gamma}\|_{L^1(I_m)}|c_m| \right. \\ & + \|p_m(v^\alpha)v^{\alpha-\gamma}H(p_{m-1}(v^\alpha\varphi^2)\varphi^2 g v^\alpha)\|_{L^1(I_m)} \\ & \left. + \|p_{m-1}(v^\alpha\varphi^2)v^{\alpha-\gamma}\varphi^2 H(p_m(v^\alpha)v^\alpha g)\|_{L^1(I_m)}\right) \\ := {} & A_1 + A_2 + A_3. \end{aligned} \tag{4.12}$$

Recalling [49, (8.21.18), p. 198]

$$|p_m(v^\alpha; x)v^{\frac{\alpha}{2}+\frac{1}{4}}(x)| \le \mathcal{C} \ne \mathcal{C}(m), \quad x \in I_m, \tag{4.13}$$

for A_1 we have

$$\begin{aligned} A_1 &\le \mathcal{C}\|v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}\|_1 \int_{-1}^{1} |g(x)p_m(v^\alpha,x)v^\alpha(x)|dx \\ &\le \mathcal{C}\|v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}\|_1 \|gv^{\frac{\alpha}{2}-\frac{1}{4}}\|_1. \end{aligned}$$

Taking into account the assumption $\gamma < \frac{\alpha}{2}+\frac{3}{4}$, we have $\frac{\alpha}{2}-\frac{1}{4}-\gamma > -1$ and then $\|v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}\|_1 \le \mathcal{C}$. Moreover, from the assumption $\gamma > \frac{\alpha}{2}+\frac{1}{4}$ we deduce $\frac{\alpha}{2}-\frac{1}{4} > \alpha-\gamma$ and then $\|gv^{\frac{\alpha}{2}-\frac{1}{4}}\|_1 \le \|gv^{\alpha-\gamma}\|_1$. Therefore, by (4.10), we get

$$A_1 \le \mathcal{C}\|gv^{\alpha-\gamma}\|_1 \le \mathcal{C}A(g). \tag{4.14}$$

Moreover, recalling that [32]

$$\int f(x)H(g;x)dx = -\int g(x)H(f;x)dx, \quad f \in L^p, \quad g \in L^q, \quad \frac{1}{p}+\frac{1}{q}=1, \quad p>1,$$

and applying (4.13), we have

$$A_2 \le \mathcal{C}\int_{-1}^{1} v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}(t)|H(p_{m-1}(v^\alpha\varphi^2)\varphi^2 v^\alpha g;t)|dt.$$

Letting $G = \operatorname{sgn}(H(p_{m-1}(v^\alpha\varphi^2)\varphi^2 v^\alpha g))$, we get

$$\begin{aligned} A_2 &\le \mathcal{C}\int_{-1}^{1} |p_{m-1}(v^\alpha\varphi^2)\varphi^2(t)v^\alpha(t)g(t)||H(v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}G;t)|dt \\ &\le \mathcal{C}\int_{-1}^{1} \left(v^{\frac{\alpha}{2}+\frac{1}{4}}(t)|g(t)| \left| \int_{-1}^{1} v^{\frac{\alpha}{2}-\frac{1}{4}-\gamma}(x)\frac{G(x)}{x-t}dx \right| \right) dt \\ &:= \int_{-1}^{1} (v^{\frac{\alpha}{2}+\frac{1}{4}}(t)|g(t)|I_1(t))dt. \end{aligned}$$

It remains to estimate $I_1(t)$. Setting $\sigma := \frac{\alpha}{2}-\frac{1}{4}-\gamma$, in virtue of the assumption $\frac{\alpha}{2}+\frac{1}{4} \le \gamma < \frac{\alpha}{2}+\frac{3}{4}$, we have $-1 < \sigma < 0$ and

$$\begin{aligned} I_1(t) &= \left| v^\sigma(t)H(G;t) + v^\sigma(t)\int_{-1}^{1} \frac{v^\sigma(x)-v^\sigma(t)}{v^\sigma(t)(x-t)}G(x)dx \right| \\ &\le v^\sigma(t)|H(G;t)| + v^\sigma(t)\int_{-1}^{1} \left| \frac{v^{-\sigma}(t)-v^{-\sigma}(x)}{v^{-\sigma}(x)(x-t)} \right| dx \\ &\le v^\sigma(t)|H(G;t)| + v^\sigma(t)\int_{-1}^{1} |x-t|^{-1-\sigma}v^\sigma(x)dx \\ &\le v^\sigma(t)(|H(G;t)|+\mathcal{C}). \end{aligned}$$

Thus, since $G \in L^\infty$ and (4.10) holds, using a result in [39] we get

$$\begin{aligned} A_2 &\le \mathcal{C}\int_{-1}^{1} |g(t)|(1+|H(G;t)|)v^{\alpha-\gamma}(t)dt \\ &\le \mathcal{C}\int_{-1}^{1} v^{\alpha-\gamma}(t)|g(t)|\log(2+v^{\alpha-\gamma}(t)|g(t)|)dt = \mathcal{C}A(g). \end{aligned} \tag{4.15}$$

Analogously we can prove that

$$A_3 \le \mathcal{C} A(g). \tag{4.16}$$

Finally, combining (4.14)–(4.16) with (4.12), (4.11) follows. □

Now we can prove Theorem 2.1.

Proof of Theorem 2.1. Since $g_m = L_m^{\alpha,\beta}(g)$ and $g \in Z_r(v)$, using (4.5) and (4.2) we get

$$\|(g - g_m)v\| \le \mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v)}. \tag{4.17}$$

Moreover, since $K_m f = L_m^{\alpha,\beta}(Kf)$, by (4.5), (4.1), and (2.8) we obtain

$$\|(Kf - K_m f)v\| \le \mathcal{C}(\log m)E_{m-1}(Kf)_v \le \frac{\mathcal{C}}{m^r}(\log m)\|fv\|. \tag{4.18}$$

By a well-known result of linear algebra we have $\sup_m \|(I + K_m)^{-1}\| < +\infty$ and

$$\operatorname{cond}(I + K_m) = \operatorname{cond}(I + K) + \mathcal{O}\left(\frac{\log m}{m^r}\right). \tag{4.19}$$

Moreover, by the identity

$$f^* - f_m^* = (I + K_m)^{-1}[(g - g_m) + (K_m - K)(I + K)^{-1}g] \tag{4.20}$$

we get

$$\begin{aligned}\|(f^* - f_m^*)v\| &\le \mathcal{C}[\|(g - g_m)v\| + \|gv\|\|K - K_m\|_{C_v \to C_v}] \\ &\le \mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v)},\end{aligned} \tag{4.21}$$

i.e., (2.10).

In order to obtain the system (2.11), we consider $f_m + K_m f_m = g_m$ and we expand $K_m f$, g_m, and f_m in the basis $\{\varphi_i\}_{i=1,\dots,m}$, with $\varphi_i = \frac{l_{m,i}(v^{\alpha,\beta})}{v(x_i)}$. Since, for every $q \in \mathbb{P}_{m-1}$, we have

$$q(x) = \sum_{i=1}^m \varphi_i(x)\gamma_i, \quad \gamma_i = q(x_i)v^{\gamma,\delta}(x_i),$$

we can write

$$f_m(y) = \sum_{i=1}^m \varphi_i(y)a_i, \quad g_m(y) = \sum_{i=1}^m \varphi_i(y)b_i, \quad b_i = v^{\gamma,\delta}(x_i)g(x_i),$$

and

$$(K_m f_m)(y) = \sum_{i=1}^m \varphi_i(y)v^{\gamma,\delta}(x_i)(Kf_m)(x_i).$$

Moreover,

$$\begin{aligned}(Kf_m)(x_i) &= \lambda\sum_{k=1}^m \frac{a_k}{v(x_k)}\int_{-1}^1 h(x, x_i)l_k^{\alpha,\beta}(x)v^{\alpha,\beta}(x)dx \\ &= \lambda\sum_{k=1}^m \lambda_k^{\alpha,\beta}\frac{a_k}{v(x_k)}S_m^{\alpha,\beta}(h(\cdot, x_i), x_k)\end{aligned}$$

and then

$$(K_m f_m)(y) = \lambda \sum_{i=1}^{m} \varphi_i(y) v^{\gamma,\delta}(x_i) \sum_{k=1}^{m} \lambda_k^{\alpha,\beta} \frac{a_k}{v(x_k)} S_m^{\alpha,\beta}(h(\cdot, x_i), x_k),$$

$\lambda_k^{\alpha,\beta}, k = 1, \dots, m$, being the Christoffel numbers. Therefore, the finite dimensional equation

$$(I + K_m) f_m = g_m$$

is equivalent to

$$\sum_{i=1}^{m} \varphi_i(y) a_i + \lambda \sum_{i=1}^{m} \varphi_i(y) v^{\gamma,\delta}(x_i) \sum_{k=1}^{m} \frac{\lambda_k^{\alpha,\beta}}{v^{\gamma,\delta}(x_k)} S_m^{\alpha,\beta}(h(\cdot, x_i), x_k) a_k = \sum_{i=1}^{m} \varphi_i(y) b_i,$$

and then (2.11) follows.

Now we prove (2.13). By (2.11), using a Marcinkievicz inequality [28] and $\lambda_k^{\alpha,\beta} \sim v^{\alpha,\beta}(x_k) \frac{\sqrt{1-x_k^2}}{m} \sim v^{\alpha,\beta}(x_k) \Delta x_k$, $\Delta x_k = x_{k+1} - x_k$, we have

$$\begin{aligned}
\|A_m\| &\le 1 + |\lambda| \max_{1 \le i \le m} v(x_i) \sum_{k=1}^{m} \frac{\lambda_k^{\alpha,\beta}}{v(x_k)} |S_m^{\alpha,\beta}(h(\cdot, x_i), x_k)| \\
&\sim 1 + |\lambda| \max_{1 \le i \le m} \sum_{k=1}^{m} v^{\alpha-\gamma,\beta-\delta}(x_k) \Delta x_k |S_m^{\alpha,\beta}(h(\cdot, x_i), x_k)| v(x_i) \\
&\le 1 + \mathcal{C} \max_{1 \le i \le m} \int_{-1}^{1} |S_m^{\alpha,\beta}(v(x_i) h(\cdot, x_i), t)| v^{\alpha-\gamma,\beta-\delta}(t) dt.
\end{aligned}$$

By the assumption (2.12), by virtue of Lemma 4.2, we get

$$\|A_m\| \le \mathcal{C}. \tag{4.22}$$

It remains to estimate $\|A_m^{-1}\|$. By virtue of the equivalence of the system (2.11) with the equation $(I + K_m) f_m = g_m$, for every $\eta = (\eta_1, \dots, \eta_m)$ there exists a unique $\theta = (\theta_1, \dots, \theta_m)$ such that $\theta = A_m^{-1} \eta$ if and only if $\tilde{\theta}(y) = (I + K_m)^{-1} \tilde{\eta}(y)$, where

$$\tilde{\theta}(y) = \sum_{i=1}^{m} \varphi_i(y) \theta_i, \quad \theta_i = (\tilde{\theta} v)(x_i), \quad \text{and} \quad \tilde{\eta}(y) = \sum_{i=1}^{m} \varphi_i(y) \eta_i, \quad \eta_i = (\tilde{\eta} v)(x_i).$$

Then, for all η, by (4.19) it results that

$$\begin{aligned}
\|A_m^{-1} \eta\|_{l^\infty} &= \|\theta\|_{l^\infty} \le \|\tilde{\theta} v\| \\
&= \|(I + K_m)^{-1} \tilde{\eta} v\| \\
&\le \|(I + K_m)^{-1}_{|\mathbb{P}_{m-1}}\| \ \|\eta\|_{l^\infty} \ \|L_m^{\alpha,\beta}\|_{C_v \to C_v} \\
&\le \mathcal{C} \|(I + K)^{-1}\| \ \|\eta\|_{l^\infty} \|L_m^{\alpha,\beta}\|_{C_v \to C_v}.
\end{aligned}$$

Using (4.4) we get

$$\|A_m^{-1}\| \le \mathcal{C} \log m. \tag{4.23}$$

Since $\operatorname{cond}(A_m) = \|A_m\| \|A_m^{-1}\|$, by (4.22) and (4.23), (2.13) follows. ∎

Proof of Theorem 2.2. We first look for a finite dimensional equation equivalent to the system (2.14). To this end, we introduce the sequence of operators $\{K_m\}_m$ defined as

$$(K_m f)(y) = L_m^{\alpha,\beta}(K^* f, y),$$

with

$$(K^* f)(y) = (K_m^* f)(y) = \lambda \int_{-1}^{1} L_m^{\alpha,\beta}(h_y, x) f(x) v^{\alpha,\beta}(x) dx,$$

and the polynomial sequence

$$g_m(y) = L_m^{\alpha,\beta}(g, y).$$

Thus, the equation we are looking for is

$$(I + K_m) f_m = g_m, \quad m = 1, 2, \ldots, \tag{4.24}$$

where f_m is the unknown polynomial of degree at most $m-1$. Expanding $K_m f_m$, g_m, and f_m in the basis $\{\varphi_i\}_{i=1,\ldots,m}, \varphi_i(x) = \frac{l_i^{\alpha,\beta}(x)}{v^{\gamma,\delta}(x_i)}$, we get the system (2.14).

Now we prove (2.18). By (2.14) we have

$$\begin{aligned}
\|A_m\| &\le 1 + |\lambda| \max_{1\le i\le m} v(x_i) \sum_{k=1}^{m} \frac{\lambda_k^{\alpha,\beta}}{v(x_k)} |h(x_k, x_i)| \\
&\sim 1 + |\lambda| \max_{1\le i\le m} \sum_{k=1}^{m} v^{\alpha-\gamma,\beta-\delta}(x_k) \Delta x_k |h(x_k, x_i) v(x_i)| \\
&\le 1 + \mathcal{C} \left(\max_{-1\le x,y\le 1} |h(x,y)| v^{\gamma,\delta}(y) \right) \int_{-1}^{1} v^{\alpha-\gamma,\beta-\delta}(x) dx.
\end{aligned}$$

By the assumptions (2.6) and (2.16), we deduce

$$\|A_m\| \le \mathcal{C}. \tag{4.25}$$

By (4.25) and (4.23) we obtain (2.18).

To prove (2.19) we will use the inequality (4.21). Taking into account (4.17), we need to estimate only $\|K - K_m\|_{C_v \to C_v}$.

Adding and subtracting $K^* f$, we have

$$\begin{aligned}
\|(Kf - K_m f)v\| &= \|(Kf - K^* f)v\| + \|(K^* f - K_m f)v\| \\
&:= A + B.
\end{aligned} \tag{4.26}$$

We first estimate A. Using (4.7), we get

$$\begin{aligned}
&|(Kf)(y) - (K^* f)(y)| v(y) \\
&\quad = v(y) \left| \lambda \int_{-1}^{1} [h_y(x) - L_m^{\alpha,\beta}(h_y, x)] v^{\alpha-\gamma,\beta-\delta}(x) (fv)(x) dx \right| \\
&\quad \le \mathcal{C} \|fv\| v(y) \int_{-1}^{1} \left| h_y(x) - L_m^{\alpha,\beta}(h_y, x) \right| v^{\alpha-\gamma,\beta-\delta}(x) dx \\
&\quad \le \mathcal{C} \|fv\| v(y) E_{m-1}(h_y).
\end{aligned}$$

By the assumption (2.17) and (4.2) we have

$$A \le \frac{\mathcal{C}}{m^r}\|fv\|. \tag{4.27}$$

Concerning B, under the assumptions (2.6), by (4.5), we obtain

$$B \le \mathcal{C}(\log m)E_{m-1}(K^*f)_v. \tag{4.28}$$

In order to estimate $E_{m-1}(K^*f)_v$ by means of the inequality (4.1), we proceed to the evaluation of $\Omega^k_\varphi(K^*f,t)_v$. Using (4.6), we get

$$\begin{aligned}|v(y)\Delta^k_{h\varphi}(K^*f)(y)| &= \left|\lambda\int_{-1}^{1} L_m^{\alpha,\beta}(v(y)\Delta^k_{h\varphi}h_y,x)v^{\alpha-\gamma,\beta-\delta}(x)(fv)(x)dx\right| \\ &\le \mathcal{C}\|fv\|\int_{-1}^{1}|L_m^{\alpha,\beta}(v(y)\Delta^k_{h\varphi}h_y,x)|v^{\alpha-\gamma,\beta-\delta}(x)dx \\ &\le \mathcal{C}\|fv\|v(y)\sup_{|x|\le 1}|\Delta^k_{h\varphi}h_x(y)|.\end{aligned}$$

Taking the supremum on $y\in[-1+4k^2h^2,1-4k^2h^2]$ first and then the supremum on $0<h\le t$, we get

$$\Omega^k_\varphi(K^*f,t)_v \le \mathcal{C}\|fv\|\sup_{|x|\le 1}\Omega^k_\varphi(h_x,t)_v \le \mathcal{C}t^r\|fv\|\sup_{|x|\le 1}\|h_x\|_{Z_r(v)}.$$

Thus, using inequality (4.1) and the assumption (2.16), (4.28) becomes

$$B \le \frac{\mathcal{C}}{m^r}(\log m)\|fv\|. \tag{4.29}$$

Combining (4.27) and (4.29) with (4.26), we get

$$\|K-K_m\|_{C_v\to C_v} \le \frac{\mathcal{C}}{m^r}(\log m). \tag{4.30}$$

Finally, substituting (4.17) and (4.30) into (4.21), we deduce (2.19). □

Proofs of section 3. We first give some notation and preliminary results.

We define in $C_{v^{\gamma,\delta}}$ the rth φ-modulus of continuity as

$$\begin{aligned}\omega^k_\varphi(f,t)_{v^{\gamma,\delta}} = \Omega^k_\varphi(f,t)_{v^{\gamma,\delta}} &+ \inf_{P\in\mathbb{P}_{k-1}}\|(f-P)v^{\gamma,\delta}\|_{C[-1,-1+4k^2t^2]} \\ &+ \inf_{P\in\mathbb{P}_{k-1}}\|(f-P)v^{\gamma,\delta}\|_{C[1-4k^2t^2,1]},\end{aligned}$$

where $0<t\le\frac{1}{2k}$. Also we use the notation $\omega_\varphi=\omega^1_\varphi$. Note that, if $f\in B_{r,q}(v^{\gamma,\delta}), 1\le q\le+\infty, r\in\mathbb{R}^+$, then $\omega^k_\varphi(f,t)_{v^{\gamma,\delta}}\sim\Omega^k_\varphi(f,t)_{v^{\gamma,\delta}}$. The following proposition holds.

PROPOSITION 4.3. *Let* $0<\alpha<1$. *Then, for* $f\in Z_r(v^{\gamma,\delta+\alpha})$ *and* γ,δ *satisfying* (3.5), *we have*

$$|v^{\alpha+\gamma,\delta}(y)(Af)(y)| \le \mathcal{C}\left[\|fv^{\gamma,\delta+\alpha}\| + \int_0^{1/2}\frac{\omega_\varphi(f,t)_{v^{\gamma,\delta+\alpha}}}{t}dt\right], \tag{4.31}$$

where $|y|\le 1$ *and* $\mathcal{C}\ne\mathcal{C}(f,y)$.

Proof. It is not hard to deduce (4.31) by [24, Theorem 3.1] (see also [3, pp. 46–47] and [30, Proof of Theorem 2.2]). ∎

PROPOSITION 4.4. *Let* $0 < \alpha < 1$. *The inequality*

$$\inf_{P\in\mathbb{P}_m} \|v^{\alpha,0}A(f-P_m)\| \leq \mathcal{C}\left[E_m(f)_{v^{0,\alpha}}\log m + \int_0^{\frac{1}{m}} \frac{\omega_\varphi^k(f,t)_{v^{0,\alpha}}}{t}dt\right] \quad \forall f \in C_{v^{0,\alpha}} \tag{4.32}$$

holds, where $1 \leq k < m$ *and* $\mathcal{C} \neq \mathcal{C}(m,f)$.

Proof. The proof of (4.32) can be found in [30, Proof of Theorem 2.2]. ∎

Proof of Lemma 3.1. Recalling (3.7), we obtain

$$\|G\|_{Z_r(v^{\alpha+\gamma,\delta})} = \|Ag\|_{Z_r(v^{\alpha+\gamma,\delta})} \leq \|Ag\|_{Z_r(v^{\alpha,0})} \leq \mathcal{C}\|g\|_{Z_r(v^{0,\alpha})}, \tag{4.33}$$

i.e., (3.8). Analogously it is possible to prove (3.9).

It remains to prove (3.10). We have

$$v^{\alpha+\gamma,\delta}(y)\|\Psi_y\|_{Z_r} = v^{\alpha+\gamma,\delta}(y)\|\Psi_y\| + v^{\alpha+\gamma,\delta}(y)\sup_{t>0}\frac{\Omega_\varphi^k(\Psi_y,t)}{t^r} := B_1 + B_2. \tag{4.34}$$

Concerning B_1, applying (4.31) we get

$$\begin{aligned} v^{\alpha+\gamma,\delta}(y)|\Psi(x,y)| &= v^{\alpha+\gamma,\delta}(y)|(Ak_x)(y)| \\ &\leq v^{\alpha+\gamma,\delta}(y)\left|k(x,y)v^{-\alpha,\alpha}(y) + \int_{-1}^{1}\frac{k(x,z)}{z-y}v^{-\alpha,\alpha}(z)dz\right| \\ &\leq \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)|k(x,y)| + \mathcal{C}\int_0^1 \frac{\omega_\varphi(k_x,t)_{v^{\gamma,\delta+\alpha}}}{t}dt. \end{aligned} \tag{4.35}$$

Recalling that

$$\omega_\varphi(k_x,t)_{v^{\gamma,\delta+\alpha}} \leq \mathcal{C}t\sup_{|y|\leq 1}\left|\varphi(y)v^{\gamma,\delta+\alpha}(y)\frac{\partial}{\partial y}k(x,y)\right|,$$

we obtain

$$B_1 \leq \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)\sup_{|x|\leq 1}|k(x,y)| + \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)\sup_{|x|\leq 1}\left|\frac{\partial}{\partial y}k(x,y)\right|. \tag{4.36}$$

Analogously, replacing k_x by $\Delta^k_{h\varphi(x)}k_x$ in (4.35), we deduce

$$\begin{aligned} v^{\alpha+\gamma,\delta}(y)|\Delta^k_{h\varphi(x)}\Psi(x,y)| &= v^{\alpha+\gamma,\delta}(y)|A(\Delta^k_{h\varphi(x)}k_x)(y)| \\ &\leq \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)|\Delta^k_{h\varphi(x)}k(x,y)| \\ &\quad + \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)\left|\Delta^k_{h\varphi(x)}\frac{\partial}{\partial y}k(x,y)\right| \end{aligned}$$

and then, taking the supremum on $x \in [-1+4k^2h^2, 1-4k^2h^2]$ as first and the supremum on $0 < h \leq t$ as second, we get

$$B_2 \leq \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)\sup_{t>0}\frac{\Omega_\varphi^k(k_y,t)}{t^r} + \mathcal{C}\sup_{|y|\leq 1} v^{\gamma,\delta+\alpha}(y)\sup_{t>0}\frac{\Omega_\varphi^k\left(\frac{\partial}{\partial y}k_y,t\right)}{t^r}. \tag{4.37}$$

Finally, combining (4.36) and (4.37) with (4.34), (3.10) follows. ∎

Proof of Proposition 3.2. We first note that (3.18) and (3.19) can be obtained by replacing g and k_x with their Lagrange polynomials based on the zeros $t_1, \dots, t_m$ of $p_m^{-\alpha,\alpha}$ and taking into account that by [25]

$$A\left[\frac{p_m^{-\alpha,\alpha}}{\cdot - t_j}\right](y) = \frac{p_m^{\alpha,-\alpha}(y) - p_m^{\alpha,-\alpha}(t_j)}{y - t_j}$$

and

$$\frac{p_m^{\alpha,-\alpha}(t_j)}{[p_m^{-\alpha,\alpha}]'(t_j)} = \frac{\sin\alpha\pi}{\pi}\lambda_{m,j}^{-\alpha,\alpha},$$

we have

$$(Al_{m,j})(x_i) = \frac{\sin\alpha\pi}{\pi}\frac{\lambda_{m,j}^{-\alpha,\alpha}}{t_j - x_i}.$$

Moreover, letting $x_i = x_{m,i}^{\alpha,-\alpha} = \cos\theta_{m,i}^{\alpha,-\alpha}$ and $t_j = t_{m,j}^{-\alpha,\alpha} = \cos\theta_{m,j}^{-\alpha,\alpha}$, from a result in [25] we have

$$\min_{i,j} |\theta_{m,i}^{\alpha,-\alpha} - \theta_{m,j}^{-\alpha,\alpha}| \geq \frac{\mathcal{C}}{m}. \tag{4.38}$$

Now we prove (3.22). Recalling that $G = Ag$ and $G_m = AL_m^{-\alpha,\alpha}(g)$ and denoting by P_{m-1} the polynomial of best approximation of g, we get

$$\begin{aligned} |G(y) - G_m(y)| &= |A[g - L_m^{-\alpha,\alpha}(g)](y)| \\ &\leq |A(g - P_{m-1})(y)| + |AL_m^{-\alpha,\alpha}(g - P_{m-1})](y)|. \end{aligned}$$

Therefore

$$\begin{aligned} v^{\alpha,0}(x_i)|G(x_i) - G_m(x_i)| \leq\ & v^{\alpha,0}(x_i)|A(g - P_{m-1})(x_i)| \\ & + \frac{1}{\pi}v^{\alpha,0}(x_i)\left|\sum_{j=1}^{m}\frac{g(t_j) - P_{m-1}(t_j)}{t_j - x_i}\lambda_{m,j}^{-\alpha,\alpha}\right| \\ :=\ & B_1 + B_2. \end{aligned} \tag{4.39}$$

Using (4.32) and (4.2) we get

$$B_1 \leq \mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v^{0,\alpha})}. \tag{4.40}$$

Recalling that [49] $\lambda_{m,j}^{-\alpha,\alpha} \sim \Delta t_j\ v^{-\alpha,\alpha}(t_j), \Delta t_j = t_{j+1} - t_j$, we obtain

$$B_2 \leq \mathcal{C}\|(g - P_{m-1})v^{0,\alpha}\|v^{\alpha,0}(x_i)\sum_{j=1}^{m}\frac{\Delta t_j}{|x_i - t_j|}v^{-\alpha,0}(t_j).$$

Moreover, since by virtue of (4.38) we have (see, for instance, [3, (5.16)])

$$\sum_{j=1}^{m}\frac{\Delta t_j}{|x_i - t_j|}v^{-\alpha,0}(t_j) \leq Cv^{-\alpha,0}(x_i)\log m,$$

we get

$$B_2 \leq \mathcal{C}(\log m) E_{m-1}(g)_{v^{0,\alpha}}$$

and, applying (4.2), we obtain

$$B_2 \leq \mathcal{C}\frac{\log m}{m^s}\|g\|_{Z_s(v^{0,\alpha})}. \tag{4.41}$$

Combining (4.39) with (4.40) and (4.41), (3.22) follows. Analogously we can prove (3.21).

Notice that from the previous discrete error estimate one can deduce an estimate of the (global) operator norm; see, e.g., the proof of Proposition 3.3. □

Proof of Proposition 3.3. Denote by $A_m a = b$ and $A_m^* \bar{a} = b^*$ the systems (3.14) and (3.20), respectively. We first show that $\mathrm{cond}(A_m) \sim \mathrm{cond}(A_m^*)$. Since, for $m > m_0$, A_m^{-1} exists, the identity

$$A_m^* = A_m[I_m + A_m^{-1}(A_m^* - A_m)]$$

holds true. Moreover, $\|A_m^{-1}\| \leq \mathcal{C}\log m$ (see the proof of Theorem 2.1) and by Proposition 3.2 we deduce

$$\|A_m - A_m^*\| \leq \mathcal{C}\frac{\log m}{m^r}. \tag{4.42}$$

Therefore, $\lim_m \|A_m^{-1}(A_m^* - A_m)\| = 0$. Consequently $(A_m^*)^{-1}$ exists and

$$\lim_m \frac{\mathrm{cond}(A_m^*)}{\mathrm{cond}(A_m)} \leq 1.$$

On the other hand, we use the identity

$$A_m = A_m^*[I_m + (A_m^*)^{-1}(A_m - A_m^*)]$$

to prove that

$$\lim_m \frac{\mathrm{cond}(A_m)}{\mathrm{cond}(A_m^*)} \leq 1$$

and, consequently,

$$\lim_m \frac{\mathrm{cond}(A_m^*)}{\mathrm{cond}(A_m)} = 1.$$

In order to estimate $f^* - f_m^{**}$ in $C_{v^{\alpha+\gamma,\delta}}$, we premise some notation. Denoting by B_z the operator B acting w.r.t. the variable z, we set

$$(Kf)(y) := \int_{-1}^{1} (Ak_x)(y) f(x) v^{\alpha,-\alpha} dx,$$

$$(K_m f)(y) := (L_m^{\alpha,-\alpha}\tilde{K}f)(y)$$

with

$$(\tilde{K}f)(y) := \int_{-1}^{1} (L_{m,x}^{\alpha,-\alpha} A_y k)(x,y) f(x) v^{\alpha,-\alpha}(x) dx$$

and

$$(K_m^* f)(y) := (L_m^{\alpha,-\alpha} K^* f)(y)$$

with

$$(K^* f)(y) := \int_{-1}^{1} (L_{m,x}^{\alpha,-\alpha} A_y L_{m,y}^{-\alpha,\alpha} k)(x,y) f(x) v^{\alpha,-\alpha}(x) dx.$$

It is not hard to prove that the finite dimensional equations

$$f_m + K_m f_m = L_m A g$$

and

$$f_m + K_m^* f_m = L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g$$

are equivalent to the systems (3.14) and (3.20), respectively, in the basis $\{\varphi_i\}_i$, with $\varphi_i = \frac{l_i^{\alpha,-\alpha}}{v^{\alpha+\gamma,\delta}(x_i)} (x_i = x_i^{\alpha,-\alpha})$.

Therefore, to estimate $\|(f^* - f_m^{**}) v^{\alpha+\gamma,\delta}\|$ by means of (4.20) it remains to prove that $K_m^* \to K$ and $L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g \to Ag$ in $C_{v^{\alpha+\gamma,\delta}}$.

It is sufficient to consider the difference $K_m - K$ in the set of the polynomial of degree at most $m-1$. Thus, for all $f_m \in \mathbb{P}_m$, we have

$$\begin{aligned} v^{\alpha+\gamma,\delta}(y) |(K f_m)(y) - (K_m^* f_m)(y)| &\le v^{\alpha+\gamma,\delta}(y) |(K f_m)(y) - (K_m f_m)(y)| \\ &\quad + v^{\alpha+\gamma,\delta}(y) |(K_m f_m)(y) - (K_m^* f_m)(y)|. \end{aligned}$$

The first addendum at the right-hand side is dominated by $\mathcal{C} \frac{\log m}{m^r} \|f_m v^{\alpha+\gamma,\delta}\|$ in view of (4.18). About the second addendum, using (4.4), we get

$$\begin{aligned} &v^{\alpha+\gamma,\delta}(y) |(K_m f_m)(y) - (K_m^* f_m)(y)| \\ &\quad = |v^{\alpha+\gamma,\delta}(y) L_m^{\alpha,-\alpha} (\tilde{K} f_m - K^* f_m, y)| \\ &\quad \le \mathcal{C} \log m \max_{i=1,\dots,m} v^{\alpha+\gamma,\delta}(x_i) |(\tilde{K} f_m)(x_i) - (K^* f_m)(x_i)| \\ &\quad = \mathcal{C} \log m \max_{i=1,\dots,m} v^{\alpha+\gamma,\delta}(x_i) \left| \int_{-1}^{1} L_m^{\alpha,-\alpha} ((A_y k)(\cdot, x_i) \right. \\ &\qquad \left. - (A_y L_{m,y}^{-\alpha,\alpha} k)(\cdot, x_i), x) f_m(x) v^{\alpha,-\alpha}(x) dx \right|. \end{aligned}$$

Applying the Gaussian quadrature rule to the last integral and using (4.42), we get

$$\begin{aligned} &v^{\alpha+\gamma,\delta}(y) |(K_m f_m)(y) - (K_m^* f_m)(y)| \\ &\quad \le \mathcal{C} \log m \max_{i=1,\dots,m} v^{\alpha+\gamma,\delta}(x_i) \left| \sum_{k=1}^{m} \lambda_k^{\alpha,-\alpha} [\Psi(x_k, x_i) - \Psi_m(x_k, x_i)] f_m(x_k) \right| \\ &\quad \le \mathcal{C} \max_{|x| \le 1} |f_m(x) v^{\alpha+\gamma,\delta}(x)| (\log m) \|A_m - A_m^*\| \\ &\quad \le \mathcal{C} \|f_m v^{\alpha+\gamma,\delta}\| \frac{\log^2 m}{m^r}. \end{aligned}$$

Thus, we have

$$\|v^{\alpha+\gamma,\delta}(K f - K_m^* f)\| \le \mathcal{C} \|f v^{\alpha+\gamma,\delta}\| \frac{\log^2 m}{m^r}. \tag{4.43}$$

In order to estimate $Ag - L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g$ we have

$$\begin{aligned} & v^{\alpha+\gamma,\delta}(y)|(Ag)(y) - (L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g)(y)| \\ & \quad \le v^{\alpha+\gamma,\delta}(y)|(Ag)(y) - (L_m^{\alpha,-\alpha} Ag)(y)| \\ & \qquad + v^{\alpha+\gamma,\delta}(y)|(L_m^{\alpha,-\alpha} Ag)(y) - (L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g)(y)|. \end{aligned}$$

Since, by (4.33), $g \in Z_r(v^{0,\alpha})$ implies $Ag \in Z_r(v^{\alpha+\gamma,\delta})$, using (4.5), (4.2), and (3.7), the first addendum at the right-hand side is dominated by $\mathcal{C}\frac{\log m}{m^r}\|g\|_{Z_r(v^{0,\alpha})}$. Moreover, about the second addendum, using (4.4) and (3.22), we get

$$\begin{aligned} & v^{\alpha+\gamma,\delta}(y)|(L_m^{\alpha,-\alpha} Ag)(y) - (L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g)(y)| \\ & \quad = |v^{\alpha+\gamma,\delta}(y) L_m^{\alpha,-\alpha}(Ag - A L_m^{-\alpha,\alpha} g)| \\ & \quad \le \mathcal{C} \log m \max_{i=1,\dots,m} v^{\alpha+\gamma,\delta}(x_i)|G(x_i) - G_m(x_i)| \\ & \quad \le \mathcal{C}\frac{\log^2 m}{m^r}\|g\|_{Z_r(v^{0,\alpha})}. \end{aligned}$$

Consequently, we get

$$\|v^{\alpha+\gamma,\delta}(Ag - L_m^{\alpha,-\alpha} A L_m^{-\alpha,\alpha} g)\| \le \mathcal{C}\frac{\log^2 m}{m^r}\|g\|_{Z_r(v^{0,\alpha})}. \tag{4.44}$$

Finally, combining (4.43) and (4.44) with (4.20), (3.24) follows. ∎

Acknowledgments. The authors are grateful to Prof. P. Junghanns for his useful remarks and thank the referees for their contribution in improving the paper.

REFERENCES

[1] K. E Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge Monogr. Appl. Comput. Math. 4, Cambridge University Press, Cambridge, UK, 1997.

[2] B. Bialecki, *Sinc quadratures for Cauchy principal value integrals*, in Numerical Integration (Bergen, 1991), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 357, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992, pp. 81–92.

[3] M. R. Capobianco, G. Mastroianni, and M. G. Russo, *Pointwise and uniform approximation of the finite Hilbert transform*, in Approximation and Optimization, Vol. I (Cluj-Napoca, 1996), Transilvania, Cluj-Napoca, 1997, pp. 45–66.

[4] M. R. Capobianco and M. G. Russo, *Uniform convergence estimates for a collocation method for the Cauchy singular integral equation*, J. Integral Equations Appl., 9 (1997), pp. 21–45.

[5] G. Criscuolo and G. Mastroianni, *On the convergence of the Gauss quadrature rules for the Cauchy principal value integrals*, Ricerche Mat., 3 (1986), pp. 45–60.

[6] G. Criscuolo and G. Mastroianni, *On the convergence of product formulas for the evaluation of certain two-dimensional Cauchy principal value integrals*, BIT, 27 (1987), pp. 72–84.

[7] G. Criscuolo and G. Mastroianni, *On the convergence of an interpolatory product rule for evaluating Cauchy principal value integrals*, Math. Comp., 48 (1987), pp. 725–735.

[8] G. Criscuolo and G. Mastroianni, *On the uniform convergence of Gaussian quadrature rules for Cauchy principal value integrals*, Numer. Math., 54 (1989), pp. 445–461.

[9] M. C. De Bonis and G. Mastroianni, *Mapping properties of some singular operators in Besov type subspaces of $C(-1,1)$*, Integral Equations Operator Theory, to appear.

[10] Z. Ditzian and V. Totik, *Moduli of Smoothness*, Springer Ser. Comput. Math. 9, Springer-Verlag, New York, 1987.

[11] Z. Ditzian and V. Totik, *Remarks on Besov spaces and best polynomial approximation*, Proc. Amer. Math. Soc., 104 (1988), pp. 1059–1066.

[12] R. Hagen, S. Roch, and B. Silbermann, *C^* Algebras and Numerical Analysis*, Monogr. Textbooks Pure Appl. Math. 236, Marcel Dekker, New York, 2001.

[13] M. Iri, S. Moriguti, and Y. Takasawa, *On a certain quadrature formula*, J. Comput. Appl. Math., 17 (1987), pp. 3–20.

[14] P. Junghanns and U. Luther, *Cauchy singular integral equations in spaces of continuous functions and methods for their numerical solution*, J. Comput. Appl. Math., 77 (1997), pp. 201–237.

[15] P. Junghanns and U. Luther, *Uniform convergence of the quadrature method for Cauchy singular integral equations with weakly singular perturbation kernels*, in Proceedings of the 3rd International Conference on Functional Analysis and Approximation Theory, Vol. II (Acquafredda di Maratea, 1996), Rend. Circ. Mat. Palermo (2) Suppl., 52 (1988), pp. 551–566.

[16] P. Junghanns and U. Luther, *Uniform convergence of a fast algorithm for a Cauchy singular integral equation*, in Proceedings of the 6th Conference of the International Linear Algebra Society (Chemnitz, 1996), Linear Algebra Appl., 275/276 (1998), pp. 327–347.

[17] P. Junghanns and B. Silbermann, *Zur Theorie der Näherungsverfahren für singuläre Integralgleichungen auf Intervallen*, Math. Nachr., 103 (1981), pp. 199–244.

[18] P. Junghanns and B. Silbermann, *Numerical Analysis of the Quadrature Method for Solving Linear and Nonlinear Singular Integral Equations*, Wiss. Schriftenreihe d. TUK, Germany, 10/1988.

[19] A. I. Kalandya, *Mathematical Methods of Two-Dimensional Elasticity*, Mir Publisher, Moscow, 1975.

[20] C. Laurita and G. Mastroianni, *Revisiting a quadrature method for Cauchy singular integral equations with a weakly singular perturbation kernel*, in Problems and Methods in Mathematical Physics (Chemnitz, 1999), Oper. Theory Adv. Appl. 121, Birkhäuser, Basel, 2001, pp. 307–326.

[21] C. Laurita and G. Mastroianni, *Condition numbers in numerical methods for Fredholm integral equations of the second kind*, J. Integral Equations Appl., 14 (2002), pp. 311–341.

[22] J. Lund and K. L. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, 1992.

[23] U. Luther and G. Mastroianni, *Fourier projections in weighted L^∞ spaces*, in Problems and Methods in Mathematical Physics (Chemnitz, 1999), Oper. Theory Adv. Appl. 121, Birkhäuser, Basel, 2001, pp. 327–351.

[24] U. Luther and M. G. Russo, *Boundedness of the Hilbert transformation in some weighted Besov type spaces*, Integral Equations Operator Theory, 36 (2000), pp. 220–240.

[25] G. Mastroianni and S. Prössdorf, *Some nodes matrices appearing in the numerical analysis for singular integral equations*, BIT, 34 (1994), pp. 120–128.

[26] G. Mastroianni and S. Prössdorf, *A quadrature method for Cauchy integral equations with weakly singular perturbation kernel*, J. Integral Equations Appl., 4 (1992), pp. 205–228.

[27] G. Mastroianni and M. G. Russo, *Lagrange interpolation in some weighted uniform spaces*, Dedicated to Professor Dragoslav S. Mitrinovic (1908–1995) (Nis, 1996), Facta Univ. Ser. Math. Inform., 12 (1997), pp. 185–201.

[28] G. Mastroianni and M. G. Russo, *Lagrange interpolation in weighted Besov spaces*, Constr. Approx., 15 (1999), pp. 257–289.

[29] G. Mastroianni and M. G. Russo, *Weighted Marcinkiewicz inequalities and boundedness of the Lagrange operator*, in Mathematical Analysis and Applications, Hadronic Press, Palm Harbor, FL, 2000, pp. 149–182.

[30] G. Mastroianni, M. G. Russo, and W. Themistoclakis, *Numerical Methods for Cauchy Singular Integral Equations in Spaces of Weighted Continuous Functions*, Oper. Theory Adv. Appl. 160, Birkhäuser, Basel, 2005, pp. 311–336.

[31] G. Mastroianni and V. Totik, *Weighted polynomial inequalities with doubling and A_∞ weights*, Constr. Approx., 16 (2000), pp. 37–71.

[32] S. G. Mikhlin and S. Prössdorf, *Singular Integral Operators*, Akademie-Verlag, Berlin, 1986.

[33] G. Monegato, *On the weights of certain quadratures for the numerical evaluation of Cauchy principal value integrals and their derivatives*, Numer. Math., 50 (1987), pp. 273–281.

[34] G. Monegato and A. Palamara Orsi, *Product formulas for Fredholm integral equations with rational kernel functions*, in Numerical Integration III 85, H. Brass and G. Hämerlin, eds., Birkhäuser, Basel, 1987, pp. 140–156.

[35] G. Monegato and S. Prössdorf, *Uniform convergence estimates for a collocation and a discrete collocation method for the generalized airfoil equation*, in Contributions to Numerical Mathematics, A.G. Argaval, ed., World Scientific, River Edge, NJ, 1993, pp. 285–299 (see also the errata corrige in the Internal Reprint No. 14 (1993) Dip. Mat. Politecnico di Torino).

[36] M. Mori, *An IMT-type double exponential formula for numerical integration*, Publ. Res. Inst. Math. Sci., 14 (1978), pp. 713–728.

[37] N. I. Muskhelishvili, *Singular Integral Equations*, Noordhoff, Groningen, 1953.

[38] P. Nevai, *Mean convergence of Lagrange interpolation* III, Trans. Amer. Math. Soc., 282 (1984), pp. 669–698.
[39] P. Nevai, *Hilbert transforms and Lagrange interpolation (letter to the editor)*, J. Approx. Theory, 60 (1990), pp. 360–363.
[40] T. Ooura and M. Mori, *The double exponential formula for oscillatory functions over the half infinite interval*, J. Comput. Appl. Math., 38 (1991), pp. 353–360.
[41] V. Z. Parton and P. I. Perlin, *Integral Equations in Elasticity*, Mir Publisher, Moscow, 1982.
[42] R. Piessens, *Modified Clenshaw-Curtis integration and applications to numerical computation of integral transforms*, in Numerical Integration (Halifax-N.S., 1986), P. Keast and G. Fairweather, eds., Reidel, Dordrecht, The Netherlands, 1987, pp. 35–51.
[43] R. Piessens and M. Branders, *Numerical solution of integral equations of mathematical physics using Chebyshev polynomials*, J. Comput. Phys., 21 (1976), pp. 178–196.
[44] H. Pollard, *The mean convergence of orthogonal series* II, Trans. Amer. Math. Soc., 63 (1948), pp. 355–367.
[45] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Akademie-Verlag, Berlin, 1991, Birkhäuser, Basel, 1991.
[46] I. H. Sloan, *A quadrature-based approach to improving the collocation method*, Numer. Math., 54 (1988), pp. 41–56.
[47] I. H. Sloan and A. Spence, *Projection methods for integral equation on the half line*, IMA J. Numer. Anal., 6 (1986), pp 153–172.
[48] I. H. Sloan and V. Thomée, *Superconvergence of the Galerkin iterates for integral equations of the second kind*, J. Integral Equations, 9 (1985), pp. 1–23.
[49] G. Szegö, *Orthogonal Polynomials*, AMS, Providence, RI, 1939.
[50] H. Takahasi and M. Mori, *Double exponential formulas for numerical integration*, Publ. Res. Inst. Math. Sci., 9 (1974), pp. 721–741.
[51] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Dover, New York, 1994.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1375–1404

ANALYSIS OF DISCONTINUOUS FINITE ELEMENT METHODS FOR GROUND WATER/SURFACE WATER COUPLING*

CLINT DAWSON†

Abstract. We derive and analyze new numerical approaches for modeling coupled ground water/surface water flow. In this coupled model, surface water flow is described by the depth-averaged shallow water equations, while ground water is modeled by saturated Darcy flow. The coupling between the two models assumes continuity of pressure and water flux across the ground water/surface water interface. The coupled model is approximated by a local discontinuous Galerkin method for ground water flow and a discontinuous Galerkin method for surface water flow. A priori error estimates are derived for this approach. A closely related approach where the well-known mixed finite element method is applied to the ground water flow equations is also described and analyzed. One advantage of these approaches is that they allow for the ground water and surface water domains to be meshed independently, under some mild restrictions.

Key words. ground water flow equations, shallow water equations, mixed finite element method, discontinuous Galerkin method, coupled method

AMS subject classifications. 65M15, 65M60, 76S05

DOI. 10.1137/050639405

1. Introduction.

Comprehensive water resource management requires a careful study of the interactions of ground water and surface water. As noted in a recent report of the United States Geological Survey [45], "Traditionally, management of water resources has focused on surface water or ground water as if they were separate entities ... Effective policies and management practices must be built on a foundation that recognizes that surface water and ground water are simply two manifestations of a single integrated resource." This article gives a number of examples of ground water/surface water interactions and the effect these interactions have on the environment and the water cycle.

In this paper, we focus on the numerical approximation of coupled ground water/surface water flow. Surface water flow models are derived from the incompressible Navier–Stokes equations and can take many forms, including two- and three-dimensional shallow water models, overland flow, and kinematic and diffusive wave models; see [42] for an overview of shallow water hydrodynamics. Ground water flow models include single phase Darcy flow and various multiphase models which account for unsaturated flow through the vadose zone, e.g., the Richards equation or a two-phase air-water model [23]. These models are fairly well understood within their respective regimes. How these models should be coupled is still a question open for debate.

The coupling of Stokes and Darcy flows has been studied mathematically and numerically in several recent papers; see, for example [27, 35, 5], where the coupling is imposed through the Beavers–Joseph–Saffman interface conditions [8, 36, 25]. Application of these conditions to the coupling of surface and ground water flow is discussed

*Received by the editors September 1, 2005; accepted for publication (in revised form) April 6, 2006; published electronically July 7, 2006. This work was supported by National Science Foundation grant DMS-0411413.

http://www.siam.org/journals/sinum/44-4/63940.html

†Center for Subsurface Modeling, 1 University Station, C0200, The University of Texas, Austin, TX 78712 (clint@ices.utexas.edu).

in Miglio, Discacciati, and Quarteroni [30], where a three-dimensional nonhydrostatic shallow water model is coupled with Darcy flow; the authors prove well-posedness of the model in the case of linear Stokes flow and formulate an iterative method to solve the coupled system.

In the engineering literature, various numerical models for coupling depth-averaged shallow water flow equations with single and multiphase ground water flow equations have been investigated; see, for example, [37, 39, 47, 46, 38]. Within these models, the coupling between surface and ground water is imposed in various ways. One approach that has been proposed is to compute an "exchange flux." This approach assumes the existence of an interfacial domain connecting the two domains, referred to as the conductance concept [2, 40]. The interfacial domain is characterized by a thickness parameter m and permeability k. The flux is then calculated as the ratio $\lambda = k/m$ multiplied by the difference between the surface water and ground water pressures at the interface. The exchange flux enters into the ground water and surface water flow equations through source terms. The drawback to this approach is that the presence of such a distinct interfacial domain has not been observed in the field [14]; thus determining k and m is problematic. The use of the conductance concept in numerical modeling of ground water/surface water coupling dates back to at least 1969 [24].

Another approach, which is often used in multiphysics problems [29, 32], is to assume continuity of normal flux and of an appropriately defined "pressure" across the interface. This is equivalent to the conductance concept if the thickness parameter m goes to zero and is also part of the Beavers–Joseph–Saffman interface conditions for coupling Darcy flow with the three-dimensional nonhydrostatic shallow water equations as discussed in [30]. This approach is studied for ground water/surface water coupling in a recent paper by Kollet and Maxwell [26] and is the approach we will take here, though the way we implement the interface conditions is different. The flow model we consider is based on the depth-averaged, two-dimensional shallow water equations coupled with a saturated ground water flow model. The depth-averaged shallow water equations are widely used for surface water flow modeling, because vertical effects are often negligible in comparison to the horizontal. Upon depth averaging, the ground water normal velocity at the ground water/surface water interface enters the shallow water continuity equation as a source term. The continuity of pressure is enforced by assuming that the ground water hydraulic head is equal to the surface water height at the interface. Thus, the ground water flow equations have a time-dependent Dirichlet boundary condition at the surface water interface, where the boundary value satisfies the shallow water equations. Unlike the Darcy/Stokes couplings mentioned above, the surface water momentum equation is not directly involved in this coupling.

Based upon this model, we derive a weak formulation and discuss the approximation of the weak solution using discontinuous and mixed finite element methods. Discontinuous Galerkin (DG) and mixed finite element (MFE) methods have been extensively studied for elliptic flow problems such as ground water flow [44, 34, 31, 15, 33, 12, 11, 10, 7, 6, 4, 3, 13, 16, 43, 41, 22, 21]. More recently, DG methods have been applied to the shallow water equations [1, 18, 19]. These methods have certain features, which have been discussed at length in these and other papers, which make them of interest for approximating the flow models under consideration here. One advantage of these methods for modeling ground water/surface water interaction is that boundary conditions are enforced weakly, which allows for flexibility in the coupling. In particular, the ground water and surface water domains can be meshed independently of each other with some minor restrictions. Another advantage of

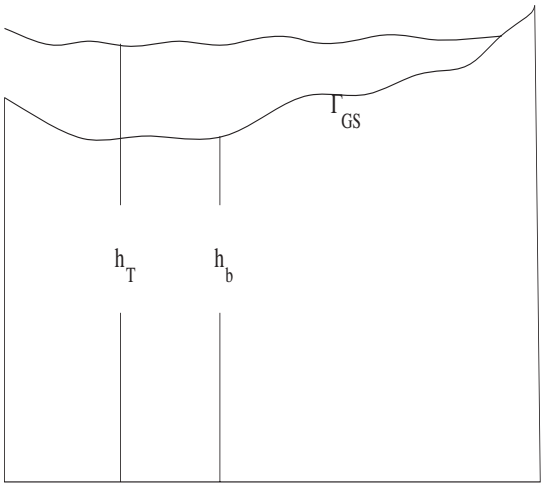


FIG. 1. *Schematic of a ground water/surface water interface.*

these methods, which has been well documented, is that they provide for locally conservative, flux-continuous flow fields. This property is important when coupling the flow equations with a transport model, for example, when considering contaminant transport [20].

In the next section, we derive the coupled flow model under consideration. We express the ground water flow equations in mixed form and consider the approximation of this model using the local discontinuous Galerkin (LDG) method as described in [17, 15]. We approximate the shallow water flow model also using a DG method as described in [19]. The coupling between the formulations is through a source term in the shallow water equations and a boundary term in the ground water flow equation. We refer to this approach as the LDG/DG method. In section 4 we derive a priori error estimates for this coupled method under very mild assumptions on how the ground water and surface water domains are discretized.

The flexibility of the LDG method allows us to easily formulate an approach based on the MFE method [33] for ground water flow, coupled with the DG method for surface water. We refer to this approach as MFE/DG, and error estimates for this approach are derived in section 5.

Finally, in section 6, we give some preliminary numerical results for the LDG/DG method on a model ground water/surface water flow problem.

2. Problem definition. Let h_b denote the bathymetric height of the ground water/surface water interface, measured relative to a reference z plane; see Figure 1. We assume that h_b is a continuous, piecewise differentiable function. Let h_T denote the total surface water height above this reference plane, and let the surface water depth $h_s = h_T - h_b$.

In the ground water domain $\Omega_{\rm g}$, let h_g denote the hydraulic head, also measured relative to the reference z-plane, and let $\mathbf{u}_g = (u_g, v_g, w_g)$ denote the ground water velocity. The ground water flow equations are given by

$$\left.\begin{array}{l} \mathbf{u}_g = -K\nabla_{(x,y,z)} h_g, \\ \nabla_{(x,y,z)} \cdot \mathbf{u}_g = f_g \end{array}\right\}, \quad (x,y,z) \in \Omega_{\rm g}, \tag{1}$$

where K is a symmetric, positive definite, hydraulic conductivity tensor and f_g models ground water source/sink terms. Here

$$\nabla_{(x,y,z)} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right).$$

Assuming continuity of pressure across the ground water/surface water interface, we have the boundary conditions

$$h_g = h_T \quad \text{on } \Gamma_{GS}, \tag{2}$$

$$h_g = h_g^D \quad \text{on } \Gamma_D, \tag{3}$$

$$\mathbf{u}_g \cdot \mathbf{n}_g = u_N \quad \text{on } \Gamma_N. \tag{4}$$

Here $\Gamma_{GS} \subset \partial\Omega_{\mathrm{g}}$ represents the ground water/surface water interface, $\Gamma_D \subset \partial\Omega_{\mathrm{g}}$ is the Dirichlet portion of the boundary external to Γ_{GS}, Γ_N is the Neumann portion of the boundary, and $\mathbf{n}_g$ is the outward normal to $\partial\Omega_{\mathrm{g}}$. We will assume that $\Gamma_D \cup \Gamma_{GS} \neq \emptyset$.

The shallow water flow equations are obtained from the three-dimensional Navier–Stokes equations under the hydrostatic pressure assumption [42]. Note that the surface Γ_{GS} is described by the bathymetry $h_b(x, y)$ for points (x, y) contained in some two-dimensional parameter space Ω_{s}; thus the normal $\mathbf{n}_g$ to this surface is proportional to the vector $(-\partial h_b/\partial x, -\partial h_b/\partial y, 1)$. The depth-integrated shallow water equations are then defined over the domain Ω_{s}. First, the shallow water continuity equation is obtained by integrating the incompressibility condition $\nabla_{(x,y,z)} \cdot \mathbf{u} = 0$ over the water depth, where $\mathbf{u} = (u_s, v_s, w_s)$ is the three-dimensional surface water velocity. Applying the kinematic boundary condition at the free surface [42] and the continuity of flux across the ground water/surface water interface results in

$$\partial_t h_s + \nabla_{(x,y)} \cdot (\mathbf{u}_s h_s) = \mathbf{u}_g \cdot (-\nabla_{(x,y)} h_b, 1), \quad (x, y) \in \Omega_{\mathrm{s}}, \quad t > 0, \tag{5}$$

where $\nabla_{(x,y)}$ is the gradient operator in x-y space. Here $\mathbf{u}_s = (\bar{u}_s, \bar{v}_s)$ is the depth-averaged horizontal surface water velocity, defined by

$$\bar{u}_s(x, y) \equiv \frac{1}{h_s} \int_0^{h_s} u_s(x, y, z) dz$$

with a similar definition for $\bar{v}_s$. In (5), we are neglecting other external sources/sinks such as rainfall and evaporation. The depth-averaged shallow water momentum equation is

$$\partial_t \mathbf{u}_s + \mathbf{u}_s \cdot \nabla_{(x,y)} \mathbf{u}_s + \tau_{bf} \mathbf{u}_s + g \nabla_{(x,y)} h_T - \mu \nabla_{(x,y)} \cdot (\nabla_{(x,y)} \mathbf{u}_s) = \mathcal{F}, \tag{6}$$

where $\tau_{bf} \geq 0$ is a bottom friction coefficient, g is gravitational acceleration, $\mu > 0$ is the horizontal eddy viscosity, and $\mathcal{F}$ represents additional forcing terms, such as wind stress, atmospheric pressure gradient, etc. For simplicity, we will assume linear bottom friction so that $\tau_{bf} = \tau_{bf}(x, y)$.

Let the domain boundary $\partial\Omega_s$ be divided into inflow and outflow regions $\partial\Omega_s = \partial\Omega_{s,in} \cup \partial\Omega_{s,out}$ defined as

$$\partial\Omega_{s,in} = \{(x, y) \in \partial\Omega_{\mathrm{s}} : \mathbf{u}_s \cdot \mathbf{n}_s < 0\}, \tag{7}$$

$$\partial\Omega_{s,out} = \{(x, y) \in \partial\Omega_{\mathrm{s}} : \mathbf{u}_s \cdot \mathbf{n}_s \geq 0\}, \tag{8}$$

where $\mathbf{n}_s$ is the outward normal to $\partial\Omega_s$, and consider the following boundary and initial conditions:

$$h_s(\cdot,0) = h_s^0, \quad \Omega_{\rm s}, \tag{9}$$

$$\mathbf{u}_s(\cdot,0) = \mathbf{u}_s^0, \quad \Omega_{\rm s}, \tag{10}$$

$$h_s = h_s^I, \quad \partial\Omega_{s,in} \times (0,T], \tag{11}$$

$$\mathbf{u}_s = \hat{\mathbf{u}}_s, \quad \partial\Omega_s \times (0,T], \tag{12}$$

where h_s^I is a specified inflow water height and $\hat{\mathbf{u}}_s$ is a specified velocity.

Therefore, the coupling between ground water and surface water flow equations in this setting occurs in the boundary conditions for ground water flow and the forcing term for the surface water elevation. Even though we have assumed a simple ground water flow model, the same type of coupling would occur if ground water flow were described by the Richards equation, for example.

We will discretize the ground water flow equation using the LDG method as described in [15]. Later we will see that the general framework of the LDG method allows us to also consider the approximation of these equations by the MFE method [33] by restricting our finite element spaces. We will discretize the primitive continuity equation (5) and momentum equation (6) also using DG methods. In particular we apply a method described in [19], where the advection term in (6) is discretized by the upwinding technique of Lesaint and Raviart [28], and the eddy viscosity terms are discretized using a nonsymmetric, interior penalty Galerkin method (NIPG) [34].

3. The LDG/DG Method. Let $\{\mathcal{T}_{\Delta_g,g}\}_{\Delta_g>0}$ denote a family of regular finite element partitions of $\Omega_{\rm g}$ such that no individual element $\Omega_{e,g}$ crosses $\partial\Omega_g$. Let $\Delta_{e,g}$ denote the element diameter with Δ_g being the maximal element diameter. We also assume that each element $\Omega_{e,g}$ is Lipschitz and affinely equivalent to one of several reference elements [9]. Similarly, let $\{\mathcal{T}_{\Delta_s,s}\}_{\Delta_s>0}$ denote a family of regular finite element partitions of $\Omega_{\rm s}$, where Δ_s is the maximal element diameter. We do not require that the partitions match up at the interface Γ_{GS}. We allow $\mathcal{T}_{\Delta_g,g}$ and $\mathcal{T}_{\Delta_s,s}$ to be nonconforming within their respective domains; i.e., element boundaries do not have to align. We assume, however, that the number of elements sharing a face is bounded independently of Δ_g or Δ_s. We assume further that each triangulation is locally quasi-uniform [9].

On each triangulation $\mathcal{T}_{\Delta_g,g}$ we will approximate $\mathbf{u}_g$ in the space $\mathcal{V}_{\Delta_g,g}$ and h_g in the space $\mathcal{W}_{\Delta_g,g}$, where

$$\mathcal{V}_{\Delta_g,g} = \{\mathbf{v} \in L^2(\Omega_{\rm g})^3 : \mathbf{v}|_{\Omega_{e,g}} \in (\mathcal{S}(\Omega_{e,g}))^3 \quad \forall\, \Omega_{e,g} \in \Omega_{\rm g}\}, \tag{13}$$

$$\mathcal{W}_{\Delta_g,g} = \{w \in L^2(\Omega_{\rm g}) : w|_{\Omega_{e,g}} \in \mathcal{S}(\Omega_{e,g}) \quad \forall\, \Omega_{e,g} \in \Omega_{\rm g}\}. \tag{14}$$

We will assume that $\mathcal{S}$ consists of complete polynomials of degree $k_g \geq 1$.

Similarly, on each triangulation $\mathcal{T}_{\Delta_s,s}$, we approximate h_s in the space $\mathcal{W}_{\Delta_s,s}$ and $\mathbf{u}_s$ in the space $\mathcal{V}_{\Delta_s,s}$, which consist of complete polynomials of degree $k_s \geq 1$ defined on each element.

In our numerical procedure defined below, $\mathbf{u}_g \approx \boldsymbol{U}_g \in \mathcal{V}_{\Delta_g,g}$, $h_g \approx H_g \in \mathcal{W}_{\Delta_g,g}$, $\mathbf{u}_s \approx \boldsymbol{U}_s \in \mathcal{V}_{\Delta_s,s}$, and $h_s \approx H_s \in \mathcal{W}_{\Delta_s,s}$.

Suppose e is an interior face in either finite element mesh; then e has two elements adjacent to it, which we label Ω_e^- and Ω_e^+. Further suppose $(\mathbf{v}, w)$ are smooth functions defined on these elements. Let $(\mathbf{v}^\pm, w^\pm)$ denote the traces of $(\mathbf{v}, w)$ on e from the interiors of the elements Ω_e^+ and Ω_e^-, respectively. Let $\mathbf{n}^-$ denote the normal vector

to e pointing from Ω_e^- to Ω^+ with a similar definition for $\mathbf{n}^+$ (hence $\mathbf{n}^+ = -\mathbf{n}^-$). We define the average $\{\cdot\}$ and the jump $[\![\cdot]\!]$ for $\mathbf{x} \in e$ as follows:

$$\{\mathbf{v}\} = (\mathbf{v}^- + \mathbf{v}^+)/2, \qquad \{w\} = (w^- + w^+)/2, \tag{15}$$

$$[\![\mathbf{v}]\!] = \mathbf{v}^+ \cdot \mathbf{n}^+ + \mathbf{v}^- \cdot \mathbf{n}^-, \qquad [\![w]\!] = w^+\mathbf{n}^+ + w^-\mathbf{n}^-. \tag{16}$$

We will use the $L^2(R)$ inner product notation $(\cdot,\cdot)_R$ for domains $R \in \mathbb{R}^d$ and the notation $\langle\cdot,\cdot\rangle_R$ to denote integration over $(d-1)$-dimensional surfaces. Let $\|\cdot\|_R$ denote the $L^2(R)$ norm on any spatial region R. The notation $\langle\cdot,\cdot,\rangle_{\mathcal{E}_{i,g}}$ denotes integration over all interior element faces in $\Omega_{\rm g}$, and $\langle\cdot,\cdot,\rangle_{\mathcal{E}_{i,s}}$ denotes integration over all interior element faces in $\Omega_{\rm s}$.

The LDG and DG weak forms of (1) and (5)–(6) are obtained as follows. Multiply the first equation in (1) by a function $\mathbf{v}_g$ which is in $(H^1(\Omega_{e,g}))^3$, integrate by parts, and sum over all elements; define

$$\begin{aligned}\mathcal{A}_g(\mathbf{u}_g, h_g, h_s; \mathbf{v}_g) &\equiv (K^{-1}\mathbf{u}_g, \mathbf{v}_g)_{\Omega_{\rm g}} - (h_g, \nabla_{(x,y,z)} \cdot \mathbf{v}_g)_{\Omega_{\rm g}} \\ &\quad + \langle \{h_g\}, [\![\mathbf{v}_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle h_g, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_N} + \langle h_s, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}}.\end{aligned} \tag{17}$$

Then by (1), (2), and (3),

$$\mathcal{A}_g(\mathbf{u}_g, h_g, h_s; \mathbf{v}_g) = -\langle h_b, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}} - \langle h_g^D, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_D}. \tag{18}$$

By the continuity of h_g, $\{h_g\} = h_g$, but this notation will be useful in defining the LDG method. Next, multiply the second equation in (1) by $w_g \in H^1(\Omega_{e,g})$. Let

$$\begin{aligned}&\mathcal{B}_g(\mathbf{u}_g, h_g, h_s; w_g) \equiv (\nabla_{(x,y,z)} \cdot \mathbf{u}_g, w_g)_{\Omega_{\rm g}} - \langle [\![\mathbf{u}_g]\!], \{w_g\}\rangle_{\mathcal{E}_{i,g}} \\ &- \langle \mathbf{u}_g \cdot \mathbf{n}_g, w_g\rangle_{\Gamma_N} + \langle \sigma_g [\![h_g]\!], [\![w_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \sigma_g(h_g - h_s), w_g\rangle_{\Gamma_{GS}} + \langle \sigma_g h_g, w_g\rangle_{\Gamma_D}.\end{aligned} \tag{19}$$

Here we add some jump and penalty terms which are necessary to stabilize the method; the penalty function $\sigma_g > 0$ will be discussed in more detail below. By (1) and the boundary conditions (2)–(4),

$$\begin{aligned}\mathcal{B}_g(\mathbf{u}_g, h_g, h_s; w_g) &= -\langle u_N, w_g\rangle_{\Gamma_N} + \langle \sigma_g h_b, w_g\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma_g h_g^D, w_g\rangle_{\Gamma_D} + (f_g, w_g)_{\Omega_{\rm g}}.\end{aligned} \tag{20}$$

Multiply (5) by a test function $w_s \in H^1(\Omega_{e,s})$, integrate by parts, and apply the boundary condition (11); define

$$\begin{aligned}\mathcal{B}_s(\mathbf{u}_g, \mathbf{u}_s, h_g, h_s; w_s) &\equiv (\partial_t h_s, w_s)_{\Omega_{\rm s}} - (\mathbf{u}_s h_s, \nabla_{(x,y)} w_s)_{\Omega_{\rm s}} + \langle \{\mathbf{u}_s\} h_s^{\uparrow}, [\![w_s]\!]\rangle_{\mathcal{E}_{i,s}} \\ &\quad - \langle \sigma_g(h_g - h_s), w_s\rangle_{\Gamma_{GS}} - \langle \mathbf{u}_g \cdot \mathbf{n}_g, w_s\rangle_{\Gamma_{GS}} + \langle \mathbf{u}_s \cdot \mathbf{n}_s h_s, w_s\rangle_{\partial\Omega_{s,out}} \\ &\quad + \langle \mathbf{u}_s \cdot \mathbf{n}_s h_s^I, w_s\rangle_{\partial\Omega_{s,in}}.\end{aligned} \tag{21}$$

Here we have added a penalty term on the interface Γ_{GS} similar to the fifth term in (19). The penalty function σ_g is the same as in the corresponding term in (19). We have also used the fact that by the definition of the domain $\Omega_{\rm s}$ and the surface integral

$$\langle \mathbf{u}_g \cdot \mathbf{n}_g, w_s\rangle_{\Gamma_{GS}} = (\mathbf{u}_g \cdot (-\nabla_{(x,y)} h_b, 1), w_s)_{\Omega_{\rm s}}.$$

In addition we define the "upwind" value of h_s on an interior element face e by

$$h_s^{\uparrow} = \begin{cases} h_s^-, & \{\boldsymbol{U}_s\} \cdot \mathbf{n}^- > 0, \\ h_s^+, & \{\boldsymbol{U}_s\} \cdot \mathbf{n}^+ > 0. \end{cases} \tag{22}$$

For the true solution h_s, $h_s^{\uparrow} = h_s$, but note that in general the upwind value is defined using the approximation $\boldsymbol{U}_s \approx \mathbf{u}_s$. Thus by (5)

$$\mathcal{B}_s(\mathbf{u}_g, \mathbf{u}_s, h_g, h_s; w_s) = -\langle \sigma_g h_b, w_s \rangle_{\Gamma_{GS}}. \tag{23}$$

Finally, as in [19], we discretize the momentum equation (6) using the Lesaint–Raviart upwinding method for the advective terms. For each element $\Omega_{e,s} \in \mathcal{T}_{\Delta_s,s}$, let

$$\partial\Omega_{e,s}^{-} = \{\mathbf{x} \in \partial\Omega_{e,s} : \{\boldsymbol{U}_s\} \cdot \mathbf{n}_e < 0\},$$

where $\mathbf{n}_e$ is the unit outward normal to $\partial\Omega_{e,s}$. This set is defined using the approximation $\boldsymbol{U}_s$. Multiply (6) by a test function $\mathbf{v}_s \in (H^2(\Omega_{e,s}))^2$ and apply the NIPG method for diffusion terms; define

$$\begin{aligned}
\mathcal{A}_s(\mathbf{u}_s, h_s; \mathbf{v}_s) \equiv\ & (\partial_t \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} + (\mathbf{u}_s \cdot \nabla_{(x,y)} \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} \\
& + \sum_{\partial\Omega_{e,s}^{-} \in \mathcal{T}_{\Delta_s,s}} \langle |\{\mathbf{u}_s\} \cdot \mathbf{n}_e| (\mathbf{u}_s^{\text{int}} - \mathbf{u}_s^{\text{ext}}), \mathbf{v}_s^{\text{int}} \rangle_{\partial\Omega_{e,s}^{-}} + (\tau_{bf} \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} \\
& + (g \nabla_{(x,y)} h_s, \mathbf{v}_s)_{\Omega_s} - \langle g [\![h_s]\!], \{\mathbf{v}_s\} \rangle_{\mathcal{E}_{i,s}} - \langle g h_s, \mathbf{v}_s \cdot \mathbf{n} \rangle_{\partial\Omega_{s,in}} \\
& + \mu (\nabla_{(x,y)} \mathbf{u}_s, \nabla_{(x,y)} \mathbf{v}_s)_{\Omega_s} - \mu \langle \{\nabla_{(x,y)} \mathbf{u}_s\}, [\![\mathbf{v}_s]\!] \rangle_{\mathcal{E}_{i,s}} + \mu \langle \{\nabla_{(x,y)} \mathbf{v}_s\}, [\![\mathbf{u}_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
& - \mu \langle \nabla_{(x,y)} \mathbf{u}_s \cdot \mathbf{n}, \mathbf{v}_s \rangle_{\partial\Omega_s} + \mu \langle \nabla_{(x,y)} \mathbf{v}_s \cdot \mathbf{n}, \mathbf{u}_s \rangle_{\partial\Omega_s} \\
& + \langle \sigma_s [\![\mathbf{u}_s]\!], [\![\mathbf{v}_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \sigma_s \mathbf{u}_s, \mathbf{v}_s \rangle_{\partial\Omega_s}.
\end{aligned} \tag{24}$$

Here $\mathbf{v}^{\text{int}}$ and $\mathbf{v}^{\text{ext}}$ denote the traces of $\mathbf{v}$ from the interior and exterior of $\Omega_{e,s}$, respectively, and when $\partial\Omega_{e,s}$ belongs to the domain boundary, we take $\mathbf{u}_s^{\text{ext}} = \hat{\mathbf{u}}_s$. Also σ_s is a positive penalty parameter. Thus, by (6), (11), and (12)

$$\begin{aligned}
\mathcal{A}_s(\mathbf{u}_s, h_s; \mathbf{v}_s) = & -(g \nabla_{(x,y)} h_b, \mathbf{v}_s)_{\Omega_s} - \langle g h_s^I, \mathbf{v}_s \cdot \mathbf{n} \rangle_{\partial\Omega_{s,in}} \\
& + \mu \langle \nabla_{(x,y)} \mathbf{v}_s \cdot \mathbf{n}, \hat{\mathbf{u}}_s \rangle_{\partial\Omega_s} + \langle \sigma_s \hat{\mathbf{u}}_s, \mathbf{v}_s \rangle_{\partial\Omega_s} + (\mathcal{F}, \mathbf{v}_s)_{\Omega_s}.
\end{aligned} \tag{25}$$

The LDG/DG method can then be stated as follows. We seek approximations $\boldsymbol{U}_g(\cdot,t) \in \mathcal{V}_{\Delta_g,g}$, $H_g(\cdot,t) \in \mathcal{W}_{\Delta_g,g}$, $\boldsymbol{U}_s(\cdot,t) \in \mathcal{V}_{\Delta_s,s}$, and $H_s(\cdot,t) \in \mathcal{W}_{\Delta_s,s}$ which satisfy for each $t > 0$

$$\mathcal{A}_g(\boldsymbol{U}_g, H_g, H_s; \mathbf{v}_g) = -\langle h_b, \mathbf{v}_g \cdot \mathbf{n}_g \rangle_{\Gamma_{GS}} - \langle h_g^D, \mathbf{v}_g \cdot \mathbf{n}_g \rangle_{\Gamma_D}, \quad \mathbf{v}_g \in \mathcal{V}_{\Delta_g,g}, \tag{26}$$

$$\begin{aligned}
\mathcal{B}_g(\boldsymbol{U}_g, H_g, H_s; w_g) = & -\langle u_N, w_g \rangle_{\Gamma_N} + \langle \sigma_g h_b, w_g \rangle_{\Gamma_{GS}} \\
& + \langle \sigma_g h_g^D, w_g \rangle_{\Gamma_D} + (f_g, w_g)_{\Omega_g}, \quad w_g \in \mathcal{W}_{\Delta_g,g},
\end{aligned} \tag{27}$$

$$\mathcal{B}_s(\boldsymbol{U}_g, \boldsymbol{U}_s, H_g, H_s; w_s) = -\langle \sigma_g h_b, w_s \rangle_{\Gamma_{GS}}, \quad w_s \in \mathcal{W}_{\Delta_s,s}, \tag{28}$$

$$\begin{aligned}
\mathcal{A}_s(\boldsymbol{U}_s, H_s; \mathbf{v}_s) = & -(g \nabla_{(x,y)} h_b, \mathbf{v}_s)_{\Omega_s} - \langle g h_s^I, \mathbf{v}_s \cdot \mathbf{n} \rangle_{\partial\Omega_{s,in}} \\
& + \mu \langle \nabla_{(x,y)} \mathbf{v}_s \cdot \mathbf{n}, \hat{\mathbf{u}}_s \rangle_{\partial\Omega_s} + \langle \sigma_s \hat{\mathbf{u}}_s, \mathbf{v}_s \rangle_{\partial\Omega_s} \\
& + (\mathcal{F}, \mathbf{v}_s)_{\Omega_s}, \quad \mathbf{v}_s \in \mathcal{V}_{\Delta_s,s}.
\end{aligned} \tag{29}$$

Furthermore, $H_s(\cdot,0)$ and $\boldsymbol{U}_s(\cdot,0)$ are defined to be the L^2 projections of h_s^0 and $\mathbf{u}_s^0$, respectively, defined by

$$(H_s(\cdot,0) - h_s^0, w_s)_{\Omega_s} = 0, \quad w_s \in \mathcal{W}_{\Delta_s,s}, \tag{30}$$

$$(\boldsymbol{U}_s(\cdot,0) - \mathbf{u}_s^0, \mathbf{v}_s)_{\Omega_s} = 0, \quad \mathbf{v}_s \in \mathcal{V}_{\Delta_s,s}. \tag{31}$$

Computationally, and in the analysis below, we redefine the sets $\partial\Omega_{s,in}$ and $\partial\Omega_{s,out}$ in (7) and (8) to coincide with where $\boldsymbol{U}_s \cdot \mathbf{n}_s < 0$ and $\boldsymbol{U}_s \cdot \mathbf{n}_s \geq 0$, respectively.

4. An a priori error estimate. In this section, we derive an a priori error estimate for the LDG/DG method (26)–(31).

To begin the estimate, we define $\pi \mathbf{u}_g \in \mathcal{V}_{\Delta_g,g}$, $\pi h_g \in \mathcal{W}_{\Delta_g,g}$, $\pi \mathbf{u}_s \in \mathcal{V}_{\Delta_s,s}$, and $\pi h_s \in \mathcal{W}_{\Delta_s,s}$ to be projections of the true solution into the approximating spaces. These projections will be specified below. Next, we define $\Psi_g = \boldsymbol{U}_g - \pi \mathbf{u}_g$, $\Theta_g = \mathbf{u}_g - \pi \mathbf{u}_g$, $\kappa_g = H_g - \pi h_g$, and $\eta_g = h_g - \pi h_g$ with similar definitions for κ_s, η_s, Ψ_s, and Θ_s. By linearity and Galerkin orthogonality, we have

$$\mathcal{A}_g(\Psi_g, \kappa_g, \kappa_s; \mathbf{v}_g) = \mathcal{A}_g(\Theta_g, \eta_g, \eta_s; \mathbf{v}_g), \tag{32}$$

$$\mathcal{B}_g(\Psi_g, \kappa_g, \kappa_s; w_g) = \mathcal{B}_g(\Theta_g, \eta_g, \eta_s; w_g). \tag{33}$$

Define

$$\mathcal{B}_{s,L}(\mathbf{u}_g, h_g, h_s; w_s) \equiv (\partial_t h_s, w_s)_{\Omega_s} - \langle \sigma_g(h_g - h_s), w_s \rangle_{\Gamma_{GS}} - \langle \mathbf{u}_g \cdot \mathbf{n}_g, w_s \rangle_{\Gamma_{GS}}.$$

Then from (23) and (28),

$$\begin{aligned}
&\mathcal{B}_{s,L}(\Psi_g, \kappa_g, \kappa_s; w_s) - (\boldsymbol{U}_s \kappa_s, \nabla_{(x,y)} w_s)_{\Omega_s} + \langle \{\boldsymbol{U}_s\} \kappa_s^{\uparrow}, [\![w_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \boldsymbol{U}_s \cdot \mathbf{n}_s \kappa_s, w_s \rangle_{\partial\Omega_{s,out}} \\
&= \mathcal{B}_{s,L}(\Theta_g, \eta_g, \eta_s; w_s) - (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s, \nabla_{(x,y)} w_s)_{\Omega_s} + \langle \mathbf{u}_s h_s - \{\boldsymbol{U}_s\} \pi h_s^{\uparrow}, [\![w_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
&\quad + \langle (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s) \cdot \mathbf{n}_s, w_s \rangle_{\partial\Omega_{s,out}} + \langle (\mathbf{u}_s - \boldsymbol{U}_s) \cdot \mathbf{n}_s h_s^I, w_s \rangle_{\partial\Omega_{s,in}}.
\end{aligned} \tag{34}$$

Define

$$\begin{aligned}
\mathcal{A}_{s,L}(\mathbf{u}_s, h_s; \mathbf{v}_s) &\equiv (\partial_t \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} + (\tau_{bf} \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} \\
&\quad + (g \nabla_{(x,y)} h_s, \mathbf{v}_s)_{\Omega_s} - \langle g [\![h_s]\!], \{\mathbf{v}_s\} \rangle_{\mathcal{E}_{i,s}} - \langle g h_s, \mathbf{v}_s \cdot \mathbf{n} \rangle_{\partial\Omega_{s,in}} \\
&\quad + \mu (\nabla_{(x,y)} \mathbf{u}_s, \nabla_{(x,y)} \mathbf{v}_s)_{\Omega_s} - \mu \langle \{\nabla_{(x,y)} \mathbf{u}_s\}, [\![\mathbf{v}_s]\!] \rangle_{\mathcal{E}_{i,s}} + \mu \langle \{\nabla_{(x,y)} \mathbf{v}_s\}, [\![\mathbf{u}_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
&\quad - \mu \langle \nabla_{(x,y)} \mathbf{u}_s \cdot \mathbf{n}, \mathbf{v}_s \rangle_{\partial\Omega_s} + \mu \langle \nabla_{(x,y)} \mathbf{v}_s \cdot \mathbf{n}, \mathbf{u}_s \rangle_{\partial\Omega_s} \\
&\quad + \langle \sigma_s [\![\mathbf{u}_s]\!], [\![\mathbf{v}_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \sigma_s \mathbf{u}_s, \mathbf{v}_s \rangle_{\partial\Omega_s}.
\end{aligned} \tag{35}$$

Then from (25) and (29) (recalling $\mathbf{u}_s^{\text{int}} = \mathbf{u}_s^{\text{ext}}$),

$$\begin{aligned}
&\mathcal{A}_{s,L}(\Psi_s, \kappa_s; \mathbf{v}_s) \\
&= \mathcal{A}_{s,L}(\Theta_s, \eta_s; \mathbf{v}_s) + (\mathbf{u}_s \cdot \nabla_{(x,y)} \mathbf{u}_s - \boldsymbol{U}_s \cdot \nabla_{(x,y)} \boldsymbol{U}_s, \mathbf{v}_s)_{\Omega_s} \\
&\quad - \sum_{\partial\Omega_{e,s}^- \in \mathcal{T}_{\Delta_s,s}} \langle |\{\boldsymbol{U}_s\} \cdot \mathbf{n}_e| (\boldsymbol{U}_s^{\text{int}} - \boldsymbol{U}_s^{\text{ext}}), \mathbf{v}_s^{\text{int}} \rangle_{\partial\Omega_{e,s}^-}.
\end{aligned} \tag{36}$$

We now set $\mathbf{v}_g = \Psi_g$, $w_g = \kappa_g$, $\mathbf{v}_s = \Psi_s$, and $w_s = \kappa_s$ above and manipulate several of the resulting terms.

First, in (36), integrating by parts,

$$\begin{aligned}
&(g \nabla_{(x,y)} \kappa_s, \Psi_s)_{\Omega_s} - \langle g [\![\kappa_s]\!], \{\Psi_s\} \rangle_{\mathcal{E}_{i,s}} - \langle g \kappa_s, \Psi_s \cdot \mathbf{n}_s \rangle_{\partial\Omega_{s,in}} \\
&= -(g \kappa_s, \nabla_{(x,y)} \cdot \Psi_s)_{\Omega_s} + \langle g \{\kappa_s\}, [\![\Psi_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle g \kappa_s, \Psi_s \cdot \mathbf{n}_s \rangle_{\partial\Omega_{s,out}}.
\end{aligned} \tag{37}$$

Thus (36) becomes

$$\begin{aligned}
&(\partial_t \Psi_s, \Psi_s)_{\Omega_s} + (\tau_{bf} \Psi_s, \Psi_s)_{\Omega_s} + \mu (\nabla_{(x,y)} \Psi_s, \nabla_{(x,y)} \Psi_s)_{\Omega_s} \\
&+ \langle \sigma_s [\![\Psi_s]\!], [\![\Psi_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \sigma_s \Psi_s, \Psi_s \rangle_{\partial\Omega_s} \\
&= \mathcal{A}_{s,L}(\Theta_s, \eta_s; \mathbf{v}_s) + (\mathbf{u}_s \cdot \nabla_{(x,y)} \mathbf{u}_s - \boldsymbol{U}_s \cdot \nabla_{(x,y)} \boldsymbol{U}_s, \mathbf{v}_s)_{\Omega_s} \\
&\quad - \sum_{\partial\Omega_{e,s}^- \in \mathcal{T}_{\Delta_s,s}} \langle |\{\boldsymbol{U}_s\} \cdot \mathbf{n}_e| (\boldsymbol{U}_s^{\text{int}} - \boldsymbol{U}_s^{\text{ext}}), \mathbf{v}_s^{\text{int}} \rangle_{\partial\Omega_{e,s}^-} \\
&\quad + (g \kappa_s, \nabla_{(x,y)} \cdot \Psi_s)_{\Omega_s} - \langle g \{\kappa_s\}, [\![\Psi_s]\!] \rangle_{\mathcal{E}_{i,s}} - \langle g \kappa_s, \Psi_s \cdot \mathbf{n}_s \rangle_{\partial\Omega_{s,out}}.
\end{aligned} \tag{38}$$

Next, in (34), integrating by parts and using the definitions of $\partial\Omega_{s,out}$ and $\partial\Omega_{s,in}$ yields

$$\begin{aligned}
&-(\boldsymbol{U}_s\kappa_s, \nabla_{(x,y)}\kappa_s)_{\Omega_s} + \langle\{\boldsymbol{U}_s\}\kappa_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle \boldsymbol{U}_s\cdot\mathbf{n}_s, \kappa_s^2\rangle_{\partial\Omega_{s,out}}\\
&= \frac{1}{2}(\nabla_{(x,y)}\cdot\boldsymbol{U}_s, \kappa_s^2)_{\Omega_s} - \frac{1}{2}\langle[\![\boldsymbol{U}_s\kappa_s^2]\!], 1\rangle_{\mathcal{E}_{i,s}} + \langle\{\boldsymbol{U}_s\}\kappa_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\quad + \frac{1}{2}\langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|, \kappa_s^2\rangle_{\partial\Omega_{s,in}} + \frac{1}{2}\langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|, \kappa_s^2\rangle_{\partial\Omega_{s,out}}. \qquad (39)
\end{aligned}$$

Using $[\![\mathbf{v}w]\!] = \{\mathbf{v}\}\cdot[\![w]\!] + [\![\mathbf{v}]\!]\{w\}$ and $\frac{1}{2}[\![w^2]\!] = [\![w]\!]\{w\}$ and the definition of $\kappa_s^{\uparrow}$, we obtain

$$\begin{aligned}
&-\frac{1}{2}\langle[\![\boldsymbol{U}_s\kappa_s^2]\!], 1\rangle_{\mathcal{E}_{i,s}} + \langle\{\boldsymbol{U}_s\}\kappa_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&= \langle\kappa_s^{\uparrow}[\![\kappa_s]\!]\{\boldsymbol{U}_s\} - \frac{1}{2}[\![\kappa_s^2]\!]\{\boldsymbol{U}_s\} - \frac{1}{2}\{\kappa_s^2\}[\![\boldsymbol{U}_s]\!], 1\rangle_{\mathcal{E}_{i,s}}\\
&= \langle(\kappa_s^{\uparrow} - \{\kappa_s\})\{\boldsymbol{U}_s\}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} - \frac{1}{2}\langle\{\kappa_s\}^2, [\![\boldsymbol{U}]\!]_s\}\rangle_{\mathcal{E}_{i,s}}\\
&= \frac{1}{2}\left[\langle|\{\boldsymbol{U}_s\}\cdot\mathbf{n}^-|, [\kappa_s^- - \kappa_s^+]^2\rangle_{\mathcal{E}_{i,s}} - \langle\{\kappa_s^2\}, [\![\boldsymbol{U}_s]\!]\rangle_{\mathcal{E}_{i,s}}\right]. \qquad (40)
\end{aligned}$$

Integration by parts gives

$$\begin{aligned}
&-(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s, \nabla_{(x,y)}\kappa_s)_{\Omega_s} + \langle(\mathbf{u}_s h_s - \{\boldsymbol{U}_s\}\pi h_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\qquad + \langle(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s)\cdot\mathbf{n}_s, \kappa_s\rangle_{\partial\Omega_{s,out}} + \langle(\mathbf{u}_s - \boldsymbol{U}_s)\cdot\mathbf{n}_s h_s^I, \kappa_s\rangle_{\partial\Omega_{s,in}}\\
&= (\nabla_{(x,y)}\cdot(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s), \kappa_s)_{\Omega_s} - \langle[\![\kappa_s(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s)]\!], 1\rangle_{\mathcal{E}_{i,s}}\\
&\qquad + \langle(\mathbf{u}_s h_s - \{\boldsymbol{U}_s\}\pi h_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle\boldsymbol{U}_s\cdot\mathbf{n}_s(\pi h_s - h_s^I), \kappa_s\rangle_{\partial\Omega_{s,in}}. \qquad (41)
\end{aligned}$$

Combining (39)–(41) with (34), we find

$$\begin{aligned}
&(\partial_t\kappa_s, \kappa_s)_{\Omega_s} - \langle\sigma_g(\kappa_g - \kappa_s), \kappa_s\rangle_{\Gamma_{GS}} - \langle\Psi_g\cdot\mathbf{n}_g, \kappa_s\rangle_{\Gamma_{GS}}\\
&\quad + \frac{1}{2}\left[\langle|\{\boldsymbol{U}_s\}\cdot\mathbf{n}^-|, [\kappa_s^- - \kappa_s^+]^2\rangle_{\mathcal{E}_{i,s}} + \langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|, \kappa_s^2\rangle_{\partial\Omega_{s,in}} + \langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|, \kappa_s^2\rangle_{\partial\Omega_{s,out}}\right]\\
&= \mathcal{B}_{s,L}(\Theta_g, \eta_g, \eta_s; \kappa_s) - \frac{1}{2}(\nabla_{(x,y)}\cdot\boldsymbol{U}_s, \kappa_s^2)_{\Omega_s} + \frac{1}{2}\langle\{\kappa_s^2\}, [\![\boldsymbol{U}_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\quad + (\nabla_{(x,y)}\cdot(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s), \kappa_s)_{\Omega_s} - \langle[\![\kappa_s(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s)]\!], 1\rangle_{\mathcal{E}_{i,s}}\\
&\quad + \langle(\mathbf{u}_s h_s - \{\boldsymbol{U}_s\}\pi h_s^{\uparrow}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle\boldsymbol{U}_s\cdot\mathbf{n}_s(\pi h_s - h_s^I), \kappa_s\rangle_{\partial\Omega_{s,in}}. \qquad (42)
\end{aligned}$$

Furthermore, by (32) and (33),

$$\begin{aligned}
&\mathcal{A}_g(\Psi_g, \kappa_g, \kappa_s; \Psi_g) + \mathcal{B}_g(\Psi_s, \kappa_g, \kappa_s; \kappa_g)\\
&= (K^{-1}\Psi_g, \Psi_g)_{\Omega_g} + \langle\kappa_s, \Psi_g\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} + \langle\sigma_g[\![\kappa_g]\!], [\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\quad + \langle\sigma_g(\kappa_g - \kappa_s), \kappa_g\rangle_{\Gamma_{GS}} + \langle\sigma_g\kappa_g, \kappa_g\rangle_{\Gamma_D}\\
&= \mathcal{A}_g(\Theta_g, \eta_g, \eta_s; \Psi_g) + \mathcal{B}_g(\Theta_g, \eta_g, \eta_s; \kappa_g). \qquad (43)
\end{aligned}$$

Adding (38), (42), and (43), we find the error equation

$$
\begin{aligned}
&||K^{-1/2}\Psi_g||^2_{\Omega_{\rm g}} + ||\sigma_g^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma_g^{1/2}\kappa_g||^2_{\Gamma_D} + ||\sigma_g^{1/2}(\kappa_g-\kappa_s)||^2_{\Gamma_{GS}}\\
&+\frac{1}{2}\left[\langle|\{\boldsymbol{U}_s\}\cdot\mathbf{n}^-|,[\kappa_s^- - \kappa_s^+]^2\rangle_{\mathcal{E}_{i,s}} + \langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|,\kappa_s^2\rangle_{\partial\Omega_{s,in}} + \langle|\boldsymbol{U}_s\cdot\mathbf{n}_s|,\kappa_s^2\rangle_{\partial\Omega_{s,out}}\right]\\
&+(\partial_t\kappa_s,\kappa_s)_{\Omega_s} + (\partial_t\Psi_s,\Psi_s)_{\Omega_s} + ||\tau_{bf}^{1/2}\Psi_s||^2_{\Omega_s} + \mu||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s}\\
&+||\sigma_s^{1/2}[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma_s^{1/2}\Psi_s||^2_{\partial\Omega_s}\\
&=\mathcal{A}_g(\Theta_g,\eta_g,\eta_s;\Psi_g) + \mathcal{B}_g(\Theta_g,\eta_g,\eta_s;\kappa_g) + \mathcal{B}_{s,L}(\Theta_g,\eta_g,\eta_s;\kappa_s) + \mathcal{A}_{s,L}(\Theta_s,\eta_s;\Psi_s)\\
&\quad-\sum_{\partial\Omega^-_{e,s}\in\mathcal{T}_{\Delta_s,s}}\langle|\{\boldsymbol{U}_s\}\cdot\mathbf{n}_e|(\boldsymbol{U}_s^{\rm int}-\boldsymbol{U}_s^{\rm ext}),\Psi_s^{\rm int}\rangle_{\partial\Omega^-_{e,s}}\\
&\quad+(\mathbf{u}_s\cdot\nabla_{(x,y)}\mathbf{u}_s - \boldsymbol{U}_s\cdot\nabla_{(x,y)}\boldsymbol{U}_s,\Psi_s)_{\Omega_{\rm s}}\\
&\quad-\frac{1}{2}(\nabla_{(x,y)}\cdot\boldsymbol{U}_s,\kappa_s^2)_{\Omega_s} + \frac{1}{2}\langle\{\kappa_s^2\},[\![\boldsymbol{U}_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\quad+(\nabla_{(x,y)}\cdot(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s),\kappa_s)_{\Omega_{\rm s}} - \langle[\![\kappa_s(\mathbf{u}_s h_s - \boldsymbol{U}_s\pi h_s)]\!],1\rangle_{\mathcal{E}_{i,s}}\\
&\quad+\langle\mathbf{u}_s h_s - \{\boldsymbol{U}_s\}\pi h_s^{\uparrow},[\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle\boldsymbol{U}_s\cdot\mathbf{n}_s(\pi h_s - h_s^I),\kappa_s\rangle_{\partial\Omega_{s,in}}\\
&\quad+(g\kappa_s,\nabla_{(x,y)}\cdot\Psi_s)_{\Omega_s} - \langle g\{\kappa_s\},[\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} - \langle g\kappa_s,\Psi_s\cdot\mathbf{n}_s\rangle_{\partial\Omega_{s,out}}\\
&\equiv\mathcal{A}_g(\Theta_g,\eta_g,\eta_s;\Psi_g) + \mathcal{B}_g(\Theta_g,\eta_g,\eta_s;\kappa_g) + \mathcal{B}_{s,L}(\Theta_g,\eta_g,\eta_s;\kappa_s) + \mathcal{A}_{s,L}(\Theta_s,\eta_s;\Psi_s)\\
&+\sum_{i=1}^{11}E_i. \qquad (44)
\end{aligned}
$$

Next, we consider the terms on the right-hand side of (44). To begin note that

$$
\begin{aligned}
\mathcal{A}_g(\Theta_g,\eta_g,\eta_s;\Psi_g) &= (K^{-1}\Theta_g,\Psi_g)_{\Omega_{\rm g}} - (\eta_g,\nabla_{(x,y,z)}\cdot\Psi_g)_{\Omega_{\rm g}} + \langle\{\eta_g\},[\![\Psi_g]\!]\rangle_{\mathcal{E}_{i,g}}\\
&\quad+\langle\eta_g,\Psi_g\cdot\mathbf{n}_g\rangle_{\Gamma_N} + \langle\eta_s,\Psi_g\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}}\\
&\equiv A_{g,1}+\cdots+A_{g,5}. \qquad (45)
\end{aligned}
$$

Adding terms and integrating by parts, we find

$$
\begin{aligned}
&\mathcal{B}_g(\Theta_g,\eta_g,\eta_s;\kappa_g) + \mathcal{B}_{s,L}(\Theta_g,\eta_g,\eta_s;\kappa_s)\\
&=(\nabla_{(x,y,z)}\cdot\Theta_g,\kappa_g)_{\Omega_{\rm g}} - \langle[\![\Theta_g]\!],\{\kappa_g\}\rangle_{\mathcal{E}_{i,g}} + \langle\Theta_g\cdot\mathbf{n}_g,\kappa_g\rangle_{\Gamma_N}\\
&\quad+\langle\sigma_g[\![\eta_g]\!],[\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle\sigma_g(\eta_g-\eta_s),\kappa_g\rangle_{\Gamma_{GS}} + \langle\sigma_g\eta_g,\kappa_g\rangle_{\Gamma_D}\\
&\quad+(\partial_t\eta_s,\kappa_s)_{\Omega_{\rm s}} - \langle\sigma_g(\eta_g-\eta_s),\kappa_s\rangle_{\Gamma_{GS}} - \langle\Theta_g\cdot\mathbf{n}_g,\kappa_s\rangle_{\Gamma_{GS}}\\
&=(\Theta_g,\nabla_{(x,y,z)}\kappa_g)_{\Omega_{\rm g}} + \langle\{\Theta_g\},[\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle\Theta_g\cdot\mathbf{n}_g,\kappa_g-\kappa_s\rangle_{\Gamma_{GS}} + \langle\Theta_g\cdot\mathbf{n}_g,\kappa_g\rangle_{\Gamma_D}\\
&\quad+\langle\sigma_g[\![\eta_g]\!],[\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle\sigma_g(\eta_g-\eta_s),\kappa_g-\kappa_s\rangle_{\Gamma_{GS}} + \langle\sigma_g\eta_g,\kappa_g\rangle_{\Gamma_D} + (\partial_t\eta_s,\kappa_s)_{\Omega_{\rm s}}\\
&\equiv B_1+\cdots+B_8. \qquad (46)
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\mathcal{A}_{s,L}(\Theta_s,\eta_s;\Psi_s) = (\partial_t\Theta_s,\Psi_s)_{\Omega_{\rm s}} + (\tau_{bf}\Theta_s,\Psi_s)_{\Omega_{\rm s}}\\
&\quad+(g\nabla_{(x,y)}\eta_s,\Psi_s)_{\Omega_{\rm s}} - \langle g[\![\eta_s]\!],\{\Psi_s\}\rangle_{\mathcal{E}_{i,s}} - \langle g\eta_s,\Psi_s\cdot\mathbf{n}\rangle_{\partial\Omega_{s,in}}\\
&\quad+\mu(\nabla_{(x,y)}\Theta_s,\nabla_{(x,y)}\Psi_s)_{\Omega_{\rm s}} - \mu\langle\{\nabla_{(x,y)}\Theta_s\},[\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} + \mu\langle\{\nabla_{(x,y)}\Psi_s\},[\![\Theta_s]\!]\rangle_{\mathcal{E}_{i,s}}\\
&\quad-\mu\langle\nabla_{(x,y)}\Theta_s\cdot\mathbf{n},\Psi_s\rangle_{\partial\Omega_s} + \mu\langle\nabla_{(x,y)}\Psi_s\cdot\mathbf{n},\Theta_s\rangle_{\partial\Omega_s}\\
&\quad+\langle\sigma_s[\![\Theta_s]\!],[\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle\sigma_s\Theta_s,\Psi_s\rangle_{\partial\Omega_s}\\
&\equiv A_{s,1}+\cdots+A_{s,12}. \qquad (47)
\end{aligned}
$$

We choose $\pi\mathbf{u}_g$, πh_g, $\pi\mathbf{u}_s$, and πh_s to be the L^2 projections of $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, and h_s into the spaces $\mathcal{V}_{\Delta_g,g}$, $\mathcal{W}_{\Delta_g,g}$, $\mathcal{V}_{\Delta_s,s}$, and $\mathcal{W}_{\Delta_s,s}$, respectively. These are defined by

$$(\pi\mathbf{u}_g - \mathbf{u}_g, \mathbf{v}_g)_{\Omega_g} = 0, \quad \mathbf{v}_g \in \mathcal{V}_{\Delta_g,g}, \tag{48}$$

$$(\pi h_g - h_g, w_g)_{\Omega_g} = 0, \quad w_g \in \mathcal{W}_{\Delta_g,g}, \tag{49}$$

$$(\pi\mathbf{u}_s - \mathbf{u}_s, \mathbf{v}_s)_{\Omega_s} = 0, \quad \mathbf{v}_s \in \mathcal{V}_{\Delta_s,s}, \tag{50}$$

$$(\pi h_s - h_s, w_s)_{\Omega_s} = 0, \quad w_s \in \mathcal{W}_{\Delta_s,s}. \tag{51}$$

We note that by our assumptions on the approximating spaces,

$$\mathbf{v}_g \in \mathcal{V}_{\Delta_g,g} \Rightarrow \nabla_{(x,y,z)} \cdot \mathbf{v}_g \in \mathcal{W}_{\Delta_g,g}$$

and

$$w_g \in \mathcal{W}_{\Delta_g,g} \Rightarrow \nabla_{(x,y,z)} w_g \in \mathcal{V}_{\Delta_g,g}.$$

Thus by (48)–(51) terms $A_{g,2} = B_1 = B_8 = A_{s,1} = 0$.

In the estimates below, we will use the trace theorem [9].

THEOREM 4.1. *Suppose that region R has a Lipschitz boundary. Then there exists a constant K^t such that*

$$||v||^2_{\partial R} \le K^t[\nu^{-1}||v||^2_R + \nu||\nabla v||^2_R] \quad \forall v \in H^1(R), \tag{52}$$

where $\nu = \operatorname{diam}(R)$.

We will also use the standard inverse inequality:

$$||\nabla w||_{\Omega_e} \le K^i \Delta_e^{-1} ||w||_{\Omega_e}, \tag{53}$$

where Ω_e is in Ω_g or Ω_s and w is in one of our finite-dimensional approximating spaces. Based on these results, for $v \in H^1(\Omega_e)$ and w in one of our approximating spaces, an argument we will use repeatedly in various ways is

$$\begin{aligned} \langle v^{\text{int}}, w^{\text{int}} \rangle_{\partial\Omega_e} &\le ||v^{\text{int}}||_{\partial\Omega_e} ||w^{\text{int}}||_{\partial\Omega_e} \\ &\le K^t[\Delta_e^{-1}||v||^2_{\Omega_e} + \Delta_e||\nabla v||^2_{\Omega_e}]^{1/2}[\Delta_e^{-1}||w||^2_{\Omega_e} + \Delta_e||\nabla w||^2_{\Omega_e}]^{1/2} \\ &\le K^t(1 + (K^i)^2)^{1/2}[\Delta_e^{-2}||v||^2_{\Omega_e} + ||\nabla v||^2_{\Omega_e}]^{1/2}||w||_{\Omega_e}. \end{aligned} \tag{54}$$

We will make the following assumptions on our solution and projections:

$$\begin{aligned} &||\mathbf{u}_s||_{L^\infty(0,T;W_1^\infty(\Omega_s))} + ||h_s||_{L^\infty(0,T;W_1^\infty(\Omega_s))} \\ &+||\pi\mathbf{u}_s||_{L^\infty(0,T;L^\infty(\Omega_s))} + ||\pi h_s||_{L^\infty(0,T;L^\infty(\Omega_s))} \le K^m, \end{aligned} \tag{55}$$

where K^m is independent of Δ_s. We will also assume that a constant $K^M \ge 2K^m$ exists, independent of Δ_s, for which

$$||\Psi_s||_{L^\infty(0,T;L^\infty(\Omega_s))} + ||\kappa_s||_{L^\infty(0,T;L^\infty(\Omega_s))} \le K^M. \tag{56}$$

We will show inductively that for $k_s, k_g > 1$ our estimate does not in fact depend on K^M.

On any face γ in the mesh $\mathcal{T}_{\Delta_g,g}(\mathcal{T}_{\Delta_s,s})$, let E_γ denote the set of elements sharing the face, and let Δ_γ be the maximum element diameter over all elements in E_γ. We will assume that

$$\sigma_{g(s)}|_\gamma = \mathcal{O}(\Delta_\gamma^{-1}). \tag{57}$$

Finally, we make the following assumption on the ground water and surface water meshes on Γ_{GS}.

Assumption GS. Let the set E_{GS} be defined as follows:

$$E_{GS} = \{\Omega_{e,g} \subset \Omega_{\rm g} : \partial\Omega_{e,g} \cap \Gamma_{GS} \neq \emptyset\}. \tag{58}$$

For elements $\Omega_{e,g}$ in the set E_{GS}, assume that $\partial\Omega_{e,g} \cap \Gamma_{GS}$, when mapped to $\Omega_{\rm s}$, intersects a finite number of elements $\Omega_{\bar{e},s}$ bounded independently of Δ_g or Δ_s and that

$$\Delta_{\bar{e},s}\Delta_{e,g}^{-1} = \mathcal{O}(1).$$

We repeatedly make use of Young's inequality

$$ab \leq \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2, \quad a,b \in \mathbb{R}, \epsilon > 0. \tag{59}$$

Furthermore, let $\mathcal{C}$ denote a generic positive constant which may depend on other constants. We will make explicit this dependence in the arguments below.

First, it is easily seen that

$$A_{g,1} \leq \mathcal{C}(K)||\Theta_g||^2_{\Omega_{\rm g}} + \epsilon_1||K^{-1/2}\Psi_g||^2_{\Omega_{\rm g}}. \tag{60}$$

Applying (54), we find

$$\begin{aligned} A_{g,3} + A_{g,4} &= \langle [\![\Psi_g]\!], \{\eta_g\}\rangle_{\mathcal{E}_{i,g}} + \langle \eta_g, \Psi_g \cdot \mathbf{n}_g\rangle_{\Gamma_N} \\ &\leq \epsilon_1||K^{-1/2}\Psi_g||^2_{\Omega_{\rm g}} + \mathcal{C}(K^t, K^i, K)\sum_{\Omega_{e,g}}[\Delta_{e,g}^{-2}||\eta_g||^2_{\Omega_{e,g}} + ||\eta_g||^2_{H^1(\Omega_{e,g})}]. \end{aligned} \tag{61}$$

Applying Theorem 4.1 to Ψ_g and Assumption GS,

$$\begin{aligned} A_{g,5} &= \langle \Psi_g \cdot \mathbf{n}_g, \eta_s\rangle_{\Gamma_{GS}} \\ &\leq \mathcal{C}\sum_{\Omega_e \in E_{GS}} ||\Psi_g||_{\partial\Omega_{e,g}}||\eta_s||_{\partial\Omega_{e,g}} \\ &\leq \epsilon_1||K^{-1/2}\Psi_g||^2_{\Omega_{\rm g}} + \mathcal{C}(K^t, K^i, K)\sum_{\Omega_{e,g}\in E_{GS}} \Delta_{e,g}^{-1}||\eta_s||^2_{\partial\Omega_{e,g}} \\ &\leq \epsilon_1||K^{-1/2}\Psi_g||^2_{\Omega_{\rm g}} + \mathcal{C}(K^t, K^i, K)\sum_{\Omega_{e,s}} \Delta_{e,s}^{-1}||\eta_s||^2_{\Omega_{e,s}}. \end{aligned} \tag{62}$$

Next, by (57) and Theorem 4.1,

$$\begin{aligned} B_2 + B_3 + B_4 &= \langle \{\Theta_g\}, [\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \Theta_g \cdot \mathbf{n}_g, \kappa_g - \kappa_s\rangle_{\Gamma_{GS}} + \langle \Theta_g \cdot \mathbf{n}_g, \kappa_g\rangle_{\Gamma_D} \\ &= \langle \sigma^{-1/2}\{\Theta_g\}, \sigma^{1/2}[\![\kappa_g]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \sigma^{-1/2}\Theta_g \cdot \mathbf{n}_g, \sigma^{1/2}(\kappa_g - \kappa_s)\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma^{-1/2}\Theta_g \cdot \mathbf{n}_g, \sigma^{1/2}\kappa_g\rangle_{\Gamma_D} \\ &\leq \mathcal{C}[\,||\sigma^{-1/2}\Theta_g||^2_{\Gamma_D} + ||\sigma^{-1/2}\Theta_g||^2_{\mathcal{E}_{i,g}} + ||\sigma^{-1/2}\Theta_g||^2_{\Gamma_{GS}}] \\ &\quad + \epsilon_2[\,||\sigma^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma^{1/2}(\kappa_g - \kappa_s)||^2_{\Gamma_{GS}} + ||\sigma^{1/2}\kappa_g||^2_{\Gamma_D}] \\ &\leq \mathcal{C}(K^t)\sum_{\Omega_{e,g}}[\,||\Theta_g||^2_{\Omega_{e,g}} + \Delta^2_{e,g}||\Theta_g||^2_{H^1(\Omega_{e,g})}] \\ &\quad + \epsilon_2[\,||\sigma^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma^{1/2}(\kappa_g - \kappa_s)||^2_{\Gamma_{GS}} + ||\sigma^{1/2}\kappa_g||^2_{\Gamma_D}]. \end{aligned} \tag{63}$$

Similarly,

$$\begin{aligned} B_5 + B_6 + B_7 \leq \mathcal{C}[\ & ||\sigma^{1/2}[\![\eta_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma^{1/2}\eta_g||^2_{\Gamma_{GS}} \\ & + ||\sigma^{1/2}\eta_s||^2_{\Gamma_{GS}} + ||\sigma^{1/2}\eta_g||^2_{\Gamma_D}] \\ & + \epsilon_2[\ ||\sigma^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma^{1/2}(\kappa_g - \kappa_s)||^2_{\Gamma_{GS}} \\ & + ||\sigma^{1/2}\kappa_g||^2_{\Gamma_D}] \\ \leq \mathcal{C}(K^t) & \sum_{\Omega_{e,g}} [\Delta^{-2}_{e,g}||\eta_g||^2_{\Omega_{e,g}} + ||\eta_g||^2_{H^1(\Omega_{e,g})}] \\ & + \epsilon_2[\ ||\sigma^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma^{1/2}(\kappa_g - \kappa_s)||^2_{\Gamma_{GS}} \\ & + ||\sigma^{1/2}\kappa_g||^2_{\Gamma_D}]. \end{aligned} \tag{64}$$

Next, consider

$$\begin{aligned} A_{s,2} + A_{s,3} + A_{s,6} &= (\tau_{bf}\Theta_s, \Psi_s)_{\Omega_s} + (g\nabla_{(x,y)}\eta_s, \Psi_s)_{\Omega_s} + \mu(\nabla_{(x,y)}\Theta_s, \nabla_{(x,y)}\Psi_s)_{\Omega_s} \\ &\leq \mathcal{C}[\ ||\tau^{1/2}_{bf}\Theta_s||^2_{\Omega_s} + ||\nabla_{(x,y)}\eta_s||^2_{\Omega_s} + \mu||\nabla_{(x,y)}\Theta_s||^2_{\Omega_s}] \\ &\quad + \frac{1}{2}||\tau^{1/2}_{bf}\Psi_s||^2_{\Omega_s} + \mathcal{C}||\Psi_s||^2_{\Omega_s} + \frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s}. \end{aligned} \tag{65}$$

By (54),

$$\begin{aligned} A_{s,4} + A_{s,5} + A_{s,8} + A_{s,10} = & -\langle g[\![\eta_s]\!], \{\Psi_s\}\rangle_{\mathcal{E}_{i,s}} - \langle g\eta_s, \Psi_s \cdot \mathbf{n}\rangle_{\partial\Omega_{s,in}} \\ & + \mu\langle\{\nabla_{(x,y)}\Psi_s\}, [\![\Theta_s]\!]\rangle_{\mathcal{E}_{i,s}} + \mu\langle\nabla_{(x,y)}\Psi_s \cdot \mathbf{n}, \Theta_s\rangle_{\partial\Omega_s} \\ \leq \mathcal{C}||\Psi_s||^2_{\Omega_s} & + \frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s} \\ & + \mathcal{C}(K^t, K^i)\sum_{\Omega_{e,s}} [\Delta^{-2}_{e,s}||\eta_s||^2_{\Omega_{e,s}} + ||\eta_s||^2_{H^1(\Omega_{e,s})}] \\ & + \mathcal{C}(K^t, K^i, \mu)\sum_{\Omega_{e,s}} [\Delta^{-2}_{e,s}||\Theta_s||^2_{\Omega_{e,s}} + ||\Theta_s||^2_{H^1(\Omega_{e,s})}]. \end{aligned} \tag{66}$$

By (57) and Theorem 4.1,

$$\begin{aligned} A_{s,7} + A_{s,9} &= -\mu\langle\{\nabla_{(x,y)}\Theta_s\}, [\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} - \mu\langle\nabla_{(x,y)}\Theta_s \cdot \mathbf{n}, \Psi_s\rangle_{\partial\Omega_s} \\ &\leq \mathcal{C}(\mu)[\ ||\sigma^{-1/2}\{\nabla_{(x,y)}\Theta_s\}||^2_{\mathcal{E}_{i,s}} + ||\sigma^{-1/2}\nabla_{(x,y)}\Theta_s||^2_{\partial\Omega_s}] \\ &\quad + \epsilon_3[\ ||\sigma^{1/2}_s[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma^{1/2}_s\Psi_s||^2_{\partial\Omega_s}] \\ &\leq \mathcal{C}(K^t, \mu)\sum_{\Omega_{e,s}}[\ ||\nabla_{(x,y)}\Theta_s||^2_{\Omega_{e,s}} + \Delta^2_{e,s}||\nabla_{(x,y)}\Theta_s||^2_{H^1(\Omega_{e,s})}] \\ &\quad + \epsilon_3[\ ||\sigma^{1/2}_s[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma^{1/2}_s\Psi_s||^2_{\partial\Omega_s}]. \end{aligned} \tag{67}$$

Furthermore,

$$\begin{aligned} A_{s,11} + A_{s,12} &= \langle\sigma_s[\![\Theta_s]\!], [\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} + \langle\sigma_s\Theta_s, \Psi_s\rangle_{\partial\Omega_s} \\ &\leq \epsilon_3[\ ||\sigma^{1/2}_s[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma^{1/2}_s\Psi_s||^2_{\partial\Omega_s}] \\ &\quad + \mathcal{C}[\ ||\sigma^{1/2}_s[\![\Theta_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma^{1/2}_s\Theta_s||^2_{\partial\Omega_s}] \\ &\leq \epsilon_3[\ ||\sigma^{1/2}_s[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma^{1/2}_s\Psi_s||^2_{\partial\Omega_s}] \\ &\quad + \mathcal{C}(K^t)\sum_{\Omega_{e,s}}[\ \Delta^{-2}_{e,s}||\Theta_s||^2_{\Omega_{e,s}} + ||\Theta_s||^2_{H^1(\Omega_{e,s})}]. \end{aligned} \tag{68}$$

Consider next

$$
\begin{aligned}
E_1 &= -\sum_{\partial\Omega^-_{e,s}\in\mathcal{T}_{\Delta_s},s} \langle |\{\boldsymbol{U}_s\}\cdot\mathbf{n}_e|(\boldsymbol{U}_s^{\text{int}}-\boldsymbol{U}_s^{\text{ext}}),\Psi_s^{\text{int}}\rangle_{\partial\Omega^-_{e,s}} \\
&\le \sum_{\partial\Omega^-_{e,s}\in\mathcal{T}_{\Delta_s},s} \langle |\{\Psi_s+\pi\mathbf{u}_s\}\cdot\mathbf{n}_e||[\![\Psi_s]\!]-[\![\Theta_s]\!]|,|\Psi_s^{\text{int}}|\rangle_{\partial\Omega^-_{e,s}} \\
&\le \mathcal{C}(K^M,K^m)\sum_{\partial\Omega^-_{e,s}\in\mathcal{T}_{\Delta_s},s}[\ ||\sigma_s^{1/2}[\![\Theta_s]\!]||_{\partial\Omega^-_{e,s}}+||\sigma_s^{1/2}[\![\Psi_s]\!]||_{\partial\Omega^-_{e,s}}]||\sigma_s^{-1/2}\Psi_s^{\text{int}}||_{\partial\Omega^-_{e,s}} \\
&\le \mathcal{C}(K^t,K^m,K^M)\sum_{\Omega_{e,s}}[\Delta_{e,s}^{-2}||\Theta_s||^2_{\Omega_{e,s}}+||\Theta_s||^2_{H^1(\Omega_{e,s})}]+\mathcal{C}(K^M,K^m,K^t,K^i)||\Psi_s||^2_{\Omega_s} \\
&\quad +\epsilon_3[\ ||\sigma_s^{1/2}[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}}+||\sigma_s^{1/2}\Psi_s||^2_{\partial\Omega_s}], \qquad (69)
\end{aligned}
$$

$$
\begin{aligned}
E_2 &= (\mathbf{u}_s\cdot\nabla_{(x,y)}\mathbf{u}_s-\boldsymbol{U}_s\cdot\nabla_{(x,y)}\boldsymbol{U}_s,\Psi_s)_{\Omega_s} \\
&= ((\mathbf{u}_s-\boldsymbol{U}_s)\cdot\nabla_{(x,y)}\mathbf{u}_s,\Psi_s)_{\Omega_s}+((\Psi_s+\pi\mathbf{u}_s)\cdot\nabla_{(x,y)}(\Theta_s-\Psi_s),\Psi_s)_{\Omega_s} \\
&\le \mathcal{C}(K^m,K^M)[\ ||\Theta_s||^2_{\Omega_s}+||\nabla_{(x,y)}\Theta_s||^2_{\Omega_s}]+\mathcal{C}(K^m,K^M,\mu^{-1})||\Psi_s||^2_{\Omega_s} \\
&\quad +\frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2, \qquad (70)
\end{aligned}
$$

$$
\begin{aligned}
E_3 &= -\frac{1}{2}(\nabla_{(x,y)}\cdot\boldsymbol{U}_s,\kappa_s^2)_{\Omega_s} \\
&= -\frac{1}{2}(\nabla_{(x,y)}\cdot(\Psi_s-\Theta_s+\mathbf{u}_s),\kappa_s^2)_{\Omega_s} \\
&\le \frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s}+\mathcal{C}(K^M)||\nabla_{(x,y)}\Theta_s||^2_{\Omega_s}+\mathcal{C}(K^m,K^M,\mu^{-1})||\kappa_s||^2_{\Omega_s}, \qquad (71)
\end{aligned}
$$

$$
\begin{aligned}
E_4 &= \frac{1}{2}\langle\{\kappa_s^2\},[\![\boldsymbol{U}_s]\!]\rangle_{\mathcal{E}_{i,s}} \\
&= \frac{1}{2}\langle\{\kappa_s^2\},[\![\Psi_s-\Theta_s]\!]\rangle_{\mathcal{E}_{i,s}} \\
&\le \mathcal{C}(K^M)[\ ||\sigma_s^{1/2}[\![\Psi_s]\!]||_{\mathcal{E}_{i,s}}+||\sigma_s^{1/2}[\![\Theta_s]\!]||_{\mathcal{E}_{i,s}}]\ ||\sigma_s^{-1/2}\{\kappa_s\}||_{\mathcal{E}_{i,s}} \\
&\le \epsilon_3||\sigma_s^{1/2}[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}}+\mathcal{C}(K^t)\sum_{\Omega_{e,s}}[\ \Delta_{e,s}^{-2}||\Theta_s||^2_{\Omega_{e,s}}+||\Theta_s||^2_{H^1(\Omega_{e,s})}] \\
&\quad +\mathcal{C}(K^M,K^t,K^i)||\kappa_s||^2_{\Omega_s}, \qquad (72)
\end{aligned}
$$

$$
\begin{aligned}
E_5 &= (\nabla_{(x,y)}\cdot(\mathbf{u}_sh_s-\boldsymbol{U}_s\pi h_s),\kappa_s)_{\Omega_s} \\
&= (\nabla_{(x,y)}\cdot(\mathbf{u}_s(h_s-\pi h_s)),\kappa_s)_{\Omega_s}+(\nabla_{(x,y)}\cdot((\mathbf{u}_s-\boldsymbol{U}_s)\pi h_s),\kappa_s)_{\Omega_s} \\
&= ((\nabla_{(x,y)}\cdot\mathbf{u}_s)\eta_s,\kappa_s)_{\Omega_s}+(\mathbf{u}_s\cdot\nabla_{(x,y)}\eta_s,\kappa_s)_{\Omega_s} \\
&\quad +(\pi h_s\nabla_{(x,y)}\cdot(\Theta_s-\Psi_s),\kappa_s)_{\Omega_s}+((\Theta_s-\Psi_s)\cdot\nabla_{(x,y)}(h_s-\eta_s),\kappa_s)_{\Omega_s} \\
&\le \mathcal{C}(K^m,K^M)||\Theta_s||^2_{H^1(\Omega_s)}+\mathcal{C}(K^m,K^M)||\eta_s||^2_{H^1(\Omega_s)}+\frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s} \\
&\quad +\mathcal{C}(K^m,K^M,\mu^{-1})||\kappa_s||^2_{\Omega_s}. \qquad (73)
\end{aligned}
$$

Using $[\![ab]\!]=\{a\}[\![b]\!]+[\![a]\!]\{b\}$ and $\{ab\}=\{a\}\{b\}+\frac{1}{4}[\![a]\!][\![b]\!]$, we find

$$
\begin{aligned}
E_6+E_7 &= -\langle[\![\kappa_s(\mathbf{u}_sh_s-\boldsymbol{U}_s\pi h_s)]\!],1\rangle_{\mathcal{E}_{i,s}}+\langle\mathbf{u}_sh_s-\{\boldsymbol{U}_s\}\pi h_s^{\uparrow},[\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} \\
&= \langle(\eta_s^{\uparrow}-\{\eta_s\})\{\boldsymbol{U}_s\}-\frac{1}{4}[\![\eta_s]\!][\![\boldsymbol{U}_s]\!],[\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} \\
&\quad +\langle\{\pi h_s\}([\![\Psi_s]\!]-[\![\Theta_s]\!])-[\![\eta_s]\!]\{\boldsymbol{U}_s\},\{\kappa_s\}\rangle_{\mathcal{E}_{i,s}}. \qquad (74)
\end{aligned}
$$

By (54)

$$\langle(\eta_s^{\uparrow} - \{\eta_s\})\{\boldsymbol{U}_s\}, [\![\kappa_s]\!]\rangle_{\mathcal{E}_{i,s}} \leq \mathcal{C}(K^m, K^M) \sum_{\Omega_{e,s}} [\ \Delta_{e,s}^{-2} ||\eta_s||^2_{\Omega_{e,s}} + ||\eta_s||^2_{H^1(\Omega_{e,s})}]$$

$$+C(K^i)||\kappa_s||^2_{\Omega_s} \tag{75}$$

with similar estimates for the above terms involving $[\![\eta_s]\!]$. An identical bound to (72) is obtained for the remaining term, $\langle\{\pi h_s\}([\![\Psi_s]\!] - [\![\Theta_s]\!], \{\kappa_s\}\rangle_{\mathcal{E}_{i,s}}$, in (74). Continuing,

$$\begin{aligned} E_8 &= \langle \boldsymbol{U}_s \cdot \mathbf{n}_s(\pi h_s - h_s^I), \kappa_s\rangle_{\partial\Omega_{s,in}} \\ &\leq \mathcal{C}(K^m, K^M)||\eta_s||^2_{\partial\Omega_{s,in}} + \epsilon_4\langle|\boldsymbol{U}_s \cdot \mathbf{n}_s|, |\kappa_s|^2\rangle_{\partial\Omega_{s,in}} \\ &\leq \mathcal{C}(K^m, K^M, K^t) \sum_{\Omega_{e,s}} [\ \Delta_{e,s}^{-1} ||\eta_s||^2_{\Omega_{e,s}} + \Delta_{e,s}||\eta_s||^2_{H^1(\Omega_{e,s})}] \\ &\quad +\epsilon_4\langle|\boldsymbol{U}_s \cdot \mathbf{n}_s|, |\kappa_s|^2\rangle_{\partial\Omega_{s,in}} \end{aligned} \tag{76}$$

and

$$\begin{aligned} E_9 &= (g\kappa_s, \nabla_{(x,y)} \cdot \Psi_s)_{\Omega_s} \\ &\leq \mathcal{C}(\mu^{-1})||\kappa_s||^2_{\Omega_s} + \frac{\mu}{8}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s}. \end{aligned} \tag{77}$$

Finally,

$$\begin{aligned} E_{10} + E_{11} &= -\langle g\{\kappa_s\}, [\![\Psi_s]\!]\rangle_{\mathcal{E}_{i,s}} - \langle g\kappa_s, \Psi_s \cdot \mathbf{n}_s\rangle_{\partial\Omega_{s,out}} \\ &\leq \mathcal{C}||\sigma_s^{-1/2}\{\kappa_s\}||_{\mathcal{E}_{i,s}}||\sigma_s^{1/2}[\![\Psi_s]\!]||_{\mathcal{E}_{i,s}} \\ &\quad +\mathcal{C}||\sigma_s^{-1/2}\kappa_s||_{\partial\Omega_{s,out}}||\sigma_s^{1/2}\Psi_s||_{\partial\Omega_{s,out}} \\ &\leq \mathcal{C}(K^t, K^i)||\kappa_s||^2_{\Omega_s} \\ &\quad +\epsilon_3[\ ||\sigma_s^{1/2}[\![\Psi_s]\!]||^2_{\mathcal{E}_{i,s}} + ||\sigma_s^{1/2}\Psi_s||^2_{\partial\Omega_{s,out}}]. \end{aligned} \tag{78}$$

By standard approximation theory, for $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, and h_s sufficiently smooth,

$$||\Theta_g||_{\Omega_{e,g}} + \Delta_{e,g}||\Theta_g||_{H^1(\Omega_{e,g})} \leq \mathcal{C}(\mathbf{u}_g)\Delta_{e,g}^{k_g+1}, \tag{79}$$

$$||\eta_g||_{\Omega_{e,g}} + \Delta_{e,g}||\eta_g||_{H^1(\Omega_{e,g})} \leq \mathcal{C}(h_g)\Delta_{e,g}^{k_g+1}, \tag{80}$$

$$||\Theta_s||_{\Omega_{e,s}} + \Delta_{e,s}||\Theta_s||_{H^1(\Omega_{e,s})} + \Delta_{e,s}^2||\Theta_s||_{H^2(\Omega_{e,s})} \leq \mathcal{C}(\mathbf{u}_s)\Delta_{e,s}^{k_s+1}, \tag{81}$$

$$||\eta_s||_{\Omega_{e,s}} + \Delta_{e,s}||\eta_s||_{H^1(\Omega_{e,s})} \leq \mathcal{C}(h_s)\Delta_{e,s}^{k_s+1}. \tag{82}$$

Using (60)–(78) to bound the right-hand side of (44), choosing ϵ_1–ϵ_4 sufficiently small, and applying (79)–(82), we find

$$\begin{aligned} &\frac{1}{2}||K^{-1/2}\Psi_g||^2_{\Omega_g} + \frac{1}{2}\partial_t||\kappa_s||^2_{\Omega_s} + \frac{1}{2}\partial_t||\Psi_s||^2_{\Omega_s} + \frac{\mu}{4}||\nabla_{(x,y)}\Psi_s||^2_{\Omega_s} \\ &+\frac{1}{2}[\ ||\sigma_g^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma_g^{1/2}\kappa_g||^2_{\Gamma_D} + ||\sigma_g^{1/2}(\kappa_g - \kappa_s)||^2_{\Gamma_{GS}}] \\ &\leq \mathcal{C}(K, K^t, K^i, K^m, K^M, \mu, \mathbf{u}_g, h_g, \mathbf{u}_s, h_s)(\Delta_g^{2k_g} + \Delta_s^{2k_s}) \\ &\quad +\mathcal{C}(K^t, K^i, K^m, K^M, \mu^{-1})[\ ||\kappa_s||^2_{\Omega_s} + ||\Psi_s||^2_{\Omega_s}]. \end{aligned} \tag{83}$$

Integrating in time from 0 to T and applying (30), (31), Gronwall's lemma, and the triangle inequality, we obtain the following error estimate.

THEOREM 4.2. *Assume* $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, *and* h_s *are sufficiently smooth so that approximation bounds* (79)–(82) *hold. Assume that the ground water and surface water meshes satisfy Assumption GS and that the penalty functions* σ_g *and* σ_s *satisfy* (57). *Then the LDG/DG method* (26)–(31) *satisfies*

$$\left(\int_0^T ||K^{1/2}\Psi_g||^2_{\Omega_g} dt\right)^{1/2}$$
$$+\left(\int_0^T [\ ||\sigma_g^{1/2}[\![\kappa_g]\!]||^2_{\mathcal{E}_{i,g}} + ||\sigma_g^{1/2}\kappa_g||^2_{\Gamma_D} + ||\sigma_g^{1/2}(\kappa_g-\kappa_s)||^2_{\Gamma_{GS}}]dt\right)^{1/2}$$
$$+||\kappa_s(\cdot,T)||_{\Omega_s} + ||\Psi_s(\cdot,T)||_{\Omega_s} \le \tilde{\mathcal{C}}(\Delta_g^{k_g} + \Delta_s^{k_s}),$$

where $\tilde{\mathcal{C}}$ *depends on* K, μ, K^t, K^i, K^M, K^m, $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, *and* h_s. *Applying the triangle inequality, we obtain*

$$\left(\int_0^T ||K^{1/2}(\mathbf{u}_g - \boldsymbol{U}_g)||^2_{\Omega_g} dt\right)^{1/2}$$
$$+||(h_s - H_s)(\cdot,T)||_{\Omega_s} + ||(\mathbf{u}_s - \boldsymbol{U}_s)(\cdot,T)||_{\Omega_s} \le \tilde{\mathcal{C}}(\Delta_g^{k_g} + \Delta_s^{k_s}).$$

We recall another inverse inequality, valid in two dimensions,

$$||w(\cdot,t)||_{L^\infty(\Omega_s)} \le \mathcal{C}^\infty \Delta_s^{-1} ||w(\cdot,t)||_{\Omega_s}, \tag{84}$$

where $\mathcal{C}^\infty$ is independent of Δ_s and $w(\cdot,t)$ is in either $\mathcal{W}_{\Delta_s,s}$ or $\mathcal{V}_{\Delta_s,s}$. Assuming that k_g and $k_s \ge 2$ and Δ_s and $\Delta_s^{-1}\Delta_g^2$ are sufficiently small, we have

$$||\kappa_s||_{L^\infty(0,T;L^\infty(\Omega_s))} \le \mathcal{C}^\infty \tilde{\mathcal{C}} \Delta_s^{-1}[\ \Delta_g^2 + \Delta_s^2] \ll K^M \tag{85}$$

and similarly for Ψ_s. Then we can remove the dependence of $\tilde{C}$ on K^M.

4.1. An estimate for H_g. The estimate (4.2) does not say anything about the accuracy of H_g. Estimates for the error $h_g - H_g$ follow from a duality argument.

In this section, let $\nabla = \nabla_{(x,y,z)}$, let $e_g = h_g - H_g$, and assume that ϕ satisfies

$$-\nabla\cdot(K\nabla\phi) = e_g, \quad \Omega_g \tag{86}$$

with the boundary conditions

$$\phi = 0, \quad \Gamma_D \cup \Gamma_{GS}, \tag{87}$$
$$K\nabla\phi\cdot\mathbf{n}_g = 0, \quad \Gamma_N. \tag{88}$$

Define $\mathbf{q} = -K\nabla\phi$. We assume problem (86)–(88) satisfies the standard elliptic regularity bounds, so that

$$||\phi||_{H^2(\Omega_g)} + ||\mathbf{q}\cdot\mathbf{n}_g||_{\Gamma_D\cup\Gamma_{GS}} \le \mathcal{C}_r ||e_g||_{\Omega_g}, \tag{89}$$

where the constant $\mathcal{C}_r$ depends on the domain Ω and the coefficient K.

Multiplying (86) by e_g and integrating by parts,

$$\begin{aligned} ||e_g||^2_{\Omega_g} &= (h_g - H_g, \nabla\cdot\mathbf{q})_{\Omega_g} \\ &= -(\nabla h_g, \mathbf{q})_{\Omega_g} + \langle h_g, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_D} + \langle h_T, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} \\ &\quad +(\nabla H_g, \mathbf{q})_{\Omega_g} - \langle [\![H_g]\!], \mathbf{q}\rangle_{\mathcal{E}_{i,g}} - \langle H_g, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_D} - \langle H_g, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}}. \end{aligned} \tag{90}$$

Let π again denote L^2 projection into the appropriate LDG approximating space. Then

$$\begin{aligned} -(\nabla h_g, \mathbf{q}) &= (\nabla h_g, \pi\mathbf{q} - \mathbf{q})_{\Omega_g} - (\nabla h_g, \pi\mathbf{q})_{\Omega_g} \\ &= (\nabla h_g - \pi(\nabla h_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g} + (K^{-1}\mathbf{u}_g, \pi\mathbf{q})_{\Omega_g}. \end{aligned} \tag{91}$$

An alternate form of (17) is obtained by integration by parts:

$$\begin{aligned} \mathcal{A}_g(\boldsymbol{U}_g, H_g, H_s; \mathbf{v}_g) = (K^{-1}\boldsymbol{U}_g, \mathbf{v}_g)_{\Omega_g} &+ (\nabla H_g, \mathbf{v}_g)_{\Omega_g} \\ &- \langle H_g, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_D} - \langle [\![H_g]\!], \{\mathbf{v}_g\}\rangle_{\mathcal{E}_{i,g}} \\ &- \langle H_g - H_s, \mathbf{v}_g \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}}. \end{aligned}$$

Therefore, by (26),

$$\begin{aligned} (\nabla H_g, \mathbf{q})_{\Omega_g} &= (\nabla H_g, \pi\mathbf{q})_{\Omega_g} \\ &= -(K^{-1}\boldsymbol{U}_g, \pi\mathbf{q})_{\Omega_g} + \langle [\![H_g]\!], \{\pi\mathbf{q}\}\rangle_{\mathcal{E}_{i,g}} \\ &\quad + \langle H_g - h_g^D, \pi\mathbf{q} \cdot \mathbf{n}_g\rangle_{\Gamma_D} + \langle H_g - H_T, \pi\mathbf{q} \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}}, \end{aligned} \tag{92}$$

where $H_T = H_s + h_b$. Combining (90)–(92) and using $[\![h_g]\!] = 0$ on $\mathcal{E}_{i,g}$, we obtain

$$\begin{aligned} ||e_g||^2_{\Omega_g} = (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q})_{\Omega_g} &+ \langle [\![H_g - h_g]\!], \{\pi\mathbf{q}\} - \mathbf{q}\rangle_{\mathcal{E}_{i,g}} \\ &+ \langle h_g - H_g, (\mathbf{q} - \pi\mathbf{q}) \cdot \mathbf{n}_g\rangle_{\Gamma_D} + \langle h_T - H_T, \mathbf{q} \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}} \\ &+ \langle H_g - H_T, (\pi\mathbf{q} - \mathbf{q}) \cdot \mathbf{n}_g\rangle_{\Gamma_{GS}} + (\nabla h_g - \pi(\nabla h_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g}. \end{aligned} \tag{93}$$

Next, consider

$$(K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q})_{\Omega_g} = (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \mathbf{q})_{\Omega_g} + (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g}. \tag{94}$$

Integrating by parts,

$$\begin{aligned} (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \mathbf{q})_{\Omega_g} &= -(K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), K\nabla\phi)_{\Omega_g} \\ &= (\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \phi)_{\Omega_g} + \langle [\![\boldsymbol{U}_g]\!], \phi\rangle_{\mathcal{E}_{i,g}} \\ &\quad + \langle \boldsymbol{U}_g \cdot \mathbf{n}_g \ - u_N, \phi\rangle_{\Gamma_N}. \end{aligned} \tag{95}$$

Furthermore,

$$(\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \phi)_{\Omega_g} = (\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \phi - \pi\phi)_{\Omega_g} + (\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \pi\phi)_{\Omega_g}, \tag{96}$$

and by (20) and (27),

$$\begin{aligned} (\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \pi\phi)_{\Omega_g} &= \langle [\![\mathbf{u}_g - \boldsymbol{U}_g]\!], \{\pi\phi\}\rangle_{\mathcal{E}_{i,g}} + \langle \mathbf{u}_g \cdot \mathbf{n}_g - \boldsymbol{U}_g \cdot \mathbf{n}_g, \pi\phi\rangle_{\Gamma_N} \\ &\quad + \langle \sigma_g [\![H_g - h_g]\!], [\![\pi\phi]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \sigma_g(H_g - H_T), \pi\phi\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma_g(H_g - h_g), \pi\phi\rangle_{\Gamma_D}. \end{aligned} \tag{97}$$

Combining (94)–(97), we find

$$\begin{aligned} (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q})_{\Omega_g} &= (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g} + (\nabla \cdot (\mathbf{u}_g - \boldsymbol{U}_g), \phi - \pi\phi)_{\Omega_g} \\ &\quad + \langle [\![\boldsymbol{U}_g - \mathbf{u}_g]\!], \phi - \{\pi\phi\}\rangle_{\mathcal{E}_{i,g}} + \langle \boldsymbol{U}_g \cdot \mathbf{n}_g - \mathbf{u}_g \cdot \mathbf{n}_g, \phi - \pi\phi\rangle_{\Gamma_N} \\ &\quad + \langle \sigma_g [\![H_g - h_g]\!], [\![\pi\phi]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \sigma_g(H_g - H_T), \pi\phi\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma_g(H_g - h_g), \pi\phi\rangle_{\Gamma_D}. \end{aligned} \tag{98}$$

Finally, substituting (98) into (93), we obtain

$$\begin{aligned}||e_g||^2_{\Omega_g} &= (K^{-1}(\mathbf{u}_g - \boldsymbol{U}_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g} + (\nabla\cdot(\mathbf{u}_g - \boldsymbol{U}_g), \phi - \pi\phi)_{\Omega_g} \\ &\quad + \langle [\![\boldsymbol{U}_g - \mathbf{u}_g]\!], \phi - \{\pi\phi\}\rangle_{\mathcal{E}_{i,g}} + \langle \boldsymbol{U}_g\cdot\mathbf{n}_g - \mathbf{u}_g\cdot\mathbf{n}_g, \phi - \pi\phi\rangle_{\Gamma_N}. \\ &\quad + \langle [\![H_g - h_g]\!], \{\pi\mathbf{q}\} - \mathbf{q}\rangle_{\mathcal{E}_{i,g}} + \langle h_g - H_g, (\mathbf{q} - \pi\mathbf{q})\cdot\mathbf{n}_g\rangle_{\Gamma_D} \\ &\quad + \langle H_g - H_T, (\pi\mathbf{q} - \mathbf{q})\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} + \langle h_T - H_T, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma_g [\![H_g - h_g]\!], [\![\pi\phi]\!]\rangle_{\mathcal{E}_{i,g}} + \langle \sigma_g(H_g - H_T), \pi\phi\rangle_{\Gamma_{GS}} \\ &\quad + \langle \sigma_g(H_g - h_g), \pi\phi\rangle_{\Gamma_D} + (\nabla h_g - \pi(\nabla h_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g} \\ &\equiv F_1 + \cdots + F_{12}. \end{aligned} \tag{99}$$

By approximation theory,

$$||\pi\mathbf{q} - \mathbf{q}||_{\Omega_{e,g}} + \Delta_{e,g}||\pi\mathbf{q} - \mathbf{q}||_{H^1(\Omega_{e,g})} \le \mathcal{C}(\mathcal{C}_r)\Delta_{e,g}||\phi||_{H^2(\Omega_{e,g})}, \tag{100}$$

$$||\pi\phi - \phi||_{\Omega_{e,g}} + \Delta_{e,g}||\pi\phi - \phi||_{H^1(\Omega_{e,g})} \le \mathcal{C}(\mathcal{C}_r)\Delta^2_{e,g}||\phi||_{H^2(\Omega_{e,g})}. \tag{101}$$

In the arguments below, let $\bar{\mathcal{C}}$ be a generic constant which depends on $\mathcal{C}_r$ and $\tilde{\mathcal{C}}$ from Theorem 4.2. Integrating (99) in time and applying Theorem 4.1 and inverse inequality (53), the bounds (100)–(101), (89), and the result of Theorem 4.2, we obtain

$$\begin{aligned}&\int_0^T [F_1 + \cdots + F_4]dt \\ &\le \int_0^T \sum_e ||K^{-1/2}(\mathbf{u}_g - \boldsymbol{U}_g)||_{\Omega_{e,g}}||\pi\mathbf{q} - \mathbf{q}||_{\Omega_{e,g}}dt \\ &\quad + \int_0^T \sum_e ||\mathbf{u}_g - \boldsymbol{U}_g||_{H^1(\Omega_{e,g})}||\phi - \pi\phi||_{\Omega_{e,g}}dt \\ &\quad + \int_0^T \sum_e [\,||\mathbf{u}_g - \boldsymbol{U}_g||^2_{\Omega_{e,g}} + \Delta^2_{e,g}||\nabla(\mathbf{u}_g - \boldsymbol{U}_g)||^2_{\Omega_{e,g}}]^{1/2} \\ &\qquad \times [\,\Delta^{-2}_{e,g}||\phi - \pi\phi||^2_{\Omega_{e,g}} + ||\nabla(\phi - \pi\phi)||^2_{\Omega_{e,g}}]^{1/2}dt \\ &\le \bar{\mathcal{C}}\Delta_g(\Delta_g^{k_g} + \Delta_s^{k_s})\left(\int_0^T ||e_g||^2_{\Omega_g}dt\right)^{1/2}. \end{aligned} \tag{102}$$

Consider

$$\begin{aligned}F_5 &= \langle [\![H_g - h_g]\!], \{\pi\mathbf{q} - \mathbf{q}\}\rangle_{\mathcal{E}_{i,g}} \\ &= \langle \sigma_g^{1/2}[\![\kappa_g - \eta_g]\!], \sigma_g^{-1/2}\{\pi\mathbf{q} - \mathbf{q}\}\rangle_{\mathcal{E}_{i,g}} \\ &\le \left(||\sigma_g^{1/2}[\![\kappa_g]\!]||_{\mathcal{E}_{i,g}} + ||\sigma_g^{1/2}[\![\eta_g]\!]||_{\mathcal{E}_{i,g}}\right)||\sigma_g^{-1/2}\{\pi\mathbf{q} - \mathbf{q}\}||_{\mathcal{E}_{i,g}}. \end{aligned} \tag{103}$$

By (57), Theorem 4.1, and (100),

$$\begin{aligned}||\sigma_g^{-1/2}\{\pi\mathbf{q} - \mathbf{q}\}||_{\mathcal{E}_{i,g}} &\le \mathcal{C}\sum_e [\,||\pi\mathbf{q} - \mathbf{q}||^2_{\Omega_{e,g}} + \Delta^2_{e,g}||\nabla(\pi\mathbf{q} - \mathbf{q})||^2_{\Omega_{e,g}}]^{1/2} \\ &\le \mathcal{C}\Delta_g||e_g||_{\Omega_g}; \end{aligned} \tag{104}$$

hence by Theorem 4.2 and estimates on η_g we find

$$\int_0^T F_5 dt \le \bar{\mathcal{C}}\Delta_g(\Delta_g^{k_g} + \Delta_s^{k_s})\left(\int_0^T ||e_g||^2_{\Omega_g}dt\right)^{1/2}. \tag{105}$$

A similar argument gives the same bound for F_6 and F_7, since $H_g - H_T = \kappa_g - \kappa_s + \eta_s - \eta_g$. For F_8,

$$\begin{aligned}\int_0^T \langle h_T - H_T, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} dt &= \int_0^T \langle \eta_s - \kappa_s, \mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}} dt \\ &\le \bar{\mathcal{C}} \int_0^T \left[||\eta_s||_{\Omega_s} + ||\kappa_s||_{\Omega_s}\right] ||e_g||_{\Omega_g} dt \\ &\le \bar{\mathcal{C}}(\Delta_g^{k_g} + \Delta_s^{k_s}) \left(\int_0^T ||e_g||^2_{\Omega_g} dt\right)^{1/2} \end{aligned} \tag{106}$$

by Theorem 4.2. Next, consider

$$\begin{aligned} F_9 &= \langle \sigma_g [\![H_g - h_g]\!], [\![\pi\phi]\!]\rangle_{\mathcal{E}_{i,g}} \\ &= \langle \sigma_g [\![\kappa_g - \eta_g]\!], [\![\pi\phi - \phi]\!]\rangle_{\mathcal{E}_{i,g}} \\ &\le \left(||\sigma_g^{1/2}[\![\kappa_g]\!]||_{\mathcal{E}_{i,g}} + ||\sigma_g^{1/2}[\![\eta_g]\!]||_{\mathcal{E}_{i,g}}\right) ||\sigma_g^{1/2}[\![\pi\phi - \phi]\!]||_{\mathcal{E}_{i,g}}. \end{aligned} \tag{107}$$

Similar to (104),

$$||\sigma_g^{1/2}[\![\pi\phi - \phi]\!]||_{\mathcal{E}_{i,g}} \le \mathcal{C}\Delta_g ||e_g||_{\Omega_g}; \tag{108}$$

therefore

$$\int_0^T F_9 \, dt \le \bar{\mathcal{C}}\Delta_g(\Delta_g^{k_g} + \Delta_s^{k_s+1/2}) \left(\int_0^T ||e_g||^2_{\Omega_g} dt\right)^{1/2}. \tag{109}$$

A similar argument gives the same bound for F_{10} and F_{11}, since $\phi = 0$ on Γ_D and Γ_{GS}. Finally,

$$\begin{aligned} F_{12} &= (\nabla h_g - \pi(\nabla h_g), \pi\mathbf{q} - \mathbf{q})_{\Omega_g} \\ &\le \bar{\mathcal{C}}\Delta_g^{k_g+1} \left(\int_0^T ||e_g||^2_{\Omega_g} dt\right)^{1/2}. \end{aligned} \tag{110}$$

Combining bounds (102)–(110) with (99), we obtain the following estimate for e_g.

THEOREM 4.3. *Under the assumptions of Theorem* 4.2 *and the regularity of the dual solution* ϕ*, we have*

$$\left(\int_0^T ||h_g - H_g)||^2 dt\right)^{1/2} \le \mathcal{C}(C_r, \tilde{C})(\Delta_g^{k_g} + \Delta_s^{k_s}).$$

This estimate unfortunately does not give the additional power of Δ_g which is obtained in the purely elliptic case considered in [15]. The problem is in the term F_8, which involves the coupling between ground water and surface water. Theorem 4.3 essentially tells us that the error in H_g is no better than the error in H_s, which is not terribly surprising.

5. The MFE/DG method. The flexibility of the LDG framework (26)–(29) easily allows for a MFE approximation of the ground water flow equations. This can be accomplished by restricting the finite element spaces $\mathcal{V}_{\Delta_g,g}$ and $\mathcal{W}_{\Delta_g,g}$ to be

standard mixed spaces; see, for example, [33, 12, 10, 11, 16]. In this section, we assume that the mesh $\mathcal{T}_{\Delta_g,g}$ is regular and conforming and denote by $\mathcal{V}^M_{\Delta_g,g}$ and $\mathcal{W}^M_{\Delta_g,g}$ MFE approximating spaces of order $k_g \geq 0$. Thus, functions in $\mathcal{W}^M_{\Delta_g,g}$ are piecewise polynomials of degree k_g, $\mathcal{V}^M_{\Delta_g,g} \subset H(\mathrm{div};\Omega_g)$, and the spaces satisfy the Babuška–Brezzi (BB) *inf-sup* condition. Furthermore,

$$\nabla_{(xyz)} \cdot \mathbf{v} \in \mathcal{W}^M_{\Delta_g,g} \quad \forall\, \mathbf{v} \in \mathcal{V}^M_{\Delta_g,g}. \tag{111}$$

For simplicity, in this section we will assume that $u_N = 0$. When this is not the case, we enforce the Neumann boundary condition by setting $\boldsymbol{U}_g \cdot \mathbf{n}_g = \tilde{u}_N$, where $\tilde{u}_N$ is the projection of u_N into a space of Lagrange multiplier functions defined on Γ_N; see, for example, [4]. The analysis below carries through with this modification at the expense of including some additional terms which have been analyzed in previous papers. Denote

$$\mathcal{V}^{M,0}_{\Delta_g,g} = \mathcal{V}^M_{\Delta_g,g} \cap \{\mathbf{v}_g : \mathbf{v}_g \cdot \mathbf{n}_g|_{\Gamma_N} = 0\}.$$

Define

$$\begin{aligned}
\mathcal{A}^M_g(\mathbf{u}_g, h_g, h_s; \mathbf{v}_g) &\equiv (K^{-1}\mathbf{u}_g, \mathbf{v}_g)_{\Omega_g} - (h_g, \nabla_{(x,y,z)} \cdot \mathbf{v}_g)_{\Omega_g} \\
&\quad + \langle h_s, \mathbf{v}_g \cdot \mathbf{n}_g \rangle_{\Gamma_{GS}}, \qquad (112)\\
\mathcal{B}^M_g(\mathbf{u}_g; w_g) &\equiv (\nabla_{(x,y,z)} \cdot \mathbf{u}_g, w_g)_{\Omega_g}, \qquad (113)\\
\mathcal{B}^M_s(\mathbf{u}_g, \mathbf{u}_s, h_s; w_s) &\equiv (\partial_t h_s, w_s)_{\Omega_s} - (\mathbf{u}_s h_s, \nabla_{(x,y)} w_s)_{\Omega_s} + \langle \{\mathbf{u}_s\} h_s^{\uparrow}, [\![w_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
&\quad - \langle \mathbf{u}_g \cdot \mathbf{n}_g, w_s \rangle_{\Gamma_{GS}} + \langle \mathbf{u}_s \cdot \mathbf{n}_s h_s, w_s \rangle_{\partial\Omega_{s,out}} \\
&\quad + \langle \mathbf{u}_s \cdot \mathbf{n}_s h_s^I, w_s \rangle_{\partial\Omega_{s,in}}. \qquad (114)
\end{aligned}$$

The MFE/DG method then is to find $\boldsymbol{U}_g \in \mathcal{V}^{M,0}_{\Delta_g,g}$, $H_g \in \mathcal{W}^M_{\Delta_g,g}$, $\boldsymbol{U}_s \in \mathcal{V}_{\Delta_s,s}$, and $H_s \in \mathcal{W}_{\Delta_s,s}$ which satisfy

$$\mathcal{A}^M_g(\boldsymbol{U}_g, H_g, H_s; \mathbf{v}_g) = -\langle h_b, \mathbf{v}_g \cdot \mathbf{n}_g \rangle_{\Gamma_{GS}} - \langle h_g^D, \mathbf{v}_g \cdot \mathbf{n}_g \rangle_{\Gamma_D}, \quad \mathbf{v}_g \in \mathcal{V}^{M,0}_{\Delta_g,g}, \tag{115}$$

$$\mathcal{B}^M_g(\boldsymbol{U}_g; w_g) = (f_g, w_g)_{\Omega_g}, \quad w_g \in \mathcal{W}^M_{\Delta_g,g}, \tag{116}$$

$$\begin{aligned}
\mathcal{B}^M_s(\mathbf{u}_g, \mathbf{u}_s, h_s; w_s) &= 0, \quad w_s \in \mathcal{W}_{\Delta_s,s}, \qquad (117)\\
\mathcal{A}_s(\boldsymbol{U}_s, H_s; \mathbf{v}_s) &= -(g\nabla_{(x,y)} h_b, \mathbf{v}_s)_{\Omega_s} - \langle g h_s^I, \mathbf{v}_s \cdot \mathbf{n} \rangle_{\partial\Omega_{s,in}} \\
&\quad + \mu \langle \nabla_{(x,y)} \mathbf{v}_s \cdot \mathbf{n}, \hat{\mathbf{u}}_s \rangle_{\partial\Omega_s} + \langle \sigma_s \hat{\mathbf{u}}_s, \mathbf{v}_s \rangle_{\partial\Omega_s} \\
&\quad + (\mathcal{F}, \mathbf{v}_s)_{\Omega_s}, \quad \mathbf{v}_s \in \mathcal{V}_{\Delta_s,s}, \quad \mathbf{v}_s \in \mathcal{V}_{\Delta_s,s}. \qquad (118)
\end{aligned}$$

Furthermore, $H_s(\cdot, 0)$ and $\boldsymbol{U}_s(\cdot, 0)$ are defined to be the L^2 projections of h_s^0 and $\mathbf{u}_s^0$, respectively.

5.1. An a priori error estimate. Proceeding as above, let $\Pi\mathbf{u}_g$, πh_g, $\pi\mathbf{u}_s$, and πh_s be projections of the true solutions. The projections πh_g, $\pi\mathbf{u}_s$, and πh_s are defined to be L^2 projections as before; $\Pi\mathbf{u}_g$ will denote the well-known "Π-projection" of $\mathbf{u}_g$ into $\mathcal{V}^{M,0}_{\Delta_g,g}$ [33], which satisfies, among other properties,

$$(\nabla_{(x,y,z)} \cdot (\Pi\mathbf{u}_g - \mathbf{u}_g), w_g)_{\Omega_g} = 0, \quad w_g \in \mathcal{W}^M_{\Delta_g,g}. \tag{119}$$

Defining Ψ_g, Θ_g, κ_g, η_g, κ_s, and η_s as above, we obtain

$$\mathcal{A}^M_g(\Psi_g, \kappa_g, \kappa_s; \mathbf{v}_g) = \mathcal{A}^M_g(\Theta_g, \eta_g, \eta_s; \mathbf{v}_g), \tag{120}$$

$$\mathcal{B}^M_g(\Psi_g; w_g) = \mathcal{B}^M_g(\Theta_g; w_g). \tag{121}$$

Define

$$\mathcal{B}^M_{s,L}(\mathbf{u}_g, h_s; w_s) \equiv (\partial_t h_s, w_s)_{\Omega_s} - \langle \mathbf{u}_g \cdot \mathbf{n}_g, w_s \rangle_{\Gamma_{GS}}. \tag{122}$$

Then

$$\begin{aligned}
&\mathcal{B}^M_{s,L}(\Psi_g, \kappa_s; w_s) - (\boldsymbol{U}_s \kappa_s, \nabla_{(x,y)} w_s)_{\Omega_s} + \langle \{\boldsymbol{U}_s\} \kappa_s^{\uparrow}, [\![w_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \boldsymbol{U}_s \cdot \mathbf{n}_s \kappa_s, w_s \rangle_{\partial \Omega_{s,out}} \\
&= \mathcal{B}^M_{s,L}(\Theta_g, \eta_s; w_s) - (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s, \nabla_{(x,y)} w_s)_{\Omega_s} + \langle \mathbf{u}_s h_s - \{\boldsymbol{U}_s\} \pi h_s^{\uparrow}, [\![w_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
&+ \langle (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s) \cdot \mathbf{n}_s, w_s \rangle_{\partial \Omega_{s,out}} + \langle (\mathbf{u}_s - \boldsymbol{U}_s) \cdot \mathbf{n}_s h_s^I, w_s \rangle_{\partial \Omega_{s,in}}.
\end{aligned} \tag{123}$$

Furthermore, (36) holds as before. Manipulating (123) as in (39)–(41) and adding the result to (36), (120), and (121), we find

$$\begin{aligned}
&||K^{-1/2} \Psi_g||^2_{\Omega_g} + (\partial_t \kappa_s, \kappa_s)_{\Omega_s} + (\partial_t \Psi_s, \Psi_s)_{\Omega_s} + ||\tau_{bf}^{1/2} \Psi_s||^2_{\Omega_s} + \mu ||\nabla_{(x,y)} \Psi_s||^2_{\Omega_s} \\
&+ \frac{1}{2} \left[\langle |\boldsymbol{U}_s \cdot \mathbf{n}^-|, [\kappa_s^- - \kappa_s^+]^2 \rangle_{\mathcal{E}_{i,s}} + \langle |\boldsymbol{U}_s \cdot \mathbf{n}_s|, \kappa_s^2 \rangle_{\partial \Omega_{s,in}} + \langle |\boldsymbol{U}_s \cdot \mathbf{n}_s|, \kappa_s^2 \rangle_{\partial \Omega_{s,out}} \right] \\
&+ \langle \sigma_s [\![\Psi_s]\!], [\![\Psi_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \sigma_s \Psi_s, \Psi_s \rangle_{\partial \Omega_s} \\
&= \mathcal{A}^M_g(\Theta_g, \eta_g, \eta_s; \Psi_g) + \mathcal{B}^M_g(\Theta_g, \eta_g, \eta_s; \kappa_g) + \mathcal{B}^M_{s,L}(\Theta_g, \eta_s; \kappa_s) + \mathcal{A}_{s,L}(\Theta_s, \eta_s; \Psi_s) \\
&\quad - \sum_{\partial \Omega^-_{e,s} \in \mathcal{T}_{\Delta_s,s}} \langle |\{\boldsymbol{U}_s\} \cdot \mathbf{n}_e| (\boldsymbol{U}_s^{\text{int}} - \boldsymbol{U}_s^{\text{ext}}), \Psi_s^{\text{int}} \rangle_{\partial \Omega^-_{e,s}} \\
&\quad + (\mathbf{u}_s \cdot \nabla_{(x,y)} \mathbf{u}_s - \boldsymbol{U}_s \cdot \nabla_{(x,y)} \boldsymbol{U}_s, \Psi_s)_{\Omega_s} \\
&\quad - \frac{1}{2} (\nabla_{(x,y)} \cdot \boldsymbol{U}_s, \kappa_s^2)_{\Omega_s} + \frac{1}{2} \langle \{\kappa_s^2\}, [\![\boldsymbol{U}_s]\!] \rangle_{\mathcal{E}_{i,s}} \\
&\quad + (\nabla_{(x,y)} \cdot (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s), \kappa_s)_{\Omega_s} - \langle [\![\kappa_s (\mathbf{u}_s h_s - \boldsymbol{U}_s \pi h_s)]\!], 1 \rangle_{\mathcal{E}_{i,s}} \\
&\quad + \langle \mathbf{u}_s h_s - \{\boldsymbol{U}_s\} \pi h_s^{\uparrow}, [\![\kappa_s]\!] \rangle_{\mathcal{E}_{i,s}} + \langle \boldsymbol{U}_s \cdot \mathbf{n}_s (\pi h_s - h_s^I), \kappa_s \rangle_{\partial \Omega_{s,in}} \\
&\quad + (g \kappa_s, \nabla_{(x,y)} \cdot \Psi_s)_{\Omega_s} - \langle g \{\kappa_s\}, [\![\Psi_s]\!] \rangle_{\mathcal{E}_{i,s}} - \langle g \kappa_s, \Psi_s \cdot \mathbf{n}_s \rangle_{\partial \Omega_{s,out}} \\
&\equiv \mathcal{A}^M_g(\Theta_g, \eta_g, \eta_s; \Psi_g) + \mathcal{B}^M_g(\Theta_g, \eta_g, \eta_s; \kappa_g) + \mathcal{B}^M_{s,L}(\Theta_g, \eta_s; \kappa_s) + \mathcal{A}_{s,L}(\Theta_s, \eta_s; \Psi_s) \\
&\quad + \sum_{i=1}^{11} E_i,
\end{aligned} \tag{124}$$

where E_1 through E_{11} are the same as in (44).

By the properties of the Π-projection and L^2 projection,

$$\mathcal{A}^M_g(\Theta_g, \eta_g, \eta_s; \Psi_g) = (K^{-1} \Theta_g, \Psi_g)_{\Omega_g} + \langle \eta_s, \Psi_g \cdot \mathbf{n}_g \rangle_{\Gamma_{GS}} \tag{125}$$

and

$$\mathcal{B}^M_g(\Theta_g; \kappa_g) + \mathcal{B}^M_{s,L}(\Theta_g, \eta_s; \kappa_s) = -\langle \Theta_g \cdot \mathbf{n}_g, \kappa_s \rangle_{\Gamma_{GS}}. \tag{126}$$

The two terms on the right-hand side of (125) are handled precisely as in (60) and (62). The term on the right-hand side of (126) is bounded by

$$\mathcal{C} ||\Theta_g \cdot \mathbf{n}_g||^2_{\Gamma_{GS}} + ||\kappa_s||^2_{\Omega_s}. \tag{127}$$

The approximation properties of the Π-projection and L^2 projection give

$$||\eta_g||_{\Omega_g} + ||\Theta_g||_{\Omega_g} + ||\Theta_g \cdot \mathbf{n}_g||_{\Gamma_{GS}} \leq \mathcal{C}(\mathbf{u}_g) \Delta_g^{k_g+1}.$$

The remaining terms on the right-hand side of (124) are handled as in section 4. We obtain the following result.

THEOREM 5.1. *Assume* $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, *and* h_s *are sufficiently smooth and that the ground water and surface water meshes satisfy Assumption GS. Then the MFE/DG method* (115)–(118) *satisfies*

$$\left(\int_0^T ||K^{1/2}(\mathbf{u}_g - \boldsymbol{U}_g)||^2 dt\right)^{1/2}$$
$$+||(h_s - H_s)(\cdot,T)||_{\Omega_s} + ||(\mathbf{u}_s - \boldsymbol{U}_s)(\cdot,T)||_{\Omega_s} \le \tilde{\mathcal{C}}(\Delta_g^{k_g+1} + \Delta_s^{k_s}),$$

where $\tilde{\mathcal{C}}$ *depends on* K, K^t, K^i, K^m, K^M, μ, $\mathbf{u}_g$, h_g, $\mathbf{u}_s$, *and* h_s.

5.2. An estimate for H_g. Let $\nabla = \nabla_{(x,y,z)}$. We consider again a dual problem

$$-\nabla \cdot (K\nabla\phi) = H_g - \pi h_g, \quad \Omega_{\mathrm{g}} \tag{128}$$

with the boundary conditions

$$\phi = 0, \quad \Gamma_D \cup \Gamma_{GS}, \tag{129}$$

$$K\nabla\phi \cdot \mathbf{n}_g = 0, \quad \Gamma_N. \tag{130}$$

Define $\mathbf{q} = -K\nabla\phi$, and let $\Pi\mathbf{q}$ be its Π-projection. Then, using properties of the Π- and L^2 projections,

$$\begin{aligned} ||\kappa_g||^2_{\Omega_{\mathrm{g}}} &= (\kappa_g, \nabla\cdot\mathbf{q})_{\Omega_{\mathrm{g}}} \\ &= (\kappa_g, \nabla\cdot\Pi\mathbf{q})_{\Omega_{\mathrm{g}}} \\ &= (H_g - h_g, \nabla\cdot\Pi\mathbf{q})_{\Omega_{\mathrm{g}}} \\ &= (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q})_{\Omega_{\mathrm{g}}} + \langle h_s - H_s, \Pi\mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}}, \end{aligned} \tag{131}$$

where in the last step we have used the orthogonality relation

$$\mathcal{A}_g^M(\mathbf{u}_g - \boldsymbol{U}_g, h_g - H_g, h_s - H_s; \mathbf{v}_g) = 0.$$

By the definition of $\mathbf{q}$ and integration by parts,

$$\begin{aligned} (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q})_{\Omega_{\mathrm{g}}} &= (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \mathbf{q})_{\Omega_{\mathrm{g}}} + (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q} - \mathbf{q})_{\Omega_{\mathrm{g}}} \\ &= (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), -K\nabla\phi)_{\Omega_{\mathrm{g}}} + (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q} - \mathbf{q})_{\Omega_{\mathrm{g}}} \\ &= (\nabla\cdot(\boldsymbol{U}_g - \mathbf{u}_g), \phi)_{\Omega_{\mathrm{g}}} + (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q} - \mathbf{q})_{\Omega_{\mathrm{g}}}. \end{aligned} \tag{132}$$

By the fact that

$$\mathcal{B}_g^M(\mathbf{u}_g - \boldsymbol{U}_g; w_g) = 0,$$

we easily see that

$$\nabla\cdot(\boldsymbol{U}_g - \Pi\mathbf{u}_g) \equiv 0,$$

and therefore

$$(\nabla\cdot(\boldsymbol{U}_g - \mathbf{u}_g), \phi)_{\Omega_{\mathrm{g}}} = (\nabla\cdot(\Pi\mathbf{u}_g - \mathbf{u}_g), \phi - \pi\phi)_{\Omega_{\mathrm{g}}}, \tag{133}$$

where $\pi\phi$ is the L^2 projection of ϕ into $\mathcal{W}^M_{\Delta_g,g}$. Substituting into (131), we obtain

$$||\kappa_g||^2_{\Omega_g} = (\nabla\cdot(\Pi\mathbf{u}_g - \mathbf{u}_g), \phi - \pi\phi)_{\Omega_g} + (K^{-1}(\boldsymbol{U}_g - \mathbf{u}_g), \Pi\mathbf{q} - \mathbf{q})_{\Omega_g} + \langle h_s - H_s, \Pi\mathbf{q}\cdot\mathbf{n}_g\rangle_{\Gamma_{GS}}. \tag{134}$$

Using the well-known estimate for the Π-projection

$$||\nabla\cdot(\mathbf{u}_g - \Pi\mathbf{u}_g)||_{\Omega_g} \le C\Delta_g^{k_g+1} \tag{135}$$

and the approximation properties of $\pi\phi$ and $\Pi\mathbf{q}$,

$$\left(\int_0^T ||\kappa_g||^2_{\Omega_g} dt\right)^{1/2} \le \mathcal{C}\left[\Delta_g^{k_g+2} + \left(\int_0^T \left[\Delta_g^2 ||K^{-1/2}(\boldsymbol{U}_g - \mathbf{u}_g)||^2_{\Omega_g} + ||h_s - H_s||^2_{\Omega_s}\right] dt\right)^{1/2}\right].$$

Therefore by Theorem 5.1 we obtain the following result.

THEOREM 5.2. *Under the assumptions of Theorem* 5.1 *and elliptic regularity of the adjoint problem* (128)–(130)*, the scheme* (115)–(117) *satisfies*

$$\left(\int_0^T ||H_g - \pi h_g||^2_{\Omega_g} dt\right)^{1/2} \le \mathcal{C}(\Delta_g^{k_g+1} + \Delta_s^{k_s}). \tag{136}$$

Using the triangle inequality, an immediate result of this theorem is the following corollary.

COROLLARY 5.3.

$$\left(\int_0^T ||H_g - h_g||^2_{\Omega_g} dt\right)^{1/2} \le \mathcal{C}(\Delta_g^{k_g+1} + \Delta_s^{k_s}). \tag{137}$$

6. Numerical results. In this section, we present some preliminary numerical results for the LDG/DG method described above. We consider a two-dimensional ground water domain Ω_g, pictured in Figure 2 with a coarse triangular discretization. The dimensions of the domain are roughly 100 by 100 cm. The top boundary of Ω_g is the ground water/surface water interface Γ_{GS}; thus Ω_s is the interval $0 < x < 100$ cm.

In Ω_s, we solve the full surface water flow equations (5)–(6) with $g = 9.81$ cm/s and $\mu = \mathcal{F} = 0$. The bottom friction coefficient $\tau_{bf} = 10\ s^{-1}$. The initial water height $h_T = 101$ cm. At the left boundary $x = 0$, we assume a total inflow height of 106 cm, which is ramped up linearly over a 60 s time period. The right boundary $x = 100$ is treated as a land boundary with $\mathbf{u}_s = 0$. Piecewise linear approximations are used for both $\mathbf{u}_s$ and h_s.

In the analysis presented above, we have not considered time discretization. This is an important issue in ground water/surface water coupling, as the temporal scales in the two regimes can be vastly different. We will examine this issue in more detail in future work. For now, the system (5)–(6) is discretized in time using an explicit second-order Runge–Kutta method with a time step of size Δt_s. Typically, Δt_s is fairly small, on the order of a second. The ground water flow equations are then

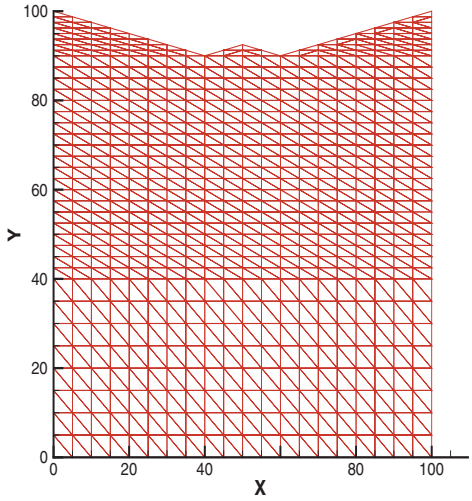


FIG. 2. *Ground water domain Ω_g with a coarse finite element mesh.*

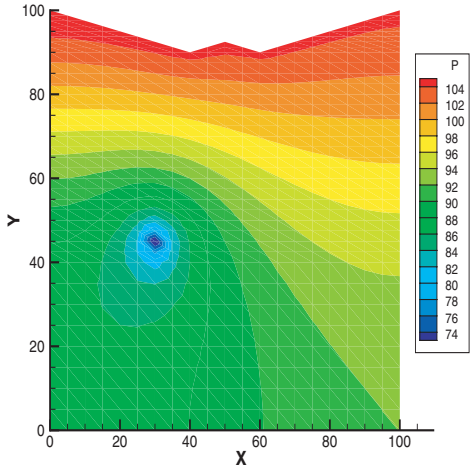


FIG. 3. *Ground water head H_g at $t = 1200$ s, computed on the mesh in Figure 2.*

solved at time steps $\Delta t_g = M\Delta t_s$ with $M \geq 1$. For the problem considered here, $\Delta t_s = .1$ s, and we have varied M to examine its effect.

In the ground water domain, the boundary conditions are no-flow ($u_N = 0$) on all boundaries except Γ_{GS}. We introduce a point sink

$$f_g(x, y) = \bar{f}\delta(x - \bar{x}, y - \bar{y}),$$

where $(\bar{x}, \bar{y}) = (30, 45)$ cm and $\bar{f} = -.3$ s^{-1}. The hydraulic conductivity $K = .00922$ cm/s. The approximations $\boldsymbol{U}_g$ and H_g are both piecewise linear.

In Figure 3, the ground water head H_g is shown at time $t = 1200$ s, computed on the triangular mesh given in Figure 2. This mesh contains 1256 elements. We can see from the head contours that flow is directed from the surface water domain into the ground water and towards the sink. In this case, the ground water flow equations were solved every 60 s or every 600 shallow water time steps. A new flux $\boldsymbol{U}_g \cdot \mathbf{n}_g$

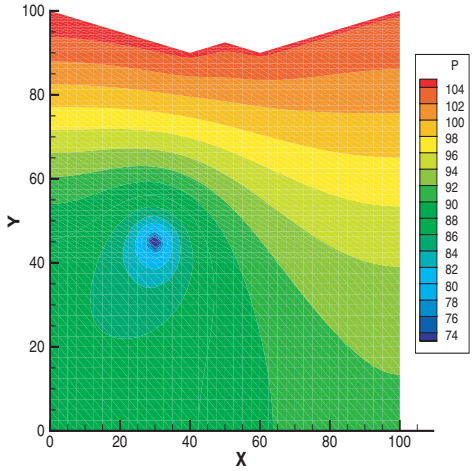


FIG. 4. *Ground water head H_g at $t = 1200$ s, computed on once refined mesh.*

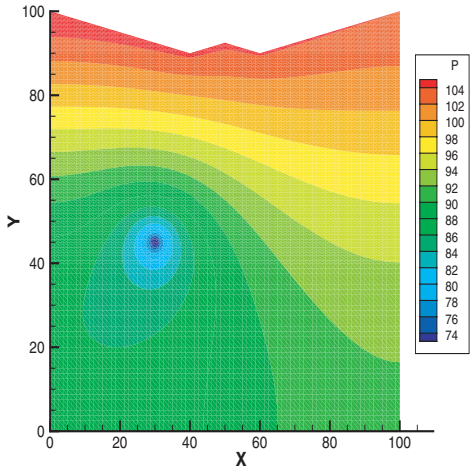


FIG. 5. *Ground water head H_g at $t = 1200$ s, computed on twice refined mesh.*

was computed at the end of each of these steps and fed back into the shallow water continuity equation (5).

The mesh was then refined twice, using edge bisection, giving meshes with 5024 and 20096 elements, respectively. On the second mesh, the ground water flow equations were solved every 30 s and, for the finest mesh, every 15 s. Contours of the solution h_g on these meshes are given in Figures 4 and 5. As observed in these figures, there is little difference between the solutions, except near the point sink. As the mesh is refined, the sink is better approximated. These results were not seen to be sensitive to the choice of M. In all of these simulations, the penalty parameter $\sigma_g = .01(\Delta_\gamma)^{-1}$ with Δ_γ defined as in (57).

The surface water heights at different times for the three different meshes are plotted in Figures 6–8. Figure 6 presents a plot of the total water height $H_T = H_s + h_b$ at time $t = 60$ s. Figures 7 and 8 are at 600 and 1200 s, respectively. In our runs, the surface water mesh aligns with the faces of the ground water mesh on Γ_{GS} mapped onto Ω_s. Thus, for the coarsest discretization, seen in Figure 2, there are 20 elements

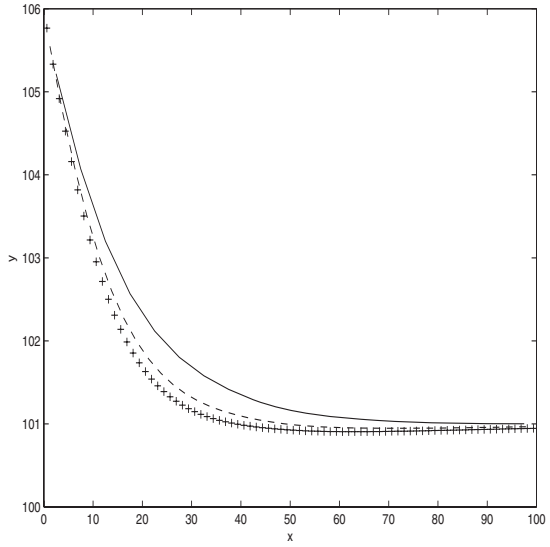


FIG. 6. *Surface water height* H_T *at* $t = 60$ s*, computed with* 20 *elements (solid),* 40 *elements* (−)*, and* 80 *elements* (+)*.*

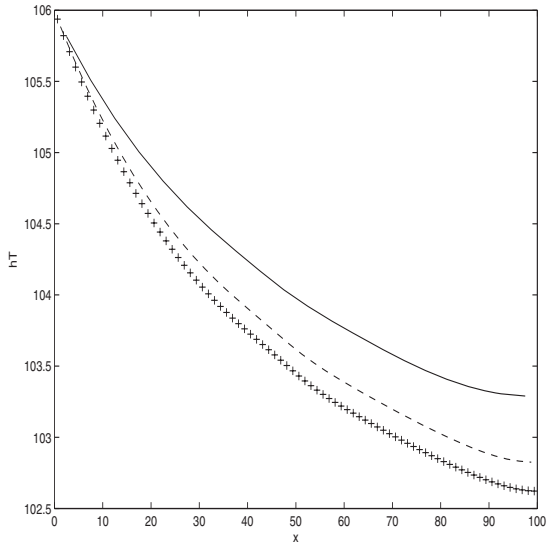


FIG. 7. *Surface water height* H_T *at* $t = 600$ s*, computed with* 20 *elements (solid),* 40 *elements* (−)*, and* 80 *elements* (+)*.*

in the surface water domain $0 < x < 100$ cm. The two finer discretizations of the domain have 40 and 80 elements in the interval, respectively. As seen in the figures, the solutions agree fairly well at the earlier time, but as time increases, the difference between the solutions also increases. The solutions with 40 and 80 elements are more similar, however, indicating that the surface water solution is starting to converge. These results suggest the need to further refine the mesh in the surface water domain, perhaps independently of the mesh used in the ground water domain. This issue will be explored further in future research.

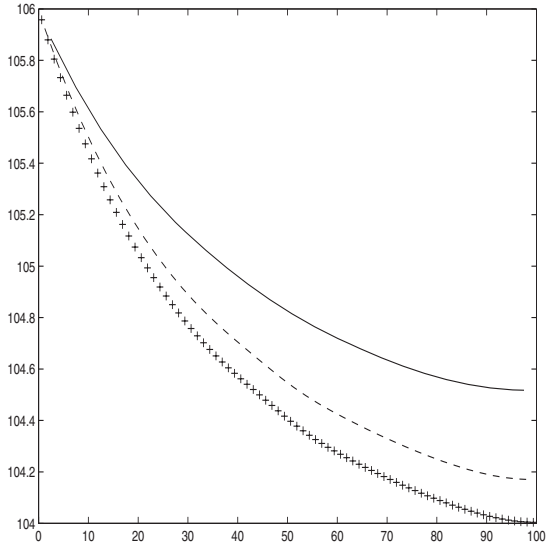


FIG. 8. *Surface water height H_T at $t = 1200$ s, computed with 20 elements (solid), 40 elements (−), and 80 elements (+).*

TABLE 1
Total ground water flux across Γ_{GS} for the three different meshes.

Mesh	Total flux
1	−360.01
2	−359.98
3	−359.99

As another check on the numerical solutions, we examined a "quantity of interest," namely, the total ground water flux across the interface Γ_{GS}. By setting $w_g \equiv 1$ in (20), we see that the true flux satisfies in this case

$$\int_0^T \int_{\Gamma_{GS}} \mathbf{u}_g \cdot \mathbf{n}_g ds dt = \int_0^T \int_{\Omega_g} f_g dx dt = \int_0^T \bar{f} dt = -.3T.$$

Our numerical solution $\boldsymbol{U}_g$ satisfies

$$\int_0^T \int_{\Gamma_{GS}} [\boldsymbol{U}_g \cdot \mathbf{n}_g + \sigma_g (H_g - H_T)] ds dt = \int_0^T \int_{\Omega_g} f_g dx dt.$$

In Table 1, we have computed

$$\int_0^T \int_{\Gamma_{GS}} \boldsymbol{U}_g \cdot \mathbf{n}_g ds dt$$

for the three meshes used to compute the solutions above with $T = 1200$. As observed in this table, the total flux agrees with the exact value of -360 to about five significant digits; the difference shows the effect of the penalty term $\sigma_g(H_g - H_T)$.

7. Conclusions. In this paper, we have analyzed DG and MFE methods for ground water/surface water coupling. The analysis gives the expected order of accuracy for the ground water velocity $\mathbf{u}_g$ and surface water velocity $\mathbf{u}_s$ but is suboptimal

for the ground water head h_g and the surface water height h_s. This appears to be unavoidable due to the coupling but deserves further study.

Our preliminary numerical results indicate that the ground water response to rapid changes in the surface water height is fairly slow, and one does not need to solve the ground water flow equations on the time scale of the surface water time step. Our results also suggest the need to use finer spatial discretization in the surface water domain than in the ground water domain. Numerical issues in temporal and spatial scaling which arise in these couplings will be the subject of future research, as will extensions to higher dimensions and to multiphase ground water flow.

REFERENCES

[1] V. Aizinger and C. N. Dawson, *Discontinuous Galerkin methods for two-dimensional flow and transport in shallow water*, Advances in Water Resources, 25 (2002), pp. 67–84.

[2] M. P. Anderson and W. W. Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*, Academic Press, San Diego, 1992.

[3] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[4] T. Arbogast, C. N. Dawson, P. T. Keenan, M. F. Wheeler, and I. Yotov, *Enhanced cell-centered finite differences for elliptic equations on general geometry*, SIAM J. Sci. Comput., 19 (1998), pp. 404–425.

[5] T. Arbogast and H. Lehr, *Homogenization of a Darcy-Stokes system modeling vuggy porous media*. Comput. Geosci., to appear.

[6] T. Arbogast, M. F. Wheeler, and I. Yotov, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[7] D. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[8] G. Beavers and D. Joseph, *Boundary conditions at a naturally impermeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.

[9] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[10] F. Brezzi, J. Douglas, Jr., R. Duran, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[11] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[12] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.

[13] Z. Cai, J. E. Jones, S. F. McCormick, and T. F. Russell, *Control-volume mixed finite element methods*, Comput. Geosci., 1 (1997), pp. 289–316.

[14] M. B. Cardenas and V. A. Zlotnik, *Three-dimensional model of modern channel bend deposits*, Water Resources Research, 39 (2003), 1141.

[15] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[16] Z. Chen and J. Douglas, Jr., *Prismatic mixed finite elements for second-order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[17] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[18] C. N. Dawson and J. Proft, *Discontinuous and coupled continuous/discontinuous Galerkin methods for the shallow water equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4721–4746.

[19] C. N. Dawson and J. Proft, *Coupled discontinuous and continuous Galerkin finite element methods for the depth-integrated shallow water equations*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 289–318.

[20] C. Dawson, S. Sun, and M. F. Wheeler, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[21] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.
[22] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.
[23] R. A. Freeze and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
[24] R. A. Freeze and R. L. Harlan, *Blue-print for a physically-based digitally simulated hydrologic response model*, J. Hydrology, 9 (1960), pp. 237–258.
[25] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.
[26] S. J. Kollet and R. M. Maxwell, *Integrated Surface-Groundwater Flow Modeling: A Free-Surface Overland Flow Boundary Condition in a Parallel Groundwater Flow Model*, preprint.
[27] W. J. Layton, F. Schieweck, and I. Yotov, *Coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 40 (2003), pp. 2195–2218.
[28] P. Lesaint and P. A. Raviart, *On a finite element method for solving the neutron transport equations*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974.
[29] Q. Lu, M. Peszynska, and M. F. Wheeler, *A parallel multiblock black-oil model in multimodel implementation*, SPE Journal, 3 (2002), pp. 278–287.
[30] E. Miglio, M. Discacciati, and A. Quarteroni, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Math., 43 (2002), pp. 57–74.
[31] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys, 146 (1998), pp. 491–519.
[32] M. Peszynska, Q. Lu, and M. F. Wheeler, *Coupling different numerical algorithms for two phase fluid flow*, in Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X, J. R. Whiteman, ed., Elsevier, Oxford, UK, 2000, pp. 205–214.
[33] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.
[34] B. Rivière, M. F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrain and discontinuous Galerkin methods for elliptic problems* I, Comput. Geosci., 3 (1999), pp. 337–360.
[35] B. Rivière and I. Yotov, *Locally conservative coupling of Stokes and Darcy flows*, SIAM J. Numer. Anal., 42 (2005), pp. 1959–1977.
[36] P. Saffman, *On the boundary condition at the surface of a porous media*, Stud. Appl. Math., 50 (1971), pp. 292–315.
[37] J. H. Schmidt and L. C. Roig, *The adaptive hydrology (adh) model: A flow and transport model for coupled surface water-groundwater analyses*, in Proceedings of Theme C, Groundwater: An Endangered Resource, XXVII IAHR Congress, A. N. Findikakis, ed., ASCE, Reston, VA, 1997, pp. 367–372.
[38] V. Singh and S. M. Bhallamudi, *Conjunctive surface-subsurface modeling of overland flow*, Advances in Water Resources, 21 (1998), pp. 567–579.
[39] C. A. Talbot, C. W. Downer, H.-C. Lin, S. E. Howington, and D. Richards, *Computational mdethods for simualting interaction between surface and subsurface hydrologic systems*, in Computational Methods in Water Resources XIV, Vol. 2, W. G. Gray, S. M. Hassanizadeh, R. J. Schotting, and G. F. Pinder, eds., Elsevier, Amsterdam, 2002, pp. 1511–1518.
[40] H. E. VanderKwaak and K. Loague, *Hydrologic-response simulations for the r-5 catchment with a comprehensive physics-based model*, Water Resources Research, 37 (2001), pp. 999–1013.
[41] A. Weiser and M. F. Wheeler, *On convergence of block-centered finite differences for elliptic problems*, SIAM J. Numer. Anal., 25 (1988), pp. 351–375.
[42] T. Weiyan, *Shallow Water Hydrodynamics*, Elsevier Oceanography Series 55, Elsevier, Amstedam, 1992.
[43] M. F. Wheeler and R. Gonzales, *Mixed finite element methods for petroleum reservoir engineering problems*, in Computing Methods in Applied Sciences and Engineering VI, R. Glowinski and J. L. Lions, eds., North–Holland, New York, 1984, pp. 639–658.
[44] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.
[45] T. C. Winter, J. W. Harvey, O. L. Franke, and W. M. Alley, *Ground Water and Surface Water: A Single Resource*. U. S. Geological Survey Circular 1139, United States Geological Survey, Denver, CO, 1998.

[46] G. T. Yeh, P. Cheng, R. Cheng, J. Lin, and W. D. Martin, *A Numerical Model Simulating Water Flow and Contaminant and Sediment Transport in Watershed Systems of* 1-*d Stream-River Network,* 2-*d Overland Regime and* 3-*d Subsurface Media (wash*123*d: Version* 1.0*)*, Technical report CHL-98-19, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, 1998.

[47] G. T. Yeh, R. Cheng, M. Li, P. Cheng, and J. Lin, *COSFLOW: A Finite Element Model Coupling One-Dimensional Canal, Two-Dimensional Overland, and Three-Dimensional Subsurface Flow*, Technical report CHL-97-20, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, 1997.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1405–1419

CONVERGENCE OF AN IMPLICIT FINITE ELEMENT METHOD FOR THE LANDAU–LIFSHITZ–GILBERT EQUATION*

SÖREN BARTELS† AND ANDREAS PROHL‡

Abstract. The Landau–Lifshitz–Gilbert equation describes the dynamics of ferromagnetism, where strong nonlinearity and nonconvexity are hard to tackle: so far, existing explicit schemes to approximate weak solutions suffer from severe time-step restrictions. In this paper, we propose an implicit fully discrete scheme and verify unconditional convergence.

Key words. ferromagnetism, Landau–Lifshitz–Gilbert equation, nonconvexity, finite elements, convergence

AMS subject classifications. 35K55, 65M12, 65M15, 68U10, 94A08

DOI. 10.1137/050631070

1. Introduction.

The phenomenological Landau–Lifshitz–Gilbert equation (LLG) describes the dynamics of ferromagnetism; let $\alpha \geq 0$ denote the damping parameter, and then the magnetization $\mathbf{m} : (0,T)\times\Omega \to S^2$, for $S^2 = \{\mathbf{x} \in \mathbb{R}^3 | \, |\mathbf{x}| = 1\}$, solves

$$\mathbf{m}_t = -\alpha\, \mathbf{m} \times (\mathbf{m} \times \Delta\mathbf{m}) + \mathbf{m} \times \Delta\mathbf{m}\,, \tag{1.1}$$

supplemented by initial and boundary conditions, $\mathbf{m}(0) = \mathbf{m}_0 \in W^{1,2}(\Omega; S^2)$, and $\partial_{\mathbf{n}}\mathbf{m} = 0$ on $(0,T)\times\partial\Omega$. A proper definition of weak solutions is given below. Limiting equations are the Heisenberg equation ($\alpha \to 0$) and heat flow for harmonic maps ($\alpha \to \infty$) (see [1, Propositions 5.1, 5.2]):

$$\mathbf{m}_t = \mathbf{m} \times \Delta\mathbf{m} \quad (\alpha \to 0)\,, \qquad \mathbf{m}_t = \Delta\mathbf{m} + |\nabla\mathbf{m}|^2\mathbf{m} \quad (\alpha \to \infty)\,. \tag{1.2}$$

The construction of convergent schemes for (1.1) is a nontrivial task, due to the nonconvex side-constraint $|\mathbf{m}| = 1$ a.e. in $(0,T)\times\Omega$, which is difficult to realize in a numerical approximation scheme. Explicit time integrators of high order coupled with occasional updates to ensure the sphere constraint are common strategies in engineering literature but suffer from nonreliable dynamics [5]. Implicit strategies to discretize LLG in time often introduce artificial damping, which prevents computed iterates from remaining on the sphere and excludes a (discrete) energy law to hold to conclude convergence. Remedies have been made, partially addressing the dual requirements of efficiency and reliability: (i) projection methods have been constructed, which independently deal with the nonconvex algebraic constraint; however, no (discrete) energy principle is available, and convergence to LLG is known only in the case of existing classical/strong solutions to LLG (see [4, 10, 9]); (ii) explicit/implicit discretizations of Ginzburg–Landau penalizations that involve an additional parameter

*Received by the editors May 9, 2005; accepted for publication (in revised form) February 16, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sinum/44-4/63107.html

†Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (sba@math.hu-berlin.de). This author's research was supported by Deutsche Forschungsgemeinschaft through the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

‡Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

$\varepsilon > 0$ are used [8, 9], which allow for a discrete energy principle, possibly for restricted choices of spatiotemporal discretization parameters; see [5]. We refer the reader to [7] for a recent review of mathematical ferromagnetism.

Recently, a first explicit scheme is proposed in [2], where also (weak sub-) convergence towards weak solutions is verified; this program is continued in [3], where $k = o\big(\alpha^2 h^{1+\frac{n}{2}}\big)$ is identified to be sufficient for stability and convergence; sharpness of these restrictions is evidenced by computational studies in [3]. From this background, we look for an implicit scheme exempt from restricting requirements for numerical parameters, and with higher flexibility with respect to (small) choices of $\alpha > 0$. The construction of our discretization is based on a reformulation of (1.1) by Gilbert (see, e.g., [7]),

$$\mathbf{m}_t + \alpha\, \mathbf{m} \times \mathbf{m}_t = (1+\alpha^2)\, \mathbf{m} \times \Delta \mathbf{m}\,.$$

Given the lowest order finite element space $\mathbf{V}_h \subset W^{1,2}(\Omega;\mathbb{R}^3)$ subordinate to a triangulation $\mathcal{T}_h$ of Ω and a time-step size $k > 0$, our approximation scheme reads as follows.

ALGORITHM 1.1. *Let* $\mathbf{m}_h^0 \in \mathbf{V}_h$. *Given* $j \geq 0$ *and* $\mathbf{m}_h^j \in \mathbf{V}_h$, *determine* $\mathbf{m}_h^{j+1} \in \mathbf{V}_h$ *from*

$$\begin{aligned}(d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h + \alpha\,(\mathbf{m}_h^j \times d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h \\ = (1+\alpha^2)(\overline{\mathbf{m}}_h^{j+1/2} \times \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\phi}_h)_h \quad \forall\, \boldsymbol{\phi}_h \in \mathbf{V}_h\,.\end{aligned}$$

Here $(\cdot,\cdot)_h$ denotes a discrete version (reduced integration) of the inner product in $L^2(\Omega;\mathbb{R}^3)$, $\tilde{\Delta}_h : W^{1,2}(\Omega;\mathbb{R}^3) \to \mathbf{V}_h$ is a discrete version of the Laplace operator, and we use $d_t\varphi^j := k^{-1}\big(\varphi^j - \varphi^{j-1}\big)$ for $j \geq 1$ and $\overline{\varphi}^{j+1/2} := \frac{1}{2}\big(\varphi^{j+1} + \varphi^j\big)$ for $j \geq 0$ and a sequence $\{\varphi^j\}_{j\geq 0}$; we refer the reader to section 2 for details.

REMARK 1.1. *The (linear) second term in Algorithm* 1.1 *is motivated by the identity*

$$\mathbf{m}_h^j \times d_t \mathbf{m}_h^{j+1} = \left(\overline{\mathbf{m}}_h^{j+1/2} - \frac{k}{2} d_t \mathbf{m}_h^{j+1}\right) \times d_t \mathbf{m}_h^{j+1} = \overline{\mathbf{m}}_h^{j+1/2} \times d_t \mathbf{m}_h^{j+1}\,.$$

It is well known that weak solutions to (1.1) solve

$$\mathbf{m}_t = \operatorname{div}\big(\mathbf{m} \times \nabla \mathbf{m}\big) + \alpha\big(\Delta \mathbf{m} + |\,\nabla \mathbf{m}\,|^2 \mathbf{m}\big)$$

in the distributional sense; cf. [1, 6]. Corresponding relations need not hold for discretizations, due to the competition of local and nonlocal aspects inherent to fully discrete finite element based methods. Lemma 6.1 below shows that solutions of Algorithm 1.1 satisfy

(1.3)
$$\begin{aligned}(d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h + \alpha\big(\nabla \overline{\mathbf{m}}_h^{j+1/2}, \nabla \boldsymbol{\phi}_h\big) = \alpha\big(|\,\nabla \overline{\mathbf{m}}_h^{j+1/2}\,|^2 \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\phi}_h\big) \\ - \big(\overline{\mathbf{m}}_h^{j+1/2} \times \nabla \overline{\mathbf{m}}_h^{j+1/2}, \nabla \boldsymbol{\phi}_h\big) + \mathrm{Corr}\end{aligned}$$

for all $\boldsymbol{\phi}_h \in \mathbf{V}_h$ and a correcting term "Corr".

Lemma 3.1 below states conservation of $|\,\mathbf{m}_h^j\,| = 1$ at the nodes of the triangulation $\mathcal{T}_h$ and verifies a discrete energy law for solutions to Algorithm 1.1. This indicates that the forcing correction term Corr serves to balance the damping effect of the implicit Euler method with employed reduced integration and local averaging

tools in Algorithm 1.1. The unconditional stability of Algorithm 1.1 allows us to prove subconvergence to a weak solution of (1.1).

The remainder of this paper is organized as follows. Preliminaries are stated in section 2. Our main result is Theorem 3.1, which verifies unconditional convergence for Algorithm 1.1; a simple fixed-point iteration is proposed in Algorithm 4.1, whose convergence is established for $k = \mathcal{O}(h^2)$, uniformly for values $\alpha \leq C$, in section 4. We discuss numerical experiments which motivate finite-time blow-up in section 5, allowing for direct comparison with results for values $\alpha = \mathcal{O}(1)$ in [3] and study of the limiting case $\alpha \to 0$. Section 6 proves (1.3) and illustrates difficulties in the construction of convergent implicit finite element schemes.

2. Preliminaries. Throughout this paper we assume that $\mathcal{T}_h$ is a quasi-uniform regular triangulation of the polygonal or polyhedral bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ into triangles or tetrahedra for $n = 2$ or $n = 3$, respectively. We define the lowest order finite element space $\mathbf{V}_h \subset W^{1,2}(\Omega;\mathbb{R}^3)$ by

$$\mathbf{V}_h = \left\{ \boldsymbol{\phi}_h \in C(\overline{\Omega};\mathbb{R}^3) : \boldsymbol{\phi}_h|_K \in \mathcal{P}_1(K;\mathbb{R}^3) \quad \forall\, K \in \mathcal{T}_h \right\},$$

where $\mathcal{P}_1(K;\mathbb{R}^3)$ denotes the set of polynomials of total degree less than or equal to one restricted to the element $K \in \mathcal{T}_h$. Given the set of nodes $\left\{ \mathbf{x}_\ell : \ell \in L \right\}$ of the triangulation $\mathcal{T}_h$, the nodal interpolation operator $\mathcal{I}_h : C(\overline{\Omega};\mathbb{R}^3) \to \mathbf{V}_h$ satisfies $\mathcal{I}_h\boldsymbol{\phi}(\mathbf{x}_\ell) = \boldsymbol{\phi}(\mathbf{x}_\ell)$ for all $\ell \in L$. Given functions $\mathbf{f}, \mathbf{g} \in L^2(\Omega;\mathbb{R}^m)$ and letting $\langle \cdot, \cdot \rangle$ denote the inner product in $\mathbb{R}^m$ we set

$$\big(\mathbf{f}, \mathbf{g}\big) = \int_\Omega \langle \mathbf{f}, \mathbf{g} \rangle \, \mathrm{d}\mathbf{x}\,.$$

For continuous functions $\boldsymbol{\phi}, \boldsymbol{\mathfrak{Z}} \in C(\overline{\Omega};\mathbb{R}^3)$ we define

$$\big(\boldsymbol{\phi},\, \boldsymbol{\mathfrak{Z}}\big)_h = \int_\Omega \mathcal{I}_h\big(\langle \boldsymbol{\phi}, \boldsymbol{\mathfrak{Z}} \rangle\big) \, \mathrm{d}\mathbf{x} = \sum_{\ell \in L} \beta_\ell \langle \boldsymbol{\phi}(\mathbf{x}_\ell), \boldsymbol{\mathfrak{Z}}(\mathbf{x}_\ell) \rangle$$

for certain weights $\beta_\ell > 0$, $\ell \in L$. If for each $\ell \in L$ we denote by $\varphi_\ell \in C(\overline{\Omega})$ the nodal basis function which is $\mathcal{T}_h$-elementwise affine and satisfies $\varphi_\ell(\mathbf{x}_\ell) = 1$ and $\varphi_\ell(\mathbf{x}_m) = 0$ for all $m \in L \setminus \{\ell\}$, then we have $\beta_\ell = \int_\Omega \varphi_\ell \, \mathrm{d}\mathbf{x}$. We define $||\boldsymbol{\phi}||_h^2 = \big(\boldsymbol{\phi},\, \boldsymbol{\phi}\big)_h$ and notice that

$$\| \boldsymbol{\phi}_h \|_{L^2}^2 \leq \| \boldsymbol{\phi}_h \|_h^2 \leq (n+2) \, \| \boldsymbol{\phi}_h \|_{L^2}^2$$

for all $\boldsymbol{\phi}_h \in V_h$. We define a discrete Laplace operator $\tilde{\Delta}_h : W^{1,2}(\Omega;\mathbb{R}^3) \to \mathbf{V}_h$ by

$$-(\tilde{\Delta}_h\boldsymbol{\phi}, \boldsymbol{\chi}_h)_h = (\nabla\boldsymbol{\phi}, \nabla\boldsymbol{\chi}_h) \quad \forall\, \boldsymbol{\chi}_h \in \mathbf{V}_h\,. \tag{2.1}$$

It is well known that there exists a constant $c_1 > 0$ such that for all $\boldsymbol{\phi}_h \in \mathbf{V}_h$ there holds

$$||\nabla\boldsymbol{\phi}_h||_{L^2} \leq c_1 h^{-1} ||\boldsymbol{\phi}_h||_{L^2}, \tag{2.2}$$

where h is the maximal mesh-size in $\mathcal{T}_h$, i.e., $h = \max\{\mathrm{diam}(K) : K \in \mathcal{T}_h\}$. Choosing $\boldsymbol{\chi}_h = \tilde{\Delta}_h\boldsymbol{\phi}_h$ in (2.1) and using (2.2) we observe that for all $\boldsymbol{\phi}_h \in \mathbf{V}_h$ there holds

$$||\tilde{\Delta}_h\boldsymbol{\phi}_h||_h^2 = -\big(\nabla\boldsymbol{\phi}_h, \nabla\tilde{\Delta}_h\boldsymbol{\phi}_h\big) \leq ||\nabla\boldsymbol{\phi}_h||_{L^2} ||\nabla\tilde{\Delta}_h\boldsymbol{\phi}_h||_{L^2} \leq c_1 h^{-1} ||\nabla\boldsymbol{\phi}_h||_{L^2} ||\tilde{\Delta}_h\boldsymbol{\phi}_h||_h. \tag{2.3}$$

Given $\boldsymbol{\phi}_h \in \mathbf{V}_h$ and a node $\mathbf{x}_\ell$ for some $\ell \in L$, we obtain from using $\boldsymbol{\chi}_h = \varphi_\ell \tilde{\Delta}_h \boldsymbol{\phi}_h(\mathbf{x}_\ell)$ in (2.1) that

$$
\begin{aligned}
|\tilde{\Delta}_h \boldsymbol{\phi}_h(\mathbf{x}_\ell)|^2 &= \beta_\ell^{-1} \big(\tilde{\Delta}_h \boldsymbol{\phi}_h, \boldsymbol{\chi}_h\big)_h = -\beta_\ell^{-1} \big(\nabla \boldsymbol{\phi}_h, \nabla \boldsymbol{\chi}_h\big) \\
&= -\beta_\ell^{-1} \sum_{\substack{m \in L : \exists K \in \mathcal{T}, \\ \mathbf{x}_m, \mathbf{x}_\ell \in K}} \langle \boldsymbol{\phi}_h(\mathbf{x}_m), \tilde{\Delta}_h \boldsymbol{\phi}_h(\mathbf{x}_\ell) \rangle (\nabla \varphi_m, \nabla \varphi_\ell) \\
&\le c_2 h^{-2} ||\boldsymbol{\phi}_h||_{L^\infty} |\tilde{\Delta}_h \boldsymbol{\phi}_h(\mathbf{x}_\ell)|,
\end{aligned} \tag{2.4}
$$

where we used (2.2), that given a node $\mathbf{x}_\ell$ the cardinality of the set $\{m \in L : \exists K, \mathbf{x}_m, \mathbf{x}_\ell \in K\}$ is bounded h-independently, and that $||\varphi_m||_{L^2} \le c \beta_\ell^{1/2}$ for all $m \in L$.

3. Unconditional convergence. We first recall the definition of a weak solution to LLG. Throughout this section we abbreviate $\Omega_T = (0,T) \times \Omega$.

DEFINITION 3.1. *Let $\mathbf{m}_0 \in W^{1,2}(\Omega; S^2)$; then $\mathbf{m}$ is called the weak solution to LLG if for all $T > 0$*

(1) *$\mathbf{m} \in W^{1,2}(\Omega_T; \mathbb{R}^3)$ such that $|\,\mathbf{m}\,| = 1$ a.e. in Ω_T;*

(2) *for all $\boldsymbol{\phi} \in C^\infty\big(\Omega_T; \mathbb{R}^3\big)$ there holds*

$$
\begin{aligned}
&\int_{\Omega_T} \langle \mathbf{m}_t, \boldsymbol{\phi} \rangle \, \mathrm{d}\mathbf{x}\mathrm{d}t + \alpha \int_{\Omega_T} \langle \mathbf{m} \times \mathbf{m}_t, \boldsymbol{\phi} \rangle \, \mathrm{d}\mathbf{x}\mathrm{d}t \\
&= \qquad -(1+\alpha^2) \int_{\Omega_T} \langle \mathbf{m} \times \nabla \mathbf{m}, \nabla \boldsymbol{\phi} \rangle \, \mathrm{d}\mathbf{x}\mathrm{d}t \,;
\end{aligned}
$$

(3) *$\mathbf{m}(0) = \mathbf{m}_0$ in the sense of traces;*

(4) *for almost all $T' \in (0,T)$ there holds*

$$
\frac{1}{2} \int_\Omega |\,\nabla \mathbf{m}(T')\,|^2 \, \mathrm{d}\mathbf{x} + \frac{\alpha}{1+\alpha^2} \int_{\Omega_{T'}} |\,\mathbf{m}_t\,|^2 \, \mathrm{d}\mathbf{x}\mathrm{d}t \le \frac{1}{2} \int_\Omega |\,\nabla \mathbf{m}_0\,|^2 \, \mathrm{d}\mathbf{x} \,.
$$

The following lemma provides discrete counterparts of (1) and (4). We remark that the well-posedness of Algorithm 1.1, i.e., the existence of a unique sequence $\{\mathbf{m}_h^j\}_{j \ge 0}$ that solves Algorithm 1.1, can be deduced from a classical argument; see, e.g., [1, sect. 3].

LEMMA 3.1. *Suppose that $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ for all $\ell \in L$. Then the sequence $\{\mathbf{m}_h^j\}_{j \ge 0}$ produced by Algorithm 1.1 satisfies for all $j \ge 0$*

$$
\begin{aligned}
&\text{(i)} \quad |\,\mathbf{m}_h^{j+1}(\mathbf{x}_\ell)\,| = 1 \quad \forall \ell \in L \,, \\
&\text{(ii)} \quad \frac{1}{2} d_t \| \nabla \mathbf{m}_h^{j+1} \|_{L^2}^2 + \frac{\alpha}{1+\alpha^2} \| \, d_t \mathbf{m}_h^{j+1} \|_h^2 = 0 \,.
\end{aligned}
$$

Proof. Verification of (i) follows from choosing $\boldsymbol{\phi}_h = \varphi_\ell \overline{\mathbf{m}}_h^{j+1/2}(\mathbf{x}_\ell) \in \mathbf{V}_h$ for $\ell \in L$ in Algorithm 1.1. In order to verify (ii), we first choose $\boldsymbol{\phi}_h = -\tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}$ and find

$$
\frac{1}{2} d_t \| \nabla \mathbf{m}_h^{j+1} \|_{L^2}^2 + \alpha \big(\overline{\mathbf{m}}_h^{j+1/2} \times d_t \mathbf{m}_h^{j+1}, \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}\big)_h = 0 \,.
$$

Choosing $\boldsymbol{\phi}_h = d_t \mathbf{m}_h^{j+1}$ yields

$$
\frac{\alpha}{1+\alpha^2} \| \, d_t \mathbf{m}_h^{j+1} \|_h^2 = \alpha \big(\overline{\mathbf{m}}_h^{j+1/2} \times \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}, d_t \mathbf{m}_h^{j+1}\big)_h .
$$

A combination of the two identities proves (ii) and completes the proof of the lemma. ∎

DEFINITION 3.2. *For* $\mathbf{x} \in \Omega$ *and* $t \in [t_j, t_{j+1})$ *define*

$$\mathbf{M}(t,\mathbf{x}) := \frac{t-t_j}{k}\mathbf{m}_h^{j+1}(\mathbf{x}) + \frac{t_{j+1}-t}{k}\mathbf{m}_h^{j}(\mathbf{x})\,,$$
$$\mathbf{M}^-(t,\mathbf{x}) := \mathbf{m}_h^{j}(\mathbf{x})\,, \quad \mathbf{M}^+(t,\mathbf{x}) := \mathbf{m}_h^{j+1}(\mathbf{x})\,, \quad \overline{\mathbf{M}}(t,\mathbf{x}) := \overline{\mathbf{m}}_h^{j+1/2}\,.$$

Given any $T' > 0$, (ii) in Lemma 3.1 may be rewritten as

$$\frac{1}{2}||\nabla \mathbf{M}^+(T')||_{L^2}^2 + \frac{\alpha}{1+\alpha^2}\int_0^{T'} ||\mathbf{M}_t||_h^2 \,\mathrm{d}t \le \frac{1}{2}||\nabla \mathbf{M}(0)||_{L^2}^2.$$

This bound yields the existence of some $\mathbf{m} \in W^{1,2}(\Omega_T;\mathbb{R}^3)$ which is the weak limit (as $k,h \to 0$) of a subsequence such that

$$\begin{gathered}\mathbf{M} \rightharpoonup \mathbf{m} \ \text{ in } W^{1,2}(\Omega_T,\mathbb{R}^3)\,,\\ \nabla\mathbf{M}^-,\, \nabla\mathbf{M}^+, \nabla\overline{\mathbf{M}} \rightharpoonup \nabla\mathbf{m} \ \text{ in } L^2(\Omega_T,\mathbb{R}^3)\,,\\ \mathbf{M}^-,\, \mathbf{M}^+,\, \overline{\mathbf{M}} \to \mathbf{m} \ \text{ in } L^2(\Omega_T,\mathbb{R}^3)\,.\end{gathered}$$

Since $|\,\mathbf{M}^-(t,\mathbf{x}_\ell)\,| = 1$ for every $\ell \in L$ and almost all $t \in (0,T)$, there holds for every $K \in \mathcal{T}_h$

$$\begin{aligned}\|\,|\,\mathbf{M}^-\,|^2 - 1\,\|_{L^2(K)} &\le Ch\,\|\,\nabla\big(|\,\mathbf{M}^-\,|^2-1\big)\,\|_{L^2(K)} = Ch\,\|\,2(\nabla\mathbf{M}^-)\mathbf{M}^-\,\|_{L^2(K)}\\ &\le 2Ch\,\|\,\nabla\mathbf{M}^-\,\|_{L^2(K)}\,,\end{aligned}$$

which implies $|\,\mathbf{M}^-\,| \to 1$ in $L^2(\Omega_T;\mathbb{R}^3)$, and hence $|\,\mathbf{m}\,| = 1$ a.e. in Ω_T. Algorithm 1.1 may be written as follows: taking $\boldsymbol{\phi}_h(t) := \mathcal{I}_h\boldsymbol{\phi}(t,\cdot)$, for $\boldsymbol{\phi} \in C^\infty(\Omega_T;\mathbb{R}^3)$, there holds

(3.1)
$$\int_0^T (\mathbf{M}_t,\boldsymbol{\phi}_h)_h\,\mathrm{d}t + \alpha\int_0^T (\mathbf{M}^-\times\mathbf{M}_t,\boldsymbol{\phi}_h)_h\,\mathrm{d}t = (1+\alpha^2)\int_0^T \big(\overline{\mathbf{M}}\times\tilde{\Delta}_h\overline{\mathbf{M}},\boldsymbol{\phi}_h\big)_h\,\mathrm{d}t\,.$$

Effects of reduced integration are controlled using the fact that for all $\boldsymbol{\chi}_h, \boldsymbol{\eta}_h \in \mathbf{V}_h$ there holds

$$|\,(\boldsymbol{\chi}_h,\boldsymbol{\eta}_h)_h - (\boldsymbol{\chi}_h,\boldsymbol{\eta}_h)\,| \le Ch\,\|\,\boldsymbol{\chi}_h\,\|_{L^2}\|\,\nabla\boldsymbol{\eta}_h\,\|_{L^2}.$$

This implies that for almost all $t \in (0,T)$ we have

$$\big|(\mathbf{M}_t,\boldsymbol{\phi}_h)_h - (\mathbf{M}_t,\boldsymbol{\phi}_h)\big| \le Ch\,\|\,\mathbf{M}_t\,\|_{L^2}\|\,\nabla\boldsymbol{\phi}_h\,\|_{L^2}$$

and allows us to prove that

$$\int_0^T (\mathbf{M}_t,\boldsymbol{\phi}_h)_h\,\mathrm{d}t \to \int_0^T (\mathbf{m}_t,\boldsymbol{\phi})\,\mathrm{d}t.$$

Using that for $\boldsymbol{\chi}_h \in \mathbf{V}_h$ and $\boldsymbol{\eta} \in C(\overline{\Omega};\mathbb{R}^3)$ there holds $(\boldsymbol{\chi}_h,\boldsymbol{\eta})_h = (\boldsymbol{\chi}_h,\mathcal{I}_h\boldsymbol{\eta})_h$, and employing a triangle inequality and standard estimates for nodal interpolation results in

$$\big|(\mathbf{M}_t,\mathbf{M}^-\times\boldsymbol{\phi}_h)_h - \big(\mathbf{M}_t,(\mathrm{Id}\pm\mathcal{I}_h)(\mathbf{M}^-\times\boldsymbol{\phi}_h)\big)\big| \le Ch\,\|\,\mathbf{M}_t\,\|_{L^2}\|\,\nabla(\mathbf{M}^-\times\boldsymbol{\phi}_h)\,\|_{L^2}\,.$$

This yields

$$\int_0^T (\overline{\mathbf{M}} \times \mathbf{M}_t, \boldsymbol{\phi}_h)_h \mathrm{d}t \to \int_0^T (\mathbf{m} \times \mathbf{m}_t, \boldsymbol{\phi}) \mathrm{d}t\,.$$

The only troublesome limit is for the last term in (3.1). We write

$$\begin{aligned}\big(\overline{\mathbf{M}} \times \tilde{\Delta}_h \overline{\mathbf{M}}, \boldsymbol{\phi}_h\big)_h &= (\overline{\mathbf{M}} \times \boldsymbol{\phi}_h, \tilde{\Delta}_h \overline{\mathbf{M}})_h = \big((\mathrm{Id} - \mathcal{I}_h)(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h), \tilde{\Delta}_h \overline{\mathbf{M}}\big)_h \\ &\quad + \big(\nabla(\mathcal{I}_h - \mathrm{Id})(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h), \nabla \overline{\mathbf{M}}\big) + \big(\nabla(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h), \nabla \overline{\mathbf{M}}\big) =: I + II + III\,.\end{aligned}$$

Control of I uses the bound $\| \tilde{\Delta}_h \boldsymbol{\chi} \|_{L^2} \le c_1 h^{-1} \| \nabla \boldsymbol{\chi} \|_{L^2}$ and estimates for nodal interpolation

$$\begin{aligned} I &\le C h^2 h^{-1} \sum_{K \in \mathcal{T}_h} \| D^2(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h) \|_{L^2(K)} \| \nabla \overline{\mathbf{M}} \|_{L^2(K)} \\ &\le C h \| \nabla \overline{\mathbf{M}} \|_{L^2} \| \nabla \boldsymbol{\phi}_h \|_{L^\infty} \| \nabla \overline{\mathbf{M}} \|_{L^2}\,.\end{aligned}$$

A similar argumentation proves

$$II \le C h \sum_{K \in \mathcal{T}_h} \| D^2\big(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h\big) \|_{L^2(K)} \| \nabla \overline{\mathbf{M}} \|_{L^2(K)} \le C h \| \nabla \overline{\mathbf{M}} \|_{L^2} \| \nabla \boldsymbol{\phi}_h \|_{L^\infty} \| \nabla \overline{\mathbf{M}} \|_{L^2}.$$

We use that given any $\boldsymbol{\mathfrak{Z}}, \boldsymbol{\chi} \in W^{1,2}(\Omega; \mathbb{R}^3)$ there holds $\langle \nabla \boldsymbol{\mathfrak{Z}}, \nabla(\boldsymbol{\mathfrak{Z}} \times \boldsymbol{\chi}) \rangle = \langle \nabla \boldsymbol{\mathfrak{Z}}, \boldsymbol{\mathfrak{Z}} \times \nabla \boldsymbol{\chi} \rangle$ to verify

$$III = \big(\nabla(\overline{\mathbf{M}} \times \boldsymbol{\phi}_h), \nabla \overline{\mathbf{M}}\big) = \big(\overline{\mathbf{M}} \times \nabla \boldsymbol{\phi}_h, \nabla \overline{\mathbf{M}}\big).$$

A combination of the last four assertions shows

$$\int_0^T \big(\overline{\mathbf{M}} \times \tilde{\Delta}_h \overline{\mathbf{M}}, \boldsymbol{\phi}_h\big)_h \,\mathrm{d}t \to \int_0^T (\mathbf{m} \times \nabla \boldsymbol{\phi}, \nabla \mathbf{m})\,\mathrm{d}t = \int_0^T \big(\nabla(\mathbf{m} \times \boldsymbol{\phi}), \nabla \mathbf{m}\big)\,\mathrm{d}t\,.$$

This proves our main theorem.

THEOREM 3.1. *Suppose* $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ *for all* $\ell \in L$, *and let* $\{\mathbf{m}_h^j\}_{j \ge 0}$ *solve Algorithm* 1.1. *Assume that* $\mathbf{m}_h^0 \to \mathbf{m}_0$ *in* $W^{1,2}(\Omega, \mathbb{R}^3)$ *for* $h \to 0$. *For* $k, h \to 0$ *there exists* $\mathbf{m} \in W^{1,2}(\Omega_T; \mathbb{R}^3)$ *such that* $\mathbf{M}$ *subconverges to* $\mathbf{m}$ *in* $W^{1,2}(\Omega_T, \mathbb{R}^3)$, *and* $\mathbf{m}$ *is a weak solution of LLG.*

4. Solving the nonlinear system. In the numerical experiments reported below we employ the following fixed-point iteration to solve the nonlinear system in Algorithm 1.1.

ALGORITHM 4.1. *Set* $\mathbf{m}_h^{j+1,0} := \mathbf{m}_h^j$ *and* $\ell := 0$.

(i) *Compute* $\mathbf{m}_h^{j+1,\ell+1} \in \mathbf{V}_h$ *such that for all* $\boldsymbol{\phi}_h \in \mathbf{V}_h$ *there holds*

(4.1)
$$\begin{aligned}\frac{1}{k}(\mathbf{m}_h^{j+1,\ell+1}, \boldsymbol{\phi}_h)_h + \frac{\alpha}{k}(\mathbf{m}_h^j \times \mathbf{m}_h^{j+1,\ell+1}, \boldsymbol{\phi}_h)_h - \frac{1+\alpha^2}{4}(\mathbf{m}_h^{j+1,\ell+1} \times \tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell}, \boldsymbol{\phi}_h)_h \\ - \frac{1+\alpha^2}{4}(\mathbf{m}_h^{j+1,\ell+1} \times \tilde{\Delta}_h \mathbf{m}_h^j, \boldsymbol{\phi}_h)_h - \frac{1+\alpha^2}{4}(\mathbf{m}_h^j \times \tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell+1}, \boldsymbol{\phi}_h)_h \\ = \frac{1}{k}(\mathbf{m}_h^j, \boldsymbol{\phi}_h)_h + \frac{1+\alpha^2}{4}(\mathbf{m}_h^j \times \tilde{\Delta}_h \mathbf{m}_h^j, \boldsymbol{\phi}_h)_h.\end{aligned}$$

(ii) *If* $||\mathbf{m}_h^{j+1,\ell+1} - \mathbf{m}_h^{j+1,\ell}||_h \le \varepsilon$, *then stop and set* $\mathbf{m}_h^{j+1} := \mathbf{m}_h^{j+1,\ell+1}$.
(iii) *Set* $\ell := \ell + 1$ *and go to* (i).

Setting $\varepsilon = 0$, the following lemma shows that the iteration converges, provided that $k \le ch^2/(1+\alpha^2)$ for an (h,k,α)-independent constant factor $c > 0$ that depends only on the geometry of $\mathcal{T}_h$.

LEMMA 4.1. *Suppose that* $|\mathbf{m}_h^j(\mathbf{x}_m)| \le c_3$ *for some* $c_3 > 0$ *and all* $m \in L$ *and that* $\gamma := \sqrt{5}(1+\alpha^2)c_1^2 c_3 h^{-2} k/4 < 1$. *Then for all* $\ell \ge 0$ *there exists a unique solution* $\mathbf{m}_h^{j+1,\ell+1}$ *to* (4.1). *For all* $\ell \ge 1$ *there holds*

$$||\mathbf{m}_h^{j+1,\ell+1} - \mathbf{m}_h^{j+1,\ell}||_h \le \Theta \frac{\gamma}{1-\gamma} ||\mathbf{m}_h^{j+1,\ell} - \mathbf{m}_h^{j+1,\ell-1}||_h, \tag{4.2}$$

provided that $\Theta := \frac{1+c_3\rho}{(1/c_3)-\rho} > 0$ *for* $\rho := (1+\alpha^2)c_2 k h^{-2}/4$. *Moreover, for all* $\ell \ge 0$ *and all* $\boldsymbol{\phi}_h \in \mathbf{V}_h$ *there holds*

$$\begin{aligned} &\big|(d_t \mathbf{m}_h^{j+1,\ell+1}, \boldsymbol{\phi}_h)_h \;+\; \alpha(\mathbf{m}_h^j \;\times\; d_t \mathbf{m}_h^{j+1,\ell+1}, \boldsymbol{\phi}_h)_h \\ &\qquad -(1 \;+\; \alpha^2)(\overline{\mathbf{m}}_h^{j+1/2,\ell+1} \;\times\; \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2,\ell+1}, \boldsymbol{\phi}_h)_h\big| \\ &\le \Theta\sqrt{5}\frac{1+\alpha^2}{4} c_1^2 h^{-2} ||\mathbf{m}_h^{j+1,\ell+1} - \mathbf{m}_h^{j+1,\ell}||_h ||\boldsymbol{\phi}_h||_h, \end{aligned}$$

where $d_t \mathbf{m}_h^{j+1,\ell+1} = k^{-1}(\mathbf{m}_h^{j+1,\ell+1} - \mathbf{m}_h^j)$ *and* $\overline{\mathbf{m}}_h^{j+1/2,\ell+1} = \frac{1}{2}(\mathbf{m}_h^{j+1,\ell+1} + \mathbf{m}_h^j)$.

By the Banach fixed-point theorem, contraction property (4.2) implies for $|\,\mathbf{m}_h^j(\mathbf{x}_m)\,| = 1$ for all $m \in L$ the existence of a unique $\mathbf{m}_h^{j+1,*} \in \mathbf{V}_h$ which solves Algorithm 1.1 for $j' := j$, and thus again satisfies Lemma 3.1.

Proof. We abbreviate $\mu = (1+\alpha^2)/4$. For $\boldsymbol{\phi}_h = \mathbf{m}_h^{j+1,\ell+1}$ the left-hand side of (4.1) is bounded from below by

$$\begin{aligned} &\frac{1}{k}||\mathbf{m}_h^{j+1,\ell+1}||_h^2 - \mu(\mathbf{m}_h^j \times \tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell+1}, \mathbf{m}_h^{j+1,\ell+1})_h \\ &\quad \ge \frac{1}{k}||\mathbf{m}_h^{j+1,\ell+1}||_h^2 - \mu||\mathbf{m}_h^j||_{L^\infty}||\tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell+1}||_h ||\mathbf{m}_h^{j+1,\ell+1}||_h \\ &\quad \ge \left(\frac{1}{k} - \mu\sqrt{5}c_1^2 c_3 h^{-2}\right) ||\mathbf{m}_h^{j+1,\ell+1}||_h^2, \end{aligned}$$

where we used $||\mathbf{m}_h^j||_{L^\infty} \le c_3$ and $||\tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell+1}||_h \le c_1^2\sqrt{5}h^{-2}||\mathbf{m}_h^{j+1,\ell+1}||_h$. Therefore, the bilinear form defined by the left-hand side of (4.1) is positive definite on $\mathbf{V}_h \times \mathbf{V}_h$ if $\gamma < 1$, and then (4.1) admits a unique solution. Let $m \in L$ be such that $||\mathbf{m}_h^{j+1,\ell}||_{L^\infty} = |\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)|$. Choosing $\boldsymbol{\phi}_h = \varphi_m \mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)$ in the equation for $\mathbf{m}_h^{j+1,\ell}$ proves

$$\begin{aligned} \frac{1}{k}|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)|^2 &\le \mu|\mathbf{m}_h^j(\mathbf{x}_m)|\,|\tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)|\,|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)| \\ &\quad + \frac{1}{k}|\mathbf{m}_h^j(\mathbf{x}_m)|\,|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)| + \mu|\mathbf{m}_h^j(\mathbf{x}_m)|\,|\tilde{\Delta}_h \mathbf{m}_h^j(\mathbf{x}_m)|\,|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)| \\ &\le c_3\mu|\tilde{\Delta}_h \mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)|\,|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)| + c_3\frac{1}{k}|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)| \\ &\quad + c_3\mu|\tilde{\Delta}_h \mathbf{m}_h^j(\mathbf{x}_m)|\,|\mathbf{m}_h^{j+1,\ell}(\mathbf{x}_m)|. \end{aligned}$$

There holds $|\tilde{\Delta}_h \boldsymbol{\phi}_h(\mathbf{x}_m)| \le c_2 h^{-2}||\boldsymbol{\phi}_h||_{L^\infty}$ for all $\boldsymbol{\phi}_h \in \mathbf{V}_h$, and hence

$$||\mathbf{m}_h^{j+1,\ell}||_{L^\infty} \le \frac{1 + c_3 k\mu c_2 h^{-2}}{(1/c_3) - k\mu c_2 h^{-2}} = \Theta.$$

Subtraction of two subsequent equations in the fixed-point iteration yields

$$\begin{aligned}
&\frac{1}{k}(\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell},\boldsymbol{\phi}_h)_h+\frac{\alpha}{k}(\mathbf{m}_h^{j}\times[\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}],\boldsymbol{\phi}_h)_h\\
&-\mu([\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}]\times\tilde\Delta_h\mathbf{m}_h^{j+1,\ell},\boldsymbol{\phi}_h)_h-\mu(\mathbf{m}_h^{j+1,\ell}\times\tilde\Delta_h[\mathbf{m}_h^{j+1,\ell}-\mathbf{m}_h^{j+1,\ell-1}],\boldsymbol{\phi}_h)_h\\
&-\mu([\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}]\times\tilde\Delta_h\mathbf{m}_h^{j},\boldsymbol{\phi}_h)_h-\mu(\mathbf{m}_h^{j}\times\tilde\Delta_h[\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}],\boldsymbol{\phi}_h)_h=0
\end{aligned}$$

for all $\boldsymbol{\phi}_h\in\mathbf{V}_h$. Choosing $\boldsymbol{\phi}_h:=\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}$ shows

$$\begin{aligned}
||\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}||_h&\le k\mu\Theta||\tilde\Delta_h[\mathbf{m}_h^{j+1,\ell}-\mathbf{m}_h^{j+1,\ell-1}]||_h\\
&\quad+c_3k\mu||\tilde\Delta_h[\mathbf{m}_h^{j+1,\ell+1}-\mathbf{m}_h^{j+1,\ell}]||_h.
\end{aligned}$$

Using $||\tilde\Delta_h\boldsymbol{\phi}_h||_h\le c_1^2\sqrt{5}h^{-2}||\boldsymbol{\phi}_h||_h$ for all $\boldsymbol{\phi}_h\in\mathbf{V}_h$ we deduce the first estimate of the lemma. In order to verify the second estimate we notice that owing to (4.1), $\mathbf{m}_h^j\times\mathbf{m}_h^j=0$, and the above estimate $||\mathbf{m}_h^{j+1,\ell+1}||_{L^\infty}\le\Theta$ there holds for all $\boldsymbol{\phi}_h\in\mathbf{V}_h$

$$\begin{aligned}
&(d_t\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h+\alpha(\mathbf{m}_h^{j}\times d_t\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h-\mu(\overline{\mathbf{m}}_h^{j+1/2,\ell+1}\times\tilde\Delta_h\overline{\mathbf{m}}_h^{j+1/2,\ell+1},\boldsymbol{\phi}_h)_h\\
&\quad=\frac{1}{k}(\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h-\frac{1}{k}(\mathbf{m}_h^{j},\boldsymbol{\phi}_h)_h\\
&\qquad+\frac{\alpha}{k}(\mathbf{m}_h^{j}\times\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h-\frac{\alpha}{k}(\mathbf{m}_h^{j}\times\mathbf{m}_h^{j},\boldsymbol{\phi}_h)_h\\
&\qquad-\ \mu(\mathbf{m}_h^{j+1,\ell+1}\times\tilde\Delta_h\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h-\mu(\mathbf{m}_h^{j+1,\ell+1}\times\tilde\Delta_h\mathbf{m}_h^{j},\boldsymbol{\phi}_h)_h\\
&\qquad-\mu(\mathbf{m}_h^{j}\times\tilde\Delta_h\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h-\mu(\mathbf{m}_h^{j}\times\tilde\Delta_h\mathbf{m}_h^{j},\boldsymbol{\phi}_h)_h\\
&\quad=\mu(\mathbf{m}_h^{j+1,\ell+1}\times\tilde\Delta_h\mathbf{m}_h^{j+1,\ell},\boldsymbol{\phi}_h)_h-\mu(\mathbf{m}_h^{j+1,\ell+1}\times\tilde\Delta_h\mathbf{m}_h^{j+1,\ell+1},\boldsymbol{\phi}_h)_h\\
&\quad\le\mu||\mathbf{m}_h^{j+1,\ell+1}||_{L^\infty}||\tilde\Delta_h(\mathbf{m}_h^{j+1,\ell}-\mathbf{m}_h^{j+1,\ell+1})||_h||\boldsymbol{\phi}_h||_h\\
&\quad\le\Theta\mu c_1^2\sqrt{5}h^{-2}||\mathbf{m}_h^{j+1,\ell}-\mathbf{m}_h^{j+1,\ell+1}||_h||\boldsymbol{\phi}_h||_h,
\end{aligned}$$

which completes the proof of the lemma. ∎

5. Numerical experiments. The implementation of Algorithms 1.1 and 4.1 was performed in MATLAB with an assemblation of the stiffness matrices in C. We set $\varepsilon=h^4$ for the termination criterion in Algorithm 4.1, and it terminated after at most five iterations in all of our experiments. The experiments are defined through the following example which is taken from [3].

EXAMPLE 5.1. *Let $\Omega=(-1/2,1/2)^2$, and let $\mathbf{m}_0:\Omega\to S^2$ be defined by*

$$\mathbf{m}_0(\mathbf{x})=\begin{cases}(0,0,-1) & \text{for } |\mathbf{x}|\ge 1/2,\\ \big(2\mathbf{x}A,A^2-|\mathbf{x}|^2\big)/\big(A^2+|\mathbf{x}|^2\big) & \text{for } |\mathbf{x}|\le 1/2,\end{cases}$$

where $A:=(1-2|\mathbf{x}|)^4/s$ for some $s>0$. The triangulations $\mathcal{T}_\ell$ used in the numerical simulations are defined through a positive integer ℓ and consist of $2^{2\ell+1}$ halved squares with edge length $h=2^{-\ell}$. Motivated by Lemma 4.1 we use $k=h^2/(10(1+\alpha^2))$. As discrete initial data we employ the nodal interpolant of $\mathbf{m}_0$; i.e., we set $\mathbf{m}^0=\mathcal{I}_{\mathcal{T}_\ell}\mathbf{m}_0$ in all experiments.

Figures 1 and 2 display snapshots of the numerical approximation provided by Algorithm 1.1 with $\alpha=1$, $s=1$, and $\ell=4$. The plots in Figure 1 display the first two components of the vector field $\mathbf{M}$ at the nodes of the triangulation (after an appropriate rescaling) and at various times. Figure 2 shows a zoom towards the

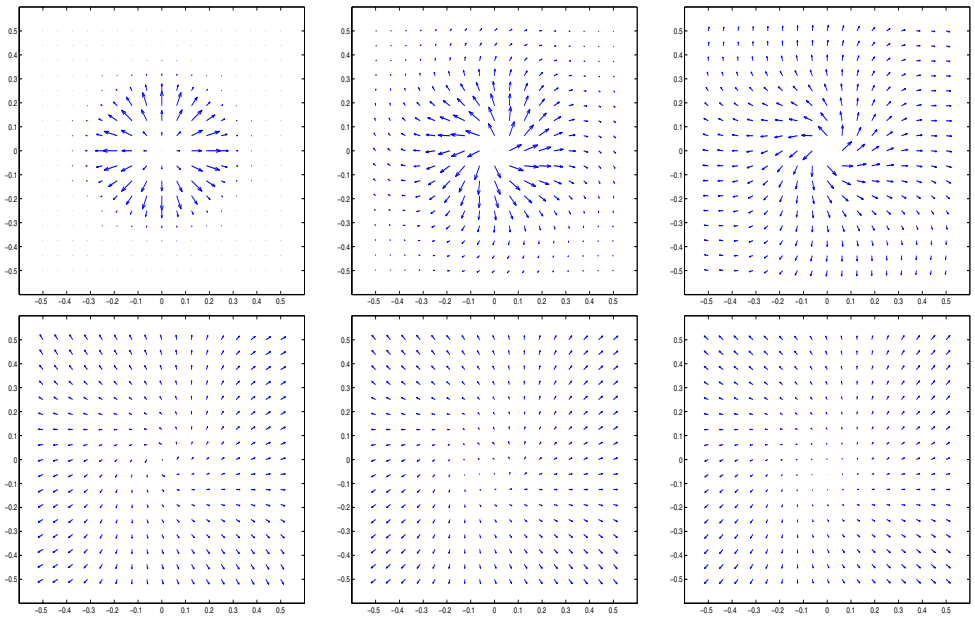


FIG. 1. *Numerical approximation* $\mathbf{M}(t,\cdot)$ *in Example* 5.1 *with* $s = 1$, $\ell = 4$, *and* $\alpha = 1$ *for* $t = 0, 0.0119, 0.0295, 0.0529, 0.0588, 0.0646$.

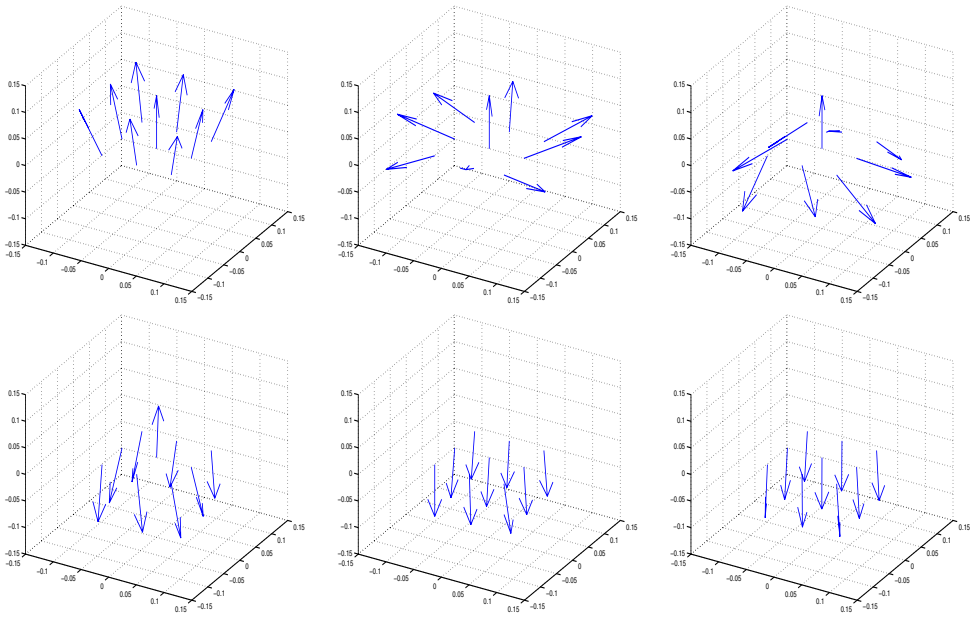


FIG. 2. *Nodal values* $\mathbf{M}(t,\mathbf{x}_m)$ *for nodes* $\mathbf{x}_m$ *close to the origin in Example* 5.1 *with* $s = 1$, $\ell = 4$, *and* $\alpha = 1$ *for* $t = 0, 0.0119, 0.0295, 0.0529, 0.0588, 0.0646$.

origin and reveals that in this experiment regularity of the exact solution cannot be expected. At time $t \approx 0.0529$ the vector at the origin points in another direction than all surrounding vectors, resulting in a large (maximal) $W^{1,\infty}$ norm.

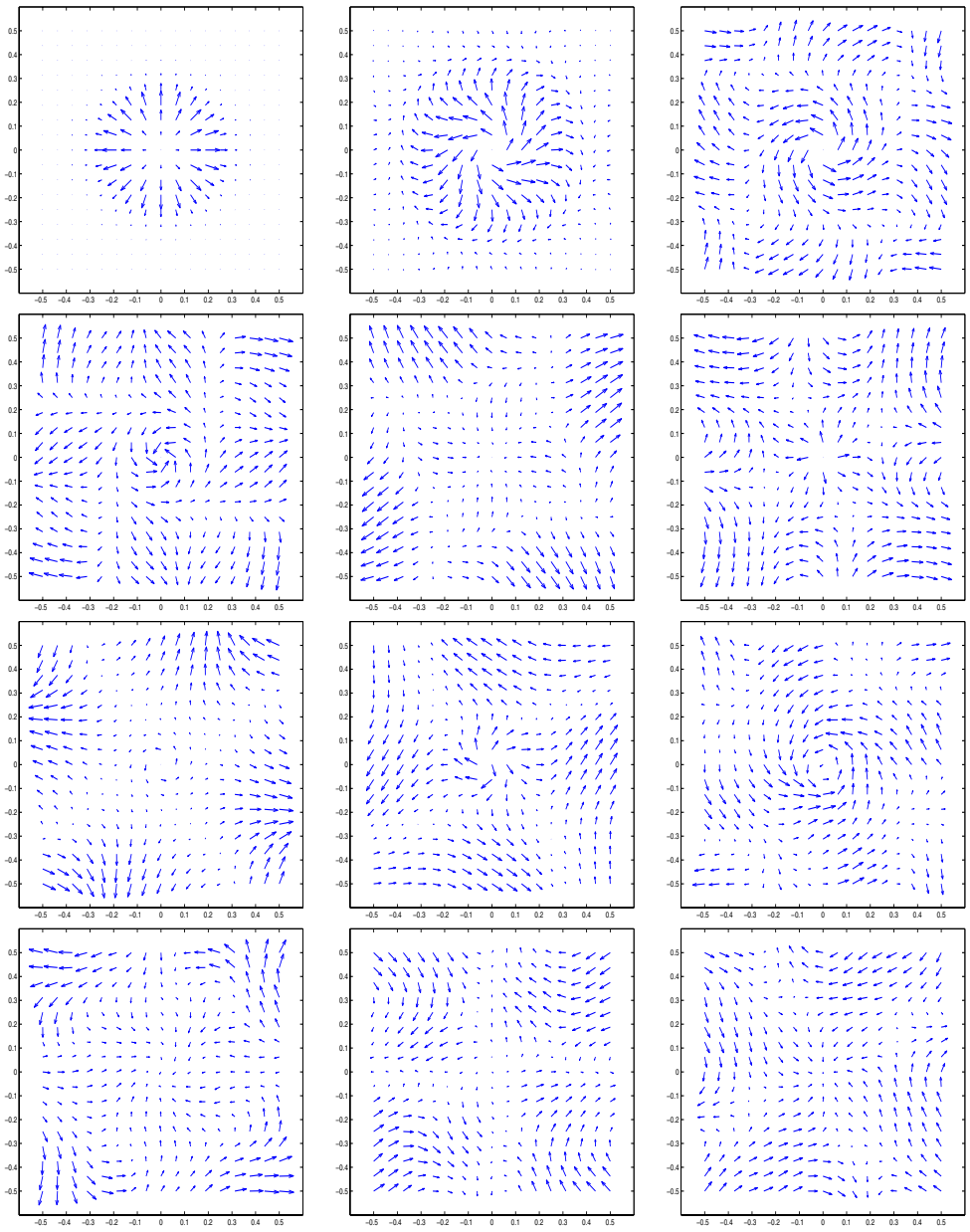


FIG. 3. *Numerical approximation* $\mathbf{M}(t,\cdot)$ *in Example* 5.1 *with* $s=1$, $\ell=4$, *and* $\alpha=1/64$ *for* $t=0,0.0102,0.0297,0.0492,0.0687,0.1078,0.1371,0.1664,0.2054,0.2347,0.2738,0.3031$.

Figures 3 and 4 show similar snapshots for $\alpha=1/64$, $s=1$, and $\ell=4$. Owing to the significantly smaller stabilization corresponding to the small value of α, the numerical solution is even less regular than in the previous experiment and fails to become stationary for times $t\leq 1/2$.

For fixed $\alpha=1$ and $s=4$ we used $\ell=4,5,6$ in Example 5.1. In Figure 5 we

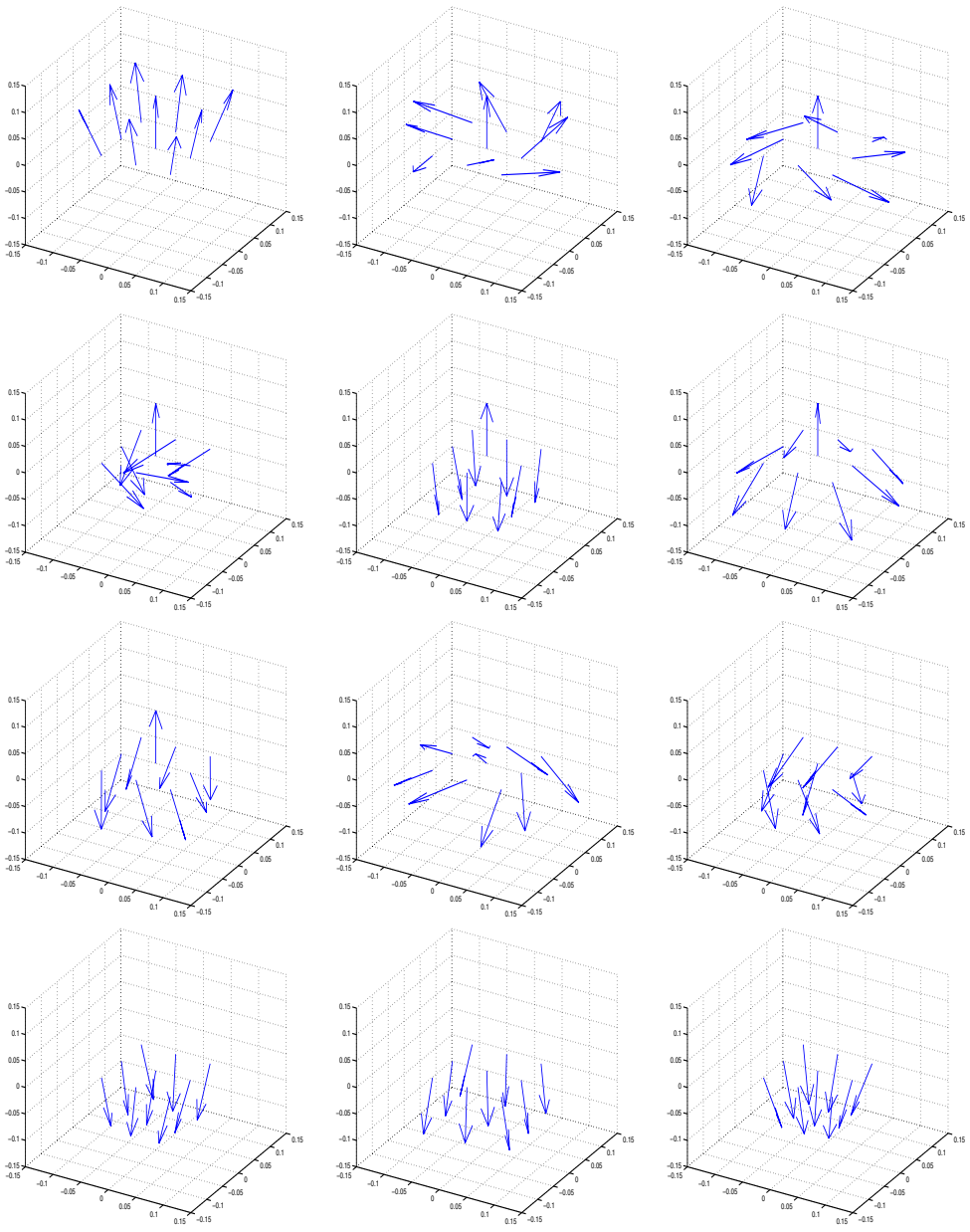


FIG. 4. *Nodal values* $\mathbf{M}(t, \mathbf{x}_m)$ *for nodes* $\mathbf{x}_m$ *close to the origin in Example* 5.1 *with* $s = 1$, $\ell = 4$, *and* $\alpha = 1/64$ *for* $t = 0, 0.0102, 0.0297, 0.0492, 0.0687, 0.1078, 0.1371, 0.1664, 0.2054, 0.2347, 0.2738, 0.3031$.

displayed the energy

$$E(\mathbf{M}(t)) = \frac{1}{2}\int_\Omega |\nabla \mathbf{M}(t,\cdot)|^2 \, \mathrm{d}\mathbf{x}$$

and the $W^{1,\infty}$ seminorm $|\mathbf{M}(t)|_{1,\infty} = ||\nabla \mathbf{M}(t)||_{L^\infty}$ as functions of t for $t \in (0, 6/100)$ for $\ell = 4, 5, 6$. For each $\ell = 4, 5, 6$ the function $t \to ||\nabla \mathbf{M}(t)||_{L^\infty}$ assumes the

maximum value $2\sqrt{2}h^{-1}$ (among functions $\boldsymbol{\phi}_h \in \mathbf{V}_h$ with $|\boldsymbol{\phi}_h(\mathbf{x}_m)| = 1$ for all nodes $\mathbf{x}_m$). We observe that for decreasing mesh-size h the blow-up time (the time at which $||\nabla \mathbf{M}(t)||_{L^\infty}$ assumes its maximum) approaches $t \approx 0.03$. In order to study the dependence of blow-up behavior on the parameter α we ran Algorithm 1.1 in Example 5.1 for fixed $\ell = 5$, $s = 1$ and for $\alpha = 1, 1/4, 1/16, 1/64, 1/256$. The plot in Figure 6 indicates that the blow-up time approaches the time $t \approx 0.06$ for decreasing α. The experimental values for $\alpha = 1/64$ and $\alpha = 1/256$ almost coincide.

We remark that the results of our experiments for $\alpha = 1, 1/4, 1/16$ are similar to the results obtained in [3] with an explicit scheme. The implicit scheme of this article allows us to use smaller values for α which lead to too restrictive conditions on the time step size for the explicit scheme of [3]. For the triangulations employed here and for $\alpha = 1$ the total runtimes of the explicit scheme (using reduced integration) and the implicit scheme are comparable. We stress, however, that for small values of α or three-dimensional problems the explicit scheme from [3] is of limited practical use.

6. Proof of (1.3).

LEMMA 6.1. *Assume that* $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ *for all* $\ell \in L$, *and let* $\{\mathbf{m}_h^j\}_{j\geq 0}$ *solve Algorithm* 1.1. *There holds for all* $\boldsymbol{\phi}_h \in \mathbf{V}_h$

$$(d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h + \alpha\big(\nabla \overline{\mathbf{m}}_h^{j+1/2}, \nabla \boldsymbol{\phi}_h\big) = \alpha\big(|\nabla \overline{\mathbf{m}}_h^{j+1/2}|^2 \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\phi}_h\big) - \big(\overline{\mathbf{m}}_h^{j+1/2} \times \nabla \overline{\mathbf{m}}_h^{j+1/2}, \nabla \boldsymbol{\phi}_h\big) + \mathrm{Corr}$$

for a correcting term $\mathrm{Corr} = \mathrm{Corr_A} + \mathrm{Corr_B}$, *with*

$$\mathrm{Corr_A} := \frac{\alpha}{2}\big(\nabla|\overline{\mathbf{m}}_h^{j+1/2}|^2, \nabla\langle \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\phi}_h\rangle\big) + \frac{\alpha^2}{1+\alpha^2}\Big(d_t \mathbf{m}_h^{j+1}, [1 - |\overline{\mathbf{m}}_h^{j+1/2}|^2]\boldsymbol{\phi}_h\Big)_h + \alpha\Big(\nabla \overline{\mathbf{m}}_h^{j+1/2}, \nabla[(1 - |\overline{\mathbf{m}}_h^{j+1/2}|^2)\boldsymbol{\phi}_h]\Big),$$

and $\mathrm{Corr_B} = \sum_{i=1}^3 \mathrm{Corr_{B_i}}$ *given in the proof below.*

Proof. Given any $\boldsymbol{\mathfrak{Z}}_h \in \mathbf{V}_h$ choose $\boldsymbol{\phi}_h = \mathcal{I}_h(\overline{\mathbf{m}}_h^{j+1/2} \times \boldsymbol{\mathfrak{Z}}_h)$ in Algorithm 1.1; then the properties of $(\cdot,\cdot)_h$ imply

$$\big(\overline{\mathbf{m}}_h^{j+1/2} \times d_t \mathbf{m}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h\big)_h + \alpha\big(\overline{\mathbf{m}}_h^{j+1/2} \times (\overline{\mathbf{m}}_h^{j+1/2} \times d_t \mathbf{m}_h^{j+1}), \boldsymbol{\mathfrak{Z}}_h\big)_h = (1+\alpha^2)\big(\overline{\mathbf{m}}_h^{j+1/2} \times (\overline{\mathbf{m}}_h^{j+1/2} \times \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}), \boldsymbol{\mathfrak{Z}}_h\big)_h. \tag{6.1}$$

Owing to $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \langle \mathbf{a}, \mathbf{c}\rangle \mathbf{b} - \langle \mathbf{a}, \mathbf{b}\rangle \mathbf{c}$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$, the second term on the left-hand side in (6.1) may be rewritten as

$$\alpha\big(\langle \overline{\mathbf{m}}_h^{j+1/2}, d_t \mathbf{m}_h^{j+1}\rangle \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h\big)_h - \alpha\big(|\overline{\mathbf{m}}_h^{j+1/2}|^2 d_t \mathbf{m}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h\big)_h = \frac{\alpha}{2}\big((d_t|\mathbf{m}_h^{j+1}|^2)\overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h\big)_h - \alpha\big(|\overline{\mathbf{m}}_h^{j+1/2}|^2 d_t \mathbf{m}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h\big)_h,$$

and the first term on the left-hand side vanishes owing to Lemma 3.1. We again use the above vector identity to recast the right-hand side of (6.1) as

$$(1+\alpha^2)\big(\langle \overline{\mathbf{m}}_h^{j+1/2}, \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}\rangle \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h\big)_h - (1+\alpha^2)\big(|\overline{\mathbf{m}}_h^{j+1/2}|^2 \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h\big)_h. \tag{6.2}$$

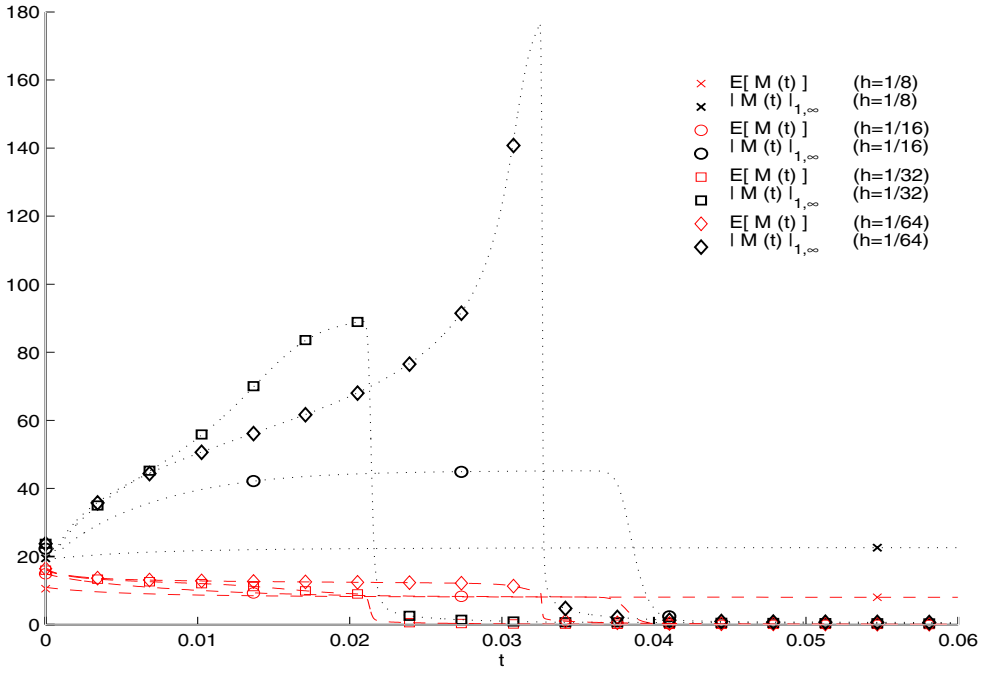


FIG. 5. *Energy and* $W^{1,\infty}$ *seminorm for decreasing mesh-sizes in Example* 5.1 *with* $\alpha = 1$ *and* $s = 4$.

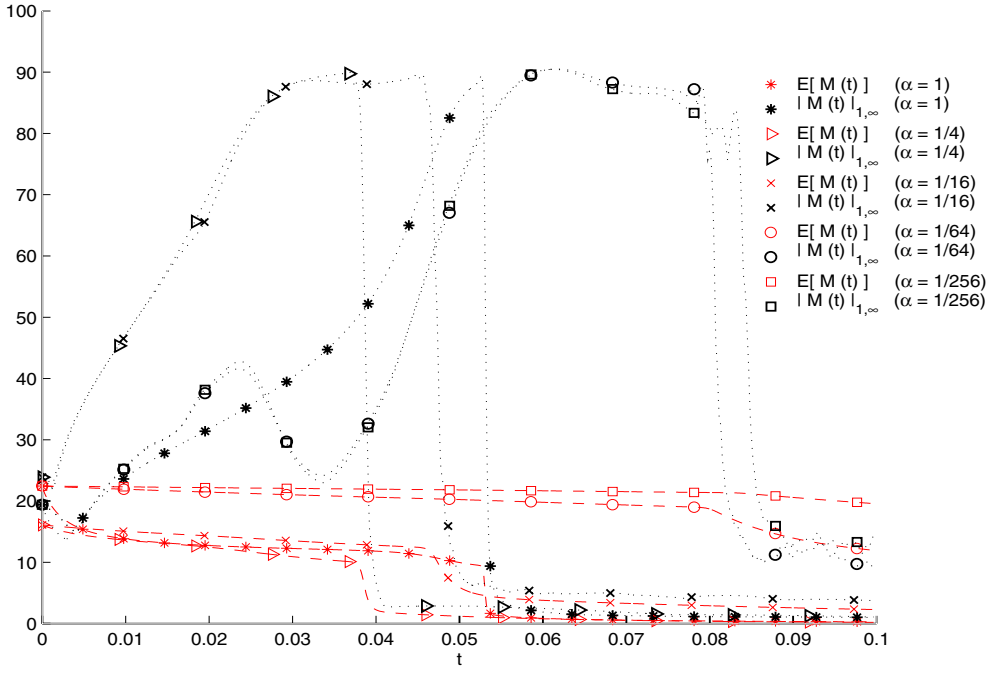


FIG. 6. *Energy and* $W^{1,\infty}$ *seminorm in Example* 5.1 *for* $\ell = 5$, $s = 1$, *and* $\alpha = 1, 1/4, 1/16, 1/64, 1/256$.

We proceed independently with arising two terms: intermitting the Lagrange interpolant for the nonlinear term in the first case to benefit from (2.1) yields

$$\begin{aligned}
&\Big(\big(\mathrm{Id}\pm\mathcal{I}_h\big)\big(\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\overline{\mathbf{m}}_h^{j+1/2}\big),\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2}\big)\Big)_h\\
&\qquad=\Big(\big(\mathrm{Id}-\mathcal{I}_h\big)\big(\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\overline{\mathbf{m}}_h^{j+1/2}\big),\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2}\big)\Big)_h\\
&+\Big(\nabla\big((\mathrm{Id}-\mathcal{I}_h)(\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\overline{\mathbf{m}}_h^{j+1/2})\big),\nabla\overline{\mathbf{m}}_h^{j+1/2}\Big)\\
&-\Big(\nabla\big(\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\overline{\mathbf{m}}_h^{j+1/2}\big),\nabla\overline{\mathbf{m}}_h^{j+1/2}\Big),
\end{aligned}$$

where the first two terms on the right-hand side are referred to as Corr_{B_1}. For the last term, we resume

$$\begin{aligned}
\big(\nabla\overline{\mathbf{m}}_h^{j+1/2},\nabla(\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\overline{\mathbf{m}}_h^{j+1/2})\big)&=\big(|\,\nabla\overline{\mathbf{m}}_h^{j+1/2}\,|^2\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\big)\\
&\quad+\frac{1}{2}\big(\nabla|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2,\nabla\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\big).
\end{aligned}$$

Similarly, we account for effects of reduced integration and local averaging inherent to the scheme for the second term in (6.2),

$$\begin{aligned}
\Big(\big(\mathrm{Id}\pm\mathcal{I}_h\big)\big(|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2\mathfrak{Z}_h\big),\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2}\Big)_h&=\Big(\big(\mathrm{Id}-\mathcal{I}_h\big)\big(|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2\mathfrak{Z}_h\big),\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2}\Big)_h\\
+\Big(\nabla\big((\mathrm{Id}-\mathcal{I}_h)(|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2\mathfrak{Z}_h)\big),\nabla\overline{\mathbf{m}}_h^{j+1/2}\Big)&-\big(\nabla(|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2\mathfrak{Z}_h),\nabla\overline{\mathbf{m}}_h^{j+1/2}\big),
\end{aligned}$$

where the first two terms on the right-hand side are gathered in Corr_{B_2}. Finally, by Algorithm 1.1 and $\langle\mathbf{a}\times\mathbf{b},\mathbf{c}\rangle=-\langle\mathbf{a}\times\mathbf{c},\mathbf{b}\rangle$, the first term in (6.1) is identical to

$$\begin{aligned}
&-\frac{1}{\alpha}\big(d_t\mathbf{m}_h^{j+1},\mathfrak{Z}_h\big)_h+\frac{1+\alpha^2}{\alpha}\big(\overline{\mathbf{m}}_h^{j+1/2}\times\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\big)_h\\
&\qquad=-\frac{1}{\alpha}\big(d_t\mathbf{m}_h^{j+1},\mathfrak{Z}_h\big)_h+\frac{1+\alpha^2}{\alpha}\big(\nabla(\overline{\mathbf{m}}_h^{j+1/2}\times\mathfrak{Z}_h),\nabla\overline{\mathbf{m}}_h^{j+1/2}\big)+\mathrm{Corr}_{B_3}
\end{aligned}$$

for $\mathrm{Corr}_{B_3}=-((\mathrm{Id}-\mathcal{I}_h)(\overline{\mathbf{m}}_h^{j+1/2}\times\mathfrak{Z}_h),\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2})_h-(\nabla\big((\mathrm{Id}-\mathcal{I}_h)(\overline{\mathbf{m}}_h^{j+1/2}\times\mathfrak{Z}_h)\big),\nabla\overline{\mathbf{m}}_h^{j+1/2})$. Reassembling (6.1) then yields to

$$\begin{aligned}
\big((1+\alpha^2|\,\mathbf{m}_h^{j+1/2}\,|^2)d_t\mathbf{m}_h^{j+1},\mathfrak{Z}_h\big)_h&=(1+\alpha^2)\big(\nabla(\overline{\mathbf{m}}_h^{j+1/2}\times\mathfrak{Z}_h),\nabla\overline{\mathbf{m}}_h^{j+1/2}\big)\\
+\alpha(1+\alpha^2)\Big[\big(|\,\nabla\overline{\mathbf{m}}_h^{j+1/2}\,|^2\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\big)&+\frac{1}{2}\big(\nabla|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2,\nabla\langle\overline{\mathbf{m}}_h^{j+1/2},\mathfrak{Z}_h\rangle\big)\\
-\big(\nabla\overline{\mathbf{m}}_h^{j+1/2},\nabla(|\,\overline{\mathbf{m}}_h^{j+1/2}\,|^2\mathfrak{Z}_h)\big)\Big]&+\alpha\Big(\mathrm{Corr}_{\mathrm{A}}+\big(1+\alpha^2\big)\,\mathrm{Corr}_{\mathrm{B}}\Big).
\end{aligned}$$

Rearranging terms then yields to the assertion. ∎

Acknowledgments. Part of the work was written when S. B. visited "Forschungsinstitut für Mathematik" (ETH Zürich) in January, 2005. S. B. gratefully acknowledges the hospitality of the Department of Mathematics at the University of Maryland at College Park.

REFERENCES

[1] F. Alouges and A. Soyeur, *On global weak solutions for Landau–Lifshitz equations: Existence and nonuniqueness*, Nonlinear Anal., 18 (1992), pp. 1071–1084.

[2] F. Alouges and P. Jaisson, *Convergence of a finite element discretization for the Landau–Lifshitz equations*, in Micromagnetism, Math. Models Methods Appl. Sci., 16 (2006), pp. 299–316.

[3] S. Bartels, J. Ko, and A. Prohl, *Numerical Approximation of Landau–Lifshitz Equations and Finite Time Blow Up of Weak Solutions*, http://www.fim.math.ethz.ch/preprints/2005/ (2005).

[4] W. E and X.-P. Wang, *Numerical methods for the Landau–Lifshitz equation*, SIAM J. Numer. Anal., 38 (2000), pp. 1647–1665.

[5] J. Fidler and T. Schrefl, *Micromagnetic modelling—the current state of the art*, J. Phys. D: Appl. Phys., 33 (2000), pp. R135–R156.

[6] B. Guo and M.-C. Hong, *The Landau–Lifshitz equation of the ferromagnetic spin chain and harmonic maps*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 311–334.

[7] M. Kružík and A. Prohl, *Recent developments in modeling, analysis, and numerics of ferromagnetism*, SIAM Rev., 48 (2006), pp. 439–483.

[8] F. Pistella and V. Valente, *Numerical stability for a discrete model in the dynamics of ferromagnetic bodies*, Methods Partial Differential Equations, 15 (1999), pp. 544–557.

[9] A. Prohl, *Computational Micromagnetism*, Teubner, Stuttgart, 2001.

[10] X.-P. Wang, C. J. García-Cervera, and W. E, *A Gauss-Seidel projection method for micromagnetics simulations*, J. Comput. Phys., 171 (2001), pp. 357–372.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1420–1440

ANALYSIS OF A MULTISCALE DISCONTINUOUS GALERKIN METHOD FOR CONVECTION-DIFFUSION PROBLEMS*

A. BUFFA†, T. J. R. HUGHES‡, AND G. SANGALLI§

Abstract. We study a multiscale discontinuous Galerkin method introduced in [T. J. R. Hughes, G. Scovazzi, P. Bochev, and A. Buffa, *Comput. Meth. Appl. Mech. Engrg.*, 195 (2006), pp. 2761–2787] that reduces the computational complexity of the discontinuous Galerkin method, seemingly without adversely affecting the quality of results. For a stabilized variant we are able to obtain the same error estimates for the convection-diffusion equation as for the usual discontinuous Galerkin method. We assess the stability of the unstabilized case numerically and find that the inf-sup constant is positive, bounded uniformly away from zero, and very similar to that for the usual discontinuous Galerkin method.

Key words. multiscale, discontinuous Galerkin, convection-diffusion

AMS subject classifications. 65N30, 65Y20

DOI. 10.1137/050640382

1. Introduction.

The discontinuous Galerkin method has undergone rapid development in recent years (see, e.g., [10] and [9]). Although it has been shown to possess advantageous properties in a number of circumstances, its practical utility has been limited by the much larger number of degrees-of-freedom it requires compared with continuous Galerkin methods [13]. This problem has persisted since the inception of the method and has only been recently addressed with the development of a multiscale discontinuous Galerkin method [17] that has the computational structure and cost of a conforming method. The new method utilizes local, element-wise problems to develop a transformation between the parameterization of the discontinuous space and a related, smaller, continuous space. The transformation enables a direct construction of the global matrix problem in terms of the degrees-of-freedom of the continuous space. In the multiscale interpretation, the continuous field is viewed as the coarse scales, and the discontinuous field is viewed as the sum of the coarse and fine scales. The discontinuous part of the solution can be determined by elementwise postprocessing of the continuous solution. In [17] it was shown numerically that the new method at least retains the quality of the discontinuous Galerkin method, and in some instances improves upon it, while at the same time it has the potential to significantly reduce computational cost. A more general framework encompassing the ideas is presented in [3].

In this paper we initiate the mathematical analysis of the method developed in [17]. In section 2 we present the boundary-value problem under consideration,

*Received by the editors September 15, 2005; accepted for publication (in revised form) March 2, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sinum/44-4/64038.html

†Istituto di Matematica Applicata e Tecnologie Informatiche del C.N.R., Via Ferrata 1, 27100 Pavia, Italy (annalisa@imati.cnr.it). This author was supported by the J. Tinsley Oden Faculty Fellowship Research Program and by the PRIN-2004 project of the Italian MIUR.

‡The University of Texas at Austin, Institute for Computational Engineering and Sciences, 1 University Station C0200, Austin, TX 78712-0027 (hughes@ices.utexas.edu). This author was supported by Sandia contract A0340.0 with the University of Texas.

§Dipartimento di Matematica "F. Casorati," Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (giancarlo.sangalli@unipv.it). This author was supported by the J. Tinsley Oden Faculty Fellowship Research Program and by the PRIN-2004 project of the Italian MIUR.

namely, convection-diffusion, and give general definitions necessary for subsequent developments. In section 3 we introduce a discontinuous Galerkin (DG) method that employs interior penalty stabilization and allows for symmetric, neutral, and skew-symmetric treatment of element interface terms corresponding to the diffusion operator. We also introduce a stabilized variant (SDG) that accounts for control of the streamline derivative on element interiors. The DG method is shown to be coercive with respect to the norm induced by its bilinear form, referred to as the DG-norm, and, likewise, the SDG method is shown to be coercive with respect to the SDG-norm induced by its bilinear form. However, the DG-norm is weak in that, in the convective limit, it controls only jumps on element interfaces. In [11], convergence of the DG method in the DG-norm was proved by utilizing the L^2-interpolant, circumventing the need for a stronger stability condition. Here we prove that the DG method is inf-sup stable with respect to the SDG-norm, and this enables us to prove its convergence in the SDG-norm by standard means.

In section 4 we present the multiscale generalizations of DG and SDG, referred to as MDG and SMDG, respectively. We define the local, elementwise problems, which amount to the DG method on individual elements with weakly enforced boundary conditions specified by the shared degrees-of-freedom of the continuous representation, and we define the "interscale transfer spaces" which emanate from the solutions of the local problems. The MDG and SMDG methods amount to the DG and SDG methods in interscale transfer spaces. We prove the inf-sup stability of the local problems in term of the SDG-norm, without streamline-derivative stabilization in the local problems. We also establish the approximation properties of the interscale transfer spaces. With these, and the fact that SDG is coercive on the discontinuous space, we are able to prove convergence and establish the same error estimates for SMDG as for SDG (and DG). Thus, the behavior of the SMDG method is completely understood. This is not the case for the easier MDG method. Indeed, the convergence proof for MDG poses an additional obstacle, namely, DG is inf-sup stable with respect to the SDG-norm on the entire discontinuous space but not necessarily inf-sup stable on the interscale transfer subspace. This problem remains open. However, a numerical assessment of the situation is made in section 5, where the inf-sup constant is calculated for a class of boundary-value problems over a range of convection and diffusion parameters, and on structured meshes. For the cases considered, we find that the MDG method is inf-sup stable with respect to the SDG-norm, and the values of the inf-sup constant are very similar to those for the DG method. These results are consistent with the numerical evaluations performed in [17]. We also assess the stability behavior of the methods in terms of the interior penalty parameter and confirm that MDG behaves in a similar fashion to DG. Results for SMDG are analogous to those for MDG and thus are omitted for brevity. Conclusions are drawn in section 6.

2. Preliminaries.

2.1. Problem description. Let Ω be a bounded *polygonal* domain in $\mathbb{R}^{n_{\mathrm{d}}}$. The strong form of the boundary value problem we are interested in is the following:

$$\begin{aligned} -\kappa\Delta\phi + \boldsymbol{a}\cdot\nabla\phi &= f \quad \text{in } \Omega, \\ \phi &= g \quad \text{on } \Gamma, \end{aligned} \tag{2.1}$$

where $\kappa \geq 0$ is the diffusion coefficient, $\boldsymbol{a}$ is the solenoidal velocity vector field defined on $\overline{\Omega}$, and $\Gamma = \partial\Omega$ is the boundary on which Dirichlet conditions are imposed. More general boundary conditions may be considered as well; see [15] and [17]. We

assume that the values of the diffusion coefficient κ and the velocity field $\boldsymbol{a}$ ensure wellposedness of (2.1). Additional assumptions on these coefficients will be set later.

2.2. General definitions. We introduce the following partition of the boundary:

$$\Gamma^- = \{x \in \Gamma : \boldsymbol{a}(x) \cdot \boldsymbol{n}(x) \leq 0\}, \tag{2.2}$$

$$\Gamma^+ = \{x \in \Gamma : \boldsymbol{a}(x) \cdot \boldsymbol{n}(x) > 0\}, \tag{2.3}$$

where $\boldsymbol{n}$ is the outward unit normal with respect to Γ. Γ^- will be referred to as the *inflow* boundary and Γ^+ as the *outflow* boundary.

Let $\{\mathcal{T}_h\}_h$ be a family of partitions of Ω into elements T. Each $\mathcal{T}_h$ is assumed to be *admissible* in the sense of Ciarlet [8], and *shape regular* (i.e., the elements verify a minimum angle condition, uniformly with respect to h). The elements $T \in \mathcal{T}_h$ are either triangles/quadrilaterals in two dimensions or tetrahedra/hexahedra in three dimensions. Let h_T denote the diameter of T and $h = \max_{T\in\mathcal{T}_h} h_T$. We denote by $\mathcal{E}_h$ the set of all edges of $\mathcal{T}_h$ (*including* edges on the boundary Γ) and by $\mathcal{E}_h^o$ the set of internal edges (*excluding* edges on the boundary Γ) and, by abuse of notation, we denote by Γ both the boundary $\partial\Omega$ and the collection of edges lying on it.

We also define a partition of the element boundary ∂T:

$$\Gamma_T^- = \{x \in \partial T : \boldsymbol{a}(x) \cdot \boldsymbol{n}(x) \leq 0\}, \tag{2.4}$$

$$\Gamma_T^+ = \{x \in \partial T : \boldsymbol{a}(x) \cdot \boldsymbol{n}(x) > 0\}. \tag{2.5}$$

Here $\Gamma_T^{\mp}$ represent the element inflow/outflow boundary, respectively, so that $\partial T = \Gamma_T = \Gamma_T^+ \cup \Gamma_T^-$.

In order to derive a discontinuous Galerkin formulation, following [1], *jumps* and *averages* for scalar and vector fields have to be defined on the edges in $\mathcal{E}_h$. Therefore, consider an interior edge $e \in \mathcal{E}_h^o$, and denote by T^+ and T^-, respectively, the downwind and upwind elements that share it, and by $\boldsymbol{n}^+$ and $\boldsymbol{n}^-$ their respective outward-pointing unit normals. Given a scalar field ν, possibly discontinuous across e, we set $\nu^{\pm} = \nu|_{T^{\pm}}$ on e and define

$$\langle \nu \rangle = \frac{1}{2}(\nu^+ + \nu^-) \qquad [\![\nu]\!] = \nu^+\boldsymbol{n}^+ + \nu^-\boldsymbol{n}^-. \tag{2.6}$$

Analogously, for a vector field $\boldsymbol{\tau}$ we set $\boldsymbol{\tau}^{\pm} = \boldsymbol{\tau}|_{T^{\pm}}$ on e and define

$$\langle \boldsymbol{\tau} \rangle = \frac{1}{2}(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-) \qquad [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}^+ \cdot \boldsymbol{n}^+ + \boldsymbol{\tau}^- \cdot \boldsymbol{n}^-. \tag{2.7}$$

The previous definitions are specialized on the edges on Γ as

$$\langle \nu \rangle = \nu, \qquad [\![\nu]\!] = \nu\ \boldsymbol{n}, \qquad \langle \boldsymbol{\tau} \rangle = \boldsymbol{\tau}, \qquad \forall e \in \Gamma. \tag{2.8}$$

We will extensively make use of the following *biased* identity (based on [1, Formula (3.3)]):

$$\begin{aligned} \sum_{T\in\mathcal{T}_h} \int_{\Gamma_T} \boldsymbol{\tau}\cdot\boldsymbol{n}\ \nu = \sum_{e\in\mathcal{E}_h^o} \left(\int_e \nu^{\pm} [\![\boldsymbol{\tau}]\!] + \int_e [\![\nu]\!] \cdot \boldsymbol{\tau}^{\mp} \right) \\ + \sum_{e\in\Gamma} \int_e \nu\ \boldsymbol{\tau}\cdot\boldsymbol{n}. \end{aligned} \tag{2.9}$$

In what follows, C is a constant, possibly different at each occurrence, which is independent of h and of the coefficients κ and $\boldsymbol{a}$. Moreover, $\alpha \lesssim \beta$ means $\alpha \leq C\beta$, while $\alpha \sim \beta$ means $\alpha \lesssim \beta$ and $\beta \lesssim \alpha$.

We suppose that κ and $\boldsymbol{a}$ are constant on each element $T \in \mathcal{T}_h$. We make use of the following notation: $\kappa_T = \kappa|_T$, $\boldsymbol{a}_T = \boldsymbol{a}|_T$, and $a_T = |\boldsymbol{a}_T|$. Finally, we assume that for any pair of elements T^+ and T^- sharing an edge,

$$\kappa_{T^+} \sim \kappa_{T^-}. \tag{2.10}$$

3. The discontinuous Galerkin method.

3.1. Method description. Given a positive index k, the following approximation space is introduced:

$$V_h = \{v \in L^2(\Omega) : v|_T \in \mathcal{P}^k(T), \quad \forall T \in \mathcal{T}_h\}, \tag{3.1}$$

where $\mathcal{P}^k(T)$ is the space of polynomials of degree at most k supported on T.

A possible DG formulation for (2.1) is as follows: find $\phi^{DG} \in V_h$ such that

$$B^{DG}(\phi^{DG}, \mu) = L^{DG}(g, f; \mu) \qquad \forall\, \mu \in V_h, \tag{3.2}$$

where

$$\begin{aligned} B^{DG}(\nu, \mu) = &- \sum_{T \in \mathcal{T}_h} \int_T \boldsymbol{\nabla}\mu \cdot (\boldsymbol{a}\nu - \kappa \boldsymbol{\nabla}\nu) \\ &+ \sum_{e \in \mathcal{E}_h^o} \int_e \left([\![\mu]\!] \cdot (\boldsymbol{a}\nu^- - \kappa^- \nabla \nu^-) + s\kappa^- \nabla \mu^- [\![\nu]\!] \right) \\ &+ \sum_{e \in \Gamma} \int_e s\kappa \nabla \mu \cdot \boldsymbol{n} \nu - \kappa \nabla \nu \cdot \boldsymbol{n} \mu \\ &+ \sum_{e \in \Gamma^+} \int_e \mu\nu \boldsymbol{a} \cdot \boldsymbol{n} + \varepsilon \sum_{e \in \mathcal{E}_h} \int_e \frac{\langle \kappa \rangle}{h_\perp} [\![\mu]\!] \cdot [\![\nu]\!], \end{aligned}$$

and

$$\begin{aligned} L^{DG}(g, f; \mu) = &\int_\Omega \mu f + \sum_{e \in \Gamma} \left(\varepsilon \int_e \frac{\langle \kappa \rangle}{h_\perp} \mu g + \int_e s\kappa \nabla \mu \cdot \boldsymbol{n} g \right) \\ &- \sum_{e \in \Gamma^-} \int_e \boldsymbol{a} \cdot \boldsymbol{n} \mu g; \end{aligned}$$

s is either -1, 0, or 1 (corresponding to symmetric, neutral, and skew-symmetric interior penalty methods), and for each $e \in \mathcal{E}_h^o$, we set $h_\perp = \frac{|T^+| + |T^-|}{2\,|e|}$, while for $e \in \Gamma$ we set $h_\perp = \frac{|T|}{|e|}$.

Remark 3.1. Notice that on each internal edge $e \in \mathcal{E}_h^o$ the normal component of the velocity field $\boldsymbol{a}$ is continuous, owing to the assumption $\operatorname{div}(\boldsymbol{a}) = 0$.

It will be useful to write the bilinear form $B^{DG}(\cdot, \cdot)$ as a sum of two contributions: the "diffusive" part and the "convective" part:

$$B^{DG}(\nu, \mu) = B^{DG}_{\mathfrak{D}}(\nu, \mu) + B^{DG}_{\mathfrak{C}}(\nu, \mu), \tag{3.3}$$

where

$$B^{DG}_{\mathfrak{D}}(\nu,\mu) = \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{\nabla}\mu\cdot\kappa\boldsymbol{\nabla}\nu - \sum_{e\in\mathcal{E}_h^o}\int_e [\![\mu]\!]\cdot\kappa^-\nabla\nu^- + s\kappa^-\nabla\mu^-[\![\nu]\!] + \sum_{e\in\Gamma}\int_e s\kappa\nabla\mu\cdot\boldsymbol{n}\nu - \kappa\nabla\nu\cdot\boldsymbol{n}\mu + \varepsilon\sum_{e\in\mathcal{E}_h}\int_e \frac{\langle\kappa\rangle}{h_\perp}[\![\mu]\!]\cdot[\![\nu]\!], \tag{3.4}$$

$$B^{DG}_{\mathfrak{C}}(\nu,\mu) = \sum_{T\in\mathcal{T}_h} -\int_T \boldsymbol{\nabla}\mu\cdot\boldsymbol{a}\nu + \sum_{e\in\mathcal{E}_h^o}\int_e [\![\mu]\!]\cdot\boldsymbol{a}\nu^- + \sum_{e\in\Gamma^+}\int_e \mu\nu\boldsymbol{a}\cdot\boldsymbol{n}. \tag{3.5}$$

We also define the DG-norm

$$\|\nu\|^2_{DG} = \|\nu\|^2_{\mathfrak{D}} + \|\nu\|^2_{\mathfrak{C}}, \tag{3.6}$$

where

$$\begin{aligned}\|\nu\|^2_{\mathfrak{D}} &= \sum_{T\in\mathcal{T}_h}\left(\kappa_T|\nu|^2_{H^1(T)} + h_T^2\kappa_T|\nu|^2_{H^2(T)}\right) + \varepsilon\sum_{e\in\mathcal{E}_h}\left(h_\perp^{-1}\|\langle\kappa\rangle[\![\nu]\!]\|^2_{L^2(e)}\right),\\ \|\nu\|^2_{\mathfrak{C}} &= \sum_{e\in\mathcal{E}_h}\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|^2_{L^2(e)}.\end{aligned} \tag{3.7}$$

The DG formulation is consistent: let ϕ be the solution of (2.1); then it is easy to verify that

$$B^{DG}(\phi,\mu) = L^{DG}(g,f;\mu) \qquad \forall\,\mu\in V_h.$$

As far as the stability is concerned, we first recall that the form $B^{DG}(\cdot,\cdot)$ is coercive with respect to the DG-norm, as stated in the next proposition.

PROPOSITION 3.2. *For each value of s, there exists positive $\bar\varepsilon$ such that, for all $\varepsilon > \bar\varepsilon$, there exists $\alpha_{DG} > 0$ such that $B^{DG}(\mu,\mu) \geq \alpha_{DG}\|\mu\|^2_{DG}$ for all $\mu\in V_h$. Moreover, α_{DG} is independent of the mesh-size h and the coefficients κ and $\boldsymbol{a}$.*

Proof. The coercivity of the convection term easily follows by integration by parts:

$$B^{DG}_{\mathfrak{C}}(\mu,\mu) \geq \frac{1}{2}\|\mu\|^2_{\mathfrak{C}}.$$

Moreover, analogously to the stability proof provided in [1], there exists $\bar\varepsilon$ such that, under the assumption $\varepsilon > \bar\varepsilon$, the coercivity of the diffusive term holds; that is,

$$B^{DG}_{\mathfrak{D}}(\mu,\mu) \geq \beta_1\|\mu\|^2_{\mathfrak{D}}. \tag{3.8}$$

Actually, when $s = 1$ (skew-symmetric case) the result holds for any $\varepsilon > 0$. ∎

The coercivity as given in Proposition 3.2 is enough to provide an estimate of the form

$$\|\phi - \phi^{DG}\|^2_{DG} \lesssim \sum_{T\in\mathcal{T}_h}\left[\left(a_T h_T^{2k+1} + \kappa_T h_T^{2k}\right)|\phi|^2_{H^{k+1}(T)}\right], \tag{3.9}$$

which can be obtained by reasoning as in [11], for example. On the other hand, if the convection dominates and the exact solution ϕ is smooth, the quantity $\|\phi - \phi^{DG}\|_{DG}$ is basically a measure of the jumps of the discrete solution. In this case the estimate (3.9) gives very little information on the error $\phi - \phi^{DG}$.

In order to improve the control of the error, we can add an SUPG (streamline-upwind Petrov–Galerkin) stabilization [7] to the DG formulation. Then, we set

$$B^{SDG}(\nu,\mu) = B^{DG}(\nu,\mu) + \sum_{T\in\mathcal{T}_h} \tau_T \int_T (\mathcal{L}_T\nu)(\boldsymbol{a}\cdot\nabla\mu), \tag{3.10}$$

$$L^{SDG}(g,f;\mu) = L^{DG}(g,f;\mu) + \sum_{T\in\mathcal{T}_h} \tau_T \int_T f(\boldsymbol{a}\cdot\nabla\mu), \tag{3.11}$$

where $\mathcal{L}_T\nu = -\kappa\Delta\nu + \boldsymbol{a}\cdot\nabla\nu$ on T and τ_T is a stabilization parameter. The combination of SUPG and DG formulations was first proposed in [18] for linear convection problems, and then in [22] for convection-diffusion problems. In particular, the method proposed in [22] is similar to the present one (the difference being that, in [22], only the convective flux is upwinded).

For the purpose of the error analysis, the required asymptotic behavior of τ_T is $\tau_T \sim \frac{h_T}{a_T}$ in the convection-dominated regime (i.e., when $\frac{\kappa_T}{h_T\,a_T} \lesssim 1$), and $\tau_T \sim \frac{h_T^2}{\kappa_T}$ in the diffusion-dominated regime (i.e., when $\frac{h_T a_T}{\kappa_T} \lesssim 1$). We simply set

$$\tau_T = \tau \min\left\{\frac{h_T}{a_T}, \frac{h_T^2}{\kappa_T}\right\}, \tag{3.12}$$

where τ is a positive real number at our disposal.

The SDG (*stabilized discontinuous Galerkin*) formulation reads as follows: find $\phi^{SDG} \in V_h$ such that

$$B^{SDG}(\phi^{SDG},\mu) = L^{SDG}(g,f;\mu) \qquad \forall\,\mu\in V_h. \tag{3.13}$$

For the theoretical analysis of the SDG scheme (3.13), we will need the SDG-norm

$$\|\nu\|_{SDG}^2 = \|\nu\|_{DG}^2 + \sum_{T\in\mathcal{T}_h} \tau_T \|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}^2 \tag{3.14}$$

and the related

$$|\!|\!|\nu|\!|\!|_{SDG}^2 = \|\nu\|_{SDG}^2 + \sum_{e\in\mathcal{E}_h^0} \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu^-\|_{L^2(e)}^2 + \sum_{T\in\mathcal{T}_h} \tau_T^{-1}\|\nu\|_{L^2(T)}^2. \tag{3.15}$$

It is immediate that the SDG formulation is consistent. Moreover, the problem (3.13) admits a unique solution under suitable assumptions, as a consequence of the following known result.

PROPOSITION 3.3. *For each value of s, there exist positive $\overline{\tau}$ and $\overline{\varepsilon}$ such that, for all $\tau < \overline{\tau}$ and $\varepsilon > \overline{\varepsilon}$, there exists $\alpha_{SDG} > 0$ such that*

$$B^{SDG}(\mu,\mu) \geq \alpha_{SDG}\|\mu\|_{SDG}^2 \qquad \forall\,\mu\in V_h, \tag{3.16}$$

where α_{SDG} is independent of the mesh-size h and the coefficients κ and $\boldsymbol{a}$. Moreover,

$$B^{SDG}(\nu,\mu) \lesssim |\!|\!|\nu|\!|\!|_{SDG}\|\mu\|_{SDG} \qquad \forall\,\nu\in V_h + H^1(\Omega)\,, \qquad \mu\in V_h. \tag{3.17}$$

Proof. We first note that, due to (3.12),

$$\sum_{T\in Th} \tau_T\|\kappa\Delta\mu\|_{L^2(T)}^2 \lesssim \sum_{T\in Th} \tau\kappa_T h_T^2 |\mu|_{H^2(T)}^2 \lesssim \|\mu\|_{\mathfrak{D}}^2. \tag{3.18}$$

Thanks to Proposition 3.2, when ε is greater than a suitable $\bar{\varepsilon}$ we have

$$B^{SDG}(\mu,\mu) \geq \alpha_{DG}\|\mu\|^2_{SDG} - \sum_{T\in\mathcal{T}_h} \tau_T \int_T (\kappa\Delta\mu)(\boldsymbol{a}\cdot\nabla\mu).$$

By the Cauchy–Schwarz inequality and (3.18), (3.16) is proved by choosing τ sufficiently small.

In order to prove (3.17), we proceed in a standard way as follows:

$$B^{SDG}(\nu,\mu) = B^{DG}_{\mathfrak{D}}(\nu,\mu) + B^{DG}_{\mathfrak{C}}(\nu,\mu) + \sum_{T\in\mathcal{T}_h} \tau_T \int_T (\boldsymbol{a}\cdot\nabla\mu)(\mathcal{L}_T\nu) = I + II + III.$$

We estimate the three terms separately. First, by reasoning similar to that of [1],

$$I = B^{DG}_{\mathfrak{D}}(\nu,\mu) \lesssim \|\nu\|_{\mathfrak{D}}\|\mu\|_{\mathfrak{D}}. \tag{3.19}$$

Second, by the Cauchy–Schwarz inequality,

$$\begin{aligned} II &= \sum_{T\in\mathcal{T}_h} -\int_T \boldsymbol{a}\cdot\nabla\mu\,\nu + \sum_{e\in\mathcal{E}^o_h}\int_e [\![\mu]\!]\cdot\boldsymbol{a}\nu^- + \sum_{e\in\Gamma^+}\int_e \mu\nu\boldsymbol{a}\cdot\boldsymbol{n} \\ &\lesssim \|\mu\|_{SDG}\left(\sum_{e\in\mathcal{E}^0_h\cup\Gamma^+} \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu^-\|^2_{L^2(e)} + \sum_{T\in\mathcal{T}_h}\tau_T^{-1}\|\nu\|^2_{L^2(T)}\right)^{1/2} \\ &\lesssim \|\mu\|_{SDG}|\!|\!|\nu|\!|\!|_{SDG}. \end{aligned} \tag{3.20}$$

Third, again by the Cauchy–Schwarz inequality, and by (3.18),

$$III = \sum_{T\in\mathcal{T}_h} \tau_T \int_T (\boldsymbol{a}\cdot\nabla\mu)(-\kappa\Delta\nu + \boldsymbol{a}\cdot\nabla\nu) \lesssim \|\mu\|_{SDG}\|\nu\|_{SDG}. \tag{3.21}$$ ∎

3.2. Error estimate. We first provide an error estimate for the SDG method (3.13).

PROPOSITION 3.4. *Let ϕ be the solution of* (2.1)*, and assume $\phi \in H^{k+1}(\Omega)$. Let ϕ^{SDG} be given by* (3.13)*. Under the assumption of Proposition* 3.3*, the following error estimate holds:*

$$\|\phi - \phi^{SDG}\|_{SDG} \lesssim \left(\sum_{T\in\mathcal{T}_h} \left(a_T h_T^{2k+1} + \kappa_T h_T^{2k}\right)|\phi|^2_{H^{k+1}(T)}\right)^{1/2}. \tag{3.22}$$

Proof. Let $\phi^I \in V_h$ be the usual nodal interpolant of ϕ. Using coercivity and continuity, and (3.16) and (3.17), together with consistency, we get

$$\begin{aligned} \alpha_{SDG}\|\phi^{SDG} - \phi^I\|^2_{SDG} &\leq B^{SDG}(\phi^{SDG} - \phi^I, \phi^{SDG} - \phi^I) \\ &= B^{SDG}(\phi - \phi^I, \phi^{SDG} - \phi^I) \\ &\lesssim |\!|\!|\phi - \phi^I|\!|\!|_{SDG}\|\phi^{SDG} - \phi^I\|_{SDG}. \end{aligned} \tag{3.23}$$

For the usual local estimates on the interpolation error $\phi - \phi^I$ we readily obtain

$$|\!|\!|\phi - \phi^I|\!|\!|_{SDG} \lesssim \left(\sum_{T\in\mathcal{T}_h} \left(\kappa_T h_T^{2k} + \tau_T a_T^2 h_T^{2k} + \tau_T^{-1} h_T^{2k+2} + \tau_T \kappa_T^2 h_T^{2k-2}\right)|\phi|^2_{H^{k+1}(T)}\right)^{1/2}.$$

When choosing the stabilization parameter τ_T according to (3.12), by direct comparison, we see that (3.22) follows. ∎

For the pure discontinuous Galerkin method (3.2), a suitable control on the streamline derivative can be obtained, as was first studied in [19] for the pure convection (scalar hyperbolic) equation. In the following result, we prove an inf-sup condition for the bilinear form $B^{DG}(\cdot,\cdot)$ with respect to the SDG-norm. This improves the stability result stated in Proposition 3.2.

THEOREM 3.5. *There exists $\bar{\varepsilon}$ such that for all $\varepsilon \geq \bar{\varepsilon}$,*

$$\inf_{\nu\in V_h} \sup_{\mu\in V_h} \frac{B^{DG}(\nu,\mu)}{\|\nu\|_{SDG}\|\mu\|_{SDG}} \geq \beta_{DG} > 0, \tag{3.24}$$

where β_{DG} is independent of h, κ, $\boldsymbol{a}$, and the domain.

Proof. Given $\nu \in V_h$, we choose $\mu = \nu + \gamma \sum_{T\in\mathcal{T}_h} \tau_T(\boldsymbol{a}\cdot\nabla\nu)_{|T} = \nu + \gamma\mu_2$, where γ is a positive parameter at our disposal. Note that $\mu \in V_h$, as the velocity field is piecewise constant on $\mathcal{T}_h$. We prove the following:

$$\|\mu\|_{SDG} \lesssim \|\nu\|_{SDG}, \tag{3.25}$$

$$B(\nu,\mu) \geq \beta\|\nu\|^2_{SDG}. \tag{3.26}$$

We start by proving (3.25). To this end, we need to estimate the different terms of $\|\mu_2\|_{SDG}$. Recall that, from (3.12),

$$\tau_T \leq \tau\frac{h_T^2}{\kappa_T} \tag{3.27}$$

and

$$\tau_T \leq \tau\frac{h_T}{a_T}. \tag{3.28}$$

Using (3.28) and a local inverse inequality, we have

$$\begin{aligned}\tau_T\|\boldsymbol{a}\cdot\nabla(\tau_T\,\boldsymbol{a}\cdot\nabla\nu)\|^2_{L^2(T)} &\leq \tau_T^3 a_T^2\|\nabla(\boldsymbol{a}\cdot\nabla\nu)\|^2_{L^2(T)}\\ &\leq (\tau C_{inv})^2\tau_T\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)},\end{aligned} \tag{3.29}$$

where C_{inv} is the constant of the local inverse inequality. From (3.29) we get

$$\sum_{T\in\mathcal{T}_h} \tau_T\|\boldsymbol{a}\cdot\nabla\mu_2\|^2_{L^2(T)} \lesssim \|\nu\|^2_{SDG}. \tag{3.30}$$

Consider an internal edge $e \in \mathcal{E}_h^o$, and denote by T^- and T^+ the adjacent upwind and downwind elements. We have

$$\begin{aligned}\|[\![\mu_2]\!]\|^2_{L^2(e)} &\lesssim \|\mu_{2|T^-}\|^2_{L^2(e)} + \|\mu_{2|T^+}\|^2_{L^2(e)}\\ &\leq \tau_{T^-}^2\|(\boldsymbol{a}\cdot\nabla\nu)_{|T^-}\|^2_{L^2(e)} + \tau_{T^+}^2\|(\boldsymbol{a}\cdot\nabla\nu)_{|T^+}\|^2_{L^2(e)}.\end{aligned} \tag{3.31}$$

Using the trace inequality,

$$\|\xi\|^2_{L^2(e)} \leq C_{tr}(h_T^{-1}\|\xi\|^2_{L^2(T)} + \|\xi\|_{L^2(T)}\|\nabla\xi\|_{L^2(T)}),$$

which holds for all $\xi \in H^1(T)$, and a local inverse inequality, we also have

$$\|(\boldsymbol{a}\cdot\nabla\nu)_{|T^\pm}\|^2_{L^2(e)} \leq C'_{inv}h_{T^\pm}^{-1}\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T^\pm)}, \tag{3.32}$$

where $C'_{inv} = C_{tr}(1 + C_{inv})$. From (3.31) and (3.32) together with (3.28), we obtain

$$\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\mu_2]\!]\|_{L^2(e)} \lesssim \tau_{T^+}^{1/2}\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T^+)} + \tau_{T^-}^{1/2}\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T^-)}.$$

Similarly, for a boundary edge $e \in \Gamma$, if $e \subset \Gamma_T$, then

$$\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\mu_2]\!]\|_{L^2(e)} \lesssim \tau_T^{1/2}\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}.$$

Summarizing, we have proved

$$\|\mu_2\|_{\mathfrak{C}}^2 \lesssim \sum_{T\in\mathcal{T}_h} \tau_T\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}^2. \tag{3.33}$$

By the inverse inequality, as in (3.29), we have

$$\kappa_T\|\nabla\mu_2\|_{L^2(T)}^2 \le \kappa_T a_T^2\tau_T^2|\nu|_{H^2(T)}^2 \lesssim C_{inv}^2\kappa_T\|\nabla\nu\|_{L^2(T)}^2. \tag{3.34}$$

On the other hand, recalling (2.10), (3.31)–(3.32) implies that, for each $e \in \mathcal{E}_h^o$,

$$\frac{\langle\kappa\rangle}{h_\perp}\|[\![\mu_2]\!]\|_{L^2(e)}^2 \lesssim \kappa_{T^+}\|\nabla\nu\|_{L^2(T^+)}^2 + \kappa_{T^-}\|\nabla\nu\|_{L^2(T^-)}^2, \tag{3.35}$$

or, for $e \in \Gamma$,

$$\frac{\langle\kappa\rangle}{h_\perp}\|[\![\mu_2]\!]\|_{L^2(e)}^2 \lesssim \kappa_T\|\nabla\nu\|_{L^2(T)}^2. \tag{3.36}$$

This proves that

$$\|\mu_2\|_{\mathfrak{D}} \le C_{\mathfrak{D}}\|\nu\|_{\mathfrak{D}} \tag{3.37}$$

where $C_{\mathfrak{D}}$ is a constant independent of the mesh-size and the problem parameters. The bounds (3.30), (3.33), and (3.37) give $\|\mu_2\|_{SDG} \lesssim \|\nu\|_{SDG}$ and finally (3.25).

We turn now to the proof of (3.26). First of all, we have

$$B_{\mathfrak{C}}^{DG}(\nu,\nu) = \frac{1}{2}\sum_{e\in\mathcal{E}_h}\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|_{L^2(e)}^2.$$

On the other hand, by integration by parts and (2.9),

$$B_{\mathfrak{C}}^{DG}(\nu,\mu_2) = \sum_{T\in\mathcal{T}_h}\left(\tau_T\int_T|\boldsymbol{a}\cdot\nabla\nu|^2 - \int_{\Gamma_T^-}\tau_T\,(\boldsymbol{a}\cdot\nabla\nu)^+\,[\![\boldsymbol{a}\nu]\!]\right), \tag{3.38}$$

and, using (3.32),

$$\begin{aligned}
\sum_{T\in\mathcal{T}_h}\int_{\Gamma_T^-}\tau_T\,(\boldsymbol{a}\cdot\nabla\nu)^+\,[\![\boldsymbol{a}\nu]\!] &\le \sum_{T\in\mathcal{T}_h}\tau_T\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}(\boldsymbol{a}\cdot\nabla\nu)^+\|_{L^2(\Gamma_T^-)} \\
&\qquad\cdot\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|_{L^2(\Gamma_T^-)} \\
&\le \frac{1}{2\lambda}\sum_{T\in\mathcal{T}_h}\tau_T^2\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}(\boldsymbol{a}\cdot\nabla\nu)^+\|_{L^2(\Gamma_T^-)}^2 \\
&\quad + \frac{\lambda}{2}\sum_{T\in\mathcal{T}_h}\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|_{L^2(\Gamma_T^-)}^2 \\
&\le \frac{\tau C'_{inv}}{2\lambda}\sum_{T\in\mathcal{T}_h}\tau_T\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}^2 \\
&\quad + \frac{\lambda}{2}\sum_{e\in\mathcal{E}_h}\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|_{L^2(e)}^2
\end{aligned}$$

for any $\lambda > 0$. Using these estimates, we have

$$
\begin{aligned}
B^{DG}_{\mathfrak{C}}(\nu,\mu) \geq & \left(1-\frac{\gamma\,\lambda}{2}\right)\sum_{e\in\mathcal{E}_h} \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|^2_{L^2(e)} \\
& + \gamma\left(1-\frac{\tau C'_{inv}}{2\lambda}\right)\sum_{T\in\mathcal{T}_h}\tau_T\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)}.
\end{aligned}
\tag{3.39}
$$

For the estimation of the diffusion part $B^{DG}_{\mathfrak{D}}(\nu,\mu)$, we use coercivity (3.8), continuity (e.g., see (3.19)) of $B^{DG}_{\mathfrak{D}}(\cdot,\cdot)$, and the estimate (3.37) to obtain

$$
\begin{aligned}
B^{DG}_{\mathfrak{D}}(\nu,\mu) &= B^{DG}_{\mathfrak{D}}(\nu,\nu) + \gamma B^{DG}_{\mathfrak{D}}(\nu,\mu_2) \\
&\geq \beta_1\|\nu\|^2_{\mathfrak{D}} - \gamma\widetilde{\beta}_2\|\mu_2\|_{\mathfrak{D}}\|\nu\|_{\mathfrak{D}} \\
&\geq (\beta_1 - \gamma C_{\mathfrak{D}}\widetilde{\beta}_2)\|\nu\|^2_{\mathfrak{D}}.
\end{aligned}
\tag{3.40}
$$

Summing equations (3.39) and (3.40), and setting $\beta_2 = C_{\mathfrak{D}}\widetilde{\beta}_2$, we obtain

$$
\begin{aligned}
B^{DG}(\nu,\mu) \geq & (\beta_1-\gamma\beta_2)\|\nu\|^2_{\mathfrak{D}} \\
& + \left(1-\frac{\gamma\,\lambda}{2}\right)\sum_{e\in\mathcal{E}_h} \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}[\![\nu]\!]\|^2_{L^2(e)} \\
& + \gamma\left(1-\frac{\tau C'_{inv}}{2\lambda}\right)\sum_{T\in\mathcal{T}_h}\tau_T\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)}.
\end{aligned}
$$

The theorem is then proved by choosing $\lambda = \tau C'_{inv}$ and $\gamma = \min\left\{\lambda^{-1}, \frac{\beta_1}{2\beta_2}\right\}$. ∎

Remark 3.6. Results analogous to those of Theorem 3.5 can be obtained for other DG formulations, such as the interior penalty method.

From Theorem 3.5 we deduce the following error estimate for the DG scheme.

COROLLARY 3.7. *Let ϕ be the solution of* (2.1), *and assume $\phi \in H^{k+1}(\Omega)$; let ϕ^{DG} be the solution of* (3.2). *We have*

$$
\|\phi-\phi^{DG}\|_{SDG} \lesssim \left(\sum_{T\in\mathcal{T}_h}\left(a_T h_T^{2k+1} + \kappa_T h_T^{2k}\right)|\phi|^2_{H^{k+1}(T)}\right)^{1/2}.
\tag{3.41}
$$

Proof. Let $\phi^I \in V_h$ be the nodal interpolant of ϕ and let $\zeta = \phi^{DG} - \phi^I$. Let $\mu \in V_h$ be a test function satisfying (3.25)–(3.26). Using consistency and Proposition 3.3,

$$
\begin{aligned}
\beta\|\zeta\|^2_{SDG} &\leq B^{DG}(\zeta,\mu) = B^{DG}(\phi-\phi^I,\mu) \\
&\lesssim |\!|\!|\phi-\phi^I|\!|\!|_{SDG}\|\mu\|_{SDG} \\
&\lesssim \left(\sum_{T\in\mathcal{T}_h}\left(a_T h_T^{2k+1} + \kappa_T h_T^{2k}\right)|\phi|^2_{H^{k+1}(T)}\right)^{1/2}\|\zeta\|_{SDG}.
\end{aligned}
$$

We deduce (3.41) by the triangle inequality. ∎

4. The multiscale discontinuous Galerkin method. In this section, we present a reduction technique, referred to as the MDG method, which was first introduced in [17]. Furthermore, a stabilized variant of this method, referred to as SMDG, will be introduced subsequently.

The main idea is the following: (i) Solve (3.2) or (3.13) on a suitable subspace of V_h preserving the stability and approximation properties; (ii) Use a multiscale paradigm and local problems to perform the elimination of degrees-of-freedom for both the test and trial spaces.

4.1. Method description. We introduce the spaces $\overline{V}_h = V_h \cap H^1(\Omega)$ and, for all $T \in \mathcal{T}_h$, $V_h(T) = V_h|_T$ (note that this is nothing other than the space of degree k polynomials on T). The *local problems* read as follows: for all $\bar{\nu} \in \overline{V}_h$, find $\nu \in V_h$ such that for all $T \in \mathcal{T}_h$,

$$b_T(\nu, \mu) = \ell_T(\bar{\nu}, f; \mu) \quad \forall\, \mu \in V_h(T), \tag{4.1}$$

where we have set

$$\begin{aligned} b_T(\nu,\mu) &= \int_T \kappa \nabla\nu\cdot\nabla\mu - \int_{\Gamma_T} (\kappa\nabla\nu\cdot\boldsymbol{n}\mu - s\kappa\boldsymbol{\nabla}\mu\cdot\boldsymbol{n}\nu) + \varepsilon\int_{\Gamma_T}\frac{\kappa}{h_\perp}\mu\nu \\ &\quad - \int_T \boldsymbol{\nabla}\mu\cdot\boldsymbol{a}\nu + \int_{\Gamma_T^+}(1+\delta)\,\mu\nu\boldsymbol{a}\cdot\boldsymbol{n}, \\ \ell_T(\bar{\nu}, f;\mu) &= -\int_{\Gamma_T^-}\mu\bar{\nu}\boldsymbol{a}\cdot\boldsymbol{n} + \delta\int_{\Gamma_T^+}\mu\bar{\nu}\boldsymbol{a}\cdot\boldsymbol{n} + \varepsilon\int_{\Gamma_T}\frac{\kappa}{h_\perp}\mu\bar{\nu} \\ &\quad + \int_{\Gamma_T} s\kappa\boldsymbol{\nabla}\mu\cdot\boldsymbol{n}\bar{\nu} + \int_T f\,\mu. \end{aligned} \tag{4.2}$$

Observe that (4.1) is a DG formulation for the local problem $\mathcal{L}_T\nu = f$ on T, with $\nu = \bar{\nu}$ on the boundary Γ_T. Comparing the local DG formulation (4.1) with the global DG formulation (3.2), notice that the former has an extra term, which depends on a new parameter $\delta > 0$. This new term is needed for implementation purposes (see [17]).

We denote by $\mathfrak{T}_h : \overline{V}_h \times L^2(\Omega) \to V_h$ the operator which associates to each $(\bar{\nu}, f) \in \overline{V}_h \times L^2(\Omega)$ the solution ν of the local problems (4.1) on each element $T \in \mathcal{T}_h$. The stability of (4.1), which is stated below (in Lemma 4.4), implies that the problems (4.1) admit unique solutions on each element $T \in \mathcal{T}_h$; that is, the operator $\mathfrak{T}_h$ is well defined. $\mathfrak{T}_h$ represents the "interscale transfer operator," and the associated "interscale transfer spaces" are the (affine) manifold

$$\mathfrak{T}_h(\overline{V}_h, f) = \left\{\mathfrak{T}_h(\bar{\nu}, f) \mid \bar{\nu} \in \overline{V}_h\right\}$$

and the (linear) manifold

$$\mathfrak{T}_h(\overline{V}_h, 0) = \left\{\mathfrak{T}_h(\bar{\nu}, 0) \mid \bar{\nu} \in \overline{V}_h\right\}.$$

With this notation, the MDG method reads as follows: find $\phi^{MDG} \in \mathfrak{T}_h(\overline{V}_h, f)$ such that

$$B^{DG}(\phi^{MDG}, \mu) = L^{DG}(g, f; \mu) \quad \forall \mu \in \mathfrak{T}_h(\overline{V}_h, 0). \tag{4.3}$$

Its stabilized version SMDG reads as follows: find $\phi^{SMDG} \in \mathfrak{T}_h(\overline{V}_h, f)$ such that

$$B^{SDG}(\phi^{SMDG}, \mu) = L^{SDG}(g, f; \mu) \quad \forall \mu \in \mathfrak{T}_h(\overline{V}_h, 0). \tag{4.4}$$

Notice that SMDG is an SUPG stabilization of MDG.

Remark 4.1. The spaces $\mathfrak{T}_h(\overline{V}_h, f)$ and $\mathfrak{T}_h(\overline{V}_h, 0)$ can be parameterized by means of the degrees-of-freedom of $\overline{V}_h$ lying on the "skeleton" $\Sigma = \cup_{e\in\mathcal{E}_h} e$.

Remark 4.2. The *MDG* method can be interpreted as a multiscale technique (see [4, 12, 14, 16]). Both trial and test discontinuous functions $\nu \in V_h$ can be split into a continuous *coarse* scale $\bar{\nu}$ plus a discontinuous *fine* scale $\nu' = \nu - \bar{\nu}$. Performing integration by parts in (4.1), we find that ν' satisfies

$$b_T(\nu', \mu) = \int_T (f - \mathcal{L}_T\bar{\nu})\mu \quad \forall \mu \in V_h(T). \tag{4.5}$$

Equation (4.5) suggests a relationship between the MDG approach and the *RFB* (residual-free bubble) approach (see, e.g., [5, 6, 21]). Consider, for the sake of simplicity, the case of lowest order approximation $k = 1$. Actually ν' in (4.5) can be understood as the DG approximation of the *exact* residual-free bubble ν^{bubble}, which satisfies $\mathcal{L}_T\nu^{bubble} = f - \mathcal{L}_T\bar{\nu}$ on T, with $\nu^{bubble} = 0$ on the boundary ∂T. A DG approximation of the *exact* residual-free bubble has also been used in the *DB* (*discontinuous bubble*) implementation of the RFB formulation (see [20]). The major difference between MDG and DB is that for the latter the space of test functions was $\overline{V}_h$ instead of $\mathfrak{T}_h(\overline{V}_h, 0)$. The relation between these two approaches deserves further investigation.

4.2. Approximation properties of $\mathfrak{T}_h(\overline{V}_h, f)$. The first step in the analysis of problems (4.3) and (4.4) is the study of the approximation properties of the interscale transfer affine space $\mathfrak{T}_h(\overline{V}_h, f)$.

THEOREM 4.3 (approximation). *Let ϕ be the solution of* (2.1); *then there exists $\nu \in \mathfrak{T}_h(\overline{V}_h, f)$ such that*

$$\|\!|\phi - \nu|\!\|_{SDG} \lesssim \left(\sum_{T\in\mathcal{T}_h} \left(a_T h_T^{2k+1} + \kappa_T h_T^{2k} \right) |\phi|^2_{H^{k+1}(T)} \right)^{1/2}. \tag{4.6}$$

Before proving Theorem 4.3, we need some lemmas. On each element $T \in \mathcal{T}_h$, we introduce the following local norm:

$$\begin{aligned} \|\nu\|^2_{SDG(T)} := {} & \kappa_T|\nu|^2_{H^1(T)} + h_T^2\kappa_T|\nu|^2_{H^2(T)} + \tau_T\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)} \\ & + \varepsilon h_T^{-1}\kappa_T\|\nu\|^2_{L^2(\Gamma_T)} + \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu\|^2_{L^2(\Gamma_T)}. \end{aligned} \tag{4.7}$$

The next lemma states that the local problems (4.1) are stable.

LEMMA 4.4 (local stability). *There exist positive $\overline{\varepsilon}$ and $\overline{\delta}$ such that for all $\varepsilon \geq \overline{\varepsilon}$ and $\delta \leq \overline{\delta}$,*

$$\inf_{\nu\in V_h(T)} \sup_{\mu\in V_h(T)} \frac{b_T(\nu,\mu)}{\|\nu\|_{SDG(T)}\|\mu\|_{SDG(T)}} \geq \beta_b > 0 \qquad \forall T \in \mathcal{T}_h, \tag{4.8}$$

and the constant β_b is independent of T, κ, and $\boldsymbol{a}$.

Proof. If $\delta = 0$, then (4.8) is a particular case of (3.24) where the domain is T (endowed with a one-element mesh) instead of Ω. Then, for $\delta = 0$, given $\nu \in V_h(T)$ there exists $\mu \in V_h(T)$ such that

$$\begin{aligned} \|\mu\|_{SDG(T)} &\leq \|\nu\|_{SDG(T)}, \\ b_T(\nu,\mu) &\geq \beta_{DG}\|\nu\|^2_{SDG(T)}. \end{aligned} \tag{4.9}$$

If $\delta \neq 0$, given $\nu \in V_h(T)$ and for the same $\mu \in V_h(T)$ as in (4.9), we have

$$\begin{aligned} &\|\mu\|_{SDG(T)} \leq \|\nu\|_{SDG(T)}, \\ &b_T(\nu,\mu) \geq \beta_{DG}\|\nu\|^2_{SDG(T)} + \int_{\Gamma_T^+} \delta\mu\nu\boldsymbol{a}\cdot\boldsymbol{n}. \end{aligned} \tag{4.10}$$

Moreover,

$$\left|\int_{\Gamma_T^+} \delta\mu\nu\boldsymbol{a}\cdot\boldsymbol{n}\right| \leq \delta\|\mu\|_{SDG(T)}\|\nu\|_{SDG(T)}.$$

Then, for $\delta \leq \overline{\delta} = \beta_{DG}/2$, from (4.10) we get

$$\begin{aligned} &\|\mu\|_{SDG(T)} \leq \|\nu\|_{SDG(T)}, \\ &b_T(\nu,\mu) \geq \frac{\beta_{DG}}{2}\|\nu\|^2_{SDG(T)}, \end{aligned} \tag{4.11}$$

which gives (4.8) for $\beta_b = \beta_{DG}/2$. ∎

The local problems are consistent: let ϕ be the solution of (2.1); then

$$b_T(\phi,\mu) = \ell_T(\phi,f,\mu) \qquad \forall \mu \in V_h, \qquad \forall T \in \mathcal{T}_h. \tag{4.12}$$

In the following lemma we state a Poincaré-like estimate for the norm $\|\cdot\|_{SDG(T)}$.

LEMMA 4.5. *For each element $T \in \mathcal{T}_h$ and each function $\nu \in H^1(T)$, the following estimate holds:*

$$\tau_T^{-1}\|\nu\|^2_{L^2(T)} \lesssim \|\nu\|^2_{SDG(T)}. \tag{4.13}$$

Proof. Fix an element $T \in \mathcal{T}_h$. Because of the definition (3.12) of τ_T, (4.13) is a consequence of the two Poincaré estimates

$$\frac{a_T}{h_T}\|\nu\|^2_{L^2(T)} \lesssim \frac{h_T}{a_T}\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)} + \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu\|^2_{L^2(\Gamma_T)}, \tag{4.14}$$

$$\frac{\kappa_T}{h_T^2}\|\nu\|^2_{L^2(T)} \lesssim \kappa_T|\nu|^2_{H^1(T)} + h_T^{-1}\|\kappa^{1/2}\nu\|^2_{L^2(\Gamma_T)}. \tag{4.15}$$

The inequality (4.15) is a consequence of the standard Poincaré inequality plus a scaling argument. Therefore, we concentrate on the less common (4.14). Let η be the solution of the problem

$$\boldsymbol{a}\cdot\nabla\eta = 1 \quad \text{on } T \quad \text{and} \quad \eta|_{\Gamma_T^-} = 0.$$

It is easy to verify that $\|\eta\|_{L^\infty(T)} \leq \frac{h_T}{a_T}$. Given $v \in H^1(T)$, we estimate $\|\cdot\|_{L^2(T)}$ as follows:

$$\begin{aligned} \|\nu\|^2_{L^2(T)} &= \int_T \nu^2\,\boldsymbol{a}\cdot\nabla\eta = -\int_T \boldsymbol{a}\cdot\nabla(\nu^2)\,\eta + \int_{\Gamma_T^+} \boldsymbol{a}\cdot\boldsymbol{n}\,\eta\,\nu^2 \\ &\leq \|\eta\|_{L^\infty(T)}(2\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}\|\nu\|_{L^2(T)} + \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu\|^2_{L^2(\Gamma_T)}) \\ &\leq \frac{h_T}{a_T}(2\|\boldsymbol{a}\cdot\nabla\nu\|_{L^2(T)}\|\nu\|_{L^2(T)} + \||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu\|^2_{L^2(\Gamma_T)}) \\ &\leq \frac{2h_T^2}{a_T^2}\|\boldsymbol{a}\cdot\nabla\nu\|^2_{L^2(T)} + \frac{1}{2}\|\nu\|^2_{L^2(T)} + \frac{h_T}{a_T}\||\boldsymbol{a}\cdot\boldsymbol{n}|^{1/2}\nu\|^2_{L^2(\Gamma_T)}. \end{aligned}$$

The inequality (4.14) follows, dividing both sides by $\frac{h_T}{a_T}$. ∎

Proof of Theorem 4.3. Let $\phi^I \in V_h$ be the nodal interpolant of ϕ and let ν be the solution of the following local problems:

$$b_T(\nu_{|T}, \mu) = \ell_T(\phi^I, f, \mu) \qquad \forall \mu \in V_h, \qquad \forall T \in \mathcal{T}_h.$$

We have $\nu \in \mathfrak{T}_h(\overline{V}_h, f)$ and will show that ν verifies the estimate (4.6). First, we prove that

$$|\!|\!| \phi - \nu |\!|\!|^2_{SDG} \lesssim \sum_{T \in \mathcal{T}_h} \|\phi - \nu\|^2_{SDG(T)}. \tag{4.16}$$

It is immediate that

$$\|\phi - \nu\|^2_{SDG} + \sum_{e \in \mathcal{E}_h^0 \cup \Gamma^+} \||\boldsymbol{a} \cdot \boldsymbol{n}|^{1/2} (\phi - \nu)^-\|^2_{L^2(e)} \lesssim \sum_{T \in \mathcal{T}_h} \|\phi - \nu\|^2_{SDG(T)}, \tag{4.17}$$

and, making use of (4.13), we also have

$$\sum_{T \in \mathcal{T}_h} \tau_T^{-1} \|\nu\|^2_{L^2(T)} \lesssim \sum_{T \in \mathcal{T}_h} \|\phi - \nu\|^2_{SDG(T)}. \tag{4.18}$$

Therefore, from (4.16) and the usual triangle inequality, we get

$$|\!|\!| \phi - \nu |\!|\!|^2_{SDG} \lesssim \sum_{T \in \mathcal{T}_h} \|\phi - \phi^I\|^2_{SDG(T)} + \sum_{T \in \mathcal{T}_h} \|\phi^I - \nu\|^2_{SDG(T)} = I + II.$$

Let us concentrate on II first. Fix a generic $T \in \mathcal{T}_h$; then consistency (4.12) implies

$$b_T(\phi - \nu, \mu) = \ell_T(\phi - \phi^I, 0; \mu) \qquad \forall \mu \in V_h(T). \tag{4.19}$$

By Lemma 4.4, there exists $\widetilde{\mu} \in V_h(T)$ such that $\|\widetilde{\mu}\|_{SDG(T)} \lesssim \|\phi^I - \nu\|_{SDG(T)}$ and

$$\begin{aligned} \|\phi^I - \nu\|^2_{SDG(T)} &\lesssim b_T(\phi^I - \nu, \widetilde{\mu}) \\ &= b_T(\phi^I - \phi, \widetilde{\mu}) + b_T(\phi - \nu, \widetilde{\mu}) \\ &= b_T(\phi^I - \phi, \widetilde{\mu}) + \ell_T(\phi - \phi^I, 0; \widetilde{\mu}). \end{aligned} \tag{4.20}$$

We have

$$\begin{aligned} b_T(\phi^I - \phi, \widetilde{\mu}) &\lesssim (\|\phi^I - \phi\|_{SDG(T)} + \tau_T^{-1} \|\phi^I - \phi\|_{L^2(T)}) \|\widetilde{\mu}\|_{SDG(T)}, \\ \ell_T(\phi - \phi^I, \widetilde{\mu}) &\lesssim \|\phi^I - \phi\|_{SDG(T)} \|\widetilde{\mu}\|_{SDG(T)}. \end{aligned} \tag{4.21}$$

Thanks to (4.20)–(4.21) and the Poincaré estimate (4.13), we obtain

$$\begin{aligned} \|\phi^I - \nu\|_{SDG(T)} &\lesssim \|\phi^I - \phi\|_{SDG(T)} + \tau_T^{-1} \|\phi^I - \phi\|_{L^2(T)} \\ &\lesssim \|\phi^I - \phi\|_{SDG(T)}. \end{aligned}$$

Squaring and summing over all the elements, we end up with

$$II \lesssim I.$$

Finally, observe that, by using the standard estimates for the interpolation error, we easily get

$$I \lesssim \left(\sum_{T \in \mathcal{T}_h} \left(a_T h_T^{2k+1} + \kappa_T h_T^{2k} \right) |\phi|^2_{H^{k+1}(T)} \right)^{1/2}.$$

This gives (4.6). ∎

4.3. Error estimate. An optimal error estimate for the SMDG method readily follows from Theorem 4.3 and Proposition 3.3.

THEOREM 4.6. *Let ϕ and ϕ^{SMDG} be the solutions of* (2.1) *and* (4.4)*, respectively. Under the same assumption as in Proposition* 3.3,

$$\|\phi - \phi^{SMDG}\|_{SDG} \lesssim \left(\sum_{T\in\mathcal{T}_h} \left(a_T h_T^{2k+1} + \kappa_T h_T^{2k}\right) |\phi|^2_{H^{k+1}(T)} \right)^{1/2}. \tag{4.22}$$

Proof. Let $\nu \in \mathfrak{T}_h(\overline{V}_h, f)$ be the approximant of ϕ given by Theorem 4.3, and let $\zeta = \phi^{SMDG} - \nu$. Linearity ensures that $\zeta \in \mathfrak{T}_h(\overline{V}_h, 0)$; that is, it is an admissible test function for (4.4). Repeating the same steps as in Proposition 3.4, we obtain the estimate

$$\|\phi - \phi^{SMDG}\|_{SDG} \lesssim |\!|\!| \phi - \nu |\!|\!|_{SDG}.$$

The statement is then proved by using Theorem 4.3. ∎

Remark 4.7. The problem of providing an optimal error estimate for MDG remains open. The error estimate (3.9) for DG, proved in [11], makes use of an interpolant which is the L^2-projection of ϕ onto V_h, which is not generally available in $\mathfrak{T}_h(\overline{V}_h, f)$. On the other hand, the stronger error estimate (3.41) we have proved in section 3, still for DG, relies on the validity of the inf-sup condition (3.24). A similar error analysis for the MDG method would need the following inf-sup condition:

$$\inf_{\nu\in\mathfrak{T}_h(\overline{V}_h,0)} \sup_{\mu\in\mathfrak{T}_h(\overline{V}_h,0)} \frac{B^{DG}(\nu,\mu)}{\|\nu\|_{SDG}\|\mu\|_{SDG}} \geq \beta_{MDG} > 0, \tag{4.23}$$

where β_{MDG} has to be independent of h. The inf-sup condition (4.23) is *not* a consequence of (3.24). One of the objectives of the next section is the numerical evaluation of the inf-sup constant β_{MDG} in (4.23).

5. Selection of parameters. The stability of the numerical schemes we have considered depends on the parameters ε (which specify the amount of *interior penalty* stabilization in all the formulations) and τ (which specifies the amount of streamline stabilization in SDG and SMDG). In this section, we want to investigate in more detail the relation between the stability of the schemes and the value of the parameters for a specific model problem. Moreover, we investigate numerically the validity of (4.23) and we demonstrate that (4.23) holds, at least for the cases covered by our numerical experiments.

We consider a square domain $\Omega = [0,1]^2$ and a uniform partition $\mathcal{T}_h$ of $N \times N$ square elements. Then, we select bilinear finite element spaces, discontinuous for V_h and globally continuous for $\bar{V}_h$. We restrict ourselves to the simplest case of constant coefficients κ and $\boldsymbol{a}$. Numerical testing of this configuration has been performed in [17]. Here, we want to measure the stability of the schemes by a numerical evaluation of the inf-sup constant

$$\inf_{\nu\in V_h} \sup_{\mu\in V_h} \frac{B(\nu,\mu)}{\|\nu\|_{SDG}\|\mu\|_{SDG}} \tag{5.1}$$

for the DG and SDG formulations (where $B(\cdot,\cdot) \equiv B^{DG}(\cdot,\cdot)$ and $B(\cdot,\cdot) \equiv B^{SDG}(\cdot,\cdot)$, resp.) and

$$\inf_{\nu\in\mathfrak{T}_h(\overline{V}_h,0)} \sup_{\mu\in\mathfrak{T}_h(\overline{V}_h,0)} \frac{B(\nu,\mu)}{\|\nu\|_{SDG}\|\mu\|_{SDG}} \tag{5.2}$$

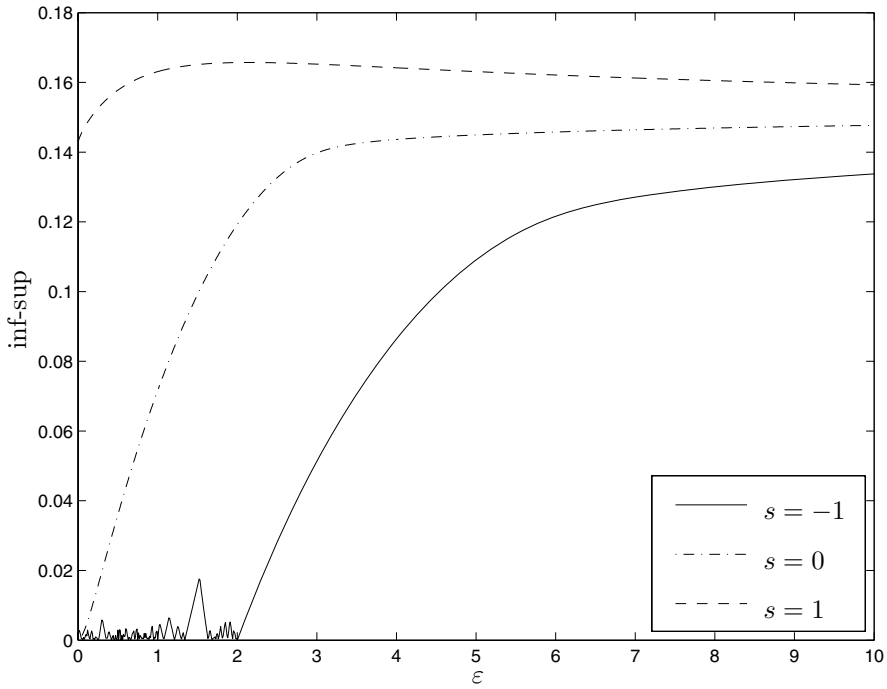


FIG. 5.1. *Inf-sup constant of the DG method versus ε.*

for the MDG and SMDG formulations (where $B(\cdot,\cdot) \equiv B^{DG}(\cdot,\cdot)$ and $B(\cdot,\cdot) \equiv B^{SDG}(\cdot,\cdot)$, resp.). The evaluation of (5.1) and (5.2) can be performed through a generalized eigenvalue computation (see, e.g., [2] for details). In what follows we assume $\delta = 0$. Very similar results are obtained with the choice $\delta = 0.01$, proposed in [17], which has advantages from the implementation standpoint.

5.1. The interior penalty parameter. First, we study the effect of ε, the amount of interior penalty stabilization. We focus on the diffusion-dominated regime, where the interior penalty term plays a role, setting $\kappa = 1$ and $\|\boldsymbol{a}\| = 10^{-10}$. The values of (5.1) and (5.2) are plotted in Figure 5.1 and 5.2, respectively, for the DG and MDG schemes (similar result are obtained for the stabilized SDG and SMDG schemes), and for a partition of 10×10 elements ($N = 10$). The symmetric version ($s = -1$), the skew-symmetric version ($s = 1$), as well as the neutral version ($s = 0$) are considered. We confirm that the skew-symmetric version is stable for all positive ϵ, while the other two formulations are unstable if the interior penalty stabilization is too small. Nevertheless, the symmetric version attains more accurate numerical solutions and is preferred (see [17]). We also observe that the MDG scheme needs less interior penalty stabilization than the DG scheme. This is not surprising: indeed, roughly speaking, in the diffusive regime, $\mathfrak{T}_h(\overline{V}_h, 0)$ is composed of functions that are *almost* continuous, and therefore the interior penalty stabilization is needed only on the boundary of Ω.

5.2. The SUPG parameter and the inf-sup stability of MDG. Second, we analyze the role of the streamline stabilization. We select, from now on, the symmetric version ($s = -1$) and we take $\varepsilon = 6$ (this gives sufficient interior penalty stabilization

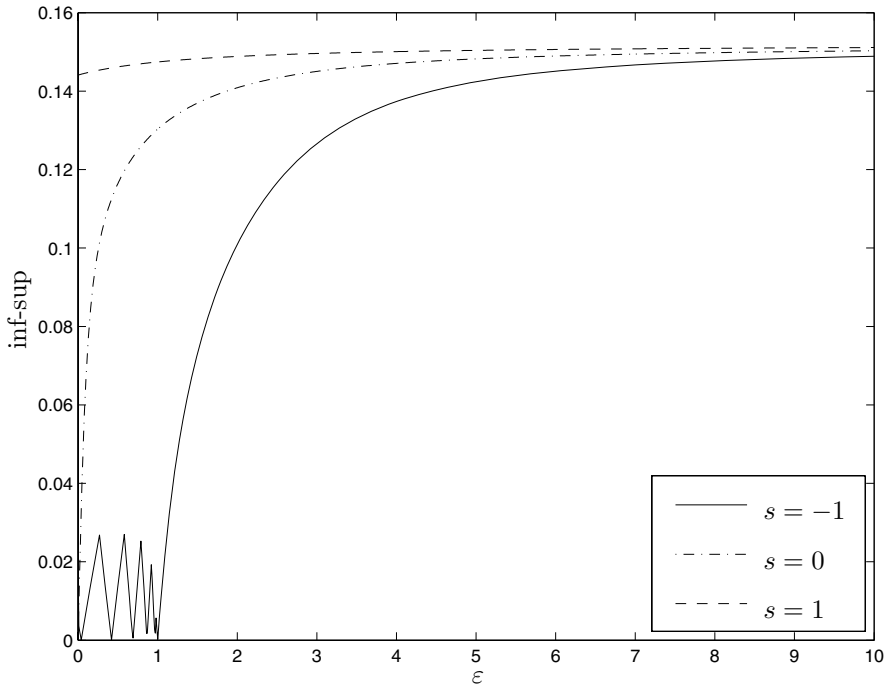


FIG. 5.2. *Inf-sup constant of the MDG method versus* ε.

to both DG and SDG, as seen in section 5.1). We know, from Theorem 3.5, that there is no need of streamline stabilization in the DG method. This is confirmed in Figure 5.3, where (5.1) is plotted for different κ and $\boldsymbol{a} = [\cos\theta, \sin\theta]$, on a grid of 10×10. We have set $\tau = 1/2$ in the definition of $\|\cdot\|_{SDG}$. The values of (5.1) are bounded away from zero, uniformly with respect to the operator coefficients. In Figure 5.4 we focus the attention on the convection-dominated regime, which is now the most interesting case: we set $\kappa = 10^{-6}$ and compute (5.1) for different $\boldsymbol{a} = [\cos\theta, \sin\theta]$ on different uniform meshes of $N \times N$ elements. We confirm that the inf-sup condition holds uniformly with respect to the mesh-size.

The major result of this section is the evaluation of the stability of the MDG scheme. Actually, the MDG scheme turns out to be stable with respect to the $\|\cdot\|_{SDG}$ for the model case considered here: in Figure 5.5 we plot the inf-sup constant (5.2) for different κ and $\boldsymbol{a} = [\cos\theta, \sin\theta]$ on the uniform 10×10 grid, while in Figure 5.6 we plot (5.2) in the convection-dominated regime ($\kappa = 10^{-6}$) for different directions of the convective field $\boldsymbol{a}$ and different uniform meshes. Our conclusion is that, at least for this model case, the MDG scheme is inf-sup stable; that is, condition (4.23) holds with β_{MDG} independent of the problem coefficients and the mesh-size. From this, and reasoning as in Theorem 4.6, we can infer the optimal error estimate for the MDG scheme:

$$\|\phi - \phi^{MDG}\|_{SDG} \lesssim \left(\sum_{T \in \mathcal{T}_h} \left(a_T h_T^{2k+1} + \kappa_T h_T^{2k} \right) |\phi|^2_{H^{k+1}(T)} \right)^{1/2}. \tag{5.3}$$

Similar plots and results are obtained for the stabilized SDG and SMDG methods, in accordance with Proposition 3.3, and are omitted.

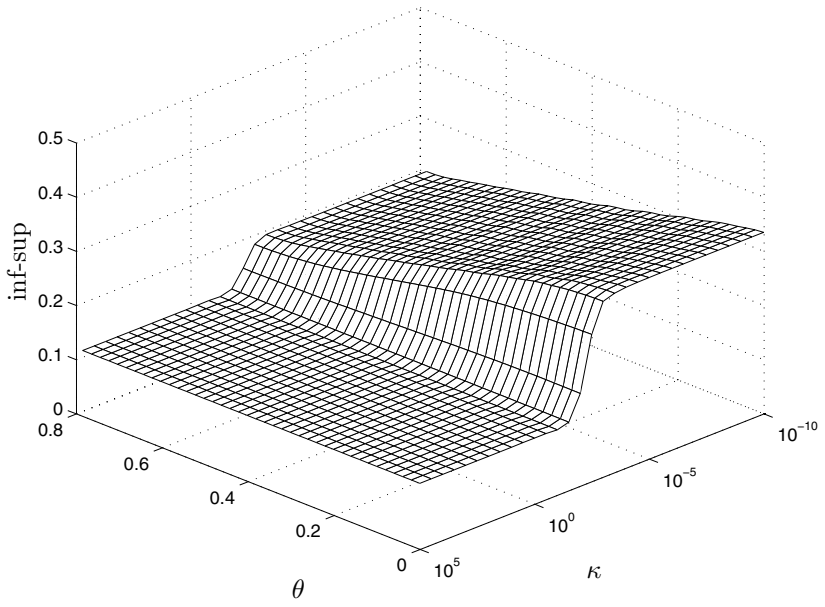


FIG. 5.3. *Inf-sup constant of the DG method versus* $\boldsymbol{a} = [\cos\theta, \sin\theta]$ *and* κ.

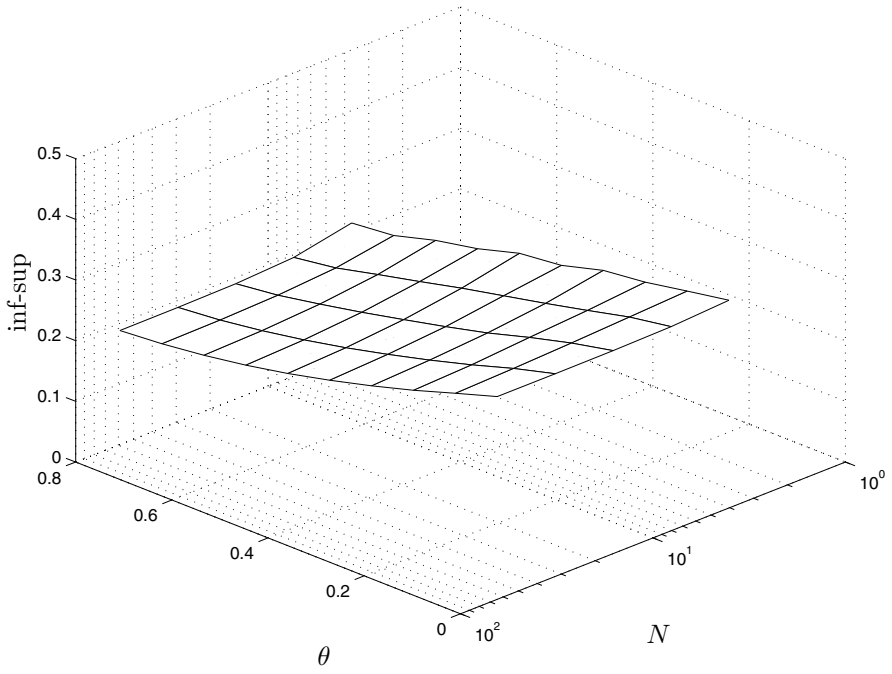


FIG. 5.4. *Inf-sup constant of the DG method versus* $\boldsymbol{a} = [\cos\theta, \sin\theta]$ *and* N.

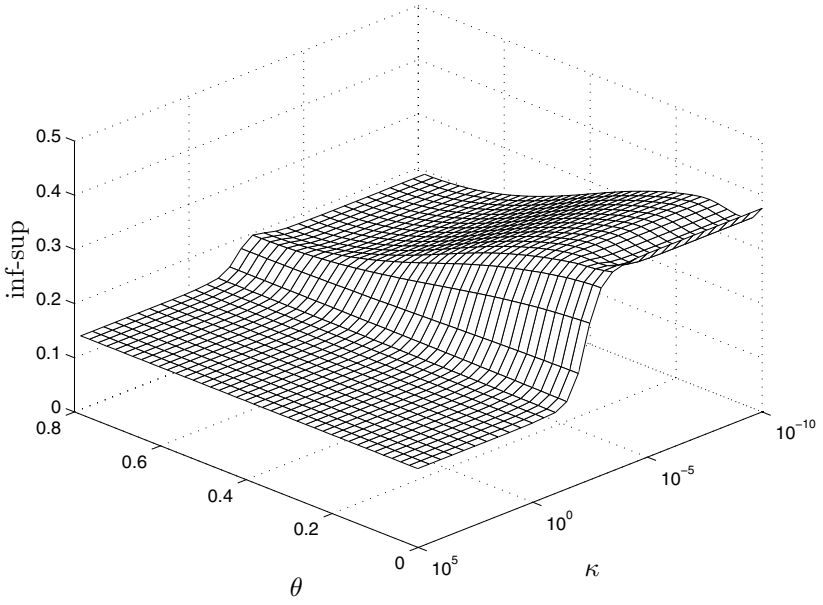


FIG. 5.5. *Inf-sup constant of the MDG method versus* $\boldsymbol{a} = [\cos\theta, \sin\theta]$ *and* κ.

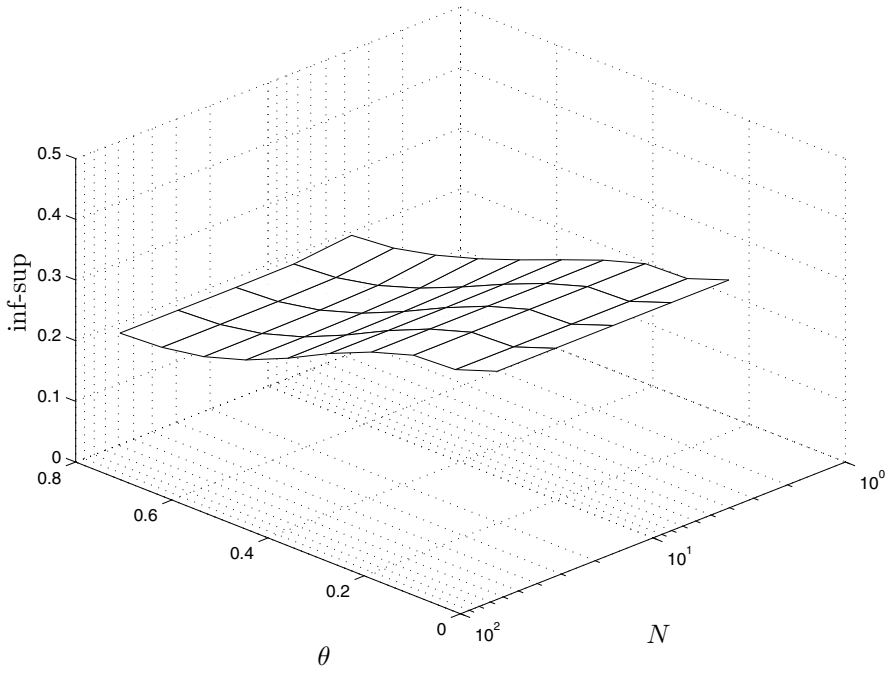


FIG. 5.6. *Inf-sup constant of the MDG method versus* $\boldsymbol{a} = [\cos\theta, \sin\theta]$ *and* N.

6. Conclusions. The mathematical analysis of the multiscale discontinuous Galerkin MDG method introduced in [17] was initiated. This method alleviates a longstanding drawback of discontinuous Galerkin methods, namely, the large size of the solution space. It utilizes local, elementwise problems to generate an interscale transfer operator, enabling the size of the matrix problem to be significantly reduced, apparently without degradation in the quality of results.

We studied MDG and a stabilized version, SMDG. We were able to characterize the approximation properties of the interscale transfer spaces. The corresponding global discontinuous Galerkin methods, DG and SDG, are inf-sup stable and coercive, respectively, with respect to the norm induced by the bilinear form of SDG. Coercivity, but not necessarily inf-sup stability, is inherited by the interscale transfer subspaces. Consequently, we were able to obtain the same error estimates for SMDG as for DG and SDG, but the situation for MDG remains open. Numerical evaluations of the inf-sup constant for MDG indicated that it was positive, bounded uniformly away from zero, and very similar to that for DG. These results are consistent with the numerical calculations performed in [17].

Acknowledgments. The first and third authors thank the Institute for Computational Engineering and Sciences (University of Texas at Austin) for kind hospitality. We thank Guglielmo Scovazzi for helpful discussions.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] K.-J. Bathe, D. Hendriana, F. Brezzi, and G. Sangalli, *Inf-sup testing of upwind methods*, Internat. J. Numer. Methods Engrg., 48 (2000), pp. 745–760.

[3] P. Bochev, T. J. R. Hughes, and G. Scovazzi, *A multiscale discontinuous Galerkin method*, in Proceedings of the 5th International Conference on Large-Scale Scientific Computing, Lecture Notes in Comput. Sci. 3743, Springer-Verlag, Berlin, 2006, pp. 84–93.

[4] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–339.

[5] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.

[6] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.

[7] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations* (FENOMECH '81, Part I, Stuttgart, 1981), Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[8] Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[9] B. Cockburn, *Discontinuous Galerkin methods for computational fluid mechanics*, in Encyclopedia of Computational Mechanics, Vol. 3, E. Stein, R. de Borst, and T. J. R. Hughes, eds., Wiley, Chichester, England, 2004, pp. 91–127.

[10] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000.

[11] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[12] T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[13] T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson, *A comparison of discontinuous and continuous Galerkin methods based on error estimates, conservation, robustness and efficiency*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000, pp. 135–146.

[14] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale method—A paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[15] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics*. VIII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[16] T. J. R. Hughes and G. Sangalli, *Variational Multiscale Analysis: The Fine-scale Green's Function, Projection, Optimization, Localization, and Stabilized Methods*, ICES Report 05-46, University of Texas, Austin, TX, 2005.

[17] T. J. R. Hughes, G. Scovazzi, P. Bochev, and A. Buffa, *A multiscale discontinuous Galerkin method with the computational structure of a continuous Galerkin method*, Comput. Meth. Appl. Mech. Engrg., 195 (2006), pp. 2761–2787.

[18] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[19] C. Johnson, and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[20] G. Sangalli, *A discontinuous residual-free bubble method for advection-diffusion problems*, J. Engrg. Math., 49 (2004), pp. 149–162.

[21] G. Sangalli, *Global and local error analysis for the residual-free bubbles method applied to advection-dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.

[22] E. Süli, P. Houston, and C. Schwab, *hp-finite element methods for hyperbolic problems*, in The Mathematics of Finite Elements and Applications, X, MAFELAP 1999, J. R. Whiteman, ed., Elsevier, Oxford, UK, 2000, pp. 143–162.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1441–1465

TRUNCATION ERRORS IN EXPONENTIAL FITTING FOR OSCILLATORY PROBLEMS*

J. P. COLEMAN† AND L. GR. IXARU‡

Abstract. A generalization of Peano's kernel theorem due to Ghizzetti and Ossicini [*Quadrature Formulae*, Birkhaüser, Basel, Switzerland, 1970] provides expressions, in the form of integrals, for the truncation errors in a variety of exponential-fitting formulae for oscillatory problems. In some circumstances this leads to an expression analogous to the Lagrange form of remainder; more generally the error can be expressed as a sum of two terms of Lagrange type. Our examples include formulae for quadrature and numerical differentiation, and linear multistep methods for ordinary differential equations. Two families of exponential-fitting quadrature formulae are investigated, one with evenly spaced abscissas and the other based on the philosophy of Gaussian quadrature. In particular, the integral representation can be used to determine the asymptotic rate of decay of the error with increasing frequency for a class of oscillatory integrands.

Key words. truncation error, exponential fitting, oscillatory problems, quadrature, Gaussian quadrature, oscillatory integrands, Filon quadrature

AMS subject classifications. 65D30, 65D32, 65D20, 65L70

DOI. 10.1137/050641752

1. Introduction.

In recent decades a sustained effort has been devoted to the construction of approximate formulae specially adapted for numerical operations on oscillatory functions. Integration of differential equations with oscillatory solutions is an area of current research interest; aspects of the problem were reviewed by Petzold, Jay, and Yen [17]. Quadrature with oscillatory integrands is of interest in its own right, for example, in connection with finite Fourier integrals. Furthermore, the efficient evaluation of such integrals is an essential requirement in the implementation of some methods for differential equations with oscillatory solutions; see Iserles [9].

Many classical approximation formulae are designed to be exact for polynomials of sufficiently low degree. However, a polynomial of low degree can provide a good approximation for a rapidly varying function only on a very short interval. In contrast, approximations based on a few terms of a more appropriate set of basis functions may provide a much better approximation. An exponential-fitting method is designed to be exact when the solution of the differential equation, or the integrand in a quadrature problem, is some suitably chosen combination of exponential functions, perhaps with polynomial terms, or products of polynomials and exponentials. Similar ideas may be applied to interpolation and numerical differentiation, as in [11]. The technique for the construction of the coefficients of such formulae is now well established, and some efficient computer codes are available; see Ixaru and Vanden Berghe [14] and the CD therein. Since we are concerned here with oscillatory problems, the arguments of the exponential functions are purely imaginary.

When the value of a function f at $x_0 + h$ is approximated by a truncated Taylor expansion about x_0, that is, by $f_K(x_0 + h) = \sum_{k=0}^{K} h^k f^{(k)}(x_0)/k!$, the resulting error

*Received by the editors October 3, 2005; accepted for publication (in revised form) March 3, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sinum/44-4/64175.html
†Department of Mathematical Sciences, University of Durham, South Road, Durham DH1 3LE, England (john.coleman@durham.ac.uk).
‡Institute of Physics and Nuclear Engineering, Division of Fundamental Physics, PO Box MG-6, Bucharest, Romania (ixaru@theory.nipne.ro).

may be expressed in the Lagrange form

$$err := f(x_0 + h) - f_K(x_0 + h) = \frac{h^{K+1}}{(K+1)!} f^{(K+1)}(\eta),$$

for some $\eta \in (x_0, x_0 + h)$, if $f^{(K+1)}(x)$ is continuous on $[x_0, x_0 + h]$. That error may also be written, less usefully, as the formal expansion $\sum_{k=K+1}^{\infty} h^k f^{(k)}(x_0)/k!$. Expressions of Lagrange type are also available for the truncation errors in some polynomial-based approximations; examples include the errors in Newton–Cotes and Gaussian quadrature formulae, and the local truncation errors in the explicit and implicit Adams formulae. Although such formulae rarely give useful quantitative bounds, they can give valuable qualitative information. In contrast to the situation for polynomial-based methods, the truncation errors of exponential-fitting methods are not well understood. Comments in the literature are mainly based on the so-called leading term of a formal expansion of the error.

Our objective in this paper is to contribute to the understanding of the truncation errors in exponential fitting methods wherever they are applied. Our starting point is the work of Ghizzetti and Ossicini [8] concerning a class of quadrature formulae which includes exponential-fitting formulae. Their results, which are summarized in section 2, provide both a technique for deriving formulae and, of more relevance to our concern, an expression for the error as an integral. The error formula is a generalization of Peano's kernel theorem. If the kernel function is of constant sign, this leads to an expression analogous to the Lagrange form of remainder. More generally, the error can be expressed as a sum of two terms of Lagrange type.

Several specific formulae obtained by exponential fitting are considered in section 3. It is found that the error is expressible as a single Lagrange term only under severe restrictions on the value of the fitted frequency, and the relationship with the concept of the "leading term" is elucidated. The next section begins with a derivation of an explicit formula for the kernel function for exponential fitting based on the $2N$ functions $x^p \exp(\pm i\omega x)$ for $p = 0, 1, \ldots, N-1$. That provides an expression for the truncation error for a particular class of exponential-fitting quadrature formulae and paves the way for analytical and numerical investigations of the errors in two families of such formulae, one with evenly spaced abscissas and the other based on the philosophy of Gaussian quadrature. Some of the methods investigated here compare favorably with a Filon–Lobatto method advocated by Iserles [9], when applied to an example which he considered.

Section 5 is devoted to a detailed study of the error in a particular two-point exponential-fitting Gaussian formula for a class of oscillatory integrands. It is shown that the integral for the truncation error can provide valuable information on the qualitative behavior of that error at large frequencies.

2. A general scheme. The work of Ghizzetti and Ossicini [8] is concerned with quadrature formulae of the form

$$\int_a^b g(x) f(x)\, dx = \sum_{i=1}^{n} \sum_{k=0}^{m-1} A_{ki} f^{(k)}(x_i) + E[f] \tag{2.1}$$

such that $E[f] = 0$ when f is a solution of a linear differential equation $L[f] = 0$ of order m. It is assumed that

$$a \le x_1 < x_2 < \cdots < x_n \le b.$$

The operator L has the form

$$L = \sum_{k=0}^{m} a_k(x) \frac{d^{m-k}}{dx^{m-k}},$$

with $a_0(x) = 1$. Smoothness conditions on the coefficients are specified in [8]. Here it will be assumed that the functions arising are as smooth as required. The adjoint differential operator L^* is defined by

$$L^*[u] = \sum_{k=0}^{m} (-1)^{m-k} \frac{d^{m-k}}{dx^{m-k}} [a_k(x)u(x)],$$

which may also be written as

$$L^*[u] = \sum_{k=0}^{m} a_k^*(x) \frac{d^{m-k}u(x)}{dx^{m-k}}.$$

The reduced operators corresponding to L are

$$L_r = \sum_{k=0}^{r} a_k(x) \frac{d^{r-k}}{dx^{r-k}} \qquad \text{for} \quad r = 0, \ldots, m-1.$$

Similarly,

$$L_r^* = \sum_{k=0}^{r} a_k^*(x) \frac{d^{r-k}}{dx^{r-k}} \qquad \text{for} \quad r = 0, \ldots, m-1.$$

For example, if

$$L = \frac{d^2}{dx^2} + a_1(x) \frac{d}{dx} + a_2(x),$$

then

$$L_1 = \frac{d}{dx} + a_1(x) \quad \text{and} \quad L_0 = 1.$$

For a sufficiently differentiable function u,

$$L^*[u] = u'' - (a_1 u)' + a_2 u;$$

thus

$$L_1^*[u] = u' - a_1 u \quad \text{and} \quad L_0^*[u] = u.$$

For any two sufficiently differentiable functions u and v, Lagrange's identity takes the form

$$vL[u] - uL^*[v] = \frac{d}{dx} \sum_{k=0}^{m-1} u^{(k)}(x) L_{m-k-1}^*[v].$$

Integration gives

$$\int_\alpha^\beta u(x)L^*[v](x)\,dx = \int_\alpha^\beta v(x)L[u](x)\,dx - \left[\sum_{k=0}^{m-1} u^{(k)}(x) L_{m-k-1}^*[v](x) \right]_{x=\alpha}^{x=\beta}.$$

In particular, this can be applied with $u = f$, $v = \phi_i$ (any solution of the differential equation $L^*[\phi] = g(x)$), $\alpha = x_i$, and $\beta = x_{i+1}$. It is convenient to define $x_0 = a$ and $x_{n+1} = b$, to allow for cases where the end-points of the integration interval are not quadrature abscissas. Let ϕ_0 and ϕ_n be the particular solutions of $L^*[\phi] = g(x)$ which satisfy the conditions

$$L^*_{m-k-1}[\phi_0](a) = 0$$

and

$$L^*_{m-k-1}[\phi_n](b) = 0$$

for $k = 0, 1, \ldots, m-1$. Then, by summing over the index i, as shown by Ghizzetti and Ossicini [8], and also on page 294 of Davis and Rabinowitz [4], we get a quadrature formula of the form (2.1). The quadrature coefficients are

$$A_{ki} = L^*_{m-k-1}\left[\phi_i(x) - \phi_{i-1}(x)\right]\Big|_{x=x_i} \tag{2.2}$$

for $k = 0, \ldots, m-1$ and $i = 1, \ldots, n$; here the functions ϕ_i, for $i = 1, \ldots, n-1$, are any solutions of the differential equation $L^*[\phi](x) = g(x)$. The error term is

$$E[f] = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} \phi_i(x) L[f](x)\, dx = \int_a^b \Phi(x) L[f](x)\, dx, \tag{2.3}$$

where

$$\Phi(x) = \phi_i(x) \qquad \text{for} \quad x_i < x < x_{i+1}, \quad i = 0, \ldots, n.$$

It is clear from the form of (2.3) that the truncation error vanishes if $L[f] = 0$.

Milne [16] approached this problem in a different way, expressing the approximating formula and its error in terms of determinants. He also expressed the truncation error in the form (2.3), but without associating the kernel Φ with solutions of the adjoint equation.

If $f^{(m)}$ is bounded, then

$$|E[f]| \le \sup_{a \le x \le b} \Big|L[f](x)\Big| \int_a^b |\Phi(x)|\, dx.$$

If $f \in C^m(a, b)$ and the kernel $\Phi(x)$ is of constant sign, the second mean-value theorem for integrals gives

$$E[f] = L[f](\eta) \int_a^b \Phi(x)\, dx \tag{2.4}$$

for some $\eta \in (a, b)$. However, $\Phi(x)$ may not be of constant sign, and even when it is that fact can be difficult to establish.

We can always write $\Phi(x) = \Phi_+(x) + \Phi_-(x)$, where

$$\Phi_+(x) := \begin{cases} \Phi(x) & \text{for all } x \text{ such that } \ \Phi(x) \ge 0, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Phi_-(x) := \begin{cases} \Phi(x) & \text{for all } x \text{ such that } \ \Phi(x) \le 0, \\ 0 & \text{otherwise} \end{cases}$$

so that the integral in (2.3) can be expressed as the sum of two integrals,

$$E[f] = \int_a^b \Phi_+(x) L[f](x)\,dx + \int_a^b \Phi_-(x) L[f](x)\,dx. \tag{2.5}$$

Assuming that $f \in C^m(a,b)$, the mean-value theorem can be applied to both integrals to give

$$E[f] = L[f](\eta_+) \int_a^b \Phi_+(x)\,dx + L[f](\eta_-) \int_a^b \Phi_-(x)\,dx \tag{2.6}$$

for some $\eta_+ , \eta_- \in (a\,, b)$.

There is an immense amount of freedom to derive formulae here. Only the functions ϕ_0 and ϕ_n are determined. Every other ϕ_i can be any solution of the inhomogeneous adjoint equation $L^*[\phi](x) = g(x)$. Equation (2.2) would allow us to calculate the quadrature coefficients corresponding to any particular choice of the functions $\phi_i(x)$, but perhaps the more interesting situation is that in which the coefficients have been determined in some other way and we want to have an expression for the truncation error.

2.1. The truncation error of a given formula. Suppose that the coefficients A_{ki} are known. The expression (2.3) for the truncation error may be used if we can find the corresponding functions ϕ_i. Theorem 2.4.1 of Ghizzetti and Ossicini [8] gives

$$\phi_i(x) = -\int_a^x K(t,x) g(t)\,dt + \sum_{k=0}^{m-1} \sum_{j=1}^{i} A_{kj} \left[\frac{\partial^k}{\partial t^k} K(t,x) \right]_{t=x_j} \tag{2.7}$$

for $i = 0, 1, \ldots, n$, with the convention that the sum does not appear when $i = 0$. Here K is the resolvent kernel corresponding to the operator L; i.e., $K(x,z)$ is the solution of $L[u](x) = 0$ such that

$$\left[\frac{\partial^k}{\partial x^k} K(x,z) \right]_{x=z} = \delta_{k,m-1} \tag{2.8}$$

for $k = 0, 1, \ldots, m-1$. It follows that

$$\phi_{i+1}(x) = \phi_i(x) + \sum_{k=0}^{m-1} A_{k,i+1} \left[\frac{\partial^k}{\partial t^k} K(t,x) \right]_{t=x_{i+1}};$$

thus it is easy to build up the ϕ-functions recursively once $K(t,x)$ and $\phi_0(x)$ are known.

3. Some examples. Here we investigate the errors in several exponential-fitting formulae by means of the general scheme presented in the previous section. We are especially interested in checking if the function $\Phi(x)$ is of constant sign. It will be seen that in most cases this is not true and therefore the Lagrange form of remainder (2.4) does not hold, but (2.6) always applies for sufficiently smooth functions.

3.1. Quadrature based on function values only. The integer m in (2.1) is the order of the differential operator L. For quadrature formulae of the form

$$\int_a^b g(x) f(x)\,dx \approx \sum_{i=1}^{n} a_i f(x_i),$$

exact for functions f which are annihilated by L, we have $A_{0i} = a_i$ for $i = 1, \ldots, n$ and, for those values of i, $A_{ki} = 0$ for $k = 1, \ldots, m-1$. Then (2.7) becomes

$$\phi_i(x) = -\int_a^x K(t,x)g(t)\,dt + \sum_{k=1}^{i} a_k K(x_k, x). \tag{3.1}$$

Example 1. The formula

$$\int_0^h f(x)\,dx \approx \frac{1-\cos\omega h}{\omega \sin\omega h}[f(0) + f(h)] \tag{3.2}$$

is exact for $\cos\omega x$ and $\sin\omega x$. This formula and its error were derived by Vanden Berghe, de Meyer, and Vanthournout [19] as an example of Ehrenmark's technique [6].

Here $g(x) = 1$, $n = 2$, and $m = 2$ with $L[f] = f'' + \omega^2 f$. Also, $x_0 = x_1 = 0$ and $x_2 = x_3 = h$. The function $K(x,t)$ may be expressed as

$$K(x,t) = A\cos\omega(x-t) + B\sin\omega(x-t).$$

The constants A and B are determined by the initial conditions (2.8) and we find that

$$K(x,t) = \frac{1}{\omega}\sin\omega(x-t). \tag{3.3}$$

Integration of $-K(t,x)$ with respect to t, from 0 to x, gives $\phi_0(x) = (1-\cos\omega x)/\omega^2$ and, from (3.1),

$$\phi_1(x) = \phi_0(x) + \frac{1-\cos\omega h}{\omega\sin\omega h}K(0,x) = \frac{1}{\omega^2} + \frac{\sin\omega(x-h) - \sin\omega x}{\omega^2\sin\omega h}.$$

To investigate the sign of $\phi_1(x)$ on $(0,h)$ it is convenient to define $y = \omega x$, $\theta = \omega h$, and

$$F(y,\theta) = \frac{\sin(y-\theta) - \sin y}{\sin\theta} + 1 = 1 - \frac{\cos(y-\theta/2)}{\cos(\theta/2)}$$

for $\theta \neq (2m+1)\pi$ for integral m. Then $F(y,\theta) = 0$ if and only if $y = 2k\pi$ or $y = 2k\pi + \theta$ for integral k. It follows that $F(y,\theta)$ does not change sign on $(0,\theta)$ when $\theta < 2\pi$. For $\theta > 2\pi$, however, $F(y,\theta)$ changes sign when $y = \theta - 2\pi$. We can conclude that, for steplengths h such that $0 < \theta < 2\pi$ and $\theta \neq \pi$, the truncation error in (3.2) for a sufficiently smooth integrand is

$$E[f] = \left[f''(\eta) + \omega^2 f(\eta)\right]\int_0^h \phi_1(x)\,dx = \frac{2}{\omega^3}\left(\frac{\theta}{2} - \tan\frac{\theta}{2}\right)\left[f''(\eta) + \omega^2 f(\eta)\right] \tag{3.4}$$

for some $\eta \in (0,h)$, a result obtained earlier by Vanden Berghe, de Meyer, and Vanthournout [19]. This analysis does not allow any such conclusion if $\theta > 2\pi$.

It is instructive to see how this result compares with what is available in the literature on exponential fitting, where the error is presented as a formal infinite series; see Ixaru and Vanden Berghe [14]. If $\mathcal{O}$ is an operator acting on some function $y(x)$ and $\mathcal{A}[h, \mathbf{a}]y(x)$ is its approximation with parameters $\mathbf{a} = [a_1, a_2, \ldots]$, then an operator $\mathcal{L}$ is introduced by

$$\mathcal{L}[h, \mathbf{a}]y(x) = (\mathcal{O} - \mathcal{A}[h, \mathbf{a}])y(x).$$

If the coefficients of the formula are determined by requiring that $\mathcal{L}[h, \mathbf{a}]y(x)$ is identically vanishing in x and h when $y(x)$ satisfies the equation $Ly = 0$, where

$$L = D^m + c_1 D^{m-1} + c_2 D^{m-2} + \cdots + c_m$$

with constant $c_1, c_2, \ldots, c_m$ and $D = d/dx$, then, formally,

$$\mathcal{L}[h, \mathbf{a}]y(x) = h^{l+m} \sum_{k=0}^{\infty} h^k T_k^*(\mathbf{z}, \mathbf{a}(\mathbf{z})) D^k L y(x),$$

where the integer l depends of the operator $\mathcal{O}$, and $\mathbf{z} = [z_1, z_2, \ldots]$ is a set of parameters depending on the coefficients c_m and the steplength h. The coefficients T_k^* can be calculated by some algebraic manipulations. In particular, T_0^* has a simple form, as can be seen in the next paragraph and more generally in section 4.2, where this quantity is denoted by T^*.

For the example under discussion we have

$$\mathcal{O}y(x) = \int_x^{x+h} y(x')dx', \qquad \mathcal{A}[h, \mathbf{a}]y(x) = h[a_1 y(x) + a_2 y(x+h)],$$

$m = 2$, $c_1 = 0$, $c_2 = \omega^2$, the vector $\mathbf{z}$ has only one component $z_1 = \theta$, the coefficients of the quadrature formula are $a_1(\theta) = a_2(\theta) = (1 - \cos\theta)/(\theta \sin\theta)$, and

$$T_0^* = \frac{1 - a_1(\theta) - a_2(\theta)}{\theta^2} = \frac{2}{\theta^3}\left(\frac{\theta}{2} - \tan\frac{\theta}{2}\right).$$

The error of the formula (3.2) is then

$$err = \mathcal{L}[h, \mathbf{a}]y(x)|_{x=0}.$$

The first term of the corresponding formal series has the same form as the right-hand side of (3.4), but with the unknown η replaced by 0. However, the new analysis clearly shows that a single term of Lagrange type is correct only if $\theta < 2\pi$.

Example 2. The formula

$$\int_{-h}^{h} f(x)\,dx \approx a_1 f(-h) + a_2 f(0) + a_3 f(h)$$

is exact for 1, x, $\cos\omega x$, and $\sin\omega x$. With $\theta = \omega h$ the coefficients are

$$a_1 = a_3 = \frac{\theta - \sin\theta}{\omega(1 - \cos\theta)}, \qquad a_2 = 2\frac{\sin\theta - \theta\cos\theta}{\omega(1 - \cos\theta)};$$

see equations (4.16) of Ixaru [11]. In this case $g(x) = 1$, and the relevant linear differential operator is

$$L = \frac{d^4}{dx^4} + \omega^2 \frac{d^2}{dx^2};$$

thus $m = 4$. The truncation error is

$$E[f] = \int_{-h}^{h} \Phi(x) L[f](x)\,dx = \int_{-h}^{0} \phi_1(x) L[f](x)\,dx + \int_0^h \phi_2(x) L[f](x)\,dx.$$

The function $K(x,t)$ may be expressed as

$$K(x,t) = A + B(x-t) + C\cos\omega(x-t) + D\sin\omega(x-t).$$

The constants A, B, C, and D are determined by the initial conditions (2.8), and we find that

$$K(x,t) = \frac{x-t}{\omega^2} - \frac{\sin\omega(x-t)}{\omega^3}.$$

Further calculation gives

$$\phi_1(x) = \frac{1}{\omega^4}\left[\frac{1}{2}\omega^2(x+h)^2 - 1 + \cos\omega(x+h) + \frac{\theta-\sin\theta}{1-\cos\theta}\left\{\sin\omega(x+h) - \omega(x+h)\right\}\right].$$

Then

$$\phi_2(x) = \phi_1(x) + a_2K(0,x),$$

and some further algebra shows that $\phi_2(x) = \phi_1(-x)$.

Numerical computations indicate that $\phi_1(x) \le 0$ on $[-h,0]$, for all values of θ for which it is defined, and if this is taken for granted, then

$$\begin{aligned} E[f] &= 2\left[f^{iv}(\eta) + \omega^2 f''(\eta)\right]\int_{-h}^{0}\phi_1(x)\,dx \\ &= \frac{h^5}{6\theta^2}\left[\frac{6}{\theta}\cot\left(\frac{\theta}{2}\right) - 3\cot^2\left(\frac{\theta}{2}\right) - 1\right]\left[f^{iv}(\eta) + \omega^2 f''(\eta)\right], \end{aligned}$$

in accordance with the "leading term" given by (4.17) of [11]. This result was first obtained by Ehrenmark [6] and rederived using ideas of Ghizzetti and Ossicini [8] in [7], where a five-point formula is also considered.

3.2. Numerical differentiation. Equation (2.1) can provide approximations for derivatives in terms of function values if the integral is removed by choosing $g(x) \equiv 0$. In that case the functions ϕ_i satisfy the homogeneous adjoint equation $L^*[\phi](x) = 0$ and, in particular, $\phi_0(x) \equiv 0$ and $\phi_n(x) \equiv 0$. Previously the inhomogeneous term gave the appropriate normalization; now we simply choose a particular value for one of the coefficients.

Example 3. The formula

$$y''(0) \approx a_1[y(h) + y(-h)] + a_2y(0), \tag{3.5}$$

with

$$a_1 = \frac{\theta}{h^2\sin\theta} \qquad \text{and} \qquad a_2 = \frac{\theta(\sin\theta - 2\cos\theta)}{h^2\sin\theta},$$

is exact for $\cos\omega x$, $\sin\omega x$, $x\cos\omega x$, and $x\sin\omega x$. Here

$$L = \left(\frac{d^2}{dx^2} + \omega^2\right)^2.$$

The formula (3.5) corresponds to (2.1) without the error term if we take $g(x) \equiv 0$, $m = 4$, $n = 3$, $x_1 = -h = -x_3$, $x_2 = 0$, $A_{01} = A_{03} = a_1$, $A_{02} = a_2$, $A_{22} = -1$, $A_{21} = 0 = A_{23}$, and $A_{1k} = 0$ for $k = 1,2,3$.

In this case

$$K(x,t) = \frac{1}{2\omega^3}\big[\sin\omega(x-t) - \omega(x-t)\cos\omega(x-t)\big],$$

and

$$\phi_1(x) = a_1 K(-h,x) = \frac{\omega(x+h)\cos\omega(x+h) - \sin\omega(x+h)}{2\omega\theta\sin\theta}.$$

Further calculation gives $\phi_2(x) = \phi_1(-x)$; thus $\Phi(x)$ is an even function.

To investigate the sign of $\phi_1(x)$ on $(-h,0)$, let $z = \omega(x+h)$ and consider

$$u(z) = z\cos z - \sin z$$

on $[0,\theta]$. This oscillatory function of increasing amplitude is negative on $(0,\theta_1)$, where $\theta_1 \approx 4.4934$ is the smallest positive root of the equation $z = \tan z$. On larger intervals $u(z)$ is not of constant sign. For $0 < \theta \le \theta_1$, when $y \in C^4[-h,h]$, the truncation error in the formula (3.5) is

$$\begin{aligned} -E[y] &= -2\left[y^{iv}(\eta) + 2\omega^2 y''(\eta) + \omega^4(\eta)\right]\int_{-h}^{0}\phi_1(x)\,dx \\ &= \frac{2h^2}{\theta^3\sin\theta}(2 + \theta\sin\theta - 2\cos\theta)\left[y^{iv}(\eta) + 2\omega^2 y''(\eta) + \omega^4 y(\eta)\right]. \end{aligned}$$

This is in agreement with the "leading term" given by (3.19) of Ixaru [11]. Here, as elsewhere, the concept of "leading term" needs to be treated with caution; when the mean-value theorem is applicable, η is unknown, and its replacement by some chosen number could give erroneous results. For example, it may be tempting to replace η by 0. If we take $y = x^6$, the "leading term" then vanishes, but the actual error is easily seen to be $-2h^4\theta\operatorname{cosec}\theta$.

It is evident from the symmetry of the formula (3.5) that it is exact for every odd function. That is also clear when the truncation error is expressed as

$$-E[y] = -\int_{-h}^{h}\Phi(x)L[y](x)\,dx,$$

since Φ is an even function and $L[y]$ involves the function y and derivatives of even order only.

3.3. Multistep methods for ODEs. As in section 3.2, we remove the integral in (2.1) by taking $g(x) \equiv 0$.

Example 4. Theorem 4 of Ixaru [11] is concerned with a two-step formula for differential equations of the form $y'' = f(x,y)$, which may be written as

$$y_{n+1} - a_0 y_n + y_{n-1} = h^2(a_1 f_{n+1} + a_2 f_n + a_3 f_{n-1}).$$

If this formula is exact for $x^p\exp(\pm i\omega x)$ with $p = 0,1,2$, the coefficients are

$$a_0 = \frac{6\cos\theta\sin\theta - 2\theta\cos^2\theta + 4\theta}{3\sin\theta + \theta\cos\theta}, \qquad a_1 = a_3 = \frac{\sin\theta}{3\sin\theta + \theta\cos\theta},$$

and

$$a_2 = \frac{4\theta\sin^2\theta - 2\cos\theta(\theta\cos\theta - \sin\theta)}{\theta(3\sin\theta + \theta\cos\theta)}.$$

The two-step formula corresponds to (2.1) with $g(x) \equiv 0$,

$$L = \left(\frac{d^2}{dx^2} + \omega^2\right)^3,$$

$m = 6$, $n = 3$, and the coefficients $A_{01} = -1 = A_{03}$, $A_{02} = a_0$, $A_{11} = 0 = A_{12} = A_{13}$, $A_{21} = A_{23} = h^2 a_1$, $A_{22} = h^2 a_2$, and $A_{ik} = 0$ for $i = 3, 4, 5$ and $k = 1, 2, 3$. In the usual way we find

$$\phi_1(x) = \frac{4\sin\theta - 2\theta\cos\theta}{4\omega^5(3\sin\theta + \theta\cos\theta)}[\sin\omega(x+h) - 3\omega(x+h)\cos\omega(x+h)] - \frac{\sin\theta}{2\omega^3(3\sin\theta + \theta\cos\theta)}(x+h)^2\sin\omega(x+h).$$

A symmetry argument shows that $\phi_2(x) = \phi_1(-x)$.

The local truncation error is

$$-E[y] = -\int_{-h}^{h} \Phi(x) L[y](x)\, dx.$$

When the sign of $\Phi(x)$ is constant on $[-h, h]$, the coefficient of $L[y](\eta)$ in the resulting expression for the truncation error is

$$-\int_{-h}^{h} \Phi(x)\, dx.$$

Evaluation shows that this is the same as the factor quoted in (5.6) of [11]. However, calculations show that the sign of $\Phi(x)$ is no longer constant on $[-h, h]$ when θ exceeds $\theta_{max} \approx 1.625$.

3.4. Interpolation. Modifying (2.1), by introducing a parameter x and taking $m = 1$, we can write

$$\int_a^b g(x,t) f(t)\, dt = \sum_{i=1}^{n} a_i(x) f(x_i) + E[f](x). \tag{3.6}$$

When g is the delta function $g(x,t) = \delta(x-t)$ this becomes

$$f(x) = \sum_{i=1}^{n} a_i(x) f(x_i) + E[f](x)$$

for $x \in (a, b)$. If the functions a_i satisfy the interpolation conditions

$$a_i(x_j) = \delta_{ij} \qquad \text{for} \quad i, j = 1, \ldots, n,$$

the truncation error in the corresponding interpolation formula is given by (2.2) as

$$E[f](x) = \int_a^b \Phi(x,t) L[f](t)\, dt,$$

with

$$\Phi(x,t) = \phi_i(x,t) \qquad \text{for} \quad x_i < t < x_{i+1}, \quad i = 0, \ldots, n,$$

and, from (2.7),

$$\phi_i(x,t) = -\int_a^t K(z,t)\delta(x-z)\,dz + \sum_{j=1}^{n} a_j(x)K(t,x)$$
$$= -K(x,t)\Theta(t-x) + \sum_{j=1}^{n} a_j(x)K(t,x),$$

where Θ is the unit step function

$$\Theta(\alpha) = \begin{cases} 0, & \alpha < 0, \\ 1, & \alpha > 0. \end{cases}$$

Interpolation involving derivative values may be accommodated in this approach by adapting (3.6) to include additional sums.

Example 5. For interpolation by a linear combination of $\cos\omega x$ and $\sin\omega x$, with two interpolation points a and b, a problem considered in [12], the canonical functions are

$$a_1(x) = \frac{\sin\omega(x-b)}{\sin\omega(a-b)}, \qquad a_2(x) = \frac{\sin\omega(x-a)}{\sin\omega(b-a)}.$$

In that case $L[f] = f''+\omega^2 f$, and the kernel function is given by (3.3). The truncation error is

$$E[f](x) = \int_a^b \phi_1(x,t)L[f](t)\,dt,$$

with

$$\phi_1(x,t) = \begin{cases} -\dfrac{\sin\omega(x-b)\sin\omega(a-t)}{\omega\,\sin\omega(b-a)}, & a < t < x, \\ -\dfrac{\sin\omega(x-a)\sin\omega(b-t)}{\omega\,\sin\omega(b-a)}, & x < t < b. \end{cases}$$

The second expression was obtained by using trigonometric identities to simplify the form of $a_1K(a,t) - K(x,t)$. In the limit as $\omega \to 0$,

$$\phi_1(x,t) \to \begin{cases} \dfrac{(x-b)(t-a)}{b-a}, & a < t < x, \\ \dfrac{(x-a)(t-b)}{b-a}, & x < t < b, \end{cases}$$

the Peano kernel for linear interpolation; see, for example, (3.7.12) of [3].

The function $\phi_1(x,t)$ is of constant sign (negative) for all x and t in (a,b) when $\omega(b-a) < \pi$. Then the truncation error of the two-point trigonometric interpolation, for $f \in C^2[a,b]$, is

$$E[f](x) = \left[f''(\eta) + \omega^2 f(\eta)\right] \int_a^b \phi_1(x,t)\,dt$$
$$= \frac{1}{\omega^2}\left[1 - \frac{\cos\omega\left\{(b+a)/2 - x\right\}}{\cos\omega\left\{(b-a)/2 - x\right\}}\right]\left[f''(\eta) + \omega^2 f(\eta)\right]$$

for some $\eta \in (a,b)$. The external factor is the same as that given in [12] and in (4.171) of [14].

Other expressions for the truncation error in mixed interpolation have been obtained in [5], [1], and [2].

4. Quadrature formulae with $L = (D^2 + \omega^2)^N$. In this section we consider a subset of the formulae (2.1), which we write as

$$\int_a^b f(x)dx = \int_{X-h}^{X+h} f(x)dx = \sum_{k=0}^{J} h^{k+1} \sum_{i=1}^{N} a_i^{(k)} f^{(k)}(X + x_i^* h) + \mathcal{L}[h, \mathbf{a}]f(X)\,, \tag{4.1}$$

where $X = (a + b)/2$ and $h = (b - a)/2$. Two families of N-point exponential-fitting quadrature formulae will be considered here, one involving values of the integrand and its first derivative, corresponding to $J = 1$, and the other involving function values only, in which case $J = 0$. For both families we take

$$L = (D^2 + \omega^2)^N,$$

where $D = d/dx$; in the notation of section 2, $m = 2N$.

4.1. The kernel function. A set of $2N$ linearly independent solutions of the differential equation

$$(D^2 + \omega^2)^N y = 0 \tag{4.2}$$

is

$$\{x^p \exp(\pm i\omega x)\}, \quad p = 0, 1, \ldots, N - 1,$$

but it is preferable to use a different set of solutions based on functions which were introduced in [10] and which appear in a slightly different notation in [14]. Let

$$\eta_{-1}(-\omega^2 x^2) := \cos \omega x,$$

and

$$\eta_0(-\omega^2 x^2) := \begin{cases} \dfrac{\sin \omega x}{\omega x} & \text{for } \omega x \neq 0, \\ 1 & \text{for } \omega x = 0. \end{cases}$$

For integers $s > 0$ the functions η_s are defined by the recurrence relation

$$\eta_s(Z) := \frac{1}{Z}[\eta_{s-2}(Z) - (2s - 1)\eta_{s-1}(Z)], \quad s = 1, 2, \ldots, \tag{4.3}$$

and have the power series expansion

$$\eta_s(Z) = 2^s \sum_{q=0}^{\infty} \frac{(q + 1)(q + 2) \cdots (q + s)\, Z^q}{(2q + 2s + 1)!}\,.$$

When $Z \to 0$ they behave as

$$\eta_s(Z) = \frac{2^s s!}{(2s + 1)!} + O\,(Z)\,. \tag{4.4}$$

These functions are related to the spherical Bessel functions by the equation

$$\eta_s(-x^2) = x^{-s} j_s(x), \quad s = 0, 1, 2, \ldots.$$

Differentiation gives

$$\eta_s'(Z) = \frac{1}{2}\eta_{s+1}(Z), \quad s = -1, 0, 1, \ldots. \tag{4.5}$$

Those properties allow us to derive an explicit expression for the resolvent kernel corresponding to the operator $L = (D^2 + \omega^2)^N$.

Let

$$y_N(x) = x^{2N-1}\eta_{N-1}(-\omega^2 x^2).$$

Then (4.5) and (4.3) give

$$y_N'(x) = x^{2N-2}\eta_{N-2}(-\omega^2 x^2),$$

and a second differentiation gives

$$(D^2 + \omega^2)y_N(x) = 2(N-1)x^{2N-3}\eta_{N-2}(-\omega^2 x^2) = 2(N-1)y_{N-1}(x).$$

In particular, $(D^2 + \omega^2)y_1(x) = 0$, with the result that

$$L[y_N](x) := (D^2 + \omega^2)^N y_N(x) = 0.$$

Furthermore, from (4.4),

$$y_N(x) = \frac{2^{N-1}(N-1)!}{(2N-1)!}x^{2N-1} + O\left(x^{2N+1}\right)$$

as $x \to 0$; thus $y_N^{(k)}(0) = 0$ for $0 \le k \le 2N-2$ and $y_N^{(2N-1)}(0) = 2^{N-1}(N-1)!$. It follows that the resolvent kernel corresponding to the operator L is

$$K(t,x) = \frac{y_N(t-x)}{2^{N-1}(N-1)!} = \frac{1}{2^{N-1}(N-1)!}(t-x)^{2N-1}\eta_{N-1}(-\omega^2(t-x)^2). \tag{4.6}$$

The kernel functions derived specifically for Examples 1, 3, and 4 of section 3 are special cases of this formula, corresponding to $N = 1$, 2, and 3, respectively.

4.2. The quadrature error. It is convenient to introduce the dimensionless variables $x^* = (x - X)/h$ and $t^* = (t - X)/h$, in which case x^*, $t^* \in [-1, 1]$. Then

$$K(t,x) = h^{2N-1}K^*(t^*, x^*),$$

where

$$K^*(t^*, x^*) = \frac{1}{2^{N-1}(N-1)!}(t^* - x^*)^{2N-1}\eta_{N-1}(-\theta^2(t^* - x^*)^2). \tag{4.7}$$

The partial derivatives of K with respect to t are

$$\frac{\partial^k}{\partial t^k}K(t,x) = h^{2N-1-k}\frac{\partial^k}{\partial t^{*k}}K^*(t^*, x^*).$$

Then $\Phi(x) = h^{2N}\Phi^*(x^*)$, with

$$\Phi^*(x) = \phi_i^*(x^*) \quad \text{for} \quad x_i^* < x^* < x_{i+1}^*, \qquad i = 0, 1, \ldots, N,$$

where $x_0^* = -1$, $x_{N+1}^* = 1$, and

$$\phi_0^*(x^*) = -\int_{-1}^{x^*} K^*(t^*, x^*)\, dt^*,$$

$$\phi_{i+1}^*(x^*) = \phi_i^*(x^*) + \sum_{k=0}^{J} a_{i+1}^{(k)} \left[\frac{\partial^k}{\partial t^{*k}} K^*(t^*, x^*)\right]_{t^*=x_{i+1}^*} \quad \text{for} \quad i = 0, 1, 2, \ldots, N-1.$$

The formulae given in the previous section are sufficient to build up all these functions in analytic form.

Let $D^* = d/dx^* = hD$. In terms of the dimensionless variables, the quadrature error is

$$err_{ef} = h \int_{-1}^{1} \Phi^*(x^*)(D^{*2} + \theta^2)^N F(x^*)\, dx^*, \tag{4.8}$$

where $F(x^*) = f(x)$.

When the integrand in (4.1) is the unit function, the quadrature error is

$$\int_{X-h}^{X+h} \Phi(x)\omega^{2N}\, dx = h\theta^{2N} \int_{-1}^{1} \Phi^*(x^*)\, dx^* =: h\theta^{2N} T^*(\theta).$$

From the quadrature formula, that error may also be expressed as

$$\int_{X-h}^{X+h} dx - h\sum_{i=1}^{N} a_i^{(0)} = h\left(2 - \sum_{i=1}^{N} a_i^{(0)}\right).$$

Therefore,

$$T^*(\theta) = \frac{2 - \sum_{i=1}^{N} a_i^{(0)}}{\theta^{2N}}.$$

The quadrature coefficients are functions of θ, as in the examples in section 3.

With $\Phi^*(x^*)$ split into $\Phi_+^*(x^*)$ and $\Phi_-^*(x^*)$ as was done earlier for $\Phi(x)$, and with

$$T_\pm^*(\theta) := \int_{-1}^{1} \Phi_\pm^*(x^*)\, dx^*,$$

the quadrature error may be expressed as

$$err_{ef} = h[T_+^*(\theta)(D^{*2} + \theta^2)^N f(\eta_+) + T_-^*(\theta)(D^{*2} + \theta^2)^N f(\eta_-)] \tag{4.9}$$

for some $\eta_+, \eta_- \in (a, b)$, which depend on θ and on the integrand f. It follows that the error can be expressed as a sum of two terms of Lagrange type for *any* θ at which the method is defined, that is, for all θ with the possible exception of a discrete set $\theta_1, \theta_2, \ldots$ of critical values. In at least some cases the simpler form in (2.4) is valid when θ is sufficiently small.

The asymptotic behaviors of $T^*(\theta)$, on one hand, and those of its components $T_\pm^*(\theta)$, on the other, are not necessarily similar. Suppose that for large values of θ the functions $T_\pm^*(\theta)$ are well described by the approximation

$$T_\pm^*(\theta) \approx \pm c(\theta)\theta^{-(2N-\bar{N})} + c_\pm(\theta)\theta^{-2N}, \tag{4.10}$$

where $0 < \bar{N} < 2N$ and the functions $c(\theta)$ and $c_{\pm}(\theta)$, with $c_{+}(\theta) \neq -c_{-}(\theta)$, are oscillating between constant limits; think, for example, of the case where $c(\theta) = c_{+}(\theta) = 1 + \cos\theta$ and $c_{-}(\theta) = -1 + \cos\theta$. Then $T^{*}_{\pm}(\theta)$ will damp out as $\theta^{-(2N-\bar{N})}$, that is, more slowly than their sum $T^{*}(\theta)$ which decays as θ^{-2N}. The determination of this $\bar{N}$ therefore represents a key issue if we want to characterize the asymptotic behavior of the error by means of (4.9).

4.3. Exponential-fitting extended Newton–Cotes rules with $J = 1$. In the N-point rule the (dimensionless) evenly spaced abscissas are

$$x^{*}_{i} = 2(i-1)/(N-1) - 1 \quad (i = 1, 2, \ldots, N), \tag{4.11}$$

and there are $2N$ coefficients to be determined, namely, $a^{(0)}_{i}$ and $a^{(1)}_{i}$ for $i = 1, \ldots, N$. The condition that $\mathcal{L}[h, \mathbf{a}]y(x)$ is identically vanishing in x and in h when $y(x)$ is a solution of the reference differential equation (4.2) is imposed with this aim. The technique is explained in [14]. For the numerical evaluation of these coefficients we used the subroutine `EFEXTQS`.

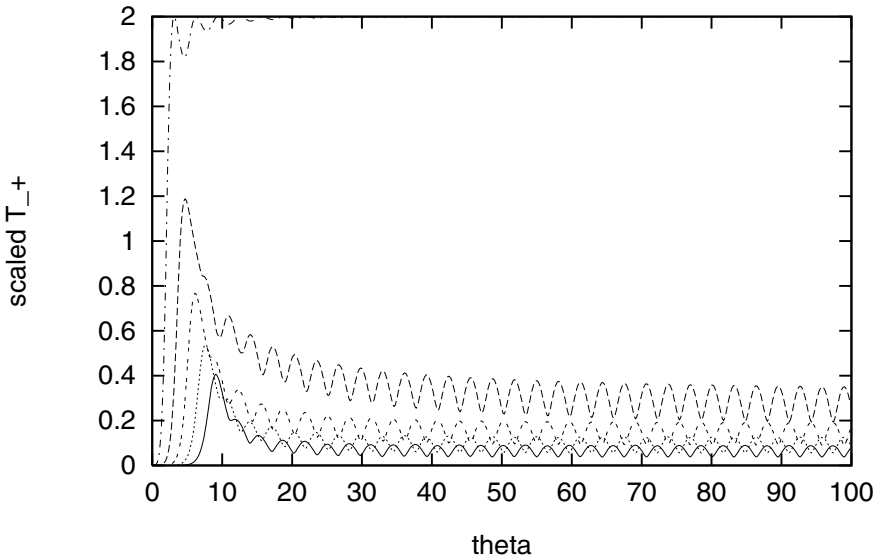


FIG. 1. *Exponential-fitting N-point extended Newton–Cotes rule with $J = 1$: variation with θ of the scaled $T^{*}_{+}(\theta)$ (the first of* (4.12) *with $\bar{N} = N - 2$) for $N = 2$ (dash and dots), $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

To determine $\bar{N}$ for each given N, trial values $0, 1, \ldots$ were considered for $\bar{N}$, and for each of these the behavior of the product $\theta^{2N-\bar{N}}T^{*}_{\pm}(\theta)$ was scanned for large θ. The desired value is that for which this product oscillates between constant limits. This kind of search was undertaken for $2 \leq N \leq 6$, and it led to the conclusion that $\bar{N} = N - 2$. To illustrate this fact, in Figures 1 and 2 we represent the dependence on θ of the functions

$$N^{\bar{N}}\theta^{2N-\bar{N}}T^{*}_{+}(\theta) \text{ and } N^{\bar{N}}\theta^{2N-\bar{N}}T^{*}_{-}(\theta), \tag{4.12}$$

respectively, for $2 \leq N \leq 6$ in Figure 1, and for $3 \leq N \leq 6$ in Figure 2, with $\bar{N} = N - 2$. The factor $N^{\bar{N}}$ was introduced to have the graphs for different N collected in a reduced number of figures.

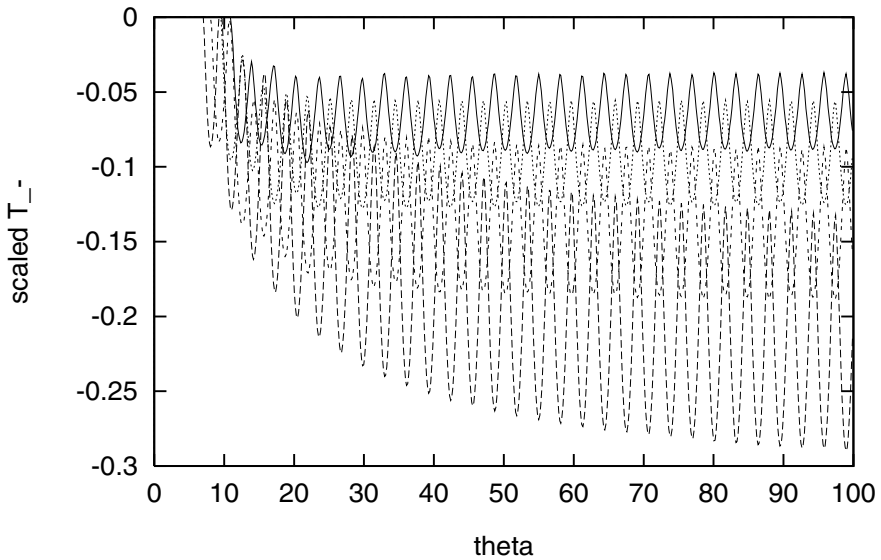


FIG. 2. *Exponential-fitting N-point extended Newton–Cotes rule with $J = 1$: variation with θ of the scaled $T_-^*(\theta)$ (the second of* (4.12) *with $\bar{N} = N - 2$) for $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

It is seen that all these curves exhibit oscillatory variation within limits which become approximately constant as $\theta \to \infty$, thus confirming that $\bar{N} = N - 2$ is the right value of $\bar{N}$. For $N = 2$ we have $\bar{N} = 0$. The reason is that $\Phi^*(x^*)$ is nonnegative everywhere on $x^* \in [-1\,, 1]$ irrespective of θ and then $\Phi_-^*(x^*) = 0$; this is also why the case $N = 2$ is absent from Figure 2.

In order to use our conclusion to predict the asymptotic behavior of the quadrature error as $\theta \to \infty$ it is necessary to specify the class of integrands to be considered. Using the dimensionless variable introduced in section 4.2 we assume that the integrand has the form

$$f(x) = F(x^*) = f_1(x^*)\cos\theta x^* + f_2(x^*)\sin\theta x^*, \tag{4.13}$$

where f_1 and f_2 are sufficiently differentiable functions independent of θ. Then

$$(D^{*2} + \theta^2) f(x) \sim 2\theta\left[f_2'(x^*)\cos\theta x^* - f_1'(x^*)\sin\theta x^*\right],$$

as $\theta \to \infty$. It follows that

$$(D^{*2} + \theta^2)^N f(x) \sim \theta^N g(x^*, \theta), \tag{4.14}$$

where $g(x^*, \theta)$ is bounded as $\theta \to \infty$.

Numerical calculations [15] indicate that $\lim_{\theta\to\infty} a_i^{(0)} = 0$ for $i = 1, 2, \ldots, N$; thus $T^*(\theta) \sim 2\theta^{-2N}$ as $\theta \to \infty$. For integrands of the assumed form the amplitude of the quadrature error given by (4.9) will, in general, damp out asymptotically as $\theta^{\bar{N}-N} = \theta^{-2}$, that is, with one and the same rate for all N. A faster decay is possible only if the unknown points $\eta_\pm^*$ have the property that, in the notation of (4.14), $g(\eta_\pm^*, \theta) = 0$.

To illustrate the quality of this prediction we take the test case

$$I(\omega) = \int_{-1}^{1} \cos[(\omega + 1/2)x]\, dx = \frac{2\sin(\omega + 1/2)}{\omega + 1/2}, \tag{4.15}$$

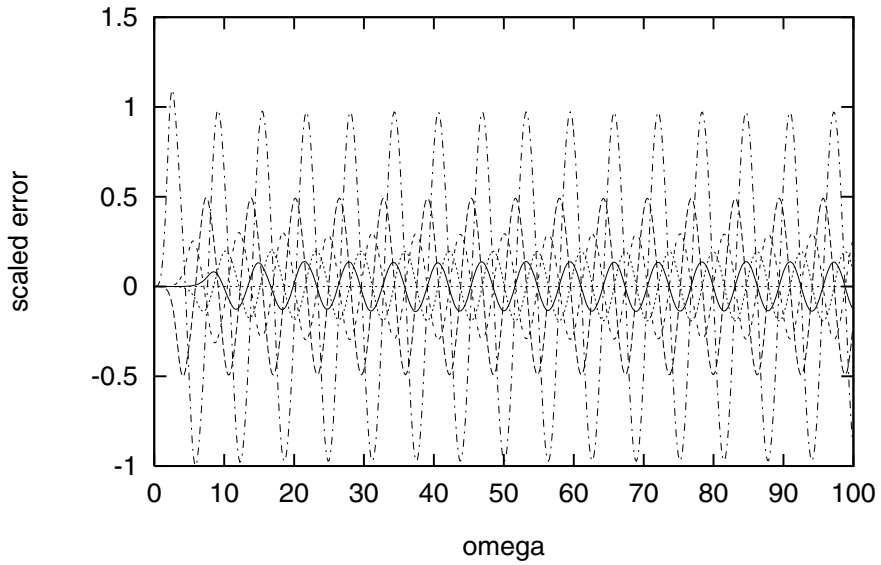


FIG. 3. *Exponential-fitting N-point extended Newton–Cotes rule with $J = 1$ for the integral* (4.15)*: variation with ω of the scaled absolute error* (4.16) *for $N = 2$ (dash and dots), $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

in which $h = 1$ and $\theta = \omega$, and in Figure 3 we present the scaled absolute error

$$N^{\bar{N}}\omega^2[I(\omega) - I_{comput}(\omega)] \tag{4.16}$$

for $2 \le N \le 6$. Each of the five curves oscillates between limits which become approximately constant at large ω, in full agreement with the prediction.

The formal series for the quadrature error, as developed in [14], is

$$\mathcal{L}[h, \mathbf{a}(\theta)]f(X) = h^{2N+1}\sum_{k=0}^{\infty} h^k T_k^*(\theta, \mathbf{a}(\theta))D^k(D^2 + \omega^2)^N f(X)\,. \tag{4.17}$$

The first term of that series is

$$lte_{ef} = hT_0^*(\theta, \mathbf{a}(\theta))(D^{*2} + \theta^2)^N f(X)\,, \tag{4.18}$$

where $T_0^*(\theta, \mathbf{a}(\theta))$ is the function denoted by $T^*(\theta)$ in section 4.2. It is clear from the discussion above that, in general, lte_{ef} damps out as θ^{-N} as $\theta \to \infty$. The authors of [14] recognized that in practice this decay rate is usually too optimistic. In fact, the two-point formula is the only one in this family of methods for which the prediction from the lte_{ef} coincides with that from the integral representation of the error. Our analysis, based on the integral representation of the error, shows why the "leading term" lte_{ef} is not a reliable predictor of the dependence of the quadrature error on θ.

4.4. Exponential-fitting Gaussian rules. These formulae are of the type (4.1) where $J = 0$, but the abscissas are no longer fixed in advance. Again $2N$ parameters are to be determined, but now they are the abscissas x_n^* and the coefficients $a_n^{(0)}$ for $n = 1, 2, \ldots, N$. All these parameters are θ-dependent (see [13]), and they were calculated numerically by the subroutine `EFGAUSS` in [14]. As for the extended Newton–Cotes formulae, numerical computations indicate that $\lim_{\theta\to\infty} a_n^{(0)}(\theta) = 0$ for all relevant values of n.

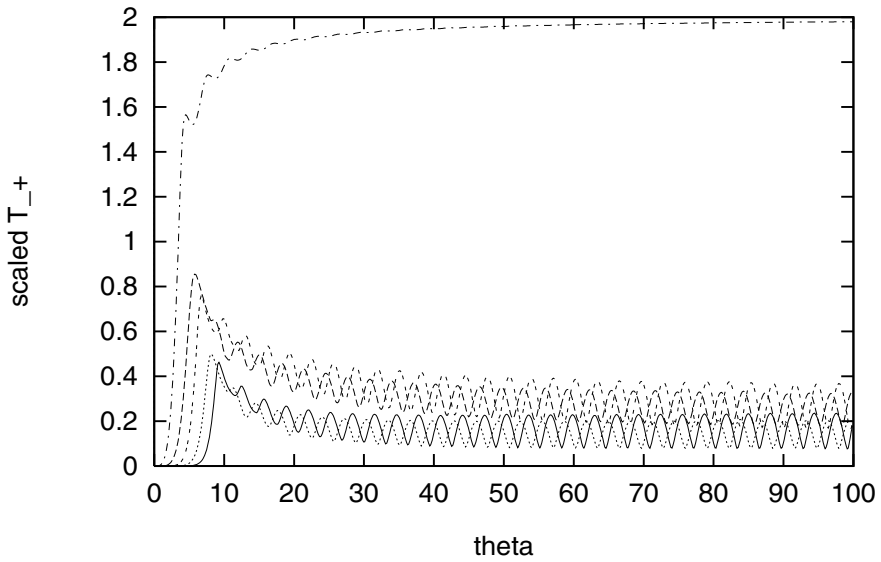


FIG. 4. *Exponential-fitting N-point Gaussian rule: variation with θ of the scaled $T_+^*(\theta)$ (the first of* (4.12) *with $\bar{N} = \lfloor (N-1)/2 \rfloor$) for $N = 2$ (dash and dots), $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

For large θ, the approximation (4.10) applies as in section 4.3, but the values of $\bar{N}$ are different. The same numerical procedure as before was used to determine $\bar{N}$, and in this case we conclude that $\bar{N} = \lfloor (N-1)/2 \rfloor$ where $\lfloor u \rfloor$ is the biggest integer less than or equal to u. For $N \leq 6$ these values are $\bar{N} = 0$ for $N = 1, 2$, $\bar{N} = 1$ for $N = 3, 4$, and $\bar{N} = 2$ for $N = 5, 6$. Thus for $N \geq 4$ they are smaller than the corresponding values for the Newton–Cotes rules. In Figures 4 and 5 we present the two scaled functions displayed in (4.12), the first for $2 \leq N \leq 6$ in Figure 4 and the second for $3 \leq N \leq 6$ in Figure 5; of course, the value $\bar{N} = \lfloor (N-1)/2 \rfloor$ is now used.

For integrands of the form (4.13) this leads to the conclusion that in general the quadrature error decays like $\theta^{\bar{N}-N}$ as $\theta \to \infty$. In particular, for $N = 3$ it decays at least as fast as θ^{-2}, as for the corresponding extended Newton–Cotes rule, but for $N = 4$ or 5 it decays at least as fast as θ^{-3}, and faster decay rates are predicted for larger values of N.

The scaled absolute error

$$N^{\bar{N}} \omega^{N-\bar{N}} [I(\omega) - I_{comput}(\omega)] \tag{4.19}$$

for the test case (4.15) is presented in Figure 6 for $2 \leq N \leq 6$. It confirms the predicted behavior of the quadrature error for this problem. In contrast, the analysis based on the "leading term" of the formal series does not distinguish between the extended Newton–Cotes rule and the corresponding Gaussian rule.

Van Daele, Vanden Berghe, and Vande Vyver [18] recently considered a variety of exponential-fitting Gaussian rules. Truncation error was not their primary concern, and they merely quoted the first nonzero terms in the corresponding formal series for the quadrature error, as developed in [14].

4.5. A numerical comparison with Filon integration. As an illustration of the effectiveness of Filon-type methods for oscillatory integrands, Iserles [9] applied

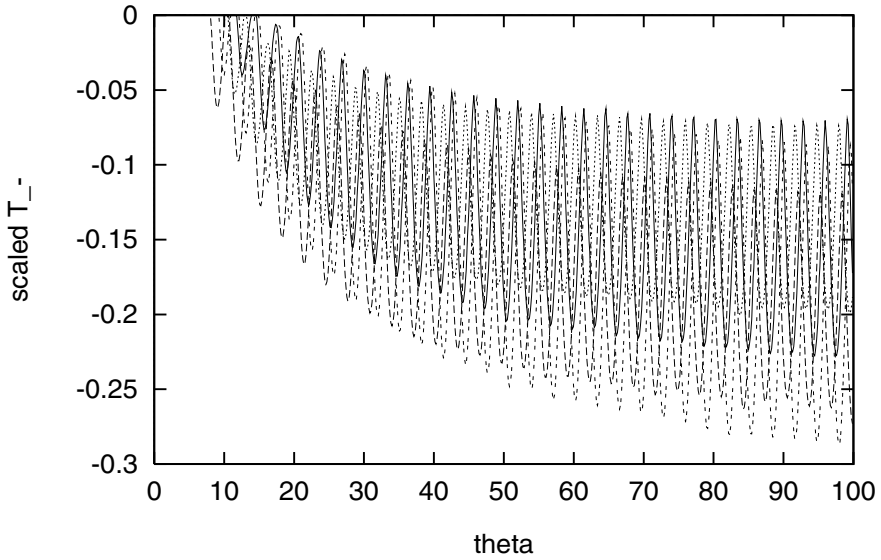


FIG. 5. *Exponential-fitting N-point Gaussian rule: variation with θ of the scaled $T_-^*(\theta)$ (the second of* (4.12) *with $\bar{N} = \lfloor (N-1)/2 \rfloor$) for $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

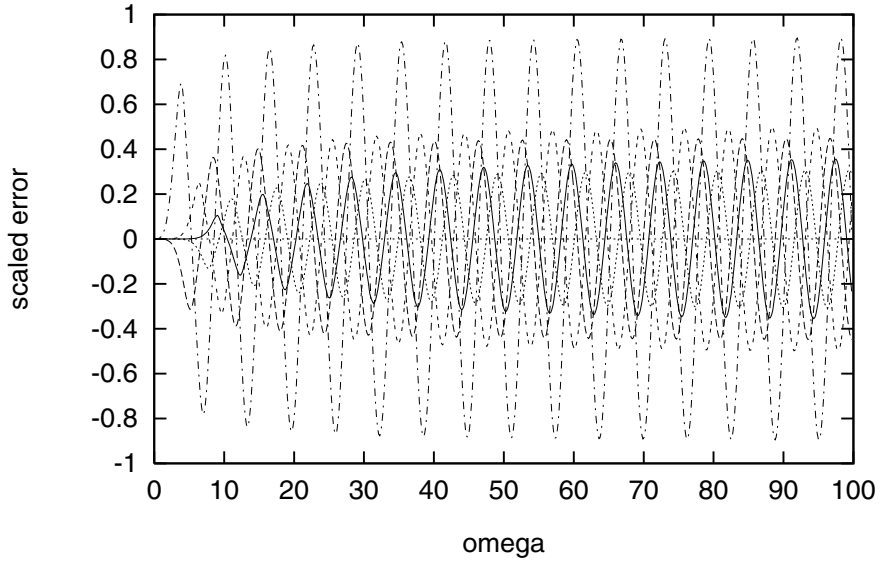


FIG. 6. *Exponential-fitting N-point Gaussian rule for the integral* (4.15): *variation with ω of the scaled absolute error* (4.19) *for $N = 2$ (dash and dots), $N = 3$ (dashes), $N = 4$ (dash pairs), $N = 5$ (dots), and $N = 6$ (solid).*

the three-point Filon–Lobatto formula to the integral

$$I = \int_0^b \exp[(1 + i\omega)x]\, dx = \frac{\exp[(1 + i\omega)b] - 1}{1 + i\omega}$$

for $b = 1/10$, denoted by h in [9]. In our notation, $h = b/2$, so $2\theta = b\omega$.

To provide a comparison with the bottom graph of Figure 3 of [9], our Figure 7 shows the variation with 2θ of the normalized absolute error $(2\theta)^2|I - I_{comput}|$ for

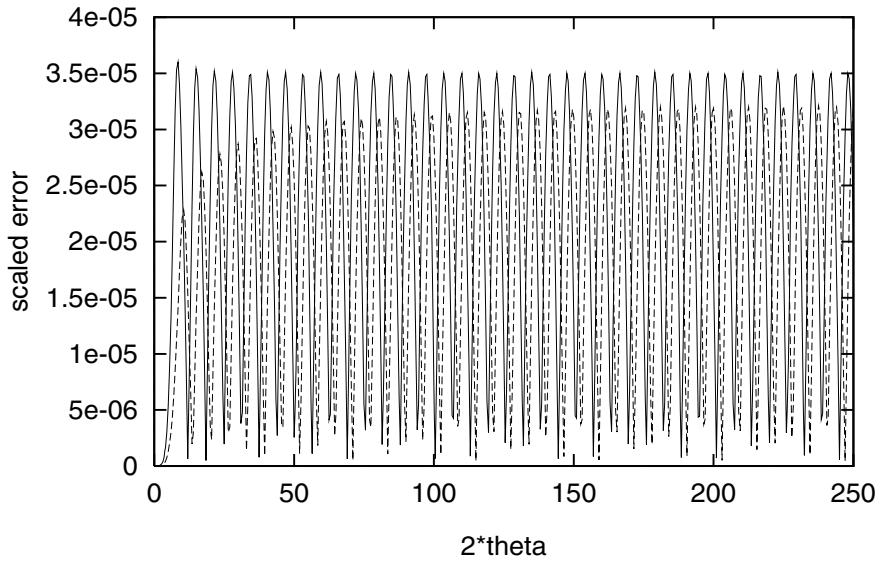


FIG. 7. *Exponential-fitting three-point rules: variation with* 2θ *of the scaled absolute error* $(2\theta)^2|I - I_{comput}|$ *for ENC (solid) and G (broken).*

the two exponential-fitting three-point rules discussed above. The solid line is for the extended Newton–Cotes rule (ENC), and the broken line corresponds to the Gaussian rule (G). To allow the detail to be seen more clearly, the plotting interval here is $[0, 250]$ rather than the larger interval used in [9]. For both methods the error envelope decays as θ^{-2} as predicted. Figure 3 of [9] shows the same rate of decay for Filon–Lobatto integration, but the amplitude of the error is greater by almost an order of magnitude than that of either of the exponential-fitting formulae.

It is also interesting to consider what happens when $N = 4$, since the predicted behaviors of the two exponential-fitting methods are different. The behavior of the extended Newton–Cotes formula is expected to be as for $N = 3$, but for the Gaussian formula the error envelope is expected to decrease at least as fast θ^{-3}. Figure 8 shows $(2\theta)^2|I - I_{comput}|$ for ENC (solid) and $(2\theta)^3|I - I_{comput}|$ for G (broken), confirming those predictions. The analysis of [9] shows that for the Filon–Lobatto method the same rate of decay is expected for $N = 4$ as for $N = 3$.

5. A quadrature formula with $L = D^2(D^2 + \omega^2)$. The way in which the expression for the error was exploited above, to draw conclusions on the behavior of the error at large θ, consisted of combining separate results on the asymptotic behaviors of $T_{\pm}^{*}(\theta)$ and $L[F](x^*)$. It worked successfully in all cases investigated in sections 4.2–4.4, but there are situations when it may fail to reveal correct asymptotic behaviors. Such a case is discussed in this section, where we also present an alternative way based on the use of the integral form of the error without invoking the mean-value theorem.

We consider the two-point exponential-fitting Gaussian rule

$$\int_{X-h}^{X+h} f(x)dx \approx h\left[a_1^{(0)} f(X + x_1^* h) + a_2^{(0)} f(X + x_2^* h)\right], \tag{5.1}$$

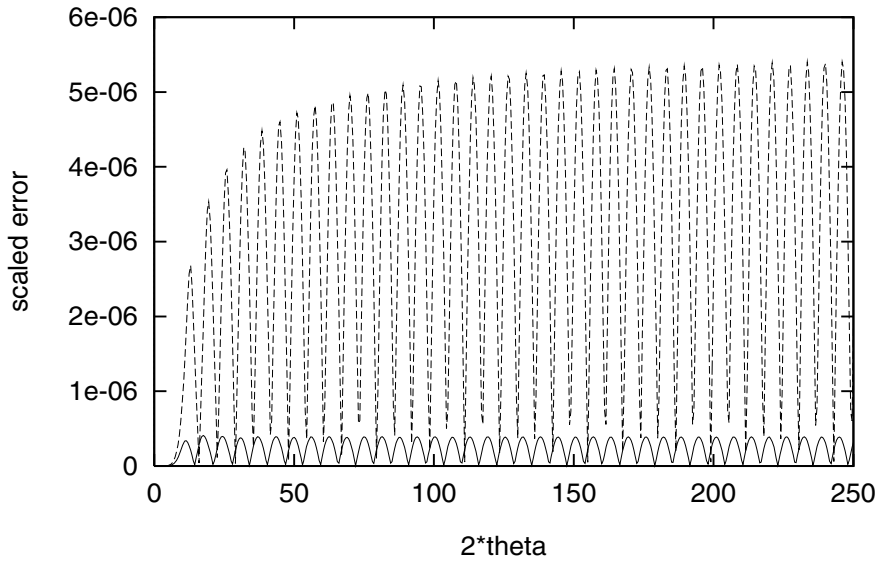


FIG. 8. *Exponential-fitting four-point rules: variation with 2θ of the scaled absolute errors $(2\theta)^2|I - I_{comput}|$ for ENC (solid) and $(2\theta)^3|I - I_{comput}|$ for G (broken).*

which is exact for 1, x, and $\exp(\pm i\omega x)$. The parameters of the required quadrature rule are

$$a_1^{(0)}(\theta) = a_2^{(0)}(\theta) = 1, \quad x_1^*(\theta) = -x_2^*(\theta) = -\arccos[\sin(\theta)/\theta]/\theta = -\arccos[\eta_0(-\theta^2)]/\theta$$

and, by considering the error when $f(x) = x^2$,

$$T_0^*(\theta, \mathbf{a}(\theta)) = T^*(\theta) = \frac{1/3 - [x_1^*(\theta)]^2}{\theta^2} = \frac{1/3 - \{\arccos[\eta_0(-\theta^2)]\}^2/\theta^2}{\theta^2}. \tag{5.2}$$

It is easy to check that, as expected, when $\theta \to 0$ this formula tends to the classical two-point Gauss–Legendre formula, i.e., $x_1^*(\theta)$, $x_2^*(\theta)$, and $T_0^*(\theta)$ tend to $-1/\sqrt{3}$, $1/\sqrt{3}$, and $1/135$, respectively. This quadrature formula was considered, with several others, in [18].

The functions which are to be integrated exactly satisfy the reference differential equation $D^2(D^2 + \omega^2)y = 0$, and the quadrature error is

$$err_{ef} = h \int_{-1}^{1} \Phi^*(x^*) D^{*2}(D^{*2} + \theta^2) F(x^*)\, dx^*, \tag{5.3}$$

where $F(x^*) = f(x)$.

The kernel function corresponding to the operator $D^2(D^2 + \omega^2)$ was derived in connection with Example 2 of section 3. In terms of the dimensionless variables introduced at the beginning of section 4.2,

$$K^*(t^*, x^*) = \frac{t^* - x^*}{\theta^2} - \frac{\sin\theta(t^* - x^*)}{\theta^3}.$$

Let $x_0^* = -1$ and $x_3^* = 1$. Then

$$\Phi^*(x^*) = \begin{cases} \phi_0^*(x^*) & \text{for } -1 < x^* < x_1^*, \\ \phi_1^*(x^*) & \text{for } x_1^* < x^* < x_2^*, \\ \phi_2^*(x^*) & \text{for } x_2^* < x^* < 1, \end{cases}$$

where

$$\begin{aligned}\phi_0^*(x^*) &= -\int_{-1}^{x^*} K^*(t^*, x^*)\, dt^* \\ &= \frac{(1+x^*)^2}{2\theta^2} - \frac{1}{\theta^4} + \frac{\cos\theta(1+x^*)}{\theta^4}, \\ \phi_1^*(x^*) &= \phi_0^*(x^*) + \frac{x_1^* - x^*}{\theta^2} - \frac{\sin\theta(x_1^* - x^*)}{\theta^3}, \\ \phi_2^*(x^*) &= \phi_1^*(x^*) + \frac{x_2^* - x^*}{\theta^2} - \frac{\sin\theta(x_2^* - x^*)}{\theta^3}.\end{aligned}$$

Using the fact that $\cos\theta x_1^* = \sin\theta/\theta$, it can be shown that ϕ_1^* is an even function and that $\phi_2^*(x^*) = \phi_0^*(-x^*)$. It follows that Φ^* is an even function.

It is easily seen that $\phi_0^*(x^*) > 0$ for $-1 < x^* \le x_1^*$; thus $\Phi^*(x^*) > 0$ for $-1 < x^* \le x_1^*$ and $x_2^* \le x^* < 1$. Numerical evidence indicates that $\Phi^*(x^*) > 0$ for all $x^* \in (-1, 1)$. If that is so, then

$$err_{ef} = hT^*(\theta)D^{*^2}(D^{*^2} + \theta^2)F(\eta) \tag{5.4}$$

for some unknown $\eta \in (-1, 1)$.

As in section 4, to investigate the asymptotic behavior of the quadrature error we assume that the integrand has the form (4.13). Then, for fixed x^*, independent of θ,

$$D^{*2}(D^{*2} + \theta^2)F(x^*) \sim 2\theta^3\,[f_1'(x^*)\sin(\theta x^*) - f_2'(x^*)\cos(\theta x^*)] \tag{5.5}$$

as $\theta \to \infty$. Equation (5.2) shows that $T^*(\theta) \sim \theta^{-2}$. Combining the two asymptotic estimates as in earlier sections we can conclude only that err_{ef} could increase like θ as $\theta \to \infty$. This estimate is not confirmed on the test integral (4.15); it is too pessimistic. Indeed, it is clear that $I(\omega) = O(1/\omega)$ as $\omega \to \infty$ and that the approximation given by the two-point formula (5.1), namely,

$$I_{comput} = \cos[(\omega + 1/2)x_1^*(\omega)] + \cos[(\omega + 1/2)x_2^*(\omega)] = 2\cos[(\omega + 1/2)x_1^*(\omega)],$$

has the same asymptotic behavior. Since the leading terms of the two expansions in powers of $1/\omega$ are different, it also follows that

$$I(\omega) - I_{comput} = O(1/\omega) \qquad \text{as} \quad \omega \to \infty.$$

In other words, the actual error decays as $1/\omega$ instead of increasing as ω, as predicted; notice that for this test case we have $h = 1$ and then $\theta = \omega$.

The reason for the discrepancy is not the forms (5.3) or (5.4) of the error but the way in which the latter was exploited. An alternative way consists of using the integral representation (5.3) directly.

In view of (5.5) and the fact that Φ^* is an even function, the asymptotic form of the quadrature error, for fixed h as $\theta \to \infty$, is

$$err_{ef} \sim 4h\theta^3 \int_{-1}^{0} \Phi^*(x^*)G(x^*)\, dx^*, \tag{5.6}$$

where

$$G(x^*) = f_{1o}'(x^*)\sin(\theta x^*) - f_{2e}'(x^*)\cos(\theta x^*),$$

f'_{1o} is the odd part of f'_1, and f'_{2e} is the even part of f'_2. This can be written as

$$4h\theta^3 \left[I_1(\theta) + I_2(\theta) \right],$$

where

$$I_1(\theta) = \int_{-1}^{0} \phi_0^*(x^*) G(x^*)\, dx^*$$

and

$$I_2(\theta) = \int_{x_1^*}^{0} [\phi_1^*(x^*) - \phi_0^*(x^*)] G(x^*)\, dx^*.$$

The integrand in I_1 may be written as a sum of three terms, corresponding to the three terms in ϕ_0^*. Because of the trigonometric factors in the function G, the Riemann–Lebesgue lemma shows that the contributions from the two terms containing θ^{-4} tend to zero faster than θ^{-4}, and we need only consider

$$\frac{1}{2\theta^2} \int_{-1}^{0} (1 + x^*)^2 \left[f'_{1o}(x^*) \sin(\theta x^*) - f'_{2e}(x^*) \cos(\theta x^*) \right] dx^*.$$

Integration by parts, noting that the odd function $f'_{1o}(x^*)$ vanishes when $x^* = 0$, allows us to write this as

$$\frac{1}{2\theta^3} \int_{-1}^{0} [g_1(x^*) \cos(\theta x^*) + g_2(x^*) \sin(\theta x^*)]\, dx^* = O(\theta^{-4});$$

the precise form of the functions g_1 and g_2 arising from the integration by parts is not of any interest.

In the integral

$$I_2(\theta) = \int_{x_1^*}^{0} \left[\frac{x_1^* - x^*}{\theta^2} - \frac{\sin \theta (x_1^* - x^*)}{\theta^3} \right] G(x^*)\, dx^*,$$

the magnitude of the second term is less than $|x_1^*|\, \theta^{-3} \|G\| = O(\theta^{-4})$ as $\theta \to \infty$. The first term in $I_2(\theta)$ is

$$\frac{1}{\theta^2} \int_{x_1^*}^{0} (x_1^* - x^*) \left[f'_{1o}(x^*) \sin(\theta x^*) - f'_{2e}(x^*) \cos(\theta x^*) \right] dx^*.$$

Integration by parts, again using the fact that $f'_{1o}(0) = 0$, reduces this to

$$\frac{1}{\theta^3} \int_{x_1^*}^{0} \left[k_1(x^*) \cos(\theta x^*) + k_2(x^*) \sin(\theta x^*) \right] dx^* = O(\theta^{-4}),$$

and the precise forms of the functions k_1 and k_2 are not important. This leads to the conclusion that for all integrands of the assumed form the error in the approximation provided by the two-point formula (5.1) tends to zero at least as fast as $1/\theta$, as $\theta \to \infty$ for fixed h, in full agreement with our observations for the test integral.

A by-product of this analysis is that the parameter η of (5.4) depends on θ in such a way that $D^{*^2}(D^{*^2} + \theta^2)F(\eta)$ increases no faster than the first power of θ as $\theta \to \infty$.

6. Conclusion. The integral representation of the truncation error, based on the work of Ghizzetti and Ossicini [8], is applicable to a wide variety of exponential-fitting methods for oscillatory problems. The detailed investigations of sections 4 and 5 are concerned with quadrature methods but the truncation errors in exponential-fitting methods for ordinary differential equations and in numerical differentiation formulae could be analyzed in the same way.

Section 4.5 shows that, for the particular example considered there, two three-point exponential-fitting formulae have the same qualitative behavior as a three-point Filon–Lobatto method advocated for such problems [9], but their errors are less by almost an order of magnitude than that of the Filon–Lobatto method. Furthermore, the errors in exponential-fitting Gaussian formulae with larger numbers of abscissas have faster decay rates, whereas the analysis of Iserles [9] shows that the decay rate of the error in Filon–Lobatto methods is not improved by increasing the number of nodes. This suggests that exponential-fitting quadrature methods, and particularly those of Gaussian type, are worthy of further investigation as practical methods for the numerical integration of oscillatory integrands.

Acknowledgment. We are grateful to the referees for their helpful comments, and particularly for an improved argument used just before (3.4).

REFERENCES

[1] A. Chakrabarti and Hamsapriye, *Derivation of a general mixed interpolation formula*, J. Comput. Appl. Math., 70 (1996), pp. 161–172.
[2] J. P. Coleman, *Mixed interpolation methods with arbitrary nodes*, J. Comput. Appl. Math., 92 (1998), pp. 69–83.
[3] P. J. Davis, *Interpolation and Approximation*, Blaisdell, Waltham, MA, 1963.
[4] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, London, 1984.
[5] H. de Meyer, J. Vanthournout, G. Vanden Berghe, and A. Vanderbauwhede, *On the error estimation for a mixed type of interpolation*, J. Comput. Appl. Math., 32 (1990), pp. 407–415.
[6] U. T. Ehrenmark, *A three-point formula for numerical quadrature of oscillatory integrals with variable frequency*, J. Comput. Appl. Math., 21 (1988), pp. 87–99.
[7] U. T. Ehrenmark, *On the error and its control in a two-parameter generalised Newton-Cotes rule*, J. Comput. Appl. Math., 75 (1996), pp. 171–195.
[8] A. Ghizzetti and A. Ossicini, *Quadrature Formulae*, Birkhaüser, Basel, Switzerland, 1970.
[9] A. Iserles, *On the numerical quadrature of highly-oscillating integrals.* I. *Fourier transforms*, IMA J. Numer. Anal., 24 (2004), pp. 365–391.
[10] L. Gr. Ixaru, *Numerical Methods for Differential Equations and Applications*, Reidel, Dordrecht, The Netherlands, 1984.
[11] L. Gr. Ixaru, *Operations on oscillatory functions*, Comput. Phys. Comm., 105 (1997), pp. 1–19.
[12] L. Gr. Ixaru, *Numerical operations on oscillatory functions*, Comput. Chem., 25 (2001), pp. 39–53.
[13] L. Gr. Ixaru and B. Paternoster, *A Gauss quadrature rule for oscillating integrands*, Comput. Phys. Comm., 133 (2001), pp. 177–188.
[14] L. Gr. Ixaru and G. Vanden Berghe, *Exponential Fitting*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
[15] J. K. Kim, R. Cools, and L. Gr. Ixaru, *Quadrature rules using first derivatives for oscillatory integrands*, J. Comput. Appl. Math., 140 (2002), pp. 479–497.
[16] W. E. Milne, *The remainder in linear methods of approximation*, J. Research Nat. Bur. Standards, 43 (1949), pp. 501–511.

[17] L. R. Petzold, L. O. Jay, and J. Yen, *Numerical solution of highly oscillatory ordinary differential equations*, in Acta Numerica 1997, Acta Numer. 6, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1997.

[18] M. Van Daele, G. Vanden Berghe, and H. Vande Vyver, *Exponentially fitted quadrature rules of Gauss type for oscillatory integrands*, Appl. Numer. Math., 53 (2005), pp. 509–526.

[19] G. Vanden Berghe, H. de Meyer, and J. Vanthournout, *On a class of modified Newton-Cotes quadrature formulae based upon mixed type interpolation*, J. Comput. Appl. Math., 3 (1990), pp. 331–349.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1466–1486

A PARALLEL METHOD FOR BACKWARD PARABOLIC PROBLEMS BASED ON THE LAPLACE TRANSFORMATION*

JINWOO LEE† AND DONGWOO SHEEN†

Abstract. A parallel method for time discretization of backward parabolic problems is proposed. The problem is reformulated to a set of Helmholtz-type problems with a parameter on a suitably chosen contour in the complex plane. After solving the resulting elliptic equations, which can be solved in parallel, we obtain a regularized solution with high frequency terms cut off by the inverse Laplace transforms without requiring the knowledge of the eigenfunctions of the differential operator. Since the regularized solution is obtained without artificial perturbation and high frequency components of the noise are suppressed, the quality of the solution is improved significantly compared to those obtained by other methods. Two different numerical inversions of Laplace transforms, with an arbitrary high order of accuracy and spectral accuracy, respectively, are used. Error estimates and numerical examples are presented.

Key words. numerical method, backward parabolic, ill-posed problem, Laplace transform, quadrature, parallel method

AMS subject classifications. 65N30, 35R25

DOI. 10.1137/050624649

1. Introduction.

We consider the following backward parabolic problem: given $u_0 \in L^2(\Omega)$, find $u = u(t) = u(\cdot, t) \in H_0^1(\Omega)$ such that

$$u_t + Au = 0 \quad \text{for } t \in (0, T], \quad \text{with } u(\cdot, 0) = u_0(\cdot) \text{ in } \Omega, \tag{1.1}$$

where $-A$ is a uniformly elliptic second-order partial differential operator on a domain Ω with a homogeneous Dirichlet boundary condition. Furthermore, we assume that A is a closed operator in the Hilbert space $L^2(\Omega)$ which generates an analytic semigroup $E(t) = e^{tA}$ and the spectrum $\sigma(-A)$ of $-A$ is contained in a sector $\{z \in \mathbb{C} : |\arg z| < \zeta\}$ for some $\zeta \in (0, \pi/2)$. We also assume that the resolvent $(zI + A)^{-1}$ of $-A$ satisfies

$$\|(zI + A)^{-1}\| \le \frac{C}{1 + |z|} \quad \text{for } z \in \Sigma_\zeta,$$

where the complementary sector Σ_ζ is given by

$$\Sigma_\zeta = \{z \in \mathbb{C} : \zeta < |\arg z| \le \pi\} \cup \{O\}.$$

Problem (1.1) is a well-known ill-posed problem in the sense that the solution does not depend continuously on the data u_0 [15, 19, 29]. However, it can be formulated as a well-posed problem, for instance, by imposing a prescribed bound on the solution at $t = T$ [19]. More precisely, given data $g \in L^2(\Omega)$ with noise, let $u^{(j)}$, $j = 1, 2$, be any two solutions of (1.1) satisfying

$$\|u^{(j)}(T)\| \le M \qquad \text{and} \qquad \|u_0^{(j)} - g\| \le \delta, \tag{1.2}$$

*Received by the editors February 17, 2005; accepted for publication (in revised form) February 16, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sinum/44-4/62464.html

†Department of Mathematics, Seoul National University, Seoul 151–747, Korea (jlee@nasc.snu.ac.kr, sheen@snu.ac.kr). The work of the second author was supported in part by Korea Research Foundation grant (KRF-2004-C00007).

where M and δ are given positive constants and $\|\cdot\|$ denotes the $L^2(\Omega)$-norm. Then it is known [29] that

$$\|u^{(1)}(t) - u^{(2)}(t)\| \le 2M^{t/T}\delta^{1-t/T} \quad \text{for } t \in (0, T], \tag{1.3}$$

which follows directly from the convexity of $F(t) = \log \|u^{(1)}(t) - u^{(2)}(t)\|^2$. We thus have continuous dependence on the data, and any numerical solution of the problem can be regarded as a kind of *regularized* solution depending on the two constraints M and δ in (1.2).

Stable numerical methods for backward parabolic problems can be applied to several practical areas such as image processing, mathematical finance, and physics. However, the ill-posedness nature of the problems requires certain types of regularization techniques. One approach to regularize the ill-posed problems is based on the use of eigenfunction expansion [17, 28, 23, 25, 30], where the eigenpairs of the corresponding elliptic operator are available. Another approach is to use the method of quasi reversibility [4, 7, 8, 12, 13, 14, 21, 27]. Other approaches include the least squares methods with Tikhonov-type regularization [2, 16, 18, 24, 26] and the use of heat kernel [9]. Buzbee and Carasso [4] introduced a method of transforming the problem (1.1) into a second-order in time problem. Later Carasso [5, 6] introduced the concept of a supplementary constraint such as that of slow evolution from the continuation boundary (SECB). The methods of quasi reversibility and Tikhonov regularization introduce artificial contamination, which is not from noises, to the numerical solutions. We thus expect to improve the solution quality if we can avoid any artificial perturbation and effectively suppress the influence of high frequency noises, which is the purpose of this paper. We also indicate that parallel algorithms have not yet been seriously addressed; however, see [24].

In this paper, we develop a parallel numerical method without any perturbation to obtain a regularized solution to problem (1.1). Instead of attacking problem (1.1) in the original space-time domain setting, we take the Laplace transform in time to have a set of complex-valued, Helmholtz-type problems with a parameter on a suitably chosen contour Γ in a control domain. After solving the resulting elliptic problems, the regularized time-domain solution with high frequency terms cut off can be recovered by applying the inverse Laplace transformation numerically. Two different choices of contour Γ will be described in detail in the next section. The first contour introduced in [32] requires no information on eigenpairs of the operator A, while the second one proposed in [22], which is more efficient, requires information on eigenvalues only. Since we obtain solutions without modifying the original problem and high frequency terms of noise are cut-off automatically, solution quality is improved significantly, especially as time goes on to the final time T (see the end of section 2.3).

The outline of the rest of the paper is as follows. In the next section, two numerical schemes are introduced based on the Laplace transformation of (1.1). In section 3, basic stability and error estimates are derived for these numerical schemes. Some numerical results are given in section 4.

2. The numerical schemes. Let us reformulate problem (1.1) by formally performing the Laplace transformation in time. Then

$$z\widehat{u} + A\widehat{u} = u_0 \text{ in } \Omega \quad \text{for } z \in \rho(-A) \tag{2.1}$$

with the homogeneous Dirichlet boundary condition on $\partial\Omega$, where $\rho(-A)$ is the resolvent set of $-A$. First, denote by $\{\phi_k\}_{k=1}^{\infty}$ and $\{\lambda_k\}_{k=1}^{\infty}$ the orthonormal eigenfunctions

of $-A$ and the corresponding eigenvalues which satisfy $0 < \operatorname{Re}\lambda_1 \le \operatorname{Re}\lambda_2 \le \cdots \to +\infty$. The solution $\widehat{u}(z)$ of (2.1) is then given in the following form:

$$\widehat{u}(z) = (zI + A)^{-1}u_0 = \sum_{k=1}^{\infty} \frac{1}{z-\lambda_k}(u_0, \phi_k)\phi_k, \tag{2.2}$$

since the solution of (1.1), if any, admits the representation

$$u(t) = \sum_{k=1}^{\infty} e^{\lambda_k t}(u_0, \phi_k)\phi_k. \tag{2.3}$$

In (2.3) we can observe that small errors on u_0 grow without bound, which is the source of ill-posedness. Among several regularization methods [17, 23] to avoid such an error growth, we shall employ a method of cutting off high frequency terms in the current paper.

From now on, denote by $\widehat{u}^{u_0}$ and $\widehat{u}^{g}$ the solutions to (2.1) with data u_0 and g, respectively. We begin by defining our regularization $u^{\Gamma,u_0}(t)$, using the Laplace inversion formula, by

$$u^{\Gamma,u_0}(t) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma} e^{zt}\widehat{u}^{u_0}(z)\, dz, \tag{2.4}$$

where, for some positive integer N,

$$\Gamma = \{z \in \mathbb{C} | \operatorname{Re}\lambda_N < \operatorname{Re} z < \operatorname{Re}\lambda_{N+1}\} \subset \rho(-A), \tag{2.5}$$

and the direction of Γ is taken such that $\operatorname{Im}(z)$ is increasing from $-\infty$ to $+\infty$. The contour Γ will be deformed subsequently for the sake of computational efficiency, and we will see that λ_i's need not be known explicitly for $\Gamma = \Gamma_1$ (see section 2.1 for the definition of Γ_1).

Notice that

$\displaystyle\sum_{k=N+1}^{\infty} \frac{1}{z-\lambda_k}(u_0, \phi_k)\phi_k$ is analytic in the half plane left of the straight line contour Γ,

which, incorporated into (2.2) and (2.4), implies that

$$\begin{aligned} u^{\Gamma,u_0}(t) &= \frac{1}{2\pi \mathrm{i}} \sum_{k=1}^{\infty} \int_{\Gamma} \frac{e^{zt}}{z-\lambda_k}(u_0, \phi_k)\phi_k\, dz \\ &= \frac{1}{2\pi \mathrm{i}} \sum_{k=1}^{N} \int_{\Gamma} \frac{e^{zt}}{z-\lambda_k}(u_0, \phi_k)\phi_k\, dz = \sum_{k=1}^{N} e^{\lambda_k t}(u_0, \phi_k)\phi_k, \end{aligned} \tag{2.6}$$

which is a spectral representation of $u^{\Gamma,u_0}(t)$ with high frequency terms cut off. The resulting $u^{\Gamma,g}(t)$ with a given data g satisfying (1.2) instead of u_0 fulfills the following stability condition:

$$\|u(t) - u^{\Gamma,g}(t)\| \le 2M^{t/T}\delta^{1-t/T} \quad \text{for } t \in (0, T] \tag{2.7}$$

if we choose N in (2.5) to be the largest integer such that

$$\operatorname{Re}\lambda_N \le \frac{1}{T}\log\frac{M}{\delta}$$

holds. See section 3 for more details.

2.1. The numerical procedure using a hyperbolic contour. We now deform the straight line contour Γ in (2.4) into the left-hand branch of a hyperbola, as in [32], with the asymptotes having slopes $\pm\kappa$, which crosses the real axis at $\gamma - \nu$, by setting

$$\Gamma_1 = \{z : z = z(\omega) = \sigma(\omega) + \mathrm{i}\kappa\omega,\ -\infty < \omega < \infty\} \subset \rho(-A), \qquad \sigma(\omega) = \gamma - \sqrt{\omega^2 + \nu^2}, \tag{2.8}$$

for suitable parameters $\nu > 0$ (usually $\nu = 0.5$) and $\kappa > 0$, where γ will be chosen such that $\gamma - \nu = (1/T)\log(M/\delta)$, provided $\gamma - \nu$ does not coincide with any $\operatorname{Re}\lambda_i$. Then we have $\operatorname{Re}\lambda_N < \gamma - \nu < \operatorname{Re}\lambda_{N+1}$ for N used in Theorem 3.1. The parameters of Γ_1 are chosen such that the eigenvalues $\lambda_1, \ldots, \lambda_N$ are to the left of the contour Γ_1, while the rest of eigenvalues are to the right of Γ_1. At the end of section 3.1 we will remark on the flexible choice of Γ_1. With the deformed contour Γ_1, the integral (2.4) can be written as

$$u^{\Gamma_1,u_0}(t) = \frac{1}{2\pi\mathrm{i}} \int_{-\infty}^{\infty} e^{z(\omega)t} \widehat{u}^{u_0}(z(\omega)) z'(\omega)\, d\omega. \tag{2.9}$$

We notice that, as $\omega \to \pm\infty$, $e^{z(\omega)t}$ goes to zero rapidly, which accelerates convergence. Next, we transform the infinite interval $(-\infty, \infty)$ for ω in (2.9) into a finite one. For this, let $\psi : (-\infty, \infty) \to (-1, 1)$ be defined by $\psi(\omega) = \tanh(\frac{\tau\omega}{2})$ as in [32]. In particular, by applying the change of variables $y = \psi(\omega)$ or, equivalently, $\omega = \psi^{-1}(y) = \frac{1}{\tau}\log\frac{1+y}{1-y}$ with the parameter $\tau > 0$ to be determined later, the integral (2.9) can be transformed into one on $(-1, 1)$ for y, where the trapezoidal rule can be applied to get

$$U_{L_1,\tau}^{\Gamma_1,u_0}(t) = \frac{1}{2\pi\mathrm{i}} \frac{1}{L_1} \sum_{j=-L_1+1}^{j=L_1-1} e^{z_j t} \widehat{u}^{u_0}(z_j) \frac{dz}{d\omega}(\omega_j) \frac{d\psi^{-1}}{dy}(y_j), \tag{2.10}$$

where

$$z_j = z(\omega_j), \quad \omega_j = \psi^{-1}(y_j) = \frac{1}{\tau}\log\frac{1+y_j}{1-y_j}, \quad \text{and } y_j = \frac{j}{L_1},\ -L_1 < j < L_1. \tag{2.11}$$

The change of variables here spreads the equidistant points y_j's in $(-1, 1)$ over $\mathbb{R}$ to ω_j's such that we have a finer grid near the origin where the integrand is relatively larger and a coarser grid where the integrand becomes relatively smaller. In [32], it is shown that this quadrature scheme has an arbitrary high order of accuracy for forward parabolic problems. In this paper we prove similarly that the same accuracy holds, although the properties of integrands are slightly different. The basic error estimate essentially shows that, for any positive integer r,

$$\|u^{\Gamma,u_0}(t) - U_{L_1,\tau}^{\Gamma_1,u_0}(t)\| \le \frac{C}{L_1^r} \left(\frac{\beta M}{\delta}\right)^{t/T} \|u_0\| \quad \text{for } t > r\tau, \tag{2.12}$$

where C depends on various parameters and $\beta > 1$ is a constant.

2.2. The numerical procedure using a union of small circles. When the eigenvalues of the operator $-A$ are known, we can deform the contour in (2.4) into a union of disjoint small circles around the eigenvalues of $-A$ [22], which enables us

to reduce computational costs significantly when the number of dominant eigenvalues is relatively small. Set $d = \min_{1\le k\le N} \min_{l\neq k} |\lambda_k - \lambda_l|$ and assume that $d > 0$. For a sufficiently small $\varepsilon \in (0, d)$ (such that the circle $|z - \lambda_k| = \varepsilon$ contains no other eigenvalues than λ_k for all k), we define the second contour Γ_2 by

$$\Gamma_2 = \bigcup_{k=1}^{N} C_k \subset \rho(-A), \quad C_k = \{z : z = \lambda_k + \varepsilon e^{\mathrm{i}\theta}, 0 \le \theta \le 2\pi\}, \tag{2.13}$$

where N is again the largest integer such that $\operatorname{Re}\lambda_N \le (1/T)\log(M/\delta)$ holds. With this Γ_2, (2.4) can be written as

$$u^{\Gamma_2,u_0}(t) = \frac{1}{2\pi i}\sum_{k=1}^{N}\int_{C_k} e^{zt}\widehat{u}^{u_0}(z)\,dz. \tag{2.14}$$

After applying the change of variables $z = \lambda_k + \varepsilon e^{\mathrm{i}\theta}$ on each circle C_k and the trapezoidal rule, we have

$$U_{L_2}^{\Gamma_2,u_0}(t) = \frac{1}{2\pi}\sum_{k=1}^{N}\frac{2\pi}{L_2}\sum_{j=0}^{L_2-1} e^{z_{k,j}t}\widehat{u}^{u_0}(z_{k,j})\varepsilon e^{\mathrm{i}\theta_j}, \tag{2.15}$$

where $z_{k,j} = \lambda_k + \varepsilon e^{\mathrm{i}\theta_j}$ and $\theta_j = \frac{2\pi}{L_2}j$ for $0 \le j \le L_2 - 1$. The efficiency of this scheme originates in that the interval of the integral could be arbitrarily small by letting ε be small (it should not be too small because of the machine precision; see section 3.2). The error estimate of this scheme results in spectral accuracy of the form

$$\|U_{L_2}^{\Gamma_2,u_0}(t) - u^{\Gamma,u_0}(t)\| = C\left(\frac{\beta' M}{\delta}\right)^{t/T}\left(\varepsilon'^{L_2} + \frac{eps}{\varepsilon}\right)\|u_0\|, \tag{2.16}$$

where C is independent of ε and L_2, *eps* is the machine precision, and $\varepsilon' = \varepsilon/d$, for $\beta' = e^{\varepsilon T} > 1$. With a double precision calculation, *eps* can be as small as about 10^{-16}.

2.3. The fully discrete schemes. The fully discrete scheme is achieved by combining the time-discretization procedure, either (2.10) or (2.15), with the finite element method for spatial approximation procedure. For this, let $(V_h)_{h>0}$ denote a family of standard piecewise linear finite element subspaces of $H_0^1(\Omega)$. Let $a(\cdot,\cdot)$ be the natural sesquilinear form associated with A. Then the finite element approximation $\widehat{u}_h^{u_0}(z) \in V_h$ to the solution $\widehat{u}^{u_0}(z)$ of (2.1) satisfies

$$z(\widehat{u}_h^{u_0}(z), v) + a(\widehat{u}_h^{u_0}(z), v) = (u_0, v) \quad \forall v \in V_h, \forall z \in \rho(-A). \tag{2.17}$$

Thus the spatially discretized approximation U_h^{Γ,u_0} to u^{Γ,u_0} in (2.4) is given by

$$U_h^{\Gamma,u_0}(t) = \frac{1}{2\pi\mathrm{i}}\int_{\Gamma} e^{zt}\widehat{u}_h^{u_0}(z)\,dz. \tag{2.18}$$

By combining the time-discretization procedure, either (2.10) or (2.15), with the finite element method the fully discretized solution is given by either

$$U_{L_1,h,\tau}^{\Gamma_1,u_0}(t) = \frac{1}{2\pi\mathrm{i}}\frac{1}{L_1}\sum_{j=-L_1+1}^{j=L_1-1} e^{z_j t}\widehat{u}_h^{u_0}(z_j)\frac{dz}{d\omega}(\omega_j)\frac{d\psi^{-1}}{dy}(y_j) \tag{2.19}$$

or

$$U_{L_2,h}^{\Gamma_2,u_0}(t) = \frac{1}{L_2}\sum_{k=1}^{N}\sum_{j=0}^{L_2-1} e^{z_{k,j}t}\widehat{u}_h^{u_0}(z_{k,j})\varepsilon e^{\mathrm{i}\theta_j}, \tag{2.20}$$

respectively. If inexact data g is given, g will be used instead of u_0, which will be analyzed in section 3.

To conclude this section, let us indicate two special features of the proposed methods. First, the methods can be implemented in parallel in the main part, solving a set of elliptic equations (2.1), of the procedures without any essential data communication among processors since they are independent of each other. This idea goes back to [11, 10] for solving wave propagation problems in the space-frequency domain setting, and it was later applied to forward parabolic problems [31, 32] and to a forward integro-differential equation with positive memory [20]. Second, the methods proposed in the current paper do not introduce any perturbation from the original problem, unlike others [2, 4, 7, 8, 12, 13, 14, 16, 18, 24, 26, 30], and therefore the solution quality of our method is much better than that of others. The trick is in *formally* reformulating the problem into problems in the Laplace transformed setting as in (2.1). The reformulated equation (2.1) is well defined and well posed, although one cannot perform the Laplace transform of the solution (2.3) since it does not exist for any $z \in \mathbb{C}$. Thus we choose to start from this equation. When it is inverted by using the Laplace inversion formula (2.4), the contour Γ has to be selected, and it gives a natural way of controlling the frequency terms. Noticing that eigenvalues of $-A$ correspond to the poles of $\widehat{u}(z)$ (see (2.2)) we see that the poles to the left of Γ are taken and those to the right of Γ are discarded (see (2.6)). In this way the regularization is performed naturally without perturbing anything, and the high frequency components of noise whose eigenvalues are bigger than λ_N cause no influence on the numerical solutions. This feature improves the quality of the solutions remarkably, which will be illustrated in section 4.

3. Stability and error estimates. In this section we analyze the stability of and error estimates of the two numerical procedures introduced in the previous section. Before going into the details of both properties, we first state and prove an error estimate between the exact solution u of (1.1) with constraints (1.2) and the regularized solution $u^{\Gamma,g}$ in (2.4) with given data g satisfying (1.2).

THEOREM 3.1. *Let $g \in L^2(\Omega)$ be given. Suppose that u is an exact solution of* (1.1) *with constraints* (1.2). *If N in* (2.5) *is chosen to be the largest integer such that* $\operatorname{Re}\lambda_N \le (1/T)\log(M/\delta) < \operatorname{Re}\lambda_{N+1}$, *then the following error bound holds:*

$$\|u(t) - u^{\Gamma,g}(t)\| \le 2M^{t/T}\delta^{1-t/T} \text{ for } t \in (0,T]. \tag{3.1}$$

Proof. Let us consider the regularized solution with exact data u_0, say $u^{\Gamma,u_0}(t)$, and N be the largest integer such that $\operatorname{Re}\lambda_N \le (1/T)\log(M/\delta) \operatorname{Re}\lambda_{N+1}$. From (1.2) and (2.3), it follows that $\sum_{k=1}^{\infty} c_k^2 e^{2\operatorname{Re}\lambda_k T} \le M^2$ with $c_k = (u_0, \phi_k)$, and therefore

$$\begin{aligned}\|u(t) - u^{\Gamma,u_0}(t)\|^2 &= \sum_{k=N+1}^{\infty} c_k^2 e^{2\operatorname{Re}\lambda_k t} = \sum_{k=N+1}^{\infty} c_k^2 e^{2\operatorname{Re}\lambda_k t} e^{2\operatorname{Re}\lambda_k T} e^{-2\operatorname{Re}\lambda_k T}\\ &\le e^{2(t-T)\operatorname{Re}\lambda_{N+1}} \sum_{k=N+1}^{\infty} c_k^2 e^{2\operatorname{Re}\lambda_k T}\\ &\le (\delta/M)^{2(1-t/T)} M^2 = \delta^{2(1-t/T)} M^{2t/T}. \end{aligned} \tag{3.2}$$

Similarly, the difference between the two regularizations with initial data u_0 and g is estimated as follows: with $g_k = (g, \phi_k)$,

$$\|u^{\Gamma,u_0}(t) - u^{\Gamma,g}(t)\|^2 = \sum_{k=1}^{N}(c_k - g_k)^2 e^{2\,\mathrm{Re}\,\lambda_k t} \le e^{2\,\mathrm{Re}\,\lambda_N t}\sum_{k=1}^{N}(c_k - g_k)^2$$
$$\le (M/\delta)^{2t/T}\delta^2 = M^{2t/T}\delta^{2(1-t/T)} \tag{3.3}$$

since we have $\sum_{k=1}^{\infty}(c_k - g_k)^2 \le \delta^2$ by (1.2). The assertion is then obtained by the triangle inequality. ∎

3.1. Analysis of the numerical procedure using a hyperbolic contour. As we mentioned in section 2.1, we put $\gamma - \nu = (1/T)\log(M/\delta)$, provided it does not coincide with any λ_i. Let

$$\Sigma_{\zeta_1,\gamma-\nu} = \{z \in \mathbb{C} : \zeta_1 < |\arg(z - \gamma + \nu)| < \pi - \zeta_1\} \cup \mathcal{N}_{\gamma-\nu} \subset \rho(-A), \tag{3.4}$$

where $\mathcal{N}_{\gamma-\nu}$ is a neighborhood of $\gamma - \nu$ that does not contain any eigenvalue of $-A$, and $\zeta_1 \in (0, \pi/2)$ is chosen such that $\Gamma_1 \subset \Sigma_{\zeta_1,\gamma-\nu}$. By (2.2), we can find a constant $B_1 > 0$, independent of z, such that

$$\|(zI + A)^{-1}\| \le \frac{B_1}{1 + |z - \gamma + \nu|} \quad \text{for } z \in \Sigma_{\zeta_1,\gamma-\nu}. \tag{3.5}$$

We notice that $B_1 = O(\eta^{-1})$, where η is a distance between the two sets $\Sigma_{\zeta_1,\gamma-\nu}$ and $\sigma(-A)$, the spectrum of $-A$. From now on set $\beta = e^{\nu T}$. We then have the following stability estimate.

THEOREM 3.2. *Let $U_{L_1,\tau}^{\Gamma_1,g}$ be the approximation defined by* (2.10) *of the regularized solution $u^{\Gamma_1,g}$ given by* (2.9)*, with u_0 replaced by g. Then, for $t > \tau$,*

$$\|U_{L_1,\tau}^{\Gamma_1,g}\| \le C(\beta M/\delta)^{t/T}\|g\|,$$

where $C = \sqrt{2}(2 - \frac{1}{L_1})\sqrt{1+\kappa^2}B_1/(\tau\pi)$.

Proof. From (2.8), (2.10), and (2.11) it follows that

$$\|U_{L_1,\tau}^{\Gamma_1,g}(t)\| \le \frac{1}{2\pi}\frac{1}{L_1}\sum_{j=-L_1+1}^{L_1-1}|e^{z_j t}|\|\widehat{u}^g(z_j)\|\left|\frac{dz}{d\omega}(\omega_j)\frac{d\psi^{-1}}{dy}(y_j)\right|$$
$$\le \frac{\sqrt{1+\kappa^2}e^{\gamma t}}{\sqrt{2}\pi L_1}\sum_{j=-L_1+1}^{L_1-1}e^{-|\omega_j|t}\left|\frac{d\psi^{-1}}{dy}(y_j)\right|\|\widehat{u}^g(z_j)\|$$
$$\le \frac{\sqrt{1+\kappa^2}e^{\gamma t}}{\sqrt{2}\pi L_1}\sqrt{\sum_{j=-L_1+1}^{L_1-1}\left(e^{-|\omega_j|t}\frac{d\psi^{-1}}{dy}(y_j)\right)^2}\sqrt{\sum_{j=-L_1+1}^{L_1-1}\|\widehat{u}^g(z_j)\|^2} \tag{3.6}$$

since $\left|\frac{dz}{d\omega}(\omega_j)\right| \le \sqrt{2}$. Now, for $j \ge 0$,

$$e^{-|\omega_j|t}\frac{d\psi^{-1}}{dy}(y_j) = e^{-\frac{t}{\tau}\log\frac{1+y_j}{1-y_j}}\frac{2}{\tau}\frac{1}{1-y_j^2} = \frac{2}{\tau}\frac{(1-y_j)^{t/\tau-1}}{(1+y_j)^{t/\tau+1}} \le \frac{2}{\tau} \quad \text{if } t > \tau. \tag{3.7}$$

The same bound holds for $j < 0$. Since $\gamma = \log(M/\delta)/T + \nu$, we have

$$e^{\gamma t} = (\beta M/\delta)^{t/T}, \tag{3.8}$$

where we recall that $\beta = e^{\nu T}$. Next, by using (2.8) and (3.5), we have, for $z \in \Sigma_{\zeta_1,\gamma-\nu}$,

$$\begin{aligned}\|\widehat{u}^g(z(\omega))\| = \|(z(\omega)I + A)^{-1}g\| &\le \frac{B_1}{1+|z(\omega)-\gamma+\nu|}\|g\| \\ &\le \frac{B_1}{1+\kappa|\omega|}\|g\| \le B_1\|g\|, \end{aligned} \tag{3.9}$$

which, combined with (3.6), completes the proof. □

Our error analysis is based on an Euler–Maclaurin-type proposition [32].

PROPOSITION 3.3 (Sheen, Sloan, and Thomeé [32]). *Let $r \ge 1$ be given and assume that $v \in C^r(\mathbb{R}; L^2(\Omega))$ and*

$$\|v^{(j)}(\omega)\| = O(e^{-r\tau|\omega|}) \quad \text{for } j \le r \text{ as } |\omega| \to \infty.$$

Furthermore, if $\|v(\omega)\| = o(e^{-\tau|\omega|})$ for $r = 1$, then we have the error estimate

$$\|Q_{L_1,\tau}(v) - I(v)\| \le C_r \frac{1}{L_1^r}\left(1 + \frac{1}{\tau^r}\right)\int_{-\infty}^{\infty} e^{r\tau|\omega|}\sum_{j=0}^{r}\|v^{(j)}(\omega)\|\,d\omega,$$

where

$$I(v) := \int_{-\infty}^{\infty} v(\omega)\,d\omega \quad \text{and} \quad Q_{L_1,\tau}(v) := \frac{1}{L_1}\sum_{j=-L_1+1}^{L_1-1} v(\omega_j)\frac{d\psi^{-1}}{dy}(y_j). \tag{3.10}$$

The proof of Proposition 3.3 is given in [32].

Proposition 3.3 implies that the formula (3.10) is of an arbitrary high order of accuracy, provided $v(\omega)$ vanishes appropriately fast at infinity. Based on Proposition 3.3, we derive an error estimate between the regularized solution $u^{\Gamma_1,g}$ given in (2.4) and its time-discretized approximation $U^{\Gamma_1,g}_{L_1,\tau}(t)$ given in (2.10) using g instead of u_0.

LEMMA 3.4. *Let $u^{\Gamma_1,g}(t)$ and $U^{\Gamma_1,g}_{L_1,\tau}(t)$ be the regularized solution defined by* (2.4) *and its approximation defined by* (2.10)*, respectively, with initial data g instead of u_0 and r a positive integer. Then, for $t > r\tau$, we have*

$$\|u^{\Gamma_1,g}(t) - U^{\Gamma_1,g}_{L_1,\tau}(t)\| \le \frac{C_{r,t}}{L_1^r}\left(\frac{\beta M}{\delta}\right)^{t/T}\|g\|, \tag{3.11}$$

where $C_{r,t} = C_r\frac{(1+\kappa^2)^{r/2}}{k}(1+t^r)(1+\frac{1}{\tau^r})(1+\log_+\frac{\kappa}{t-r\tau})$ and $\log_+ x = \max(0,\log x)$.

Proof. Set $v(\omega,t) = \frac{1}{2\pi i}e^{z(\omega)t}\widehat{u}^g(z(\omega))z'(\omega)$. Then, from (2.9) and (2.10), with u_0 replaced by g, it follows that

$$u^{\Gamma_1,g}(t) - U^{\Gamma_1,g}_{L_1,\tau}(t) = I(v(\cdot,t)) - Q_{L_1,\tau}(v(\cdot,t)).$$

Our aim is to apply Proposition 3.3, and for this we need to bound the derivatives of the function $\widehat{u}^g(z)$ on Γ_1. Since $\frac{d^j}{dz^j}(zI+A)^{-1} = (-1)^j j!(zI+A)^{-j-1}$, (3.9) and an induction on j imply that

$$\left\|\frac{d^j}{dz^j}\widehat{u}^g(z)\right\| \le \frac{C_j}{1+\kappa|\omega|}\|g\| \quad \text{for } z \in \Gamma_1. \tag{3.12}$$

The Leibniz rule is then applied to obtain

$$\left\| \frac{\partial^j}{\partial \omega^j} v(t,\omega) \right\| \le C_r \left(1+t^r\right) \frac{(1+\kappa^2)^{r/2} e^{t\sigma(\omega)}}{1+\kappa|\omega|} \|g\| \quad \text{for } j \le r, \quad \omega \in \mathbb{R},$$

where $C_r > 0$ is a constant depending on $\|\frac{d^j}{d\omega^j}\sigma(\omega)\|_{L_\infty(\mathbb{R})}$ for $j \le r$. Since $\sigma(\omega) \approx -|\omega|$ for large $|\omega|$ the assumptions of Proposition 3.3 are thus satisfied if $t > r\tau$, and the proposition implies that

$$\|u^{\Gamma_1,g}(t) - U^{\Gamma_1,g}_{L_1,\tau}(t)\| \le C_r L_1^{-r} \left(1+\tau^{-r}\right)\left(1+t^r\right) e^{\gamma t} \int_{-\infty}^{\infty} \frac{(1+\kappa^2)^{r/2} e^{-|\omega|(t-r\tau)}}{1+\kappa|\omega|}\, d\omega\, \|g\|,$$

since we have $\sigma(\omega) \le \gamma - |\omega|$. To bound the integral that remains, notice that $\int_0^\infty e^{-\omega t}(1+\omega)^{-1}\, d\omega \le C(1+\log_+(1/t))$, which can be verified easily by arithmetic calculations. Then (3.8) is used to complete the proof of the lemma. ∎

Remark 3.5. The parameter r appears only in the theorem, not in the method. Its implication is that the larger the r, the faster the convergence. And the above estimate is valid at least for $t > \tau$. In the numerical examples in section 4, we choose $\tau = 1/2$.

Next we deal with the space-discretization error.

LEMMA 3.6. *Let $U^{\Gamma_1,g}_{L_1,\tau}(t)$ and $U^{\Gamma_1,g}_{L_1,h,\tau}(t)$ be defined as in* (2.10) *and* (2.19)*, respectively, using g instead of u_0. Then, for $t > \tau$, we have*

$$\|U^{\Gamma_1,g}_{L_1,\tau}(t) - U^{\Gamma_1,g}_{L_1,h,\tau}(t)\| \le Ch^2(\beta M/\delta)^{t/T}\|g\|, \tag{3.13}$$

where $C = C\sqrt{1+\kappa^2}\left(\frac{2L_1-1}{L_1\tau}\right)$.

Proof. Combining (2.10), (2.19), and (2.8) we have

$$\begin{aligned}\|U^{\Gamma_1,g}_{L_1,\tau}(t) - U^{\Gamma_1,g}_{L_1,h,\tau}(t)\| &\le \frac{1}{2\pi}\frac{1}{L_1}\sum_{j=-L_1+1}^{L_1-1} |e^{z_j t}| \|\widehat{u}^g(z_j) - \widehat{u}^g_h(z_j)\| \left|\frac{dz}{d\omega}(\omega_j)\frac{d\psi^{-1}}{dy}(y_j)\right| \\ &\le C\sqrt{1+\kappa^2}h^2\|g\|e^{\gamma t}\frac{1}{L_1}\sum_{j=-L_1+1}^{L_1-1} e^{-|\omega_j|t}\left|\frac{d\psi^{-1}}{dy}(y_j)\right| \qquad (3.14)\end{aligned}$$

since, for h small (see [33]),

$$\|\widehat{u}^g(z) - \widehat{u}^g_h(z)\| \le Ch^2\|g\| \qquad \text{for } z \in \Gamma_1.$$

Owing to (3.7),

$$\frac{1}{L_1}\sum_{j=-L_1+1}^{L_1-1} e^{-|\omega_j|t}\left|\frac{d\psi^{-1}}{dy}(y_j)\right| \le \frac{2}{\tau}\left(\frac{2L_1-1}{L_1}\right).$$

Then (3.8) is used to complete the proof. ∎

Finally, combining Theorem 3.1, Lemma 3.4, Lemma 3.6, and the triangle inequality, we obtain the main result of this subsection.

THEOREM 3.7. *Let $u(t)$ be an exact solution of* (1.1)*, $g \in L^2(\Omega)$ be given satisfying* (1.2)*, and $U^{\Gamma_1,g}_{L_1,h,\tau}(t)$ be our fully discretized approximation to $u(t)$ defined by* (2.19)*. Then we have, for any integer $r \ge 1$,*

$$\begin{aligned}\|u(t) - U^{\Gamma_1,g}_{L_1,h,\tau}(t)\| &\le 2M^{t/T}\delta^{1-t/T} \\ &\quad + C\left(\frac{\beta M}{\delta}\right)^{t/T}\left(\frac{(1+\kappa^2)^{r/2}}{\kappa}\frac{1}{L_1^r} + \sqrt{1+\kappa^2}h^2\right)\|g\| \quad \textit{for} \quad r\tau < t < T.\end{aligned}$$

Remark 3.8. The choice of contour Γ_1 is flexible by letting the parameter κ variable with the asymptotes having slopes ± 1 and the real axis cut at $\gamma - \nu$. Indeed, in (2.8), let $\kappa = \kappa(\omega)$ be chosen such that the contour Γ_1 can be nearly parallel to the imaginary axis until it meets the line $y = \pm \tan\zeta$ if the eigenvalue $\operatorname{Re}\lambda_{N-1}$ is close to $\operatorname{Re}\lambda_N$. Inside the sector Σ_ζ the contour can be deformed analytically to have asymptotes with slopes ± 1 in order to have fast convergence. Corresponding to the contour

$$\Gamma_1 = \{z : z = z(\omega) = \sigma(\omega) + \mathrm{i}\kappa(\omega)\omega, -\infty < \omega < \infty\} \subset \rho(-A),$$
$$\sigma(\omega) = \gamma - \sqrt{\omega^2 + \nu^2},$$

suitable modifications in the analysis can be carried out accordingly.

3.2. Analysis of the numerical procedure using a union of small circles. Let us turn to analyzing the case of the second numerical scheme under the assumption that the eigenvalues of $-A$ are known. We shall see that the quadrature scheme (2.15) with Γ_2 is of spectral accuracy. Since Γ_2 is a compact subset of $\rho(-A)$, we may assume that

$$\|(zI + A)^{-1}\| \le B_2 \quad \text{for } z \in \Gamma_2, \tag{3.15}$$

with B_2 independent of z. From (2.2) and (2.13) it follows that $B_2 = O(\varepsilon^{-1})$.

Set $\beta' = e^{\varepsilon T}$. We then have the following stability estimate.

THEOREM 3.9. *Let $U_{L_2}^{\Gamma_2,u_0}(t)$ be the approximation defined by* (2.15) *of the regularized solution* (2.14)*. Then we have*

$$\|U_{L_2}^{\Gamma_2,u_0}(t)\| \le CN(\beta' M/\delta)^{t/T}\|u_0\| \quad \textit{for } t > 0.$$

Proof. By using (2.2), (2.15), and (3.15), a direct estimation leads to

$$\left\|U_{L_2}^{\Gamma_2,u_0}(t)\right\| = \left\|\sum_{k=1}^{N}\frac{1}{L_2}\sum_{j=0}^{L_2-1} e^{z_{k,j}t}\widehat{u}(z_{k,j})\varepsilon e^{\mathrm{i}\theta_j}\right\|$$
$$\le B_2 N\varepsilon e^{(\lambda_N+\varepsilon)t}\|u_0\| \le CNe^{(\operatorname{Re}\lambda_N+\varepsilon)t}\|u_0\|,$$

owing to the fact $B_2 = O(\varepsilon^{-1})$. Finally, since $\operatorname{Re}\lambda_N \le \log(M/\delta)/T$, one has

$$e^{(\operatorname{Re}\lambda_N+\varepsilon)t} \le (\beta' M/\delta)^{t/T}, \tag{3.16}$$

which proves the theorem. ∎

Recalling that $d = \min_{1\le k\le N}\min_{l\ne k}|\lambda_k - \lambda_l|$ and $\varepsilon < d$ (see section 2.2), we have the following lemma.

LEMMA 3.10. *Let $u^{\Gamma_2,g}(t)$ and $U_{L_2}^{\Gamma_2,g}(t)$ be the regularized solution defined by* (2.14) *and its approximation defined by* (2.15)*, respectively, with g instead of u_0 and L_2 a positive integer. Then we have, for some $C > 0$ independent of ε and L_2,*

$$\|u^{\Gamma_2,g}(t) - U_{L_2}^{\Gamma_2,g}(t)\| \le CN\left(\frac{\beta' M}{\delta}\right)^{t/T}\left(\varepsilon'^{L_2} + \frac{eps}{\varepsilon}\right)\|g\| \quad \textit{for } t > 0, \tag{3.17}$$

where eps is the machine precision, and $\varepsilon' = \varepsilon/d$.

Proof. Let $F(z) = e^{zt}\widehat{u}^g(z)$. By (2.2), $F(z)$ has simple poles at $z = \lambda_k$ and thus has a Laurent series expansion of the form

$$F(z) = \sum_{m=-1}^{\infty} c_{k,m}(z-\lambda_k)^m \quad \text{for } |z-\lambda_k| < d, \tag{3.18}$$

where $c_{k,m} = \frac{1}{2\pi \mathrm{i}} \int_{C_k} \frac{F(z)}{(z-\lambda_k)^{m+1}}\, dz$. Due to Cauchy's residue theorem, we get

$$u^{\Gamma_2,g}(t) = \sum_{k=1}^{N} c_{k,-1}. \tag{3.19}$$

Plugging (3.18) into (2.15) with u_0 replaced by g and rearranging the summand, we have

$$\begin{aligned} U_{L_2}^{\Gamma_2,g}(t) &= \sum_{k=1}^{N} \frac{1}{L_2} \sum_{m=-1}^{\infty} c_{k,m}\varepsilon^{m+1} \left(\sum_{j=0}^{L_2-1} (e^{\mathrm{i}\theta_{m+1}})^j \right) \\ &= \sum_{k=1}^{N} \left(c_{k,-1} + c_{k,L_2-1}\varepsilon^{L_2} + c_{k,2L_2-1}\varepsilon^{2L_2} + \cdots \right). \end{aligned} \tag{3.20}$$

The last equality follows from the fact that if $m+1$ is a multiple of L_2, the sum with index j becomes L_2; otherwise, it is $\frac{1-(e^{\mathrm{i}\theta_{m+1}})^{L_2}}{1-e^{\mathrm{i}\theta_{m+1}}}$, which is 0. Under the assumption of no machine round-off error, one then has the following type of bound: $\|u^{\Gamma_2,g}(t) - U_{L_2}^{\Gamma_2,g}(t)\| \le CN\varepsilon^{L_2}$. However, the extra term eps/ε in (3.17) will be included due to the round-off errors which become significant if ε is too small. For the derivation of this and various examples which show the validity of this estimation, we refer the reader to [22].

Now we calculate $c_{k,m}$ explicitly to get the final error form. By inserting (2.2) into $c_{k,m}$, for $m \ge 0$, we get $c_{k,m} = (1/2\pi\mathrm{i}) \int_{C_k} G(z)/(z-\lambda_k)^{m+1} dz$, where $G(z) = \sum_{l\neq k} \frac{e^{zt}}{z-\lambda_l} g_l\phi_l$, and $g_l = (g, \phi_l)$. Notice that $G(z)$ is analytic inside C_k and recall the definition of d. Since $\beta' = e^{\varepsilon T} > 1$, we have, by Cauchy's integral theorem for derivatives, the Leibniz rule, and (3.16), that

$$\begin{aligned} \|c_{k,m}\| &= \frac{1}{m!}\|G^{(m)}(\lambda_k)\| = \left\| \sum_{r=0}^{m} \frac{1}{m!} \binom{m}{r} t^{m-r} e^{\lambda_k t} (-1)^r r! \sum_{l\neq k} \frac{g_l\phi_l}{(\lambda_k-\lambda_l)^{r+1}} \right\| \\ &\le t^m e^{\mathrm{Re}\,\lambda_N t} \left\| \sum_{r=0}^{m} \frac{t^{-r}}{(m-r)!} \sum_{l\neq k} \frac{g_l\phi_l}{|\lambda_k-\lambda_l|^{r+1}} \right\| \\ &\le t^m e^{\mathrm{Re}\,\lambda_N t} \frac{1}{d} \sum_{r=0}^{m} \frac{1}{(m-r)!} \left(\frac{1}{dt}\right)^r \|g\| \\ &= \frac{1}{d} e^{\mathrm{Re}\,\lambda_N t} \sum_{s=0}^{m} \frac{(dt)^s}{s!} \frac{1}{d^m} \|g\| \\ &\le \frac{1}{d^{m+1}} e^{dt} \left(\frac{\beta' M}{\delta}\right)^{t/T} \|g\| \quad \text{for } k \le N. \end{aligned} \tag{3.21}$$

The last estimate combined with (3.19) and (3.20) leads to

$$\begin{aligned}\|u^{\Gamma_2,g}(t) - U_{L_2}^{\Gamma_2,g}(t)\| &= \sum_{k=1}^{N}\sum_{m=1}^{\infty} \|c_{k,mL_2-1}\varepsilon^{mL_2}\| \\ &\le N\sum_{m=1}^{\infty}\left(\frac{\varepsilon}{d}\right)^{mL_2} e^{dt}\left(\frac{\beta' M}{\delta}\right)^{t/T}\|g\| \\ &= N\frac{(\varepsilon')^{L_2}}{1-(\varepsilon')^{L_2}} e^{dt}\left(\frac{\beta' M}{\delta}\right)^{t/T}\|g\| \\ &\le CN\left(\frac{\beta' M}{\delta}\right)^{t/T}\left(\varepsilon'^{L_2} + \frac{eps}{\varepsilon}\right)\|g\|,\end{aligned}$$

with $C = \frac{e^{dT}}{1-(\varepsilon')^{L_2}}$, where the machine precision truncation was considered in the last estimate. This completes the proof. ∎

Remark 3.11. Inequality (3.21) is the worst case estimate. Although there exist concrete examples that show that the above estimate is sharp, most experiments similar to the examples in section 4 exhibit that d acts as 1.

Remark 3.12. For a fixed L_2, we should have $(\frac{\varepsilon}{d})^{L_2} + \frac{eps}{\varepsilon} \ge 2(\frac{eps}{d})^{L_2/(L_2+1)}$ with equality holding for $\varepsilon = d^{L_2/(L_2+1)} eps^{1/(L_2+1)}$, which is an optimal choice of ε for a fixed L_2. With this ε the error bound tends to $C \cdot eps/d$ as L_2 tends to ∞.

Remark 3.13. In our case of computing $F(z)$ numerically, the smaller the ε one chooses, the more computational costs one needs in order to achieve a given tolerance. Thus Remark 3.12 says that any choice of ε and L_2 such that $(\varepsilon/d)^{L_2}$ is less than a given tolerance is economic, provided $\varepsilon > d^{L_2/(L_2+1)} eps^{1/(L_2+1)}$. In the examples in section 4 we regard d as 1.

Next we consider the space-discretization error.

LEMMA 3.14. *Using g instead of u_0, let $U_{L_2}^{\Gamma_2,g}(t)$ and $U_{L_2,h}^{\Gamma_2,g}(t)$ be as in* (2.15) *and* (2.20)*, respectively. Then we have*

$$|U_{L_2}^{\Gamma_2,g}(t) - U_{L_2,h}^{\Gamma_2,g}(t)\| \le CNh^2(\beta' M/\delta)^{t/T}\|g\| \quad for\ t > 0. \tag{3.22}$$

Proof. The lemma is a consequence of the estimate $\|\widehat{u}^g(z) - \widehat{u}_h^g(z)\| \le Ch^2\|g\|$ for $z \in \Gamma_2$ and Theorem 3.9 with u_0 replaced by g. ∎

We are, finally, in a position to state the main result of this subsection.

THEOREM 3.15. *Let $g \in L^2(\Omega)$ be given data satisfying* (1.2)*, $u(t)$ be an exact solution of* (1.1)*, and $U_{L_2,h}^{\Gamma_2,g}$ be the fully discretized approximation to $u(t)$ defined by* (2.20)*. Then we have*

$$\|u(t) - U_{L_2,h}^{\Gamma_2,g}(t)\| \le 2M^{t/T}\delta^{1-t/T} + CN\left(\frac{\beta' M}{\delta}\right)^{t/T}\left(\varepsilon'^{L_2} + \frac{eps}{\varepsilon} + h^2\right)\|g\|, \tag{3.23}$$

for $0 < t < T$, with $\varepsilon' = \varepsilon/d$.

Proof. The proof is just a combination of Theorem 3.1, Lemma 3.10, Lemma 3.14, and the triangle inequality. ∎

4. Numerical examples. Examples 1, 2, and 3 have been chosen to illustrate the convergence theory developed in section 3. The complicated solution profiles produced in Examples 4 and 5 demonstrate the high quality of the regularized numerical

solutions compared with that of other recently proposed numerical methods [2, 24]. Parallel performance is reported in Example 6.

In each example the initial data with noise are generated by adding a perturbation to the exact initial data. Let x_i, $1 \le i \le J$, be a uniform partition of Ω. For each i, let $rd(x_i)$ be a pseudorandom number selected from $(-1, 1)$. Then define $rd : \Omega \to [-1, 1]$ by linear interpolation of $rd(x_i)$ and set $per(x) := \delta \cdot rd(x)/|\Omega|^{d/2}$, where d and δ denote the dimension of Ω and amplitude, respectively.

Example 1. Let $\Omega = (0, \pi)$ and $T = 4$. Then consider the following backward parabolic problem:

$$u_t + u_{xx} = 0 \quad \text{in } \Omega \times (0, T), \tag{4.1a}$$

$$u = 0 \quad \text{on } \partial\Omega \times (0, T), \tag{4.1b}$$

$$u_0(x) = e^{-4} \sin x + e^{-16} \sin 2x \quad \text{for } x \in \Omega, \tag{4.1c}$$

with the exact solution $u(x, t) = e^{t-4} \sin x + e^{4(t-4)} \sin 2x$. Notice that the eigenpairs are $\phi_k = \sqrt{2/\pi} \sin kx$, and $\lambda_k = k^2, k = 1, 2, \ldots$. In this case we have $\|u(\cdot, T)\| = 1.89$. Set $M = 2$ and $\delta = 10^{-2}$ in (1.2). Then $\lambda_N \le (1/T) \log(M/\delta) = 1.32$, by which $\lambda_1 = 1$ is the largest eigenvalue bounded. Thus we have $N = 1$ in Theorem 3.1.

In implementation, (2.19) is applied with the contour Γ_1 given in (2.8) with the parameters chosen in the following fashion without requiring information on eigenpairs:

- we take $\nu = 0.5$ and $\tau = 0.5$;
- γ is chosen such that $\gamma - \nu = (1/T) \log(M/\delta) = 1.32$ at which the contour crosses the real axis; in this case, $\gamma = 1.82$.

In practice where the exact eigenvalues are not known, we recommend that γ is chosen such that $\gamma - \nu = (1/T) \log(M/\delta)$ if it does not coincide with any eigenvalue of $-A$. If $(1/T) \log(M/\delta)$ happens to be an eigenvalue of $-A$ (this occurs with probability 0), replace $(1/T) \log(M/\delta)$ with a slightly larger number so that it is not an eigenvalue and assign the resulting value as $\gamma - \nu$ to choose γ such that (3.4) is satisfied.

Now we would like to verify Theorem 3.1, or equivalently Theorem 3.7, provided the term $2M^{t/T}\delta^{1-t/T}$ is dominant. To use (2.19), $L_1 = 64$ and $h = \pi/600$ are chosen so that the term $2M^{t/T}\delta^{1-t/T}$ is dominant in (3.15).

TABLE 1

L^2 errors for Example 1: *actually computed L^2 errors and the predicted L^2 error bounds for $\delta = 10^{-2}$ and 10^{-4}.*

t	L^2 errors with $\delta = 10^{-2}$		L^2 errors with $\delta = 10^{-4}$	
	Computed	Predicted	Computed	Predicted
1	0.586E-02	0.752E-01	0.592E-04	0.238E-02
2	0.159E-01	0.283E+00	0.450E-03	0.283E-01
3	0.490E-01	0.106E+01	0.230E-01	0.336E+00
4	0.126E+01	0.400E+01	0.125E+01	4.000E+00

Table 1 shows the computational results for the L^2 errors and the error bounds, $2M^{t/T}\delta^{1-t/T}$, predicted by the Theorem 3.1. To illustrate the effects of different noise levels, we also showed errors with $\delta = 10^{-4}$, leaving the other parameters fixed except $\gamma = (1/T) \log(M/\delta) + \nu = 2.98$. Observe that the computational results in Table 1 are better than those predicted bounds by the theorem. However, it is well known [5, 17] that the estimate (3.1) is sharp, and one cannot improve this

without introducing a supplementary constraint such as that of slow evolution from the continuation boundary (SECB) [5, 6].

Example 2. In this example we report the L^2 error estimates for the first method (2.19) of time discretization to solve the same problem as in Example 1 with the data u_0 replaced by $u_0 = e^{-4}\sin x$ without adding noise. This enables the term $2M^{t/T}\delta^{1-t/T}$ in (3.15) to be neglected. In this case, $1/T\log(M/\delta)$ is infinity, which implies that the contour should contain all eigencomponents of the solution. The solution $u(x,t) = e^{t-4}\sin x$ contains 1 eigencomponent, and thus $\gamma - \nu$ should be greater than 1. Therefore we assigned $\gamma = 1.82, \nu = 0.5$, and $\tau = 0.5$ as in Example 1. Keeping the term $C(\beta M/\delta)^{t/T}/L_1^r$ to be dominant in (3.15) by choosing a sufficiently small $h(= \pi/5000)$, we vary L_1 to see the error behavior. Table 2 summarizes the L^2 errors. The values in the parentheses denote the error reduction ratios defined by $\log_2(e_{L_1/2}/e_{L_1})$, where e_{L_1} is the L^2 error with L_1 in (3.15). Notice that the rates of convergence, while erratic, are asymptotically as large as the order $t/\tau = 2t$ predicted by Theorem 3.7. That is, the error reduction ratios at time t become larger than $2t$ as L_1 is chosen sufficiently large.

TABLE 2
L^2 errors (and their reduction ratios) for Example 2.

$t\backslash L_1$	4	8	16	32
1.0	0.144E-02	0.645E-03(1.16)	0.305E-05(7.72)	0.419E-06(2.86)
2.0	0.120E-01	0.154E-02(2.97)	0.613E-05(7.97)	0.109E-07(9.13)
3.0	0.112E+00	0.196E-02(5.84)	0.192E-04(6.67)	0.455E-07(8.73)
4.0	0.842E+00	0.162E-01(5.70)	0.791E-04(7.68)	0.164E-06(8.91)

Example 3. Now we examine the second method (2.20) for the same problem as in Example 2. Thus parameters in (2.13) are chosen such that $N = 1$ and $\lambda_1 = 1$. Since the time-discretization errors for this method become small very rapidly, it is not easy to make the term $\varepsilon^{L_2} + eps/\varepsilon$ be dominant in (3.23). However, in [22] the theoretical and computational results have already been shown for this scheme without space variable to show the rate $\varepsilon^{L_2} + eps/\varepsilon$. In the current example, instead of varying ε and L_2, we thus fix $L_2 = 4$ and $\varepsilon = 10^{-2} > eps^{(1/L_2+1)} \approx 10^{-3}$ (see Remark 3.13) in (3.23) and vary h eight times smaller at each step which will reduce errors 1/64 times. This would verify that the space-discretization errors are dominant compared to their time-discretization counterparts.

TABLE 3
L^2 errors (and their reduction ratios) for Example 3.

$t\backslash h$	1/30	1/240	1/1920	1/15360
1.0	0.665E-04	0.976E-06(6.09)	0.154E-07(5.99)	0.510E-08(1.59)
2.0	0.330E-03	0.496E-05(6.05)	0.790E-07(5.97)	0.124E-07(2.67)
3.0	0.131E-02	0.200E-04(6.04)	0.328E-06(5.93)	0.202E-07(4.02)
4.0	0.471E-02	0.722E-04(6.03)	0.126E-05(5.84)	0.395E-07(5.00)

The numerical results in Table 3 provide confirmation of the behavior predicted by Theorem 3.15, where the predicted space-discretization error-reduction ratio is 6 since we choose h to be 1/8 times smaller at each step and the predicted time-discretization error $\varepsilon^{L_2} + eps/\varepsilon$ is about 10^{-8}. In the columns with $h = 1/240$ and $h = 1/1920$, the

apparent reduction ratios are more or less in agreement with the theory, while the ratios in the last column are less than 6 since the time-discretization errors, which are about 10^{-8}, are actively deteriorating the numerical solutions.

We now proceed to deal with two more complicated examples than the previous ones: the example employed in [24] which has severely oscillatory data and a similar example which appeared in [1, 2] with nonsmooth data. Consider the backward parabolic problem:

$$u_t + cu_{xx} = 0, \quad \Omega \times (0,T), \tag{4.2a}$$

$$u = 0, \quad \partial\Omega, \tag{4.2b}$$

with given noisy data $g \in L^2(\Omega)$. Let $\Omega = (0,l), l = 1$, or π. Here c is a positive constant that controls the diffusion speed. In this case, eigenvalues are $c(k\pi/l)^2, k = 1, 2, \ldots$. Thus when we apply Γ_2, in (2.13) $\lambda_k = c(k\pi/l)^2$, and N is chosen by the maximum integer k such that $c(k\pi/l)^2 \le (1/T)\log(M/\delta)$.

Example 4. Consider the problem (4.2) with $l = 1$, $T = 16$, $c = 1/87960$, and $u(x,T) = \sin(25\pi x^2)$. Since no analytic expression for u_0 is available, we solved a forward problem starting from $u(\cdot,T)$ using the method introduced in [32] to approximate u_0 as well as $u(\cdot,T/4), u(\cdot,T/2)$, $u(\cdot,3T/4), u(\cdot,7T/8)$, and $u(\cdot,15T/16)$, with errors less than 10^{-7}, which are served as reference solutions for (4.2).

The constant c is chosen so that the size of the oscillations in u_0 close to the right endpoint is less than 1 percent of that of $u(T)$ at the right endpoint. In [24], a similar data u_0 is used such that the size of oscillations in u_0 close to the right endpoint is less than 10 percent of their initial size. (Indeed, in [24] the constant $c = 1$ is used with very small time duration, with which the solutions agree with ours. Observe that any c and T in (4.2) give equivalent solutions if cT is constant.) Once u_0 is calculated, multiplicative noise is added (see Figure 1(c)) to make g such that $g = u_0(1 + per(x))$ with $\delta = 10^{-3}$, which should be stressed by comparing this with the choice $\delta = 10^{-6}$ in [24].

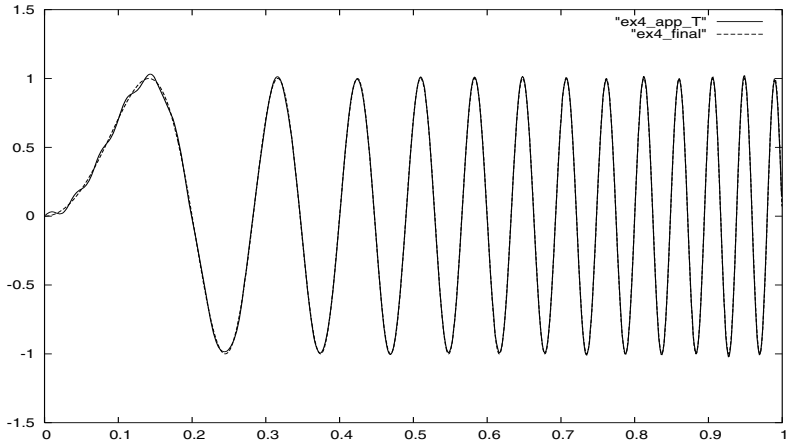
We have $\|u(\cdot,T)\| = M \approx 0.69$, and $(1/T)\log(M/\delta) \approx 0.41$. Thus when applying Γ_1, we choose $\nu = 0.5, \tau = 0.5$, and $\gamma = 0.91$. Also we used the discretization parameters $L_1 = 200$ and $h = 1/1000$. When we apply Γ_2, we choose $L_2 = 3, \varepsilon = eps^{1/(L_2+1)} \approx 10^{-4}$, and $h = 1/5000$. Figures 1(a) and 1(b) show the exact and the computed solutions at $t = T$ based on the contours Γ_1 and Γ_2, respectively.

Table 4 shows the predicted L^2 error bounds by Theorem 3.7 and the computed errors at various t values for the two contours Γ_1 and Γ_2. In order to compare with the method proposed in [24], computational results in [24] are also presented. The results from [24] are given in L^∞ errors, and thus we calculate L^∞ errors also in the parenthesis, although they are not much different since solutions are smooth.

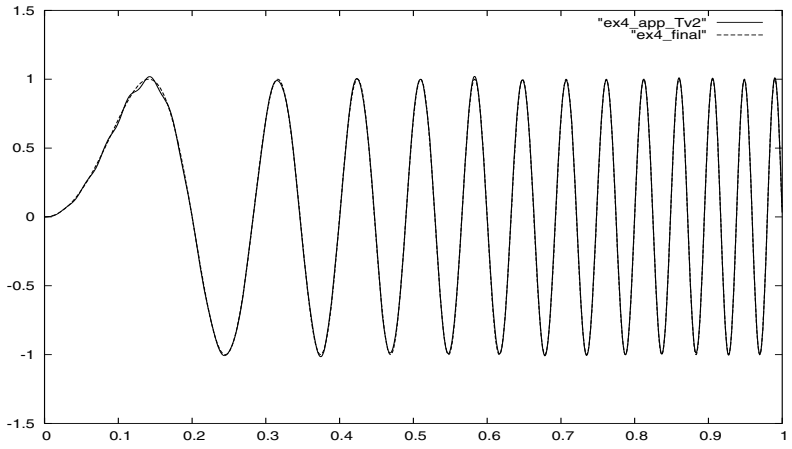
We observe that the method [24] gives better results at $t = T/4$, but as we proceed to the final time our methods provide better solutions. Such an observation is expected for the following reasons. First, our methods recover information on eigenpairs without artificial contaminant since they do not perturb the original differential equation. Second, high frequency components of noise (larger than λ_N) do not affect our numerical solutions since they are automatically cut off in implementation.

We remark that the noise amplitude of our data g is 10^3 times as big as that in [24], and the loss of information on given data is worse than that of [24] by 10 times. With such a bad data our methods recover solutions relatively well.

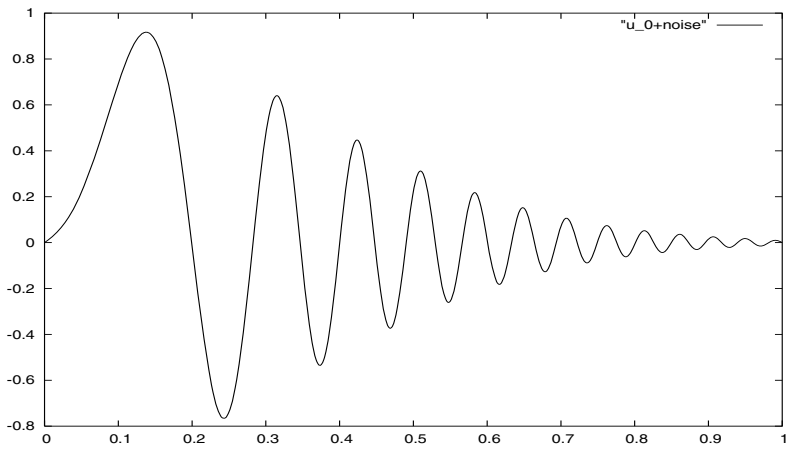
We should also remark that the computed errors are much better than the predicted bounds given in Theorem 3.7. Even at $t = T$ when we lose continuous depen-



(a)



(b)



(c)

FIG. 1. (a) *Exact and computed solutions at* $t = T$ *using the contour* Γ_1, (b) *the contour* Γ_2, *and* (c) *the perturbed initial data profile with noise.*

TABLE 4
Comparison of errors for Example 4.

Time	Predicted errors	$L^2(L^\infty)$ errors using the contour Γ_1	$L^2(L^\infty)$ errors using the contour Γ_2	L^∞ errors from [24]
T/4	1.02E-02	1.56E-04 (6.83E-04)	1.46E-03 (2.42E-03)	6.25E-05
T/2	5.25E-02	5.88E-04 (2.16E-03)	1.57E-03 (3.64E-03)	1.09E-03
3T/4	2.69E-01	2.63E-03 (8.04E-03)	2.92E-03 (9.20E-03)	1.40E-02
7T/8	6.09E-01	5.71E-03 (1.62E-02)	5.65E-03 (1.69E-02)	4.44E-02
15T/16	9.16E-01	8.47E-03 (2.30E-02)	8.19E-03 (2.36E-02)	7.77E-02
T	1.38E+00	1.26E-02 (3.30E-02)	1.20E-02 (3.33E-02)	-

dence on the data theoretically, the computed L^∞ errors are of only 3.30 and 3.33 percent. This is typically the case when the diffusion coefficient is small, and the very high frequency components of the noise do not play a dominant role. This can be explained theoretically if we introduce further constraints, SECB [5, 6], which would be treated in a future work.

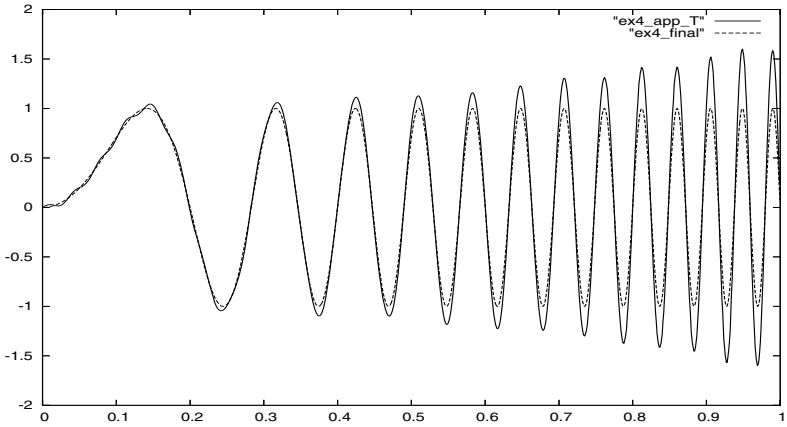
Finally, we would like to discuss our choice of $(c, T) = (1/87960, 16)$, which corresponds to $(c, T) = (1, 0.0001819)$. Although our T with $c = 1$ is bigger than that of [24], it is still small. While making "reasonable recovery," how large a T can one choose? It depends on the definition of "reasonable recovery" and which method one uses. Let us fix the L^2-norm of noises; say it is about 10^{-3}. We can say, for example, if the L^2 error of the recoverd image is less than 0.2, it is a "reasonable recovery." With our method, we can make $(c, T) = (1, 0.0002018)$. Recovered and the exact solutions are shown in Figure 2(a), where we observe highly tilted values as x approaches 1, which come from noises. Thus it is still challenging to make T as large as possible with such highly oscillating profiles.

Example 5. Now we consider (4.2) with $l = \pi$, $T = 4$, and $c = 1/32$, and the piecewise linear solution at $t = T$ given by

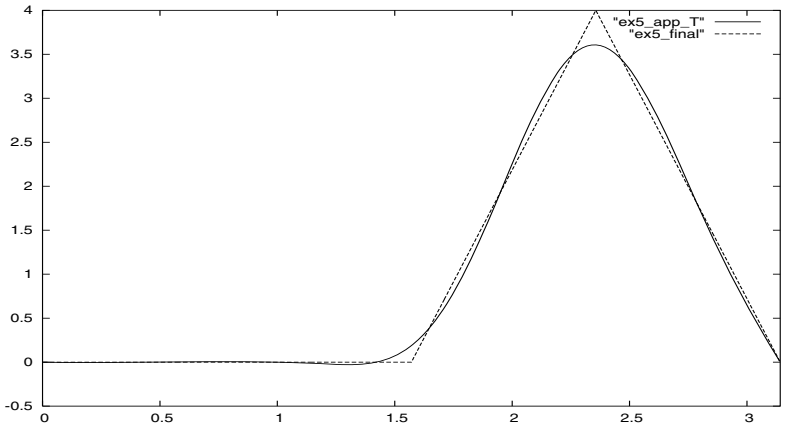
$$u(\cdot, T) = \begin{cases} 0, & 0 \le x \le \frac{\pi}{2}, \\ \frac{16}{\pi}x - 8, & \frac{\pi}{2} \le x \le \frac{3}{4}\pi, \\ 16 - \frac{16}{\pi}x, & \frac{3}{4}\pi \le x \le \pi. \end{cases} \tag{4.3}$$

As in Example 4, we integrate forward in time starting from $u(T)$ to generate u_0, $u(T/4)$, $u(T/2)$, and $u(3T/4)$, which are served as reference solutions for the problem. Additive noises are introduced by $g = u_0 + per$ with $\delta = 10^{-3}$. In this case $\|u(T)\| = M \approx 2.89$, and we choose $\gamma = (1/T)\log(M/\delta) + \nu \approx 2.49, L_1 = 100$, and $h = \pi/1000$ for Γ_1. For Γ_2 we take $L_2 = 4, \varepsilon = 10^{-2} > eps^{1/(L_2+1)} \approx 10^{-3}$, and $h = \pi/1000$. Table 5 shows L^2 errors, and Figure 2(b) presents the exact and the computed solution profiles at $t = T$ based on the contour Γ_1. Although the authors in [1, 2] stated a pessimistic opinion about approximating $u(T) \notin C^1$, our methods recover this profile in relatively good shape even with noisy data. The noise used for Example 5 is plotted in Figure 2(c).

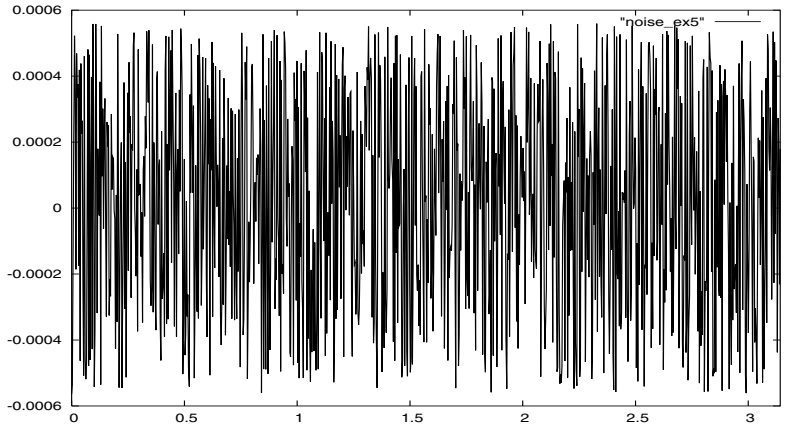
Example 6. Finally, in order to observe the parallel performance of the proposed method, we solve a spatially three-dimensional (3D) problem version of (4.1)



(a)



(b)



(c)

FIG. 2. (a) *Numerical and exact solutions with bigger* T *in Example* 4 *(*L^2 *error* $= 0.19$*),* (b) *profiles of computed and exact solutions of Example* 5 *using* Γ_1 *at* $t = T$*,* (c) *per*(x) *used for Example* 5.

TABLE 5
Predicted and computed errors for Example 5.

Time	Predicted errors	L^2 errors using the contour Γ_1	L^2 errors using the contour Γ_2
T/4	1.47E-02	7.05E-05	2.96E-04
T/2	1.08E-01	3.12E-04	8.51E-04
3T/4	7.89E-01	4.61E-03	7.13E-03
T	5.79E+00	1.48E-01	1.54E-01

TABLE 6
The computing time in seconds and speedup for Example 6. *The relative* $L^2(\Omega)$ *error of the approximation is* 5.99×10^{-3}.

# of processor(p)	1	2	4	8	16	32
Computing time	635.9	327.7	164.9	83.1	41.7	21.1
Speedup	1	1.9	3.9	7.7	15.2	30.1

($\Omega = (0,1)^3$ and $T = 1$) with $u_0(x) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$. In this case, the exact solution is $u(x,t) = e^{3\pi^2 t}\sin(\pi x)\sin(\pi y)\sin(\pi z)$. For a fully discretized numerical solution, (2.19) is applied with $L_1 = 32$. $\gamma = 31.5$, and $\nu = \tau = 0.5$ are used for defining Γ_1 (see Example 2 for these selections). To obtain $\widehat{u}_h^{u_0}(z_j)$ ($j = 1, \ldots, L_1 - 1$) in (2.19), we solved (2.1) using the Q_1-conforming (*e.g., trilinear*) finite element space V_h based on the $64 \times 64 \times 64$ uniform hexahedron triangulation of Ω (see, for instance, [3]). Then in order to calculate $U_{L_1,h,\tau}^{\Gamma_1,u_0}(1)$, p processors ($p = 1, 2, 4, 8, 16, 32$) are employed to compute $L_1(=32)$ independent elliptic problems (2.1) for $z_j, j = 0, \ldots, L_1$, by evenly distributing them to p processors. For example if $p = 32$, each processor solves only one equation. The computing clock time in seconds from solving linear equations derived from (2.17) until obtaining $U_{L_1,h,\tau}^{\Gamma_1,u_0}(1)$ required with p processors are reported in Table 6. The speedup defined by (computing time for 1 processor)/(computing time for p processors) is also presented and a nearly perfect speedup is observed. The result is expected since the most time-consuming part in obtaining $\widehat{u}_h^{u_0}(z_j)$ for $j = 0, \ldots, L_1 - 1$ can be solved independently without data communication. Finally, we remark that obviously such a perfect speedup is expected when $p \leq L_1$. The computation of this example was carried out using a parallel machine whose nodes are based on the IBM PowerPC970 (2.2GHz) 2-way CPUs with 1 Giga Byte Myrinet network link.

5. Conclusion. In this paper, a parallel method has been proposed to solve backward parabolic problems. The algorithm is based on the Laplace transformation in time of the original time-dependent problems on a suitable contour in the complex plane. The time dependence of the resulting Laplace transformed problems is thus suppressed. After solving elliptic problems for each point on the complex contour, a numerical inversion of Laplace transformed solutions will recover the time-dependent solutions.

The proposed scheme to solve parabolic problems backwards in time does not introduce an artificial parameter in order to regularize the numerical solutions. Theories and numerical examples show the proposed method gives optimal stability without perturbing anything, resulting in improved quality of the regularized solutions.

Additionally our scheme is highly scalable for parallel implementaion in the realm of time discretization.

Acknowledgments. The authors wish to express thanks to the anonymous referees for their valuable suggestions to improve the original version. Thanks also go to Dr. Taeyoung Ha for providing a parallel 3D finite element code.

REFERENCES

[1] K. A. Ames, G. W. Clark, J. F. Epperson, and S. F. Oppenheimer, *A comparison of regularizations for an ill-posed problem*, Math. Comp., 67 (1998), pp. 1451–1471.

[2] K. A. Ames and J. F. Epperson, *A kernel-based method for the approximate solutions of backward parabolic problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1357–1390.

[3] S. Brenner and R. Scott, *The Methematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[4] B. L. Buzbee and A. Carasso, *On the numerical computation of parabolic problems for preceding times*, Math. Comp., 27 (1973), pp. 237–266.

[5] A. S. Carasso, *Overcoming Hölder continuity in ill-posed continuation problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1535–1557.

[6] A. S. Carasso, *Logarithmic convexity and the "slow evolution" constraint in ill-posed initial value problems*, SIAM J. Math. Anal., 30 (1999), pp. 479–496.

[7] G. Clark and S. Oppenheimer, *Quasireversibility methods for nonwell-posed problems*, Electron. J. Differential Equations, 1994 (1994), pp. 1–9.

[8] D. Colton and J. Wimp, *The construction of solutions to the heat equation backward in time*, Math. Methods Appl. Sci., 1 (1979), pp. 32–39.

[9] J. Douglas, Jr., and J. R. Cannon, *The approximation of harmonic and parabolic functions on half-spaces from interior data*, in Numerical Analysis of Partial Differential Equations Symposium, Edizioni Cremonese, Rome, 1968, pp. 193–230.

[10] J. Douglas, Jr., J. E. Santos, and D. Sheen, *Approximation of scalar waves in the space–frequency domain*, Math. Models Methods Appl. Sci., 4 (1994), pp. 509–531.

[11] J. Douglas, Jr., J. E. Santos, D. Sheen, and Lynn S. Bennethum, *Frequency domain treatment of one–dimensional scalar waves*, Math. Models Methods Appl. Sci., 3 (1993), pp. 171–194.

[12] L. Elden, *Time discretization in the backward solution of parabolic equations.* I, Math. Comp., 39 (1982), pp. 53–68.

[13] L. Elden, *Time discretization in the backward solution of parabolic equations.* II, Math. Comp., 39 (1982), pp. 69–84.

[14] R. E. Ewing, *The approximation of certain parabolic equations backward in time by Sobolev equations*, SIAM J. Math. Anal., 6 (1975), pp. 283–294.

[15] J. Hadamard, *Lectures on the Cauchy Problems in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, 1923.

[16] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Survey Math. Indust., 3 (1993), pp. 253–315.

[17] W. Höhn, *Finite elements for parabolic equations backwards in time*, Numer. Math., 40 (1982), pp. 207–227.

[18] H. Houde and H. Gang, *Stabilized numerical approximations of the backward problem of a parabolic equation*, A Journal of Chinese Universities, 10 (2001), pp. 182–192.

[19] F. John, *Continuous dependence on data for solutions with a prescribed bound*, Comm. Pure Appl. Math., 13 (1960), pp. 551–585.

[20] K. Kwon and D. Sheen, *A parallel method for the numerical solution of integro-differential equation with positive memory*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4641–4658.

[21] R. Lattes and J. L. Lions, *The Method of Quasi-Reversibility: Applications to Partial Differential Equations*, Elsevier, New York, 1969.

[22] J. Lee and D. Sheen, *An accurate numerical inversion of Laplace transforms based on the location of their poles*, Comput. Math. Appl., 48 (2004), pp. 1415–1423.

[23] P. Manselli and K. Miller, *Dimensionality reduction methods for efficient numerical solution, backward in time, of parabolic equations with variable coefficients*, SIAM J. Math. Anal., 11 (1980), pp. 147–159.

[24] J. M. Marbán and C. Palencia, *A new numerical method for backward parabolic problems in the maximum-norm setting*, SIAM J. Numer. Anal, 40 (2002), pp. 1405–1420.

[25] K. Miller, *Three circle theorems in partial differential equations and applications to improperly posed problems*, Arch. Rational Mech. Anal., 16 (1964), pp. 126–154.

[26] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J.

Math. Anal., 1 (1970), pp. 52–74.
[27] K. Miller, *Stabilized quasi-reversibility and other nearly-best-possible methods for nonwell-posed problems interior data*, in Proceedings of the Sympoisum on NonWell-Posed Problems and Logarithmic Convexity, Lecture Notes in Math. 316, Springer-Verlag, Berlin, 1973, pp. 161–176.
[28] K. Moszyński, *Approximation with frequency filter for backward parabolic equations*, Numer. Math., 88 (2001), pp. 159–183.
[29] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, SIAM, Philadelphia, 1975.
[30] T. I. Seidman, *Optimal filtering for the backward heat equation*, SIAM J. Numer. Anal., 33 (1996), pp. 162–170.
[31] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature*, Math. Comp., 69 (2000), pp. 177–195.
[32] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic equations based on Laplace transformation and quadrature*, IMA J. Numer. Anal., 23 (2003), pp. 269–299.
[33] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1487–1513

ANISOTROPIC MESH REFINEMENT: THE CONDITIONING OF GALERKIN BOUNDARY ELEMENT MATRICES AND SIMPLE PRECONDITIONERS*

IVAN G. GRAHAM† AND WILLIAM MCLEAN‡

Abstract. In this paper we obtain upper and lower bounds on the spectrum of the stiffness matrix arising from a finite element Galerkin approximation (using nodal basis functions) of a bounded, symmetric bilinear form which is elliptic on a Sobolev space of real index $m \in [-1, 1]$. The key point is that the finite element mesh is required to be neither quasi-uniform nor shape-regular, so that our theory allows anisotropic meshes often used in practice. (However, we assume that the polynomial degree of the elements is fixed.) Our bounds indicate the ill-conditioning which can arise from anisotropic mesh refinement. In addition we obtain spectral bounds for the diagonally scaled stiffness matrix, which indicate the improvement provided by this simple preconditioning. For the special case of boundary integral operators on a two-dimensional screen in $\mathbb{R}^3$, numerical experiments show that our bounds are sharp. We find that diagonal scaling essentially removes the ill-conditioning due to mesh degeneracy, leading to the same asymptotic growth in the condition number as arises for a quasi-uniform mesh refinement. Our results thus generalize earlier work by Bank and Scott [*SIAM J. Numer. Anal.*, 26 (1989), pp. 1383–1394] and Ainsworth, McLean, and Tran [*SIAM J. Numer. Anal.*, 36 (1999), pp. 1901–1932] for the shape-regular case.

Key words. symmetric elliptic problems, boundary integral equations, Galerkin approximation, anisotropic refinement, spectral bounds, diagonal scaling, condition number estimates

AMS subject classifications. 65N38, 65N30, 65N22, 65F10, 65F35

DOI. 10.1137/040621247

1. Introduction.

Edge and corner singularities are characteristic features of solutions to three-dimensional (3D) elliptic boundary value problems and, in both finite element and boundary element methods, are commonly dealt with by some kind of local mesh refinement. Typically, an edge singularity is strongly *anisotropic*: the lack of smoothness occurs only in directions normal to the edge. For this reason, the local mesh refinement should also be anisotropic if we are to minimize the number of degrees of freedom used to achieve a sufficiently small error in, say, the energy norm. Extra refinement is not needed parallel to an edge, except maybe in the vicinity of a corner. The meshes that result from such local refinement are certainly not quasi-uniform and usually even fail to be *shape-regular* because elements near edges but away from any corner may have a very large aspect ratio.

In this paper we investigate the influence of such meshes on the condition number of the stiffness matrix arising in the Galerkin approximation of a class of symmetric elliptic problems. Our general framework includes as special cases the single layer and the hypersingular boundary integral equations for the Laplacian on the surface of a 3D Lipschitz domain or on a Lipschitz screen as well as the Dirichlet problem for second-order symmetric elliptic PDEs. We obtain general bounds for the spectrum of

*Received by the editors December 21, 2004; accepted for publication (in revised form) January 19, 2006; published electronically August 7, 2006. This work was supported by UK Engineering and Physical Sciences Research Council grant GR/S43399/01.

http://www.siam.org/journals/sinum/44-4/62124.html

†Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK (I.G.Graham@bath.ac.uk).

‡School of Mathematics, University of New South Wales, Sydney, NSW 2052, Australia (w.mclean@unsw.edu.au).

the resulting Galerkin matrix in terms of quantities which depend on the geometry of the elements and the particular basis functions utilized.

In addition we study in detail two model problems—the weakly singular and hypersingular equations for the Laplacian on a rectangular screen, discretized using classical tensor-product power-graded meshes. For these model problems our general estimates yield explicit bounds in terms of the number of degrees of freedom and the strength of the power grading. We show by numerical experiments that our estimates are sharp and, moreover, exhibit a strong increase in the condition number as the maximum aspect ratio of the elements increases.

These results have practical implications for the performance of iterative techniques such as conjugate gradients, which are often used as solvers for the dense linear systems which arise in these methods (usually combined with a fast matrix-vector multiplication such as fast multipole [12] or panel-clustering [8]). Efficient solvers require effective preconditioners, and as a first step in this direction we also analyze in detail the use of diagonal scaling. We obtain general estimates for the spectrum of the diagonally scaled matrix and again investigate this in fine detail for the special cases of the model problems mentioned above.

Throughout the paper Γ will denote either a bounded, d-dimensional Lipschitz surface in $\mathbb{R}^{d+1}$, for $d = 2$, or a bounded Lipschitz domain in $\mathbb{R}^d$, for $d = 2$ or 3. In the former case, the surface Γ may be open or closed. B will denote a bounded and *symmetric* bilinear form such that, for some Sobolev index m satisfying $|m| \le 1$, $B : \widetilde{H}^m(\Gamma) \times \widetilde{H}^m(\Gamma) \to \mathbb{R}$, and

$$c\|v\|^2_{\widetilde{H}^m(\Gamma)} \le B(v,v) \le C\|v\|^2_{\widetilde{H}^m(\Gamma)} \quad \text{for all } v \in \widetilde{H}^m(\Gamma), \tag{1.1}$$

where c and C are positive constants. Thus the energy space for B is equivalent to the Sobolev space $\widetilde{H}^m(\Gamma)$. (Here we are working with standard Sobolev spaces on Γ; see section 3 for more detail.)

We shall consider approximations of the following variational problem: Find $u \in \widetilde{H}^m(\Gamma)$ such that

$$B(u,v) = \langle f, v\rangle_\Gamma \quad \text{for all } v \in \widetilde{H}^m(\Gamma), \tag{1.2}$$

where, with $d\sigma$ denoting the usual surface element on Γ,

$$\langle f, v\rangle_\Gamma = \int_\Gamma f v \, d\sigma.$$

By the Lax–Milgram lemma, (1.2) has a unique solution $u \in \widetilde{H}^m(\Gamma)$ for each $f \in H^{-m}(\Gamma)$.

Within this abstract framework we can treat not only some boundary element methods, in particular with $m = \pm 1/2$, but also finite element methods for symmetric H^1 elliptic PDEs with homogeneous Dirichlet boundary conditions.

To approximate the solution u of (1.2), we introduce a finite-dimensional subspace $X \subseteq \widetilde{H}^m(\Gamma)$ and then apply Galerkin's method, seeking $u_X \in X$ such that

$$B(u_X, v) = \langle f, v\rangle_\Gamma \quad \text{for all } v \in X. \tag{1.3}$$

In this paper we are concerned only with the h-version of the finite element method in which X is a space of piecewise polynomials of fixed degree with respect to a family of increasingly refined partitions (or meshes) $\{\mathcal{P}\}$ on Γ. The partition $\mathcal{P}$ contains

elements $K \in \mathcal{P}$ which have diameter h_K and diameter of largest inscribed ball ρ_K. We will introduce a basis for X consisting of nodal basis functions $\{\phi_j : j \in \mathcal{N}\}$, where $\mathcal{N}$ is a suitable index set with cardinality N. We will define the allowable partitions, elements, and basis functions more precisely in section 3. Writing $u_X = \sum_{k\in\mathcal{N}} \alpha_k \phi_k$, inserting into (1.3), and choosing $v = \phi_j$ for each $j \in \mathcal{N}$ leads to the $N \times N$ linear system

$$\boldsymbol{B}\boldsymbol{\alpha} = \boldsymbol{f}, \tag{1.4}$$

with a symmetric positive definite matrix $\boldsymbol{B} = [B(\phi_k, \phi_j)]$, a solution vector $\boldsymbol{\alpha} = [\alpha_k]$, and a right-hand side vector $\boldsymbol{f} = [\langle f, \phi_j\rangle]$.

The conditioning of $\boldsymbol{B}$ in the case of shape-regular mesh refinement (i.e., $h_K \lesssim \rho_K$ for all $K \in \mathcal{P}$) was investigated by Ainsworth, McLean, and Tran in [1, 2], where the condition number estimate

$$\operatorname{cond}(\boldsymbol{B}) \lesssim \left(\frac{h_{\max}}{h_{\min}}\right)^{d-2m} N^{2|m|/d} \quad \text{for } 2|m| < d \tag{1.5}$$

was proved. Here, $h_{\max} = \max_{K\in\mathcal{P}} h_K$ and $h_{\min} = \min_{K\in\mathcal{P}} h_K$. (For matrices $\boldsymbol{B}$ with positive spectrum, $\operatorname{cond}(\boldsymbol{B}) := \lambda_{\max}(\boldsymbol{B})/\lambda_{\min}(\boldsymbol{B})$, where $\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\min}(\boldsymbol{B})$ denote the largest and smallest eigenvalues of $\boldsymbol{B}$, respectively. The symbols $\lesssim$ and $\simeq$ indicate (in)equality up to a hidden constant, independent of the mesh; see section 3.) In the limiting cases $2m = -d$ and $2m = d$ an additional logarithmic factor occurs in the bound (1.5).

For quasi-uniform meshes, $h_{\max} \simeq h_{\min} \simeq h$ so the bound (1.5) gives the well-known result that $\operatorname{cond}(\boldsymbol{B}) = O(N^{2|m|/d}) = O(h^{-2|m|})$. However, this bound deteriorates if the global mesh ratio $h_{\max}/h_{\min}$ becomes large, and the deterioration becomes stronger as the Sobolev index m becomes more negative. Fortunately, such additional growth in the condition number is easily eliminated by *diagonal scaling*. In fact, let $\boldsymbol{D}$ denote the diagonal matrix formed from $\boldsymbol{B}$ by setting all the off-diagonal entries to zero, and put

$$\boldsymbol{B}' = \boldsymbol{D}^{-1/2}\boldsymbol{B}\boldsymbol{D}^{-1/2}. \tag{1.6}$$

Then it is shown in [1] that in the shape-regular case we have

$$\operatorname{cond}(\boldsymbol{D}^{-1}\boldsymbol{B}) = \operatorname{cond}(\boldsymbol{B}') \lesssim N^{2|m|/d} \quad \text{for } 2|m| < d. \tag{1.7}$$

We remark that $\boldsymbol{B}' = [B(\phi'_j, \phi'_k)]$ is just the Galerkin matrix that arises if we scale the nodal basis so that $\phi'_j = \phi_j/\sqrt{B(\phi_j, \phi_j)}$ has unit energy; i.e., $B(\phi'_j, \phi'_j) = 1$.

This paper obtains bounds analogous to (1.5) and (1.7) for the case when the $\{\mathcal{P}\}$ is no longer required to be shape-regular, and each partition $\mathcal{P}$ may contain elements K for which the aspect ratio h_K/ρ_K approaches infinity as the mesh is refined.

In particular we show that in the case of the weakly singular and hypersingular boundary integral operators (and except possibly for some logarithmic factors), diagonal scaling removes the ill-conditioning produced by the high aspect ratios, restoring the rate of growth of the condition number (in terms of the number of degrees of freedom) to essentially what it would be for a quasi-uniform mesh with the same number of degrees of freedom.

We remark that our results not only generalize the results [1, 2] to more general meshes, but they also generalize some of the earlier results of Bank and Scott [3],

who obtained the analogous result for H^1 finite elements and shape-regular mesh sequences.

The layout of this paper is as follows. In section 2 we motivate the theory by describing the results for the weakly singular and hypersingular operators in detail, without proofs. In section 3 we set the theoretical scene by describing the class of finite (boundary) elements which we shall consider (which allow degenerate meshes), and we introduce the corresponding nodal bases. A key step in the theory in [1, 2] is the proving of estimates for Sobolev norms of nodal basis functions. In section 4 we extend these estimates to the case of non–shape-regular meshes. Here we make essential use of recently derived inverse estimates for finite element functions on anisotropic meshes [7]. In section 5 we obtain general bounds on the spectra of $\boldsymbol{B}$ and $\boldsymbol{B}'$ in terms of the geometry of the elements and the Sobolev norms of the nodal basis functions. For the case of power graded meshes and the weakly singular and hypersingular operators these lead to quantitative spectral estimates which are tested in the numerical experiments in section 6. These eperiments show that the results for $\boldsymbol{B}'$ are not completely sharp. Sharper results for special cases, which explain the numerical results, are proved in section 7. Finally, section 8 presents some additional numerical results using a more complicated family of meshes.

2. Examples.

2.1. Two integral equations. The *weakly singular* (or *single-layer*) boundary integral equation,

$$\frac{1}{4\pi}\int_\Gamma \frac{u(y)}{|x-y|}\,d\sigma_y = f(x) \quad \text{for } x\in\Gamma, \tag{2.1}$$

arises, for example, in the solution of the Dirichlet problem for the Laplace equation in the region exterior to Γ. This equation (2.1) may be written in the form (1.2), with

$$B(u,v) = \frac{1}{4\pi}\iint_{\Gamma\times\Gamma} \frac{u(y)v(x)}{|x-y|}\,d\sigma_x\,d\sigma_y. \tag{2.2}$$

Then B satisfies the norm equivalence (1.1) for $m=-1/2$.

The *hypersingular* integral equation,

$$-\frac{1}{4\pi}\int_\Gamma \left(\frac{\partial}{\partial\nu_x}\frac{\partial}{\partial\nu_y}\frac{1}{|x-y|}\right)u(y)\,d\sigma_y = f(x) \quad \text{for } x\in\Gamma, \tag{2.3}$$

arises, for example, in the solution of the Neumann problem for the Laplace equation. (Here $\partial/\partial\nu_x$ denotes the normal derivative at $x\in\Gamma$, and the integral is defined as the finite part integral in the sense of Hadamard.) The integration by parts procedure of Nédélec [10], [9, Theorem 9.15] allows us to write the associated bilinear form as

$$B(u,v) = \frac{1}{4\pi}\iint_{\Gamma\times\Gamma} \frac{\mathbf{curl}_\Gamma u(x)\cdot\mathbf{curl}_\Gamma v(y)}{|x-y|}\,d\sigma_x\,d\sigma_y, \tag{2.4}$$

where $\mathbf{curl}_\Gamma$ denotes the surface curl operator. The norm equivalence (1.1) holds for $m=+1/2$. If the surface Γ is flat, then $\mathbf{curl}_\Gamma$ can be replaced by the two-dimensional (2D) gradient operator ∇.

It is well known that the solutions of (2.1) and (2.3) in general exhibit singular behavior near the edges and corners of Γ. In particular near the interior of an edge, the solution u of (2.1) typically has a singularity of order $O(\rho^{\alpha-1})$ as $\rho\to 0$, where ρ

is the distance of a point from the edge and $\alpha > 1/2$ depends on the angle subtended by the boundary Γ near the edge. The solution of (2.3) typically has a singularity of order $O(\rho^\alpha)$. More complicated behavior appears near corners. The full detail is well known; see, e.g., [4, 5, 11].

2.2. Power-graded meshes. For the h-version of the boundary integral method, it is common to approximate (2.1) and (2.3) using power-graded meshes. To describe these, first consider the special case of a flat, square screen

$$\Gamma = \{\, x \in \mathbb{R}^3 : 0 < x_1 < 1, 0 < x_2 < 1, x_3 = 0 \,\}, \tag{2.5}$$

and think of Γ as a subset of $\mathbb{R}^2$ by writing $x = (x_1, x_2) = (x_1, x_2, 0)$ for $x \in \Gamma$. Choose a *grading exponent* $\beta \geq 1$ and define a mesh on the interval $(0, 1)$ by

$$t_j = \begin{cases} \dfrac{1}{2}\left(\dfrac{2j}{n}\right)^\beta & \text{if } 0 \leq j \leq n/2, \\ 1 - t_{n-j} & \text{if } n/2 < j \leq n. \end{cases} \tag{2.6}$$

For $\beta = 1$ the mesh is uniform, but as β increases from 1, the points are more concentrated at each end of the interval. The length $\Delta t_j = t_j - t_{j-1}$ of the jth interval satisfies

$$\Delta t_j \simeq \frac{1}{n}\left(\frac{j}{n}\right)^{\beta-1} \simeq \Delta t_{n-j} \quad \text{for } 1 \leq j \leq n/2. \tag{2.7}$$

We construct the corresponding product mesh with n^2 elements on Γ with vertices

$$t_{(j_1,j_2)} = (t_{j_1}, t_{j_2}) \quad \text{for } 0 \leq j_1 \leq n \text{ and } 0 \leq j_2 \leq n. \tag{2.8}$$

Elements K near any corner are shape-regular with $h_K \simeq (1/n)^\beta \simeq \rho_K$. Away from the boundary they are also shape-regular with $h_K \simeq 1/n \simeq \rho_K$. However, near the middle of an edge we have $h_K \simeq 1/n$ and $\rho_K \simeq (1/n)^\beta$; hence if $\beta > 1$, then degeneracy occurs with the maximum aspect ratio for the elements growing like $n^{\beta-1}$; see Figure 1. This construction can be generalized to other polyhedral surfaces; see, e.g., [11, 5] and Example 5.5 below.

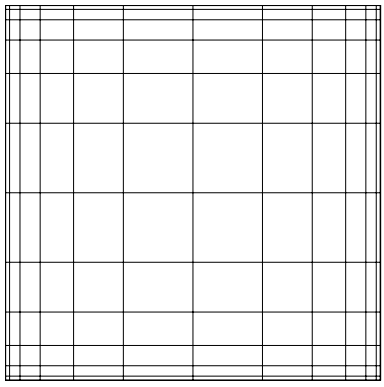


FIG. 1. *Power-graded tensor-product mesh with $\beta = 3$ and $N = 14^2$ elements.*

2.3. Condition number estimates. As an illustration of the results which we shall prove later in this paper, let us suppose we apply the Galerkin method to the weakly singular equation (2.1), with Γ given by (2.5) and with the subspace $X \subset \widetilde{H}^{-1/2}(\Gamma)$ chosen to be the space of piecewise-constant functions on the mesh (2.8). The dimension of X is $N = n^2$. In Theorems 5.7 and 7.4 we will prove that in this case the Galerkin matrix $\boldsymbol{B}$ satisfies the spectral bounds $\lambda_{\max}(\boldsymbol{B}) \lesssim N^{-1}$ and $\lambda_{\min}(\boldsymbol{B}) \gtrsim N^{-3\beta/2}$, whereas the diagonally scaled Galerkin matrix $\boldsymbol{B}'$ satisfies

$$\lambda_{\max}(\boldsymbol{B}') \lesssim N^{1/2} \times \begin{cases} 1 & \text{if } 1 \le \beta < 2, \\ (1+\log N)^{1/2} & \text{if } \beta = 2, \\ (1+\log N)^2 & \text{if } \beta > 2 \end{cases}$$

and

$$\lambda_{\min}(\boldsymbol{B}') \gtrsim \begin{cases} 1 & \text{if } \beta = 1, \\ (1+\log N)^{-1} & \text{if } \beta > 1. \end{cases}$$

Hence, $\operatorname{cond}(\boldsymbol{B})$ grows like $N^{(3\beta/2)-1}$, whereas $\operatorname{cond}(\boldsymbol{B}')$ essentially grows like $N^{1/2}$, which is the rate of growth in the case of shape-regular meshes.

On the other hand, suppose we solve (2.3), with Γ given by (2.5) and with $X \subset \widetilde{H}^{1/2}(\Gamma)$ chosen to be the space of continuous piecewise-bilinear functions on the mesh (2.8) which vanish at the boundary of Γ. The dimension of X is $N = (n-1)^2 = O(n^2)$. In Theorems 5.8 and 7.5, we shall prove that

$$\lambda_{\max}(\boldsymbol{B}) \lesssim N^{-1/2} \quad \text{and} \quad \lambda_{\min}(\boldsymbol{B}) \gtrsim \begin{cases} N^{-1} & \text{for } 1 \le \beta < 2, \\ N^{-1}(1+\log N)^{-1} & \text{for } \beta = 2, \\ N^{-\beta/2} & \text{for } \beta > 2, \end{cases}$$

whereas $\boldsymbol{B}'$ satisfies

$$\lambda_{\max}(\boldsymbol{B}') \lesssim 1 \quad \text{and} \quad \lambda_{\min}(\boldsymbol{B}') \gtrsim N^{-1/2} \times \begin{cases} 1 & \text{for } 1 \le \beta < 2, \\ (1+\log N)^{-1/2} & \text{for } \beta = 2, \\ (1+\log N)^{-2} & \text{for } \beta > 2. \end{cases}$$

Thus, in this case the condition number of $\boldsymbol{B}$ grows like $N^{(\beta-1)/2}$ (for $\beta > 2$), whereas the condition number of $\boldsymbol{B}'$ again essentially grows only like $N^{1/2}$ (for any $\beta \ge 1$), which is again the rate of growth in the shape-regular case.

3. General framework. In this section we set up the theoretical apparatus in which the general spectral estimates of section 5 will be proved. Recall that Γ denotes either a bounded (open or closed) d-dimensional Lipschitz surface in $\mathbb{R}^{d+1}$, for $d = 2$, or a bounded Lipschitz domain in $\mathbb{R}^d$, for $d = 2$ or 3.

We define the Sobolev spaces $H^s(\Gamma)$, $\widetilde{H}^s(\Gamma)$, $|s| \le 1$ in the usual way; see, for example, [9] for details. In particular, when Γ is a Lipschitz domain or an open Lipschitz surface, $u \in \widetilde{H}^1(\Gamma)$ implies that u has vanishing trace on the boundary of Γ. For $0 < s < 1$, $\widetilde{H}^s(\Gamma)$ interpolates between $L_2(\Gamma)$ and $\widetilde{H}^1(\Gamma)$. In any case, $H^{-s}(\Gamma)$ is the dual of $\widetilde{H}^s(\Gamma)$ and $\widetilde{H}^{-s}(\Gamma)$ is the dual of $H^s(\Gamma)$ for all $|s| \le 1$. When Γ is a closed surface, $\widetilde{H}^s(\Gamma) = H^s(\Gamma)$ for all $|s| \le 1$. (Higher order Sobolev spaces can be defined on domains and on smooth enough surfaces, but we do not need these here.)

In what follows we will also be interested in Sobolev norms of various functions defined over subdomains $\hat{\Gamma} \subset \Gamma$. Since different equivalent norms for $H^s(\hat{\Gamma})$ or $\widetilde{H}^s(\hat{\Gamma})$ might scale differently with the size of $\hat{\Gamma}$, we follow the notation used in [1] and write $|||u|||_{H^s(\hat{\Gamma})}$ and $|||u|||_{\widetilde{H}^s(\hat{\Gamma})}$ to indicate the specific norms obtained for $|s| \le 1$ by real interpolation and duality, starting from the usual norm in $L_2(\hat{\Gamma})$ and the Sobolev norms

$$|||u|||^2_{H^1(\hat{\Gamma})} = \|u\|^2_{L_2(\hat{\Gamma})} + |u|^2_{H^1(\hat{\Gamma})} \quad \text{and} \quad |||u|||^2_{\widetilde{H}^1(\hat{\Gamma})} = |u|^2_{H^1(\hat{\Gamma})} = \sum_{|\alpha|=1} \|\partial^\alpha u\|^2_{L_2(\hat{\Gamma})}.$$

Note that $|\cdot|_{H^1(\hat{\Gamma})}$ is only a seminorm on $H^1(\hat{\Gamma})$ but is a norm on $\widetilde{H}^1(\hat{\Gamma})$. (The distinction between $\|\cdot\|_{H^s(\hat{\Gamma})}$ and $|||\cdot|||_{H^s(\hat{\Gamma})}$ is significant only when $\hat{\Gamma}$ is a proper subset of Γ. In what follows we will freely interchange $\|\cdot\|_{H^s(\Gamma)}$ and $|||\cdot|||_{H^s(\Gamma)}$.)

For later use, we recall [11, Lemma 3.2], [1, Theorem 4.1] that *if* $\Gamma_1, \Gamma_2, \ldots, \Gamma_N$ *is a partitioning of a bounded Lipschitz domain* Γ *into nonoverlapping Lipschitz domains, then for* $|s| \le 1$,

$$|||v|||^2_{\widetilde{H}^s(\Gamma)} \le \sum_{j=1}^N |||v|||^2_{\widetilde{H}^s(\Gamma_j)} \quad \textit{and} \quad \sum_{j=1}^N |||u|||^2_{H^s(\Gamma_j)} \le |||u|||^2_{H^s(\Gamma)}. \tag{3.1}$$

We also note that (see [1, eq. (4.1)])

$$\|u\|_{H^s(\Gamma)} \lesssim \|u\|_{\widetilde{H}^s(\Gamma)} \quad \text{if } u \in \widetilde{H}^s(\Gamma) \cap L_2(\Gamma), \text{ for all } |s| \le 1, \tag{3.2}$$

and that [9, p. 320]

$$|||v|||_{\widetilde{H}^s(\hat{\Gamma})} \lesssim |||v|||^{1-s}_{L_2(\hat{\Gamma})} |||v|||^s_{\widetilde{H}^1(\hat{\Gamma})} \quad \text{and} \quad |||v|||_{\widetilde{H}^{-s}(\hat{\Gamma})} \lesssim |||v|||^{1-s}_{L_2(\hat{\Gamma})} |||v|||^s_{\widetilde{H}^{-1}(\hat{\Gamma})} \quad \text{for } 0 < s < 1. \tag{3.3}$$

As mentioned in section 1, we will be considering a family of partitions $\{\mathcal{P}\}$ of Γ. Each partition $\mathcal{P}$ consists of relatively open, pairwise-disjoint finite elements $K \subset \Gamma$ with the property $\overline{\Gamma} = \cup\{\overline{K} : K \in \mathcal{P}\}$. For each $K \in \mathcal{P}$, h_K denotes its diameter and ρ_K the diameter of the largest sphere whose intersection with $\overline{\Gamma}$ lies entirely inside $\overline{K}$. Also, for any measurable subset S of Γ, $|S|$ denotes its d-dimensional measure.

In order to impose a simple geometric character on the mesh $\mathcal{P}$, we assume that each $K \in \mathcal{P}$ is diffeomorphic to a simple reference element. More precisely, let $\hat{\sigma}^d$ denote the unit simplex and $\hat{\kappa}^d = [0,1]^d$ the unit cube in $\mathbb{R}^d$. Thus, $\hat{\sigma}^2$ is the triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$ and $\hat{\sigma}^3$ is the tetrahedron with vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$.

We assume that for each $K \in \mathcal{P}$, there exists a reference element $\hat{K} = \hat{\sigma}^d$ or $\hat{\kappa}^d$ and a bijective map $\chi_K : \hat{K} \to K$, with both χ_K and χ_K^{-1} smooth. (Here, for simplicity, "smooth" means $\mathcal{C}^\infty$.) Each element has vertices and edges, defined to be the images of the vertices and edges of the corresponding unit element under χ_K. In the 3D case, the element also has faces, comprising the images of the faces of the unit element. We assume each partition is *conforming*; i.e., for each K, $K' \in \mathcal{P}$ with $K \ne K'$, the intersection $\overline{K} \cap \overline{K'}$ is allowed to be either empty, a vertex, an edge, or (when $d = 3$) a face of both K and K'. The requirement that χ_K is smooth ensures that edges of Γ ($d = 2$) and edges of $\partial\Gamma$ ($d = 3$) are confined to edges of elements $K \in \mathcal{P}$.

Let J_K denote the $3 \times d$ Jacobian of χ_K. Then

$$\int_K f(\mathbf{x})dx = \int_{\hat{K}} f(\chi_K(\hat{\mathbf{x}}))g_K(\hat{\mathbf{x}})d\hat{x}, \quad \text{where} \quad g_K := \left(\det J_K^T J_K\right)^{1/2}. \tag{3.4}$$

In addition to the assumption that χ_K and χ_K^{-1} are smooth, we also require the following assumption on J_K.

ASSUMPTION 3.1. *There exist positive constants D, E such that*

$$D^{-1}|K|^2 \le \det\big(J_K(\hat{\mathbf{x}})^T J_K(\hat{\mathbf{x}})\big) \le D|K|^2, \tag{3.5a}$$

$$E\rho_K^2 \le \lambda_{\min}\big(J_K(\hat{\mathbf{x}})^T J_K(\hat{\mathbf{x}})\big), \tag{3.5b}$$

uniformly for $\hat{\mathbf{x}} \in \hat{K}$, $K \in \mathcal{P}$, and $\mathcal{P} \in \mathcal{F}$.

Assumption 3.1 holds, for example, when K is a planar triangle ($d = 2$) or a tetrahedron ($d = 3$) and χ_K is affine. It is also satisfied by bilinear maps from the unit square to quadrilaterals ($d = 2$), provided that the quadrilaterals are not too far from parallelograms. These and other examples are explored in [7].

Assumption 3.1 describes the quality of the maps which take the unit element $\hat{K}$ to each K. We also need assumptions on how the size and shape of neighboring elements in our mesh may vary. Here we impose *only very weak local conditions which require the meshes to be neither quasi-uniform nor shape-regular.* In addition, we need a uniform bound on the number of elements that touch the ith node $\mathbf{x}_i$.

ASSUMPTION 3.2. *There exist positive constants F, G, H and an integer M such that for all $\mathcal{P} \in \mathcal{F}$,*

$$h_K \le F\ h_{K'}\,, \quad \rho_K \le G\ \rho_{K'}\,, \quad |K| \le H\ |K'| \quad \text{for all } K, K' \in \mathcal{P} \text{ with } \overline{K} \cap \overline{K'} \ne \emptyset, \tag{3.6a}$$

$$\text{and also} \qquad \max_{i \in \mathcal{N}} \#\{K \in \mathcal{P} : \mathbf{x}_i \in \overline{K}\} \ \le\ M. \tag{3.6b}$$

Note that condition (3.6a) requires that h_K and ρ_K do not vary too rapidly between neighboring elements. This allows elements with a large aspect ratio, provided that their immediate neighbors have a comparable aspect ratio. It is clear that the power meshes (2.8) satisfy Assumption 3.2.

From now on, if $A(\mathcal{P})$ and $B(\mathcal{P})$ are two mesh-dependent quantities, then the inequality $A(\mathcal{P}) \lesssim B(\mathcal{P})$ will mean that there is a constant C independent of $\mathcal{P}$, such that $A(\mathcal{P}) \le CB(\mathcal{P})$. ($C$ may depend on D, E, F, G, or M.) Also the notation $A(\mathcal{P}) \simeq B(\mathcal{P})$ will mean that $A(\mathcal{P}) \lesssim B(\mathcal{P})$ and $B(\mathcal{P}) \lesssim A(\mathcal{P})$.

For an integer $\ell \ge 0$ and a reference element $\hat{K} \in \{\hat{\sigma}^d, \hat{\kappa}^d\}$, we define

$$\mathbb{P}^\ell(\hat{K}) = \begin{cases} \text{polynomials of total degree } \le \ell \text{ on } \hat{K} & \text{if } \hat{K} = \hat{\sigma}^d, \\ \text{polynomials of coordinate degree } \le \ell \text{ on } \hat{K} & \text{if } \hat{K} = \hat{\kappa}^d \end{cases}$$

and the finite element spaces

$$\mathcal{S}_0^\ell(\mathcal{P}) = \{u \in L^\infty(\Gamma) : u \circ \chi_K \in \mathbb{P}^\ell(\hat{K}), K \in \mathcal{P}\} \quad \text{for } \ell \ge 0,$$
$$\mathcal{S}_1^\ell(\mathcal{P}) = \{u \in C^0(\Gamma) : u \circ \chi_K \in \mathbb{P}^\ell(\hat{K}), K \in \mathcal{P}\} \quad \text{for } \ell \ge 1.$$

Finally we introduce suitable bases for these spaces. In this paper we consider standard nodal bases defined as follows. Let $d(\ell)$ denote the dimension of $\mathbb{P}^\ell(\hat{K})$

and choose a set of nodes $\{\hat{\mathbf{x}}_p : p = 1, \ldots, d(\ell)\} \subset \overline{\hat{K}}$ with the property that each $\hat{u} \in \hat{\mathbb{P}}^\ell(\hat{K})$ is uniquely determined by its values at the $\hat{\mathbf{x}}_p$. Then there are basis functions $\{\hat{\phi}_p\ , p = 1, \ldots, d(\ell)\}$ with the property $\hat{\phi}_p(\hat{\mathbf{x}}_q) = \delta_{p,q}$. From these we can define basis functions on the open set K (implicitly) by $\phi_{p,K} \circ \chi_K = \hat{\phi}_p$. We extend $\phi_{p,K}$ to $\overline{K}$ by continuity and then (discontinuously) to the whole of Γ by zero. If we introduce the nodes $\mathbf{x}_{p,K} := \chi_K(\hat{\mathbf{x}}_p) \in \overline{K}$, then clearly

$$\phi_{p,K}(\mathbf{x}_{q,K'}) = \delta_{(p,K),(q,K')} \quad \text{for} \quad p, q = 1, \ldots, d(\ell), \quad K, K' \in \mathcal{P}. \tag{3.7}$$

The functions

$$\{\phi_{p,K} : \ p = 1, \ldots d(\ell), \ K \in \mathcal{P}\} \tag{3.8}$$

then constitute a suitable basis of $\mathcal{S}_0^\ell(\mathcal{P})$. When $\ell = 0$ we have the simple piecewise constant functions, and the nodes $\mathbf{x}_K = \mathbf{x}_{1,K}$ can be chosen as the centroids of each K.

For $\mathcal{S}_1^\ell(\mathcal{P})$, we require further that if two elements K and K' share a common edge e, then this edge is parametrized *equally from both sides*. More precisely, we require that if $\chi_K^{-1}(e) = \hat{e}$ and $\chi_{K'}^{-1}(e) = \hat{e}'$, then there exists an affine mapping $\gamma : \hat{e} \to \hat{e}'$ such that χ_K and $\chi_{K'} \circ \gamma$ coincide pointwise on $\hat{e}$. We assume that the points $\mathbf{x}_{p,K}$ and $\mathbf{x}_{p,K'}$ restricted to e coincide and that the values of u at these points are sufficient to determine uniquely $u|_e$ on e. (This condition is satisfied in the simplest case when χ_K and $\chi_{K'}$ are both affine maps.) In this case any $u \in \mathcal{S}_1^\ell(\mathcal{P})$ is determined uniquely by its values at the set of global nodes $\{\mathbf{x}_{p,K} : \ p = 1, \ldots, d(\ell), \ K \in \mathcal{P}\}$, where coincident nodes on the boundary of more than one element now constitute a single degree of freedom. Denoting this set more abstractly by $\{\mathbf{x}_k : k \in \mathcal{N}\}$ for some suitable index set of nodes (or degrees of freedom) $\mathcal{N}$, our basis for $\mathcal{S}_1^\ell(\mathcal{P})$ is

$$\{\phi_k : k \in \mathcal{N}\}, \tag{3.9}$$

where $\phi_k \in \mathcal{S}_1^\ell(\mathcal{P})$ is the unique function satisfying

$$\phi_k(\mathbf{x}_{k'}) = \delta_{k,k'} \quad \text{for all } k, \, k' \in \mathcal{N}. \tag{3.10}$$

A simple example is the space of the continuous bilinear functions on a mesh of quadilaterals, with nodes chosen to be the vertices of the elements.

Clearly the basis (3.8) may be written in the abstract form (3.9) by allowing the set $\mathcal{N}$ to contain double indices of the form (p, K). With this notation, (3.10) follows from (3.7). Moreover, in any case,

$$\Gamma_k \subseteq \bigcup \left\{\overline{K} : \mathbf{x}_k \in \overline{K}\right\}, \quad \text{where } \ \Gamma_k := \operatorname{supp} \phi_k. \tag{3.11}$$

Throughout the rest of the paper N denotes the cardinality of the nodal set $\mathcal{N}$.

4. Sobolev norm of a nodal basis function. We now establish some technical estimates needed in the next section in our proofs of the spectral bounds for $\boldsymbol{B}$ and $\boldsymbol{B}'$. For these we need the following notation. For $k \in \mathcal{N}$, define

$$\begin{aligned} h_k &= \text{average of those } h_K \text{ for which } x_k \in \overline{K}, \\ \rho_k &= \text{average of those } \rho_K \text{ for which } x_k \in \overline{K}, \end{aligned}$$

and note that the second inequality in (3.6a) implies that

$$\min_{x_k\in\overline{K}}\rho_K \lesssim \rho_k \lesssim \max_{x_k\in\overline{K}}\rho_K \quad \text{for } k\in\mathcal{N}. \tag{4.1}$$

The following theorem is closely related to [7, Theorems 3.2 and 3.6].

THEOREM 4.1. *Let $\{\phi_k\}_{k\in\mathcal{N}}$ be a nodal basis for $\mathcal{S}_i^\ell(\mathcal{P})\subset\widetilde{H}^m(\Gamma)$, where $i=0$ or 1.*

(i) *If $0\le m\le 1$, then $|||\phi_k|||_{\widetilde{H}^m(\Gamma_k)}\lesssim\rho_k^{-m}\|\phi_k\|_{L_2(\Gamma_k)}$.*

(ii) *If $-1\le m\le 0$, then $\rho_k^{-m}\|\phi_k\|_{L_2(\Gamma_k)}\lesssim|||\phi_k|||_{H^m(\Gamma_k)}$.*

Proof. The proof follows the same lines as the proofs of [7, Theorems 3.2 and 3.6], in which the same result is proved on the whole domain Γ. To get the proof of the present result, one has to check only that the arguments in [7] remain true if the global norm $|||\cdot|||_{\widetilde{H}^s(\Gamma)}$ is replaced by the local norm $|||\cdot|||_{\widetilde{H}^s(\Gamma_k)}$. We recall that the latter norm is obtained for $s\in(0,1)$ by interpolation between the norms $\|\cdot\|_{L_2(\Gamma_k)}$ and $|\cdot|_{H^1(\Gamma_k)}$ and then by duality for $s\in[-1,0]$. Now, following the arguments in [7, Theorem 3.2], it is easily seen that

$$|\phi_k|^2_{H^1(\Gamma_k)}=\sum_{\substack{K\in\mathcal{P}\\K\subset\Gamma_k}}\int_K|\nabla\phi_k|^2\lesssim\rho_k^{-2}\sum_{\substack{K\in\mathcal{P}\\K\subset\Gamma_k}}\|\phi_k\|^2_{L_2(K)}=\rho_k^{-2}\|\phi_k\|^2_{L_2(\Gamma_k)}.$$

The proof of (i) for $m=1$ follows directly, and result (i) then follows by interpolation (3.3).

For (ii), note first that when $m\in[-1,0]$, the definition of the dual space implies

$$|||\phi_k|||_{H^m(\Gamma_k)}\ge\frac{|(\phi_k,w)_{\Gamma_k}|}{|||w|||_{\widetilde{H}^{-m}(\Gamma_k)}} \tag{4.2}$$

for any $w\in\widetilde{H}^{-m}(\Gamma_k)$, not identically zero. The proof is completed by constructing a test function $w\in\widetilde{H}^{-m}(\Gamma_k)$ with the properties

$$|(\phi_k,w)_{\Gamma_k}| \;\gtrsim\; \rho_k^2\|\phi_k\|^2_{L_2(\Gamma_k)}, \tag{4.3}$$

$$|||w|||_{\widetilde{H}^{-m}(\Gamma_k)} \;\lesssim\; \rho_k^{2+m}\|\phi_k\|_{L_2(\Gamma_k)}. \tag{4.4}$$

The required construction of w is given in the proof of [7, Theorem 3.6], where the estimates (4.3) and (4.4) with $m=-1$ are established (see [7, eqs. (3.14), (3.15)] and put $\alpha=0$ and $k=1$). The proof of (4.4) for $m\in[-1,0]$ is obtained by establishing it for $m=0$ and then interpolating with $m=-1$. To establish (4.4) for $m=0$ one has to look closely at the argument in [7, Theorem 3.6]. For any $K\in\mathcal{P}$, $K\subset\Gamma_k$, it is shown that there exists a subset $t(K)\subset K$ and a function $P_{t(K)}\in\widetilde{H}^1(K)$ such that $\phi_k|_{t(K)}$ is one-signed and such that $\|P_{t(K)}\|_{L_2(K)}\simeq|t(K)|^{1/2}$. The hidden constants in this estimate are independent of k,K and the mesh. Then the w which satisfies (4.3) and (4.4) with $m=-1$ is defined as $w=\sum_{K\subset\Gamma_k}\rho_K^2\operatorname{sign}(\phi_k|_{t(K)})\inf_{x\in t(K)}|\phi_k(x)|P_{t(K)}$. Then

$$\begin{aligned}\|w\|^2_{L_2(\Gamma_k)}&=\sum_{K\subset\Gamma_k}\rho_K^4\Big(\inf_{x\in t(K)}|\phi_k(x)|\Big)^2\|P_{t(K)}\|^2_{L_2(t(K))}\\&\lesssim\rho_k^4\sum_{K\subset\Gamma_k}\Big(\inf_{x\in t(K)}|\phi_k(x)|\Big)^2|t(K)|\lesssim\rho_k^4\|\phi_k\|^2_{L_2(\Gamma_k)},\end{aligned}$$

as required. ∎

Theorem 4.1 is the key component in the proof of the next result, which in turn is a partial generalization of [1, Lemma 4.7] and [1, Theorem 4.8].

THEOREM 4.2. *Let $k \in \mathcal{N}$.*

(i) *If $1 \le p \le \infty$, then $\|\phi_k\|_{L_p(\Gamma)} = \|\phi_k\|_{L_p(\Gamma_k)} \simeq |\Gamma_k|^{1/p}$.*

(ii) *If $0 \le m \le 1$, then $\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim |||\phi_k|||^2_{\widetilde{H}^m(\Gamma_k)} \lesssim |\Gamma_k|\rho_k^{-2m}$.*

(iii) *If $-1 \le m \le 0$, then $|\Gamma_k|\rho_k^{-2m} \lesssim |||\phi_k|||^2_{H^m(\Gamma_k)} \lesssim \|\phi_k\|^2_{H^m(\Gamma)}$.*

(iv) *If $0 \le 2m < d$, then $\|\phi_k\|^2_{H^m(\Gamma)} \gtrsim |\Gamma_k|^{1-2m/d}$.*

(v) *If $-d < 2m \le 0$, then $\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim |\Gamma_k|^{1-2m/d}$.*

Proof. (i) By the definition of ϕ_k,

$$\|\phi_k\|^p_{L_p(\Gamma_k)} = \sum_{\substack{K\in\mathcal{P}\\ K\subset\Gamma_k}} \int_K |\phi_k|^p.$$

For a typical $K \subset \Gamma_k$, recall (3.4) and write $\int_K |\phi_k|^p = \int_{\hat{K}} |\hat{\phi}_k|^p g_K \simeq |K| \int_{\hat{K}} |\hat{\phi}_k|^p \simeq |K|$ by Assumption 3.1. Now sum over all elements $K \subset \Gamma_k$ to obtain the result.

The left-hand inequality in (ii) follows directly from the left-hand inequality in (3.1), while the right-hand inequality in (ii) follows from part (i) of Theorem 4.1 and part (i) of the present theorem.

Similarly, in part (iii), we use the right-hand inequality in (3.1) and part (ii) of Theorem 4.1, combined with part (i) of the present theorem.

To prove (iv), we put $p = 2d/(d-2m) \in [2,\infty)$ and apply the Sobolev imbedding theorem together with part (i) above to obtain

$$\|\phi_k\|^2_{H^m(\Gamma)} \gtrsim \|\phi_k\|^2_{L_p(\Gamma)} \simeq |\Gamma_k|^{2/p} = |\Gamma_k|^{1-2m/d}.$$

Part (v) follows using a dual imbedding: For $q = 2d/(d-2m) \in (1,2]$,

$$\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \|\phi_k\|^2_{L_q(\Gamma)} \simeq |\Gamma_k|^{2/q} = |\Gamma_k|^{1-2m/d}. \qquad \square$$

5. Bounds on the extremal eigenvalues. In this section we obtain general bounds on the spectra of the matrices $\boldsymbol{B}$ and $\boldsymbol{B}'$ which were defined in (1.4) and (1.6). Since $\boldsymbol{B}$ is symmetric, these may be obtained by estimating the Rayleigh quotient $\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha}/\boldsymbol{\alpha}^T\boldsymbol{\alpha}$ from above and from below. For a typical $v \in X$ we write

$$v = \sum_{k\in\mathcal{N}} v_k, \quad \text{where } v_k = \alpha_k\phi_k \text{ and } \alpha_k = v(x_k). \tag{5.1}$$

Then, since

$$\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha} = B(v,v) \simeq \|v\|^2_{\widetilde{H}^m(\Gamma)} \quad \text{and} \quad \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \sum_{k\in\mathcal{N}} v(x_k)^2,$$

if we show the bounds

$$\lambda_X \sum_{k\in\mathcal{N}} v(x_k)^2 \lesssim \|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \Lambda_X \sum_{k\in\mathcal{N}} v(x_k)^2 \quad \text{for all } v \in X, \tag{5.2}$$

then we have the estimates $\lambda_{\max}(\boldsymbol{B}) \lesssim \Lambda_X$ and $\lambda_{\min}(\boldsymbol{B}) \gtrsim \lambda_X$.

Similarly, for the diagonally scaled matrix $\boldsymbol{B}'$, note that

$$\boldsymbol{\alpha}^T\boldsymbol{D}\boldsymbol{\alpha} = \sum_{k\in\mathcal{N}} \alpha_k^2 B(\phi_k,\phi_k) = \sum_{k\in\mathcal{N}} B(v_k,v_k) \simeq \sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^m(\Gamma)}.$$

Thus if we can show that

$$\lambda'_X \sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \Lambda'_X \sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^m(\Gamma)} \quad \text{for all } v \in X, \tag{5.3}$$

then it will follow that $\lambda_{\max}(\boldsymbol{B}') \lesssim \Lambda'_X$ and $\lambda_{\min}(\boldsymbol{B}') \gtrsim \lambda'_X$.

For each element $K \in \mathcal{P}$, let $\mathcal{N}(K) = \{\, k \in \mathcal{N} : \Gamma_k \cap K \neq \emptyset \,\}$. Our assumptions on the family of partitions $\{\mathcal{P}\}$ imply that

$$\text{the cardinality of } \mathcal{N}(K) \text{ for } K \in \mathcal{P} \text{ is bounded independently of } \mathcal{P} \tag{5.4}$$

and that for each $\mathcal{P}$ the index set $\mathcal{N}$ may be partitioned into disjoint subsets $\mathcal{N}_1$, $\mathcal{N}_2, \ldots, \mathcal{N}_L$ having the property

$$\text{interior}(\operatorname{supp}\phi_k) \cap \text{interior}(\operatorname{supp}\phi_{k'}) = \emptyset \quad \text{if } k,\, k' \in \mathcal{N}_\ell \text{ and } k \neq k', \tag{5.5}$$

in such a way that L is bounded independently of $\mathcal{P}$.

In Lemmas 5.1 and 5.2 we will obtain bounds on the spectra of $\boldsymbol{B}$ and $\boldsymbol{B}'$, some of which involve the quantities

$$\Phi_{m,k} := \frac{|\Gamma_k|\rho_k^{-2m}}{\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)}}\,, \quad k \in \mathcal{N}. \tag{5.6}$$

Simple bounds on $\Phi_{m,k}$ may be obtained by employing Theorem 4.2 and (3.2) to obtain

$$\Phi_{m,k} \lesssim 1 \text{ for } -1 \le m \le 0 \qquad \text{and} \qquad \Phi_{m,k} \gtrsim \frac{|\Gamma_k|^{2m/d}}{\rho_k^{2m}} \text{ for } -d < 2m \le 0, \tag{5.7}$$

$$\Phi_{m,k} \gtrsim 1 \text{ for } 0 \le m \le 1 \qquad \text{and} \qquad \Phi_{m,k} \lesssim \frac{|\Gamma_k|^{2m/d}}{\rho_k^{2m}} \text{ for } 0 \le 2m < d. \tag{5.8}$$

(Note that in the shape-regular case, (5.8) and (5.7) are sharp estimates, since $|\Gamma_k|^{1/d} \simeq h_k \simeq \rho_k$, and thus $\Phi_{m,k} \simeq 1$ for $2|m| < d$.)

In the next two lemmas we shall decompose an arbitrary $v \in X \subset \widetilde{H}^m(\Gamma)$ as in (5.1).

LEMMA 5.1. *For $-1 \le m \le 0$ and $-d < 2m$, we have*

$$\min_{k\in\mathcal{N}} |\Gamma_k|\rho_k^{-2m} \;\lesssim\; \lambda_{\min}(\boldsymbol{B}) \;\le\; \lambda_{\max}(\boldsymbol{B}) \;\lesssim\; \left(\sum_{k\in\mathcal{N}} |\Gamma_k|^{1-d/2m}\right)^{-2m/d},$$

$$\min_{k\in\mathcal{N}} \Phi_{m,k} \;\lesssim\; \lambda_{\min}(\boldsymbol{B}') \;\le\; \lambda_{\max}(\boldsymbol{B}') \;\lesssim\; \left(\sum_{k\in\mathcal{N}} |\Gamma_k|\rho_k^{-d}\right)^{-2m/d}.$$

The lower bounds continue to hold if the hypothesis is weakened to just $-1 \le m \le 0$.

Proof. With $q = 2d/(d-2m) \in (1,2]$, and using the dual Sobolev embedding, we obtain $\|v\|_{\widetilde{H}^m(\Gamma)} \lesssim \|v\|_{L_q(\Gamma)}$. Now, using Hölder's inequality and property (5.5), we obtain

$$\begin{aligned} \|v\|_{\widetilde{H}^m(\Gamma)} &\lesssim \|v\|_{L_q(\Gamma)} \lesssim \sum_{\ell=1}^{L} \left\| \sum_{k\in\mathcal{N}_\ell} v_k \right\|_{L_q(\Gamma)} \\ &\lesssim \left(\sum_{\ell=1}^{L} \left\| \sum_{k\in\mathcal{N}_\ell} v_k \right\|^q_{L_q(\Gamma)} \right)^{1/q} \lesssim \left(\sum_{k\in\mathcal{N}} \|v_k\|^q_{L_q(\Gamma)} \right)^{1/q}. \end{aligned} \tag{5.9}$$

Moreover from Hölder's inequality with $1/p + q/2 = 1$ and with w_k denoting any positive weight, we have

$$\sum_{k\in\mathcal{N}} \|v_k\|^q_{L_q(\Gamma)} \le \left(\sum_{k\in\mathcal{N}} \left(w_k^{q/2}\right)^p\right)^{1/p} \left(\sum_{k\in\mathcal{N}} \left(w_k^{-q/2}\|v_k\|^q_{L_q(\Gamma)}\right)^{2/q}\right)^{q/2}$$
$$\lesssim \left(\sum_{k\in\mathcal{N}} w_k^{pq/2}\right)^{1/p} \left(\sum_{k\in\mathcal{N}} w_k^{-1}\|v_k\|^2_{L_q(\Gamma)}\right)^{q/2}.$$

Combining this with (5.9), we obtain

$$\|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \left(\sum_{k\in\mathcal{N}} w_k^{pq/2}\right)^{2/(pq)} \sum_{k\in\mathcal{N}} w_k^{-1}\|v_k\|^2_{L_q(\Gamma)}. \tag{5.10}$$

Note that $2/q = 1 - 2m/d$, $p = 1 - d/(2m)$, and $pq/2 = -d/(2m)$. By choosing $w_k = |\Gamma_k|^{2/q}$ in (5.10) and applying Theorem 4.2 (i), we obtain

$$\|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \left(\sum_{k\in\mathcal{N}} |\Gamma_k|^{1-d/2m}\right)^{-2m/d} \sum_{k\in\mathcal{N}} v(x_k)^2,$$

which, together with (5.2) implies the upper bound for $\lambda_{\max}(\boldsymbol{B})$.

On the other hand, with $w_k = |\Gamma_k|^{2/q-1}\rho_k^{2m}$, it follows from Theorem 4.2(i), (iii) that

$$w_k^{-1}\|v_k\|^2_{L_q(\Gamma)} \simeq v(x_k)^2|\Gamma_k|\rho_k^{-2m} \lesssim v(x_k)^2\|\phi_k\|^2_{H^m(\Gamma)} = \|v_k\|^2_{H^m(\Gamma)}.$$

Since $w_k^{pq/2} = |\Gamma_k|^{p-pq/2}(\rho_k^{2m})^{pq/2} = |\Gamma_k|\rho_k^{-d}$, (5.10) leads to

$$\|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \left(\sum_{k\in\mathcal{N}} |\Gamma_k|\rho_k^{-d}\right)^{-2m/d} \sum_{k\in\mathcal{N}} \|v_k\|^2_{H^m(\Gamma)}.$$

The upper bound for $\lambda_{\max}(\boldsymbol{B}')$ follows by (3.2) and (5.3).

Now consider the lower bounds. Given $k \in \mathcal{N}$, choose an element $K \in \mathcal{P}$ such that $x_k \in \overline{K}$. Then, with $\hat{v} = v \circ \chi_K$, and using equivalence of norms on finite-dimensional spaces, we have

$$v(x_k)^2 \le \|v\|^2_{L_\infty(K)} = \|\hat{v}\|^2_{L_\infty(\hat{K})} \simeq \|\hat{v}\|^2_{L_2(\hat{K})} \simeq |K|^{-1}\|v\|^2_{L_2(K)}. \tag{5.11}$$

Moreover, using [7, Theorem 3.6, Remark 3.8] applied on the single element K, we obtain $\|v\|_{L_2(K)} \lesssim \rho_K^m |||v|||_{H^m(K)}$. Combining this with (5.11) and using Assumption 3.2, we get

$$v(x_k)^2 \lesssim |K|^{-1}\rho_K^{2m}|||v|||^2_{H^m(K)} \lesssim |\Gamma_k|^{-1}\rho_k^{2m}|||v|||^2_{H^m(K)}. \tag{5.12}$$

Hence using (5.4) and (3.1),

$$\sum_{k\in\mathcal{N}} v(x_k)^2 \lesssim \left(\max_{k\in\mathcal{N}} |\Gamma_k|^{-1}\rho_k^{2m}\right) \sum_{K\in\mathcal{P}} |||v|||^2_{H^m(K)} \lesssim \left(\max_{k\in\mathcal{N}} |\Gamma_k|^{-1}\rho_k^{2m}\right)\|v\|^2_{H^m(\Gamma)},$$

and the lower bound for $\lambda_{\min}(\boldsymbol{B})$ follows by (3.2). To obtain the lower bound for $\lambda_{\min}(\boldsymbol{B}')$, we use the definition (5.6) of $\Phi_{m,k}$ combined with (5.12) to obtain

$$\|v_k\|^2_{\widetilde{H}^m(\Gamma)} = v(x_k)^2\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} = \Phi_{m,k}^{-1}\left[v(x_k)^2|\Gamma_k|\rho_k^{-2m}\right] \lesssim \Phi_{m,k}^{-1}|||v|||^2_{H^m(K)}. \tag{5.13}$$

Then the required estimate follows by summing over k and using (3.1). ∎

LEMMA 5.2. *For $0 \le m \le 1$ and $2m < d$, we have*

$$\left(\sum_{k\in\mathcal{N}} |\Gamma_k|^{1-d/2m}\right)^{-2m/d} \lesssim \lambda_{\min}(\boldsymbol{B}) \le \lambda_{\max}(\boldsymbol{B}) \lesssim \max_{k\in\mathcal{N}} |\Gamma_k|\rho_k^{-2m},$$

$$\left(\sum_{k\in\mathcal{N}} |\Gamma_k|\rho_k^{-d}\right)^{-2m/d} \lesssim \lambda_{\min}(\boldsymbol{B}') \le \lambda_{\max}(\boldsymbol{B}') \lesssim \max_{k\in\mathcal{N}} \Phi_{m,k}.$$

The upper bounds continue to hold if the hypothesis is weakened to just $0 \le m \le 1$.

Proof. Using the decomposition (5.1) of v, we have

$$\begin{aligned}\|v\|^2_{\widetilde{H}^m(\Gamma)} &= \Big\|\sum_{\ell=1}^L \sum_{k\in\mathcal{N}_\ell} v_k\Big\|^2_{\widetilde{H}^m(\Gamma)} \le \left(\sum_{\ell=1}^L \Big\|\sum_{k\in\mathcal{N}_\ell} v_k\Big\|_{\widetilde{H}^m(\Gamma)}\right)^2 \\ &\le L\sum_{\ell=1}^L \Big\|\sum_{k\in\mathcal{N}_\ell} v_k\Big\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \sum_{k\in\mathcal{N}} |||v_k|||^2_{\widetilde{H}^m(\Gamma_k)},\end{aligned} \tag{5.14}$$

where we used the left-hand inequality in (3.1) and the property (5.5). By Theorem 4.2(ii),

$$|||v_k|||^2_{\widetilde{H}^m(\Gamma_k)} = v(x_k)^2|||\phi_k|||^2_{\widetilde{H}^m(\Gamma_k)} \lesssim v(x_k)^2|\Gamma_k|\rho_k^{-2m}. \tag{5.15}$$

Substituting this into (5.14) yields

$$\|v\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \left(\max_{j\in\mathcal{N}} |\Gamma_j|\rho_j^{-2m}\right) \sum_{k\in\mathcal{N}} v(x_k)^2,$$

which, recalling (5.2), implies the upper bound for $\lambda_{\max}(\boldsymbol{B})$.

To obtain the upper bound for $\lambda_{\max}(\boldsymbol{B}')$, we use (5.15) to write

$$|||v_k|||^2_{\widetilde{H}^m(\Gamma_k)} \lesssim v(x_k)^2|\Gamma_k|\rho_k^{-2m} = v(x_k)^2\Phi_{m,k}\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} = \Phi_{m,k}\|v_k\|^2_{\widetilde{H}^m(\Gamma)}.$$

Then we combine this with (5.14) and (5.3) to obtain the result.

Now we consider the lower bounds. Let $p = 2d/(d-2m) \in [2,\infty)$ so that $\|v\|_{L_p(\Gamma)} \lesssim \|v\|_{H^m(\Gamma)}$. Clearly $v(x_k)^2 \le \|v\|^2_{L_\infty(K)}$ for some $K \in \mathcal{P}$, $K \subseteq \Gamma_k$. Also $v\circ\chi_K = \hat{v}$, with $\hat{v} \in \mathbb{P}^\ell(\hat{K})$, and by equivalence of norms on finite-dimensional spaces, combined with Assumptions 3.1 and 3.2, we have

$$v(x_k)^2 \le \|v\|^2_{L_\infty(K)} = \|\hat{v}\|^2_{L_\infty(\hat{K})} \simeq \|\hat{v}\|^2_{L_p(\hat{K})} \simeq |K|^{-2/p}\|v\|^2_{L_p(K)} \lesssim |\Gamma_k|^{-2/p}\|v\|^2_{L_p(\Gamma_k)}. \tag{5.16}$$

Hence, by Hölder's inequality with $2/p + 1/q = 1$,

$$\sum_{k\in\mathcal{N}} v(x_k)^2 \lesssim \left(\sum_{j\in\mathcal{N}} \left(|\Gamma_j|^{-2/p}\right)^q\right)^{1/q} \left(\sum_{k\in\mathcal{N}} \|v\|^p_{L_p(\Gamma_k)}\right)^{2/p} \lesssim \left(\sum_{j\in\mathcal{N}} |\Gamma_j|^{-2q/p}\right)^{1/q} \|v\|^2_{L_p(\Gamma)},$$

where we used the property (5.4). Now, since $2/p = 1 - 2m/d$ we have $1/q = 2m/d$ and

$$\sum_{k\in\mathcal{N}} v(x_k)^2 \lesssim \left(\sum_{j\in\mathcal{N}} |\Gamma_j|^{1-d/2m}\right)^{2m/d} \|v\|^2_{H^m(\Gamma)},$$

which, in view of (3.2) and (5.2), proves the lower bound for $\lambda_{\min}(\boldsymbol{B})$.

To estimate $\lambda_{\min}(\boldsymbol{B}')$, we use Theorem 4.2(ii) and (5.16) to obtain

$$\|v_k\|^2_{\widetilde{H}^m(\Gamma)} = v(x_k)^2\|\phi_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim v(x_k)^2|\Gamma_k|\rho_k^{-2m} \lesssim |\Gamma_k|^{1-2/p}\rho_k^{-2m}\|v\|^2_{L_p(\Gamma_k)}. \tag{5.17}$$

Thus, recalling $1 - 2/p = 2m/d$ and employing again (5.4),

$$\begin{aligned}&\sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^m(\Gamma)} \lesssim \sum_{k\in\mathcal{N}} |\Gamma_k|^{2m/d}\rho_k^{-2m}\|v\|^2_{L_p(\Gamma_k)}\\ &\lesssim \left(\sum_{j\in\mathcal{N}} \left(|\Gamma_j|^{2m/d}\rho_j^{-2m}\right)^q\right)^{1/q}\left(\sum_{k\in\mathcal{N}} \|v\|^p_{L_p(\Gamma_k)}\right)^{2/p} \lesssim \left(\sum_{j\in\mathcal{N}} |\Gamma_j|\rho_j^{-d}\right)^{2m/d}\|v\|^2_{H^m(\Gamma)},\end{aligned}$$

which, again using (3.2) and (5.3), gives the lower bound for $\lambda_{\min}(\boldsymbol{B}')$. ∎

REMARK 5.3. *The left-hand side of the first inequality in Lemma* 5.2 *and the right-hand side of the first inequality in Lemma* 5.1 *should be interpreted as the appropriate limit when* $m \to 0$. *Observe that these lemmas reproduce several known results as special cases. For example, putting* $m = 0$, *we obtain estimates for the "mass matrix" corresponding to an operator of order* 0*:*

$$\min_{k\in\mathcal{N}} |\Gamma_k| \lesssim \lambda_{\min}(\boldsymbol{B}) \le \lambda_{\max}(\boldsymbol{B}) \lesssim \max_{k\in\mathcal{N}} |\Gamma_k| \quad \textit{and} \quad 1 \lesssim \lambda_{\min}(\boldsymbol{B}') \le \lambda_{\max}(\boldsymbol{B}') \lesssim 1.$$

These are well-known, at least for the shape-regular case (i.e., $|\Gamma_k| \sim \rho_k^d \sim h_k^d$*). Moreover, in the shape-regular case for general* m *and* d, *it is easy to see that Lemmas* 5.2 *and* 5.1 *imply the previously proved estimate* (1.7). *Lemmas* 5.2 *and* 5.1 *can be combined with* (5.8) *and* (5.7) *to obtain general spectral estimates in the non–shape-regular case, in terms of the computable quantities* $|\Gamma_k|$ *and* ρ_k *and generic mesh-independent constants. In what follows we shall illustrate the uses of these estimates for operators of general order* m, *but (since boundary integral equations is our main application), we shall restrict this illustration to the case*

$$d = 2 \quad \textit{and} \quad |\Gamma_k| \sim h_k\rho_k \quad \textit{for each} \quad k \in \mathcal{N}. \tag{5.18}$$

Then we have the following corollary for operators of general order m.

COROLLARY 5.4. (i) *Assume* (5.18). *For* $-1 < m \le 0$,

$$\begin{aligned}\min_{k\in\mathcal{N}}(h_k\rho_k^{1-2m}) \;\lesssim\; \lambda_{\min}(\boldsymbol{B}) \;\le\; \lambda_{\max}(\boldsymbol{B}) \;\lesssim\; \left(\sum_{k\in\mathcal{N}} (h_k\rho_k)^{1-1/m}\right)^{-m},\\ \min_{k\in\mathcal{N}}(h_k\rho_k^{-1})^m \;\lesssim\; \lambda_{\min}(\boldsymbol{B}') \;\le\; \lambda_{\max}(\boldsymbol{B}') \;\lesssim\; \left(\sum_{k\in\mathcal{N}} h_k\rho_k^{-1}\right)^{-m}.\end{aligned}$$

(ii) *For* $0 \le m < 1$,

$$\left(\sum_{k\in\mathcal{N}} (h_k\rho_k)^{1-1/m}\right)^{-m} \lesssim \lambda_{\min}(\boldsymbol{B}) \le \lambda_{\max}(\boldsymbol{B}) \lesssim \max_{k\in\mathcal{N}} \left(h_k\rho_k^{1-2m}\right),$$
$$\left(\sum_{k\in\mathcal{N}} h_k\rho_k^{-1}\right)^{-m} \lesssim \lambda_{\min}(\boldsymbol{B}') \le \lambda_{\max}(\boldsymbol{B}') \lesssim \max_{k\in\mathcal{N}} \left(h_k\rho_k^{-1}\right)^m.$$

The hypersingular and weakly singular examples considered in section 2.1 are then obtained from the special cases $m = 1/2$ and $m = -1/2$. These estimates can be applied to any mesh specified by the user. To illustrate its use on a typical class of meshes, let us consider the following example.

EXAMPLE 5.5. *Suppose* Γ *is a plane convex polygon with perimeter* γ. *For some fixed* $\delta > 0$, *let* $\gamma^{\parallel}$ *be the inscribed polygon, each of whose edges* $e^{\parallel}$ *is parallel to and a perpendicular distance* δ *from a corresponding edge* e *of* γ. *To mesh* Γ, *extend each* $e^{\parallel}$ *in a straight line at each end until it touches* γ. *For* δ *sufficiently small, this subdivides* Γ *into near-vertex rhombi, near-edge trapezia, and an inner polygon. For each* e, *draw* $n-1$ *parallel lines inside* Γ, *a perpendicular distance* $(i/n)^\beta$ *from* e, *for* $i = 1, \ldots, n-1$. *(The last of these lines is an extension of* $e^{\parallel}$.*) This defines a mesh of quadrilaterals on each of the rhombi. For each near-edge trapezium, introduce a quadrilateral mesh by subdividing each of its parallel sides uniformly with* n *subintervals and draw straight lines between corresponding points. Finally subdivide the interior polygon* $\Gamma^{\rm int}$ *with a quasi-uniform mesh with* $O(n^2)$ *(triangular or quadrilateral) elements, whose nodes on* γ *coincide with the nodes already specified. (See Figure* 2.*) This mesh has* $N = O(n^2)$ *elements, and the mesh in Figure* 1 *is a particular case.*

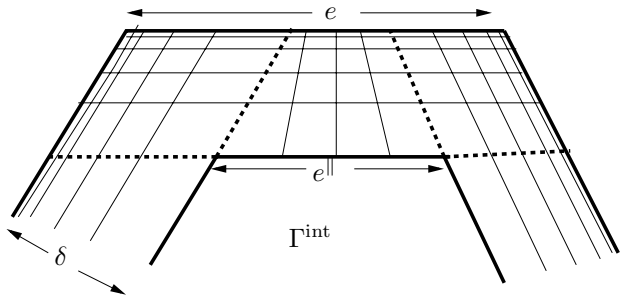


FIG. 2. *Section of mesh on polygon* Γ, *depicting two near-vertex rhombi and a near-edge trapezium.*

LEMMA 5.6. *For the class of meshes specified in Example* 5.5 *we have*

$$\sum_{k\in\mathcal{N}} h_k\rho_k^{-1} \lesssim \begin{cases} N & \text{for } 1 \le \beta < 2, \\ N(1+\log N) & \text{for } \beta = 2, \\ N^{\beta/2} & \text{for } \beta > 2, \end{cases}$$

(5.19)

$$\sum_{k\in\mathcal{N}} (\rho_k h_k)^{-1} \lesssim \begin{cases} N^2, & \text{for } 1 \le \beta < 2, \\ N^2(1+\log N)^2, & \beta = 2, \\ N^\beta, & \text{for } \beta > 2. \end{cases}$$

Proof. For the quasi-uniform mesh on $\Gamma^{\rm int}$ the required estimates follow from the standard inequalities $\rho_k \gtrsim h_k \gtrsim N^{-1/2}$. Therefore we have to consider only the near-vertex rhombi and the near-edge trapezia.

Any typical near-vertex rhombus is the image of the unit square $[0,1]^2$ under an invertible affine map. Moreover, the mesh on any near-vertex rhombus can be obtained by applying this affine map to the tensor product mesh with vertices $\big((i/n)^\beta, (j/n)^\beta\big)$. Without loss of generality, we can estimate the quantities (5.19) for this unit square because mesh-independent constants are not important. In this case, with $t_j = (j/n)^\beta$ we have $\Delta t_j = t_j - t_{j-1} \simeq n^{-1}(j/n)^{\beta-1}$ and so we have

$$\begin{aligned}\sum_{k\in\mathcal{N}} h_k \rho_k^{-1} &\simeq \sum_{i=1}^n \sum_{j=1}^i \frac{\Delta t_i}{\Delta t_j} \simeq \sum_{i=1}^n \sum_{j=1}^i \frac{(i/n)^{\beta-1}}{(j/n)^{\beta-1}} \\ &= n^2 \sum_{i=1}^n \left(\frac{i}{n}\right)^{\beta-1} \frac{1}{n} \sum_{j=1}^i \left(\frac{j}{n}\right)^{1-\beta} \frac{1}{n} \simeq n^2 \int_{1/n}^1 s^{\beta-1} \int_{1/n}^s t^{1-\beta}\, dt\, ds,\end{aligned}$$

from which the left-hand inequality in (5.19) follows (on recalling that $N \sim n^2$). Similarly,

$$\begin{aligned}\sum_{k\in\mathcal{N}} (\rho_k h_k)^{-1} &\simeq \sum_{i=1}^n \sum_{j=1}^n \frac{1}{\Delta t_i\, \Delta t_j} \simeq \sum_{i=1}^n \sum_{j=1}^i n^2 \left(\frac{i}{n}\right)^{1-\beta} \left(\frac{j}{n}\right)^{1-\beta} \\ &\simeq n^4 \sum_{i=1}^n \left(\frac{i}{n}\right)^{1-\beta} \frac{1}{n} \sum_{j=1}^i \left(\frac{j}{n}\right)^{1-\beta} \frac{1}{n} \simeq n^4 \int_{1/n}^1 s^{1-\beta} \int_{1/n}^s t^{1-\beta}\, dt\, ds,\end{aligned}$$

from which the right-hand inequality in (5.19) follows.

The meshes on the near-edge trapezia can be obtained as images of the unit square under a nonsingular bilinear map, meshed with the tensor-product mesh $\big(i/n, (j/n)^\beta\big)$, and the estimates (5.19) are then obtained analogously to those above. ∎

The following theorems now follow by combining Corollary 5.4 with Lemma 5.6.

THEOREM 5.7. *Consider the weakly singular boundary integral equation* (2.1) *on the polygon* Γ *discretized as in Example* 5.5. *Then for conforming boundary elements of any degree in* $\widetilde{H}^{-1/2}(\Gamma)$ *and with the nodal basis introduced in section* 3,

1. *the Galerkin matrix* $\boldsymbol{B}$ *satisfies the spectral bounds* $\lambda_{\max}(\boldsymbol{B}) \lesssim N^{-1}$ *and* $\lambda_{\min}(\boldsymbol{B}) \gtrsim N^{-3\beta/2}$, *and*
2. *the diagonally scaled Galerkin matrix* $\boldsymbol{B}'$ *satisfies*

$$\lambda_{\max}(\boldsymbol{B}') \lesssim \begin{cases} N^{1/2} & \text{for } 1 \le \beta < 2, \\ N^{1/2}(1+\log N)^{1/2} & \text{for } \beta = 2, \\ N^{\beta/4} & \text{for } \beta > 2, \end{cases} \qquad \text{and} \quad \lambda_{\min}(\boldsymbol{B}') \gtrsim N^{-(\beta-1)/4}.$$

Proof. Elementary estimates for the meshes in Example 5.5 yield, for each $k \in \mathcal{N}$,

$$N^{-\beta/2} \lesssim \rho_k \le h_k \lesssim N^{-1/2}. \tag{5.20}$$

We apply Corollary 5.4 with $m = -1/2$. The bounds for $\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\min}(\boldsymbol{B})$ and the lower bound for $\lambda_{\min}(\boldsymbol{B}')$ follow immediately from (5.20), whereas the upper bound for $\lambda_{\max}(\boldsymbol{B}')$ follows from Lemma 5.6. ∎

THEOREM 5.8. *Consider the hypersingular boundary integral equation* (2.3) *on the polygonal screen* Γ *discretized as in Example* 5.5. *For conforming boundary elements of any degree in* $\widetilde{H}^{1/2}(\Gamma)$ *and with the nodal basis introduced in section* 3,

1. *the Galerkin matrix* $\boldsymbol{B}$ *satisfies the spectral bounds* $\lambda_{\max}(\boldsymbol{B}) \lesssim N^{-1/2}$ *and*

$$\lambda_{\min}(\boldsymbol{B}) \gtrsim \begin{cases} N^{-1} & \text{for } 1 \le \beta < 2, \\ N^{-1}(1+\log N)^{-1} & \text{for } \beta = 2, \\ N^{-\beta/2} & \text{for } \beta > 2, \end{cases}$$

2. *the diagonally scaled Galerkin matrix* $\boldsymbol{B}'$ *satisfies* $\lambda_{\max}(\boldsymbol{B}') \lesssim N^{(\beta-1)/4}$ *and*

$$\lambda_{\min}(\boldsymbol{B}') \gtrsim \begin{cases} N^{-1/2} & \text{for } 1 \le \beta < 2, \\ N^{-1/2}(1+\log N)^{-1/2} & \text{for } \beta = 2, \\ N^{-\beta/4} & \text{for } \beta > 2. \end{cases}$$

Proof. Note that the condition that the finite element space is $\widetilde{H}^{1/2}(\Gamma)$-conforming implies that it must be chosen from the class $\{v \in \mathcal{S}_1^\ell(\mathcal{P}) : v|_{\partial\Gamma} = 0\}$ for some $\ell \ge 1$.

We apply Corollary 5.4 with $m = 1/2$. The upper bounds for $\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\max}(\boldsymbol{B}')$ follow immediately from (5.20), whereas the lower bounds follow from Lemma 5.6. ∎

6. Numerical experiments. In this section we report some numerical experiments with the integral equations from section 2 on the square screen (2.5), with the power-graded meshes (2.8).

First we consider the weakly singular equation discretized using piecewise-constant basis functions. For $\beta = 2$ and $\beta = 3$, Tables 1 and 2 show the extremal eigenvalues and the condition numbers of $\boldsymbol{B}$ and of $\boldsymbol{B}'$. From one row of the table to the next, the number of subintervals along each axis doubles, so the number of degrees of freedom N increases by a factor of 4. For each of the six quantities under investigation, the left-hand column shows the quantity itself whereas the right-hand column gives the apparent exponent μ such that the quantity is proportional to N^μ. (To compute μ, we simply divide the logarithm of the ratio of successive values by $\log 4$.) The observed exponent values indicate that the estimates of Theorem 5.7 are sharp for $\boldsymbol{B}$ but not for $\boldsymbol{B}'$. However, the improved spectral bounds for $\boldsymbol{B}'$, proved

TABLE 1
Weakly singular integral equation (2.1) *on the screen* (2.5) *with* $\beta = 2$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\operatorname{cond}(\boldsymbol{B})$	
64	9.89E-02	-0.896	7.54E-05	-2.991	1.31E+03	2.095
256	2.57E-02	-0.973	1.18E-06	-3.000	2.18E+04	2.026
1024	6.48E-03	-0.993	1.84E-08	-3.000	3.52E+05	2.007
4096	1.62E-03	-0.998	2.88E-10	-3.000	5.64E+06	2.002
16384	4.06E-04	-1.000	4.50E-12	-3.000	9.03E+07	2.000
Theorem 5.7	$\lesssim N^{-1}$		$\gtrsim N^{-3}$		$\lesssim N^2$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\operatorname{cond}(\boldsymbol{B}')$	
64	7.09E+00	0.477	3.53E-01	-0.060	2.01E+01	0.537
256	1.40E+01	0.492	3.20E-01	-0.071	4.38E+01	0.563
1024	2.79E+01	0.497	2.71E-01	-0.121	1.03E+02	0.618
4096	5.58E+01	0.499	2.31E-01	-0.115	2.42E+02	0.615
16384	1.12E+02	0.500	1.99E-01	-0.105	5.59E+02	0.605
Theorem 5.7	$\lesssim N^{1/2}(1+\log N)^{1/2}$		$\gtrsim N^{-1/4}$		$\lesssim N^{3/4}(1+\log N)^{1/2}$	
Theorem 7.4	$\lesssim N^{1/2}(1+\log N)^{1/2}$		$\gtrsim (1+\log N)^{-1}$		$\lesssim N^{1/2}(1+\log N)^{3/2}$	

TABLE 2
Weakly singular integral equation (2.1) *on the screen* (2.5) *with* $\beta = 3$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\mathrm{cond}(\boldsymbol{B})$	
64	1.78E-01	-0.755	1.28E-06	-4.496	1.39E+05	3.741
256	4.90E-02	-0.932	2.50E-09	-4.500	1.96E+07	3.568
1024	1.26E-02	-0.982	4.89E-12	-4.500	2.57E+09	3.518
4096	3.16E-03	-0.995	9.54E-15	-4.500	3.31E+11	3.505
16384	7.91E-04	-0.999	1.86E-17	-4.500	4.25E+13	3.502
Theorem 5.7	$\lesssim N^{-1}$		$\gtrsim N^{-9/2}$		$\lesssim N^{7/2}$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\mathrm{cond}(\boldsymbol{B}')$	
64	6.43E+00	0.463	3.25E-01	-0.096	1.98E+01	0.559
256	1.26E+01	0.488	2.45E-01	-0.204	5.16E+01	0.692
1024	2.51E+01	0.496	1.86E-01	-0.200	1.35E+02	0.696
4096	5.02E+01	0.499	1.48E-01	-0.165	3.39E+02	0.663
16384	1.00E+02	0.499	1.22E-01	-0.137	8.21E+02	0.637
Theorem 5.7	$\lesssim N^{3/4}$		$\gtrsim N^{-1/2}$		$\lesssim N^{5/4}$	
Theorem 7.4	$\lesssim N^{1/2}(1+\log N)^2$		$\gtrsim (1+\log N)^{-1}$		$\lesssim N^{1/2}(1+\log N)^3$	

later in Theorem 7.4, appear to be sharp up to logarithmic factors. We remark that $\beta = 3$ gives the optimal convergence rate $O(N^{-3})$ for the capacitance of Γ, when piecewise constant elements are used (see, e.g., [6]).

Our second experiment is for the hypersingular equation discretized using continuous piecewise-bilinear basis functions. Tables 3 and 4 give our results for $\beta = 2$ and $\beta = 3$, which indicate that the estimates in Theorem 5.8 are (essentially) sharp for $\boldsymbol{B}$ but not for $\boldsymbol{B}'$. However, the improved spectral bounds for $\boldsymbol{B}'$, proved later in Theorem 7.5, appear to be sharp up to logarithmic factors.

TABLE 3
Hypersingular integral equation (2.3) *on the screen* (2.5) *with* $\beta = 2$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\mathrm{cond}(\boldsymbol{B})$	
49	1.00E-01	-0.342	2.27E-02	-0.798	4.41E+00	0.456
225	5.78E-02	-0.397	5.69E-03	-1.000	1.02E+01	0.602
961	3.14E-02	-0.440	1.42E-03	-1.000	2.21E+01	0.560
3969	1.65E-02	-0.465	3.56E-04	-1.000	4.63E+01	0.535
16129	8.49E-03	-0.479	8.89E-05	-1.000	9.54E+01	0.521
Theorem 5.8	$\lesssim N^{-1/2}$		$\gtrsim N^{-1}(1+\log N)^{-1}$		$\lesssim N^{1/2}(1+\log N)$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\mathrm{cond}(\boldsymbol{B}')$	
49	1.64E+00	0.156	5.97E-01	0.055	2.74E+00	0.101
225	1.84E+00	0.084	4.28E-01	-0.240	4.29E+00	0.323
961	1.94E+00	0.041	2.18E-01	-0.487	8.92E+00	0.528
3969	1.99E+00	0.018	1.09E-01	-0.497	1.82E+01	0.515
16129	2.02E+00	0.008	5.48E-02	-0.499	3.68E+01	0.508
Theorem 5.8	$\lesssim N^{1/4}$		$\gtrsim N^{-1/2}(1+\log N)^{-1/2}$		$\lesssim N^{3/4}(1+\log N)^{1/2}$	
Theorem 7.5	$\lesssim 1$		$\gtrsim N^{-1/2}(1+\log N)^{-1/2}$		$\lesssim N^{1/2}(1+\log N)^{1/2}$	

TABLE 4
Hypersingular integral equation (2.3) *on the screen* (2.5) *with* $\beta = 3$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\operatorname{cond}(\boldsymbol{B})$	
49	1.38E-01	-0.348	1.52E-02	-1.189	9.12E+00	0.840
225	8.47E-02	-0.354	1.90E-03	-1.500	4.47E+01	1.146
961	4.72E-02	-0.422	2.37E-04	-1.500	1.99E+02	1.078
3969	2.51E-02	-0.456	2.96E-05	-1.500	8.46E+02	1.044
16129	1.30E-02	-0.473	3.70E-06	-1.500	3.51E+03	1.027
Theorem 5.8	$\lesssim N^{-1/2}$		$\gtrsim N^{-3/2}$		$\lesssim N$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\operatorname{cond}(\boldsymbol{B}')$	
49	1.68E+00	0.104	5.40E-01	0.043	3.11E+00	0.061
225	1.87E+00	0.078	4.30E-01	-0.165	4.36E+00	0.243
961	1.96E+00	0.033	2.72E-01	-0.330	7.21E+00	0.363
3969	2.00E+00	0.015	1.37E-01	-0.494	1.46E+01	0.509
16129	2.02E+00	0.006	6.87E-02	-0.498	2.94E+01	0.505
Theorem 5.8	$\lesssim N^{1/2}$		$\gtrsim N^{-1/2}(1+\log N)^{-2}$		$\lesssim N(1+\log N)^2$	
Theorem 7.5	$\lesssim 1$		$\gtrsim N^{-1/2}(1+\log N)^{-2}$		$\lesssim N^{1/2}(1+\log N)^2$	

The remainder of the paper is devoted to explaining our numerical results for $\boldsymbol{B}'$.

7. Sharper results for special cases.

7.1. Improved spectral bounds for $\boldsymbol{B}'$. For each of the model problems of section 2, the observed rate of growth for $\operatorname{cond}(\boldsymbol{B}')$ is slower than the rate predicted by the results proved in section 5 if the mesh grading is sufficiently strong, more precisely, if $\beta > 2$. However, the next lemma leads to bounds that are sharp to within logarithmic factors for $\lambda_{\max}(\boldsymbol{B}')$ (weakly singular case) and for $\lambda_{\min}(\boldsymbol{B}')$ (hypersingular case). Recall that γ denotes the perimeter of the open surface Γ.

LEMMA 7.1. *Let* $d_k = \sup_{x\in\Gamma_k} \operatorname{dist}(x,\gamma)$ *and assume that* $d_{\min} := \min_{k\in\mathcal{N}} d_k$ *is sufficiently small.*

1. *For the weakly singular boundary integral equation* (2.1) *on an open surface discretized with conforming finite elements of any degree in* $\widetilde{H}^{-1/2}(\Gamma)$,
$$\lambda_{\max}(\boldsymbol{B}') \lesssim \left(\log\frac{1}{d_{\min}}\right)^2 \max_{j\in\mathcal{N}} \frac{d_j}{\rho_j}.$$
2. *For the hypersingular boundary integral equation* (2.3) *on an open surface discretized with conforming finite elements of any degree in* $\widetilde{H}^{1/2}(\Gamma)$,
$$\lambda_{\min}(\boldsymbol{B}') \gtrsim \left(\log\frac{1}{d_{\min}}\right)^{-2} \min_{j\in\mathcal{N}} \frac{\rho_j}{d_j}.$$

Proof. First we prove part 2. Let $v \in X \subset \widetilde{H}^{1/2}(\Gamma)$ and decompose v as in (5.1). Taking $p = 2$ in (5.17) we have

$$\|v_k\|^2_{\widetilde{H}^m(\Gamma)} \le \rho_k^{-2m} \|v\|^2_{L_2(\Gamma_k)} \quad \text{for } 0 \le m \le 1 \text{ and } 0 \le 2m < d. \tag{7.1}$$

Now define $w(x)$ by $w(x) = \operatorname{dist}(x,\gamma)$. It can be shown [9, Lemma 3.32] that

$$\|vw^{-s}\|^2_{L_2(\Gamma)} = \int_\Gamma w(x)^{-2s} v(x)^2\,dx \lesssim \frac{1}{(\frac12 - s)^2} \|v\|^2_{H^s(\Gamma)} \quad \text{for } \tfrac14 \le s < \tfrac12, \tag{7.2}$$

where the hidden constant is independent of $s \in [\frac{1}{4}, \frac{1}{2})$. Since $w(x) \lesssim d_k$ for $x \in \Gamma_k$, using (7.1) with $m = 1/2$, we obtain

$$\|v_k\|^2_{\widetilde{H}^{1/2}(\Gamma)} \lesssim \rho_k^{-1}\|v\|^2_{L_2(\Gamma_k)} \lesssim \rho_k^{-1}\|(d_k/w)^s v\|^2_{L_2(\Gamma_k)} = \frac{d_k^{2s}}{\rho_k}\|vw^{-s}\|^2_{L_2(\Gamma_k)}, \quad s > 0.$$

Hence, using (7.2), we have

$$\begin{aligned}\sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^{1/2}(\Gamma)} &\lesssim \left(\max_{j\in\mathcal{N}} \frac{d_j^{2s}}{\rho_j}\right) \sum_{k\in\mathcal{N}} \|vw^{-s}\|^2_{L_2(\Gamma_k)} \\ &\lesssim \left(\max_{j\in\mathcal{N}} \frac{d_j^{2s}}{\rho_j}\right)\|vw^{-s}\|^2_{L_2(\Gamma)} \lesssim \frac{1}{(\frac{1}{2}-s)^2}\left(\max_{j\in\mathcal{N}} \frac{d_j^{2s}}{\rho_j}\right)\|v\|^2_{H^s(\Gamma)},\end{aligned}$$

where the hidden constants are independent of $s \in [\frac{1}{4}, \frac{1}{2})$. Now with $\epsilon := (\log 1/d_{\min})^{-1}$, it follows that $d_j^{-2\epsilon} \lesssim 1$ for all $j \in \mathcal{N}$. Hence putting $s = \frac{1}{2} - \epsilon$, we obtain

$$\sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^{1/2}(\Gamma)} \lesssim \left(\log \frac{1}{d_{\min}}\right)^2 \left(\max_{j\in\mathcal{N}} \frac{d_j}{\rho_j}\right)\|v\|^2_{H^{1/2}(\Gamma)}.$$

The estimate in part 2 follows at once.

To prove part 1, we apply a duality argument. Suppose $\frac{1}{4} \le s < \frac{1}{2}$. Then applying Cauchy–Schwarz together with (7.2) we obtain

$$\big|\langle v, \psi\rangle_{L_2(\Gamma)}\big| = \big|\langle vw^s, \psi w^{-s}\rangle_{L_2(\Gamma)}\big| \lesssim \|vw^s\|_{L_2(\Gamma)}\|\psi w^{-s}\|_{L_2(\Gamma)} \lesssim \frac{1}{\frac{1}{2}-s}\|vw^s\|_{L_2(\Gamma)}\|\psi\|_{H^s(\Gamma)},$$

and hence $\|v\|_{\widetilde{H}^{-s}(\Gamma)} \lesssim (\frac{1}{2}-s)^{-1}\|vw^s\|_{L_2(\Gamma)}$. Recalling (5.1) and (5.4), this implies that

(7.3)
$$\|v\|^2_{\widetilde{H}^{-1/2}(\Gamma)} \lesssim \|v\|^2_{\widetilde{H}^{-s}(\Gamma)} \lesssim (\tfrac{1}{2}-s)^{-2}\|vw^s\|^2_{L_2(\Gamma)} \lesssim (\tfrac{1}{2}-s)^{-2}\sum_{k\in\mathcal{N}} \|v_k w^s\|^2_{L_2(\Gamma)}.$$

Taking $p = 2$ in Theorem 4.2(i), and then using Theorem 4.2(iii) and (3.2), we see that

$$\begin{aligned}\|v_k w^s\|^2_{L_2(\Gamma)} &\lesssim d_k^{2s}\alpha_k^2\|\phi_k\|^2_{L_2(\Gamma)} \simeq \frac{d_k^{2s}}{\rho_k}\,\alpha_k^2|\Gamma_k|\rho_k \\ &\lesssim \frac{d_k^{2s}}{\rho_k}\,\alpha_k^2\|\phi_k\|^2_{\widetilde{H}^{-1/2}(\Gamma)} = \frac{d_k^{2s}}{\rho_k}\,\|v_k\|^2_{\widetilde{H}^{-1/2}(\Gamma)}, \quad \text{where } \alpha_k = v(x_k).\end{aligned} \tag{7.4}$$

So, combining (7.3) and (7.4) and putting $s = \frac{1}{2} - \epsilon$ with $\epsilon = (\log 1/d_{\min})^{-1}$, as above, we obtain

$$\|v\|^2_{\widetilde{H}^{-1/2}(\Gamma)} \lesssim \left(\log \frac{1}{d_{\min}}\right)^2 \left(\max_{j\in\mathcal{N}} \frac{d_j}{\rho_j}\right) \sum_{k\in\mathcal{N}} \|v_k\|^2_{\widetilde{H}^{-1/2}(\Gamma)},$$

which proves part 1. ∎

7.2. Improved bounds for $\Phi_{m,k}$. For the rest of the paper, we restrict our attention to piecewise-constant and continuous piecewise-bilinear basis functions on power-graded tensor-product meshes as defined in section 2.2.

We saw in (5.7) that $\Phi_{m,k} \lesssim 1$ if $-1 \le m \le 0$. The next lemma gives a sharp two-sided bound for the special case that occurs in our numerical experiments.

LEMMA 7.2. *For the piecewise-constant nodal basis on a rectangular mesh,*

$$1 \gtrsim \Phi_{-1/2,k} \gtrsim \frac{1}{1+\log(h_k/\rho_k)}.$$

Proof. We may assume without loss of generality that $\Gamma_k = [-h_k/2, h_k/2] \times [-\rho_k/2, \rho_k/2]$. For brevity we omit the subscript k for the remainder of the proof. Define the one-dimensional (1D) piecewise-constant basis function

$$\psi(x,h) = \begin{cases} 1 & \text{for } -h/2 < x < h/2, \\ 0 & \text{otherwise,} \end{cases}$$

and write the tensor-product basis function as $\phi(x) = \psi(x_1,h)\psi(x_2,\rho)$. Recalling (5.7), we see that the result will follow from the upper bound

$$\|\phi\|^2_{\widetilde{H}^{-1/2}(\Gamma)} \lesssim h\rho^2\left(1+\log\frac{h}{\rho}\right). \tag{7.5}$$

Denote the 2D Fourier transform of ϕ by

$$\hat\phi(\xi) = \int_{\mathbb{R}^2} e^{-i2\pi\xi\cdot x}\phi(x)\,dx = \hat\psi(\xi_1,h)\hat\psi(\xi_2,\rho),$$

where $\hat\psi$, the 1D Fourier transform of ψ, is given by

$$\hat\psi(\xi_1,h) = \int_{-h/2}^{h/2} e^{-i2\pi\xi_1 x_1}\,dx_1 = h\,\mathrm{sinc}(\xi_1 h), \qquad \mathrm{sinc}(z) = \begin{cases} \dfrac{\sin\pi z}{\pi z}, & z \ne 0, \\ 1, & z = 0. \end{cases}$$

Note that $|\hat\psi(\xi,h)| \le \min(h, |\xi|^{-1})$. We have the norm equivalence

$$\begin{aligned} \|\phi\|^2_{\widetilde{H}^{-1/2}(\Gamma)} = \|\phi\|^2_{\widetilde{H}^{-1/2}(\mathbb{R}^2)} &\simeq \int_{\mathbb{R}^2} (1+|\xi|^2)^{-1/2}|\hat\phi(\xi)|^2\,d\xi \\ &= \int_{-\infty}^{\infty} |\hat\psi(\xi_1,h)|^2 \int_{-\infty}^{\infty} (1+\xi_1^2+\xi_2^2)^{-1/2}|\hat\psi(\xi_2,\rho)|^2\,d\xi_2\,d\xi_1 \\ &=: I_1 + I_2 + I_3 + I_4 + I_5, \end{aligned}$$

with

$$I_1 = \int_{\substack{-\infty<\xi_1<\infty,\\ |\xi_2|>\rho^{-1}}}, \quad I_2 = \int_{\substack{|\xi_1|<h^{-1},\\ |\xi_2|<h^{-1}}}, \quad I_3 = \int_{\substack{|\xi_1|<h^{-1}\\ h^{-1}<|\xi_2|<\rho^{-1}}},$$

$$I_4 = \int_{\substack{|\xi_1|>h^{-1},\\ |\xi_2|<h^{-1}}}, \quad I_5 = \int_{\substack{|\xi_1|>h^{-1}\\ h^{-1}<|\xi_2|<\rho^{-1}}}.$$

By Plancherel's theorem,

$$\begin{aligned} I_1 &\le \int_{-\infty}^{\infty} |\hat\psi(\xi_1,h)|^2\,d\xi_1 \int_{|\xi_2|>\rho^{-1}} |\xi_2|^{-1}|\hat\psi(\xi_2,\rho)|^2\,d\xi_2 \\ &\le 2\int_{-\infty}^{\infty} |\psi(x_1,h)|^2\,dx_1 \int_{\rho^{-1}}^{\infty} \xi_2^{-3}\,d\xi_2 = h\rho^2. \end{aligned}$$

Using polar coordinates we find that

$$I_2 \le \int_{-h^{-1}}^{h^{-1}} h^2 \int_{-h^{-1}}^{h^{-1}} (1+\xi_1^2+\xi_2^2)^{-1/2}\rho^2 \, d\xi_2 \, d\xi_1 \le h^2\rho^2 \int_0^{h^{-1}\sqrt{2}} (1+r^2)^{-1/2} 2\pi r \, dr \le 2\pi\sqrt{2}\, h\rho^2,$$

and simple estimation gives

$$I_3 \le 4\int_0^{h^{-1}} h^2 \, d\xi_1 \int_{h^{-1}}^{\rho^{-1}} \xi_2^{-1}\rho^2 \, d\xi_2 = 4h\rho^2 \int_{h^{-1}}^{\rho^{-1}} \frac{d\xi_2}{\xi_2} = 4h\rho^2 \log\frac{h}{\rho},$$

$$I_4 \le 4\int_{h^{-1}}^{\infty} \xi_1^{-2} \int_0^{h^{-1}} \xi_1^{-1}\rho^2 \, d\xi_2 \, d\xi_1 = 4h^{-1}\rho^2 \int_{h^{-1}}^{\infty} \xi_1^{-3} \, d\xi_1 = 2h\rho^2,$$

$$I_5 \le 4\int_{h^{-1}}^{\infty} \xi_1^{-2} \, d\xi_1 \int_{h^{-1}}^{\rho^{-1}} \xi_2^{-1}\rho^2 \, d\xi_2 = 4h\rho^2 \log\frac{h}{\rho}. \qquad \square$$

LEMMA 7.3. *Let Γ be the square screen (2.5). For the continuous, piecewise-bilinear nodal basis on the power-graded mesh with vertices defined by (2.8), we have*

$$\Phi_{m,k} \simeq 1 \quad \textit{for } 0 \le m \le 1.$$

Proof. Without loss of generality, we may assume that x_k is the origin and that $\Gamma_k = [-h_-, h_+] \times [-\rho_-, \rho_+]$ with $h_\pm \simeq h_k$ and $\rho_\pm \simeq \rho_k$. We define the 1D continuous, piecewise-linear basis function on the interval $(-h_-, h_+)$,

$$\psi(x, h_+, h_-) = \begin{cases} 1 + \dfrac{x}{h_-} & \text{for } -h_- < x < 0, \\ 1 - \dfrac{x}{h_+} & \text{for } 0 < x < h_+, \\ 0 & \text{otherwise,} \end{cases}$$

and write the bilinear basis function on Γ_k by $\phi(x) = \psi(x_1, h_+, h_-)\psi(x_2, \rho_+, \rho_-)$. Recalling (5.8), we see that the result will follow from the lower bound

$$\|\phi\|^2_{\widetilde{H}^m(\Gamma)} \gtrsim h\rho^{1-2m}. \tag{7.6}$$

Since $\|\phi\|^2_{L_2(\Gamma)} \simeq h\rho$ and $|\phi|^2_{H^1(\Gamma)} \simeq h\rho(h^{-2} + \rho^{-2})$, the cases $m = 0$ and $m = 1$ are obvious. If $0 < m < 1$, then we use the norm equivalence

$$\|\phi\|^2_{\widetilde{H}^m(\Gamma)} = \|\phi\|^2_{\widetilde{H}^m(\mathbb{R}^2)} \simeq \int_{\mathbb{R}^2} (1 + |\xi|^2)^m |\hat{\phi}(\xi)|^2 \, d\xi,$$

where

$$\hat{\phi}(\xi) = \hat{\psi}(\xi_1, h_+, h_-)\hat{\psi}(\xi_2, \rho_+, \rho_-)$$

and

$$\hat{\psi}(\xi, h_+, h_-) = \frac{1}{(2\pi\xi)^2} \left\{ \frac{1 - e^{i2\pi\xi h_-}}{h_-} + \frac{1 - e^{-i2\pi\xi h_+}}{h_+} \right\}.$$

Since $(1 + |\xi|^2)^m \ge |\xi_2|^{2m}$ we see that

$$\|\phi\|^2_{\widetilde{H}^m(\mathbb{R}^2)} \gtrsim \int_{\mathbb{R}^2} |\xi_2|^{2m} |\hat{\phi}(\xi)|^2 \, d\xi = \int_{-\infty}^{\infty} \left|\hat{\psi}(\xi_1, h_+, h_-)\right|^2 d\xi_1 \int_{-\infty}^{\infty} |\xi_2|^{2m} \left|\hat{\psi}(\xi_2, \rho_+, \rho_-)\right|^2 d\xi_2, \tag{7.7}$$

and by Plancherel's theorem,

$$\int_{-\infty}^{\infty} \bigl|\hat{\psi}(\xi_1, h_+, h_-)\bigr|^2 \, d\xi_1 = \int_{-\infty}^{\infty} \bigl|\psi(x_1, h_+, h_-)\bigr|^2 \, dx_1 = \frac{h_+ + h_-}{3}. \tag{7.8}$$

Now define $h = (h_+ + h_-)/2$, $\Delta h = (h_+ - h_-)/2$, $\rho = (\rho_+ + \rho_-)/2$, and $\Delta\rho = (\rho_+ - \rho_-)/2$, so that $h_\pm = h \pm \Delta h$ and $\rho_\pm = \rho \pm \Delta\rho$. Using the substitution $\xi_2 = t/\rho$ in (7.7), we see that

$$\|\phi\|^2_{\widetilde{H}^m(\mathbb{R}^2)} \gtrsim h \int_{-\infty}^{\infty} |t/\rho|^{2m} \bigl|\hat{\psi}(t/\rho, \rho_+, \rho_-)\bigr|^2 \, \frac{dt}{\rho} = h\rho^{1-2m} \int_{-\infty}^{\infty} |t|^{2m} \bigl|\rho^{-1}\hat{\psi}(t/\rho, \rho_+, \rho_-)\bigr|^2 \, dt.$$

Putting $\epsilon = \Delta\rho/\rho = (\rho_+ - \rho_-)/(\rho_+ + \rho_-) \in (-1, 1)$, a simple calculation gives

$$\rho^{-1}\hat{\psi}(t/\rho, \rho_+, \rho_-) = (1-\epsilon) f_+[(1-\epsilon)t] + (1+\epsilon) f_-[(1+\epsilon)t], \text{ where } f_\pm(t) = \frac{1 - e^{\pm i 2\pi t}}{(2\pi t)^2}.$$

Since $f_\pm(t) = \mp i/(2\pi t) + O(1)$ as $t \to 0$ and $f_\pm(t) = O(t^{-2})$ as $t \to \infty$, the integral

$$I(\epsilon) = \int_{-\infty}^{\infty} |t|^{2m} \bigl|\rho^{-1}\hat{\psi}(t/\rho, \rho_+, \rho_-)\bigr|^2 \, dt$$

is analytic for $|\epsilon| < 1$, and (since $\rho_- = \rho_+ = \rho$ when $\epsilon = 0$) we have

$$I(0) = \int_{-\infty}^{\infty} |t|^{2m} \left(\frac{\sin \pi t}{\pi t}\right)^4 dt \simeq 1 \quad \text{for } 0 \le m \le 1.$$

The lower bound (7.6) follows for $0 < m < 1$ because $\max_{k\in\mathcal{N}} \epsilon_k \to 0$ as $N \to \infty$. ∎

7.3. Sharper versions of the theorems in section 5.

THEOREM 7.4. *Consider the weakly singular boundary integral equation* (2.1) *on the square screen* (2.5). *Then for piecewise-constant nodal basis functions with the mesh* (2.8),

$$\lambda_{\max}(\boldsymbol{B}') \lesssim N^{1/2} \times \begin{cases} 1 & \text{if } 1 \le \beta < 2, \\ (1 + \log N)^{1/2} & \text{if } \beta = 2, \\ (1 + \log N)^2 & \text{if } \beta > 2, \end{cases}$$

and

$$\lambda_{\min}(\boldsymbol{B}') \gtrsim \begin{cases} 1 & \text{if } \beta = 1, \\ (1 + \log N)^{-1} & \text{if } \beta > 1. \end{cases}$$

Proof. Note that $d_{\min}$ (defined in Lemma 7.1) satisfies $d_{\min} = n^{-\beta} \simeq N^{-\beta/2}$. Note also that

$$\max_{k\in\mathcal{N}} \frac{d_k}{\rho_k} = \max_{1\le j\le n/2} \frac{t_j}{\Delta t_j} \simeq \max_{1\le j\le n/2} \frac{(j/n)^\beta}{n^{-1}(j/n)^{\beta-1}} = \max_{1\le j\le n/2} j \simeq n \simeq N^{1/2}. \tag{7.9}$$

Hence, from part 1 of Lemma 7.1 we obtain

$$\lambda_{\max}(\boldsymbol{B}') \lesssim N^{1/2}(1 + \log N)^2 \quad \text{for all } \beta \ge 1.$$

We can combine this with Theorem 5.7 to obtain the required bounds on $\lambda_{\max}(\boldsymbol{B}')$. The proof is completed by using Lemmas 5.1 and 7.2 to obtain

$$\lambda_{\min}(\boldsymbol{B}') \gtrsim \min_{k\in\mathcal{N}} \Phi_{-1/2,k} \simeq \frac{1}{1+\log(n^{-1}/n^{-\beta})} \simeq \frac{1}{1+(\beta-1)\log n}$$
$$\simeq \begin{cases} 1 & \text{for } \beta = 1, \\ (1+\log N)^{-1} & \text{for } \beta > 1. \end{cases} \qquad \square$$

THEOREM 7.5. *Consider the hypersingular boundary integral equation* (2.3) *on the square screen* (2.5). *Then with conforming piecewise bilinear nodal basis functions on the mesh* (2.8),

$$\lambda_{\max}(\boldsymbol{B}') \lesssim 1 \quad \textit{and} \quad \lambda_{\min}(\boldsymbol{B}') \gtrsim N^{-1/2} \times \begin{cases} 1 & \textit{for } 1 \le \beta < 2, \\ (1+\log N)^{-1/2} & \textit{for } \beta = 2, \\ (1+\log N)^{-2} & \textit{for } \beta > 2. \end{cases}$$

Proof. Lemmas 5.1 and 7.3 imply that $\lambda_{\max}(\boldsymbol{B}') \lesssim \max_{k\in\mathcal{N}} \Phi_{1/2,j} \lesssim 1$. For $\beta > 2$ we sharpen the bound in Theorem 5.8 by using part 2 of Lemma 7.1. In fact, $d_{\min} \simeq n^{-\beta} \simeq N^{-\beta/2}$ and, by (7.9), $\min_{k\in\mathcal{N}} \rho_k/d_k \simeq N^{-1/2}$. Hence $\lambda_{\min}(\boldsymbol{B}') \gtrsim N^{-1/2}/(1+\log N)^2$, completing the proof. $\square$

8. Numerical experiments with a different family of meshes. To conclude, we present some numerical results for the weakly singular equation (2.1) over the nonconvex, polygonal screen

$$\Gamma = (-1,1)^2 \setminus ([0,1]\times[-1,0]). \tag{8.1}$$

The meshes are constructed by a pseudoadaptive procedure that starts with a uniform mesh and then selectively bisects elements so that the relation between h_K, ρ_K and the distance to the nearest edge or corner is equivalent to that for a power-graded mesh with a chosen grading exponent $\beta \ge 1$. For simplicity, we grade only into the edges and vertex of the re-entrant corner; i.e., we ignore the other four edges and five corners. Figure 3 shows a typical mesh. Strictly speaking, our theory does not

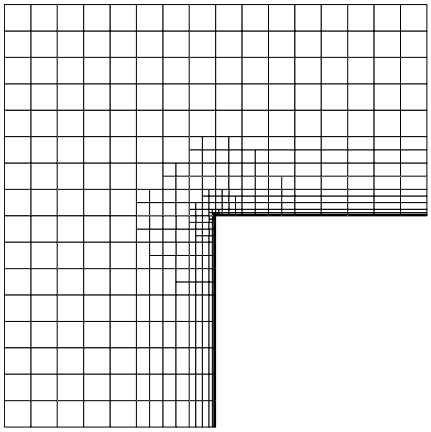


FIG. 3. *Anisotropic mesh on nonconvex screen, produced by a pseudoadaptive procedure,* $N = 488$, $\beta = 3$.

TABLE 5
Weakly singular integral equation (2.1) *on the nonconvex screen* (8.1) *with* $\beta = 2$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\operatorname{cond}(\boldsymbol{B})$	
73	2.82E-01		9.15E-04		3.08E+02	
296	7.15E-02	-0.981	1.43E-05	-2.971	5.00E+03	1.990
1185	1.79E-02	-0.999	2.23E-07	-2.998	8.01E+04	1.999
4743	4.47E-03	-0.999	3.49E-09	-2.999	1.28E+06	1.999
18958	1.12E-03	-1.000	5.46E-11	-3.002	2.05E+07	2.001
Expected	$\lesssim N^{-1}$		$\gtrsim N^{-3}$		$\lesssim N^{2}$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\operatorname{cond}(\boldsymbol{B}')$	
73	8.40E+00		3.61E-01		2.33E+01	
296	1.67E+01	0.491	3.11E-01	-0.106	5.37E+01	0.597
1185	3.33E+01	0.497	2.67E-01	-0.111	1.25E+02	0.608
4743	6.65E+01	0.499	2.30E-01	-0.108	2.89E+02	0.607
18958	1.33E+02	0.500	2.00E-01	-0.101	6.65E+02	0.601
Expected	$\lesssim N^{1/2}(1+\log N)^{1/2}$		$\gtrsim (1+\log N)^{-1}$		$\lesssim N^{1/2}(1+\log N)^{3/2}$	

TABLE 6
Weakly singular integral equation (2.1) *on the nonconvex screen* (8.1) *with* $\beta = 3$.

N	$\lambda_{\max}(\boldsymbol{B})$		$\lambda_{\min}(\boldsymbol{B})$		$\operatorname{cond}(\boldsymbol{B})$	
111	2.81E-01		1.44E-05		1.95E+04	
488	6.66E-02	-0.972	2.82E-08	-4.213	2.37E+06	3.241
2017	1.67E-02	-0.976	5.50E-11	-4.396	3.03E+08	3.421
8095	4.21E-03	-0.991	1.07E-13	-4.489	3.92E+10	3.499
Expected	$\lesssim N^{-1}$		$\gtrsim N^{-9/2}$		$\lesssim N^{7/2}$	
N	$\lambda_{\max}(\boldsymbol{B}')$		$\lambda_{\min}(\boldsymbol{B}')$		$\operatorname{cond}(\boldsymbol{B}')$	
111	1.03E+01		2.72E-01		3.78E+01	
488	2.12E+01	0.489	2.04E-01	-0.194	1.04E+02	0.683
2017	4.29E+01	0.497	1.60E-01	-0.169	2.67E+02	0.666
8095	8.75E+01	0.498	1.31E-01	-0.145	6.53E+02	0.643
Expected	$\lesssim N^{1/2}(1+\log N)^{2}$		$\gtrsim (1+\log N)^{-1}$		$\lesssim N^{1/2}(1+\log N)^{3}$	

cover this example because we require conforming meshes, although in practical terms there is no need to enforce any interelement continuity condition at the hanging nodes because we use discontinuous (piecewise-constant) nodal basis functions. Tables 5 and 6 give our numerical results using meshes with $\beta = 2$ and $\beta = 3$. The asymptotic behavior of the extremal eigenvalues and the condition numbers is essentially the same as we observed previously, in Tables 1 and 2, for simple tensor-product, power-graded meshes on a square screen with the same choices of β.

Acknowledgment. We thank Dr. E. Georgoulis for very helpful comments.

REFERENCES

[1] M. AINSWORTH, W. MCLEAN, AND T. TRAN, *The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling*, SIAM J. Numer. Anal., 36 (1999), pp. 1901–1932.

[2] M. AINSWORTH, W. MCLEAN, AND T. TRAN, *Diagonal scaling of stiffness matrices in the Galerkin boundary element method*, ANZIAM J., 42 (2000), pp. 141–150.

[3] R. E. BANK AND L. R. SCOTT, *On the conditioning of finite element equations with highly refined meshes*, SIAM J. Numer. Anal., 26 (1989), pp. 1383–1394.

[4] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.
[5] J. Elschner, *The double layer potential operator over polyhedral domains. II. Spline Galerkin methods*, Math. Methods Appl. Sci., 15 (1992), pp. 23–37.
[6] I. G. Graham, W. Hackbusch, and S. A. Sauter, *Hybrid Galerkin boundary elements on degenerate meshes*, in Mathematical Aspects of Boundary Element Methods, Chapman & Hall/CRC Res. Notes Math. 414, M. Bonnet, A.-M. Sändig, and W. L. Wendland, eds., Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 140–151.
[7] I. G. Graham, W. Hackbusch, and S. A. Sauter, *Finite elements on degenerate meshes: Inverse-type inequalities and applications*, IMA J. Numer. Anal., 25 (2005), pp. 379–407.
[8] W. Hackbusch and Z. Nowak, *On the fast matrix multiplication in the boundary element method by panel clustering*, Numer. Math., 54 (1989), pp. 463–491.
[9] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[10] J. C. Nédélec, *Integral equations with non-integrable kernels*, Integral Equations Operator Theory, 4 (1982), pp. 563–572.
[11] T. von Petersdorff, *Randwertprobleme der Elastizitätstheorie für Polyeder—Singularitäten und Approximation mit Randelementmethoden*, Dissertation, Technische Hochschule Darmstadt, Darmstadt, Germany, 1989.
[12] V. Rokhlin, *Rapid solution of the integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1514–1534

TWO-LEVEL SCHWARZ ALGORITHMS WITH OVERLAPPING SUBREGIONS FOR MORTAR FINITE ELEMENTS*

HYEA HYUN KIM† AND OLOF B. WIDLUND†

Abstract. Preconditioned conjugate gradient methods based on two-level overlapping Schwarz methods often perform quite well. Such a preconditioner combines a coarse space solver with local components which are defined in terms of subregions that form an overlapping covering of the region on which the elliptic problem is defined. Precise bounds on the rate of convergence of such iterative methods have previously been obtained in the case of conforming lower order and spectral finite elements as well as in a number of other cases. In this paper, this domain decomposition algorithm and analysis are extended to mortar finite elements. It is established that the condition number of the relevant iteration operator is independent of the number of subregions and varies with the relative overlap between neighboring subregions linearly as in the conforming cases previously considered.

Key words. domain decomposition, elliptic finite element problems, preconditioned conjugate gradients, mortar finite elements, overlapping Schwarz algorithms

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/050635857

1. Introduction.

In this paper, the well-known two-level Schwarz method (see, e.g., [15, Chap. 3]) is extended to mortar finite element methods. Mortar finite element methods were first introduced in [7]. They are nonconforming finite element methods based on a partitioning, not necessarily geometrically conforming, of the region Ω into substructures Ω_i. Thus, in three dimensions, vertices and edges of one substructure can fall in the interior of edges and/or faces of its neighbors and in two dimensions vertices can divide edges of neighboring substructures. In each of the substructures, we choose a conforming standard finite element or a spectral element method without much regard for its neighbors. Even if the substructures are geometrically conforming, e.g., when the set of substructures forms a regular finite element triangulation, the local finite element meshes need not. We can also use spectral finite element spaces of different order in different substructures, and we can mix finite elements and spectral elements as well. In this paper, we will work out a theory only for the case of piecewise linear mortar finite elements; we treat both the more conventional mortar finite elements and those introduced by Wohlmuth [16, 17].

We note that Achdou and Maday have considered a related problem in [1]. However, in their paper, the principal issue is to establish the convergence and best possible error bounds for finite element methods based on overlapping subdomains. Typically, the meshes in the regions common to two or more overlapping subdomains do not match, and mortar conditions are used to introduce a weak continuity between the boundary values of one component of the finite element solution and the interior val-

*Received by the editors July 12, 2005; accepted for publication (in revised form) February 14, 2006; published electronically August 7, 2006.
http://www.siam.org/journals/sinum/44-4/63585.html

†Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012 (hhk2@cims.nyu.edu, http://cims.nyu.edu/~hhk2, widlund@cs.nyu.edu, http://cs.nyu.edu/cs/faculty/widlund/index.html). The first author was supported in part by the Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DE-FG02-00ER25053 and in part by the Post-doctoral Fellowship Program of Korea Science and Engineering Foundation (KOSEF). The second author was supported in part by the U.S. Department of Energy under contract DE-FC02-01ER25482.

ues of different components along the boundary of the first subdomain. In the final subsection of their paper a convergence result similar to ours, and that for standard conforming elements, is formulated and established. We note that in our paper, we instead consider overlapping Schwarz methods for the standard mortar methods. For references to earlier work by Cai, Dryja, and Sarkis, which is related to Achdou's and Maday's work, see the reference section of [1].

Finally, a word about the history of this project. The second author worked on algorithms of this kind almost ten years ago; the work was then not completed, but some results were presented in a talk at the 1996 ECCOMAS conference in Paris. The basic idea of using three independent decompositions of the region, including one for a conforming finite element space on a regular coarse grid, was inspired by a paper by Chan, Smith, and Zou [9]. Around the same time, Dan Stefanica conducted numerical experiments which demonstrated that there is very little difference in the performance of the two-level overlapping Schwarz method for a mortar case and a regular conforming finite element case if the subdomains and the overlap are chosen similarly. The work then lay dormant until it recently was reexamined by the present authors; many details have now been added and a more complete theory has now been developed.

2. The elliptic problem and mortar finite element methods. To simplify the notation, we consider only Poisson's equation. As usual, we formulate our elliptic problem as follows: find $u \in V$, such that

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx = f(v) \quad \forall v \in V. \tag{2.1}$$

The definition of $V \subset H^1(\Omega)$ incorporates the boundary conditions, and the region Ω is assumed to be bounded and polyhedral; a homogeneous Dirichlet condition is imposed on a nonempty subset $\partial\Omega_D$ of the boundary $\partial\Omega$ of Ω and a natural boundary condition is given on $\partial\Omega_N = \partial\Omega \backslash \partial\Omega_D$. (Inhomogeneous Neumann boundary data can be incorporated into the right-hand side of (2.1).) It is well known that the bilinear form $a(\cdot,\cdot)$ is self-adjoint, elliptic, and bounded in $V \times V$. Our analysis is equally valid for two and three dimensions. The bilinear form $a(u,v)$ is directly related to the Sobolev space $H^1(\Omega)$ that is defined by the seminorm and norm

$$|u|^2_{H^1(\Omega)} = a(u,u) \quad \text{and} \quad \|u\|^2_{H^1(\Omega)} = |u|^2_{H^1(\Omega)} + \|u\|^2_{L^2(\Omega)},$$

respectively.

The discretization of an elliptic, second order problem starts by partitioning the computational domain Ω into a union of nonoverlapping substructures, $\{\Omega_i\}_{i=1}^I$, and an interface Γ, defined by $(\cup_{i\neq j}\partial\Omega_i \cap \partial\Omega_j) \setminus \partial\Omega_D$, which is a set of points that belong to the boundaries of at least two substructures. The restriction to an individual substructure Ω_i, of the mortar finite element space considered in detail in this paper, will just be a standard piecewise linear finite element space defined on a quasi-uniform mesh. The meshes of two neighboring substructures do not necessarily match on their common interface, and the elements of the discrete space V^h are typically discontinuous across the interface Γ. Instead of pointwise continuity, the interface jumps are made orthogonal to a carefully chosen space of trial functions. In our work, we primarily consider the second generation mortar element methods for which continuity is not even imposed at the vertices or wire baskets (the union of the edges and vertices) of the substructures. Even if the meshes match across the interface between adjacent

substructures, the mortar finite element functions will not, generally, be pointwise continuous.

This weak continuity is introduced in terms of a set of *mortars* $\{\gamma_m\}_{m=1}^M$ obtained by selecting open edges/faces of the substructures such that

$$\overline{\Gamma} = \cup_{m=1}^M \overline{\gamma}_m, \quad \gamma_m \cap \gamma_n = \emptyset \ \text{ if } m \neq n.$$

Each edge/face and mortar γ_m is viewed as belonging to just one substructure. The remaining edges/faces are the *nonmortars* and are denoted by δ_n. The restrictions of the triangulations of the different substructures to the mortars and nonmortars typically will not match and are denoted by γ_m^h and δ_n^h, respectively; discontinuous mortar finite element functions have two different traces on the interface Γ given by one-sided limits of finite element functions defined on the individual substructures. The continuity across the interface of a conforming finite element method is replaced by weak continuity across the individual nonmortars: for each n, we define a space of test functions $M(\delta_n)$ given by the restriction, to the nonmortar δ_n, of the finite element space defined on the substructure of which δ_n is an edge/face. In two dimensions, the elements of $M(\delta_n)$ are subject to the constraints that they are constant in the first and last mesh intervals of δ_n^h. In three dimensions, the value of a test function of $M(\delta_n)$ at a node on $\partial\delta_n$ is given by a fixed convex combination of nodal values at its next neighbors in δ_n; cf. Belgacem and Maday [5]. We will call this the standard Lagrange multiplier space. In the spectral case, we would use as test functions polynomials of a degree two less.

Lagrange multiplier spaces with dual bases have been developed by Wohlmuth [16, 17]. Each basis function associated with these Lagrange multiplier spaces is supported on a few mesh intervals just as for the standard Lagrange multiplier spaces. They are discontinuous and lead to a diagonal matrix instead of the mass matrix appearing in the standard mortar matching condition. Our algorithm and our proofs can be applied both to the standard and dual Lagrange multiplier spaces, and $M(\delta_n)$ can therefore represent either the standard or the dual Lagrange multiplier space.

In this paper, we consider partitions $\{\Omega_i\}_{i=1}^I$, where the Ω_i are geometrically nonconforming. We assume that $\{\Omega_i\}_{i=1}^I$ form a regular partition of Ω, i.e., the size of Ω_i is comparable to that of its neighboring substructures. We will impose some assumptions on the meshes and the Lagrange multiplier space $M(\delta_n)$. A nonmortar $\delta_n \subset \partial\Omega_i$ can be partitioned into several edges/faces $\{\delta_{n,j}\}_j$ by mortar neighbors $\Omega_{m(n,j)}$ with boundaries which intersect $\partial\Omega_i$ along $\delta_{n,j}$, i.e., $\delta_{n,j} = \partial\Omega_{m(n,j)} \cap \partial\Omega_i$. We will use the following assumptions on the meshes and the Lagrange multiplier space in some of our work.

ASSUMPTION 1. *Each subpartition $\delta_{n,j}$ of a nonmortar is the union of entire elements.*

ASSUMPTION 2. *The Lagrange multiplier space $M(\delta_{n,j})$ are defined on each edge/face of the partition $\delta_{n,j}$ individually. Standard or dual Lagrange multiplier spaces are thus given on each $\delta_{n,j}$ which inherits the triangulation from δ_n^h. The Lagrange multiplier space $M(\delta_n)$ on δ_n is then defined by*

$$M(\delta_n) = \prod_{\delta_{n,j}} M(\delta_{n,j}).$$

With these assumptions, mortar methods provide a best approximation even for geometrically nonconforming partitions. Without them, an additional factor $|\log(h)|$ will appear in the error bound; see [2]. See also [3, 4, 5, 7], where error bounds of

the same type as for standard conforming methods are derived. We will first analyze two-level overlapping Schwarz algorithms for mortar methods under Assumptions 1 and 2 and later derive a slightly weaker result after removing these assumptions.

The *mortar projection* π_n maps all of $L_2(\delta_n)$ onto the finite element space defined on the nonmortar mesh δ_n^h. For two dimensions and for a given $w \in L_2(\delta_n)$ with given values at v_{n_1} and v_{n_2}, the endpoints of δ_n, we define $\pi_n(w, w^{(n)}(v_{n_1}), w^{(n)}(v_{n_2}))$ on δ_n^h by

$$\int_{\delta_n} (w - \pi_n(w, w^{(n)}(v_{n_1}), w^{(n)}(v_{n_2})))\psi ds = 0 \quad \forall \psi \in M(\delta_n). \tag{2.2}$$

We note that only the values at the interior nodes of δ_n are determined by this condition; the values $w^{(n)}(v_{n_1})$ and $w^{(n)}(v_{n_2})$ are genuine degrees of freedom. Similarly, for three dimensions, the values in the interior of δ_n are determined not only by the values on the part of Γ opposite the nonmortar, but also by the nodal values on $\partial\delta_n$.

As when working with other nonconforming methods, the original bilinear form $a(\cdot,\cdot)$ is replaced by $a^\Gamma(\cdot,\cdot)$ defined as the sum of the contributions from the individual substructures to $a(\cdot,\cdot)$:

$$a^\Gamma(u_h, v_h) = \sum_{i=1}^{I} a_{\Omega_i}(u_h, v_h). \tag{2.3}$$

For $u_h = v_h$, we obtain the square of what is often called a *broken norm*. The norm has been broken along Γ and it is finite for any element of the mortar space even if it is discontinuous across Γ. The resulting discrete variational problem gives rise to a linear system with a symmetric, positive definite matrix.

After these preparations, the mortar finite element space V^h, and the problem as a whole, can be fully defined. The discrete problem is then the following: find $u \in V^h$ such that

$$a^\Gamma(u, v) = f^\Gamma(v) \quad \forall v \in V^h, \tag{2.4}$$

where $a^\Gamma(u, v)$ is defined in (2.3) and, similarly, $f^\Gamma(v)$ is the sum of contributions from the different substructures.

3. The Dryja–Widlund algorithm. We now describe the additive Schwarz method introduced in Dryja and Widlund [10]; cf. also Smith, Bjørstad, and Gropp [14, Chap. 5] and, for many details, Toselli and Widlund [15, Chap. 3]. This additive Schwarz method for an overlapping subdomain partition performs quite well even for partitions with small overlap as first established in Dryja and Widlund [11]. The condition number bound given in [11] has also been proven to be optimal by Brenner [8]. We now use two additional decompositions of the region Ω, in addition to the set of substructures $\{\Omega_i\}$, used to define the mortar finite element problem, namely, a set of overlapping subregions $\{\widetilde{\Omega}_j\}$ and an independent coarse mesh $\{\tau_l^H\}$. Let X_i^h be the finite element space on the substructure Ω_i equipped with a quasi-uniform triangulation $\mathcal{T}^h(\Omega_i)$. Throughout this paper, we will impose the following assumptions on these partitions.

ASSUMPTION 3. *The diameter H_i of a substructure Ω_i is comparable to the diameter H of any triangle τ_l^H that intersects it.*

ASSUMPTION 4. *The diameter H_i of a substructure Ω_i satisfies*

$$H_i \leq C\widetilde{H}_j,$$

where $\widetilde{H}_j$ is the diameter of any subregion $\widetilde{\Omega}_j$ that intersects it.

ASSUMPTION 5. *The mesh sizes of the substructures that intersect along a common edge/face are comparable.*

The $\widetilde{\Omega}_j$ can be quite arbitrary; a local subspace V_j will be associated with each of them, essentially by making all genuine degrees of freedom associated with nodes outside $\widetilde{\Omega}_j$ equal to zero. More precisely, the space V_j is given by

$$V_j = \left\{ v \in \prod_{i=1}^{I} X_i^h \ : \ v(x) = 0 \text{ for } x \in \Omega \setminus \widetilde{\Omega}_j, \text{ or } x \in \delta_n \right\}, \quad j = 1, \ldots, N,$$

where δ_n denotes any nonmortar edges/faces. The space V_0 is V^H, the space of continuous, piecewise linear functions on an independent coarse mesh given by its elements τ_l^H. We further impose zero Dirichlet conditions, on the elements of V_j, on $\partial\widetilde{\Omega}_j \cap \partial\Omega_D$ and on the elements of V_0, on $\partial\Omega_D$.

We note that the overlap can be quite small. Thus, if no degrees of freedom are shared between neighboring subregions, the overlap is on the order of h, the diameter of the elements of the fine discretization. Our analysis applies in this case as well, in which case our Schwarz method corresponds to a block Jacobi preconditioner augmented by a coarse solver.

It is now appropriate essentially to follow the description and analysis of Schwarz methods given in Smith, Bjørstad, and Gropp [14] and Toselli and Widlund [15]. Our iterative method is given in terms of $N + l$ finite element spaces $V_j^h, j = 0, \ldots, N$, which are subspaces of V^h and are associated with the space V_j:

$$V_j^h = I^m(V_j).$$

The interpolation operator $I^m : \prod_{i=1}^{I} C(\Omega_i) \rightarrow V^h$ is defined by

$$I^m(u) = \sum_{i=1}^{I} \left(I_i^h(u) + \sum_{\delta_n \subset \partial\Omega_i} \widetilde{\pi}_n \left(I_{m(\delta_n)}^h(u) - I_i^h(u) \right) \right), \tag{3.1}$$

where $I_i^h(u)$ is the nodal value interpolant in the space X_i^h and $\widetilde{\pi}_n(w)$ is the zero extension of $\pi_n(w)$ to $\overline{\Omega}_i$. Here $\pi_n(w)$ denotes $\pi_n(w, 0, 0)$; see (2.2). (In the following, we will use the simple notation $\pi_n(w)$ instead of $\pi_n(w, 0, 0)$.) It has been shown that $\pi_n(w)$ is L^2-stable but not H^1-stable; see [17, Chap. 1]. We recall that δ_n denotes a nonmortar edge/face of $\partial\Omega_i$ and that $\{\delta_{n,j}\}_j$ is the partition of δ_n described in section 2, i.e., $\delta_{n,j} = \partial\Omega_{m(n,j)} \cap \partial\Omega_i$. The interpolant $I_{m(\delta_n)}^h(u)$ is defined by

$$I_{m(\delta_n)}^h(u) = I_{m(n,j)}^h(u) \ \text{ on } \delta_{n,j},$$

and it can thus be discontinuous across the boundaries of $\delta_{n,j}$. The mortar finite element space V^h can then be represented as the sum

$$V^h = V_0^h + V_1^h + \cdots + V_N^h \,. \tag{3.2}$$

Remark 1. The local spaces $\{V_j^h\}_{j=1}^N$, in which our Schwarz algorithm will be considered, consist of functions defined on the whole domain Ω, and not just on the subregion $\widetilde{\Omega}_j$, as in the standard Schwarz algorithms described in [15, Chap. 3]. Therefore the trivial extension operator from V_j^h to V^h will not appear in our algorithm.

We note that the support of each function in V_j^h is contained in the union of the substructures Ω_i that intersect the subregion $\widetilde{\Omega}_j$.

It is often more economical to use approximate rather than exact solvers for the subspace problems. The approximate solvers can be described in terms of inner products $\tilde{a}_j(\cdot,\cdot)$ defined on $V_j^h \times V_j^h$. One assumption that needs to be checked for each of them is the existence of a constant ω such that

$$a^{\Gamma}(u,u) \leq \omega \tilde{a}_j(u,u) \quad \forall u \in V_j^h \ . \tag{3.3}$$

In terms of matrices, this inequality becomes a one-sided bound of a submatrix of the stiffness matrix, given by $a^{\Gamma}(\cdot,\cdot)$ and V_j^h, in terms of the matrix given by $\tilde{a}_j(\cdot,\cdot)$.

A projection-like operator $T_j : V^h \to V_j^h$ is now defined for each j by

$$\tilde{a}_j(T_j u, \phi_h) = a^{\Gamma}(u, \phi_h) \quad \forall \phi_h \in V_j^h \ . \tag{3.4}$$

It is easy to show that the operator T_j is positive semidefinite and symmetric with respect to $a^{\Gamma}(\cdot,\cdot)$ and that the minimal constant ω in (3.3) is $\|T_j\|_a$, i.e.,

$$\|T_j\|_a \leq \omega; \tag{3.5}$$

see [15, Chap. 2]. Additive and multiplicative Schwarz methods can now be defined straightforwardly in terms of polynomials of the operators T_j. We note that if exact solvers, and thus genuine projections P_j, are used, then $\omega = 1$. The operator relevant to an additive Schwarz operator is $T = \sum_{j=0}^{N} T_j$. In the case of no coarse space and the local spaces forming a direct sum, this operator is a block-Jacobi operator, with one block for each subspace.

In order to estimate the rate of convergence of our special, or any other, additive Schwarz methods, we need upper and lower bounds for the spectrum of the operator relevant in the conjugate gradient iteration. A lower bound can be obtained by using the following lemma; see, e.g., Zhang [18], Smith, Bjørstad, and Gropp [14], or Toselli and Widlund [15, Chap. 2].

LEMMA 1. *Let T_j be the operators defined in* (3.4) *and let $T = T_0 + T_1 + \cdots + T_N$. Then*

$$a(T^{-1}u, u) = \min_{u = \sum u_j} \sum \tilde{a}_j(u_j, u_j), \ u_j \in V_j^h.$$

Therefore, if a representation, $u = \sum u_j$, can be found, such that

$$\sum \tilde{a}_j(u_j, u_j) \leq C_0^2 a(u,u) \quad \forall u \in V^h,$$

then

$$\lambda_{min}(T) \geq C_0^{-2}.$$

For the algorithms considered in this paper, and many other domain decomposition algorithms, it is easy to show that there is an upper bound for T which is proportional to ω.

In this paper, our results are formulated only for additive algorithms and with exact solvers for the subdomain problems. The corresponding bounds for the multiplicative variants, etc., can easily be worked out using the general Schwarz theory; see, e.g., Smith, Bjørstad, and Gropp [14] or Toselli and Widlund [15, Chap. 2].

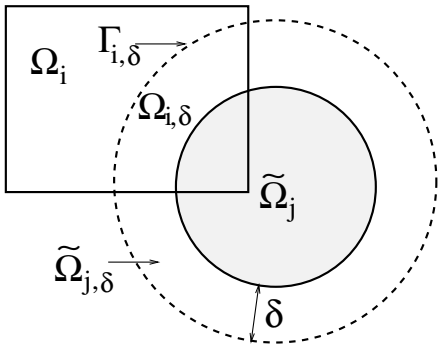


FIG. 1. *The substructure Ω_i intersects the subregion $\widetilde{\Omega}_j$ (interior of the dashed circle); $\widetilde{\Omega}_{j,\delta}$ (the part between the dashed and the solid circles) is the support of $\nabla\theta_j$, $\Omega_{i,\delta}$ is the part of $\widetilde{\Omega}_{j,\delta}$ which belongs to Ω_i, and $\Gamma_{i,\delta}$ is a part of the boundary of $\Omega_{i,\delta}$ that divides Ω_i into two parts.*

4. The lower bound. We will find a lower bound of the two-level Schwarz algorithm. A stable decomposition of $u = \sum_{j=0}^{N} u_j$ will be provided with

$$C_0^2 = C \max_{j=1,\ldots,N} \{(1 + H_j/\delta_j)\}.$$

Here H_j is the diameter of the subregion $\widetilde{\Omega}_j$ and δ_j is the overlapping width of $\widetilde{\Omega}_j$, i.e., the minimal width of the subset of $\widetilde{\Omega}_j$ which is common to some neighbors, and C is a constant independent of the mesh sizes, the subregion diameters, and the number of subregions; see Figure 1. We first assume that the five assumptions hold and later derive a bound for C_0^2 with an additional $\log(H/h)$ factor for the general case for which Assumptions 1 and 2 are removed.

4.1. Technical tools. In this section, we will collect a number of technical tools that are used in proving our main results. Some of the tools can be borrowed directly from Toselli and Widlund [15, Chap. 3], but some work also needs to be done that is directly related to the mortar finite element method.

As before, $\Omega \subset R^d, d = 2$ or 3, is a bounded, polygonal region, $\{\Omega_i\}_{i=1}^I$ is a nonoverlapping decomposition of Ω into substructures, and $\{\widetilde{\Omega}_j\}_{j=1}^N$ that of a set of overlapping subregions. Let $\{\widetilde{\theta}_j\}_{j=1}^N$ be a partition of unity for the overlapping partition $\{\widetilde{\Omega}_j\}_{j=1}^N$ of Ω, with the following properties (see, e.g., [15, sect. 3.2]):

$$\begin{aligned}
&0 \le \tilde{\theta}_j(x) \le 1, \quad x \in \widetilde{\Omega}_j,\\
&\operatorname{supp}(\widetilde{\theta}_j) \subset \widetilde{\Omega}_j,\\
&\sum_{j=1}^{N} \widetilde{\theta}_j = 1,\\
&|\nabla\widetilde{\theta}_j| \le \frac{C}{\delta_j}.
\end{aligned}$$

We will employ a modified partition of unity θ_j obtained by interpolating $\widetilde{\theta}_j$ on the triangulations $\{\mathcal{T}^h(\Omega_i)\}_{i=1}^I$. The θ_j will be discontinuous across substructure interfaces. However, we can easily check that the modified partition of unity $\{\theta_j\}_{j=1}^N$ has the same properties as $\{\widetilde{\theta}_j\}_{j=1}^N$ when restricted to any substructure Ω_i because these properties hold for each elements of $\{\mathcal{T}^h(\Omega_i)\}_{i=1}^I$.

We now consider the case in Figure 1. The substructure Ω_i intersects the subregion $\widetilde{\Omega}_j$. We denote the support of $\nabla\theta_j$ by $\widetilde{\Omega}_{j,\delta}$ and the intersection of Ω_i and $\widetilde{\Omega}_{j,\delta}$ by $\Omega_{i,\delta}$. As in Figure 1, we select $\Gamma_{i,\delta}$ as a part of the boundary of $\Omega_{i,\delta}$ that divides the domain Ω_i into two parts. We will prove the following lemma, which is similar to [15, Lem. 3.10].

LEMMA 2. *Let u be an arbitrary element of $H^1(\Omega_i)$. Then*

$$\|u\|^2_{L^2(\Omega_{i,\delta})} \le C\,\delta^2\big((1+H_i/\delta)|u|^2_{H^1(\Omega_i)} + 1/(H_i\delta)\|u\|^2_{L^2(\Omega_i)}\big),$$

where H_i denotes the diameter of Ω_i and δ is the overlapping width of $\widetilde{\Omega}_j$, a subregion that intersects Ω_i.

Proof. Let us cover $\Omega_{i,\delta}$ by shape-regular patches $\{P_l\}_l$ with $O(\delta)$ diameters. We may assume that the $P_{l,\Gamma}$ $(:= \partial P_l \cap \Gamma_{i,\delta})$ have positive measure. By using a Friedrichs inequality (see Toselli and Widlund [15, Lem. A.17]) for each patch P_l and summing over all patches, we obtain

$$\|u\|^2_{L^2(\Omega_{i,\delta})} \le C\left(\delta^2|u|^2_{H^1(\Omega_{i,\delta})} + \delta\|u\|^2_{L^2(\Gamma_{i,\delta})}\right). \tag{4.1}$$

From the embedding $H^{1/2}(\Gamma_{i,\delta}) \subset L^2(\Gamma_{i,\delta})$, a trace theorem, and a scaling argument, we obtain

$$\begin{aligned}
\|u\|^2_{L^2(\Gamma_{i,\delta})} &= H_i^{d-1}\|\widehat{u}\|^2_{L^2(\widehat{\Gamma}_{i,\delta})}\\
&\le CH_i^{d-1}\|\widehat{u}\|^2_{H^{1/2}(\widehat{\Gamma}_{i,\delta})}\\
&\le CH_i^{d-1}\|\widehat{u}\|^2_{H^1(\widehat{\Omega}_{i,1})}\\
&= CH_i^{d-1}\left(|\widehat{u}|^2_{H^1(\widehat{\Omega}_i)} + \|\widehat{u}\|^2_{L^2(\widehat{\Omega}_i)}\right)\\
&= CH_i^{d-1}\left(H_i^{2-d}|u|^2_{H^1(\Omega_i)} + H_i^{-d}\|u\|^2_{L^2(\Omega_i)}\right).
\end{aligned}$$

Here the hat designates a dilated domain with diameter 1 or a function defined on the scaled domain, and $\Omega_{i,1}$ is a part of Ω_i divided by $\Gamma_{i,\delta}$. By combining the above estimate with (4.1), the desired bound follows. ∎

We also have the following generalized Poincaré–Friedrichs inequality (see Nečas [13]).

LEMMA 3. *Let Φ be a seminorm on $H^1(\Omega)$ with the following properties:*

(1) $\Phi(\phi) \le C_1\|\phi\|_{1,\Omega} \quad \forall\phi \in H^1(\Omega)$.

(2) *For a constant function c, $\Phi(c) = 0$ iff $c = 0$.*

Then we have a generalized Poincaré–Friedrichs inequality for $H^1(\Omega)$,

$$\|\phi\|_{0,\Omega} \le CH^{d/2}\left(H^{(2-d)/2}|\phi|_{1,\Omega} + H^{k(\Phi)}\Phi(\phi)\right) \quad \forall\phi \in H^1(\Omega),$$

where d is the dimension of the domain Ω, H is the diameter of Ω, and the constant C is independent of H; $\Phi(\phi)$ is homogeneous of degree $k(\Phi)$, i.e., $k(\Phi)$ is the real number which makes $H^{k(\Phi)}\Phi(\phi)$ invariant to scaling.

A prime example is provided by

$$\Phi(\phi) = \left|\int_\gamma \phi\,ds\right| \quad \forall\phi \in H^1(\Omega).$$

Then the two assumptions of Lemma 3 hold for $\Phi(\phi)$ and the application of the Poincaré–Friedrichs inequality for ϕ with a zero average on γ, i.e., $\Phi(\phi) = 0$, gives

$$\|\phi\|_{0,\Omega} \leq CH|\phi|_{1,\Omega}. \tag{4.2}$$

We will now consider two cases. In the first, the meshes and Lagrange multipliers satisfy Assumptions 1 and 2 on the nonconformity of the subdomain partition $\{\Omega_i\}_{i=1}^I$. In the second, we will drop these assumptions. In the latter case, the Lagrange multiplier space $M(\delta_n)$ is then a standard or dual Lagrange multiplier space defined on the triangulation δ_n^h, without partitioning it into $\{\delta_{n,j}\}_j$. The following approximation properties hold for both the standard and the dual Lagrange multiplier spaces; see [7, 12, 17].

LEMMA 4. *Let $0 < \alpha \leq 1/2$. For $v \in H^\alpha(\delta_{n,j})$, there exists a $\psi \in M(\delta_{n,j})$ such that*

$$\|v - \psi\|_{0,\delta_{n,j}} \leq Ch^\alpha |v|_{H^\alpha(\delta_{n,j})},$$

where h denotes the diameter of the elements of the nonmortar δ_n.

LEMMA 5. *Let $0 < \alpha \leq 1/2$. For $v \in H^\alpha(\delta_n)$, there exists $\psi \in M(\delta_n)$ such that*

$$\|v - \psi\|_{(H^\alpha(\delta_n))'} \leq Ch^{2\alpha} |v|_{H^\alpha(\delta_n)},$$

where h denotes the diameter of the nonmortar elements and $(H^\alpha(\delta_n))'$ is the dual space of $H^\alpha(\delta_n)$.

LEMMA 6. *Let the meshes and Lagrange multiplier spaces satisfy Assumptions* 1 *and* 2. *Then, for $v = (v_1, \ldots, v_I) \in V^h$, we have*

$$\|v_i - v_j\|_{0,\delta_{n,j}} \leq Ch_i^{1/2} \left(|v_i|_{1,\Omega_i} + |v_j|_{1,\Omega_j} \right),$$

where Ω_i and Ω_j are the nonmortar and mortar substructures of the interface $\delta_{n,j} = \partial\Omega_i \cap \partial\Omega_j$.

Proof. We have

$$\begin{aligned} \|v_i - v_j\|_{0,\delta_{n,j}}^2 &= \int_{\delta_{n,j}} (v_i - v_j)(v_i - v_j - \psi)\, ds \\ &\leq \|v_i - v_j\|_{0,\delta_{n,j}} \|v_i - v_j - \psi\|_{0,\delta_{n,j}}. \end{aligned}$$

This inequality holds for an arbitrary $\psi \in M(\delta_{n,j})$. Applying Lemma 4 with $\alpha = 1/2$ and a trace theorem, we obtain

$$\min_{\psi \in M_{\delta_{n,j}}} \|v_i - v_j - \psi\|_{0,\delta_{n,j}} \leq Ch_i^{1/2} \left(|v_i|_{1,\Omega_i} + |v_j|_{1,\Omega_j} \right). \qquad \square$$

We now consider a general case without the extra Assumptions 1 and 2 on the meshes and Lagrange multiplier spaces. The set of nonmortars $\{\delta_n\}_n$ is selected from the edges/faces of the subdomain partition, and the Lagrange multiplier spaces $M(\delta_n)$ are defined on the finite elements associated with the nonmortar interfaces δ_n. We recall that any nonmortar edge/face $\delta_n \subset \partial\Omega_i$ is partitioned into

$$\overline{\delta}_n = \cup_j \overline{\delta}_{n,j}, \quad \delta_{n,j} = \delta_n \cap \partial\Omega_{n_j}.$$

The mortar matching condition is then

$$\int_{\delta_n} (v_{i(n)} - \phi)\psi\, ds = 0 \quad \forall \psi \in M(\delta_n), \tag{4.3}$$

where ϕ is given by $\phi = v_{n_j}$ on $\delta_{n,j}$. We see that $\phi \in H^{1/2-\epsilon}(\delta_n)$ for any $0 < \epsilon \le 1/2$. Moreover the following estimate holds for ϕ; see [6].

LEMMA 7. *Let each subdomain Ω_{n_j} be scaled by H_i, the diameter of the subdomain Ω_i. Then, for any $0 < \epsilon \le 1/2$, we have*

$$\sqrt{\epsilon}\|\phi\|_{H^{1/2-\epsilon}(\delta_n)} \le C \sum_j \|v_{n_j}\|_{1,\Omega_{n_j}},$$

where ϕ is given by $\phi = v_{n_j}$ on $\delta_n \cap \partial\Omega_{n_j}$.

In the general case, without Assumptions 1 and 2, the space V^h consists of functions $v = (v_1, \ldots, v_I)$ satisfying the mortar matching condition (4.3) on each nonmortar edge/face δ_n. Let us denote by $\{\psi_l\}_l$ a basis for the Lagrange multiplier space $M^h(\delta_n)$. We also select $\{\psi_{j_k}\}_k$ from $\{\psi_l\}_l$ such that supp $(\psi_{j_k}) \subset \delta_{n,j}$ $(= \delta_n \cap \partial\Omega_{n_j})$, and set $\psi_{n,j} = \sum_k \psi_{j_k}$; we assume that at least one such ψ_{j_k} exists for every $\delta_{n,j}$. We will then show that the L^2-norm of the jump across δ_n is bounded by the sum of H^1-seminorms of the functions on the subdomains Ω_k for which $\partial\Omega_k$ intersects δ_n with a positive measure.

LEMMA 8. *Let $\delta_n \subset \partial\Omega_i$ be a nonmortar edge/face. For the general case, without Assumptions 1 and 2, we have*

$$\|v_i - \phi\|_{0,\delta_n} \le C h_i^{1/2} \left(\log \frac{H_i}{h_i}\right)^{1/2} \left(|v_i|_{1,\Omega_i} + \sum_j |v_{n_j}|_{1,\Omega_{n_j}}\right)$$

for $v = (v_1, \ldots, v_I) \in V^h$, where ϕ is given by $\phi = v_{n_j}$ on $\delta_{n,j}$.

Proof. We first dilate Ω_i and Ω_{n_j} so that the diameter of Ω_i is 1. The triangles/tetrahedra of each subdomain are then also scaled by the diameter H_i. We obtain

$$\begin{aligned}
\|v_i - \phi\|^2_{0,\delta_n} &= \int_{\delta_n} (v_i - \phi)(v_i - \phi - \psi)\, ds \\
&\le \|v_i - \phi\|_{H^{1/2-\epsilon}(\delta_n)} \|v_i - \phi - \psi\|_{(H^{1/2-\epsilon}(\delta_n))'} \\
&\le C\|v_i - \phi\|_{H^{1/2-\epsilon}(\delta_n)} h_i^{2(1/2-\epsilon)} |v_i - \phi|_{H^{1/2-\epsilon}(\delta_n)} \\
&\le C h_i^{1-2\epsilon} \|v_i - \phi\|^2_{H^{1/2-\epsilon}(\delta_n)}. \qquad (4.4)
\end{aligned}$$

Here $\psi \in M(\delta_n)$ is the best approximation and h_i is the scaled mesh size. We have also used the mortar matching condition and Lemma 5 for the function $v_i - \phi \in H^{1/2-\epsilon}(\delta_n)$.

We now define

$$\tilde{v}_i = v_i - c_{ij}, \quad \tilde{\phi} = v_{n_j} - c_{ij} \quad \text{on } \delta_{n,j},$$

where

$$c_{ij} = \frac{\int_{\delta_{n,j}} v_i \psi_{n,j}\, ds}{\int_{\delta_{n,j}} \psi_{n,j}\, ds} = \frac{\int_{\delta_{n,j}} v_{n_j} \psi_{n,j}\, ds}{\int_{\delta_{n,j}} \psi_{n,j}\, ds}.$$

The equality above holds because of the mortar matching condition for $v = (v_1, \ldots, v_I) \in V^h$ and the fact that the function $\psi_{n,j} \in M^h(\delta_n)$ is supported in $\delta_{n,j}$. We also have

$$\tilde{v}_i - \tilde{\phi} = v_i - \phi \quad \text{in } L^2(\delta_n), \quad \tilde{v}_i - \tilde{\phi} \in H^{1/2-\epsilon}(\delta_n).$$

From these properties and by applying (4.4) to $\widetilde{v}_i - \widetilde{\phi}_i$, we obtain

$$\|v_i - \phi\|^2_{0,\delta_n} \leq C h_i^{1-2\epsilon} \|\tilde{v}_i - \tilde{\phi}\|^2_{H^{1/2-\epsilon}(\delta_n)}.$$

Applying Lemma 7 to $\tilde{v}_i$ and $\tilde{\phi}$ gives

$$\|v_i - \phi\|_{0,\delta_n} \leq C h_i^{1/2-\epsilon} \epsilon^{-1/2} \sum_j \left(\|v_i - c_{ij}\|_{1,\Omega_i} + \|v_{n_j} - c_{ij}\|_{1,\Omega_{n_j}} \right).$$

Let

$$\Phi_{ij}(w) = \left| \int_{\delta_{n,j}} w \, \psi_{n,j} \, ds \right|.$$

Since $\psi_{n,j}$ is bounded from above by a constant, independent of the mesh parameters, and $\int_{\delta_{n,j}} \psi_{n,j} \, ds > 0$, $\Phi_{ij}(w)$ satisfies the two properties of Lemma 3; positivity of the integral also holds for the dual Lagrange multiplier case. By applying Lemma 3 to $v_i - c_{ij}$ and $v_{n_j} - c_{ij}$, with the seminorm Φ_{ij}, we obtain

$$\|v_i - c_{ij}\|_{1,\Omega_i} \leq C |v_i|_{1,\Omega_i}, \quad \|v_{n_j} - c_{ij}\|_{1,\Omega_{n_j}} \leq C |v_{n_j}|_{1,\Omega_{n_j}}.$$

Therefore,

$$\|v_i - \phi\|_{0,\delta_n} \leq C h_i^{1/2-\epsilon} \epsilon^{-1/2} \left(|v_i|_{1,\Omega_i} + \sum_j |v_{n_j}|_{1,\Omega_{n_j}} \right).$$

Letting $\epsilon = 1/|\log h_i|$ gives $\log(h_i^{-\epsilon}) = 1$ and results in the bound

$$\|v_i - \phi\|_{0,\delta_n} \leq C h_i^{1/2} |\log h_i|^{1/2} \left(|v_i|_{1,\Omega_i} + \sum_j |v_{n_j}|_{1,\Omega_{n_j}} \right). \tag{4.5}$$

By considering the scaling, we find

$$\|v\|_{0,\delta_n} = H_i^{(d-1)/2} \|\hat{v}\|_{0,\hat{\delta}_n}, \quad |v|_{1,\Omega_i} = H_i^{(d-2)/2} |\hat{v}|_{1,\hat{\Omega}_i}. \tag{4.6}$$

Here $\hat{\delta}_n$ and $\hat{\Omega}_i$ denote the scaled domains and $\hat{v}$ denotes the function defined on the scaled set $\hat{\delta}_n$ or $\hat{\Omega}_i$. We then obtain

$$\begin{aligned}
\|v_i - \phi\|_{0,\delta_n} &= H_i^{(d-1)/2} \|\hat{v}_i - \hat{\phi}\|_{0,\hat{\delta}_n} \\
&\leq C H_i^{(d-1)/2} \hat{h}_i^{1/2} |\log \hat{h}_i|^{1/2} \left(|\hat{v}_i|_{1,\hat{\Omega}_i} + \sum_j |\hat{v}_{n_j}|_{1,\hat{\Omega}_{n_j}} \right) \\
&\leq C H_i^{(d-1)/2} H_i^{-(d-2)/2} \hat{h}_i^{1/2} |\log \hat{h}_i|^{1/2} \left(|v_i|_{1,\Omega_i} + \sum_j |v_{n_j}|_{1,\Omega_{n_j}} \right) \\
&\leq C H_i^{1/2} \left(\frac{h_i}{H_i} \right)^{1/2} \left(\log \frac{H_i}{h_i} \right)^{1/2} \left(|v_i|_{1,\Omega_i} + \sum_j |v_{n_j}|_{1,\Omega_{n_j}} \right).
\end{aligned}$$

Here we have used (4.5), (4.6) and that $\hat{h}_i = h_i/H_i$. ∎

4.2. The stability of a certain interpolation operator. Let $I^H : V^h \to V^H$ be a stable quasi interpolant in both the H^1- and L^2-norms in the following sense:

$$\sum_{i=1}^{I} |I^H u|^2_{1,\Omega_i} \le C \sum_{i=1}^{I} |u|^2_{1,\Omega_i},$$
$$\sum_{i=1}^{I} \frac{1}{H_i^2} \|u - I^H u\|^2_{0,\Omega_i} \le C \sum_{i=1}^{I} |u|^2_{1,\Omega_i},$$

where H_i denotes the diameter of Ω_i. We then obtain the same bound for $u_0 = I^m(I^H u)$.

LEMMA 9. *Let $u_0 = I^m(I^H u)$ for $u \in V^h$. Then u_0 satisfies*

$$\sum_{i=1}^{I} |u_0|^2_{1,\Omega_i} \le C \sum_{i=1}^{I} |u|_{1,\Omega_i},$$
$$\sum_{i=1}^{I} \frac{1}{H_i^2} \|u - u_0\|^2_{0,\Omega_i} \le C \sum_{i=1}^{I} |u|^2_{1,\Omega_i}.$$

Proof. We find, using (3.1), that

$$|u_0|^2_{1,\Omega_i} \le C \left\{ |I_i^h(I^H u)|^2_{1,\Omega_i} + \sum_{\delta_n \subset \partial\Omega_i} \left| \widetilde{\pi}_n \left(I^h_{m(\delta_n)}(I^H u) - I_i^h(I^H u) \right) \right|^2_{1,\Omega_i} \right\}.$$

From the H^1-stability of the nodal value interpolant I_i^h for functions in V^H (see [15, Lem. 3.8]), the first term above is bounded by

$$|I_i^h(I^H u)|^2_{1,\Omega_i} \le C |I^H u|^2_{1,\Omega_i}. \tag{4.7}$$

We estimate the second term by

$$\begin{aligned} &\left| \widetilde{\pi}_n \left(I^h_{m(\delta_n)}(I^H u) - I_i^h(I^H u) \right) \right|^2_{1,\Omega_i} \\ &\quad \le C h_i^{-1} \left\| \pi_n \left(I^h_{m(\delta_n)}(I^H u) - I_i^h(I^H u) \right) \right\|^2_{0,\delta_n} \qquad (4.8) \\ &\quad \le C h_i^{-1} \left\{ \|I^h_{m(\delta_n)}(I^H u) - I^H u\|^2_{0,\delta_n} + \|I_i^h(I^H u) - I^H u\|^2_{0,\delta_n} \right\} \\ &\quad \le C h_i^{-1} \left\{ \sum_{\delta_{n,j} \subset \delta_n} h_{m(n,j)} |I^H u|^2_{1,\Omega_{m(n,j)}} + h_i |I^H u|^2_{1,\Omega_i} \right\}, \qquad (4.9) \end{aligned}$$

where $\delta_{n,j} = \partial\Omega_{m(n,j)} \cap \partial\Omega_i$. We have used an inverse inequality, the stability of π_n in $L^2(\delta_n)$, and the approximation property of the nodal value interpolation operator for $I^H u \in V^H$ provided by [15, Lem. 3.8]. Adding (4.7) and (4.9) over all nonmortar sides and subdomains and using Assumption 5 and the H^1-stability of the coarse interpolation operator I^H, we obtain

$$\sum_{i=1}^{I} |u_0|^2_{1,\Omega_i} \le C \sum_{i=1}^{I} |u|^2_{1,\Omega_i}.$$

We now estimate

$$\begin{aligned}&\|u-u_0\|^2_{0,\Omega_i}\\ &\le C\left\{\left\|u-I^h_i(I^Hu)\right\|^2_{0,\Omega_i}+\sum_{\delta_n\subset\partial\Omega_i}\left\|\widetilde{\pi}_n\left(I^h_{m(\delta_n)}(I^Hu)-I^h_i(I^Hu)\right)\right\|^2_{0,\Omega_i}\right\}.\end{aligned}\tag{4.10}$$

The first term is bounded by

$$\begin{aligned}\|u-I^h_i(I^Hu)\|^2_{0,\Omega_i}&\le 2\|u-I^Hu\|^2_{0,\Omega_i}+2\|I^h_i(I^Hu)-I^Hu\|^2_{0,\Omega_i}\\ &\le C\left\{\|u-I^Hu\|^2_{0,\Omega_i}+h_i^2|I^Hu|^2_{1,\Omega_i}\right\}.\end{aligned}$$

By using (4.8) and (4.9), we bound the second term of (4.10) as follows:

$$\begin{aligned}&\left\|\widetilde{\pi}_n\left(I^h_{m(\delta_n)}(I^Hu)-I^h_i(I^Hu)\right)\right\|^2_{0,\Omega_i}\\ &\le Ch_i\left\|\pi_n\left(I^h_{m(\delta_n)}(I^Hu)-I^h_i(I^Hu)\right)\right\|^2_{0,\delta_n}\\ &\le C\left(h_i^2|I^Hu|^2_{1,\Omega_i}+\sum_{\delta_{n,j}\subset\delta_n}h_ih_{m(n,j)}|I^Hu|^2_{1,\Omega_{m(n,j)}}\right)\\ &\le C\left(H_i^2|I^Hu|^2_{1,\Omega_i}+\sum_{\delta_{n,j}\subset\delta_n}H^2_{m(n,j)}|I^Hu|^2_{1,\Omega_{m(n,j)}}\right).\end{aligned}$$

In (4.10), summing the second term over the nonmortar sides gives

$$\|u-u_0\|^2_{0,\Omega_i}\le C\left(\|u-I^Hu\|^2_{0,\Omega_i}+\sum_{\delta_n\subset\partial\Omega_i}\sum_{|\partial\Omega_l\cap\delta_n|>0}H_l^2|I^Hu|^2_{1,\Omega_l}\right).$$

From the assumption that the diameter of Ω_i is comparable to those of its neighbors Ω_l, a coloring argument, and the L^2- and H^1-stability of the interpolation I^Hu, we obtain the second bound of the lemma. ∎

We now introduce our coarse interpolation operator $I^H:V^h\to V^H$. Let K be a triangle/tetrahedron in the coarse triangulation of Ω. Each vertex y_l of the triangle belongs to at least one substructure Ω_k (or to $\partial\Omega_k$) of the nonoverlapping partition. We denote the subdomain containing the vertex y_l by Ω_l. The set w_K is the union of the elements in T^H, the boundary of which intersects the boundary of the given element K. We consider a case as in Figure 2. The interpolation is defined by the values

$$(I^Hu)(y_l)=\frac{1}{|w_{y_l}|}\int_{w_{y_l}}u\,dx,$$

where $w_{y_l}=w_K\cap\Omega_l$ and $|w_{y_l}|$ denotes the volume of w_{y_l}. In the following, we show that this coarse interpolation operator is stable in both the H^1- and L^2-norms.

LEMMA 10. *The coarse interpolant $I^H:V^h\to V^H$ satisfies*

$$\sum_{i=1}^I\frac{1}{H_i^2}\|u-I^Hu\|^2_{0,\Omega_i}\le C\sum_{i=1}^I|u|^2_{1,\Omega_i},$$

$$\sum_{i=1}^I|I^Hu|^2_{1,\Omega_i}\le C\sum_{i=1}^I|u|^2_{1,\Omega_i}.$$

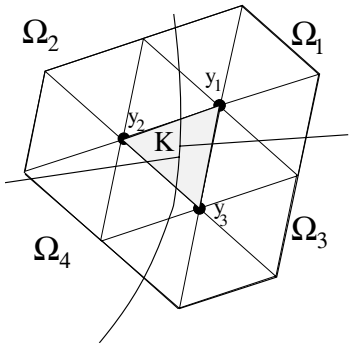


FIG. 2. *The region w_K divided by a geometrically nonconforming subdomain partition.*

Proof. We first estimate

$$\|I^H u\|_{0,K}^2 \le C \sum_{l=1}^{3} |(I^H u)(y_l)|^2 \|\phi_l\|_{0,K}^2 \le C \sum_{l=1}^{3} \|u(x)\|_{0,w_{y_l}}^2 \frac{|K|}{|w_{y_l}|}, \tag{4.11}$$

where ϕ_l is the nodal basis function of the vertex y_l of the coarse triangle K. In general, we can have more than one subdomain Ω_k which intersects K and does not contain any vertices of K. For simplicity, we assume that we have only one such subdomain and denote it by Ω_4 (see Figure 2).

Let us denote by c_l the average of u over the subdomain Ω_l, and by K_l the common part of K and Ω_l, and let

$$c_l = \frac{1}{|\Omega_l|} \int_{\Omega_l} u \, dx, \quad K_l = K \cap \Omega_l \quad \forall\, l = 1, \ldots, 4. \tag{4.12}$$

We then obtain

$$\begin{aligned} \|u - I^H u\|_{0,K}^2 &= \|u - c_1 - I^H(u - c_1)\|_{0,K}^2 \\ &\le 2\|u - c_1\|_{0,K}^2 + 2\|I^H(u - c_1)\|_{0,K}^2 \\ &\le C \left\{ \|u - c_1\|_{0,K}^2 + \sum_{l=1}^{3} \|u - c_1\|_{0,w_{y_l}}^2 \frac{|K|}{|w_{y_l}|} \right\} \qquad (4.13) \\ &\le C \left\{ \sum_{l=1}^{3} \|u - c_1\|_{0,w_{y_l}}^2 + \|u - c_1\|_{0,K_4}^2 \right\}. \qquad (4.14) \end{aligned}$$

Here we have used the identity $I^H(c_1) = c_1$, the estimate (4.11), and the fact that the factor $|K|/|w_{y_l}|$ is bounded from above independently of any mesh parameters.

From the Poincaré inequality and Assumption 3, we have

$$\|u - c_l\|_{0,w_{y_l}}^2 \le C H_K^2 \, |u|_{1,\Omega_l}^2, \quad l = 1, 2, 3.$$

We now consider

$$\|u - c_1\|_{0,w_{y_2}}^2 \le 2\|u - c_2\|_{0,w_{y_2}}^2 + 2\|c_2 - c_1\|_{0,w_{y_2}}^2.$$

Let

$$c_{12} = \frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} u|_{\Omega_1} \, ds = \frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} u|_{\Omega_2} \, ds,$$

where Γ_{12} is the common edge/face of Ω_1 and Ω_2. The identity follows from the mortar matching condition for the function u. We then have

$$\|c_2 - c_1\|^2_{0,w_{y_2}} \leq C \left\{ |c_2 - c_{12}|^2 + |c_1 - c_{12}|^2 \right\} |w_{y_2}|.$$

The first term in the above equation is written as

$$\begin{aligned} c_2 - c_{12} &= \frac{1}{|\Omega_2|} \int_{\Omega_2} u_2 \, dx - \frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} u_2 \, ds \\ &= \frac{1}{|\Omega_2|} \int_{\Omega_2} \left(u_2 - \frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} u_2 \, ds \right) dx, \end{aligned}$$

where $u_2 = u|_{\Omega_2}$. Let

$$\widetilde{u}_2 = u_2 - \frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} u_2.$$

Applying the Poincaré inequality to $\widetilde{u}_2$ and using the Hölder inequality, we obtain

$$|c_2 - c_{12}|^2 \leq C H_2^{2-d} |u|^2_{1,\Omega_2}.$$

Similarly, we obtain

$$|c_1 - c_{12}|^2 \leq C H_1^{2-d} |u|^2_{1,\Omega_1}.$$

We then have

$$\|c_2 - c_1\|^2_{0,w_{y_2}} \leq C H_K^2 (|u|^2_{1,\Omega_1} + |u|^2_{1,\Omega_2}).$$

Here we have used that $|w_{y_2}| \leq H_K^d$ for $d = 2, 3$ and Assumption 3. The estimate of the remaining terms in (4.14) can be done similarly and gives

$$\|u - I^H u\|^2_{0,K} \leq C H_K^2 \sum_{l=1}^{4} |u|^2_{1,\Omega_l}. \tag{4.15}$$

By summing the above inequality over all K which intersect Ω_i, we obtain

$$\begin{aligned} \frac{1}{H_i^2} \|u - I^H u\|^2_{0,\Omega_i} &\leq \frac{1}{H_i^2} \sum_{K \cap \Omega_i \neq \emptyset} \|u - I^H u\|^2_{0,K} \\ &\leq C \frac{1}{H_i^2} \sum_{K \cap \Omega_i \neq \emptyset} H_K^2 \left(\sum_{\Omega_l \cap K \neq \emptyset} |u|^2_{1,\Omega_l} \right). \end{aligned}$$

The fact that the H_i is comparable to H_K and a coloring argument give the first estimate of the lemma. We note that we also have the following estimate from (4.13) and (4.15):

$$\|u - c_1\|^2_{0,K} \leq C H_K^2 \sum_{l=1}^{4} |u|_{1,\Omega_l}. \tag{4.16}$$

We now estimate

$$\begin{aligned}|I^H u|^2_{1,K} &= |I^H u - c_1|^2_{1,K} \\ &\le C H_K^{-2} \|I^H u - c_1\|^2_{0,K} \\ &\le C H_K^{-2} \left(\|I^H u - u\|^2_{0,K} + \|u - c_1\|^2_{0,K} \right),\end{aligned}$$

where c_1 is the constant defined in (4.12). We have used an inverse inequality. By using (4.15) and (4.16), we obtain

$$|I^H u|^2_{1,K} \le C \sum_{l=1}^{4} |u|^2_{1,\Omega_l}.$$

The second estimate of the lemma follows by summing the above term over all triangles K and a coloring argument. ∎

Remark 2. For the general case, without Assumptions 1 and 2, we choose

$$c_{12} = \frac{\int_{\Gamma_{12}} u|_{\Omega_1} \psi_{12}\, ds}{\int_{\Gamma_{12}} \psi_{12}\, ds} = \frac{\int_{\Gamma_{12}} u|_{\Omega_2} \psi_{12}\, ds}{\int_{\Gamma_{12}} \psi_{12}\, ds},$$

where ψ_{12} is the sum of the basis functions for $M^h(\delta_n)$ that are supported in Γ_{12}. The identity holds for $u \in V^h$. The arguments in the proof of Lemma 10 can also be applied to this general case and give the same bounds.

LEMMA 11. *Under Assumptions* 1 *and* 2, *and for* $u \in V^h$, *there exists a stable decomposition*

$$u = u_0 + u_1 + \cdots + u_N$$

such that

$$\sum_{i=0}^{N} a^{\Gamma}(u_i, u_i) \le C \max_{i=1,\ldots,N} \left\{ \left(1 + \frac{H_i}{\delta_i} \right) \right\} a^{\Gamma}(u, u),$$

where H_i *and* δ_i *denote the diameter of the subregion* $\widetilde{\Omega}_i$ *and the overlapping width of* $\widetilde{\Omega}_i$.

Proof. We take $u_0 = I^m(I^H(u))$ using the interpolants I^m and I^H provided in Lemmas 9 and 10. We then define

$$u_i = I^m(\widetilde{u}_i), \quad \widetilde{u}_i = \theta_i(u - u_0) \quad \text{for } i = 1, \ldots, N.$$

From $u - u_0 \in V^h$ and $\sum_{i=1}^{N} \theta_i = 1$, we see that

$$u - u_0 = I^m(u - u_0) = \sum_{i=1}^{N} u_i.$$

The function u_i is supported as in Figure 3 and can be written as

$$u_i = I^m(\widetilde{u}_i) = \sum_{l=1}^{6} \left(I^h_{k_l}(\widetilde{u}_i) + \sum_{\delta_n \subset \partial\Omega_{k_l}} \widetilde{\pi}_{\delta_n} \left(I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i) \right) \right).$$

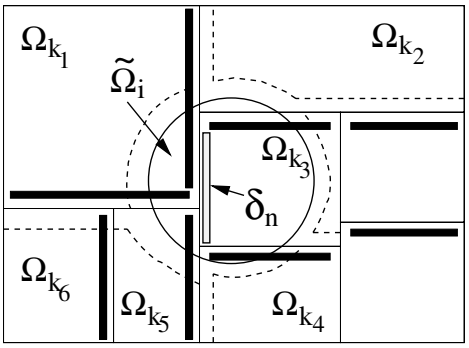


FIG. 3. *Nonconforming subdomain partition: mortar sides of interfaces (black bars), support of the functions $u_i \in V_i^h\,(= I^m(V_i))$ corresponding to the overlapping subdomain $\widetilde{\Omega}_i$ (interior of the dotted line); the subdomain Ω_{k_3} meets Ω_{k_1} and Ω_{k_5} along the nonmortar interface δ_n.*

Here $I^h_{m(\delta_n)}(\widetilde{u}_i)$, on δ_n in Figure 3, is given by

$$I^h_{m(\delta_n)}(\widetilde{u}_i) = \begin{cases} I^h_{k_1}(\widetilde{u}_i) \text{ on } \delta_{n,1} = \partial\Omega_{k_1} \cap \delta_n, \\ I^h_{k_5}(\widetilde{u}_i) \text{ on } \delta_{n,5} = \partial\Omega_{k_5} \cap \delta_n. \end{cases}$$

We will now prove that

$$\sum_{i=1}^{N} a^{\Gamma}(u_i, u_i) \le C \max_{i=1,\ldots,N} \left\{ \left(1 + \frac{H_i}{\delta_i}\right) \right\} a^{\Gamma}(u, u).$$

The required bound then follows by combining with Lemma 9. We consider

$$\begin{aligned} a^{\Gamma}(u_i, u_i) &= \sum_{l=1}^{6} |u_i|^2_{1,\Omega_{k_l}} \\ &= \sum_{l=1}^{6} \left| I^h_{k_l}(\widetilde{u}_i) + \sum_{\delta_n \subset \partial\Omega_{k_l}} \widetilde{\pi}_{\delta_n} \left(I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i) \right) \right|^2_{1,\Omega_{k_l}}. \end{aligned} \tag{4.17}$$

We note that $\widetilde{u}_i|_{\Omega_{k_l}}$ is a continuous and piecewise quadratic function defined on $T^h(\Omega_{k_l})$. From [15, Lem. 3.9], we have

$$|I^h_{k_l}(\widetilde{u}_i)|^2_{1,\Omega_{k_l}} \le C|\widetilde{u}_i|^2_{1,\Omega_{k_l}}. \tag{4.18}$$

For the second term of (4.17), we obtain

$$\begin{aligned} \left| \widetilde{\pi}_{\delta_n} \left(I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i) \right) \right|^2_{1,\Omega_{k_l}} &\le C h_{k_l}^{-2} h_{k_l} \left\| \pi_{\delta_n} \left(I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i) \right) \right\|^2_{0,\delta_n} \\ &\le C h_{k_l}^{-1} \left\| I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i) \right\|^2_{0,\delta_n}. \end{aligned} \tag{4.19}$$

Here we have used an inverse inequality, the quasi uniformity of the triangulation in the subdomain Ω_{k_l}, and the L^2-continuity of the mortar projection π_{δ_n}. We now

consider the term $\|I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_l}(\widetilde{u}_i)\|^2_{0,\delta_n}$ for δ_n and $l = 3$ in Figure 3:

$$\begin{aligned}
&\left\|I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_3}(\widetilde{u}_i)\right\|^2_{0,\delta_n} \\
&= \left\|I^h_{k_1}(\widetilde{u}_i) - I^h_{k_3}(\widetilde{u}_i)\right\|^2_{0,\delta_{n,1}} + \left\|I^h_{k_5}(\widetilde{u}_i) - I^h_{k_3}(\widetilde{u}_i)\right\|^2_{0,\delta_{n,5}} \\
&\le C\left(\left\|I^h_{k_1}(\widetilde{u}_i) - \widetilde{u}_i|_{\Omega_{k_1}}\right\|^2_{0,\delta_{n,1}} + \left\|I^h_{k_3}(\widetilde{u}_i) - \widetilde{u}_i|_{\Omega_{k_3}}\right\|^2_{0,\delta_{n,1}} + \left\|\widetilde{u}_i|_{\Omega_{k_1}} - \widetilde{u}_i|_{\Omega_{k_3}}\right\|^2_{0,\delta_{n,1}}\right. \\
&\quad \left. + \left\|I^h_{k_5}(\widetilde{u}_i) - \widetilde{u}_i|_{\Omega_{k_5}}\right\|^2_{0,\delta_{n,5}} + \left\|I^h_{k_3}(\widetilde{u}_i) - \widetilde{u}_i|_{\Omega_{k_3}}\right\|^2_{0,\delta_{n,5}} + \left\|\widetilde{u}_i|_{\Omega_{k_5}} - \widetilde{u}_i|_{\Omega_{k_3}}\right\|^2_{0,\delta_{n,5}}\right),
\end{aligned}$$

where $\delta_{n,j} = \partial\Omega_{k_j} \cap \partial\Omega_{k_3}$ for $j = 1, 5$.

Let $w = u - u_0$. We now consider

$$\begin{aligned}
&\|\widetilde{u}_i|_{\Omega_{k_1}} - \widetilde{u}_i|_{\Omega_{k_3}}\|^2_{0,\delta_{n,1}} \\
&= \|I^h_{k_1}(\widetilde{\theta}_i)w|_{\Omega_{k_1}} - I^h_{k_3}(\widetilde{\theta}_i)w|_{\Omega_{k_3}}\|^2_{0,\delta_{n,1}} \\
&\le C\left(\sum_{l=1,3} \|(I^h_{k_l}(\widetilde{\theta}_i) - \widetilde{\theta}_i)w|_{\Omega_{k_l}}\|^2_{0,\delta_{n,1}} + \|\widetilde{\theta}_i(w|_{\Omega_{k_1}} - w|_{\Omega_{k_3}})\|^2_{0,\delta_{n,1}}\right). \qquad (4.20)
\end{aligned}$$

Using the approximation property of the nodal value interpolant, $\|\nabla\widetilde{\theta}_i\|_\infty \le C/\delta_i$, and a trace theorem, the first term above can be estimated:

$$\begin{aligned}
\|(I^h_{k_l}(\widetilde{\theta}_i) - \widetilde{\theta}_i)w|_{\Omega_{k_l}}\|^2_{0,\delta_{n,1}} &\le \|I^h_{k_l}(\widetilde{\theta}_i) - \widetilde{\theta}_i\|^2_{0,\delta_{n,1}}\|w|_{\Omega_{k_l}}\|^2_{0,\delta_{n,1}} \\
&\le Ch_{k_l}|\widetilde{\theta}_i|^2_{1,\Omega_{k_l}}\|w\|^2_{1,\Omega_{k_l}} \\
&\le Ch_{k_l}\frac{1}{\delta_i^2}|\Omega_{k_l,\delta_i}|\|w\|^2_{1,\Omega_{k_l}},
\end{aligned}$$

where $|\Omega_{k_l,\delta_i}|$ denotes the volume of the set Ω_{k_l,δ_i}, which is the support of $\nabla\widetilde{\theta}_i$ contained in Ω_{k_l}. In general, we have $|\Omega_{k_1,\delta_i}| \le C\delta_i^{d-1}H_{k_l}$ with $d = 2$ or 3. Using this, we obtain

$$\|(I^h_{k_l}(\widetilde{\theta}_i) - \widetilde{\theta}_i)w|_{\Omega_{k_l}}\|^2_{0,\delta_{n,1}} \le Ch_{k_l}\left(1 + \frac{H_{k_l}}{\delta_i}\right)\left(|w|^2_{1,\Omega_{k_l}} + \frac{1}{H^2_{k_l}}\|w\|^2_{0,\Omega_{k_l}}\right). \qquad (4.21)$$

Using Lemma 6, the second term in (4.20) is bounded by

$$\begin{aligned}
&\left\|\widetilde{\theta}_i(w|_{\Omega_{k_j}} - w|_{\Omega_{k_3}})\right\|^2_{0,\delta_{n,j}} \\
&\le C\left\|\widetilde{\theta}_i\right\|^2_{\infty,\delta_{n,j}}\left\|w|_{\Omega_{k_j}} - w|_{\Omega_{k_3}}\right\|^2_{0,\delta_{n,j}} \\
&\le Ch_{k_3}\left(|w|^2_{1,\Omega_{k_j}} + |w|^2_{1,\Omega_{k_3}}\right), \quad j = 1, 5. \qquad (4.22)
\end{aligned}$$

Combining (4.20) with (4.21) and (4.22), and the approximation property of the nodal interpolation operators $I^h_{k_j}$, $j = 1, 3, 5$, for the functions $\widetilde{u}_i$, which are continu-

ous and piecewise quadratic on $T^h(\Omega_{k_j})$, lead to the following estimate:

(4.23)
$$\left\|I^h_{m(\delta_n)}(\widetilde{u}_i) - I^h_{k_3}(\widetilde{u}_i)\right\|^2_{0,\delta_n} \leq Ch_{k_3}\left(\sum_{j=1,3,5}|\widetilde{u}_i|^2_{1,\Omega_{k_j}} + \left(1+\frac{H_i}{\delta_i}\right)\sum_{j=1,3,5}\left(|w|^2_{1,\Omega_{k_j}} + \frac{1}{H^2_{k_j}}\|w\|^2_{0,\Omega_{k_j}}\right)\right),$$

where H_i is the diameter of the subregion $\widetilde{\Omega}_i$. Here we have used Assumptions 4 and 5.

Combining the estimates in (4.23), (4.19), and (4.18) with (4.17), we obtain

$$a^\Gamma(u_i,u_i) \leq C\left(\sum_{l\in\mathcal{S}_i}|\widetilde{u}_i|^2_{1,\Omega_{k_l}} + \left(1+\frac{H_i}{\delta_i}\right)\sum_{l\in\mathcal{S}_i}\left(|u-u_0|^2_{1,\Omega_{k_l}} + \frac{1}{H_{k_l}}\|u-u_0\|^2_{0,\Omega_{k_l}}\right)\right),$$

where $\mathcal{S}_i = \{l \,:\, \Omega_{k_l}\cap\widetilde{\Omega}_i \neq \emptyset\}$, the set of indices k_l of the substructures which intersect the subregion $\widetilde{\Omega}_i$. The first term of the above equation is estimated as follows:

$$\begin{aligned}|\widetilde{u}_i|^2_{1,\Omega_{k_l}} &= \left\|\nabla\left(\widetilde{\theta}_i(u-u_0)\right)\right\|^2_{0,\Omega_{k_l}}\\ &\leq C\left\{\int_{\Omega_{k_l}}\left|(u-u_0)\nabla\widetilde{\theta}_i\right|^2 dx + \int_{\Omega_{k_l}}\left|\widetilde{\theta}_i\nabla(u-u_0)\right|^2 dx\right\}\\ &\leq C\left\{\frac{1}{\delta_i^2}\int_{\Omega_{k_l,\delta_i}}(u-u_0)^2\,dx + |u-u_0|^2_{1,\Omega_{k_l}}\right\},\end{aligned}$$

where Ω_{k_l,δ_i} is the support of $\nabla\widetilde{\theta}_i$ contained in Ω_{k_l}. We then obtain by applying Lemma 2 to $\int_{\Omega_{k_l,\delta_i}}(u-u_0)^2\,dx$:

$$\frac{1}{\delta_i^2}\int_{\Omega_{k_l,\delta_i}}(u-u_0)^2\,dx \leq C\left(\left(1+\frac{H_{k_l}}{\delta_i}\right)|u-u_0|^2_{1,\Omega_{k_l}} + \frac{1}{H_{k_l}\delta_i}\|u-u_0\|^2_{0,\Omega_{k_l}}\right).$$

Using Assumption 4, we have

$$a^\Gamma(u_i,u_i) \leq C\left(1+\frac{H_i}{\delta_i}\right)\left(\sum_{l\in\mathcal{S}_i}|u-u_0|^2_{1,\Omega_{k_l}} + \sum_{l\in\mathcal{S}_i}\frac{1}{H^2_{k_l}}\|u-u_0\|^2_{0,\Omega_{k_l}}\right).$$

By summing the above estimate over all the subregions $\widetilde{\Omega}_i$, using a coloring argument and the estimates in Lemma 9, we obtain

$$\begin{aligned}\sum_{i=1}^N a^\Gamma(u_i,u_i) &\leq C\max_{i=1,\ldots,N}\left\{\left(1+\frac{H_i}{\delta_i}\right)\right\}\left(\sum_{l=1}^N|u-u_0|^2_{1,\Omega_l} + \sum_{l=1}^N\frac{1}{H_l^2}\|u-u_0\|^2_{0,\Omega_l}\right)\\ &\leq C\max_{i=1,\ldots,N}\left\{\left(1+\frac{H_i}{\delta_i}\right)\right\}a^\Gamma(u,u). \qquad \square\end{aligned}$$

Remark 3. In the above lemma, we use Assumption 5, which states that the mesh sizes are comparable between neighboring subdomains. On any interface of two

subdomains, denote by h_m and h_{nm} the mesh sizes of the mortar subdomain and the nonmortar subdomain, respectively. If they satisfy

$$h_m \leq C h_{nm}, \tag{4.24}$$

then the result of Lemma 11 holds without the assumption of comparable meshes between neighboring subdomains. However, condition (4.24) is the opposite of the one considered in previous work on the mortar methods; see [17, sect. 1.5.3].

By combining the bound in Lemma 11 with Lemma 1 and the upper bound (3.5), we obtain the following condition number bound.

THEOREM 1. *With Assumptions* 1 *and* 2*, the two-level additive algorithm satisfies*

$$\kappa\left(\sum_{i=0}^{N} T_i\right) \leq C \max_{i=1,\ldots,N}\left\{\left(1+\frac{H_i}{\delta_i}\right)\right\},$$

where C depends on the constant ω in (3.5)*.*

For the general case, we bound the term in (4.22) by using Lemma 8:

$$\sum_{j=1,5} \|\widetilde{\theta}_i(w|_{\Omega_{k_j}} - w|_{\Omega_{k_3}})\|^2_{0,\delta_{n,j}} \leq C h_{k_3} \log\left(\frac{H_{k_3}}{h_{k_3}}\right) \sum_{j=1,3,5} |w|^2_{1,\Omega_{k_j}}.$$

This gives the bound in the general case:

$$\sum_{i=0}^{N} a^{\Gamma}(u_i,u_i) \leq C \max_{i=1,\ldots,N}\left\{\left(1+\frac{H_i}{\delta_i}\right) \max_{\Omega_{k_l}\cap\, \mathrm{supp}(V_i^h)\neq\emptyset}\left\{\log\left(\frac{H_{k_l}}{h_{k_l}}\right)\right\}\right\} a^{\Gamma}(u,u),$$

where $\mathrm{supp}(V_i^h)$ denotes the support of the functions in the space V_i^h. By combining this bound with Lemma 1 and the upper bound (3.5), we obtain the following condition number bound.

THEOREM 2. *Without Assumptions* 1 *and* 2*, the two-level additive algorithm satisfies*

$$\kappa\left(\sum_{i=0}^{N} T_i\right) \leq C \max_{i=1,\ldots,N}\left\{\left(1+\frac{H_i}{\delta_i}\right) \max_{\Omega_{k_l}\cap\, supp(V_i^h)\neq\emptyset}\left\{log\left(\frac{H_{k_l}}{h_{k_l}}\right)\right\}\right\},$$

where C depends on the constant ω in (3.5)*.*

REFERENCES

[1] Y. ACHDOU AND Y. MADAY, *The mortar element method with overlapping subdomains*, SIAM J. Numer. Anal., 40 (2002), pp. 601–628.

[2] F. B. BELGACEM, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[3] F. B. BELGACEM, *Discretisations 3D Non Conformes pour la Méthode de Decomposition de Domaine des Elément avec Joints: Analyse Mathématique et Mise en Œvre pour le Probleme de Poisson*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, Tech. report HI-72/93017, Electricité de France, 1993.

[4] F. B. BELGACEM AND Y. MADAY, *Adaptation de la méthode des éléments avec joints au couplage spectral élńents finis en dimension* 3*. Etude de l'érreur pour l'équation de Poisson*, Tech. report HI-72/7095, Electricité de France, 1992.

[5] F. B. BELGACEM AND Y. MADAY, *The mortar element method for three dimensional finite elements*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.

[6] C. Bernardi and Y. Maday, *Spectral, spectral element, and mortar element methods*, in Theory and Numerics of Differential Equations (Durham, 2000), Universitext, Springer-Verlag, Berlin, 2001, pp. 1–57.

[7] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Collège de France Seminar, H. Brezis and J.-L. Lions, eds., Pitman Res. Notes Math. Ser. 299, Longman Scientific, Harlow, 1994, pp. 13–51. This paper appeared as a technical report about five years earlier.

[8] S. C. Brenner, *Lower bounds of two-level additive Schwarz preconditioners with small overlap*, SIAM J. Sci. Comput., 21 (2000), pp. 1657–1669.

[9] T. F. Chan, B. F. Smith, and J. Zou, *Overlapping Schwarz methods on unstructured meshes using nonmatching coarse grids*, Numer. Math., 73 (1996), pp. 149–167.

[10] M. Dryja and O. B. Widlund, *An Additive Variant of the Schwarz Alternating Method for the Case of Many Subregions*, Tech. report 339, also Ultracomputer Note 131, Department of Computer Science, Courant Institute, New York, 1987.

[11] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[12] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.

[13] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.

[14] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[15] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer-Verlag, Berlin, 2005.

[16] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[17] B. I. Wohlmuth, *Discretization methods and iterative solvers based on domain decomposition*, Lecture Notes in Comput. Sci. Engrg. 17, Springer-Verlag, Berlin, 2001.

[18] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1535–1555

SUBSPACE TRUST-REGION METHODS FOR LARGE BOUND-CONSTRAINED NONLINEAR EQUATIONS*

STEFANIA BELLAVIA† AND BENEDETTA MORINI†

Abstract. Trust-region methods for solving large bound-constrained nonlinear systems are considered. These allow for spherical or elliptical trust regions where the search for an approximate solution is restricted to a low-dimensional space. A general formulation for these methods is introduced and global and superlinear/quadratic convergence is shown under standard assumptions. Viable approaches for implementation in conjunction with Krylov methods are discussed and the practical performance of the resulting algorithms is shown.

Key words. bound-constrained nonlinear systems, subspace trust-region methods, inexact Newton step, Krylov subspace methods

AMS subject classifications. 65H10, 65F10, 90C06

DOI. 10.1137/040611951

1. Introduction.

A number of applications arising in chemical engineering [18, 21], power engineering [37], and PDE-constrained optimization [4] are naturally stated as large constrained nonlinear systems. In particular, systems where the variables are subjected to lower and upper bounds are fairly general because sets of algebraic equations and inequalities and the Karush–Kuhn–Tucker (KKT) systems can be cast in such form.

This paper is concerned with the development of a trust-region method for solving large bound-constrained nonlinear systems

$$F(x) = 0, \qquad x \in \Omega. \tag{1.1}$$

Here $F : X \to \mathbb{R}^n$ is a continuously differentiable mapping, $X \subseteq \mathbb{R}^n$ is an open set containing the feasible region Ω, and Ω is an n-dimensional box, $\Omega = \{x \in \mathbb{R}^n : l \leq x \leq u\}$. These inequalities are meant componentwise and $l \in (\mathbb{R} \cup -\infty)^n$, $u \in (\mathbb{R} \cup \infty)^n$.

The development of globally convergent methods for large unconstrained nonlinear systems has received a great deal of attention; see, e.g., [2, 6, 7, 15, 23, 26, 27]. The methods proposed suggest that the search directions employed in the global strategy might belong to a low-dimensional subspace as such directions may often be computed cheaply. Therefore, they avoid the factorization of the matrices involved and consider the combination of global strategies such as linesearch techniques and model trust-region algorithms with Krylov methods. The resulting procedures belong to the framework of the inexact Newton methods [12].

The main computational effort in trust-region methods is the solution of the so-called trust-region problem, which is to find the minimizer of some model of the objective function within a region where the model adequately reflects the objective function. For small and medium problems, solving the trust-region problem relies on

*Received by the editors July 20, 2004; accepted for publication (in revised form) February 27, 2006; published electronically August 7, 2006. This work was supported by MIUR, Rome, Italy, through "Progetti di Ricerca Interesse Nazionale" (PRIN) and by INDAM-GNCS, Italy.
http://www.siam.org/journals/sinum/44-4/61195.html

†Dipartimento di Energetica "S. Stecco," Università di Firenze, via C. Lombroso 6/17, 50134 Firenze, Italy (stefania.bellavia@unifi.it, benedetta.morini@unifi.it).

matrix factorization. When n is large, several authors have suggested restricting the search for an approximate solution to such a problem to a low-dimensional subspace. Thus, the full space trust-region problem is replaced with a subspace trust-region problem and a large overhead of computing is avoided; see [10] and the references therein. Proposed approaches include the truncated conjugate gradient method [32, 33], the truncated Lanczos approach [9, 17], the two-dimensional subspace minimization [5, 8], and a subspace dogleg method [6, 7].

Numerical methods for problem (1.1) differ from the procedures for unconstrained nonlinear systems in several respects. They are augmented with strategies that enforce feasibility of the iterates. In addition, major modifications in the globalization techniques are necessary. We are aware of the trust-region methods [1, 3, 16, 34, 35] tailored for small- and medium-size problems and of the procedures [28, 29] appropriate for large systems. Focusing on the approaches for large problems, Qi, Tong, and Li [28] propose an active set projected trust-region algorithm for bound-constrained nonlinear systems. As a result of the active set strategy, the trust-region problem may be of reduced dimension, which is potentially cheaper when the method is applied to large problems. The method by Qi, Qi, and Sun [29] concerns the solution of the KKT systems. The trust-region problem is built around those components of the current iterates which are far from the boundary of the positive orthant, and it is solved by the truncated conjugate gradient method.

In this paper we introduce a prototype subspace trust-region method for large bound-constrained nonlinear systems. Our proposal is to investigate the idea of solving the trust-region problem in a small subspace while still attaining global and local fast convergence. Both spherical and elliptical trust-regions are allowed. To ensure global convergence properties we use a generalized Cauchy step. Fast local convergence relies on mild conditions on the subspace and is independent of the way of computing an approximate trust-region solution. At each iteration the trial step used to compute the new iterate is a linear combination of the generalized Cauchy step and the approximate trust-region solution.

The general scheme proposed serves as a paradigm for some specific implementations. In particular, the theoretical results obtained suggest ways to implement it by using Krylov solvers [30]. The first proposal is a two-dimensional subspace strategy. The second is a dogleg subspace strategy in conjunction with the iterative linear solver GMRES [31]. Both strategies compute an approximate solution of the related subspace trust-region problem with a low computational cost and require matrix-vector products only. In this regard, we remark that the computation of the generalized Cauchy point calls for the product of the transpose of the Jacobian of F with vectors. Thus, the proposed strategies cannot be implemented in a matrix-free manner, i.e., without computing the whole Jacobian matrix. On the other hand if the Jacobian of F is not available, these products can be effectively computed by using software for automatic differentiation [36].

We mention that [6, 7] propose a matrix-free Newton-GMRES dogleg strategy for unconstrained nonlinear systems where the Cauchy point is replaced by the steepest descent direction in a space generated by GMRES. But, in our opinion, this approach cannot be easily extended to constrained problems.

In section 2 we describe the main features of a trust-region method for problem (1.1), and in section 3 we propose a prototype method for large problems. In section 4 we provide global and local convergence properties. In section 5 we discuss ways in which an implementation of our procedure may be developed. Finally, in

section 6 we provide computational experiments showing the practical performance of our algorithm.

1.1. Notation. Throughout the paper we use the following notation. For any mapping $F : X \to \mathbb{R}^n$, differentiable at a point $x \in X \subset \mathbb{R}^n$, the Jacobian matrix of F at x is denoted by $F'(x)$ and $F(x_k)$ is denoted by F_k. To represent the ith component of x, the symbol $(x)_i$ is used but, when clear from the context, the brackets are omitted. For any vector $y \in \mathbb{R}^n$, the 2-norm is denoted by $\|y\|$ and the open ball with center y and radius ρ is indicated by $B_\rho(y)$, i.e., $B_\rho(y) = \{x : \|x - y\| < \rho\}$. The identity matrix of dimension n is denoted by I.

2. Preliminaries. In this section we provide the essential features of a trust-region method for the solution of (1.1). The sequence $\{x_k\}$ generated is expected to converge to a point which solves the optimization problem

$$\min_{x\in\Omega} f(x) = \min_{x\in\Omega} \frac{1}{2}\|F(x)\|^2. \tag{2.1}$$

In fact, the solutions to (1.1) solve the constrained minimization problem (2.1). A solution x^* of (2.1) satisfies

$$D^{-2}(x^*)\nabla f(x^*) = 0, \tag{2.2}$$

where $\nabla f(x) = F'(x)^T F(x)$, $D(x)$ is the diagonal scaling matrix

$$D(x) = \operatorname{diag}(|v_1(x)|^{-1/2}, |v_2(x)|^{-1/2}, \ldots, |v_n(x)|^{-1/2}), \tag{2.3}$$

and

$$v_i(x) = \begin{cases} x_i - u_i & \text{if} \quad (\nabla f(x))_i < 0 \quad \text{and} \quad u_i < \infty, \\ x_i - l_i & \text{if} \quad (\nabla f(x))_i > 0 \quad \text{and} \quad l_i > -\infty, \\ \min\{x_i - l_i,\, u_i - x_i\} & \text{if} \quad (\nabla f(x))_i = 0 \quad \text{and} \quad l_i > -\infty \text{ or } u_i < \infty, \\ 1 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$; see [11].

Numerical methods for (1.1) need well-angled directions to handle the bounds. In particular, search directions biased toward the interior of Ω are required. This way, sufficiently large steps in these directions are allowed before violating the constraints. If x_k lies in Ω, the following scaled gradient of f is well angled with respect to the bounds:

$$d_k = -D_k^{-2}\nabla f_k.$$

This is due to the fact that D_k^{-2} penalizes the step ∇f_k, preventing a step directly toward a boundary point. Moreover, by (2.2) d_k monitors the progress toward a solution of problem (2.1).

In a framework for (1.1), we consider the trust-region problem

$$\min_{p\in\mathbb{R}^n}\{m_k(p) : \quad \|G_k\, p\| \le \Delta_k\}, \tag{2.4}$$

where Δ_k is the current trust-region radius, m_k is the quadratic model for f at x_k,

$$m_k(p) = \frac{1}{2}\|F_k + F_k' p\|^2,$$

and $G_k = G(x_k) \in \mathbb{R}^{n\times n}$ with $G : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ defined as

$$G(x) = I \qquad \text{or} \qquad G(x) = D(x).$$

The first choice of G is used in [28, 34, 35] and yields the standard spherical trust-region problem. The choice $G(x) = D(x)$, $x \in \mathbb{R}^n$, has been considered in [1, 3] and gives rise to an elliptical trust region. In this case, for decreasing values of Δ_k, the solution to problem (2.4) tends to become parallel to d_k.

For all the proposed methods, the iterates x_k are forced to belong to Ω in order to deal with problems where F is not defined outside Ω. Moreover, since D is not defined on the boundary of Ω, the methods given in [1, 3] generate strictly feasible iterates, $x_k \in \text{int}(\Omega) = \{x \in \mathbb{R}^n : \ l < x < u\}$.

To find the next iterate, a key role is played by a so-called generalized Cauchy step $p_c(\Delta_k)$ depending on the scaled gradient d_k. The vector $p_c(\Delta_k)$ has the form

$$p_c(\Delta_k) = \tau_k \, d_k, \tag{2.5}$$

and it is such that $x_k + p_c(\Delta_k) \in \text{int}(\Omega)$. The value of τ_k is fixed as follows. Consider

$$\hat{\tau}_k = \min\left\{ \frac{\|D_k^{-1}\nabla f_k\|^2}{\|F_k' D_k^{-2}\nabla f_k\|^2}, \frac{\Delta_k}{\|G_k D_k^{-2}\nabla f_k\|} \right\} = \underset{\|\tau G_k d_k\| \le \Delta_k}{\text{argmin}} \ m_k(\tau d_k). \tag{2.6}$$

If $x_k + \hat{\tau}_k d_k \in \text{int}(\Omega)$, we let $\tau_k = \hat{\tau}_k$ in (2.5). Otherwise we let λ_k be the stepsize along d_k to the boundary, i.e.,

$$\lambda_k = \min_{1\le i\le n} \Lambda_i, \quad \text{where} \quad \Lambda_i = \begin{cases} \max\{\frac{l_i - (x_k)_i}{(d_k)_i}, \frac{u_i - (x_k)_i}{(d_k)_i}\} & \text{if } (d_k)_i \neq 0, \\ \infty & \text{if } (d_k)_i = 0, \end{cases}$$

and set τ_k smaller than λ_k. Summarizing, the parameter τ_k is given by

$$\tau_k = \begin{cases} \hat{\tau}_k & \text{if } x_k + \hat{\tau}_k d_k \in \text{int}(\Omega), \\ \theta\lambda_k, & \theta \in (0,\, 1), \text{ otherwise.} \end{cases} \tag{2.7}$$

In [11], it has been shown that global convergence to a first-order stationary point of (2.1) depends on obtaining, at each iteration, at least as much decrease in m_k as a fixed fraction of the decrease attained by the generalized Cauchy step $p_c(\Delta_k)$. In particular, letting $p(\Delta_k)$ be the step taken to update x_k, $p(\Delta_k)$ must satisfy the following condition

$$\rho_c(p(\Delta_k)) = \frac{m_k(0) - m_k(p(\Delta_k))}{m_k(0) - m_k(p_c(\Delta_k))} \ge \beta_1 \tag{2.8}$$

for a given constant $\beta_1 \in (0, 1)$.

Finally, as for the unconstrained problems, the sufficient improvement condition

$$\rho_f(p(\Delta_k)) = \frac{f(x_k) - f(x_k + p(\Delta_k))}{m_k(0) - m_k(p(\Delta_k))} \ge \beta_2 \tag{2.9}$$

is required to hold for a given constant $\beta_2 \in (0, 1)$. Namely, if (2.9) is satisfied, then $p(\Delta_k)$ is accepted, the new iterate $x_{k+1} = x_k + p(\Delta_k)$ is formed, and the trust-region radius may be increased. Otherwise, $p(\Delta_k)$ is rejected and Δ_k is shrunk.

3. A paradigm method for large-scale problems. In this section we present a general trust-region scheme for large bound-constrained nonlinear systems. Since the main source of computational effort of a trust-region algorithm is the work for solving problem (2.4), we replace (2.4) by the following subspace trust-region problem:

$$\min_{p\in S_k}\{m_k(p): \quad \|G_k\, p\| \le \Delta_k\}. \tag{3.1}$$

In fact, for a small subspace S_k of $\mathbb{R}^n$ the solution of (3.1) can be computed cheaply. At each iteration our scheme includes the choice of the subspace S_k, the solution of the subspace trust-region problem (3.1), and the construction of a step which combines the generalized Cauchy step and the subspace trust-region solution.

Our subspace model trust-region approach is based upon finding a small-dimension subspace S_k of $\mathbb{R}^n$ such that the minimum value of m_k on S_k is a fraction of $m_k(0)$. In particular, we fix S_k so that

$$p_k = \operatorname*{argmin}_{p\in S_k} m_k(p), \qquad m_k(p_k) \le \eta_k^2\, m_k(0), \tag{3.2}$$

with $\eta_k \in [0, 1)$. Clearly, by (3.2)

$$F_k'\, p_k = -F_k + r_k, \qquad \|r_k\| \le \eta_k \|F_k\|, \tag{3.3}$$

i.e., p_k is an inexact Newton step for the problem $F(x) = 0$; see [12].

To provide a flexible scheme, here we deliberately do not specify how to determine S_k and the solution $p_{tr}(\Delta_k)$ to (3.1). In section 5 we will show that these tasks can be readily implemented in different ways.

In regard to the problem of finding the actual step from x_k, our aim is to find a step $p(\Delta_k)$ producing a strictly feasible iterate $x_{k+1} = x_k + p(\Delta_k)$ and a sufficient reduction in the values of both the model function m_k and the objective function f.

To maintain strict feasibility we employ the interior modification of the projection onto Ω proposed in [19]. Namely, we form the vector

$$\bar{p}_{tr}(\Delta_k) = \alpha_k(P(x_k + p_{tr}(\Delta_k)) - x_k), \tag{3.4}$$

where $\alpha_k \in (0, 1)$ and $P(x)$ is the classical projection of x onto Ω, i.e., $(P(x))_i = \max\{l_i, \min\{x_i, u_i\}\}$, $i = 1, \ldots, n$. Clearly, $x_k + \bar{p}_{tr}(\Delta_k) \in \operatorname{int}(\Omega)$ and $(p_{tr}(\Delta_k))_i$ and $(\bar{p}_{tr}(\Delta_k))_i$ have the same sign. In fact, the point $x_k + \bar{p}_{tr}(\Delta_k)$ is the classical projection of $x_k + p_{tr}(\Delta_k)$ onto Ω, followed by a small step toward the interior of Ω. We remark that

$$\|\bar{p}_{tr}(\Delta_k)\| < \|p_{tr}(\Delta_k)\|. \tag{3.5}$$

Then we follow along the lines of [3] and seek a vector of the form

$$p(\Delta_k) = t\, p_c(\Delta_k) + (1 - t)\bar{p}_{tr}(\Delta_k), \qquad t \in [0, 1), \tag{3.6}$$

satisfying (2.8). Specifically, if $\rho_c(\bar{p}_{tr}(\Delta_k)) \ge \beta_1$, we take $t = 0$. Otherwise, since $m_k(t\, p_c(\Delta_k) + (1 - t)\, \bar{p}_{tr}(\Delta_k))$ is a quadratic function in t and $\rho_c(p_c(\Delta_k)) = 1$, it is easy to see that there exists $t \in (0, 1)$ such that $p(\Delta_k)$ satisfies $\rho_c(p(\Delta_k)) = \beta_1$. Next, we summarize the process for finding $p(\Delta_k)$.

ALGORITHM I. Finding a step that satisfies the model decrease (2.8).
Input parameters: $x_k \in \mathrm{int}(\Omega)$, $\bar{p}_{tr}(\Delta_k)$, $p_c(\Delta_k)$.

1. Set $t = 0$.
2. If $\rho_c(\bar{p}_{tr}(\Delta_k)) < \beta_1$
   Compute $u_1 = F_k' \, p_c(\Delta_k)$, $u_2 = F_k' \bar{p}_{tr}(\Delta_k)$, $u = u_1 - u_2$, $z = -F_k - u_2$,
   $$w = (z^T u)^2 - 2\,\|u\|^2 \left(F_k^T(u_2 - \beta_1 u_1) + \|u_2\|^2/2 - \|u_1\|^2/2\right).$$
   Set $t = (\, z^T u - w^{\frac{1}{2}})/\|u\|^2$.
3. Compute $p(\Delta_k)$ by (3.6).

We point out that the inclusion of $p_c(\Delta_k)$ in (3.6) ensures the existence of a vector $p(\Delta_k)$ satisfying (2.8). Then it enables the method to converge globally. Furthermore, the use of $\bar{p}_{tr}(\Delta_k)$ in (3.6) and suitable choices of the sequences $\{\eta_k\}$ and $\{\alpha_k\}$ yield rapid local convergence. The convergence analysis provided in the next section will highlight these features.

Below we summarize the overall procedure named the subspace interior affine trust-region (SIATR) method.

SIATR METHOD.
Input parameters: the starting point $x_0 \in \mathrm{int}(\Omega)$, the function G, $\Delta_{min} > 0$, the initial trust-region size $\bar{\Delta}_0 \geq \Delta_{min}$, β_1, β_2, δ, $\theta \in (0, 1)$.

For $k = 0, 1, \ldots$

1. Set $\Delta_k = \bar{\Delta}_k/\delta$.
2. Choose $\alpha_k \in (0, 1)$, $\eta_k \in [0, 1)$.
3. Repeat
   3.1 Set $\Delta_k = \delta\Delta_k$.
   3.2 Find $S_k \subset \mathbb{R}^n$ s.t. (3.2) holds.
   3.3 Compute the solution $p_{tr}(\Delta_k)$ to (3.1).
   3.4 Form $\bar{p}_{tr}(\Delta_k)$ by (3.4).
   3.5 Compute $p_c(\Delta_k)$ by (2.5) and (2.7).
   3.6 Find $p(\Delta_k)$ by Algorithm I.
   Until $\rho_f(p(\Delta_k)) \geq \beta_2$
4. Set $x_{k+1} = x_k + p(\Delta_k)$.
5. Choose $\bar{\Delta}_{k+1} \geq \Delta_{min}$.

In the SIATR method, $\bar{\Delta}_k$ is the initial value of the trust-region radius at the kth iteration. To develop our convergence analysis, we force $\bar{\Delta}_k$ to be greater than or equal to a fixed threshold $\Delta_{min} > 0$ for all $k \geq 0$. In this regard, we remark that the strategy for choosing $\bar{\Delta}_{k+1}$ does not affect our convergence results and that in step 5 it can be chosen following classical strategies based on the agreement between the model function m_k and the function f at iteration k.

4. Convergence analysis. In this section we develop a theoretical foundation for the SIATR method. We assume the following.

Assumption 1. F' is Lipschitz continuous in $L = \cup_{k=0}^{\infty} \{x \in X \ : \ \|x - x_k\| \leq r\}$, $r > 0$, with constant $2\gamma_L$.

Assumption 2. $\|F'\|$ is bounded above on L and $\chi_J = \sup_{x \in L} \|F'(x)\|$.

We begin studying the features of $p(\Delta_k)$. First, if $\|\nabla f_k\| \neq 0$, then condition (2.9) is met after a finite number of repetitions of step 3. This can be proved following along the lines of [3, Lemma 3.2]. Second, by Algorithm I we have $\rho_c(p(\Delta_k)) \geq \rho_c(\bar{p}_{tr}(\Delta_k))$ i.e.,

$$m_k(p(\Delta_k)) \leq m_k(\bar{p}_{tr}(\Delta_k)). \tag{4.1}$$

Third, the decrease attained in the value of m_k by $p(\Delta_k)$ is given in the following result.

LEMMA 4.1. *Assume that* $\|\nabla f_k\| \neq 0$. *If* $p(\Delta_k)$ *satisfies* (2.8), *then*

$$m_k(0) - m_k(p(\Delta_k)) \geq \frac{1}{2}\beta_1 \|D_k^{-1}\nabla f_k\| \min\left\{\frac{\Delta_k}{\|G_k D_k^{-1}\|}, \frac{\|D_k^{-1}\nabla f_k\|}{\|F_k' D_k^{-1}\|^2}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\right\}.$$

Proof. To prove the thesis we provide a lower bound for $m_k(0) - m_k(p_c(\Delta_k))$. By (2.5) we know that $p_c(\Delta_k)$ takes the form $p_c(\Delta_k) = \tau_k d_k$. Suppose $\tau_k = \hat{\tau}_k$ with $\hat{\tau}_k = \Delta_k/\|G_k D_k^{-2}\nabla f_k\|$. Since (2.6) implies $\hat{\tau}_k \leq \|D_k^{-1}\nabla f_k\|^2/\|F_k' D_k^{-2}\nabla f_k\|^2$, we get

$$m_k(0) - m_k(p_c(\Delta_k)) = \hat{\tau}_k\left(\|D_k^{-1}\nabla f_k\|^2 - \frac{1}{2}\hat{\tau}_k\|F_k' D_k^{-2}\nabla f_k\|^2\right) \geq \frac{1}{2}\hat{\tau}_k\|D_k^{-1}\nabla f_k\|^2$$

$$\geq \frac{1}{2}\frac{\Delta_k}{\|G_k D_k^{-1}\|}\|D_k^{-1}\nabla f_k\|. \tag{4.2}$$

In the case that either $\tau_k = \hat{\tau}_k = \|D_k^{-1}\nabla f_k\|^2/\|F_k' D_k^{-2}\nabla f_k\|^2$ or $\tau_k = \theta\lambda_k$, we know from [1, Lemma 3.3] that

$$m_k(0) - m_k(p_c(\Delta_k)) \geq \frac{1}{2}\frac{\|D_k^{-1}\nabla f_k\|^2}{\|F_k' D_k^{-1}\|^2}, \tag{4.3}$$

$$m_k(0) - m_k(p_c(\Delta_k)) \geq \frac{1}{2}\,\theta\,\frac{\|D_k^{-1}\nabla f_k\|^2}{\|\nabla f_k\|_\infty}, \tag{4.4}$$

respectively. From (2.8), (4.2), (4.3), and (4.4) the thesis follows. ∎

Now we can formalize the global convergence properties of the SIATR method. They essentially derive from forcing (2.8) and can be easily proved following along the lines of [3, Theorem 3.1] and using Lemma 4.1.

THEOREM 4.1. *If the sequence* $\{x_k\}$ *generated by the SIATR method is bounded, then all the limit points of* $\{x_k\}$ *are stationary points for the problem* (2.1), *i.e.,*

$$\lim_{k\to\infty} \|D_k^{-1}\nabla f_k\| = 0.$$

Further, if there exists a limit point $x^* \in \operatorname{int}(\Omega)$ *of* $\{x_k\}$ *such that* $F'(x^*)$ *is nonsingular, then* $\|F_k\| \to 0$ *and all the accumulation points of* $\{x_k\}$ *solve problem* (1.1).

Moreover, if there exists a limit point $x^* \in \Omega$ such that $F(x^*) = 0$ and $F'(x^*)$ is invertible, then $\{x_k\}$ converges to x^*. To prove this fact, we first recall some technical results.

LEMMA 4.2. *Let* $x^* \in \Omega$ *be a limit point of the sequence* $\{x_k\}$ *generated by the SIATR method such that* $F(x^*) = 0$ *and* $F'(x^*)$ *is nonsingular. Let* $K_1 = 2\,\|F'(x^*)\|$, $K_2 = 2\,\|F'(x^*)^{-1}\|$, *and* $\mu = \max\{K_1, K_2\}/2$ *and let* $\Gamma \in (0, 1/\mu)$ *be given. Then there exists* $\rho > 0$ *so that if* $x \in B_\rho(x^*)$, *then* $x \in L$ *and*

$$\|x - x^*\| \leq K_2\,\|F(x)\|, \tag{4.5}$$

$$\|F(x)\| \leq K_1\,\|x - x^*\|, \tag{4.6}$$

$$\|F'(x)^{-1}\| \leq K_2, \tag{4.7}$$

$$\|F(x) - F(z) - F'(z)(x - z)\| \leq \Gamma\|x - z\|^2 \quad \textit{for all } z \in B_\rho(x^*). \tag{4.8}$$

Proof. The existence of $\rho > 0$ so that if $\|x - x^*\| \leq \rho$, then $x \in L$ is shown in [3, Lemma 3.3]. Conditions (4.5)–(4.7) follows from [22, Lemma 4.3.1]. Finally, (4.8) is given in [25, Lemma 3.2.10]. ∎

The next theorem shows the convergence of the sequence $\{x_k\}$.

THEOREM 4.2. *Assume that x^* is a limit point of the sequence $\{x_k\}$ generated by the SIATR method such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular. Then $\{x_k\}$ converges to x^*.*

Proof. First, note that, by (4.5), there exists a neighborhood of x^* where $\|F(x)\| > 0$ if $x \neq x^*$. This implies that x^* is an isolated solution of (1.1). Moreover, since the sequence $\{\|F_k\|\}$ is monotone decreasing, it is convergent. Then the assumption $F(x^*) = 0$ implies $\|F_k\| \to 0$ and every limit point of $\{x_k\}$ is a solution of (1.1). Then, as the point x^* is an isolated solution of (1.1), it is an isolated limit point of the sequence $\{x_k\}$.

Let ρ be as in Lemma 4.2, let $\{x_{k_j}\}$ be a subsequence such that $x_{k_j} \to x^*$, and let j_0 be the index such that $x_{k_j} \in B_\rho(x^*) \cap \Omega$ when $k_j \geq k_{j_0}$. Note that $\lim_{k_j \to \infty} \|\nabla f_{k_j}\| = 0$. Assume $k_j \geq k_{j_0}$. To prove the thesis, we examine the asymptotic behavior of $p_c(\Delta_{k_j})$ and $\bar{p}_{tr}(\Delta_{k_j})$. By (2.5) and $\tau_k \leq \frac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|F'_{k_j} D_{k_j}^{-2}\nabla f_{k_j}\|^2}$, we obtain the following:

$$\begin{aligned} \|p_c(\Delta_{k_j})\| &\leq \frac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|D_{k_j}^{-2}\nabla f_{k_j}\|}\|{F'_{k_j}}^{-1}\|^2 = \frac{\nabla f_{k_j}^T (D_{k_j}^{-2}\nabla f_{k_j})}{\|D_{k_j}^{-2}\nabla f_{k_j}\|}\|{F'_{k_j}}^{-1}\|^2 \\ &\leq \|\nabla f_{k_j}\|\|{F'_{k_j}}^{-1}\|^2. \end{aligned} \tag{4.9}$$

By (4.7) and $\lim_{k_j \to \infty} \|\nabla f_{k_j}\| = 0$, we conclude that $\lim_{k_j \to \infty} \|p_c(\Delta_{k_j})\| = 0$.

Regarding $p_{tr}(\Delta_{k_j})$, by construction $m_k(p_{tr}(\Delta_{k_j})) \leq m_k(0)$ i.e., letting $\hat{r}_{k_j} = F'_{k_j} p_{tr}(\Delta_{k_j}) + F_{k_j}$ we have $\|\hat{r}_{k_j}\| \leq \|F_{k_j}\|$. Then

$$\|p_{tr}(\Delta_{k_j})\| = \|{F'}_{k_j}^{-1}(-F_{k_j} + \hat{r}_{k_j})\| \leq 2\,\|{F'}_{k_j}^{-1}\|\,\|F_{k_j}\| \leq 2\,K_2\,\|F_{k_j}\|.$$

Thus, $\|F_k\| \to 0$ and (3.5) yield $\lim_{k_j \to \infty} \|\bar{p}_{tr}(\Delta_{k_j})\| = 0$.

Hence, we have $\lim_{k_j \to \infty} \|p(\Delta_{k_j})\| = 0$ and using [24, Lemma 4.10] we conclude that $\{x_k\}$ converges to x^*. ∎

We now move on to discuss the convergence rate issues. We make the additional hypothesis $\|G_k p_k\| \to 0$ as $k \to \infty$. In practice, this condition may fail to hold only when $G_k = D_k$ and x^* belongs to the boundary of Ω. On the other hand, it is guaranteed when $G_k = I$ or when $G_k = D_k$ and x^* lies in the interior of Ω. To show this, note that by (3.3) and (4.7) we get

$$\|p_k\| = \|{F'}_k^{-1}(-F_k + r_k)\| \leq (1 + \eta_k)\,\|{F'}_k^{-1}\|\,\|F_k\| \leq 2\,K_2\,\|F_k\|, \tag{4.10}$$

and this implies that $\|p_k\| \to 0$ as $k \to \infty$. Also, it is easy to see that $\|D_k\| \leq \sqrt{2/\bar{\rho}}$ whenever $x_k \in B_{\bar{\rho}/2}(x^*)$ with $x^* \in \mathrm{int}(\Omega)$ and $\bar{\rho}$ sufficiently small that $B_{\bar{\rho}}(x^*) \subset \mathrm{int}(\Omega)$ [1, Corollary 3.1]. Then $\|D_k p_k\| \to 0$ as $k \to \infty$ when $x^* \in \mathrm{int}(\Omega)$.

First, we prove that eventually, for Δ_k equal to the initial trust-region radius $\bar{\Delta}_k$, the trust-region constraint in (3.1) becomes inactive, i.e., $p_{tr}(\bar{\Delta}_k)$ is the minimizer p_k of m_k on S_k. Moreover, we study the features of $p_{tr}(\bar{\Delta}_k)$, $\bar{p}_{tr}(\bar{\Delta}_k)$, and $p(\bar{\Delta}_k)$ when x_k is sufficiently near to x^*. Then we will show how the choice of η_k's and α_k's affects the convergence rate of the SIATR method.

From now on, with γ_L and χ_J as in Assumptions 1 and 2 and K_1, K_2, and Γ as

in Lemma 4.2, we let

$$K^* = \|F'(x^*)\| \, \|F'(x^*)^{-1}\|, \tag{4.11}$$

$$\nu = 8\, K^* \, (K_2 \chi_J + 1), \tag{4.12}$$

$$\delta_k = K_2 \Gamma \nu^2 \|x_k - x^*\| + 4\, K^* \, \eta_k, \tag{4.13}$$

$$\psi_k = \chi_J \delta_k + \gamma_L \nu^2 \|x_k - x^*\| + K_1 (1 - \alpha_k), \tag{4.14}$$

$$\sigma_k = \max\{\psi_k,\, K_2 (\, \Gamma \nu^2 \|x_k - x^*\| + \psi_k)\}. \tag{4.15}$$

LEMMA 4.3. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that either*

- *$G_k = I$, $k \geq 0$, or*
- *$G_k = D_k$, $k \geq 0$, and $\|D_k p_k\| \to 0$ as $k \to \infty$.*

Then, for ρ as in Lemma 4.2, there exists $\rho_1 \leq \rho$ such that for all $x_k \in B_{\rho_1}(x^*) \cap \text{int}(\Omega)$,

$$p_{tr}(\bar{\Delta}_k) = p_k, \tag{4.16}$$

where p_k is given in (3.2) and $\bar{\Delta}_k$ is the initial trust-region radius at kth iteration. Further, when $x_k \in B_{\rho_1}(x^*) \cap \text{int}(\Omega)$ we have

$$\|F_k + F_k' p_{tr}(\bar{\Delta}_k)\| \leq K_1 \, \eta_k \|x_k - x^*\|, \tag{4.17}$$

$$\|\bar{p}_{tr}(\bar{\Delta}_k)\| < \|p_{tr}(\bar{\Delta}_k)\| \leq \nu \|x_k - x^*\|, \tag{4.18}$$

$$\|p(\bar{\Delta}_k)\| \leq \nu \|x_k - x^*\|. \tag{4.19}$$

Proof. The relationship (4.16) is proved by using the fact that $\bar{\Delta}_k \geq \Delta_{min}$, i.e., $\bar{\Delta}_k$ is bounded below from zero for each $k \geq 0$. Let $G_k = I$ for all $k \geq 0$. By (4.10) $\lim_{k \to \infty} \|p_k\| = 0$. Then, there exists $\rho_1 \leq \rho$ such that $\|p_k\| \leq \bar{\Delta}_k$ when $x_k \in B_{\rho_1}(x^*) \cap \text{int}(\Omega)$. Since p_k is feasible for the trust-region problem (3.1), the thesis (4.16) follows. Now consider the case $G_k = D_k$ for all $k \geq 0$. The assumption $\lim_{k \to \infty} \|D_k p_k\| = 0$ implies that there exists $\rho_1 \leq \rho$ such that p_k solves the trust-region problem (3.1) whenever $x_k \in B_{\rho_1}(x^*) \cap \text{int}(\Omega)$ and (4.16) again follows.

The remaining results are proved independently of the form of G_k. By (4.16) and (3.2) we obtain $\|F_k + F_k' p_{tr}(\bar{\Delta}_k)\| \leq \eta_k \|F_k\|$. Thus, (4.6) implies (4.17).

The result (4.18) is derived noting that by (3.5), (4.16), (4.10), (4.6), and (4.11) we get

$$\|\bar{p}_{tr}(\bar{\Delta}_k)\| < \|p_{tr}(\bar{\Delta}_k)\| \leq 2\, K_2 \, \|F_k\| \leq 8\, K^* \, \|x_k - x^*\|. \tag{4.20}$$

Then, by (4.12) relation (4.18) follows.

Finally, (4.9), (4.7), (4.6), and Assumption 2 yield $\|p_c(\bar{\Delta}_k)\| \leq 4 K^* K_2 \chi_J \|x_k - x^*\|$. Hence, by (3.6) and (4.20)

$$\|p(\bar{\Delta}_k)\| \leq \|p_c(\bar{\Delta}_k)\| + \|\bar{p}_{tr}(\bar{\Delta}_k)\| \leq 8\, K^* \, (K_2 \, \chi_J + 1) \, \|x_k - x^*\|.$$

This, along with (4.12), proves (4.19). ∎

LEMMA 4.4. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that either*

- *$G_k = I$, $k \geq 0$, or*
- *$G_k = D_k$, $k \geq 0$, and $\|D_k p_k\| \to 0$ as $k \to \infty$.*

Then, for ρ_1 as in Lemma 4.3 and $x_k \in B_{\rho_2}(x^*) \cap \mathrm{int}(\Omega)$, $\rho_2 \le \rho_1/(1+\nu)$, we have

$$\|F_k + F_k' \bar{p}_{tr}(\bar{\Delta}_k)\| \le \sigma_k \|x_k - x^*\|, \tag{4.21}$$

$$\|x_k + \bar{p}_{tr}(\bar{\Delta}_k) - x^*\| \le \sigma_k \|x_k - x^*\|, \tag{4.22}$$

where $\bar{\Delta}_k$ is the initial trust-region radius at the kth iteration.

Proof. Let $0 < \rho_2 \le \rho_1/(1+\nu)$ and let k be sufficiently large to have $x_k \in B_{\rho_2}(x^*) \cap \mathrm{int}(\Omega)$. To begin, note that from (4.5) and (4.8), any vector $x_k + q \in B_{\rho_1}(x^*)$ satisfies

$$\begin{aligned} \|x_k + q - x^*\| &\le K_2(\|F(x_k+q) - F_k - F_k' q\| + \|F_k + F_k' q\|) \\ &\le K_2(\Gamma \|q\|^2 + \|F_k + F_k' q\|). \end{aligned} \tag{4.23}$$

From (4.18) we get $\|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| \le \|x_k - x^*\| + \|p_{tr}(\bar{\Delta}_k)\| \le \rho_2(1+\nu) \le \rho_1$. Analogously, $\|x_k + \bar{p}_{tr}(\bar{\Delta}_k) - x^*\| \le \rho_1$. Hence, $x_k + p_{tr}(\bar{\Delta}_k)$ and $x_k + \bar{p}_{tr}(\bar{\Delta}_k)$ belong to $B_{\rho_1}(x^*)$. Further, from (4.23), (4.17), and (4.18) we get

$$\|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| \le \delta_k \|x_k - x^*\|, \tag{4.24}$$

where δ_k is given in (4.13).

Now, letting $\hat{p} = P(x_k + p_{tr}(\bar{\Delta}_k)) - x_k$, we derive an upper bound for $\|F_k + F_k' \hat{p}\|$. First, we note that $\|\hat{p}\| \le \|p_{tr}(\bar{\Delta}_k)\|$ and recall the nonexpansivity of the projection operator $P(\cdot)$, i.e., $\|x_k + \hat{p} - x^*\| \le \|x_k + p_{tr}(\bar{\Delta}_k) - x^*\|$. Then, by

$$\begin{aligned} F_k + F_k' \hat{p} &= F(x_k + \hat{p}) - F(x^*) + \int_0^1 (F'(x_k) - F'(x_k + t\hat{p}))\hat{p}\, dt \\ &= \int_0^1 F'(x^* + t(x_k + \hat{p} - x^*))(x_k + \hat{p} - x^*) dt \\ &\quad + \int_0^1 (F'(x_k) - F'(x_k + t\hat{p}))\hat{p}\, dt \end{aligned} \tag{4.25}$$

and using Assumptions 1–2, (4.24), and (4.18), we obtain

$$\begin{aligned} \|F_k + F_k' \hat{p}\| &\le \chi_J \|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| + \gamma_L \|p_{tr}(\bar{\Delta}_k)\|^2 \\ &\le (\chi_J \delta_k + \gamma_L \nu^2 \|x_k - x^*\|) \|x_k - x^*\|. \end{aligned} \tag{4.26}$$

Further, from (3.4), (4.26), (4.6), and (4.14) we get

$$\|F_k + F_k' \bar{p}_{tr}(\bar{\Delta}_k)\| \le \alpha_k \|F_k + F_k' \hat{p}\| + (1 - \alpha_k)\|F_k\| \tag{4.27}$$

$$\le \psi_k \|x_k - x^*\|, \tag{4.28}$$

and (4.15) yields the thesis (4.21).

Finally (4.22) is derived by (4.23), (4.18), (4.28), and (4.15), as follows:

$$\|x_k + \bar{p}_{tr}(\bar{\Delta}_k) - x^*\| \le K_2(\Gamma \nu^2 \|x_k - x^*\| + \psi_k)\, \|x_k - x^*\|. \qquad \square$$

4.1. Superlinear convergence. In this section we show that if $\eta_k \to 0$ and $\alpha_k \to 1$ as $k \to \infty$, then eventually the step $\bar{p}_{tr}(\bar{\Delta}_k)$ satisfies both conditions (2.8) and (2.9). Then, for k sufficiently large, $p(\bar{\Delta}_k) = \bar{p}_{tr}(\bar{\Delta}_k)$ and the actual step is not biased toward the generalized Cauchy step. Moreover,

- $x_k \to x^*$ superlinearly;
- $x_k \to x^*$ quadratically if $\eta_k = O(\|F_k\|)$ and $\alpha_k = 1 - O(\|F_k\|)$ as $k \to \infty$.

THEOREM 4.3. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that $\eta_k \to 0$, $\alpha_k \to 1$, as $k \to \infty$, and either*

- *$G_k = I$, $k \geq 0$, or*
- *$G_k = D_k$, $k \geq 0$, and $\|D_k p_k\| \to 0$ as $k \to \infty$.*

Then, eventually, $\bar{p}_{tr}(\bar{\Delta}_k)$ satisfies (2.8) and (2.9) and the sequence $\{x_k\}$ converges to x^* superlinearly. Moreover, if $\eta_k = O(\|F_k\|)$, $\alpha_k = 1 - O(\|F_k\|)$ as $k \to \infty$, the convergence rate is quadratic.

Proof. Let ζ be such that

$$\zeta < \min\left\{ \rho_2, \frac{1-\beta_2}{2K_2\Gamma\nu^2} \right\}, \tag{4.29}$$

with ρ_2 as in Lemma 4.4, K_2 and Γ as in Lemma 4.2, and ν as in (4.12). Let k be sufficiently large to have $x_k \in B_\zeta(x^*) \cap \text{int}(\Omega)$ and

$$\sigma_k < \frac{1}{K_2} \min\left\{ \sqrt{1-\beta_1}, \frac{1}{2} \right\}, \tag{4.30}$$

where σ_k is given in (4.15). This condition is met for k sufficiently large since

$$\sigma_k = O(\|x_k - x^*\| + \eta_k + (1-\alpha_k)), \qquad k \to \infty. \tag{4.31}$$

First, we prove that $\bar{p}_{tr}(\bar{\Delta}_k)$ satisfies (2.8), i.e., $p(\bar{\Delta}_k) = \bar{p}_{tr}(\bar{\Delta}_k)$. By (2.8), (4.21), (4.5), and (4.30), it follows that

$$\rho_c(\bar{p}_{tr}(\bar{\Delta}_k)) \geq 1 - \frac{\|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2}{\|F_k\|^2} \geq 1 - \sigma_k^2 K_2^2 > \beta_1. \tag{4.32}$$

Second, we prove that $\rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) \geq \beta_2$. Note that

$$\rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) = 1 - \frac{\|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2}{\|F_k\|^2 - \|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2} \tag{4.33}$$

and

$$\begin{aligned} \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2 &= \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k)) - F_k - F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2 \\ &+ 2(F(x_k + \bar{p}_{tr}(\bar{\Delta}_k)) - F_k - F_k'\bar{p}_{tr}(\bar{\Delta}_k))^T \\ &\times (F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)). \end{aligned}$$

Note that (4.29) implies $\zeta < 1/(2K_2\Gamma\nu^2)$ and (4.30) implies $2\sigma_k < 1/K_2$. Then by (4.8), (4.21), and (4.18) we get

$$\begin{aligned} \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2 &\leq \Gamma^2\|\bar{p}_{tr}(\bar{\Delta}_k)\|^4 + 2\sigma_k\Gamma\|\bar{p}_{tr}(\bar{\Delta}_k)\|^2\|x_k - x^*\| \\ &\leq \Gamma\nu^2(\Gamma\nu^2\|x_k - x^*\| + 2\sigma_k)\|x_k - x^*\|^3 \\ &\leq \frac{3\Gamma\nu^2}{2K_2}\|x_k - x^*\|^3. \end{aligned}$$

Further, from (4.5), (4.21), and (4.30) we get

$$\|F_k\|^2 - \|F_k + F_k'\bar{p}_{tr}(\bar{\Delta}_k)\|^2 \geq \left(\frac{1}{K_2^2} - \sigma_k^2\right)\|x_k - x^*\|^2 > \frac{3}{4K_2^2}\|x_k - x^*\|^2.$$

Therefore, by (4.33) and (4.29) we get

$$\rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) \geq 1 - 2K_2\Gamma\nu^2\|x_k - x^*\| > \beta_2.$$

Hence, $x_{k+1} = x_k + \bar{p}_{tr}(\bar{\Delta}_k)$ and from (4.22) we conclude $\|x_{k+1} - x^*\| \leq \sigma_k\|x_k - x^*\|$. The form of σ_k given in (4.31) ensures superlinear convergence rate if $\eta_k \to 0$ and $\alpha_k \to 1$ as $k \to \infty$. Moreover, if $\eta_k = O(\|F_k\|)$ and $1 - \alpha_k = O(\|F_k\|)$, by (4.5)–(4.6), we get $\sigma_k = O(\|x_k - x^*\|)$, and this yields the quadratic convergence rate. ∎

4.2. Linear convergence and norm weighted analysis. In this section we characterize the convergence order of the SIATR method dropping the assumption $\eta_k \to 0$ for $k \to \infty$. In particular, we let $\eta_k \leq \eta_{max} < \bar{\eta} < 1$, $k \geq 0$, and provide convergence results which are in accordance with those of inexact Newton methods for unconstrained nonlinear systems [12].

Following along the lines of the previous analysis, in the next theorem we show that if $\bar{\eta}$ is sufficiently small, then the sequence $\{x_k\}$ converges at a linear rate.

THEOREM 4.4. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that either*

- *$G_k = I$, $k \geq 0$, or*
- *$G_k = D_k$, $k \geq 0$, and $\|D_k p_k\| \to 0$ as $k \to \infty$.*

Then there exists $\bar{\eta} < 1$ such that if $\eta_k \leq \eta_{max} < \bar{\eta}$ for $k \geq 0$ and $\alpha_k \to 1$, as $k \to \infty$, then eventually $\bar{p}_{tr}(\bar{\Delta}_k)$ satisfies (2.8)–(2.9) and the sequence $\{x_k\}$ converges to x^* linearly.

Proof. Let k be sufficiently large to have $x_k \in B_\zeta(x^*) \cap \mathrm{int}(\Omega)$ with ζ given in (4.29). Since we intend to proceed as in Theorem 4.3, we need to ensure (4.30). Further, to provide linear convergence rate, σ_k given in (4.15) must be such that $\sigma_k < 1$.

Note that from (4.13)–(4.15) it follows that

$$\psi_k \leq \nu^2(\gamma_L + K_2\Gamma\chi_J)\|x_k - x^*\| + K_1(1 - \alpha_k) + 4\chi_J K^*\eta_{max},$$
$$\sigma_k \leq \max\{1,\, K_2\}\,\psi_k + K_2\Gamma\nu^2\|x_k - x^*\|.$$

Since $\lim_{k\to\infty}\alpha_k = 1$, $\lim_{k\to\infty}\|x_k - x^*\| = 0$, letting

$$\bar{\eta} < \min\left\{1,\, \frac{1}{4\,K^*\,\chi_J\max\{1,\,K_2\}},\, \frac{\min\{\sqrt{1-\beta_1},\,\frac{1}{2}\}}{4\,K^*\,K_2\,\chi_J\max\{1,\,K_2\}}\right\}, \tag{4.34}$$

we can take $\zeta > 0$ sufficiently small such that $\sigma_k < \min\{1,\, \frac{1}{K_2}\min\{\sqrt{1-\beta_1},\,\frac{1}{2}\}\}$ for $x_k \in B_\zeta(x^*) \cap \mathrm{int}(\Omega)$ and k sufficiently large. Thus, both $\sigma_k < 1$ and (4.30) hold, and proceeding as in Theorem 4.3, we conclude that $\bar{p}_{tr}(\bar{\Delta}_k)$ satisfies (2.8)–(2.9) and $x_{k+1} = x_k + \bar{p}_{tr}(\bar{\Delta}_k)$. Finally, (4.22) yields linear convergence rate. ∎

Trivially, (4.34) implies that $\bar{p}_{tr}(\bar{\Delta}_k)$ is not guaranteed to satisfy (2.8) for any $\bar{\eta} < 1$. On the other hand, if no other condition is imposed than requiring the sequence $\{\eta_k\}$ to be uniformly bounded away from 1, the linear convergence of the sequence $\{x_k\}$ depends on the norm used. Introducing the weighted norm

$$\|\cdot\|_* = \|F'(x^*)\cdot\|,$$

we prove that $\{x_k\}$ converges to x^* linearly in the sense that $\|x_{k+1} - x^*\|_* \leq \bar{\eta}\,\|x_k - x^*\|_*$ for any $\bar{\eta} < 1$. To prove this result we need to provide some useful bounds employing the weighted norm. With μ and Γ as in Lemma 4.2, we let

$$\omega_1 = 1/(1 - \Gamma\mu), \qquad \omega_2 = 1 + \Gamma\mu. \tag{4.35}$$

LEMMA 4.5. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that either*

- $G_k = I$, $k \geq 0$, *or*
- $G_k = D_k$, $k \geq 0$, *and* $\|D_k p_k\| \to 0$ *as* $k \to \infty$.

Then there exists $\rho_3 > 0$ so that if $x_k \in B_{\rho_3}(x^*) \cap \text{int}(\Omega)$, then

$$\|x_k - x^*\|_* \leq \omega_1 \|F_k\|, \tag{4.36}$$

$$\|F_k\| \leq \omega_2 \|x_k - x^*\|_*, \tag{4.37}$$

$$\|F_k + F_k' p_{tr}(\bar{\Delta}_k)\| \leq \eta_k \omega_2 \|x_k - x^*\|_*, \tag{4.38}$$

$$\|p_{tr}(\bar{\Delta}_k)\| \leq \nu_* \|x_k - x^*\|_*, \tag{4.39}$$

$$\|p(\bar{\Delta}_k)\| \leq \nu_* \|x_k - x^*\|_*, \tag{4.40}$$

with $\nu_* = 2\, K_2 \omega_2\, (K_2\, \chi_J + 1)$.

Proof. Note that

$$\frac{1}{\mu}\|y\| \leq \|y\|_* \leq \mu \|y\| \quad \text{for all } y \in \mathbb{R}^n. \tag{4.41}$$

Let ρ_1 be as in Lemma 4.3 and Γ as in Lemma 4.2. From the continuity of F' there exists $0 < \rho_3 < \rho_1$, so that

$$\|F'(x) - F'(x^*)\| \leq \Gamma, \qquad \|F'(x)^{-1} - F'(x^*)^{-1}\| \leq \Gamma \tag{4.42}$$

for $x \in B_{\rho_3}(x^*) \cap \text{int}(\Omega)$. In what follows, let $x_k \in B_{\rho_3}(x^*) \cap \text{int}(\Omega)$.

By (4.42), (4.41), and (4.35) we obtain (4.36) as follows:

$$\begin{aligned}
\|F_k\| &= \|F'(x^*)(x^* - x_k) - \int_0^1 (F'(x^* + t(x_k - x^*)) - F'(x^*))(x_k - x^*)\, dt\|, \\
&\geq \left| \|x_k - x^*\|_* - \left\| \int_0^1 (F'(x^* + t(x_k - x^*)) - F'(x^*))(x_k - x^*)\, dt \right\| \right| \\
&\geq (1 - \Gamma\mu)\|x_k - x^*\|_*.
\end{aligned}$$

Further, (4.37) follows from

$$\begin{aligned}
\|F(x)\| &\leq \int_0^1 \|F'(x^* + t(x_k - x^*)) - F'(x^*)\|\, \|x_k - x^*\| + \|F'(x^*)(x_k - x^*)\| dt \\
&\leq (\Gamma\mu + 1)\|x_k - x^*\|_*
\end{aligned}$$

and (4.35).

Since from Lemma 4.3 it follows that $p_{tr}(\bar{\Delta}_k) = p_k$, by (3.2) and (4.37) we trivially obtain (4.38). Moreover, from (4.10) and (4.37) we get $\|p_{tr}(\bar{\Delta}_k)\| \leq 2K_2\omega_2\|x_k - x^*\|_*$, and this yields (4.39). Finally, (4.9), Assumption 2, (4.7), and (4.37) yield $\|p_c(\bar{\Delta}_k)\| \leq K_2^2\, \chi_J\, \omega_2 \|x_k - x^*\|_*$. Then (4.40) is obtained by (3.6) and (3.5) as follows:

$$\|p(\bar{\Delta}_k)\| \leq \|p_c(\bar{\Delta}_k)\| + \|\bar{p}_{tr}(\bar{\Delta}_k)\| \leq 2\omega_2\, K_2\, (K_2\, \chi_J + 1)\, \|x_k - x^*\|_*. \qquad \square$$

Next we establish conditions under which $x_{k+1} = x_k + p(\bar{\Delta}_k)$ for k sufficiently large and for any $\bar{\eta} < 1$. Then the linear convergence rate of $\{x_k\}$ in the weighted norm is shown.

THEOREM 4.5. *Assume that there exists a solution x^* of (1.1) such that $F'(x^*)$ is nonsingular and that the sequence $\{x_k\}$ generated by the SIATR method converges to x^*. Suppose that $\eta_k \le \eta_{max} < \bar{\eta} < 1$, $k \ge 0$, $\alpha_k \to 1$, as $k \to \infty$ and either*

- *$G_k = I$, $k \ge 0$, or*
- *$G_k = D_k$, $k \ge 0$, and $\|D_k p_k\| \to 0$ as $k \to \infty$,*

If, for k sufficiently large,

$$P(x_k + p_{tr}(\bar{\Delta}_k)) - x_k = p_{tr}(\bar{\Delta}_k), \tag{4.43}$$

then $p(\bar{\Delta}_k)$ satisfies (2.9) and the sequence $\{x_k\}$ converges to x^* linearly in the sense that

$$\|x_{k+1} - x^*\|_* \le \bar{\eta}\, \|x_k - x^*\|_*.$$

Proof. Let $0 < \rho_4 \le \rho_3/(1+\nu)$, where ρ_3 is as in the proof of Lemma 4.5 and ν is given in (4.12). Let k be sufficiently large to have $x_k \in B_{\rho_4}(x^*) \cap \mathrm{int}(\Omega)$. From (4.19) it follows that $x_k + p(\bar{\Delta}_k)$ belongs to $B_{\rho_3}(x^*) \cap \mathrm{int}(\Omega)$. Let

$$\bar{\epsilon}_k = \omega_2(\alpha_k\, \eta_k + 1 - \alpha_k), \tag{4.44}$$

$$\epsilon_k = \omega_1(\Gamma \nu_*^2 \|x_k - x^*\|_* + \bar{\epsilon}_k). \tag{4.45}$$

By hypothesis, $\hat{p} = P(x_k + p_{tr}(\bar{\Delta}_k)) - x_k = p_{tr}(\bar{\Delta}_k)$. Then by (4.1), (4.27), (4.38), (4.37), and (4.44) it follows that

$$\begin{aligned} \|F_k + F_k' p(\bar{\Delta}_k)\| &\le \|F_k + F_k' \bar{p}_{tr}(\bar{\Delta}_k)\| \le \omega_2(\alpha_k \eta_k + 1 - \alpha_k)\|x_k - x^*\|_* \\ &= \bar{\epsilon}_k\, \|x_k - x^*\|_*. \end{aligned} \tag{4.46}$$

Using (4.36) and proceeding as in (4.23), we obtain

$$\|x_k + q - x^*\|_* \le \omega_1 \|F(x_k + q)\| \le \omega_1(\Gamma \|q\|^2 + \|F_k + F_k' q\|)$$

for any vector $x_k + q \in B_{\rho_3}(x^*)$. Then (4.40), (4.46), and (4.45) give

$$\|x_k + p(\bar{\Delta}_k) - x^*\|_* \le \epsilon_k\, \|x_k - x^*\|_*. \tag{4.47}$$

Note that $\omega_1 \to 1$, $\omega_2 \to 1$ as $\Gamma \to 0$. Since $\alpha_k \to 1$ as $k \to \infty$, there exist Γ and ζ sufficiently small such that

$$\omega_1 \omega_2 < \frac{\bar{\eta}}{\eta_{max}}, \qquad \bar{\epsilon}_k < \frac{\bar{\eta}}{\omega_1} < \frac{1}{\omega_1}, \qquad \text{and} \qquad \epsilon_k < \bar{\eta} \tag{4.48}$$

whenever $x_k \in B_\zeta(x^*) \cap \mathrm{int}(\Omega)$ for k sufficiently large. As a consequence, eventually $p(\bar{\Delta}_k)$ satisfies (2.9). In fact, following along the lines of Theorem 4.3 and using (4.46), (4.40), and (4.36), we obtain

$$\begin{aligned} \rho_f(p(\bar{\Delta}_k)) &= 1 - \frac{\|F(x_k + p(\bar{\Delta}_k))\|^2 - \|F_k + F_k' p(\bar{\Delta}_k)\|^2}{\|F_k\|^2 - \|F_k + F_k' p(\bar{\Delta}_k)\|^2} \\ &\ge 1 - \frac{\Gamma^2 \|p(\bar{\Delta}_k)\|^4 + 2\Gamma \bar{\epsilon}_k \|p(\bar{\Delta}_k)\|^2 \|x_k - x^*\|_*}{(\frac{1}{\omega_1^2} - \bar{\epsilon}_k^2)\|x_k - x^*\|_*^2} \\ &\ge 1 - \frac{\Gamma \nu_*^2 (\Gamma \nu_*^2 \|x_k - x^*\|_* + 2\bar{\epsilon}_k)\|x_k - x^*\|_*}{\frac{1}{\omega_1^2} - \bar{\epsilon}_k^2}. \end{aligned}$$

Hence for k sufficiently large, $\rho_f(p(\bar{\Delta}_k)) \geq \beta_2$, $x_{k+1} = x_k + p(\bar{\Delta}_k)$, and from (4.47) and (4.48) the proof is completed. □

We conclude this section by noting that the assumption (4.43) is guaranteed whenever x^* belongs to the interior of Ω, as $\|p_{tr}(\bar{\Delta}_k)\|$ tends to zero, while it is an additional condition when x^* lies on the boundary of Ω. This assumption allows us to obtain (4.47) where $\epsilon_k < \bar{\eta}$ as $k \to \infty$. This yields linear convergence in the weighted norm for any $\bar{\eta} < 1$. On the contrary, if $P(x_k + p_{tr}(\bar{\Delta}_k)) - x_k \neq p_{tr}(\bar{\Delta}_k)$, proceeding as in Lemma 4.4, we cannot derive a bound on $\|x_k + p(\bar{\Delta}_k) - x^*\|_*$ with analogous properties. In fact, we only manage to get a bound of the form (4.47) where $\epsilon_k < \bar{\eta}\,\chi_J K_2$ for k sufficiently large.

5. Applications. To develop viable approaches for large-scale problems, this section discusses the two issues left unspecified in the description of the SIATR method: the choice of the subspace S_k and the way of solving the subspace trust-region problem (3.1).

Since there is no finite method of determining the exact solution of (3.1), an approximation to it is used. Remarkably, it is easy to see that the convergence properties of the SIATR method take place using an approximate solution $p_{tr}(\Delta_k)$ to (3.1) which satisfies the following two mild conditions:

(a) $m_k(p_{tr}(\Delta_k)) \leq m_k(0)$;

(b) $p_{tr}(\Delta_k) = p_k$ when p_k is feasible for (3.1).

The key is that global convergence is provided by $p_c(\Delta_k)$ and rapid local convergence is ensured if eventually p_k is the solution of (3.1).

We outline a subspace dogleg strategy for solving (3.1) approximately. Let $S_k = span\{s_1, s_2, \ldots, s_r\}$, $S_k^G = span\{G_k s_1, G_k s_2, \ldots, G_k s_r\}$. Once an orthonormal basis $W \in \mathbb{R}^{n\times r}$ for S_k^G has been constructed, a vector $p \in S_k$ is such that $G_k p = Wq$ for some $q \in \mathbb{R}^r$, and instead of (3.1), one can consider the spherical trust-region problem

$$\min_{q\in\mathbb{R}^r} \{\phi_k(q) : \quad \|q\| \leq \Delta_k\}, \tag{5.1}$$

where ϕ_k is the quadratic model on $\mathbb{R}^r$

$$\phi_k(q) = \frac{1}{2}\|F_k + F_k' G_k^{-1} W q\|^2. \tag{5.2}$$

Let $q_{tr}(\Delta_k)$ be the dogleg solution to (5.1); see [10]. Its evaluation calls for the Cauchy point $q_c(\Delta_k)$ for (5.1) and for the vector

$$q_k = \operatorname*{argmin}_{q\in\mathbb{R}^r} \phi_k(q). \tag{5.3}$$

We remark that $q_k = W^T G_k p_k$ with p_k given in (3.2). Therefore, if one of the two vectors p_k and q_k is known, the other one can be trivially evaluated.

Once $q_{tr}(\Delta_k)$ has been computed, coming back into the original space, the vector

$$p_{tr}(\Delta_k) = G_k^{-1} W q_{tr}(\Delta_k) \tag{5.4}$$

is built. It approximately solves (3.1) and satisfies both (a) and (b). In fact, (a) is straightforward and (b) is due to the fact that $G_k p_k \in S_k^G$, i.e., $WW^T G_k p_k = G_k p_k$. This yields $\|q_k\| = \|W^T G_k p_k\| = \|G_k p_k\|$. Then q_k is feasible for (5.1) if p_k is feasible for (3.1). Consequently, $q_{tr}(\Delta_k) = q_k$ whenever $\|q_k\| \leq \Delta_k$, and this implies $p_{tr}(\Delta_k) = p_k$ whenever $\|G_k p_k\| \leq \Delta_k$.

This discussion leads to the subspace dogleg strategy sketched below.

ALGORITHM II. A subspace dogleg strategy for (3.1).
Input parameters $x_k \in \mathrm{int}(\Omega)$, $\Delta_k > 0$, $\eta_k \in [0,1)$, G_k, ∇f_k.

1. Choose a subspace $S_k = span\{s_1, s_2, \ldots, s_r\}$ such that (3.2) holds.
2. Find an orthonormal basis $W \in \mathbb{R}^{n\times r}$ for $S_k^G = span\{G_k s_1, G_k s_2, \ldots, G_k s_r\}$.
3. Compute the vector $q_k \in \mathbb{R}^r$ satisfying (5.3).
4. Compute the Cauchy step $q_c(\Delta_k) = -\hat{\mu}_k W^T G_k^{-1} \nabla f_k$ with

$$\begin{aligned} \hat{\mu}_k &= \operatorname*{argmin}_{\|\mu W^T G_k^{-1}\nabla f_k\| \le \Delta_k} \phi_k(-\mu W^T G_k^{-1}\nabla f_k) \\ &= \min\left\{ \frac{\|W^T G_k^{-1}\nabla f_k\|^2}{\|F_k' G_k^{-1} W W^T G_k^{-1}\nabla f_k\|^2}, \frac{\Delta_k}{\|W^T G_k^{-1}\nabla f_k\|} \right\}. \end{aligned} \tag{5.5}$$

5. Find the dogleg solution $q_{tr}(\Delta_k)$ to (5.1):

$$q_{tr}(\Delta_k) = \begin{cases} q_k & \text{if } \|q_k\| \le \Delta_k, \\ q_c(\Delta_k) & \text{if } \|q_c(\Delta_k)\| = \Delta_k, \\ sq_k + (1-s)q_c(\Delta_k), s \in (0,1), \text{ s.t. } \|q_{tr}(\Delta_k)\| = \Delta_k & \text{otherwise.} \end{cases}$$

6. Compute $p_{tr}(\Delta_k)$ by (5.4).

We recall that (3.2) holds whenever the subspace S_k contains an inexact Newton step p_k^I for the problem $F(x) = 0$ such that

$$F_k' p_k^I = -F_k + r_k, \qquad \|r_k\| \le \eta_k \|F_k\|. \tag{5.6}$$

Our main purpose now is to show that Krylov subspace methods for solving (5.6) provide the way to perform steps 1–3 of the above algorithm effectively. The resulting methods belong to the class of trust-region Newton–Krylov methods [6, 7, 15]. Moreover, thanks to our convergence results, the linear system (5.6) can be solved with an accuracy that increases as the solution is approached, and oversolving can be avoided by choosing suitable sequences $\{\eta_k\}$; see [27].

Two-dimensional subspace minimization. A possible approach consists in determining p_k^I by a Krylov method and fixing

$$S_k = span\{p_k^I, G_k^{-2}\nabla f_k\}.$$

This way, Algorithm II sketches a two-dimensional subspace trust-region strategy. Step 2 requires one step of the Gram–Schmidt procedure to compute W. The least-squares problem $\min_{q\in\mathbb{R}^2} \|F_k + F_k' G_k^{-1} W q\|^2$ in step 3 can be solved without much effort either by the QR factorization of the $n \times 2$ matrix $F_k' G_k^{-1} W$ or by solving the normal equations.

In the case $G_k = D_k$, $k \ge 0$, the vector $p_{tr}(\Delta_k)$ produces as much decrease in the quadratic model m_k as the generalized Cauchy point $p_c(\Delta_k)$. To show this fact, note that by $d_k \in S_k$ and $G_k d_k = -D_k^{-1}\nabla f_k \in S_k^G$, it trivially follows that $D_k^{-1}\nabla f_k = W W^T D_k^{-1}\nabla f_k$. Consequently, it is easy to see that $G_k^{-1} W q_c(\Delta_k) = \hat{\tau}_k d_k$, where $\hat{\tau}_k$ is given in (2.6) and

$$m_k(p_{tr}(\Delta_k)) = \phi_k(q_{tr}(\Delta_k)) \le \phi_k(q_c(\Delta_k)) = m_k(p_c(\Delta_k)),$$

i.e., $p_{tr}(\Delta_k)$ satisfies (2.8).

GMRES subspace strategy. Another implementation of the subspace dogleg strategy can be proposed in connection with the GMRES method [31]. GMRES shows a certain optimality among all Krylov methods, BICGSTAB, TFQMR etc., commonly used in the solution of general linear systems. In practice, it minimizes the residual norm $(2\,m_k(p))^{\frac{1}{2}} = \|F_k + F_k'\,p\|$ over all corrections in the current Krylov subspace. Due to this property it is possible to define subspace dogleg strategies using information provided by GMRES; see, e.g., [6, 7, 20].

For the sake of clarity, we sketch the application of GMRES to $F_k'p = -F_k$ in order to solve (5.6). For details we refer to [31]. Given $p_k^0 \in \mathbb{R}^n$, GMRES generates a sequence of iterates $\{p_k^m\}$, $p_k^m \in \mathbb{R}^n$, $m \geq 0$, until $\|F_k'\,p_k^m + F_k\| \leq \eta_k \|F_k\|$. Then $p_k^I = p_k^m$ is set. Each vector p_k^m solves the least-squares problem

$$\min_{p \in p_k^0 + K_m} \|F_k + F_k'\,p\|, \tag{5.7}$$

where K_m is the Krylov subspace $K_m = span\{r_k^0, F_k' r_k^0, (F_k')^2 r_k^0, \ldots, (F_k')^{m-1} r_k^0\}$ and $r_k^0 = -F_k'\,p_k^0 - F_k$. To accomplish the solution of (5.7), GMRES computes an orthonormal basis $V_m = [v_1, \ldots, v_m] \in \mathbb{R}^{n \times m}$ of K_m by the Arnoldi process. It is known that $F_k' V_m = V_{m+1} H_m$, where $V_{m+1} = [v_1, \ldots, v_{m+1}] \in \mathbb{R}^{n \times (m+1)}$ is the orthonormal basis for K_{m+1} and $H_m \in \mathbb{R}^{(m+1) \times m}$ is an Hessenberg matrix. The QR factorization of H_m is required and cheaply performed for all $m \geq 0$.

Typically the restarted procedure GMRES(m_M) is applied. Suppose the vector p_k^I is generated using the GMRES(m_M) method. By construction, p_k^I minimizes $m_k(p)$ within the affine subspace $p_k^0 + K_m$. Then if $p_k^0 = 0$, it is convenient to perform steps 1–3 of the subspace dogleg strategy as follows. Setting

$$S_k = K_m,$$

trivially, we get $p_k = p_k^I$. Moreover, $W = V_m$ if $G_k = I$, $k \geq 0$, while for the other choice of G the matrix W has to be computed. Finally, step 3 is completed by setting $q_k = W^T G_k p_k$.

If $p_k^0 \neq 0$, let

$$S_k = span\{v_1, \ldots, v_m, p_k^0\}.$$

If $G_k = I$, $k \geq 0$, an orthonormal basis W for S_k is easily obtained adding one column to the matrix V_m. Such a column can be computed with one step of the Gram–Schmidt procedure. Otherwise, the whole matrix W must be computed. To evaluate q_k we compute the solution p_k to problem (3.2) and let $q_k = W^T G_k p_k$. In particular, letting $T_k = [V_m, p_k^0] \in \mathbb{R}^{n \times r}$, $r = m + 1$, p_k has the form

$$p_k = T_k y_k, \qquad y_k = \operatorname*{argmin}_{y \in \mathbb{R}^r} \|F_k + F_k'\,T_k\,y\|.$$

If the columns of $F_k'\,T_k$ are linearly independent, i.e., the columns of T_k are linearly independent, the solution y_k is unique. Moreover, the Cholesky factorization $R_k^T R_k$ of the matrix $(F_k'\,T_k)^T F_k'\,T_k \in \mathbb{R}^{r \times r}$ can be computed, exploiting the QR factorization of the Hessenberg matrix H_m provided by GMRES and solving one upper triangular system of dimension m; see [6]. A complication to this global strategy arises when $F_k' T_k$ is ill-conditioned. We can monitor this occurrence estimating the condition number of the small matrix R_k. If such a number is greater than a fixed threshold, we may perturb the quadratic model $\|F_k + F_k'\,T_k\,y\|^2$ following the strategy given in [13, p. 151].

Finally, we make some comments on the use of preconditioning techniques. Let P^{-1} be the preconditioner employed. Since our stopping criterion for computing the inexact Newton step p_k^I is based on the unpreconditioned residuals, we focus on the linear system $F_k' P^{-1} s = -F_k$ with $p = P^{-1} s$. Without loss of generality, we concentrate on the case where a null initial guess for GMRES, $s_k^0 = 0$, is chosen, restart is not used, and the choice $G_k = I$, $k \geq 0$, is adopted. Then the Krylov space generated by GMRES has the form

$$K_m^p = \{r_k^0, (F_k' P^{-1}) r_k^0, (F_k' P^{-1})^2 r_k^0, \ldots, (F_k' P^{-1})^{m-1} r_k^0\},$$

where $r_k^0 = -F_k$. Clearly, the vector $s_k^I \in K_m^p$ such that $\|F_k' P^{-1} s_k^I + F_k\| \leq \eta_k \|F_k\|$ gives rise to the vector $p_k^I = P^{-1} s_k^I$ satisfying (5.6).

To design the subspace dogleg strategy, it is convenient to set $S_k = span\{P^{-1} v_1, P^{-1} v_2, \ldots, P^{-1} v_m\}$, where $v_1, \ldots, v_m$ is the orthonormal basis of K_m^p computed by GMRES. This way, $p_k^I \in S_k$ and $p_k = p_k^I$ holds. Moreover, computing an orthonormal basis for S_k requires the application of an orthonormalization procedure.

6. Numerical experiments. In this section, some computational results are discussed to illustrate the viability of our proposals. They have been selected to show the behavior of both of the subspace dogleg strategies on problems with different features and are not meant to be exhaustive.

We implemented the SIATR method with spherical trust regions, $G_k = I$, $k \geq 0$, in a Fortran code. We refer to the SIATR method with the two-dimensional subspace strategy as `SIATR-2D` and to the SIATR method with the GMRES subspace strategy as `SIATR-G`.

The inexact Newton step p_k^I was computed using the iterative linear solver GMRES with null initial guess [31]. Restart was not employed and a maximum of 50 GMRES iterations was allowed. If after 50 GMRES iterations condition (5.6) had not been met, our algorithm continued with p_k^I given by the last computed GMRES iterate.

In all runs, we set $\bar{\Delta}_0 = 1$, $\Delta_{min} = \sqrt{\epsilon_m}$, where ϵ_m denotes the machine precision, $\beta_1 = 0.1$, $\beta_2 = 0.25$, $\theta = 0.99995$. For the computation of the projected step (3.4) we used $\alpha_k = \max\{0.95, 1 - \|F_k\|\}$, $k > 0$. The strategy for updating Δ_k was the same as in [3, p. 17]

Here we report numerical experiments performed with machine precision $\epsilon_m = 2 \times 10^{-16}$ on a 1.7 GHz Intel Pentium M with 512K cache. Convergence is declared when $\|F_k\| \leq 10^{-8}$, while failure is declared when a maximum number of 200 iterations are performed. We conducted experiments with two choices of the forcing terms η_k proposed in [27]: *Choice* 1 and *Choice* 2. In our implementation we use $\eta_0 = 0.9$ and the safeguards suggested in [27, p. 305]. In Choice 2 we set the parameters as $\gamma = 0.9$ and $\alpha = 2$.

For the purpose of this presentation we consider two representative test problems: the seven diagonal problem [23] and the Bratu complementarity problem [14]. The first problem was solved with $n = 5000$. It has more than one solution, and in order to approximate the positive solution, we formulated it as a bound-constrained nonlinear system with $l_i = 0$, $u_i = \infty$, $i = 1, \ldots, n$. The Bratu complementarity problem was reformulated as a system of $n = 5 \times 10^4$ smooth bound-constrained nonlinear equations with $l_i = 0$ and $u_i = \infty$ for $i = 1, \ldots, n$. It depends on a parameter λ and we considered the value $\lambda = 6$.

The two problems display different features: the seven diagonal problem has the solution in the interior of Ω and preconditioning is not required. On the other hand

TABLE 6.1
Performance with Choice 1 *and Choice* 2 *of* η_k*'s.*

		Seven diagonal						Bratu					
		SIATR-2D			SIATR-G			SIATR-2D			SIATR-G		
η_k	l	NIT	NF	NLIT	NIT	NF	NLIT	NIT	NF	NLIT	NIT	NF	NLIT
Choice 1	0	34	35	377	37	38	514	30	31	197	28	29	257
	1	*	*	*	40	41	632	30	31	373	29	30	297
	2	20	21	65	20	21	65	30	31	432	30	31	443
	3	38	41	104	35	37	89	*	*	*	*	*	*
Choice 2	0	22	23	157	24	25	243	18	19	355	15	16	195
	1	62	66	1284	27	28	251	19	20	406	16	17	304
	2	7	8	65	7	8	67	19	20	358	20	21	408
	3	33	38	112	29	35	106	*	*	*	*	*	*

the solution of the Bratu problem lies on the boundary of the feasible set and a preconditioner is needed to speed up the convergence of GMRES. In our runs we employed the ILU(0) preconditioner. For both problems we used four initial guesses: $x_0^{(l)} = 10^{l-2}$ with $l = 0, 1, 2, 3$.

In reporting the numerical results we provide the following: the parameter l used to form the initial guess; the number NIT of nonlinear iterations; the number NF of function evaluations; and the number NLIT of linear iterations performed by GMRES. An asterisk indicates a failure.

Table 6.1 shows the performance of the subspace dogleg strategies achieved with both choices of η_k's. It gives an indication of the computational cost in terms of nonlinear iterations and GMRES iterations. The SIATR-G method seems to perform slightly better in terms of nonlinear iterations and GMRES iterations. This behavior could be predictable since in the SIATR-G method we search for an approximate trust-region solution within a subspace of dimension larger than two. On the other hand, as we pointed out in the previous section, performing the GMRES Subspace strategy may be more expensive than performing the two-dimensional subspace strategy. Hence, for the runs where the two approaches are comparable in terms of linear and nonlinear iterations, the overall cost of the SIATR-2D method may be expected to be lower. However, comparing the two strategies in terms of execution time, it comes out that the extra work for the GMRES subspace strategy has a minor impact on the overall performance of the SIATR-G method.

7. Conclusions. We have introduced a prototype trust-region method for large bound-constrained nonlinear systems. The proposed method involves the solution of the subspace trust-region problems (3.1). The crucial issue in such an approach is the choice of the subspaces S_k. We have investigated this point, showing how to choose such subspaces in order to ensure both strong convergence properties and practical viability in the large scale setting. The convergence results provided are in accordance with those of inexact Newton methods for unconstrained nonlinear systems. To our knowledge, this is the first contribution in interior methods for general large-scale nonlinear systems with simple bounds.

We have implemented algorithms that are based on the proposed paradigm. They are consistent with the convergence theory and involve Krylov methods to construct the subspaces S_k. The numerical results provided indicate that our subspace trust-region approaches are a promising tool for large-scale computation.

Acknowledgment. We would like to thank one of the referees for carefully reading the manuscript and for several suggestions that improved the presentation of this work.

REFERENCES

[1] S. Bellavia, M. Macconi, and B. Morini, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Appl. Numer. Math., 44 (2003), pp. 257–280.
[2] S. Bellavia and B. Morini, *A globally convergent Newton-GMRES subspace method for systems of nonlinear equations*, SIAM J. Sci. Comput., 23 (2001), pp. 940–960.
[3] S. Bellavia and B. Morini, *An interior global method for nonlinear systems with simple bounds*, Optim. Methods Software, 20 (2005), pp. 1–22.
[4] L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, *Large-Scale PDE-Constrained Optimization*, Lect. Notes Comput. Sci. Eng. 30, Springer, Berlin, 2003.
[5] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1–23.
[6] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.
[7] P. N. Brown and Y. Saad, *Convergence theory of nonlinear Newton–Krylov algorithms*, SIAM J. Optim., 4 (1994), pp. 297–330.
[8] R. H. Byrd, R. B. Schnabel, and M. H. Schultz, *Approximate solution of the trust-region problem by minimization over two-dimensional subspaces*, Math. Program., 40 (1988), pp. 247–263.
[9] D. Calvetti and L. Reichel, *Gauss quadrature applied to trust region computations*, Numer. Algorithms, 34 (2003), pp. 85–102.
[10] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[11] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.
[12] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
[13] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[14] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.
[15] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.
[16] R. Fletcher and S. Leyffer, *Filter-type algorithms for solving systems of algebraic equations and inequalities*, in High Performance Algorithms and Software for Nonlinear Optimization, G. Di Pillo and A. Murli, eds., Kluwer Academic, Norwell, MA, 2003, pp. 259–278.
[17] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[18] A. Hasnat and S. Roy, *Microphase-enhanced reactions: Simultaneous effects on ion coupling and counterion binding*, Ind. Eng. Chem. Res., 38 (1999), pp. 4571–4578.
[19] M. Heinkenschloss, M. Ulbrich, and S. Ulbrich, *Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumptions*, Math. Program., 86 (1999), pp. 615–635.
[20] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust-region SQP algorithms*, SIAM J. Optim., 12 (2001), pp. 283–302.
[21] V. A. Juvekar, C. V. Anoop, S. K. Pattanayek, and V. M. Naik, *A continuum model for polymer adsorption at the solid-liquid interface*, Macromolecules, 32 (1999), pp. 863–873.
[22] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.
[23] L. Luksan, *Inexact trust-region method for large sparse systems of nonlinear equations*, J. Optim. Theory Appl., 81 (1994), pp. 569–591.
[24] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[25] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[26] R. P. Pawlowski, J. N. Shadid, J. P. Simonis, and H. F. Walker, *Globalization techniques for Newton–Krylov methods and applications to the fully coupled solution of the Navier–Stokes equations*, SIAM Rev., to appear.

[27] M. Pernice and H. F. Walker, *NITSOL: A Newton iterative solver for nonlinear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 302–318.

[28] L. Qi, X. J. Tong, and D. H. Li, *An active-set projected trust region algorithm for box-constrained nonsmooth equations*, J. Optim. Theory Appl., 120 (2004), pp. 627–649.

[29] H. Qi, L. Qi, and D. Sun, *Solving Karush–Kuhn–Tucker systems via the trust region and the conjugate gradient methods*, SIAM J. Optim., 14 (2003), pp. 439–463.

[30] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[31] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[32] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[33] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, New York, 1981, pp. 57–88.

[34] X. J. Tong and L. Qi, *On the convergence of a trust-region method for solving constrained nonlinear equations with degenerate solutions*, J. Optim. Theory Appl., 123 (2004), pp. 187–211.

[35] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2001), pp. 889–917.

[36] A. Walter, A. Griewank, and A. Best, *Multiple vector-Jacobian products are cheap*, Appl. Numer. Math., 30 (1999), pp. 367–377.

[37] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*, John Wiley, New York, 1996.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1556–1583

ON THE STABILITY OF EVOLUTION GALERKIN SCHEMES APPLIED TO A TWO-DIMENSIONAL WAVE EQUATION SYSTEM*

M. LUKÁČOVÁ-MEDVIĎOVÁ†, G. WARNECKE‡, AND Y. ZAHAYKAH‡

Abstract. The subject of the paper is the analysis of stability of the evolution Galerkin (EG) methods for the two-dimensional wave equation system. We apply von Neumann analysis and use the Fourier transformation to estimate the stability limits of both the first and the second order EG methods.

Key words. hyperbolic systems, wave equation, evolution Galerkin schemes, discrete Fourier transformation, amplification matrix, CFL condition

AMS subject classifications. 65L05, 65M06, 65M12, 65M25, 35L45, 35L65

DOI. 10.1137/040615882

1. Introduction.

Evolution Galerkin (EG) methods were proposed to approximate first order hyperbolic problems. These schemes were introduced by Lin, Morton, and Süli; see, e.g., [8] for scalar problems and [9] for one-dimensional systems. The first generalization to two-dimensional systems was made in [23] by Ostkamp for the wave equation system as well as for the Euler equations of gas dynamics. In [13] Lukáčová-Medvid'ová, Morton, and Warnecke studied systematically approximate evolution operators and constructed new EG schemes with better accuracy and stability properties. Further EG schemes as well as the approximate evolution operator of the solution for the wave equation system in three space dimensions were derived in [28]. These methods and their finite volume versions were applied to the linearized Euler equations and Maxwell equations [16]. Higher order finite volume EG (FVEG) methods have been introduced and studied in [12], [14], [15], and [17]. In [11], [15], [6] the FVEG schemes have been generalized to fully nonlinear systems of hyperbolic conservation laws, such as the Euler equations of gas dynamics, shallow water equations, and the shallow water magnetohydrodynamic equations. For hyperbolic conservation laws with source terms, the so-called well-balanced FVEG schemes are proposed in [20]. In general, the FVEG schemes produce very accurate numerical solutions within CPU time comparable to some other well-known finite volume methods. In particular, genuinely multidimensional features, such as oblique shocks, are resolved very well; cf., e.g., [11], [15]. For example, it has been shown in [15] that the global error of the second order FVEG scheme using (7.1)–(7.3) and (8.1)–(8.3) is approximately six times smaller than the error of the Lax–Wendroff (rotated Richtmyer) scheme as well as of the second order wave propagation algorithm of LeVeque [7].

The FVEG methods belong to the class of the so-called genuinely multidimensional schemes. The goal is to have a method which approximates possibly all of the

*Received by the editors September 28, 2004; accepted for publication (in revised form) February 27, 2006; published electronically August 7, 2006. This research was supported by the VolkswagenStiftung Agency, by Deutsche Forschungsgemeinschaft grant Wa 633/6-2, and partially by the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sinum/44-4/61588.html

†TU Hamburg-Harburg, Arbeitsbereich Mathematik, Schwarzenbergstrasse 95, 21 073 Hamburg, Germany (lukacova@tu-harburg.de).

‡Institut für Analysis und Numerik, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39 106 Magdeburg, Germany (Gerald.Warnecke@mathematik.uni-magdeburg.de, Yousef.Zahaykah @mathematik.uni-magdeburg.de).

infinitely many directions of wave propagation. The reader is referred, for example, to [3], [4], [7], [22], [25] for other genuinely multidimensional schemes.

The main objective of this paper is the analysis of the stability of the evolution Galerkin schemes. In [13] we have proven that the EG schemes are conditionally stable. However, the precise stability limits were not computed there. The goal of this paper is to find stability limits by analysis of the spectrum of the corresponding discrete operators for the EG as well as FVEG schemes; cf. [11], [13]. First, we consider the so-called EG3 scheme for the wave equation system in two space dimensions. We apply the discrete Fourier transform to obtain the amplification matrix. It turns out that its structure is too complex in order to derive precise stability limits theoretically. Anyway, we find theoretical stability estimates for a simplified problem. This is then compared with the experimental estimate of the spectrum of the amplification matrix of the EG3 scheme.

Further, we derive amplification matrices for the first- and the second order FVEG schemes based on the approximate evolution operators. The spectral radius of the amplification matrices is estimated experimentally by a built-in MATLAB procedure. Hence the stability limit of the schemes is estimated numerically.

The outline of this paper is as follows: in the next section we survey the general theory that we used to derive the exact integral equations. The exact integral equations as well as the approximate evolution operators for the two-dimensional wave equation system are given in section 3. In section 4 we recall the evolution Galerkin schemes. In section 5 we introduce the discrete Fourier transform as well as the spectral norm that serve as tools in our analysis. In section 6 we present the derivation of a stability condition for a simplified problem and compare the theoretical estimate, which we obtained by means of the Fourier analysis, with the experimental limit of the original problem. In section 7 we consider the first order finite volume schemes based on the approximate evolution operator E_Δ^{const}. We determine the amplification matrices and estimate their stability limits. Finally in section 8 we determine the amplification matrices of the second order finite volume schemes based on the approximate evolution operator E_Δ^{bilin} and estimate the stability limits.

2. General theory. In this section we recall the exact integral equations for a general linear hyperbolic system using the concept of bicharacteristics. Consider a general form of linear hyperbolic system

$$\mathbf{U}_t + \sum_{k=1}^{d} \mathcal{A}_k \mathbf{U}_{x_k} = 0, \;\; \mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d, \tag{2.1}$$

where the coefficient matrices $\mathcal{A}_k$, $k = 1, \ldots, d$, are elements of $\mathbb{R}^{p\times p}$ and the dependent variables are $\mathbf{U} = (u_1, \ldots, u_p)^T = \mathbf{U}(\mathbf{x}, t) \in \mathbb{R}^p$. Let $\mathcal{A}(\mathbf{n}) = \sum_{k=1}^{d} n_k \mathcal{A}_k$ be the *pencil matrix*, where $\mathbf{n} = (n_1, \ldots, n_d)^T$ is a unit vector in $\mathbb{R}^d$. Since the system (2.1) is hyperbolic the matrix $\mathcal{A}(\mathbf{n})$ has p real eigenvalues λ_k, $k = 1, \ldots, p$, and p corresponding linearly independent right eigenvectors $\mathbf{r}_k = \mathbf{r}_k(\mathbf{n})$, $k = 1, \ldots, p$. Let $\mathcal{R} = [\mathbf{r}_1 | \mathbf{r}_2 | \ldots | \mathbf{r}_p]$ be the matrix of right eigenvectors. We define the characteristic variable $\mathbf{W} = \mathbf{W}(\mathbf{n})$ as $\partial \mathbf{W}(\mathbf{n}) = \mathcal{R}^{-1} \partial \mathbf{U}$. Since the system (2.1) has constant coefficient matrices $\mathcal{A}_k$ we have $\mathbf{W} = \mathcal{R}^{-1}\mathbf{U}$ or $\mathbf{U} = \mathcal{R}\,\mathbf{W}$.

Transforming system (2.1) by multiplying it with $\mathcal{R}^{-1}$ from the left we get

$$\mathcal{R}^{-1}\mathbf{U}_t + \sum_{k=1}^{d} \mathcal{R}^{-1}\mathcal{A}_k \mathcal{R}\, \mathcal{R}^{-1}\mathbf{U}_{x_k} = 0. \tag{2.2}$$

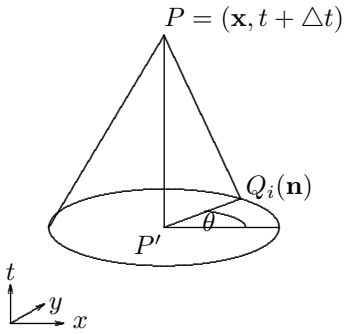


FIG. 1. *Bicharacteristics along the Mach cone through P and $Q_i(\mathbf{n})$, $d = 2$.*

Let $\mathcal{B}_k = \mathcal{R}^{-1}\mathcal{A}_k\mathcal{R} = (b_{ij}^k)_{i,j=1}^p$, where $k = 1, 2, \ldots, d$; then the system (2.2) can be rewritten in the following form using the characteristic variables:

$$\mathbf{W}_t + \sum_{k=1}^{d} \mathcal{B}_k \mathbf{W}_{x_k} = 0.$$

Now we decompose $\mathcal{B}_k$ into the diagonal part $\mathcal{D}_k$ and the remaining part $\mathcal{B}'_k$, i.e., $\mathcal{B}_k = \mathcal{D}_k + \mathcal{B}'_k$. We obtain

$$\mathbf{W}_t + \sum_{k=1}^{d} \mathcal{D}_k \mathbf{W}_{x_k} = -\sum_{k=1}^{d} \mathcal{B}'_k \mathbf{W}_{x_k} =: \mathbf{S}. \tag{2.3}$$

The ith bicharacteristic corresponding to the ith equation of (2.3) is defined by

$$\frac{d\mathbf{x}_i}{d\tilde{t}} = \mathbf{b}_{ii}(\mathbf{n}) = (b_{ii}^1, b_{ii}^2, \ldots, b_{ii}^d)^T,$$

where $i = 1, \ldots, p$. The diagonal entries b_{ii}^k of the matrices $\mathcal{B}_k$, $k = 1, \ldots, d$, $i = 1, \ldots, p$, create the ray velocity vector $\mathbf{b}_{ii}$; cf. [1]. We consider the bicharacteristics backwards in time and set the initial conditions $\mathbf{x}_i(t + \Delta t, \mathbf{n}) = \mathbf{x}$ for all $\mathbf{n} \in \mathbb{R}^d$ and $i = 1, \ldots, p$, i.e., $\mathbf{x}_i(\tilde{t}, \mathbf{n}) = \mathbf{x} - \mathbf{b}_{ii}(\mathbf{n})(t + \Delta t - \tilde{t})$.

We will integrate the ith equation of the system (2.3) from the point $P \equiv (\mathbf{x}, t + \Delta t) \in \mathbb{R}^p \times \mathbb{R}_+$ down to the point $Q_i(\mathbf{n}) = (\mathbf{x}_i(t, \mathbf{n}), t) = (\mathbf{x} - \Delta t \mathbf{b}_{ii}, t)$, where the bicharacteristic hits the plane at time t; see Figure 1. Note that bicharacteristics are straight lines because the system is linear and has constant coefficients. Now the ith equation reads

$$\frac{\partial w_i}{\partial t} + \sum_{k=1}^{d} b_{ii}^k \frac{\partial w_i}{\partial x_k} = -\left(\sum_{j=1, i\neq j}^{d} \left(b_{ij}^1 \frac{\partial w_j}{\partial x_1} + b_{ij}^2 \frac{\partial w_j}{\partial x_2} + \cdots + b_{ij}^d \frac{\partial w_j}{\partial x_d} \right) \right) = S_i. \tag{2.4}$$

Taking a vector $\sigma_i = (b_{ii}^1, b_{ii}^2, \ldots, b_{ii}^d, 1)$, we can define the directional derivative

$$\frac{dw_i}{d\sigma_i} = \left(\frac{\partial w_i}{\partial x_1}, \frac{\partial w_i}{\partial x_2}, \ldots, \frac{\partial w_i}{\partial x_d}, \frac{\partial w_i}{\partial t} \right) \cdot \sigma_i = \frac{\partial w_i}{\partial t} + b_{ii}^1 \frac{\partial w_i}{\partial x_1} + b_{ii}^2 \frac{\partial w_i}{\partial x_2} + \cdots + b_{ii}^d \frac{\partial w_i}{\partial x_d}.$$

Hence the ith equation (2.4) can be rewritten as follows:

$$\frac{dw_i}{d\sigma_i} = S_i = -\sum_{j=1, i\neq j}^{d} \left(b_{ij}^1 \frac{\partial w_j}{\partial x_1} + b_{ij}^2 \frac{\partial w_j}{\partial x_2} + \cdots + b_{ij}^d \frac{\partial w_j}{\partial x_d} \right).$$

Integration from P to $Q_i(\mathbf{n})$ gives

$$w_i(P) - w_i(Q_i(\mathbf{n})) = S_i', \tag{2.5}$$

where

$$S_i' = \int_t^{t+\Delta t} S_i(\mathbf{x}_i(\tilde{t},\mathbf{n}),\tilde{t},\mathbf{n})d\tilde{t} = \int_0^{\Delta t} S_i(\mathbf{x}_i(t+\Delta t-\tau,\mathbf{n}),t+\Delta t-\tau,\mathbf{n})\mathrm{d}\tau.$$

The reverse transformation of (2.5) into a system written in the original physical variables is done by multiplication with $\mathcal{R}$ from the left and $(d-1)$-dimensional integration of the variable $\mathbf{n}$ over the unit sphere O in $\mathbb{R}^d$. This leads to the integral representation of the solution in the point $\mathbf{x}$ at time $t+\Delta t$:

$$\mathbf{U}(P) = \mathbf{U}(\mathbf{x},t+\Delta t) = \frac{1}{|O|}\int_O \mathcal{R}(\mathbf{n}) \begin{pmatrix} w_1(Q_1(\mathbf{n}),\mathbf{n}) \\ w_2(Q_2(\mathbf{n}),\mathbf{n}) \\ w_3(Q_3(\mathbf{n}),\mathbf{n}) \\ \vdots \\ w_p(Q_p(\mathbf{n}),\mathbf{n}) \end{pmatrix} \mathrm{d}O + \tilde{\mathbf{S}}, \tag{2.6}$$

where

$$\tilde{\mathbf{S}} = (\tilde{S}_1,\tilde{S}_2,\ldots,\tilde{S}_p)^T = \frac{1}{|O|}\int_O \mathcal{R}(\mathbf{n})\mathbf{S}'\mathrm{d}O = \frac{1}{|O|}\int_O\int_0^{\Delta t} \mathcal{R}(\mathbf{n})\mathbf{S}(t+\Delta t-\tau,\mathbf{n})\mathrm{d}\tau\mathrm{d}O$$

and $|O|$ corresponds to the measure of the domain of integration.

3. Exact integral equations and approximate evolution operators for the wave equation system. In this section we illustrate the application of the general theory of bicharacteristics for the two-dimensional system of wave equations. We recall the exact integral equations and present their possible approximation, the so-called EG3 approximate evolution operator. Consider the two-dimensional wave equation system

$$\begin{aligned} \phi_t + c(u_x + v_y) &= 0, \\ u_t + c\phi_x &= 0, \\ v_t + c\phi_y &= 0, \end{aligned} \tag{3.1}$$

where c is a given positive constant representing the speed of sound. We will recall here the exact integral equations derived in [13]. Let $P = (x,y,t+\Delta t)$, $P' = (x,y,t)$, $Q = (x + c\Delta t\cos\theta, y + c\Delta t\sin\theta, t) = (\mathbf{x} + c\Delta t\mathbf{n}(\theta), t)$ and let the so-called source term be given as

$$S = c\left[u_x\sin^2\theta - (u_y+v_x)\sin\theta\cos\theta + v_y\cos^2\theta\right]; \tag{3.2}$$

then *exact integral equations* for the wave equation system (3.1) are given as

$$\phi_P = \frac{1}{2\pi}\int_0^{2\pi} (\phi_Q - u_Q\,\cos\theta - v_Q\,\sin\theta)\,\mathrm{d}\theta + \tilde{S}_1, \tag{3.3}$$

$$u_P = \frac{1}{2}u_{P'} + \frac{1}{2\pi}\int_0^{2\pi} (-\phi_Q\,\cos\theta + u_Q\,\cos^2\theta + v_Q\,\sin\theta\,\cos\theta)\,\mathrm{d}\theta + \tilde{S}_2, \tag{3.4}$$

$$v_P = \frac{1}{2}v_{P'} + \frac{1}{2\pi}\int_0^{2\pi} (-\phi_Q\,\sin\theta + u_Q\,\cos\theta\,\sin\theta + v_Q\,\sin^2\theta)\,\mathrm{d}\theta + \tilde{S}_3, \tag{3.5}$$

where

$$\tilde{S}_1 = \frac{-1}{2\pi}\int_0^{2\pi}\int_0^{\triangle t} S(\mathbf{x}+c\tau\mathbf{n}(\theta), t+\Delta t-\tau,\theta)\,\mathrm{d}\tau\,\mathrm{d}\theta,$$

$$\tilde{S}_2 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\triangle t} \cos\theta S(\mathbf{x}+c\tau\mathbf{n}(\theta), t+\Delta t-\tau,\theta)\,\mathrm{d}\tau\,\mathrm{d}\theta$$

$$-\frac{1}{2\pi}\int_0^{2\pi}\int_0^{\triangle t} \left[c\phi_x(\mathbf{x}, t+\Delta t-\tau)\sin^2\theta - c\phi_y(\mathbf{x}, t+\Delta t-\tau)\sin\theta\cos\theta\right]\,\mathrm{d}\tau\,\mathrm{d}\theta,$$

$$\tilde{S}_3 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\triangle t} \sin\theta S(\mathbf{x}+c\tau\mathbf{n}(\theta), t+\Delta t-\tau,\theta)\,\mathrm{d}\tau\,\mathrm{d}\theta$$

$$-\frac{1}{2\pi}\int_0^{2\pi}\int_0^{\Delta t} \left[c\phi_y(\mathbf{x}, t+\Delta t-\tau)\cos^2\theta - c\phi_x(\mathbf{x}, t+\Delta t-\tau)\sin\theta\cos\theta\right]\,\mathrm{d}\tau\,\mathrm{d}\theta.$$

The above integral equations give us an implicit formulation of the solution at the point $P = (x, y, t^{n+1})$. In order to obtain an explicit numerical scheme it is necessary to use numerical quadrature rules in order to approximate the time integral from 0 to Δt. Using the backward rectangle rule leads to an $\mathcal{O}(\Delta t^2)$ approximation of the time integrals appearing in $\tilde{S}_1$, $\tilde{S}_2$, and $\tilde{S}_3$. Further we use the following result [13, Lemma 2.1]:

$$\Delta t\int_0^{2\pi} S(t,\theta)\mathrm{d}\theta = \int_0^{2\pi}(u\cos\theta + v\sin\theta)\mathrm{d}\theta,$$

$$\Delta t\int_0^{2\pi} S(t,\theta)\cos\theta\mathrm{d}\theta = \int_0^{2\pi}(u\cos 2\theta + v\sin 2\theta)\mathrm{d}\theta,$$

$$\Delta t\int_0^{2\pi} S(t,\theta)\sin\theta\mathrm{d}\theta = \int_0^{2\pi}(u\sin 2\theta + v\cos 2\theta)\mathrm{d}\theta. \tag{3.6}$$

Note that these formulae allow us to replace the derivatives of our dependent variables in S by the variables themselves. Rectangle rule approximation for the time integral and (3.6) yield the so-called EG3 approximate evolution operator.

Approximate evolution operator for EG3.

$$\phi_P = \frac{1}{2\pi}\int_0^{2\pi}(\phi_Q - 2u_Q\cos\theta - 2v_Q\sin\theta)\mathrm{d}\theta + O(\Delta t^2), \tag{3.7}$$

$$u_P = \frac{1}{2}u_{P'} + \frac{1}{2\pi}\int_0^{2\pi}(-2\phi_Q\cos\theta + u_Q(3\cos^2\theta - 1) + 3v_Q\sin\theta\cos\theta)\mathrm{d}\theta + O(\Delta t^2), \tag{3.8}$$

$$v_P = \frac{1}{2}v_{P'} + \frac{1}{2\pi}\int_0^{2\pi}(-2\phi_Q\sin\theta + 3u_Q\sin\theta\cos\theta + v_Q(3\sin^2\theta - 1))\mathrm{d}\theta + O(\Delta t^2). \tag{3.9}$$

We refer the reader to [13, 28] for other approximate evolution operators EG1, EG2, EG4. In what follows we will concentrate on the stability analysis of the EG3 scheme, for which the best numerical results have been obtained; see [13]. The stability analysis for other schemes can be done in an analogous way.

4. Evolution Galerkin schemes. In this section we describe EG schemes in the finite difference framework as well as FVEG schemes. The main idea behind EG schemes is the following. Transported quantities are evolved in time along the bicharacteristics and then projected onto a finite element space. These methods connect finite element methods with the theory of bicharacteristics. In the finite volume framework the approximate operators are used only in order to compute fluxes on cell interfaces. Thus, instead of one-dimensional Riemann solvers, which work only in the normal directions to the cell interfaces, we compute the approximate solution at cell interfaces by a multidimensional evolution operator. This can be considered as a predictor step. In the corrector step the finite volume update is made.

Consider a mesh in $\mathbb{R}^2$, which consists of the square mesh cells

$$\begin{aligned}\Omega_{kl} &= \left[\left(k-\frac{1}{2}\right)h, \left(k+\frac{1}{2}\right)h\right] \times \left[\left(l-\frac{1}{2}\right)h, \left(l+\frac{1}{2}\right)h\right] \\ &= \left[x_k - \frac{h}{2}, x_k + \frac{h}{2}\right] \times \left[y_l - \frac{h}{2}, y_l + \frac{h}{2}\right],\end{aligned}$$

where $h > 0$ is the mesh size parameter, and $k, l \in \mathbb{Z}$. Let us denote by $E(s) : (L^2(\mathbb{R}^2))^p \to (L^2(\mathbb{R}^2))^p$ the exact evolution operator for a general hyperbolic system (2.1), i.e.,

$$\mathbf{U}(., t+s) = E(s)\mathbf{U}(., t). \tag{4.1}$$

We suppose that S_h^m is a finite element space consisting of piecewise polynomials of degree $m \geq 0$ with respect to the square mesh. Assume a constant time step, i.e., $t_n = n\Delta t$. Let $\mathbf{U}^n$ be an approximation in the space S_h^m to the exact solution $\mathbf{U}(., t_n)$ at time $t_n \geq 0$. We consider $E_\tau : (L^2(\mathbb{R}^2))^p \to (L^2(\mathbb{R}^2))^p$ to be a suitable approximate evolution operator for $E(\tau)$. In practice we will use restrictions of E_τ to the subspace S_h^m for $m \geq 0$. Then we can define the general class of EG methods as follows.

DEFINITION 4.1. *Starting from some initial data $\mathbf{U}^0 \in S_h^m$ at time $t = 0$, an* EG method *is recursively defined by means of*

$$\mathbf{U}^{n+1} = P_h E_\tau \mathbf{U}^n, \tag{4.2}$$

where P_h is the L^2-projection given by the integral averages in the following way:

$$P_h \mathbf{U}^n|_{\Omega_{kl}} = \frac{1}{|\Omega_{kl}|} \int_{\Omega_{kl}} \mathbf{U}(x, y, t_n) \mathrm{d}x\mathrm{d}y. \tag{4.3}$$

In this paper we will limit our considerations to the cases where $m = 0$. In this case the integrals that we obtain from the projection are evaluated either exactly using the fact that the approximate values $\mathbf{U}^n$ are piecewise constant or by means of some numerical quadrature rules. Using piecewise constants, the resulting schemes will only be of first order accuracy, even when E_τ is approximated to a higher order. Higher order accuracy can be obtained either by taking $m > 0$ or by inserting a recovery stage R_h before the evolution step in (4.2) to give

$$\mathbf{U}^{n+1} = P_h E_\tau R_h \mathbf{U}^n. \tag{4.4}$$

Here we have denoted by $R_h : S_h^m \to S_h^r$ a recovery operator, $r > m \geq 0$, and consider our approximate evolution operator E_τ on S_h^r. To implement (4.4), rather complex three-dimensional integrals need to be evaluated exactly. This approach seems to be hardly feasible for efficient derivation and implementation of higher order methods. A simplification that we used is to apply the multidimensional evolution only on the cell interfaces. This latter approach leads to the FVEG methods.

DEFINITION 4.2. *Starting from some initial data* $\mathbf{U}^0 \in S_h^m$, *the* finite volume evolution Galerkin (FVEG) method *is recursively defined by means of*

$$\mathbf{U}^{n+1} = \mathbf{U}^n - \frac{1}{h}\int_0^{\Delta t} \sum_{j=1}^{2} \delta_{x_j} \mathbf{f}_j(\tilde{\mathbf{U}}^{n+\frac{\tau}{\Delta t}}) \mathrm{d}\tau, \tag{4.5}$$

where $\delta_{x_j}\mathbf{f}_j(\tilde{\mathbf{U}}^{n+\frac{\tau}{\Delta t}})$ *represents an approximation to the edge flux difference and* δ_x *is defined by* $\delta_x v(x) = v(x+\frac{h}{2}) - v(x-\frac{h}{2})$. *The cell boundary value* $\tilde{\mathbf{U}}^{n+\frac{\tau}{\Delta t}}$ *is evolved using the approximate evolution operator* E_τ *to* $t_n + \tau$ *and averaged along the cell boundary, i.e.,*

$$\tilde{\mathbf{U}}^{n+\frac{\tau}{\Delta t}} = \sum_{k,l\in\mathbb{Z}} \left(\frac{1}{|\partial\Omega_{kl}|} \int_{\partial\Omega_{kl}} E_\tau R_h \mathbf{U}^n dS \right) \chi_{kl}, \tag{4.6}$$

where χ_{kl} *is the characteristic function of* $\partial\Omega_{kl}$.

For more details on higher order FVEG schemes, see [10], [15], [17], where the error analysis as well as numerical experiments are presented. Using the L^2-projection (4.3), the approximate evolution operator E_τ, and (4.5), (4.6), the EG and FVEG schemes can be written in the finite difference form

$$\mathbf{U}_{kl}^{n+1} = \mathbf{U}_{kl}^n + \sum_{r=1-}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \mathbf{U}_{k+r,l+s}^n, \tag{4.7}$$

where

$$\mathcal{C}_{rs} = \begin{pmatrix} \alpha_{rs}^1 & \beta_{rs}^1 & \gamma_{rs}^1 \\ \alpha_{rs}^2 & \beta_{rs}^2 & \gamma_{rs}^2 \\ \alpha_{rs}^3 & \beta_{rs}^3 & \gamma_{rs}^3 \end{pmatrix}. \tag{4.8}$$

Here the entries $\alpha_{rs}^m, \beta_{rs}^m, \gamma_{rs}^m$, $m = 1, 2, 3$, are chosen appropriately according to the approximate evolution operator E_τ used. In the appendix the stencil matrices α^m, β^m, and γ^m, $m = 1, 2, 3$, are displayed for some EG schemes.

5. Basic tools. As we mentioned above our stability considerations are based on Fourier analysis. We first recall some basic concepts; see, e.g., [24]. Let $\{\psi_{kl}^n\}_{k,l=-\infty}^{\infty}$ be a two-dimensional sequence in ℓ_2.

DEFINITION 5.1. *The discrete Fourier transformation of* $\{\psi_{kl}^n\} \in \ell_2$ *is the function* $\hat{\psi^n} \in L_2\big(\big[-\frac{\pi}{h}, \frac{\pi}{h}\big] \times \big[-\frac{\pi}{h}, \frac{\pi}{h}\big]\big)$ *defined by*

$$\hat{\psi}^n(\xi,\eta) = h^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \psi_{kl}^n \exp^{-ih(k\xi+l\eta)}.$$

Similarly to the continuous Fourier transform, we have both an inverse formula and Parseval's identity.

LEMMA 5.2 (inverse formula). *If* $\{\psi_{kl}^n\} \in \ell_2$ *and* $\hat{\psi}^n$ *is the discrete Fourier transform of* $\{\psi_{kl}^n\}$*, then*

$$\psi_{kl}^n = \frac{1}{4\pi^2} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \hat{\psi}^n(\xi,\eta)\, \exp^{ih(k\xi+l\eta)}\, \mathrm{d}\xi\, \mathrm{d}\eta.$$

LEMMA 5.3 (Parseval's identity). *If* $\{\psi_{kl}^n\} \in \ell_2$ *and* $\hat{\psi}^n$ *is the discrete Fourier transform of* $\{\psi_{kl}^n\}$*, then*

$$||\hat{\psi}^n|| = ||\psi_{kl}^n||,$$

where the first norm is the L_2*-norm on* $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right] \times \left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$ *and the second norm is the* ℓ_2*-norm.*

Hence we have the following result.

LEMMA 5.4. *The sequence* $\{\psi_{kl}^n\}$ *is bounded in* ℓ_2 *if and only if the sequence* $\{\hat{\psi}^n\}$ *is bounded in* $L_2\left(\left[-\frac{\pi}{h}, \frac{\pi}{h}\right] \times \left[-\frac{\pi}{h}, \frac{\pi}{h}\right]\right)$.

In order to study the stability of linear numerical schemes the Fourier transform is used. This leads to a bound on the spectral radius of the so-called amplification matrix. The spectral radius of a square complex matrix $\mathcal{A}$ with eigenvalues λ_i is defined to be

$$\rho(\mathcal{A}) = \max_i |\lambda_i|. \tag{5.1}$$

The spectral norm of the matrix $\mathcal{A}$ is defined as

$$||\mathcal{A}|| = \sup_{\mathbf{x}\neq 0} \frac{||\mathcal{A}\mathbf{x}||}{||\mathbf{x}||}. \tag{5.2}$$

The norms on the right-hand side of (5.2) are the Euclidean norms of the vectors $\mathcal{A}\mathbf{x}$ and $\mathbf{x}$, respectively. Note that for the spectral norm, as for any matrix norm, we always have $||\mathcal{A}|| \geq \rho(\mathcal{A})$.

6. Estimate of the stability limit. In [13, Lemma 5.1] Lukáčová-Medvid'ová, Morton, and Warnecke proved the following stability result for EG schemes. There exists $\nu_{max} < 1$ such that EG schemes for the two-dimensional wave equation system (3.1) are stable for any ν such that $0 \leq \nu \leq \nu_{max}$, where $\nu = c\frac{\Delta t}{h}$ is the CFL number. The goal of this section is to estimate ν_{max} for the EG3 scheme by means of a von Neumann stability analysis. We refer to [2] for a related approach used to estimate stability limits of other finite volume schemes for the Maxwell equations. Analogous calculations can be done also for other EG schemes of type EG1, EG2, EG4 as well as for the FVEG schemes. First we apply the discrete Fourier transform to both sides of (4.7):

$$\hat{\mathbf{U}}^{n+1} = \hat{\mathbf{U}}^n + h^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \left(\sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \mathbf{U}_{k+rl+s}^n \right) \exp^{-ih(k\xi+l\eta)}. \tag{6.1}$$

By making the change of variables $k' = k + r$ and $l' = l + s$ we get

$$h^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \left(\sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \mathbf{U}_{k+rl+s}^n \right) \exp^{-ih(k\xi+l\eta)}$$

$$= \sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \exp^{ih(r\xi+s\eta)} \left(h^2 \sum_{k'=-\infty}^{\infty} \sum_{l'=-\infty}^{\infty} \mathbf{U}_{k'l'}^n \exp^{-ih(k'\xi+l'\eta)} \right)$$

$$= \sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \exp^{ih(r\xi+s\eta)} \hat{\mathbf{U}}^n. \tag{6.2}$$

Thus, using this expression in (6.1), we get

$$\hat{\mathbf{U}}^{n+1} = \left(\mathcal{I} + \sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \exp^{ih(r\xi+s\eta)} \right) \hat{\mathbf{U}}^n, \tag{6.3}$$

where $\mathcal{I}$ is the identity matrix. The coefficient of $\hat{\mathbf{U}}^n$ in (6.3),

$$\mathcal{T}(\xi,\eta) = \mathcal{I} + \sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \exp^{ih(r\xi+s\eta)}, \tag{6.4}$$

is called the amplification matrix of the finite difference scheme (4.7). Applying recursively the result of (6.3) $n+1$ times yields

$$\hat{\mathbf{U}}^{n+1} = \left(\mathcal{I} + \sum_{r=-1}^{1} \sum_{s=-1}^{1} \mathcal{C}_{rs} \exp^{ih(r\xi+s\eta)} \right)^{n+1} \hat{\mathbf{U}}^0 = \mathcal{T}^{n+1}(\xi,\eta)\hat{\mathbf{U}}^0. \tag{6.5}$$

We note that if $||\mathcal{T}(\xi,\eta)|| \le 1$, then $||\hat{\mathbf{U}}^{n+1}|| \le ||\hat{\mathbf{U}}^0||$, which means that the $\{\hat{\mathbf{U}}^n\}$ is L^2-stable. Consider the EG3 scheme, i.e., the numerical scheme based on equations (3.7)–(3.9); see also the stencil matrices in the appendix. After some calculation we obtain the entries of the amplification matrix $\mathcal{T}(\xi,\eta)$:

$$T_{11}(\xi,\eta) = 1 + \frac{\nu^2}{\pi} - \frac{4\nu}{\pi} + \frac{\nu^2}{\pi}\cos(h\xi)\cos(h\eta) + \left(\frac{2\nu}{\pi} - \frac{\nu^2}{\pi}\right)(\cos(h\xi) + \cos(h\eta)),$$

$$T_{12}(\xi,\eta) = -i\left(\frac{4\nu^2}{3\pi}\sin(h\xi)\cos(h\eta) + \left(\nu - \frac{4\nu^2}{3\pi}\right)\sin(h\xi)\right),$$

$$T_{13}(\xi,\eta) = -i\left(\frac{4\nu^2}{3\pi}\cos(h\xi)\sin(h\eta) + \left(\nu - \frac{4\nu^2}{3\pi}\right)\sin(h\eta)\right),$$

$$T_{22}(\xi,\eta) = 1 - \frac{2\nu}{\pi} + \frac{\nu^2}{2\pi} + \frac{\nu^2}{2\pi}\cos(h\xi)\cos(h\eta) + \left(\frac{2\nu}{\pi} - \frac{\nu^2}{2\pi}\right)\cos(h\xi) - \frac{\nu^2}{2\pi}\cos(h\eta),$$

$$T_{23}(\xi,\eta) = \frac{-3\nu^2}{8}\sin(h\xi)\sin(h\eta),$$

$$T_{33}(\xi,\eta) = 1 - \frac{2\nu}{\pi} + \frac{\nu^2}{2\pi} + \frac{\nu^2}{2\pi}\cos(h\xi)\cos(h\eta) + \left(\frac{2\nu}{\pi} - \frac{\nu^2}{2\pi}\right)\cos(h\eta) - \frac{\nu^2}{2\pi}\cos(h\xi),$$

$$T_{21}(\xi,\eta) = T_{12}(\xi,\eta), \qquad T_{31}(\xi,\eta) = T_{13}(\xi,\eta), \qquad T_{32}(\xi,\eta) = T_{23}(\xi,\eta).$$

Using the substitutions $S_\xi = \sin(h\xi)$, $s_\xi = \sin(\frac{h\xi}{2})$, $S_\eta = \sin(h\eta)$, and $s_\eta = \sin(\frac{h\eta}{2})$, we can write the amplification matrix $\mathcal{T} = \mathcal{T}(\xi,\eta)$ as

$$\mathcal{T} = \begin{pmatrix} C_{11} & -i\nu C_\xi & -i\nu C_\eta \\ -i\nu C_\xi & C_{22} & \nu^2 C_{\xi\eta} \\ -i\nu C_\eta & \nu^2 C_{\xi\eta} & C_{33} \end{pmatrix},$$

where

$$
\begin{aligned}
C_{11} &= 1 - \frac{4\nu}{\pi}(s_\xi^2 + s_\eta^2) + \frac{4\nu^2}{\pi} s_\xi^2 s_\eta^2,\\
C_\xi &= S_\xi\left(1 - \frac{8\nu}{3\pi} s_\eta^2\right),\\
C_\eta &= S_\eta\left(1 - \frac{8\nu}{3\pi} s_\xi^2\right),\\
C_{\xi\eta} &= \frac{-3}{8} S_\xi S_\eta,\\
C_{22} &= 1 - \frac{4\nu}{\pi} s_\xi^2 + \frac{2\nu^2}{\pi} s_\xi^2 s_\eta^2,\\
C_{33} &= 1 - \frac{4\nu}{\pi} s_\eta^2 + \frac{2\nu^2}{\pi} s_\xi^2 s_\eta^2.
\end{aligned}
$$

Set

$$
\mathcal{E} = \begin{pmatrix} i & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix};
$$

then

$$
\mathcal{Q} = \begin{pmatrix} C_{11} & -\nu C_\xi & -\nu C_\eta \\ \nu C_\xi & C_{22} & \nu^2 C_{\xi\eta} \\ \nu C_\eta & \nu^2 C_{\xi\eta} & C_{33} \end{pmatrix} = \mathcal{E}^{-1}\mathcal{T}\mathcal{E},
$$

which means that $\mathcal{T}$ and $\mathcal{Q}$ are similar matrices and thus have the same eigenvalues. Moreover, the matrix $\mathcal{Q}$ can be decomposed as

$$
\mathcal{Q} = \mathcal{I} - \nu\,(\mathcal{D} + \mathcal{C}) + \nu^2 \tilde{\mathcal{C}},
$$

where

$$
\mathcal{D} = \begin{pmatrix} d+f & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & f \end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix} 0 & C_\xi & C_\eta \\ -C_\xi & 0 & 0 \\ -C_\eta & 0 & 0 \end{pmatrix}, \quad \tilde{\mathcal{C}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & C_{\xi\eta} \\ 0 & C_{\xi\eta} & 0 \end{pmatrix},
$$

$$
d = \frac{4}{\pi} s_\xi^2 - \frac{2\nu}{\pi} s_\xi^2 s_\eta^2 = \frac{2}{\pi} s_\xi^2 (2 - \nu s_\eta^2), \quad f = \frac{4}{\pi} s_\eta^2 - \frac{2\nu}{\pi} s_\xi^2 s_\eta^2 = \frac{2}{\pi} s_\eta^2 (2 - \nu s_\xi^2).
$$

Let

$$
\mathcal{H} = \mathcal{I} - \nu(\mathcal{D} + \mathcal{C}), \tag{6.6}
$$

and let $||.||_\star$ be an operator norm such that $||\mathcal{H}||_\star = ||\mathcal{J}||_\infty$, where $\mathcal{J}$ is a scaled Jordan normal form of $\mathcal{H}$ having $\epsilon \ll 1$ on the first off-diagonal and eigenvalues of $\mathcal{H}$ on the diagonal. According to [5] we know that the following property holds: $\rho(\mathcal{H}) < 1$ if and only if $||\mathcal{H}||_\star < 1$. Since all norms in finite-dimensional spaces are equivalent we have

$$
||\mathcal{Q} - (\mathcal{I} - \nu(\mathcal{D} + \mathcal{C}))||_\star \le c\,||\mathcal{Q} - (\mathcal{I} - \nu(\mathcal{D} + \mathcal{C}))|| = c\,\nu^2 |C_{\xi\eta}| = O(\nu^2). \tag{6.7}
$$

Thus, using (6.7) we obtain

$$
\rho(\mathcal{Q}) \le ||Q||_\star \le ||\mathcal{H}||_\star + c\nu^2, \tag{6.8}
$$

and it suffices to study the spectrum of matrix $\mathcal{H}$, since $\rho(\mathcal{H}) < 1$ if and only if $||\mathcal{H}||_\star < 1$.

Further, since $\mathcal{H} = \mathcal{I} - \nu(\mathcal{D} + \mathcal{C})$ and due to the form of $\mathcal{D}, \mathcal{C}$ it can be shown readily that $||\mathcal{H}||_\infty = |1 + c_1\nu + c_2\nu^2|$ for some constants $c_1, c_2 \in \mathbb{R}$. Since all norms in finite-dimensional spaces are equivalent, there exist $k_1, k_2 > 0$, such that for any matrix $\mathcal{M} \in \mathbb{R}^{(3,3)}$

$$k_1||\mathcal{M}||_\infty \le ||\mathcal{M}||_\star \le k_2||\mathcal{M}||_\infty .$$

In particular, if $\mathcal{M} = \mathcal{H}$, we have for any $\nu > 0$

$$k_1|1 + c_1\nu + c_2\nu^2| \le ||\mathcal{H}||_\star \le k_2|1 + c_1\nu + c_2\nu^2|. \tag{6.9}$$

Thus, it is clear that $||\mathcal{H}||_\star$ depends at most quadratically on ν, i.e., $||\mathcal{H}||_\star = |1 + c_3\nu + c_4\nu^2|$. Now, if we assume that $||\mathcal{H}||_\star < 1$, the linear term has to be negative, i.e., $c_3 < 0$, as otherwise $||\mathcal{H}||_\star$ cannot be strictly less than 1 for $\nu \le 1$. Thus, if $||\mathcal{H}||_\star < 1$, we can find small enough ν such that $||\mathcal{H}||_\star + c\nu^2 \le 1$ and $\rho(Q) \le 1$ due to (6.8).

Unfortunately, we cannot give any quantitative estimate on ν since we do not know how large the constant c in (6.8) is. However, we know from [13] that there exists $\nu_{\max} > 0$ such that for all $\nu \in (0, \nu_{\max}]$ we have $\rho(Q) \le 1$.

In what follows we will study the spectrum of the matrix $\mathcal{H}$ and find ν such that $\rho(\mathcal{H}) < 1$. Note that for all $(\xi, \eta) \in [\frac{-\pi}{h}, \frac{\pi}{h}] \times [\frac{-\pi}{h}, \frac{\pi}{h}]$ the entries of $\mathcal{H}$ are bounded. We need to estimate the spectral radius of $\mathcal{H}$ for all choices of ξ, η, and ν, $0 < \nu \le 1$.

First of all, it is easy to see that in a special case when $\xi = 0 = \eta$ we have $d = f = C_\xi = C_\eta = C_{\xi\eta} = 0$ and $\mathcal{Q} = \mathcal{I} = \mathcal{H}$. Thus, trivially $\rho(\mathcal{H}) = \rho(\mathcal{Q}) = 1$ for any ν. Therefore in what follows it suffices to study the case when $\xi \ne 0$ or $\eta \ne 0$.

Since $0 \le s_\xi^2 \le 1$ and $0 \le s_\eta^2 \le 1$ and $\nu \le 1$, then $d \ge 0$ and $f \ge 0$. Now the matrices $\mathcal{D}, \mathcal{C}$ are real and $\mathcal{C}$ is skewsymmetric. Hence $\mathcal{D} + \mathcal{C}$ has either three real eigenvalues or one real eigenvalue and two complex conjugate eigenvalues.

Consider a *real eigenvalue*, say $\lambda = \lambda_r$. Let $\mathbf{v} = (v_1, v_2, v_3)$ be the corresponding eigenvector; then $\mathbf{v}^T(\mathcal{D} + \mathcal{C})\mathbf{v} = \mathbf{v}^T\lambda_r\mathbf{v}$. Since $\mathcal{C}$ is skewsymmetric, then $\mathbf{v}^T\mathcal{C}\mathbf{v} = 0$. Hence we get

$$(d + f - \lambda_r)v_1^2 + (d - \lambda_r)v_2^2 + (f - \lambda_r)v_3^2 = 0. \tag{6.10}$$

The coefficients in (6.10) cannot all have the same sign for v_1^2, v_2^2, $v_3^2 > 0$. Therefore, we get the estimates

$$0 \le \min(d, f) \le \lambda_r \le d + f. \tag{6.11}$$

Let μ_r be a real eigenvalue of $\mathcal{H}$; then $\mu_r = 1 - \nu\lambda_r$. Hence $|\mu_r| < 1$ is equivalent to $-1 < 1 - \nu\lambda_r < 1$. According to (6.11) we assume now that

$$\lambda_r > 0; \tag{6.12}$$

the case $\lambda_r = 0$ will be treated separately later; cf. (6.22). Further,

$$1 - \frac{4\nu}{\pi}\left(s_\xi^2 + s_\eta^2\right) + \frac{4\nu^2}{\pi}s_\xi^2 s_\eta^2 \le 1 - \nu\lambda_r < 1.$$

To ensure that $|\mu_r| < 1$ we need

$$1 - \frac{4\nu}{\pi}\left(s_\xi^2 + s_\eta^2\right) + \frac{4\nu^2}{\pi}s_\xi^2 s_\eta^2 > -1.$$

The last inequality reads

$$\nu^2\left(\frac{4}{\pi}s_\xi^2 s_\eta^2\right)-\nu\left(\frac{4}{\pi}\left(s_\xi^2+s_\eta^2\right)\right)+2>0. \tag{6.13}$$

It suffices to bound ν so that $2-\nu\frac{4}{\pi}(s_\xi^2+s_\eta^2)>0$. Since $(s_\xi^2+s_\eta^2)\le 2$, this is true if

$$\nu<\frac{\pi}{4}\approx 0.7854. \tag{6.14}$$

Now let us assume that μ_c is a *complex eigenvalue* of $\mathcal{H}$. Then $\mu_c=1-\nu\lambda_c$, where λ_c is a complex eigenvalue of the matrix $\mathcal{D}+\mathcal{C}$. This implies that

$$|\mu_c|^2=1-2\nu\mathrm{Re}(\lambda_c)+\nu^2|\lambda_c|^2.$$

Thus $|\mu_c|^2<1$ is equivalent to $\nu^2|\lambda_c|^2-2\nu\mathrm{Re}(\lambda_c)<0$. Since $\lambda_r>0$ (cf. (6.12)), we have

$$\nu^2\lambda_r|\lambda_c|^2-2\nu\lambda_r\mathrm{Re}(\lambda_c)<0. \tag{6.15}$$

Let $b=C_\xi$ and $c=C_\eta$. It is well known that

$$\begin{aligned}\det(\mathcal{D}+\mathcal{C})&=d^2f+f^2d+b^2f+c^2d=\lambda_r|\lambda_c|^2,\\ \mathrm{Tr}(\mathcal{D}+\mathcal{C})&=2(d+f)=\lambda_r+\lambda_c+\bar{\lambda_c}=\lambda_r+2\mathrm{Re}(\lambda_c).\end{aligned}$$

Hence (6.15) reads

$$p(\lambda_r)=\lambda_r^2-2(d+f)\lambda_r+\nu(d^2f+f^2d+b^2f+c^2d)<0. \tag{6.16}$$

Let us consider the polynomial

$$p=p(\lambda)=\lambda^2-2(d+f)\lambda+\nu(d^2f+f^2d+b^2f+c^2d).$$

The discriminant of p gives

$$\begin{aligned}\Delta&=4(d+f)^2-4\nu(d^2f+f^2d+b^2f+c^2d)\\ &=4(d^2+f^2)+8fd-4\nu(d^2f+f^2d+b^2f+c^2d).\end{aligned}$$

It suffices to show that the following inequality holds:

$$8fd-4\nu(d^2f+f^2d+b^2f+c^2d)>0,$$

which leads to $\Delta>4(d^2+f^2)\ge 0$. Now

$$\begin{aligned}8fd=\frac{32}{\pi^2}s_\xi^2s_\eta^2(2-\nu s_\eta^2)(2-\nu s_\xi^2)&=\frac{32}{\pi^2}s_\xi^2s_\eta^2(4-2\nu(s_\xi^2+s_\eta^2)+\nu^2s_\xi^2s_\eta^2)\\ &\ge\frac{32}{\pi^2}s_\xi^2s_\eta^2(4-2\nu(s_\xi^2+s_\eta^2)).\end{aligned}$$

Note that the last inequality is strict if $\xi\neq 0$ and $\eta\neq 0$. Hence,

$$8fd\ge\frac{32}{\pi^2}s_\xi^2s_\eta^2(4-4\nu)=\frac{128}{\pi^2}s_\xi^2s_\eta^2(1-\nu). \tag{6.17}$$

If $\xi\neq 0$ and $\eta\neq 0$, the inequality in (6.17) is strict.

Further, we have

$$d^2 f = \frac{8}{\pi^3} s_\xi^4 s_\eta^2 (2-\nu s_\eta^2)^2 (2-\nu s_\xi^2) \le \frac{64}{\pi^3} s_\xi^4 s_\eta^2 \le \frac{64}{\pi^3}, \tag{6.18}$$

$$b^2 f = S_\xi^2 (1 - \frac{8\nu^2}{3\pi} s_\eta^2)^2 \frac{2}{\pi} s_\eta^2 (2-\nu s_\xi^2) \le \frac{4}{\pi} S_\xi^2 s_\eta^2 \le \frac{4}{\pi}. \tag{6.19}$$

Again, note that in the case that either $\xi = 0$ or $\eta = 0$, the inequality in (6.18) is strict.

Analogously, we obtain

$$f^2 d \le \frac{64}{\pi^3} \qquad \text{and} \qquad c^2 d \le \frac{4}{\pi}.$$

Therefore,

$$-4\nu(d^2 f + f^2 d + b^2 f + c^2 d) \ge -4 \frac{128 + 8\pi^2}{\pi^3} \nu; \tag{6.20}$$

if $\xi = 0$ or $\eta = 0$, the above inequality is strict.

Combining (6.17) and (6.20) we get

$$\begin{aligned} 8df - 4\nu(d^2 f + f^2 d + b^2 f + c^2 d) &> \frac{128}{\pi^2} s_\xi^2 s_\eta^2 (1-\nu) - 4\left(\frac{128+8\pi^2}{\pi^3}\right)\nu \\ &= \frac{128}{\pi^2} s_\xi^2 s_\eta^2 - \nu\left(\frac{128}{\pi^2} s_\xi^2 s_\eta^2 + 4\left(\frac{128+8\pi^2}{\pi^3}\right)\right) \ge 0. \end{aligned}$$

The last inequality implies

$$\nu \le \frac{\frac{128}{\pi^2} s_\xi^2 s_\eta^2}{4\left(\frac{128+8\pi^2}{\pi^3}\right) + \frac{128}{\pi^2} s_\xi^2 s_\eta^2} \le \frac{\frac{128}{\pi^2}}{4\left(\frac{128+8\pi^2}{\pi^3}\right) + \frac{128}{\pi^2} s_\xi^2 s_\eta^2}.$$

Since

$$4\left(\frac{128+8\pi^2}{\pi^3}\right) + \frac{128}{\pi^2} s_\xi^2 s_\eta^2 \ge 4\left(\frac{128+8\pi^2}{\pi^3}\right)$$

we then have

$$\frac{1}{4\left(\frac{128+8\pi^2}{\pi^3}\right) + \frac{128}{\pi^2} s_\xi^2 s_\eta^2} \le \frac{1}{4\left(\frac{128+8\pi^2}{\pi^3}\right)}.$$

Therefore we get

$$\nu \le \frac{\frac{128}{\pi^2}}{\frac{4(128+8\pi^2)}{\pi^3}} = \frac{32\pi}{128+8\pi^2} \approx 0.4858. \tag{6.21}$$

Thus we have obtained a sufficient condition on ν for $\Delta > 0$. For $\nu \le 0.4858$ we have $\Delta > 4(d^2 + f^2) \ge 0$.

Since $\lambda_r > 0$ (cf. (6.12)), $p(\lambda)$ has two distinct real roots r_1 and r_2, where

$$r_1 = (d+f) - \frac{\sqrt{\Delta}}{2}, \qquad r_2 = (d+f) + \frac{\sqrt{\Delta}}{2}.$$

Inequality (6.11) gives $\lambda_r < r_2$. To show that $\lambda_r > r_1$ note that from $\Delta > 4(d^2+f^2)$ we have $r_1 < (d+f) - \sqrt{d^2+f^2}$. Furthermore $\sqrt{d^2+f^2} \geq \max(d,f)$. Therefore

$$r_1 < (d+f) - \sqrt{d^2+f^2} \leq (d+f) - \max(d,f) = \min(d,f) \leq \lambda_r.$$

Hence $\lambda_r \in (r_1, r_2)$. This implies that $p(\lambda_r) < 0$, as we wished to show; cf. (6.16).

Moreover, since $\lambda_r \geq 0$ (cf. (6.11)), we need to consider the case $\lambda_r = 0$. Then either $d = 0$ or $f = 0$. Suppose $d = \frac{2}{\pi} s_\xi^2 (2 - \nu s_\eta^2) = 0$; the case $f = 0$ is analogous. Then we have that $s_\xi = 0$, $\xi = 0$, and

$$\mathcal{D} + \mathcal{C} = \begin{pmatrix} \frac{4}{\pi} s_\eta^2 & 0 & S_\eta \\ 0 & 0 & 0 \\ -S_\eta & 0 & \frac{4}{\pi} s_\eta^2 \end{pmatrix}. \tag{6.22}$$

Note that in this case $\mathcal{Q} = \mathcal{H}$, $\mu_r = 1$ and we need to find the condition on ν to ensure that $\rho(\mathcal{H}) = \rho(\mathcal{Q}) \leq 1$. The eigenvalues of $\mathcal{D} + \mathcal{C}$ are 0, $\frac{4}{\pi} s_\eta^2 \pm i S_\eta$. Now $|\mu_c|^2 = |1 - \nu\lambda_c|^2 \leq 1$ gives

$$\left(1 - \nu\left(\frac{4}{\pi} s_\eta^2 + i S_\eta\right)\right)\left(1 - \nu\left(\frac{4}{\pi} s_\eta^2 - i S_\eta\right)\right) = \left(\left(1 - \frac{4}{\pi}\nu s_\eta^2\right)^2 + \nu^2 S_\eta^2\right)$$
$$= 1 - \frac{8\nu}{\pi} s_\eta^2 + \frac{16\nu^2}{\pi^2} s_\eta^4 + \nu^2 S_\eta^2 \leq 1.$$

This leads to

$$-\frac{8\nu}{\pi} s_\eta^2 + \nu^2 \left(\frac{16}{\pi^2} s_\eta^4 + S_\eta^2\right) \leq 0.$$

Suppose $s_\eta \neq 0$, as otherwise $\eta = 0$ and $\xi = 0$, which is a special case that has already been considered above. Then we have

$$-\frac{8}{\pi} + \nu\left(\frac{16}{\pi^2} s_\eta^2 + \left(\frac{S_\eta}{s_\eta}\right)^2\right) \leq 0,$$
$$\nu\left(\frac{16}{\pi^2} s_\eta^2 + \left(\frac{S_\eta}{s_\eta}\right)^2\right) \leq \frac{8}{\pi}.$$

The last inequality yields

$$\nu \leq \frac{\frac{8}{\pi}}{\left(\frac{16}{\pi^2} s_\eta^2 + \left(\frac{S_\eta}{s_\eta}\right)^2\right)}. \tag{6.23}$$

Since

$$\frac{16}{\pi^2} s_\eta^2 + \left(\frac{S_\eta}{s_\eta}\right)^2 \leq 4,$$

it suffices to take ν such that

$$\nu \leq \frac{2}{\pi} \approx 0.6366. \tag{6.24}$$

Finally (6.14), (6.21), and (6.24) imply that if

$$\nu \leq \frac{32\pi}{128 + 8\pi^2} \approx 0.4858, \tag{6.25}$$

then either $\rho(\mathcal{Q}) = \rho(\mathcal{H}) = 1$ or $\rho(\mathcal{H}) < 1$. Hence we have proved the following result.

LEMMA 6.1. *Consider the EG3 scheme. Then there exists ν small enough such that $\rho(\mathcal{T}) \leq 1$, where $\mathcal{T}$ is the amplification matrix of the discrete operator representing the EG3 scheme.*

More precisely, we know that $\rho(\mathcal{T}) \leq \|\mathcal{H}\|_* + O(\nu^2)$, where $\mathcal{H}$ is the matrix defined in (6.6). *Moreover, if $\nu \leq \frac{32\pi}{128+8\pi} \approx 0.4858$, then $\rho(\mathcal{H}) < 1$ and $\|\mathcal{H}\|_* < 1$, except for the special case when $\rho(\mathcal{H}) = \rho(\mathcal{T}) = 1$. Otherwise, there is ν small enough such that $\rho(\mathcal{T}) \leq 1$.*

TABLE 1
Stability limit using $\rho_{\xi,\eta}(\mathcal{T}(\xi,\eta))$.

$\frac{c\Delta t}{h}$	$\rho_{\xi,\eta}(\mathcal{T}(\xi,\eta))$ for $EG3$
0.10	1.000000000000000
0.20	1.000000000000000
0.30	1.000000000000000
0.40	1.000000000000000
0.50	1.000000000000000
0.58	1.000000000000000
0.59	1.000003244461521
0.60	1.000112236111448
0.70	1.008474049696319

In Table 1 we have estimated the stability limit of the scheme EG3 using the standard MATLAB procedure `eig` for the eigenvalues of the matrix $\mathcal{T}$. Note that our theoretical result $\nu \approx 0.4858$ for a simplified problem gives a stronger estimate on the CFL number than the experimental results for the original problem. They show that the EG3 scheme stays stable up to $\nu = 0.58$. In Figure 2 (left) we plot the eigenvalues of the matrix $\mathcal{H}$ as well as the unit circle. A similar plot with different scale is shown in Figure 2 (right). In Figure 3 we show, using different scales, the eigenvalues of the amplification matrix corresponding to the first order EG3 scheme. We illustrate that it is possible to include all eigenvalues inside the unit circle for the CFL number up to 0.58. Throughout the paper in order to plot the eigenvalues of amplification matrices we have used 100×100 values of $h\xi$, $h\eta \in [-\pi, \pi]$.

7. Approximate evolution operator E_Δ^{const} for piecewise constant data. In [11], Lukáčová-Medvid'ová, Morton and Warnecke proposed new approximate evolution operators E_Δ^{const} and E_Δ^{bilin} for the two-dimensional wave equation system and for the Euler equations of gas dynamics. Extensive numerical experiments presented in [11] indicate that these new operators improve the stability of the FVEG schemes considerably; i.e., in particular, our numerical tests indicated that they have a larger stability range than the EG3 method, which we considered in section 6. We will show that for special choices of discretization techniques stability limits close to the natural limit of 1 can be achieved. Numerical experiments, presented in [11], for these FVEG schemes confirm high accuracy as well as good multidimensional behavior of the new FVEG schemes. The key idea of the development of these new operators was to exploit the fact that the exact explicit solution to the one-dimensional wave equation system is available. Our new approximate operators are constructed in such a way

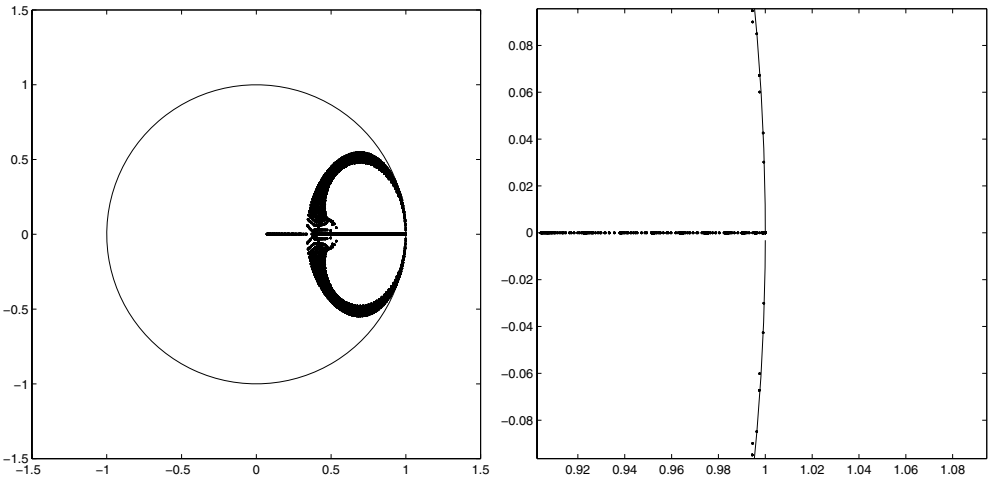


FIG. 2. *Eigenvalues of the matrix* $\mathcal{H}$*,* $CFL = 0.48$*.*

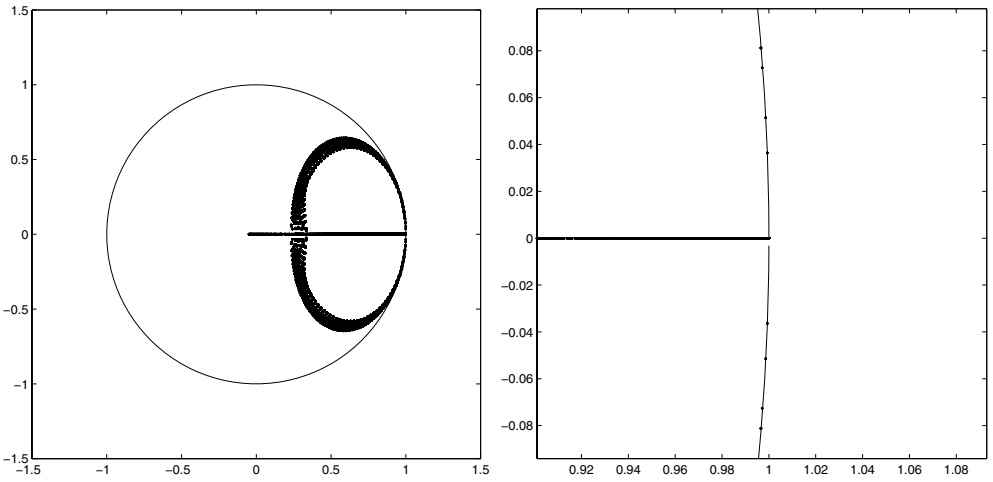


FIG. 3. *Eigenvalues of the amplification matrix of the first order EG*3 *scheme for* $CFL = 0.58$*.*

that this exact solution is reproduced exactly for given one-dimensional data. Thus, the approximate evolution operator E_Δ^{const} calculates exactly any one-dimensional wave which is represented by a piecewise constant data and propagates either in the x- or the y-direction. An analogous situation holds for the operator E_Δ^{bilin} and approximated waves by means of continuous piecewise bilinear data. The approximate evolution operator E_Δ^{const} for piecewise constant data reads (cf. [11])

$$\phi_P = \frac{1}{2\pi}\int_0^{2\pi} (\phi_Q - u_Q \mathrm{sgn}(\cos\theta) - v_Q \mathrm{sgn}(\sin\theta))\mathrm{d}\theta, \tag{7.1}$$

$$u_P = \frac{1}{2\pi}\int_0^{2\pi} \left(-\phi_Q \mathrm{sgn}(\cos\theta) + u_Q\left(\frac{1}{2} + \cos^2\theta\right) + v_Q \sin\theta\cos\theta\right)\mathrm{d}\theta, \tag{7.2}$$

$$v_P = \frac{1}{2\pi}\int_0^{2\pi} \left(-\phi_Q \mathrm{sgn}(\sin\theta) + u_Q \sin\theta\cos\theta + v_Q\left(\frac{1}{2} + \sin^2\theta\right)\right)\mathrm{d}\theta. \tag{7.3}$$

Integrations from 0 to 2π around the sonic circle in (7.1)–(7.3) are evaluated exactly. In this way all of the infinitely many directions of wave propagation are taken into account explicitly. For the cell interface integration along $\partial\Omega$ in (4.6) we have two possibilities. These edge integrals can be computed either exactly or numerically. Exact cell interface integration yields, e.g., for the vertical edge, the intermediate values

$$\begin{aligned}
\tilde{\Phi}_{edge}^{n+\frac{1}{2}} &= \left(1+\frac{\nu}{2\pi}\delta_y^2\right)\mu_x\Phi^n - \left(\frac{1}{2}+\frac{\nu}{4\pi}\delta_y^2\right)\delta_x U^n - \frac{\nu}{\pi}\mu_x\mu_y\delta_y V^n,\\
\tilde{U}_{edge}^{n+\frac{1}{2}} &= -\left(\frac{1}{2}+\frac{\nu}{4\pi}\delta_y^2\right)\delta_x\Phi^n + \left(1+\frac{5\nu}{12\pi}\delta_y^2\right)\mu_x U^n + \frac{\nu}{6\pi}\delta_x\mu_y\delta_y V^n,
\end{aligned} \tag{7.4}$$

where $\mu_x f(x) = \frac{1}{2}\left(f\left(x+\frac{h}{2}\right)+f\left(x-\frac{h}{2}\right)\right)$, $\delta_x^2 f(x) = f(x+h)-2f(x)+f(x-h)$.

The stencil matrices of this FVEG scheme are given in the appendix. Another possibility for evaluating the cell interface integrals is to use some numerical quadrature. In this way, further simplification in the derivation of the scheme can be made. Instead of the two-dimensional integrals along the cell interfaces and around the sonic circle, only the sonic circle integrals need to be evaluated exactly. In our experiments we used the trapezoidal rule and Simpson's rule for the cell interface integration. Thus, we need to determine $\tilde{\mathbf{U}}^{n+\frac{1}{2}}$:

$$\begin{aligned}
\tilde{\Phi}_{vertex}^{n+\frac{1}{2}} &= \mu_x\mu_y\Phi^n - \frac{1}{2}\mu_y\delta_x U^n - \frac{1}{2}\mu_x\delta_y V^n,\\
\tilde{U}_{vertex}^{n+\frac{1}{2}} &= -\frac{1}{2}\mu_y\delta_x\Phi^n + \mu_x\mu_y U^n + \frac{1}{4\pi}\delta_x\delta_y V^n,\\
\tilde{\Phi}_{midpoint}^{n+\frac{1}{2}} &= \mu_x\Phi^n - \frac{1}{2}\delta_x U^n,\\
\tilde{U}_{midpoint}^{n+\frac{1}{2}} &= -\frac{1}{2}\delta_x\Phi^n + \mu_x U^n.
\end{aligned} \tag{7.5}$$

The stencil matrices of the FVEG scheme with trapezoidal and Simpson quadratures for the cell interface integration are given in the appendix.

Analogously to section 6, we can show that the amplification matrix $\mathcal{T}$ of the first order FVEG scheme with exact edge integrals is similar to the matrix

$$\mathcal{Q} = \mathcal{I} - \nu\left(\mathcal{D}+\mathcal{C}\right) + \nu^2\tilde{\mathcal{C}},$$

where the matrix $\mathcal{D}$ is defined, as before, with

$$d = 2s_\xi^2\left(1-\frac{2\nu}{\pi}s_\eta^2\right), \qquad f = 2s_\eta^2\left(1-\frac{2\nu}{\pi}s_\xi^2\right).$$

The matrices $\mathcal{C}$ and $\tilde{\mathcal{C}}$ are given as

$$\mathcal{C} = \begin{pmatrix} 0 & C_\xi - \frac{\nu}{3\pi}S_\xi s_\eta^2 & C_\eta - \frac{\nu}{3\pi}S_\eta s_\xi^2 \\ -C_\xi & 0 & 0 \\ -C_\eta & 0 & 0 \end{pmatrix}, \qquad \tilde{\mathcal{C}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & C_{\xi\eta} \\ 0 & C_{\xi\eta} & 0 \end{pmatrix},$$

where

$$\begin{aligned}
C_\xi &= S_\xi\left(1-\frac{2\nu}{\pi}s_\eta^2\right), \qquad C_\eta = S_\eta\left(1-\frac{2\nu}{\pi}s_\xi^2\right),\\
C_{\xi\eta} &= \frac{-1}{\pi}S_\xi S_\eta.
\end{aligned}$$

Since the matrix $\mathcal{C}$ is not skewsymmetric, it is now not possible to carry out an analysis similar to the one in section 6 in order to estimate the stability limit. Instead we use the MATLAB procedure `eig` to estimate the stability limit. The results are given in Table 2. In column 2 we present the stability limit of the first order FVEG scheme with exact edge integrals. The stability limit of this scheme is improved considerably: the scheme is stable approximately up to CFL = 0.89. Column 3 demonstrates that the first order scheme based on the trapezoidal rule is stable up to the natural stability limit 1. Column 4 shows that the stability of the first order scheme based on Simpson's rule is also increased: the scheme is stable approximately up to CFL = 0.75.

TABLE 2
Stability limit using $\rho_{\xi,\eta}(\mathcal{T}(\xi,\eta))$.

$\frac{c\Delta t}{h}$	Exact	Trapezoidal	Simpson
0.70	1.0000000000	1.0000000000	1.0000000000
0.75	1.0000000000	1.0000000000	1.0000000000
0.76	1.0000000000	1.0000000000	1.0266666667
0.80	1.0000000000	1.0000000000	
0.89	1.0000000000	1.0000000000	
0.90	1.0007993640	1.0000000000	
1.00		1.0000000000	
1.01		1.0200000000	

In Figure 4 we plot, using different scales, the eigenvalues of the amplification matrices corresponding to the first order FVEG schemes based on the operator (7.1)–(7.3). The top two panels are obtained using exact integration along cell interfaces; in the middle panels the trapezoidal rule was used to approximate interface integrals. In the bottom two panels we have plotted eigenvalues of the amplification matrix of the FVEG3 scheme with Simpson's quadrature for the cell interface integrals. Analogously to the previous section, it is possible to include all eigenvalues into the unit disc.

8. Approximate evolution operator E_{Δ}^{bilin} for piecewise bilinear data. In this section we investigate the stability of the second order finite volume schemes proposed by Lukáčová-Medvid'ová, Morton, and Warnecke in [11]. These schemes are based on the approximate evolution operator E_{Δ}^{bilin}, which is given as follows:

$$\phi_P = \left(1-\frac{\pi}{2}\right)\phi'_P + \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{\pi}{2}\phi_Q - 2u_Q\cos\theta - 2v_Q\sin\theta\right)\mathrm{d}\theta + O(\Delta t^2), \tag{8.1}$$

$$u_P = \left(1-\frac{\pi}{4}\right)u'_P + \frac{1}{2\pi}\int_0^{2\pi}\Big(-2\cos\theta\phi_Q + \frac{\pi}{2}u_Q\left(3\cos^2\theta-1\right) + \frac{3\pi}{2}v_Q\sin\theta\cos\theta\Big)\mathrm{d}\theta + O(\Delta t^2), \tag{8.2}$$

$$v_P = \left(1-\frac{\pi}{4}\right)v'_P + \frac{1}{2\pi}\int_0^{2\pi}\Big(-2\sin\theta\phi_Q + \frac{3\pi}{2}u_Q\sin\theta\cos\theta + \frac{\pi}{2}v_Q\left(3\sin^2\theta-1\right)\Big)\mathrm{d}\theta + O(\Delta t^2). \tag{8.3}$$

Analogously to E_{Δ}^{const}, this approximate evolution operator is designed so that it computes any one-dimensional linear plane wave propagating in the x- or y-direction exactly; for more details see [11]. In order to obtain second order finite volume schemes

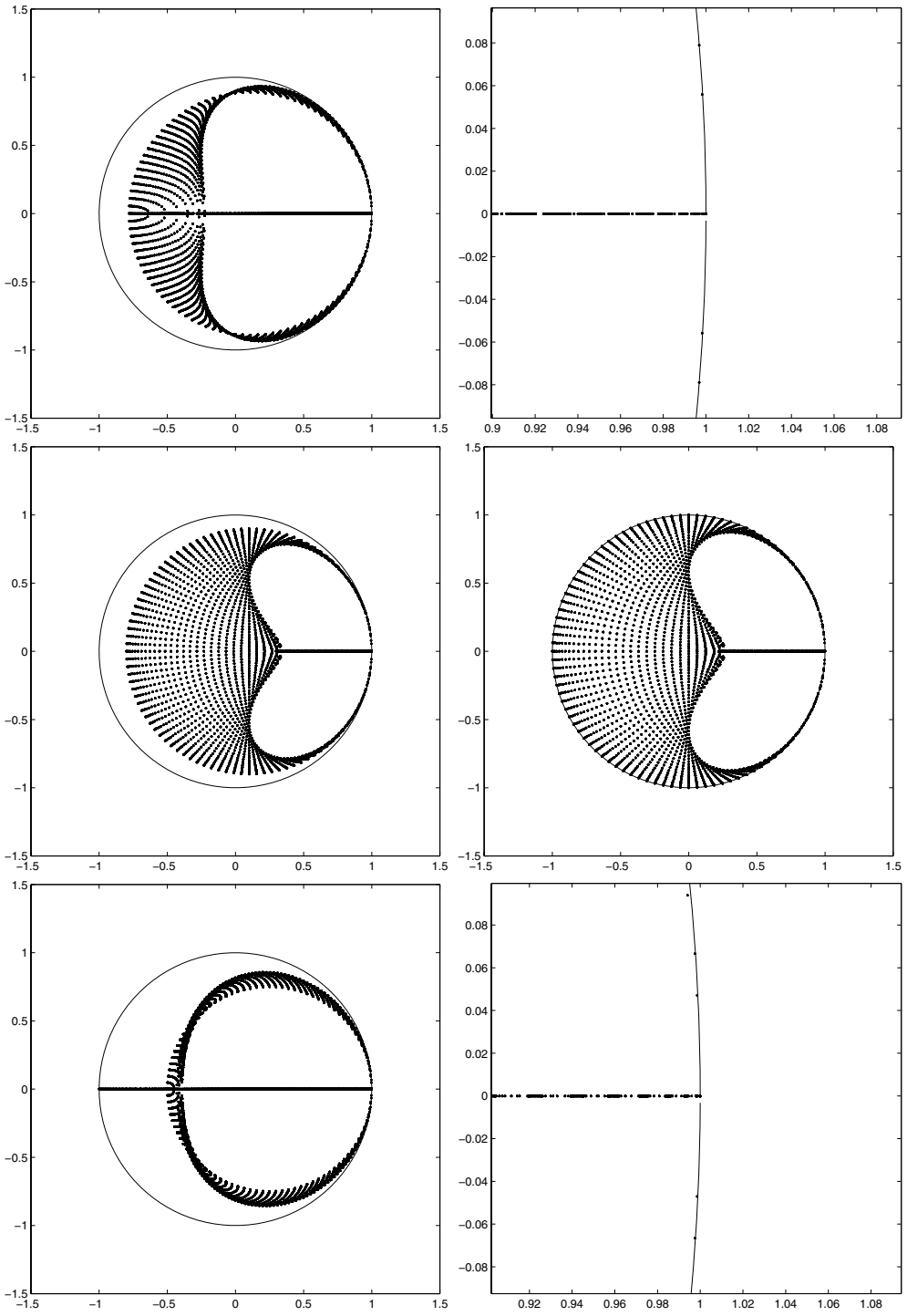


FIG. 4. *Eigenvalues of the amplification matrices; top: exact interface integration for the* $CFL = 0.89$ (*right zoom*)*; middle: interface integrals approximated using the trapezoidal rule for the* $CFL = 0.9, 1.0$.*; bottom: interface integrals approximated by Simpson's rule for the* $CFL = 0.75$ (*right zoom*).

we carry out a recovery stage before applying the approximate evolution operator; see Definition 4.2. The following two types of bilinear recoveries have been considered in [11]:

$$R_h^C\mathbf{U}\big|_{\Omega_{kl}} = \Big(\mu_x^2\mu_y^2 + \frac{x-x_k}{h}\mu_x\mu_y^2\delta_x + \frac{y-y_l}{h}\mu_x^2\mu_y\delta_y + \frac{(x-x_k)(y-y_l)}{h^2}\mu_x\mu_y\delta_x\delta_y\Big)\mathbf{U}_{kl}, \tag{8.4}$$

$$R_h^D\mathbf{U}\big|_{\Omega_{kl}} = \Big(1 + \frac{x-x_k}{h}\mu_x\mu_y^2\delta_x + \frac{y-y_l}{h}\mu_x^2\mu_y\delta_y + \frac{(x-x_k)(y-y_l)}{h^2}\mu_x\mu_y\delta_x\delta_y\Big)\mathbf{U}_{kl}. \tag{8.5}$$

Note that the recovery (8.4) is continuous, while the recovery (8.5) is discontinuous and conservative. We use the midpoint rule to approximate the time integral in (4.5). Denoting the cell interface intermediate value that is computed in the predictor step (4.6) by $\tilde{\mathbf{U}}^{n+\frac{1}{2}}$, we obtain the following schemes:

$$\begin{aligned}
&\text{scheme A} && \tilde{\mathbf{U}}^{n+\frac{1}{2}} = E_\Delta^{bilin}R_h^C\mathbf{U}^n + E_\Delta^{const}(1-\mu_x^2\mu_y^2)\mathbf{U}^n,\\
&\text{scheme B} && \tilde{\mathbf{U}}^{n+\frac{1}{2}} = E_\Delta^{bilin}R_h^C\mathbf{U}^n,\\
&\text{scheme C} && \tilde{\mathbf{U}}^{n+\frac{1}{2}} = E_\Delta^{bilin}R_h^D\mathbf{U}^n.
\end{aligned}$$

Each of these schemes has two further types according to the evaluation of the cell interface integrals. We used the subscripts $1, 2$ to distinguish between them. Thus, 1 corresponds to Simpson's rule and 2 to the trapezoidal rule. For example, for the scheme C_2 the predicted values along the right cell interface are

$$\begin{aligned}
\tilde{\Phi}^{n+\frac{1}{2}} &= \left[1+\left(\frac{-\pi}{32}+\frac{\nu}{16}\right)\delta_x^2\mu_y^2+\left(\frac{-\pi}{32}+\frac{\nu}{16}\right)\delta_y^2\mu_x^2+\left(\frac{\pi}{32}-\frac{\nu}{8}+\frac{\nu^2}{32}\right)\delta_x^2\delta_y^2\right]\mu_x\mu_y\Phi^n\\
&+\left[\frac{-2}{\pi}+\left(\frac{1}{2\pi}-\frac{\nu}{8}\right)\mu_x^2\mu_y^2+\left(\frac{1}{8\pi}-\frac{\nu}{16\pi}\right)\delta_y^2\mu_x^2+\left(\frac{-1}{2\pi}+\frac{\nu}{8}+\frac{\nu}{4\pi}-\frac{\nu^2}{6\pi}\right)\mu_x^2\delta_y^2\right]\delta_x\mu_yU^n\\
&+\left[\frac{-2}{\pi}+\left(\frac{1}{8\pi}-\frac{\nu}{16\pi}\right)\delta_x^2\mu_y^2+\left(\frac{1}{2\pi}-\frac{\nu}{8}\right)\mu_x^2\mu_y^2+\left(\frac{-1}{2\pi}+\frac{\nu}{8}+\frac{\nu}{4\pi}-\frac{\nu^2}{6\pi}\right)\mu_y^2\delta_x^2\right]\delta_y\mu_xV^n,
\end{aligned}$$

$$\begin{aligned}
\tilde{U}^{n+\frac{1}{2}} &= \left[\frac{-2}{\pi}+\left(\frac{1}{2\pi}-\frac{\nu}{8}\right)\mu_x^2\mu_y^2+\left(\frac{1}{8\pi}-\frac{\nu}{16\pi}\right)\delta_y^2\mu_x^2\right.\\
&\qquad\left.+\left(\frac{-1}{2\pi}+\frac{\nu}{8}+\frac{\nu}{4\pi}-\frac{\nu^2}{6\pi}\right)\mu_x^2\delta_y^2\right]\delta_x\mu_y\Phi^n\\
&+\left[1+\left(\frac{-\pi}{64}+\frac{\nu}{16}\right)\delta_x^2\mu_y^2-\frac{\pi}{64}\delta_y^2\mu_x^2+\left(\frac{\pi}{64}-\frac{\nu}{16}+\frac{\nu^2}{64}\right)\delta_x^2\delta_y^2\right]\mu_x\mu_yU^n\\
&+\left[\frac{1}{8}+\left(\frac{1}{16}-\frac{\nu}{8}+\frac{\pi\nu^2}{64}\right)\mu_x^2\mu_y^2\right]3\delta_x\delta_yV^n,
\end{aligned}$$

with the equation for $\tilde{V}^{n+\frac{1}{2}}$ that is analogous to that for $\tilde{U}^{n+\frac{1}{2}}$. Further, we can express analogously the predicted values for the other cell interfaces as well as for other schemes. Substituting the predicted values in the corrector step (4.5) yields, for

all second order finite volume schemes FVEG-A, B, C,

(8.6)
$$\mathbf{U}_{kl}^{n+1} = \mathbf{U}_{kl}^{n} + \sum_{r=-1}^{1}\sum_{s=-1}^{1} \mathcal{C}_{rs}\mathbf{U}_{k+rl+s}^{n} + \mathcal{C}_{rs}^{x}\mathbf{U}_{x_{k+rl+s}}^{n} + \mathcal{C}_{rs}^{y}\mathbf{U}_{y_{k+rl+s}}^{n} + \mathcal{C}_{rs}^{xy}\mathbf{U}_{xy_{k+rl+s}}^{n},$$

where $\mathcal{C}_{rs}^{x}$, $\mathcal{C}_{rs}^{y}$, and $\mathcal{C}_{rs}^{xy}$ are the coefficient matrices corresponding to the approximation of x-, y-, and xy-slopes. Moreover,

$$\mathbf{U}_{x_{k+rl+s}}^{n} = \mu_x\mu_y^2\delta_x\mathbf{U}_{k+rl+s}^{n}, \qquad \mathbf{U}_{y_{k+rl+s}}^{n} = \mu_x^2\mu_y\delta_y\mathbf{U}_{k+rl+s}^{n},$$
$$\mathbf{U}_{xy_{k+rl+s}}^{n} = \mu_x\mu_y\delta_x\delta_y\mathbf{U}_{k+rl+s}^{n}.$$

Applying a von Neumann analysis and the Fourier transforms we obtain the amplification matrices $\mathcal{T}$. It should be pointed out that their structure is too complicated to apply estimates of the spectral radius similar to those in section 6 for the first order EG3 scheme. Anyway, we can use the standard MATLAB procedure to determine the eigenvalues of $\mathcal{T}$. The corresponding stability limits for the FVEG schemes are given in Table 3.

TABLE 3
Stability limits of the second order FVEG schemes.

	Trapezoidal rule	Simpson's rule
Scheme A	0.94	0.75
Scheme B	0.78	-
Scheme C	0.78	0.58

We should note that all CFL limits given in Table 3 have also been confirmed by various numerical experiments. In Figures 5 and 6 we plot, using different scales, the eigenvalues of the amplification matrices corresponding to the second order FVEG schemes: scheme A_i, B_i, and C_i, where $i = 1, 2$. Similarly to the previous cases, these plots indicate that all eigenvalues are included in the unit disc. Note that different quadrature rules have a considerable effect on the form of the spectrum of the resulting amplification matrix. For example, Simpson's quadrature rule for cell interface integrals yields the spectrum which is more compact. Further, it follows from Figure 5 that the second order FVEG scheme based on the operator (8.1)–(8.3) with the continuous nonconservative recovery (8.4) with Simpson's rule, i.e., scheme B_1, is unconditionally unstable. This fact has also been confirmed by other numerical tests for the wave equation system with discontinuous solution. We have found for all CFL numbers, no matter how small they were chosen, instabilities in the solution for fine enough meshes. We would like to remark that the loss of stability of EG schemes under numerical integration has been also observed even for the scalar linear hyperbolic equation by Morton, Priestly, and Süli [21].

Remarks. The derivation of the EG and FVEG schemes might be considered at first sight a rather complex task. In fact, the finite difference formulation (4.7), (4.8) which uses the stencil matrices (cf. appendix) is used only for theoretical analysis. In practice we implement the approximate evolution operators (3.7)–(3.9), (7.1)–(7.3), or (8.1)–(8.3) directly. Thus, for example, in the FVEG scheme the flux integrals along cell interfaces are approximated by the Simpson or the trapezoidal rule and the only complexity lies in the implementation of the exact integrals of the type

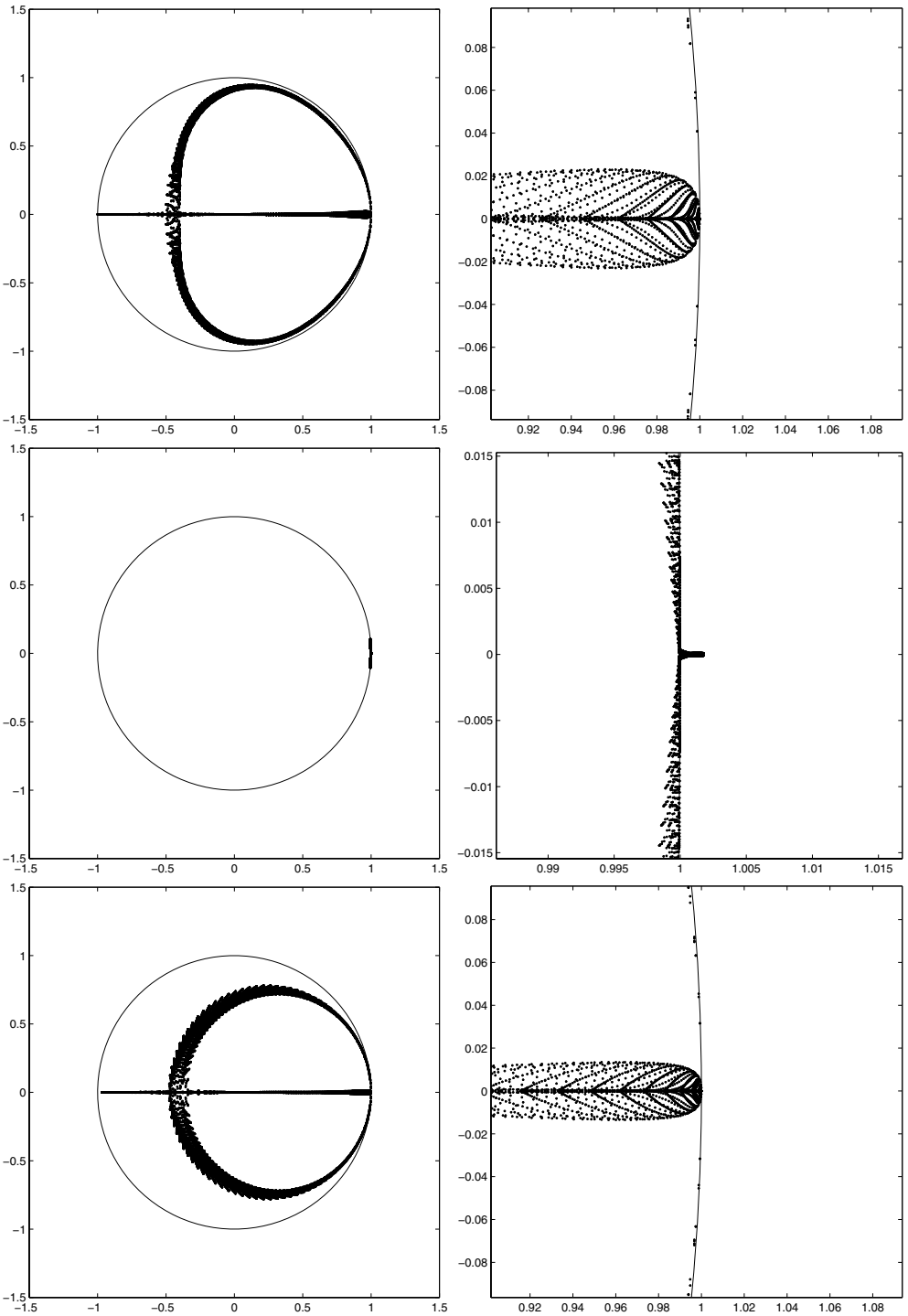


FIG. 5. *Eigenvalues corresponding to the amplification matrices of scheme* A_1 ($CFL = 0.75$), *scheme* B_1 ($CFL = 0.1$), *and scheme* C_1 ($CFL = 0.58$).

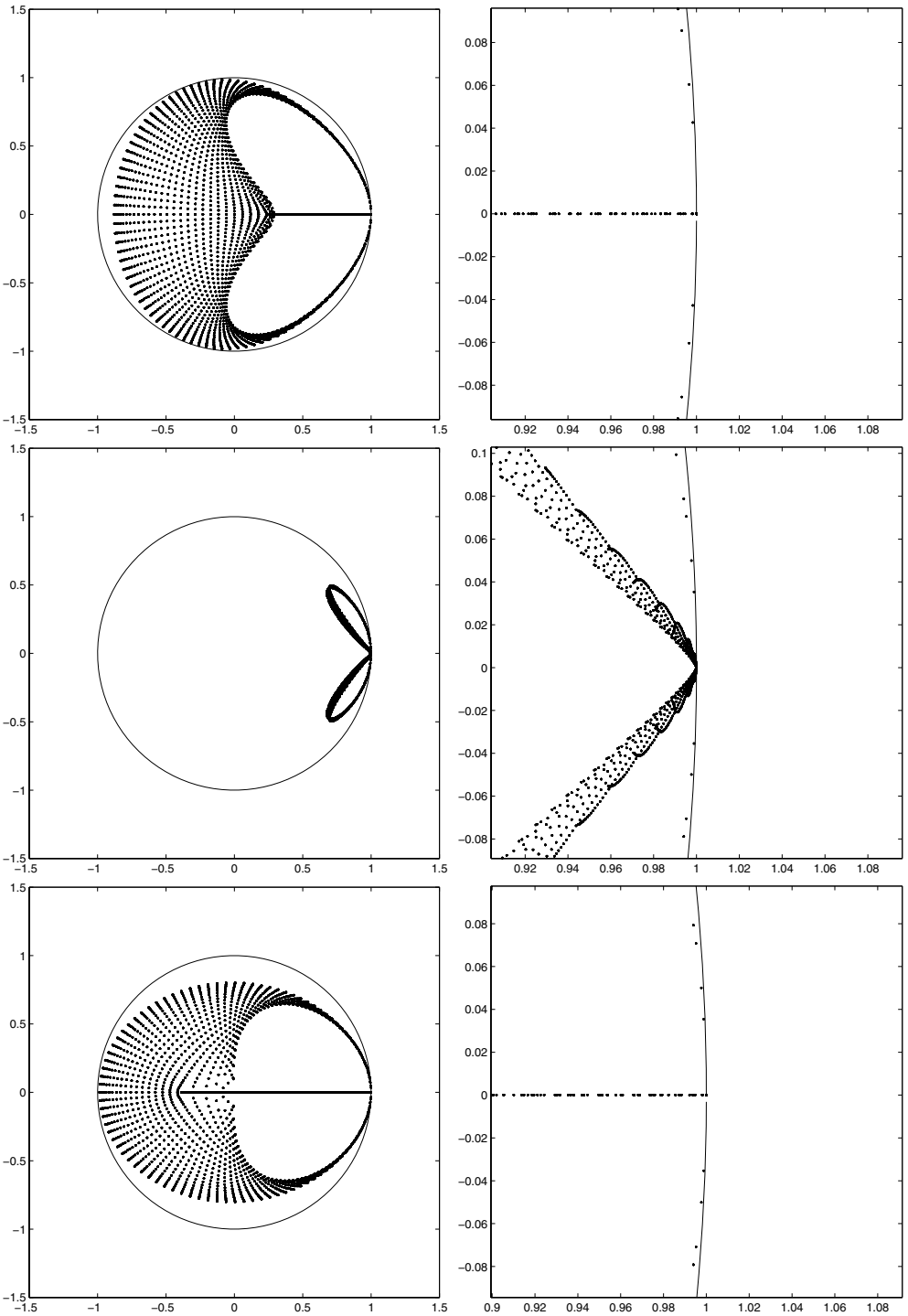


FIG. 6. *Eigenvalues corresponding to the amplification matrices of scheme* A_2 ($CFL = 0.94$), *scheme* B_2 ($CFL = 0.78$), *and scheme* C_2 ($CFL = 0.78$).

$\int_\alpha^\beta \cos^n\theta \sin^m\theta \mathrm{d}\theta$, where $n, m \geq 0$ integers, and $\alpha, \beta \in [0, 2\pi]$ are corresponding angles according to a position of the (slanted) Mach cone. Alternatively, the integrals along the sonic circle, i.e., for θ from 0 to 2π, can be approximated by a suitable numerical quadrature, which further simplifies the implementation of the FVEG schemes.

Based on our knowledge of the EG and FVEG schemes for regular rectangular two-dimensional meshes, some further generalizations have been done. In particular, the FVEG schemes with the approximate evolution operators (3.7)–(3.9), (7.1)–(7.3), and (8.1)–(8.3) have been generalized to three-dimensional problems using regular cubic meshes; see [19]. Further, in [26], [27] the FVEG schemes have been generalized for unstructured triangular meshes. Of course, the stability analysis of the EG schemes on general unstructured meshes is much more involved. One possible way would be to apply the energy analysis in the $L^2(\mathbb{R}^2)$-norm, i.e., to show at least the weak L^2-stability $\|P_h E_{\Delta t}\|_{L^2(\mathbb{R}^2)\to L^2(\mathbb{R}^2)} \leq 1 + C\Delta t$. This is a nontrivial problem, which can be considered in the future.

The absorbing boundary conditions have been studied for the EG schemes extensively in [18]. We have considered simple extrapolation boundary conditions, the characteristic boundary conditions, as well as a perfectly matched layer approach. Numerical experiments reported in [18] indicate that the best results are obtained by combining the EG and FVEG schemes with the perfectly matched layer technique. We have observed no influence of these boundary conditions on stability of the EG schemes. It would be interesting to investigate this question theoretically in the future.

Conclusion. In this paper we have studied the stability of various EG schemes by a von Neumann analysis. The schemes are applied to the linear wave equation system. First, we have discussed theoretical stability estimates of the EG3 schemes, the most favorable one among the finite difference EG schemes considered. Due to the complex structure of the amplification matrix we were able to give theoretical stability estimates only for a simplified problem. Further, we analyzed experimentally the spectral radius of amplification matrix of the EG3 scheme. The experimental analysis indicates that the scheme is stable up to the CFL number 0.58. The stability of the FVEG schemes was studied experimentally, too. It has been shown that new quadratures in time for time integrals in the exact evolution operator, which were proposed in our recent paper [11], improve the stability limits considerably. For example, if the trapezoidal rule is used for the cell interface integrals, the CFL number is 1 and 0.94 for the first and the second order FVEG schemes, respectively. On the other hand, Simpson's quadrature rule reduces the stability range slightly.

Appendix.

EG3 scheme. For the discrete form of the scheme, see (4.4), (4.7), (4.8), and (3.7)–(3.9) for the approximate evolution operator EG3.

$$\alpha^1 := \begin{Bmatrix} \frac{\nu^2}{4\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \\ \frac{\nu}{\pi} - \frac{\nu^2}{2\pi} & -\frac{4\nu}{\pi} + \frac{\nu^2}{\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{2\pi} \\ \frac{\nu^2}{4\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \end{Bmatrix}, \quad \beta^1 := \begin{Bmatrix} \frac{\nu^2}{3\pi} & 0 & \frac{-\nu^2}{3\pi} \\ \frac{\nu}{2} - \frac{2\nu^2}{3\pi} & 0 & -\frac{\nu}{2} + \frac{2\nu^2}{3\pi} \\ \frac{\nu^2}{3\pi} & 0 & -\frac{\nu^2}{3\pi} \end{Bmatrix},$$

$$\gamma^1 := \begin{Bmatrix} -\frac{\nu^2}{3\pi} & -\frac{\nu}{2} + \frac{2\nu^2}{3\pi} & -\frac{\nu^2}{3\pi} \\ 0 & 0 & 0 \\ \frac{\nu^2}{3\pi} & \frac{\nu}{2} - \frac{2\nu^2}{3\pi} & \frac{\nu^2}{3\pi} \end{Bmatrix}, \quad \alpha^2 := \begin{Bmatrix} \frac{\nu^2}{3\pi} & 0 & -\frac{\nu^2}{3\pi} \\ \frac{\nu}{2} - \frac{2\nu^2}{3\pi} & 0 & -\frac{\nu}{2} + \frac{2\nu^2}{3\pi} \\ \frac{\nu^2}{3\pi} & 0 & \frac{-\nu^2}{3\pi} \end{Bmatrix},$$

$$\beta^2 := \left\{ \begin{array}{ccc} \frac{\nu^2}{8\pi} & -\frac{\nu^2}{4\pi} & \frac{\nu^2}{8\pi} \\ \frac{\nu}{\pi} - \frac{\nu^2}{4\pi} & -\frac{2\nu}{\pi} + \frac{\nu^2}{2\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{4\pi} \\ \frac{\nu^2}{8\pi} & -\frac{\nu^2}{4\pi} & \frac{\nu^2}{8\pi} \end{array} \right\}, \ \gamma^2 := \left\{ \begin{array}{ccc} -\frac{3\nu^2}{32} & 0 & \frac{3\nu^2}{32} \\ 0 & 0 & 0 \\ \frac{3\nu^2}{32} & 0 & -\frac{3\nu^2}{32} \end{array} \right\},$$

$$\alpha^3 := \left\{ \begin{array}{ccc} -\frac{\nu^2}{3\pi} & -\frac{\nu}{2} + \frac{2\nu^2}{3\pi} & -\frac{\nu^2}{3\pi} \\ 0 & 0 & 0 \\ \frac{\nu^2}{3\pi} & +\frac{\nu}{2} - \frac{2\nu^2}{3\pi} & \frac{\nu^2}{3\pi} \end{array} \right\}, \ \beta^3 := \left\{ \begin{array}{ccc} -\frac{3\nu^2}{32} & 0 & \frac{3\nu^2}{32} \\ 0 & 0 & 0 \\ \frac{3\nu^2}{32} & 0 & -\frac{3\nu^2}{32} \end{array} \right\},$$

$$\gamma^3 := \left\{ \begin{array}{ccc} \frac{\nu^2}{8\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{4\pi} & \frac{\nu^2}{8\pi} \\ -\frac{\nu^2}{4\pi} & -\frac{2\nu}{\pi} + \frac{\nu^2}{2\pi} & -\frac{\nu^2}{4\pi} \\ \frac{\nu^2}{8\pi} & \frac{\nu}{\pi} - \frac{\nu^2}{4\pi} & \frac{\nu^2}{8\pi} \end{array} \right\}.$$

FVEG scheme with E_{Δ}^{const} operator using exact cell interface integration. For the discrete form of the scheme, see (4.5)–(4.8), and see (7.1)–(7.3) for the approximate evolution operator E_{Δ}^{const}.

$$\alpha^1 := \left\{ \begin{array}{ccc} \frac{\nu^2}{2\pi} & \frac{\nu}{2} - \frac{\nu^2}{\pi} & \frac{\nu^2}{2\pi} \\ \frac{\nu}{2} - \frac{\nu^2}{\pi} & -2\nu + \frac{2\nu^2}{\pi} & \frac{\nu}{2} - \frac{\nu^2}{\pi} \\ \frac{\nu^2}{2\pi} & \frac{\nu}{2} - \frac{\nu^2}{\pi} & \frac{\nu^2}{2\pi} \end{array} \right\}, \quad \beta^1 := \left\{ \begin{array}{ccc} \frac{7\nu^2}{24\pi} & 0 & \frac{-7\nu^2}{24\pi} \\ \frac{\nu}{2} - \frac{7\nu^2}{12\pi} & 0 & \frac{-\nu}{2} + \frac{7\nu^2}{12\pi} \\ \frac{7\nu^2}{24\pi} & 0 & \frac{-7\nu^2}{24\pi} \end{array} \right\},$$

$$\gamma^1 := \left\{ \begin{array}{ccc} \frac{-7\nu^2}{24\pi} & \frac{-\nu}{2} + \frac{7\nu^2}{12\pi} & \frac{-7\nu^2}{24\pi} \\ 0 & 0 & 0 \\ \frac{7\nu^2}{24\pi} & \frac{\nu}{2} - \frac{7\nu^2}{12\pi} & \frac{7\nu^2}{24\pi} \end{array} \right\}, \quad \alpha^2 := \left\{ \begin{array}{ccc} \frac{\nu^2}{4\pi} & 0 & \frac{-\nu^2}{4\pi} \\ \frac{\nu}{2} - \frac{\nu^2}{2\pi} & 0 & \frac{-\nu}{2} + \frac{\nu^2}{2\pi} \\ \frac{\nu^2}{4\pi} & 0 & \frac{-\nu^2}{4\pi} \end{array} \right\},$$

$$\beta^2 := \left\{ \begin{array}{ccc} \frac{\nu^2}{4\pi} & -\frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \\ \frac{\nu}{2} - \frac{\nu^2}{2\pi} & -\nu + \frac{\nu^2}{\pi} & \frac{\nu}{2} - \frac{\nu^2}{2\pi} \\ \frac{\nu^2}{4\pi} & -\frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \end{array} \right\}, \ \gamma^2 := \left\{ \begin{array}{ccc} \frac{-\nu^2}{4\pi} & 0 & \frac{\nu^2}{4\pi} \\ 0 & 0 & 0 \\ \frac{\nu^2}{4\pi} & 0 & \frac{-\nu^2}{4\pi} \end{array} \right\},$$

$$\alpha^3 := \left\{ \begin{array}{ccc} \frac{-\nu^2}{4\pi} & \frac{-\nu}{2} + \frac{\nu^2}{2\pi} & \frac{-\nu^2}{4\pi} \\ 0 & 0 & 0 \\ \frac{\nu^2}{4\pi} & \frac{\nu}{2} - \frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \end{array} \right\}, \quad \beta^3 := \left\{ \begin{array}{ccc} \frac{-\nu^2}{4\pi} & 0 & \frac{\nu^2}{4\pi} \\ 0 & 0 & 0 \\ \frac{\nu^2}{4\pi} & 0 & \frac{-\nu^2}{4\pi} \end{array} \right\},$$

$$\gamma^3 := \left\{ \begin{array}{ccc} \frac{\nu^2}{4\pi} & \frac{\nu}{2} - \frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \\ -\frac{\nu^2}{2\pi} & -\nu + \frac{\nu^2}{\pi} & -\frac{\nu^2}{2\pi} \\ \frac{\nu^2}{4\pi} & \frac{\nu}{2} - \frac{\nu^2}{2\pi} & \frac{\nu^2}{4\pi} \end{array} \right\}.$$

FVEG with E_Δ^{const} operator using Simpson's quadrature for cell interface integration. For the discrete form of the scheme, see (4.5), (4.6) with Simpson's quadrature, (4.7), (4.8), and (7.1)–(7.3) for the approximate evolution operator E_Δ^{const}.

$$\alpha^1 := \begin{Bmatrix} \frac{\nu}{12} & \frac{\nu}{3} & \frac{\nu}{12} \\ \frac{\nu}{3} & -\frac{5\nu}{3} & \frac{\nu}{3} \\ \frac{\nu}{12} & \frac{\nu}{3} & \frac{\nu}{12} \end{Bmatrix}, \quad \beta^1 := \begin{Bmatrix} \frac{\nu}{24}\left(1+\frac{1}{\pi}\right) & 0 & -\frac{\nu}{24}\left(1+\frac{1}{\pi}\right) \\ \frac{\nu}{24}\left(10-\frac{2}{\pi}\right) & 0 & -\frac{\nu}{24}\left(10-\frac{2}{\pi}\right) \\ \frac{\nu}{24}\left(1+\frac{1}{\pi}\right) & 0 & -\frac{\nu}{24}\left(1+\frac{1}{\pi}\right) \end{Bmatrix},$$

$$\gamma^1 := \begin{Bmatrix} -\frac{\nu}{24}\left(1+\frac{1}{\pi}\right) & -\frac{\nu}{24}\left(10-\frac{2}{\pi}\right) & -\frac{\nu}{24}\left(1+\frac{1}{\pi}\right) \\ 0 & 0 & 0 \\ \frac{\nu}{24}\left(1+\frac{1}{\pi}\right) & \frac{\nu}{24}\left(10-\frac{2}{\pi}\right) & \frac{\nu}{24}\left(1+\frac{1}{\pi}\right) \end{Bmatrix}, \quad \alpha^2 := \begin{Bmatrix} \frac{\nu}{24} & 0 & -\frac{\nu}{24} \\ \frac{10\nu}{24} & 0 & -\frac{10\nu}{24} \\ \frac{\nu}{24} & 0 & -\frac{\nu}{24} \end{Bmatrix},$$

$$\beta^2 := \begin{Bmatrix} \frac{\nu}{24} & -\frac{2\nu}{24} & \frac{\nu}{24} \\ \frac{10\nu}{24} & -\frac{20\nu}{24} & \frac{10\nu}{24} \\ \frac{\nu}{24} & -\frac{2\nu}{24} & \frac{\nu}{24} \end{Bmatrix}, \quad \gamma^2 := \begin{Bmatrix} -\frac{\nu}{24} & 0 & \frac{\nu}{24} \\ 0 & 0 & 0 \\ \frac{\nu}{24} & 0 & -\frac{\nu}{24} \end{Bmatrix},$$

$$\alpha^3 := \begin{Bmatrix} -\frac{\nu}{24} & -\frac{10\nu}{24} & -\frac{\nu}{24} \\ 0 & 0 & 0 \\ \frac{\nu}{24} & \frac{10\nu}{24} & \frac{\nu}{24} \end{Bmatrix}, \quad \beta^3 := \begin{Bmatrix} -\frac{\nu}{24} & 0 & \frac{\nu}{24} \\ 0 & 0 & 0 \\ \frac{\nu}{24} & 0 & -\frac{\nu}{24} \end{Bmatrix},$$

$$\gamma^3 := \begin{Bmatrix} \frac{\nu}{24} & \frac{10\nu}{24} & \frac{\nu}{24} \\ -\frac{2\nu}{24} & -\frac{20\nu}{24} & -\frac{2\nu}{24} \\ \frac{\nu}{24} & \frac{10\nu}{24} & \frac{\nu}{24} \end{Bmatrix}.$$

FVEG with E_Δ^{const} operator using the trapezoidal quadrature for cell interface integration. For the discrete form of the scheme, see (4.5), (4.6) with the trapezoidal quadrature, (4.7), (4.8), and (7.1)–(7.3) for the approximate evolution operator E_Δ^{const}.

$$\alpha^1 := \begin{Bmatrix} \frac{\nu}{4} & 0 & \frac{\nu}{4} \\ 0 & -\nu & 0 \\ \frac{\nu}{4} & 0 & \frac{\nu}{4} \end{Bmatrix}, \quad \beta^1 := \begin{Bmatrix} \frac{\nu}{8}\left(1+\frac{1}{\pi}\right) & 0 & -\frac{\nu}{8}\left(1+\frac{1}{\pi}\right) \\ \frac{\nu}{8}\left(2-\frac{2}{\pi}\right) & 0 & -\frac{\nu}{8}\left(2-\frac{2}{\pi}\right) \\ \frac{\nu}{8}\left(1+\frac{1}{\pi}\right) & 0 & -\frac{\nu}{8}\left(1+\frac{1}{\pi}\right) \end{Bmatrix},$$

$$\gamma^1 := \begin{Bmatrix} -\frac{\nu}{8}\left(1+\frac{1}{\pi}\right) & -\frac{\nu}{8}\left(2-\frac{2}{\pi}\right) & -\frac{\nu}{8}\left(1+\frac{1}{\pi}\right) \\ 0 & 0 & 0 \\ \frac{\nu}{8}\left(1+\frac{1}{\pi}\right) & \frac{\nu}{8}\left(2-\frac{2}{\pi}\right) & \frac{\nu}{8}\left(1+\frac{1}{\pi}\right) \end{Bmatrix}, \quad \alpha^2 := \begin{Bmatrix} \frac{\nu}{8} & 0 & -\frac{\nu}{8} \\ \frac{2\nu}{8} & 0 & -\frac{2\nu}{8} \\ \frac{\nu}{8} & 0 & -\frac{\nu}{8} \end{Bmatrix},$$

$$\beta^2 := \begin{Bmatrix} \frac{\nu}{8} & -\frac{2\nu}{8} & \frac{\nu}{8} \\ \frac{2\nu}{8} & -\frac{4\nu}{8} & \frac{2\nu}{8} \\ \frac{\nu}{8} & -\frac{2\nu}{8} & \frac{\nu}{8} \end{Bmatrix}, \quad \gamma^2 := \begin{Bmatrix} -\frac{\nu}{8} & 0 & \frac{\nu}{8} \\ 0 & 0 & 0 \\ \frac{\nu}{8} & 0 & -\frac{\nu}{8} \end{Bmatrix},$$

$$\alpha^3 := \begin{Bmatrix} -\frac{\nu}{8} & -\frac{2\nu}{8} & -\frac{\nu}{8} \\ 0 & 0 & 0 \\ \frac{\nu}{8} & \frac{2\nu}{8} & \frac{\nu}{8} \end{Bmatrix}, \quad \beta^3 := \begin{Bmatrix} -\frac{\nu}{8} & 0 & \frac{\nu}{8} \\ 0 & 0 & 0 \\ \frac{\nu}{8} & 0 & -\frac{\nu}{8} \end{Bmatrix},$$

$$\gamma^3 := \begin{Bmatrix} \frac{\nu}{8} & \frac{2\nu}{8} & \frac{\nu}{8} \\ -\frac{2\nu}{8} & -\frac{4\nu}{8} & -\frac{2\nu}{8} \\ \frac{\nu}{8} & \frac{2\nu}{8} & \frac{\nu}{8} \end{Bmatrix}.$$

REFERENCES

[1] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience, Wiley, New York, 1948; reprinted by Springer-Verlag, New York, 1985.

[2] S. Depeyre, *A stability analysis for finite volume schemes applied to the Maxwell system*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 443–458.

[3] M. Fey, *Multidimensional upwinding, Part* I. *The method of transport for solving the Euler equations*, J. Comput. Phys., 143 (1998), pp. 159–180.

[4] M. Fey, *Multidimensional upwinding, Part* II. *Decomposition of the Euler equations into advection equations*, J. Comput. Phys., 143 (1998), pp. 181–199.

[5] G. Hämmerlin and K.-H. Hoffmann, *Numerical Mathematics*, Undergrad. Texts Math., Readings in Mathematics, Springer-Verlag, New York, 1991.

[6] T. Kröger and M. Lukáčová-Medviďová, *An evolution Galerkin scheme for the shallow water magnetohydrodynamic equations in two space dimensions*, J. Comput. Phys., 206 (2005), pp. 122–149.

[7] R. J. LeVeque, *Wave propagation algorithms for multidimensional hyperbolic systems*, J. Comput. Phys., 131 (1997), pp. 327–353.

[8] P. Lin, K. W. Morton, and E. Süli, *Euler characteristic Galerkin scheme with recovery*, M2AN Math. Model. Numer. Anal., 27 (1993), pp. 863–894.

[9] P. Lin, K. W. Morton, and E. Süli, *Characteristic Galerkin schemes for scalar conservation laws in two and three space dimensions*, SIAM J. Numer. Anal., 34 (1997), pp. 779–796.

[10] M. Lukáčová-Medviďová, K. W. Morton, and G. Warnecke, *Finite volume evolution Galerkin methods for Euler equations of gas dynamics*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 425–434.

[11] M. Lukáčová-Medviďová, K. W. Morton, and G. Warnecke, *Finite volume evolution Galerkin methods for hyperbolic systems*, SIAM J. Sci. Comput., 26 (2004), pp. 1–30.

[12] M. Lukáčová-Medviďová, K. W. Morton, and G. Warnecke, *High resolution finite volume evolution Galerkin schemes for multidimensional conservation laws*, in Numerical Mathematics and Advanced Applications, Proceedings of the 3rd European Conference (ENUMATH'99), P. Neittaanmäki, T. Tiihonen, and P. Tarvainen, eds., World Scientific, Singapore, 2000, pp. 633–640.

[13] M. Lukáčová-Medviďová, K. W. Morton, and G. Warnecke, *Evolution Galerkin methods for hyperbolic systems in two space dimensions*, Math. Comput., 69 (2000), pp. 1355–1384.

[14] M. Lukáčová-Medviďová, K. W. Morton, and G. Warnecke, *Evolution Galerkin methods for multidimensional hyperbolic systems*, in Proceedings of ECCOMAS 2000, Barcelona, 2000, pp. 1–14.

[15] M. Lukáčová-Medviďová, J. Saibertová, and G. Warnecke, *Finite volume evolution Galerkin methods for nonlinear hyperbolic systems*, J. Comput. Phys., 183 (2002), pp. 533–562.

[16] M. Lukáčová-Medviďová, J. Saibertová, G. Warnecke, and Y. Zahaykah, *On evolution Galerkin methods for the Maxwell and the linearized Euler equations*, Appl. Math., 49 (2004), pp. 415–439.

[17] M. Lukáčová-Medviďová, G. Warnecke, and Y. Zahaykah, *Third order finite volume evolution Galerkin (FVEG) methods for two-dimensional wave equation system*, J. Numer. Math., 11 (2003), pp. 235-251.

[18] M. Lukáčová-Medviďová, G. Warnecke, and Y. Zahaykah, *On the boundary conditions for EG-methods applied to the two-dimensional wave equation systems*, ZAMM Z. Angew. Math. Mech., 84 (2004), pp. 237–251.

[19] M. Lukáčová-Medviďová, G. Warnecke, and Y. Zahaykah, *Finite volume evolution Galerkin (FVEG) methods for three-dimensional wave equation system*, submitted.

[20] M. Lukáčová-Medviďová, S. Noelle, and M. Kraft, *Well-balanced finite volume evolution Galerkin methods for the shallow water equations*, J. Comput. Phys., to appear.

[21] K. W. Morton, A. Priestly, and E. Süli, *Convergence of the Lagrange-Galerkin method with non-exact integration*, M2AN Math. Model. Numer. Anal., 22 (1988), pp. 625–653.

[22] S. Noelle, *The MOT-ICE: A new high-resolution wave-propagation algorithm for multidimensional systems of conservative laws based on Fey's method of transport*, J. Comput. Phys., 164 (2000), pp. 283–334.

[23] S. Ostkamp, *Multidimensional characteristic Galerkin schemes and evolution operators for hyperbolic systems*, Math. Methods Appl. Sci., 20 (1997), pp. 1111–1125.

[24] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, Interscience, John Wiley, New York, London, Sydney, 1967.

[25] P. Roe, *Discrete models for the numerical analysis of time-dependent multidimensional gas dynamics*, J. Comput. Phys., 63 (1986), pp. 458–476.

[26] Qurrat-ul-Ain, *Multidimensional Schemes for Hyperbolic Conservation Laws on Triangular Meshes*, dissertation, University of Magdeburg, Germany, 2005.

[27] Qurrat-ul-Ain, G. Warnecke, and Y. Zahaykah, *On the Finite Volume Evolution Galerkin (FVEG) Methods for Two-Dimensional First Order Hyperbolic Systems on Structured and Unstructured Triangular Meshes*, in preparation.

[28] Y. Zahaykah, *Evolution Galerkin Schemes and Discrete Boundary Conditions for Multidimensional First Order Systems*, dissertation, University of Magdeburg, Germany, 2002.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1584–1611

UNCONDITIONAL STABILITY OF CORRECTED EXPLICIT-IMPLICIT DOMAIN DECOMPOSITION ALGORITHMS FOR PARALLEL APPROXIMATION OF HEAT EQUATIONS*

HAN-SHENG SHI† AND HONG-LIN LIAO‡

Abstract. A class of corrected explicit-implicit domain decomposition (CEIDD) methods is investigated for the parallel approximation of linear heat equations. Explicit-implicit domain decomposition (EIDD) methods are computationally and communicationally efficient for each time step but always suffer from small time step size restrictions. By adding an interface correction step to Kuznetsov's EIDD, the one-dimensional CEIDD procedure achieves unconditional stability without discarding the time-stepwise efficiency of the EIDD method. In order to maintain the virtues of the CEIDD method and improve the flexibility in domain partitioning, for solving multidimensional problems, special zigzag-shaped interfaces are suggested in the CEIDD method. Based on noncrossover and crossover types of zigzag interfaces, the resulting CEIDD-ZI algorithms are studied for two strategies of subdomain partition. By the energy method, it shows that the proposed algorithms, including their degenerate cases—the corrected explicit hopscotch schemes—are convergent in the discrete H^1 seminorm and L^2 norm. Numerical experiments confirm the results in our analysis.

Key words. heat equations, nonoverlapping domain decomposition, zigzag-shaped interior boundary, explicit-implicit difference method, stability, convergence

AMS subject classifications. 65M06, 65M12, 68Y05, 65M55

DOI. 10.1137/040609215

1. Introduction.

We propose a class of corrected explicit-implicit domain decomposition (CEIDD) algorithms for the numerical solution of heat equations

$$u_t = \sum_{m=1}^{d} \frac{\partial}{\partial x_m}\left(a^m(x)\frac{\partial u}{\partial x_m}\right) + f(x,t), \quad (x,t) \in \Omega \times (0,T], \tag{1.1}$$

together with initial and Dirichlet boundary conditions

$$u(x,0) = u^0(x), \quad x \in \Omega, \tag{1.2}$$

$$u(x,t) = u_b(x,t), \quad (x,t) \in \partial\Omega \times (0,T], \tag{1.3}$$

where $\Omega = (0,1)^d$ with $d = 1$ or 2, and the spatial variable $x = (x_m)_{m=1}^d \in \mathbf{R}^d$. The functions $a^m(x)$, $f(x,t)$ are smooth and $a^m(x) \geq a_0$ for a positive constant a_0.

The spatial domain Ω is discretized uniformly with spacing $h = 1/N$ for each spatial variable, where N is an even integer, although different spacing for different spatial variables may be considered in a similar way. And selecting a time step τ and an integer J so that $J\tau = T$, the time interval $(0,T]$ is also discretized uniformly. As usual, the discrete space grid, its boundary, and the discrete time-spatial mesh are denoted by Ω_h, $\partial\Omega_h$, and Ω_h^τ, respectively.

*Received by the editors June 1, 2004; accepted for publication (in revised form) March 23, 2006; published electronically August 7, 2006.
http://www.siam.org/journals/sinum/44-4/60921.html

†Department of Applied Mathematics and Physics, Institute of Sciences, PLA University of Science and Technology, Nanjing, 211101, P.R. China (liaohl2003@yahoo.com.cn).

‡Department of Mathematics, Southeast University, Nanjing, 210096, P.R. China, and Department of Applied Mathematics and Physics, Institute of Sciences, PLA University of Science and Technology, Nanjing, 211101, P.R. China.

For parallel solutions of time-dependent PDE problems, an interesting approach is to use explicit-implicit alternating algorithms on nonoverlapping domain decomposition; see [4, 5, 6, 8] for related discussions. Kuznetsov [8] proposed a mixed scheme, where the stable implicit scheme is used inside each subdomain while the explicit Euler scheme is applied to obtain the interface solutions on the new time level. Once the interface values are available, the whole problem is completely decoupled and thus can be computed in parallel. Due to the stability constraint, the method is not stable unless time step size $\tau \leq h^2/2$. A similar hybrid scheme was studied by Dawson, Du, and Dupont [4], where instead of using the same spacing h as for the interior points where the fully implicit scheme is applied, a larger spacing h_D is used at interface points where the one-directional implicit predictor scheme is applied. Correspondingly, the stability constraint becomes $\tau \leq h_D^2/2$. These globally noniterative explicit-implicit domain decomposition (EIDD) algorithms are computationally and communicationally efficient for each time step when compared with Schwarz-type domain decomposition elliptic solvers [2, 3] incorporated into implicit temporal discretizations; however, they suffer from stability-related time step restrictions, while Schwarz methods could maintain the good stability of implicit temporal discretization schemes.

Recently, many investigators have turned to improve the stability of EIDD methods because the time-stepwisely efficient EIDD procedures will process great potential for large-scale parallel simulations on distributed memory computers if they are free from time step size restrictions. A penalized EIDD algorithm proposed by Black [1] achieves numerically verified unconditional stability by employing a stable Du Fort–Frankel-type scheme as the explicit predictor; however, it makes the algorithm inconsistent unless $\tau/h \to 0$. Thus consistency comes only after paying a price of restricting time step size $\tau = \mathcal{O}(h^2)$ to achieve a first order temporal accuracy, a restriction quantitatively similar to the restriction on the EIDD methods.

An alternative approach for improving the stability is to apply an implicit correction technique to EIDD algorithms. The idea of implicit correction is to replace the interface predictor value by a new solution on the interface boundaries computed by an implicit corrector scheme. In 1998, Qian and Zhu [9] investigated experimentally the implicit correction technique for Kuznetsov's EIDD method [8] for the one-dimensional (1-D) heat equation and showed the better stability by comparing its results with several other methods. As the generalizations of Qian's algorithm [9], the stabilized EIDD methods proposed by Zhuang and Sun [11] exhibit unconditional stability experimentally for a wide range of multidimensional parabolic problems, from heat equation to convection-diffusion to nondissipative convection-diffusion. However, (P1) both attempts failed to consider the mathematical proofs of the improved stability. Furthermore, for multidimensional problems, the methods in [11] (P2) suffer from a parallel time overload for the interface correction step because some elliptic solver should be used to obtain the interface solutions, and (P3) decrease the flexibility of domain partitioning due to the noncrossover assumption of interior boundaries.

To remedy the disadvantages (P1)–(P3), in this paper, a new class of CEIDD algorithms is reported. First, a generic parallel version of the CEIDD algorithm for computing the solution u_h^k is presented below (also see [11]).

The parallel CEIDD algorithm.

0. Determine the shape of man-made interior interfaces Γ_h and the strategy of domain partitioning. Then divide the entire domain Ω_h into $p = \prod_{m=1}^{d} p_m$ subdomains $\Omega_{h,1}, \ldots, \Omega_{h,p}$ with interface boundary between $\Omega_{h,i}$ and $\Omega_{h,j}$ denoted by Γ_{ij} $(1 \leq i < j < p)$. Assign subdomain $\Omega_{h,i}$ and interface Γ_{ij} to processor P_i. Go to step 1 with $k := 1$.

1. On each processor P_i, compute all of the interface predictors $\tilde{u}_h^k$ on Γ_{ij} in parallel using an explicit scheme and then pass $\tilde{u}_h^k$ from P_i to P_j for the subdomain operation. Go to step 2.
2. Compute solutions u_h^k in parallel inside each subdomain $\Omega_{h,i}$ using any unconditionally stable scheme with the predictors $\tilde{u}_h^k$ computed at step 1 as the interface boundary conditions of Dirichlet type. Then pass part of subdomain data u_h^k from P_j to P_i for the correction computation. Go to step 3.
3. On each processor P_i, throw away the interface predictors obtained at step 1 and bring back u_h^{k-1} on Γ_{ij}. Compute the interface correctors u_h^k on Γ_{ij} in parallel using a fully implicit scheme. If $k = J$, stop the procedure; if not, return to step 1 for the next time level iteration with $k := k + 1$.

In the above algorithm, we consider noncrossover and crossover types of zigzag-shaped interfaces and the corresponding strategies of domain decomposition in step 0. The forward Euler scheme is used at step 1 and the backward Euler scheme is applied at steps 2–3. In other words, the implicit correction technique [9] and special zigzag-shaped interfaces are added to Kuznetsov's EIDD method [8]. The resulting method is denoted by CEIDD-ZI (CEIDD based on zigzag-shaped interfaces). Since the interface between two subdomains is only one point in the 1-D case, the 1-D CEIDD-ZI algorithm is also denoted by CEIDD.

For practical implementations, we always assume that $2 \le p \ll |\Omega_h|$ so that the parallel algorithms are of coarse granularity, where $|\Omega_h|$ is the number of unknown points inside Ω_h. However, for theoretical approaches, it is reasonable to assume that

(1.4) *each subdomain has at least one interior grid point.*

For a negative example, let $p = |\Omega_h|$; then each subdomain has no interior point and step 2 of the algorithm can be omitted. Actually, step 1 is also not needed because the interface predictors $\tilde{u}_h^k$ at Γ_h are thrown away at step 3. Therefore, an elliptic solver has to be employed to find the solutions u_h^k on $\Gamma_h (= \Omega_h)$ at step 3. The solver is carried out on only one processor with other processors standing idle. In such a case, however, CEIDD algorithms are unconditionally stable since it essentially equals a sequential backward Euler scheme. On the other hand, if there exists an empty subdomain, i.e., a subdomain has no interior grid points, the interface nearby can be deleted to make the new subdomain contain at least one interior point.

Under the subdomain-width assumption (1.4), we obtain following results.

1. The CEIDD-ZI methods add negligible computation cost to EIDD algorithms (Propositions 3.1 and 5.1).
2. The CEIDD-ZI methods are unconditionally stable (Theorems 4.3 and 6.3).
3. With an order of $\mathcal{O}\big(\sqrt{1 + \sum_{m=1}^{d}(p_m - 1)h^{-1}(\tau + h^2)}\big)$, the numerical solutions of the CEIDD-ZI methods converge to the exact solution of continuous problem (Theorems 4.4 and 6.4).

For the linear problem considered in the present paper, the unconditional stability of a domain decomposition algorithm is theoretically proved in the sense that the numerical solutions of the algorithm are continuously dependent on the initial value $u^0(x)$ and the outer-forced term $f(x,t)$ without any restrictions of the mesh ratio $r = \tau/h$ and the strategy of domain decomposition.

The context will be organized as follows. The next section presents some notation and auxiliary lemmas. The detailed presentation of the CEIDD method for the 1-D problem is given in section 3. Section 4 has witnessed the rigorous study of the stability and convergence of the CEIDD method based on multisubdomain decompo-

sition. Designing noncrossover and crossover types of zigzag-line interior boundary, we present two-dimensional (2-D) CEIDD-ZI procedures for two strategies of domain decomposition in section 5 together with analysis of computation and communication overhead of the correction. Section 6 provides the theoretical analysis of the two parallel CEIDD-ZI approaches. Numerical experiments are addressed in section 7, and some comments, including the three-dimensional (3-D) extension of CEIDD-ZI algorithms, are presented in the concluding section.

2. Notation and some auxiliary lemmas. For 1-D problems, $\Omega_h^\tau = \{(ih, k\tau) | 1 \le i \le N-1, 0 \le k \le J\}$. With v a mesh function on Ω_h^τ, we denote $\partial_t v^k = (v^k - v^{k-1})/\tau$, $\delta_x v_{i-\frac{1}{2}} = (v_i - v_{i-1})/h$, and

$$\Delta_h v_i = (a_{i+\frac{1}{2}} \delta_x v_{i+\frac{1}{2}} - a_{i-\frac{1}{2}} \delta_x v_{i-\frac{1}{2}})/h = (a_{i+\frac{1}{2}} v_{i+1} - 2\bar{a}_i v_i + a_{i-\frac{1}{2}} v_{i-1})/h^2,$$

where $a_{i-\frac{1}{2}} = a(x_{i-\frac{1}{2}}), \bar{a}_i = (a_{i+\frac{1}{2}} + a_{i-\frac{1}{2}})/2$. Supposing that grid functions $u, v \in \Omega_h$, we define the discrete inner product by $\langle u, v \rangle = h \sum_{i=1}^{N-1} u_i v_i$, the discrete L^2 norm by $\|v\| = \sqrt{\langle v, v \rangle}$, and the H^1 seminorm by

$$|v|_{a,1} = \|\sqrt{a}\delta_x v\| \equiv \sqrt{h \sum_{i=1}^{N} a_{i-\frac{1}{2}} (\delta_x v_{i-\frac{1}{2}})^2}.$$

In the 2-D case, $\Omega_h^\tau = \{(ih, jh, k\tau) | 1 \le i, j \le N-1, 0 \le k \le J\}$. The difference operators are defined similarly, such as $\delta_{x_1} v_{i-\frac{1}{2},j} = (v_{i,j} - v_{i-1,j})/h$,

$$\Delta_{h,1} v_{ij} = (a^1_{i+\frac{1}{2},j} \delta_{x_1} v_{i+\frac{1}{2},j} - a^1_{i-\frac{1}{2},j} \delta_{x_1} v_{i-\frac{1}{2},j})/h,$$

and $\Delta_h v_{ij} = \Delta_{h,1} v_{ij} + \Delta_{h,2} v_{ij}$. We also denote that $\bar{a}^1_{ij} = (a^1_{i+\frac{1}{2},j} + a^1_{i-\frac{1}{2},j})/2$, $\bar{a}^2_{ij} = (a^2_{i,j+\frac{1}{2}} + a^2_{i,j-\frac{1}{2}})/2$, and $\bar{a}^{12}_{ij} = \bar{a}^1_{ij} + \bar{a}^2_{ij}$. The inner product is defined by $\langle u, v \rangle = h^2 \sum_{i,j=1}^{N-1} u_{ij} v_{ij}$, the discrete L^2 norm by $\|v\| = \sqrt{\langle v, v \rangle}$, and the H^1 seminorm by

$$|v|_{a,1} = \sqrt{\|\sqrt{a^1}\delta_{x_1} v\|^2 + \|\sqrt{a^2}\delta_{x_2} v\|^2}.$$

Throughout this paper, the discrete energy method is employed to establish the stability and convergence of parallel CEIDD-ZI algorithms. Some preparatory lemmas are given below.

LEMMA 2.1. *Let $v \in \Omega_h^\tau$. Then it holds that*

(a) $2v^k \partial_t v^k = \partial_t [(v^k)^2] + \tau (\partial_t v^k)^2$;

(b) $2v^{k-1} \partial_t v^k = \partial_t [(v^k)^2] - \tau (\partial_t v^k)^2$.

Proof. Equality (a) can be derived directly from

$$2v^k (v^k - v^{k-1}) = (v^k)^2 - (v^{k-1})^2 + (v^k - v^{k-1})^2,$$

and (b) similarly. ∎

LEMMA 2.2. *Assume that $v \in \Omega_h^\tau$, and $v(x) = 0$ for $x \in \partial\Omega_h$. Then*

(a) $\langle v, \Delta_h v \rangle = -|v|^2_{a,1}$;

(b) $2\langle \partial_t v^k, \Delta_h v^k \rangle = -\partial_t (|v^k|^2_{a,1}) - \tau |\partial_t v^k|^2_{a,1}$.

Proof. The zero boundary condition and Abel integrate formulation of parts lead to (a). Equality (b) can be derived from (a) and Lemma 2.1(a). ∎

LEMMA 2.3 (discrete Gronwall inequality). *Assume that* $\{G^k|k \geq 0\}$ *is a nonnegative sequence,* $\sigma \geq 0$, $\Phi^0 \geq 0$, *and that* F^k *satisfies* $F^0 \leq \Phi^0$,

$$F^k \leq \Phi^0 + \sigma \sum_{l=0}^{k-1} F^l + \sum_{l=0}^{k} G^l, \quad k \geq 1.$$

Then F^k *satisfies*

$$F^k \leq e^{k\sigma}\left(\Phi^0 + \sum_{l=0}^{k} G^l\right), \quad k \geq 0.$$

Proof. An induction argument gives the result; see also [10] for a detail. ∎

3. Parallel CEIDD algorithm for 1-D problems. We decompose the domain $(0,1)$ into p nonoverlapping subdomains. In general, p is related to the problem size and the number of processors in the computer platform, and the subdomains may be of different widths; however, for theoretical considerations, the subdomain-width assumption (1.4) implies that $2 \leq p \leq N/2$, and that there are $(p-1)$ interfaces $\Gamma_h = \bigcup_{\alpha=1}^{p-1} \Gamma_{h,\alpha}$, where

$$\Gamma_{h,\alpha} = \{i_\alpha h|\ 4 \leq i_\alpha + 2 \leq i_{\alpha+1} \leq N-2\}. \tag{3.1}$$

Then Ω_h is decoupled into p nonoverlapping subdomains $\Omega_{h,1} = \{ih|\ 0 < i < i_1\}$, $\Omega_{h,\xi} = \{ih|\ i_{\xi-1} < i < i_\xi\}\,(\xi = 2, \ldots, p-1)$, and $\Omega_{h,p} = \{ih|\ i_{p-1} < i < N\}$, where $\Gamma_{h,\alpha}\,(1 \leq \alpha \leq p-1)$ is the common interface of subdomains $\Omega_{h,\alpha}$ and $\Omega_{h,\alpha+1}$.

Suppose that u_i^k is the numerical approximation of exact solution $u(ih, k\tau)$ on Ω_h^τ; the 1-D CEIDD method on general p subdomains is described below.

THE 1-D PARALLEL CEIDD ALGORITHM ON p SUBDOMAINS.

0. We have p processors denoted as $P_\xi\,(\xi = 1, \ldots, p)$. Assign subdomain $\Omega_{h,\xi}$ to processor P_ξ and interface $\Gamma_{h,\alpha}$ to processor $P_\alpha\,(\alpha = 1, \ldots, p-1)$. To start the procedure, the initial condition (1.2) is discretized as

$$u_i^0 = u^0(ih) \ \text{ on } \ \Omega_h. \tag{3.2}$$

Set $k = 1$ and go to step 1.

1. On each processor P_α, we obtain the predictor value $\tilde{u}_{i_\alpha}^k$ in parallel by applying the explicit Euler scheme to (1.1), that is,

$$\frac{\tilde{u}_{i_\alpha}^k - u_{i_\alpha}^{k-1}}{\tau} = \Delta_h u_{i_\alpha}^{k-1} + f_{i_\alpha}^{k-1} \ \text{ at } \ \Gamma_{h,\alpha}. \tag{3.3}$$

Then pass $\tilde{u}_{i_\alpha}^k$ from P_α to $P_{\alpha+1}$ and go to step 2.

2. The implicit Euler scheme is adopted on each subdomain $\Omega_{h,\xi}$, where the predictor values $\tilde{u}_h^k$ at Γ_h are served the Dirichlet boundary condition, viz.,

$$\begin{cases} \partial_t u_i^k = \Delta_h u_i^k + f_i^k & \text{on } \ \Omega_{h,\xi} \\ u_i^k = \tilde{u}_i^k & \text{at } \ \partial\Omega_{h,\xi} \cap \Gamma_h \\ u_i^k = u_b(ih, k\tau) & \text{at } \ \partial\Omega_{h,\xi} \cap \partial\Omega_h \end{cases} \quad \big(\xi = 1, 2, \ldots, p\big), \tag{3.4}$$

where $\partial\Omega_{h,\xi} \cap \Gamma_h = \{ih|i = i_{\xi-1} \ \text{ and/or } \ i = i_\xi\}$, $\partial\Omega_{h,\xi} \cap \partial\Omega_h$ is always empty except $\partial\Omega_{h,1} \cap \partial\Omega_h = \{0\}$, and $\partial\Omega_{h,p} \cap \partial\Omega_h = \{Nh\}$. Thus, the Thomas algorithm can be applied in parallel on each processor to find the subdomain solutions. Then pass the subdomain data $u_{i_\alpha+1}^k$ from $P_{\alpha+1}$ to P_α for the correction computation. Go to step 3.

3. On each processor P_α, throw away the interface predictors $\tilde{u}^k_{i_\alpha}$ obtained at step 1 and bring back $u^{k-1}_{i_\alpha}$ at $\Gamma_{h,\alpha}$. We obtain the interface solution $u^k_{i_\alpha}$ in parallel by using the implicit Euler scheme,

$$\partial_t u^k_{i_\alpha} = \Delta_h u^k_{i_\alpha} + f^k_{i_\alpha} \quad \text{at} \quad \Gamma_{h,\alpha}. \tag{3.5}$$

If $k = J$, stop the procedure; else, return to step 1 with $k := k + 1$.

For each interface boundary $\Gamma_{h,\alpha}$ ($\alpha = 1, 2, \ldots, p-1$), the interface-related steps (steps 1 and 3) are carried out on the assigned processor P_α, and then data transfer operations between P_α and $P_{\alpha+1}$ are necessary for each time level, as is the case with any EIDD method; however, the implicit correction step of the CEIDD method adds zero communication cost to EIDD algorithms because the data communication at step 2 is also necessary for computing the interface predictors at new time level t^{k+1} in EIDD algorithms.

On the other hand, given the solutions on subdomains, the implicit scheme (3.5) obtains the interface solution explicitly,

$$(1 + 2\bar{a}_{i_\alpha} r)u^k_{i_\alpha} = u^{k-1}_{i_\alpha} + r\left(a_{i_\alpha - \frac{1}{2}} u^k_{i_\alpha - 1} + a_{i_\alpha + \frac{1}{2}} u^k_{i_\alpha + 1}\right) + \tau f^k_{i_\alpha}, \quad 1 \le \alpha \le p-1,$$

where mesh ratio $r = \tau/h^2$. That is to say, the implicit correction (step 3) of the CEIDD algorithm adds negligible computation cost to any EIDD algorithm. Thus the CEIDD method maintains the efficiency in computation and communication of the EIDD methods, as the following proposition states.

PROPOSITION 3.1. *The implicit correction step of the CEIDD algorithm* (3.2)–(3.5) *adds zero communication and negligible computation cost to EIDD algorithms.*

Furthermore, in the next section, we will show that the proposed CEIDD algorithm is unconditionally stable. To end this section, we mention a special CEIDD procedure here for completeness. If there are $\frac{N}{2}$ subdomains, or each subdomain has only one interior point, then Kuznetsov's EIDD method is just the so-called explicit hopscotch scheme [7] and the CEIDD algorithm degenerates to a corrected explicit hopscotch (CEH) scheme, which is a small-block explicit method since the subdomain solution can be computed explicitly by using (3.4). Specifically, for computing solution u^k_h from the data at time level t^{k-1}, the CEH procedure can be described as the following point-related schemes, for $1 \le i \le N-1$:

$$\tilde{u}^k_i = u^{k-1}_i + \tau \Delta_h u^{k-1}_i + \tau f^{k-1}_i, \qquad i \in \text{even}, \tag{3.6}$$

$$(1 + 2\bar{a}_i r)u^k_i = u^{k-1}_i + r(a_{i-\frac{1}{2}} \tilde{u}^k_{i-1} + a_{i+\frac{1}{2}} \tilde{u}^k_{i+1}) + \tau f^k_i, \quad i \in \text{odd}, \tag{3.7}$$

$$(1 + 2\bar{a}_i r)u^k_i = u^{k-1}_i + r(a_{i-\frac{1}{2}} u^k_{i-1} + a_{i+\frac{1}{2}} u^k_{i+1}) + \tau f^k_i, \quad i \in \text{even}, \tag{3.8}$$

where the notation $\tilde{u}^k_0$ and $\tilde{u}^k_N$ is used in (3.7) with complementary definitions: $\tilde{u}^k_0 \equiv u^k_0$ and $\tilde{u}^k_N \equiv u^k_N$. Obviously, the CEH scheme is a fast CEIDD algorithm; however, as revealed later, it suffers from more loss of accuracy compared with the coarsely granular algorithm in the case of $2 < p \ll N$.

4. Theoretical analysis of the 1-D CEIDD algorithm. In this section, the prior estimations of the solution are established by the energy method. And the theoretical results on stability and convergence of the CEIDD algorithm are obtained by using those prior estimations.

LEMMA 4.1 (H^1-estimation). *Let $v^k_h = \{v^k_i | 0 \le i \le N, 0 \le k \le J\}$ satisfy the CEIDD algorithm on p subdomains decoupled by the interfaces Γ_h defined by* (3.1)*:*

$$\frac{\tilde{v}^k_i - v^{k-1}_i}{\tau} = \Delta_h v^{k-1}_i + \tilde{g}^{k-1}_i \quad \text{at} \quad \Gamma_h, \quad 1 \le k \le J, \tag{4.1}$$

$$\begin{cases} \partial_t v_i^k = \Delta_h v_i^k + g_i^k & on \quad \Omega_{h,\xi} \\ v_i^k = \tilde{v}_i^k & at \quad \Gamma_h \end{cases} \quad \big(\xi = 1, 2, \ldots, p\big), \quad 1 \le k \le J, \tag{4.2}$$

$$\partial_t v_i^k = \Delta_h v_i^k + g_i^k \qquad at \quad \Gamma_h, \quad 1 \le k \le J, \tag{4.3}$$

$$v_i^0 = \phi_i, \qquad 0 \le i \le N, \quad 1 \le k \le J, \tag{4.4}$$

$$v_0^k = v_N^k = 0, \qquad 0 \le k \le J. \tag{4.5}$$

Then it holds that

$$E_1^k \le \frac{3}{2} e^{\frac{3}{2}k\tau} \left[E_1^0 + \tau \sum_{l=1}^{k} \big(\|g^l\|^2 + |||g^l|||_{I,1}^2 \big) \right], \quad 1 \le k \le J, \tag{4.6}$$

for $0 < \tau \le 1/3$, *where*

$$E_1^k = |v^k|_{a,1}^2 + 2\tau^2 h^{-1} \sum_{\alpha=1}^{p-1} \bar{a}_{i_\alpha} (\Delta_h v_{i_\alpha}^k)^2,$$

$$|||g^k|||_{I,1}^2 = \sum_{\alpha=1}^{p-1} \left[2\bar{a}_{i_\alpha} \tau h^{-1} (\tilde{g}_{i_\alpha}^{k-1})^2 + \frac{1}{2} \big(h + 4\bar{a}_{i_\alpha} h^{-1} \big) (g_{i_\alpha}^k - \tilde{g}_{i_\alpha}^{k-1})^2 \right].$$

Proof. To make the calculations more transparent and the notation simpler, we are going to write out the proof of this lemma only for two subdomains $\Omega_{h,\xi}$ $(\xi = 1, 2)$ decoupled by the one-point interface $\Gamma_{h,1} = \{i_1 h\}$. The proof in the general p-subdomain case is similar to the 2-subdomain case, but the notation is messier.

For the 2-subdomain case, the interface schemes (4.1) and (4.3) lead to

$$(\tilde{v}_{i_1}^k - v_{i_1}^k) = -\tau^2 (\partial_t \Delta_h v_{i_1}^k) - \tau (g_{i_1}^k - \tilde{g}_{i_1}^{k-1}),$$

and the solution schemes (4.2)–(4.3) read as

$$\partial_t v_i^k = \Delta_h v_i^k + g_i^k, \quad 0 < i < i_1 - 1, \quad 1 \le k \le J, \tag{4.7}$$

$$\partial_t v_{i_1-1}^k = \Delta_h v_{i_1-1}^k + g_{i_1-1}^k + a_{i_1-\frac{1}{2}} (\tilde{v}_{i_1}^k - v_{i_1}^k)/h^2, \quad 1 \le k \le J, \tag{4.8}$$

$$\partial_t v_{i_1}^k = \Delta_h v_{i_1}^k + g_{i_1}^k, \quad 1 \le k \le J, \tag{4.9}$$

$$\partial_t v_{i_1+1}^k = \Delta_h v_{i_1+1}^k + g_{i_1+1}^k + a_{i_1+\frac{1}{2}} (\tilde{v}_{i_1}^k - v_{i_1}^k)/h^2, \quad 1 \le k \le J, \tag{4.10}$$

$$\partial_t v_i^k = \Delta_h v_i^k + g_i^k, \quad i_1 + 1 < i < N, \quad 1 \le k \le J. \tag{4.11}$$

Multiplying (4.7)–(4.11), respectively, by $2h\partial_t v_i^k$, $2h\partial_t v_{i_1-1}^k$, $2h\partial_t v_{i_1}^k$, $2h\partial_t v_{i_1+1}^k$, and $2h\partial_t v_i^k$, summing i, and adding the resulting equalities, we get

$$2\langle \partial_t v^k, \partial_t v^k \rangle = 2\langle \partial_t v^k, \Delta_h v^k \rangle + 2\langle \partial_t v^k, g^k \rangle + Q_{1i_1}^k.$$

By using Lemma 2.2(b) and inequality

$$2\langle \partial_t v^k, g^k \rangle \le \|\partial_t v^k\|^2 + \|g^k\|^2,$$

it follows that

$$\partial_t (|v^k|_{a,1}^2) \le -\|\partial_t v^k\|^2 - \tau |v^k|_{a,1}^2 + \|g^k\|^2 + Q_{1i_1}^k \le \|g^k\|^2 + Q_{1i_1}^k. \tag{4.12}$$

And, for the interface term $Q_{1i_1}^k$, it holds that

$$Q_{1i_1}^k = 2h^{-1} (\tilde{v}_{i_1}^k - v_{i_1}^k) \big(a_{i_1-\frac{1}{2}} \partial_t v_{i_1-1}^k + a_{i_1+\frac{1}{2}} \partial_t v_{i_1+1}^k \big)$$

$$
\begin{aligned}
&= 2h^{-1}(\tilde{v}_{i_1}^k - v_{i_1}^k)\left(h^2\partial_t\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}\partial_t v_{i_1}^k\right)\\
&= -2h^{-1}\Big[\tau^2(\partial_t\Delta_h v_{i_1}^k) + \tau(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})\Big]\Big(h^2\partial_t\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}g_{i_1}^k\Big)\\
&= I_1 + I_2 + I_3 + I_4 + I_5,
\end{aligned}
$$

where, applying Lemma 2.1 and the well-known Young inequality,

$$
\begin{aligned}
I_1 &= -2\tau^2h^{-1}(\partial_t\Delta_h v_{i_1}^k)\left(h^2\partial_t\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}\Delta_h v_{i_1}^k\right)\\
&= -2\tau^2h(\partial_t\Delta_h v_{i_1}^k)^2 - 2\bar{a}_{i_1}\tau^2h^{-1}\partial_t[(\Delta_h v_{i_1}^k)^2] - 2\bar{a}_{i_1}\tau^3h^{-1}(\partial_t\Delta_h v_{i_1}^k)^2,\\
I_2 &= -4\bar{a}_{i_1}\tau^2h^{-1}(\partial_t\Delta_h v_{i_1}^k)g_{i_1}^k \le 2\bar{a}_{i_1}\tau^3h^{-1}(\partial_t\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k)^2,\\
I_3 &= -2\tau h(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})(\partial_t\Delta_h v_{i_1}^k) \le 2\tau^2h(\partial_t\Delta_h v_{i_1}^k)^2 + \frac{1}{2}h(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})^2,\\
I_4 &= -4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})(\Delta_h v_{i_1}^k)\\
&\le 2\bar{a}_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})^2,\\
I_5 &= -4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})g_{i_1}^k = -4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k)^2 + 4\bar{a}_{i_1}\tau h^{-1}g_{i_1}^k\tilde{g}_{i_1}^{k-1}\\
&\le -2\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau h^{-1}(\tilde{g}_{i_1}^{k-1})^2.
\end{aligned}
$$

Thus, adding I_m from $m = 1$ to 5, we obtain

$$
Q_{1i_1}^k \le -2\bar{a}_{i_1}\tau^2h^{-1}\partial_t[(\Delta_h v_{i_1}^k)^2] + 2\bar{a}_{i_1}\tau^2h^{-1}(\Delta v_{i_1}^k)^2 + |||g^k|||_{I,1}^2. \tag{4.13}
$$

From (4.13) and (4.12), it follows that

$$
\partial_t E_1^k \le E_1^k + \|g^k\|^2 + |||g^k|||_{I,1}^2, \qquad 1 \le k \le J,
$$

which leads to

$$
(1-\tau)E_1^k \le E_1^0 + \tau\sum_{l=1}^{k-1}E_1^l + \tau\sum_{l=1}^{k}\left(\|g^l\|^2 + |||g^l|||_{I,1}^2\right), \qquad 1 \le k \le J.
$$

Supposing that $0 < \tau \le 1/3$, we have

$$
E_1^k \le \frac{3}{2}E_1^0 + \frac{3\tau}{2}\sum_{l=1}^{k-1}E_1^l + \frac{3\tau}{2}\sum_{l=1}^{k}\left(\|g^l\|^2 + |||g^l|||_{I,1}^2\right), \qquad 1 \le k \le J.
$$

Then Lemma 2.3 implies prior estimation (4.6), and the proof is complete. ∎

LEMMA 4.2 (L^2-estimation). *Let $v_h^k = \{v_i^k | 0 \le i \le N, 0 \le k \le J\}$ be the solution of the CEIDD algorithm* (4.1)–(4.5). *Then it holds that*

$$
\begin{aligned}
E_2^k \le{}& \frac{3}{2}(J+1)\exp\left(\frac{3}{2}(1+T)k\tau\right)E_2^0\\
&+ \frac{3T}{2}\exp\left(\frac{3}{2}(1+T)k\tau\right)\left[4\tau\sum_{l=0}^{k}E_1^l + \tau\sum_{l=1}^{k}(\|g^l\|^2 + |||g^l|||_{I,2}^2)\right]
\end{aligned} \tag{4.14}
$$

for $0 < \tau \le 1/3, 1 \le k \le J$, where

$$
E_2^k = \tau\sum_{l=0}^{k}\|v^l\|^2 + \tau^3h\sum_{\alpha=1}^{p-1}\sum_{l=0}^{k}(\Delta_h v_{i_\alpha}^l)^2 + 2\tau^2h^{-1}\sum_{\alpha=1}^{p-1}\bar{a}_{i_\alpha}(v_{i_\alpha}^k)^2,
$$

$$|||g^1|||_{I,2}^2 = \sum_{\alpha=1}^{p-1} \left\{ 2\bar{a}_{i_\alpha}\tau^2 h^{-1}\Big[(g_{i_\alpha}^1)^2 + 2(\tilde{g}_{i_\alpha}^0)^2\Big] + h(g_{i_\alpha}^1 - \tilde{g}_{i_\alpha}^0)^2 + 2\bar{a}_{i_\alpha}h^{-1}(\tilde{g}_{i_\alpha}^0)^2 \right\},$$

$$\begin{aligned}|||g^k|||_{I,2}^2 = &\sum_{\alpha=1}^{p-1} 2\bar{a}_{i_\alpha}\tau^2 h^{-1}\Big[(g_{i_\alpha}^k)^2 + 2(g_{i_\alpha}^{k-1})^2 + 2(\tilde{g}_{i_\alpha}^{k-1})^2\Big] \\ &+ \sum_{\alpha=1}^{p-1} \Big[h(g_{i_\alpha}^k - \tilde{g}_{i_\alpha}^{k-1})^2 + 2\bar{a}_{i_\alpha}h^{-1}(g_{i_\alpha}^{k-1} - \tilde{g}_{i_\alpha}^{k-1})^2\Big], \quad k \geq 2.\end{aligned}$$

Proof. As in the proof of Lemma 4.1, we are going to write out this proof only for two subdomains $\Omega_{h,\xi}\,(\xi = 1, 2)$ decoupled by the one-point interface $\Gamma_{h,1} = \{i_1 h\}$ and suppose that v_h^k satisfies (4.7)–(4.11) and (4.4)–(4.5). For the sake of succinctness, we also use the notation

$$G^k \equiv \|v^k\|^2 + \tau^2 h(\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau^2 h^{-1}\partial_t(v_{i_1}^k)^2;$$

then $G^k = \partial_t E_2^k$ for $1 \leq k \leq J$.

Multiplying (4.7)–(4.11), respectively, by $2hv_i^k$, $2hv_{i_1-1}^k$, $2hv_{i_1}^k$, $2hv_{i_1+1}^k$, and $2hv_i^k$, summing i, and adding the resulting equalities, we get

$$2\left\langle v^k, \partial_t v^k \right\rangle = 2\left\langle v^k, \Delta_h v^k \right\rangle + 2\left\langle v^k, g^k \right\rangle + Q_{2i_1}^k.$$

Owing to

$$2\left\langle v^k, \partial_t v^k \right\rangle = \partial_t(\|v^k\|^2) + \tau\|\partial_t v^k\|^2$$

and

$$2\langle v^k, g^k\rangle \leq \|v^k\|^2 + \|g^k\|^2,$$

the above equality becomes

$$\partial_t(\|v^k\|^2) \leq \|v^k\|^2 + \|g^k\|^2 + Q_{2i_1}^k, \tag{4.15}$$

where the interface term $Q_{2i_1}^k$ reads

$$Q_{2i_1}^k = 2h^{-1}(\tilde{v}_{i_1}^k - v_{i_1}^k)\left(h^2\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}v_{i_1}^k\right).$$

Recalling the expression of $(\tilde{v}_{i_1}^k - v_{i_1}^k)$ and the scheme (4.9), we obtain that

$$\begin{aligned} Q_{2i_1}^k &= 2h^{-1}(\tilde{v}_{i_1}^k - v_{i_1}^k)\left(h^2\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}\tau\partial_t v_{i_1}^k + 2\bar{a}_{i_1}v_{i_1}^{k-1}\right) \\ &= 2h^{-1}(\tilde{v}_{i_1}^k - v_{i_1}^k)\left(h^2\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}\tau\Delta_h v_{i_1}^k + 2\bar{a}_{i_1}v_{i_1}^{k-1} + 2\bar{a}_{i_1}\tau g_{i_1}^k\right) \\ &= -2\tau^2 h^{-1}(2\bar{a}_{i_1}\tau + h^2)(\partial_t\Delta_h v_{i_1}^k)\Delta_h v_{i_1}^k - 4\bar{a}_{i_1}\tau^2 h^{-1}(\partial_t\Delta_h v_{i_1}^k)v_{i_1}^{k-1} \\ &\quad - 4\bar{a}_{i_1}\tau^3 h^{-1}(\partial_t\Delta_h v_{i_1}^k)g_{i_1}^k - 2\tau h^{-1}(2\bar{a}_{i_1}\tau + h^2)(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})\Delta_h v_{i_1}^k \\ &\quad - 4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})v_{i_1}^{k-1} - 4\bar{a}_{i_1}\tau^2 h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})g_{i_1}^k. \end{aligned} \tag{4.16}$$

For a complete analysis of the term $-4\bar{a}_{i_1}\tau^2 h^{-1}(\partial_t\Delta_h v_{i_1}^k)v_{i_1}^{k-1}$ of (4.16), two cases of k will be discussed: one is $k \geq 2$, the other $k = 1$.

(a) If $k \geq 2$, by using (4.9) and Lemma 2.1, we derive that

$$\begin{aligned}
&-2\tau(\partial_t\Delta_h v_{i_1}^k)v_{i_1}^{k-1} = -2(\Delta_h v_{i_1}^k - \Delta_h v_{i_1}^{k-1})v_{i_1}^{k-1}\\
&= -2(\partial_t v_{i_1}^k - \partial_t v_{i_1}^{k-1})v_{i_1}^{k-1} + 2(g_{i_1}^k - g_{i_1}^{k-1})v_{i_1}^{k-1}\\
&= -\partial_t(v_{i_1}^k)^2 + \tau(\partial_t v_{i_1}^k)^2 + \partial_t(v_{i_1}^{k-1})^2 + \tau(\partial_t v_{i_1}^{k-1})^2 + 2(g_{i_1}^k - g_{i_1}^{k-1})v_{i_1}^{k-1}\\
&= -\tau\partial_t\left[\partial_t(v_{i_1}^k)^2\right] + \tau(\Delta_h v_{i_1}^k + g_{i_1}^k)^2 + \tau(\Delta_h v_{i_1}^{k-1} + g_{i_1}^{k-1})^2 + 2(g_{i_1}^k - g_{i_1}^{k-1})v_{i_1}^{k-1}.
\end{aligned}$$

Then interface term (4.16) follows

$$Q_{2i_1}^k = I_6 + I_7 + I_8 + I_9 + I_{10} + I_{11},$$

in which I_m $(m = 6, 7, \ldots, 11)$ are treated in detail as follows:

$$\begin{aligned}
I_6 &= -2\tau^2 h^{-1}(2\bar{a}_{i_1}\tau + h^2)(\partial_t\Delta_h v_{i_1}^k)\Delta_h v_{i_1}^k - 2\bar{a}_{i_1}\tau^2 h^{-1}\partial_t\left[\partial_t(v_{i_1}^k)^2\right]\\
&\quad + 2\bar{a}_{i_1}\tau^2 h^{-1}(\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau^2 h^{-1}(\Delta_h v_{i_1}^{k-1})^2\\
&= -\tau^2 h\partial_t(\Delta_h v_{i_1}^k)^2 - 2\bar{a}_{i_1}\tau^2 h^{-1}\partial_t\left[\partial_t(v_{i_1}^k)^2\right] + 4\bar{a}_{i_1}\tau^2 h^{-1}(\Delta_h v_{i_1}^{k-1})^2\\
&\quad - \tau^3 h(\partial_t\Delta_h v_{i_1}^k)^2 - 2\bar{a}_{i_1}\tau^4 h^{-1}(\partial_t\Delta_h v_{i_1}^k)^2,\\
I_7 &= -4\bar{a}_{i_1}\tau^3 h^{-1}(\partial_t\Delta_h v_{i_1}^k)g_{i_1}^k \leq 2\bar{a}_{i_1}\tau^4 h^{-1}(\partial_t\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau^2 h^{-1}(g_{i_1}^k)^2,\\
I_8 &= -2\tau h(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})\Delta_h v_{i_1}^k \leq \tau^2 h(\Delta_h v_{i_1}^k)^2 + h(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})^2,\\
I_9 &= 4\bar{a}_{i_1}\tau^2 h^{-1}\left[g_{i_1}^k\Delta_h v_{i_1}^k + g_{i_1}^{k-1}\Delta_h v_{i_1}^{k-1} - (g_{i_1}^k - \tilde{g}_{i_1}^{k-1})\Delta_h v_{i_1}^k\right]\\
&= 4\bar{a}_{i_1}\tau^2 h^{-1}\left(\tilde{g}_{i_1}^{k-1}\Delta_h v_{i_1}^k + g_{i_1}^{k-1}\Delta_h v_{i_1}^{k-1}\right)\\
&\leq 2\bar{a}_{i_1}\tau^2 h^{-1}\left[(\Delta_h v_{i_1}^k)^2 + (\Delta_h v_{i_1}^{k-1})^2\right] + 2\bar{a}_{i_1}\tau^2 h^{-1}\left[(g_{i_1}^{k-1})^2 + (\tilde{g}_{i_1}^{k-1})^2\right],\\
I_{10} &= 4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k - g_{i_1}^{k-1})v_{i_1}^{k-1} - 4\bar{a}_{i_1}\tau h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})v_{i_1}^{k-1}\\
&\leq 2\bar{a}_{i_1}\tau^2 h^{-1}(v_{i_1}^{k-1})^2 + 2\bar{a}_{i_1}h^{-1}(g_{i_1}^{k-1} - \tilde{g}_{i_1}^{k-1})^2,\\
I_{11} &= 2\bar{a}_{i_1}\tau^2 h^{-1}\left[(g_{i_1}^k)^2 + (g_{i_1}^{k-1})^2\right] - 4\bar{a}_{i_1}\tau^2 h^{-1}(g_{i_1}^k - \tilde{g}_{i_1}^{k-1})g_{i_1}^k\\
&\leq 2\bar{a}_{i_1}\tau^2 h^{-1}\left[(g_{i_1}^{k-1})^2 + (\tilde{g}_{i_1}^{k-1})^2\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
Q_{2i_1}^k \leq &-\tau^2 h\partial_t(\Delta_h v_{i_1}^k)^2 - 2\bar{a}_{i_1}\tau^2 h^{-1}\partial_t\left[\partial_t(v_{i_1}^k)^2\right]\\
&+ \tau^2 h(\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau^2 h^{-1}(v_{i_1}^{k-1})^2\\
&+ 2\bar{a}_{i_1}\tau^2 h^{-1}(\Delta_h v_{i_1}^k)^2 + 6\bar{a}_{i_1}\tau^2 h^{-1}(\Delta_h v_{i_1}^{k-1})^2 + |||g^k|||_{I,2}^2, \quad k \geq 2.
\end{aligned} \tag{4.17}$$

Substituting (4.17) into (4.15) and using Lemma 4.1, we obtain

$$\begin{aligned}
\partial_t G^k \leq &\|v^k\|^2 + \tau^2 h(\Delta_h v_{i_1}^k)^2 + 2\bar{a}_{i_1}\tau^2 h^{-1}(v_{i_1}^{k-1})^2\\
&+ E_1^k + 3E_1^{k-1} + \|g^k\|^2 + |||g^k|||_{I,2}^2, \qquad k \geq 2,
\end{aligned}$$

which implies

$$\begin{aligned}
G^n - G^1 \leq &\tau\sum_{l=2}^{n}\|v^l\|^2 + \tau^3 h\sum_{l=2}^{n}(\Delta_h v_{i_1}^l)^2 + 2\bar{a}_{i_1}\tau^3 h^{-1}\sum_{l=1}^{n-1}(v_{i_1}^l)^2\\
&+ \tau\sum_{l=2}^{n}(E_1^l + 3E_1^{l-1}) + \tau\sum_{l=2}^{n}(\|g^l\|^2 + |||g^l|||_{I,2}^2), \qquad k \geq 2.
\end{aligned} \tag{4.18}$$

(b) If $k=1$, by using (4.9) and Lemma 2.1, we have

$$\begin{aligned}&-2\tau(\partial_t\Delta_h v_{i_1}^1)v_{i_1}^0=-2(\Delta_h v_{i_1}^1-\Delta_h v_{i_1}^0)v_{i_1}^0\\&=-2(\partial_t v_{i_1}^1)v_{i_1}^0+2g_{i_1}^1v_{i_1}^0+2(\Delta_h v_{i_1}^0)v_{i_1}^0\\&\le-\partial_t(v_{i_1}^1)^2+\tau(\Delta_h v_{i_1}^1+g_{i_1}^1)^2+2g_{i_1}^1v_{i_1}^0+\tau(\Delta_h v_{i_1}^0)^2+\tau^{-1}(v_{i_1}^0)^2.\end{aligned}$$

Then, by taking $k=1$ in (4.16), we can arrive at

$$Q_{2i_1}^1\le I_{12}+I_{13}+I_{14}+I_{15}+I_{16}+I_{17},$$

where,

$$\begin{aligned}I_{12}&=-2\tau^2h^{-1}(2\bar a_{i_1}\tau+h^2)(\partial_t\Delta_h v_{i_1}^1)\Delta_h v_{i_1}^1-2\bar a_{i_1}\tau h^{-1}\partial_t\left[(v_{i_1}^1)^2\right]\\&\quad+2\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^1)^2+2\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^0)^2\\&=-\tau^2h\partial_t(\Delta_h v_{i_1}^1)^2-2\bar a_{i_1}\tau h^{-1}\partial_t\left[(v_{i_1}^1)^2\right]+4\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^0)^2\\&\quad-\tau^3h(\partial_t\Delta_h v_{i_1}^1)^2-2\bar a_{i_1}\tau^4h^{-1}(\partial_t\Delta_h v_{i_1}^1)^2,\\I_{13}&=-4\bar a_{i_1}\tau^3h^{-1}(\partial_t\Delta_h v_{i_1}^1)g_{i_1}^1\le2\bar a_{i_1}\tau^4h^{-1}(\partial_t\Delta_h v_{i_1}^1)^2+2\bar a_{i_1}\tau^2h^{-1}(g_{i_1}^1)^2,\\I_{14}&=-2\tau h(g_{i_1}^1-\tilde g_{i_1}^0)\Delta_h v_{i_1}^1\le\tau^2h(\Delta_h v_{i_1}^1)^2+h(g_{i_1}^1-\tilde g_{i_1}^0)^2,\\I_{15}&=4\bar a_{i_1}\tau^2h^{-1}\left[g_{i_1}^1\Delta_h v_{i_1}^1-(g_{i_1}^1-\tilde g_{i_1}^0)\Delta_h v_{i_1}^1\right]\\&=4\bar a_{i_1}\tau^2h^{-1}\tilde g_{i_1}^0\Delta_h v_{i_1}^1\le2\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^1)^2+2\bar a_{i_1}\tau^2h^{-1}(\tilde g_{i_1}^0)^2,\\I_{16}&=4\bar a_{i_1}\tau h^{-1}g_{i_1}^1v_{i_1}^0-4\bar a_{i_1}\tau h^{-1}(g_{i_1}^1-\tilde g_{i_1}^0)v_{i_1}^0+2\bar a_{i_1}h^{-1}(v_{i_1}^0)^2\\&\le2\bar a_{i_1}\tau^2h^{-1}(v_{i_1}^0)^2+2\bar a_{i_1}h^{-1}(v_{i_1}^0)^2+2\bar a_{i_1}h^{-1}(\tilde g_{i_1}^0)^2,\\I_{17}&=2\bar a_{i_1}\tau^2h^{-1}(g_{i_1}^1)^2-4\bar a_{i_1}\tau^2h^{-1}(g_{i_1}^1-\tilde g_{i_1}^0)g_{i_1}^1\le2\bar a_{i_1}\tau^2h^{-1}(\tilde g_{i_1}^0)^2.\end{aligned}$$

Therefore,

$$\begin{aligned}Q_{2i_1}^1\le&-\tau^2h\partial_t(\Delta_h v_{i_1}^1)^2-2\bar a_{i_1}\tau h^{-1}\partial_t(v_{i_1}^1)^2\\&+\tau^2h(\Delta_h v_{i_1}^1)^2+2\bar a_{i_1}\tau^2h^{-1}(v_{i_1}^0)^2+2\bar a_{i_1}h^{-1}(v_{i_1}^0)^2\\&+2\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^1)^2+4\bar a_{i_1}\tau^2h^{-1}(\Delta_h v_{i_1}^0)^2+|||g^1|||_{I,2}^2.\end{aligned}\tag{4.19}$$

Inserting (4.19) into (4.15) and multiplying the resulting inequality by τ, we obtain

$$\begin{aligned}G^1\le&\tau\|v^1\|^2+\tau^3h(\Delta_h v_{i_1}^1)^2+2\bar a_{i_1}\tau^3h^{-1}(v_{i_1}^0)^2\\&+\|v^0\|^2+\tau^2h(\Delta_h v_{i_1}^0)^2+2\bar a_{i_1}\tau h^{-1}(v_{i_1}^0)^2\\&+\tau(E_1^1+2E_1^0)+\tau(\|g^1\|^2+|||g^1|||_{I,2}^2).\end{aligned}\tag{4.20}$$

Now we combine the result of (a) with (b) to continue this proof. Adding (4.20) to (4.18), we get

$$\partial_tE_2^n=G^n\le E_2^n+\tau^{-1}E_2^0+\tau\sum_{l=0}^{n-1}E_2^l+4\tau\sum_{l=0}^{n}E_1^l+\tau\sum_{l=1}^{n}(\|g^l\|^2+|||g^l|||_{I,2}^2)$$

for $1\le n\le J$. Furthermore, summing n from 1 to k, we have

$$(1-\tau)E_2^k\le\tau(1+T)\sum_{l=0}^{k-1}E_2^l+(k+1)E_2^0+4\tau T\sum_{l=0}^{k}E_1^l+\tau T\sum_{l=1}^{k}\|g^l\|_*^2,$$

where $\|g^k\|_*^2 = \|g^k\|^2 + |||g^k|||_{I,2}^2$. Supposing that $0 < \tau \le 1/3$; it follows

$$E_2^k \le \frac{3(1+T)\tau}{2}\sum_{l=0}^{k-1} E_2^l + \frac{3(k+1)}{2}E_2^0 + 6\tau T\sum_{l=0}^{k} E_1^l + \frac{3T\tau}{2}\sum_{l=1}^{k}\|g^l\|_*^2.$$

This implies (4.14), due to Lemma 2.3, and this proof is complete. □

Although not explicitly stated, the above proofs are true of the CEH scheme with $p = N/2$. In the analogous proofs, the schemes (4.2)–(4.3) could be rewritten as

$$\begin{aligned} &\partial_t u_i^k = \Delta_h v_i^k + g_i^k && \text{(at interface points)},\\ &\partial_t u_i^k = \Delta_h v_i^k + g_i^k + a_{i-\frac{1}{2}} w_{i-1}^k + a_{i+\frac{1}{2}} w_{i+1}^k && \text{(at interior points)}, \end{aligned}$$

where $w_i^k = (\tilde v_i^k - v_i^k)/h^2$ is defined at interface points and $w_0^k = w_N^k = 0$. Correspondingly, the interface terms $Q_{1i_1}^k$, $Q_{2i_1}^k$ would be replaced with $\sum_{\alpha=1}^{N/2-1} Q_{1i_\alpha}^k$ and $\sum_{\alpha=1}^{N/2-1} Q_{2i_\alpha}^k$, but the other parts of the proofs remain the same.

Obviously, Lemmas 4.1 and 4.2 imply the following stability theorem.

THEOREM 4.3 (stability). *Under the reasonable assumption* (1.4)*, the* 1*-D CEIDD algorithm* (3.2)–(3.5) *is unconditionally stable with respect to the* H^1 *seminorm and the* L^2 *norm.*

Before we end this section, let us briefly discuss the convergence of the CEIDD method. Let $e_i^k = u(ih, k\tau) - u_i^k$ be the error of solution u_i^k computed by the 1-D CEIDD algorithm (3.2)–(3.5), and let $\tilde e_{i_\alpha}^k = u(i_\alpha h, k\tau) - \tilde u_{i_\alpha}^k$ be the error of predictor value $\tilde u_{i_\alpha}^k$ at interface boundaries Γ_h. And the local truncation errors of the backward and forward Euler schemes are denoted by r_i^k, $\tilde r_i^{k-1}$.

It is easy to verify that the error equations of the 1-D CEIDD algorithm (3.2)–(3.5) is the same as (4.1)–(4.5) by assuming $\phi_i \equiv 0$ and replacing the variables $v_i^k, \tilde v_i^k, g_i^k, \tilde g_i^{k-1}$ with $e_i^k, \tilde e_i^k, r_i^k, \tilde r_i^{k-1}$, respectively. Thus, (4.6) of Lemma 4.1 takes the form

$$E_1^k(e) \le \frac{3\tau}{2} e^{\frac{3}{2}k\tau}\sum_{l=1}^{k}\left(\|r^l\|^2 + |||r^l|||_{I,1}^2\right), \quad 1 \le k \le J,$$

and L^2 estimation (4.14) of Lemma 4.2 holds in the form

$$E_2^k(e) \le \frac{3T}{2}\exp\left(\frac{3}{2}(1+T)k\tau\right)\left[4\tau\sum_{l=1}^{k} E_1^l(e) + \tau\sum_{l=1}^{k}(\|r^l\|^2 + |||r^l|||_{I,2}^2)\right].$$

Throughout the paper, any subscripted c, which is fixed in value, will denote a generic positive constant that is dependent on the exact solution $u(x,t)$ and coefficients $a(x)$ but independent of time step τ, spacing h, and the number of subdomains p. By the standard truncation error analysis, we get $|r_i^k| \le c_1(\tau + h^2)$ and $|\tilde r_i^k| \le c_1(\tau + h^2)$. If $a(x)$ is bounded by c_2, it holds that

$$\|r^k\|^2 + |||r^k|||_{I,1}^2 \le 2c_1^2(1+5c_2)\left(1 + (p-1)h^{-1}\right)(\tau + h^2)^2$$

and

$$|e^k|_{a,1} \le \sqrt{E_1^k(e)} \le c_3\sqrt{1 + (p-1)h^{-1}}(\tau + h^2), \quad 1 \le k \le J, \tag{4.21}$$

where $c_3 = \sqrt{3Te^{\frac{3}{2}T}c_1^2(1+5c_2)}$. Similarly,

$$\|r^k\|^2 + |||r^k|||_{I,2}^2 \le 2c_1^2(2+9c_2)\left(1+(p-1)h^{-1}\right)(\tau+h^2)^2$$

and

$$\sqrt{\tau\sum_{l=1}^{k}\|e^l\|^2} \le \sqrt{E_2^k(e)} \le c_4\sqrt{1+(p-1)h^{-1}}(\tau+h^2), \quad 1 \le k \le J, \tag{4.22}$$

where $c_4 = \sqrt{3T^2\exp\left(\frac{3}{2}(1+T)T\right)(2c_1^2+9c_1^2c_2+2c_3^2)}$. Thus, we obtain the following theorem from (4.21) and (4.22).

THEOREM 4.4 (convergence). *If the solution of problem* (1.1)–(1.3) *is sufficiently smooth, the numerical solution of the CEIDD algorithm* (3.2)–(3.5) *converges to the exact solution of* (1.1)–(1.3) *with an order of* $\mathcal{O}(\sqrt{1+(p-1)h^{-1}}(\tau+h^2))$ *with respect to the* H^1 *seminorm and the* L^2 *norm when spacing* h *is sufficiently small and time step size* $\tau = \mathcal{O}(p^{-1/2}h^{1/2+\epsilon})$ *for* $\epsilon > 0$.

We note that, for the coarsely granular algorithm in the case of $2 \le p \ll N$, the accuracy of the CEIDD algorithm is $\mathcal{O}(h^{-1/2}(\tau+h^2))$ as $\tau = \mathcal{O}(h^{1/2+\epsilon})$. Compared with sequential implicit Euler method ($p = 1$), there is a loss of $h^{-1/2}$. However, for the CEH scheme with $p = h^{-1}/2$ (see section 3), the accuracy decreases to $\mathcal{O}(h^{-1}(\tau+h^2))$ with $\tau = \mathcal{O}(h^{1+\epsilon})$, in which a loss of h^{-1} is seen. As for the cause, we return to the upper bounds of $E_1^k(e)$ and $E_2^k(e)$ and observe that the loss of accuracy can be ascribed to the interface-related terms $2\bar{a}_{i_\alpha}h^{-1}(r_{i_\alpha}^k - \tilde{r}_{i_\alpha}^{k-1})^2$ of $|||r^k|||_{I,1}$ and $2\bar{a}_{i_\alpha}h^{-1}(r_{i_\alpha}^{k-1} - \tilde{r}_{i_\alpha}^{k-1})^2$ of $|||r^k|||_{I,2}$. To make this clearer, we refer to the terms I_4 in the proof of Lemma 4.1 and I_{10} in Lemma 4.2. Therefore, it stands to reason that the accumulated truncation errors at interface nodes, introduced by the explicit prediction and implicit correction steps, affect the global accuracy of our CEIDD algorithm.

On the other hand, noticing that $I_{16}(e) = 2\bar{a}_{i_1}h^{-1}(e_{i_1}^0)^2$ (since $e_{i_1}^0 = 0$) in the analogous proof of Lemma 4.2 for L^2 convergence, we can get

$$|||r^1|||_{I,2}^2 = \sum_{\alpha=1}^{p-1}\left[2\bar{a}_{i_\alpha}\tau^2h^{-1}\left((r_{i_\alpha}^1)^2 + (\tilde{r}_{i_\alpha}^0)^2\right) + h(r_{i_\alpha}^1 - \tilde{r}_{i_\alpha}^0)^2\right].$$

Compared with the earlier definition, this formulation removes the main error term. (Actually, the arguments would be true of the 2-D parallel algorithms described in the next two sections but are omitted there.) Thus, an improved accuracy of the CEIDD algorithm would be obtained by using our prior estimations (4.6) and (4.14) if it holds that $|r_{i_\alpha}^k - \tilde{r}_{i_\alpha}^{k-1}| \le c_1\tau h^2$ and $|r_{i_\alpha}^{k-1} - \tilde{r}_{i_\alpha}^{k-1}| \le c_1\tau h^2$; see Example 1 for a 2-D instance. Specifically, we can get

$$|e^k|_{a,1} \le c_5\sqrt{1+(p-1)\tau h^{-1}}(\tau+h^2), \quad 1 \le k \le J,$$

and

$$\sqrt{\tau\sum_{l=1}^{k}\|e^l\|^2} \le c_6\sqrt{1+(p-1)\tau^2h^{-1}}(\tau+h^2), \quad 1 \le k \le J.$$

Thus, the convergence rate reaches $\mathcal{O}(\tau+h^2)$ even for the CEH method.

5. Zigzag interfaces and CEIDD-ZI algorithms for 2-D problems. In the first half of this paper, we proved that the parallel CEIDD procedure is free from stability-related time step restriction and maintains the efficiency in computation and communication of the EIDD methods. For solving 2-D problems (1.1)–(1.3) on $\Omega = (0,1)^2$, we are in a different situation, where the interior interfaces are no longer any nodes but some directional lines such as grid lines. As the implicit correction technique (always, the fully implicit scheme) is applied at points along those straight-line interfaces, an elliptic solver should be employed to compute the solutions [11], and then an extra parallel time would be suffered. Moreover, for large-scale parallel simulations, domain partitioning flexibility is always of the essence, and some crossover interfaces would be necessary. When those straight-line boundaries cross into each other inside the computational domain, however, the situation becomes worse because the globalized transfers of data for computing interface solutions suffer from sustained parallel-time overload.

It seems that the 2-D generalization of the CEIDD algorithm would not inherit the virtues of the 1-D CEIDD method; nevertheless, an alternative approach will do in which the straight-line boundaries are tailored to the special lines of zigzag shape. In this and the following sections, we show that CEIDD-ZI (CEIDD based on zigzag-shaped interfaces) algorithms maintain the virtues of the CEIDD method and improve the flexibility in domain partitioning as well as localize the communication of data.

Under the subdomain-width condition (1.4), we assume that $2 \le p_1 \le N/2$ and $2 \le p_2 \le N/2$. The zigzag-line interfaces, $\Gamma_h = \Gamma_h^1 \cup \Gamma_h^2$, consist of $(p_1 - 1)$ interior boundaries $\Gamma_h^1 = \bigcup_{\alpha=1}^{p_1-1} \Gamma_{h,\alpha}^1$ in the first spatial direction with

$$\Gamma_{h,\alpha}^1 = \{(i_\alpha h, jh) |\ i_\alpha = m_\alpha + \mathrm{mod}(j,2), 3 \le m_\alpha + 2 \le m_{\alpha+1} \le N-3\}, \tag{5.1}$$

and $(p_2 - 1)$ interior boundaries $\Gamma_h^2 = \bigcup_{\beta=1}^{p_2-1} \Gamma_{h,\beta}^2$ in the second with

$$\Gamma_{h,\beta}^2 = \{(ih, j_\beta h) |\ j_\beta = n_\beta + \mathrm{mod}(i,2), 3 \le n_\beta + 2 \le n_{\beta+1} \le N-3\}. \tag{5.2}$$

Then, by those interior interfaces Γ_h, the discrete domain Ω_h is decomposed into $p = p_1 \times p_2$ subdomains $\Omega_{\xi\eta}\,(\xi = 1,2,\ldots,p_1, \eta = 1,2,\ldots,p_2)$. Two examples on the two strategies of domain partitioning are depicted graphically in Figures 5.1–5.2. Denoting that $\gamma_{\xi\eta}^1 = \partial\Omega_{\xi\eta} \cap \partial\Omega_{\xi+1,\eta}$, the common boundary of $\Omega_{\xi\eta}$ and $\Omega_{\xi+1,\eta}$, and $\gamma_{\xi\eta}^2 = \partial\Omega_{\xi\eta} \cap \partial\Omega_{\xi,\eta+1}$, it holds that $\Gamma_{h,\alpha}^1 = \bigcup_{\eta=1}^{p_2} \gamma_{\alpha\eta}^1$ and $\Gamma_{h,\beta}^2 = \bigcup_{\xi=1}^{p_1} \gamma_{\xi\beta}^2$.

We have $p = p_1 \times p_2$ processors denoted as $P_{\xi\eta}$. Similar to the 1-D case, we assign $\Omega_{\xi\eta}$ and $\gamma_{\xi\eta}^1 \cup \gamma_{\xi\eta}^2$ to processor $P_{\xi\eta}$. Suppose that u_{ij}^k is the numerical approximation of the exact solution $u(ih, jh, k\tau)$ on Ω_h^τ; the 2-D CEIDD-ZI method for computing the solution $u_h^k = \{u_{ij}^k | 0 \le i,j \le N\}$ from u_h^{k-1} is described below compactly:

$$\frac{\tilde{u}_{ij}^k - u_{ij}^{k-1}}{\tau} = \Delta_h u_{ij}^{k-1} + f_{ij}^{k-1} \quad \text{on} \quad \Gamma_h, \tag{5.3}$$

$$\begin{cases} \partial_t u_{ij}^k = \Delta_h u_{ij}^k + f_{ij}^k & \text{on} \quad \Omega_{\xi\eta} \\ u_{ij}^k = \tilde{u}_{ij}^k & \text{on} \quad \Gamma_h \end{cases} \begin{pmatrix} \xi = 1,2,\ldots,p_1 \\ \eta = 1,2,\ldots,p_2 \end{pmatrix}, \tag{5.4}$$

$$\partial_t u_{ij}^k = \Delta_h u_{ij}^k + f_{ij}^k \quad \text{on} \quad \Gamma_h, \tag{5.5}$$

$$u_{ij}^0 = u^0(ih, jh) \quad \text{on} \quad \Omega_h, \tag{5.6}$$

$$u_{ij}^k = u_b(ih, jh, k\tau) \quad \text{on} \quad \partial\Omega_h. \tag{5.7}$$

Based on the two strategies of domain decomposition, we distinguish the CEIDD-ZI method on $p_1 \times 1$ subdomains from that on $p_1 \times p_2$ subdomains $(p_2 > 1)$ and denote

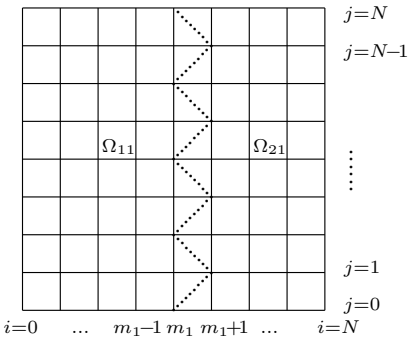


FIG. 5.1. *Strategy* 1: *The zigzag-line interface* Γ_h (*dashed lines*) *divides* Ω_h *into* 2×1 *subdomains for* $p_1 = 2$ *and* $p_2 = 1$.

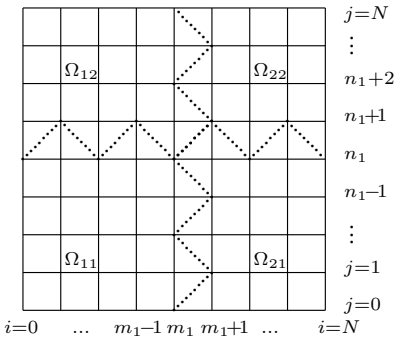


FIG. 5.2. *Strategy* 2: *The zigzag-line interfaces* Γ_h (*dashed lines*) *divide* Ω_h *into* 2×2 *subdomains for* $p_1 = p_2 = 2$.

the former as CEIDD-ZI1 and the later as CEIDD-ZI2; see Figures 5.1–5.2 for the relevant instances.

As for the CEIDD-ZI1 method, $\Gamma_h = \Gamma_h^1$. We compute interface predictor values $\tilde{u}_h^k(\Gamma_h^1)$ by the explicit scheme (5.3); then we get the subdomain solutions $u_h^k(\Omega_{\xi\eta})$ by applying an elliptic solver such as SOR and PCG to implicit scheme (5.4). Once the subdomain solutions are available, the interface solutions $u_h^k(\Gamma_h^1)$ are obtained explicitly by the implicit scheme (5.5) for $1 \le j \le N-1$ and $1 \le \alpha \le p_1 - 1$:

$$\begin{aligned}\left(1 + 2r\bar{a}^{12}_{i_\alpha,j}\right) u^k_{i_\alpha,j} = u^{k-1}_{i_\alpha,j} &+ r\left(a^1_{i_\alpha-\frac{1}{2},j} u^k_{i_\alpha-1,j} + a^1_{i_\alpha+\frac{1}{2},j} u^k_{i_\alpha+1,j}\right) \\ &+ r\left(a^2_{i_\alpha,j-\frac{1}{2}} u^k_{i_\alpha,j-1} + a^2_{i_\alpha,j+\frac{1}{2}} u^k_{i_\alpha,j+1}\right) + \tau f^k_{i_\alpha,j}.\end{aligned}$$

Being similar to the 1-D case, the CEIDD-ZI1 method maintains the efficiency in computation and communication of the 2-D EIDD algorithms.

As for the CEIDD-ZI2 method, we assume that the crossover interfaces $\Gamma^1_{h,\alpha}$, $\Gamma^2_{h,\beta}$ satisfy the intersectant conditions

$$\mathrm{mod}(n_\beta, 2) = \mathrm{mod}(m_\alpha, 2), \quad 1 \le \alpha \le p_1 - 1, 1 \le \beta \le p_2 - 1, \tag{5.8}$$

so any two intersecting zigzag-line interfaces have two common points: $(m_\alpha h, n_\beta h)$ and $(m_\alpha h + h, n_\beta h + h)$ if m_α is even, or $(m_\alpha h, n_\beta h + h)$ and $(m_\alpha h + h, n_\beta h)$ if m_α is odd. Figure 5.2 shows the former case for $\alpha = \beta = 1$. If the intersectant conditions (5.8) are true, the implicit Euler scheme (5.5) applied at the intersecting interfaces Γ_h computes interface solutions explicitly and the data communication is localized. Thus, the CEIDD-ZI2 method also maintains the efficiency in computation and communication of the 2-D EIDD algorithms.

On the contrary, if the intersectant conditions (5.8) do not hold, any two intersecting boundaries have no common grid points; see Figure 5.3 for a simple case. The fully implicit corrector scheme (5.5) introduces some points, such as the "$\circ$" points in Figure 5.3, where the interface solutions should be updated by an elliptic solver on one processor. So, compared with the above case shown in Figure 5.2, more operations of data transferring will be needed for each time level. Then an extra parallel time for data transfers and interface correction is added to the 2-D EIDD algorithms. Thus we arrive at the following conclusion.

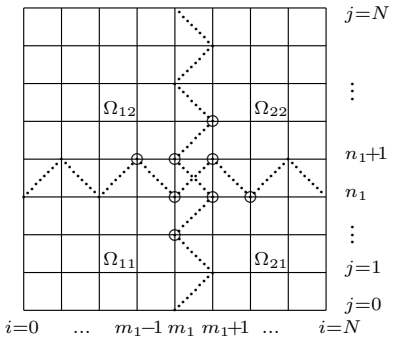


FIG. 5.3. *An elliptic solver is necessary to compute interface solutions on the intersecting zigzag-line interfaces which do not satisfy the intersectant conditions* (5.8).

PROPOSITION 5.1. *Under the intersectant conditions* (5.8)*, the implicit correction step of the CEIDD-ZI algorithm on general $p_1 \times p_2$ subdomains adds zero communication and negligible computation cost to the* 2*-D EIDD algorithms.*

The localization of data communication is important for the two parallel CEIDD-ZI procedures; however, we will prefer CEIDD-ZI2 to CEIDD-ZI1 for large-scale parallel simulations on distributed memory computers because the former has a better efficiency in communication. Given parallel machine number, namely, fixed p, the total data communication cost T_{comm} per time step is an increasing function of the total "length" of the interior boundaries. To carry out a quantitative analysis, we assume that two data transferring operations (one occurs in step 1 to compute subdomain solutions, and the other occurs in step 2 to obtain interface corrector and predictor values) are carried out by $p-1$ processors simultaneously with almost equal load. Suppose that $p = p_1 \times p_2$ and $p_1 \geq p_2 > 1$; the total communication time at each time step satisfies

$$T_{comm,1} = 2\lambda \frac{(p-1)N}{p-1} + \mu$$

for the CEIDD-ZI1 method, and

$$T_{comm,2} = 2\lambda \frac{(p_1-1)N + (p_2-1)N}{p-1} + \mu$$

for the CEIDD-ZI2 method, where λ is some system-dependent data transferring parameter and μ is the communication startup overload. For an efficiency comparison in communication, we define a ratio $\rho = \rho(p_2)$,

$$\rho(p_2) = \frac{T_{comm,1} - T_{comm,2}}{T_{comm,1}} = \frac{2\lambda N}{(2\lambda N + \mu)(p-1)} \left(p - \frac{p}{p_2} - p_2 + 1 \right). \tag{5.9}$$

Obviously, $\rho(1) = 0$ and $\rho(p_2)$ is an increasing function for $1 < p_2 \leq \sqrt{p}$ so that we obtain the following results.

PROPOSITION 5.2. *For given $p = p_1 \times p_2$ processors, the communication cost at each time level of the CEIDD-ZI*2 *algorithm is less than that of the CEIDD-ZI*1 *procedure, and for $1 < p_2 \leq \sqrt{p}$, the larger the value of p_2, the less the cost.*

Similar to the 1-D CEIDD method, if the subdomain solutions are obtained explicitly, the parallel CEIDD-ZI procedures also degenerate to a 2-D CEH scheme. As

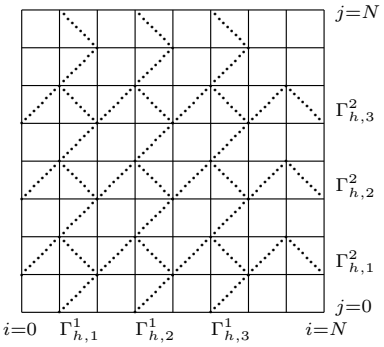


FIG. 5.4. *CEH scheme on* 4×4 *subdomains.*

the special version of the CEIDD-ZI algorithms, the scheme would always be constructed on $\frac{N}{2} \times \frac{N}{2}$ subdomains; see Figure 5.4 for an instance. The unconditional stability of the CEH scheme will be verified in the next section. On the other hand, the CEH method always can be described as the following point-related schemes:

$$\tilde{u}_{ij}^k = u_{ij}^{k-1} + \tau \Delta_h u_{ij}^{k-1} + \tau f_{ij}^{k-1}, \qquad (i+j) \in \text{odd}, \tag{5.10}$$

$$\begin{aligned}(1 + 2\bar{a}_{ij}^{12} r) u_{ij}^k &= u_{ij}^{k-1} + r(a^1_{i-\frac{1}{2},j} \tilde{u}^k_{i-1,j} + a^1_{i+\frac{1}{2},j} \tilde{u}^k_{i+1,j}) \\ &\quad + r(a^2_{i,j-\frac{1}{2}} \tilde{u}^k_{i,j-1} + a^2_{i,j+\frac{1}{2}} \tilde{u}^k_{i,j+1}) + \tau f_{ij}^k, \quad (i+j) \in \text{even},\end{aligned} \tag{5.11}$$

$$\begin{aligned}(1 + 2\bar{a}_{ij}^{12} r) u_{ij}^k &= u_{ij}^{k-1} + r(a^1_{i-\frac{1}{2},j} u^k_{i-1,j} + a^1_{i+\frac{1}{2},j} u^k_{i+1,j}) \\ &\quad + (a^2_{i,j-\frac{1}{2}} u^k_{i,j-1} + a^2_{i,j+\frac{1}{2}} u^k_{i,j+1}) + \tau f_{ij}^k, \quad (i+j) \in \text{odd},\end{aligned} \tag{5.12}$$

where $1 \le i, j \le N-1$.

6. Stability and convergence of CEIDD-ZI algorithms. In this section, H^1 and L^2 prior estimations for difference solutions of the CEIDD-ZI methods are derived by analogy with the 1-D case in section 4.

LEMMA 6.1 (prior estimation). *Let* $v_h^k = \{v_{ij}^k | 0 \le i, j \le N, 0 \le k \le J\}$ *satisfy the CEIDD-ZI*1 *algorithm on* $p_1 \times 1$ *subdomains decoupled by the zigzag-line interfaces* Γ_h^1*:*

$$\frac{\tilde{v}_{ij}^k - v_{ij}^{k-1}}{\tau} = \Delta_h v_{ij}^{k-1} + \tilde{g}_{ij}^{k-1} \quad on \;\; \Gamma_h^1, \quad 1 \le k \le J, \tag{6.1}$$

$$\begin{cases} \partial_t v_{ij}^k = \Delta_h v_{ij}^k + g_{ij}^k & on \;\; \Omega_{\xi,1} \;\; (\xi = 1, 2, \ldots, p_1), \\ v_{ij}^k = \tilde{v}_{ij}^k & on \;\; \Gamma_h^1 \end{cases} \quad 1 \le k \le J, \tag{6.2}$$

$$\partial_t v_{ij}^k = \Delta_h v_{ij}^k + g_{ij}^k \qquad on \;\; \Gamma_h^1, \quad 0 \le k \le J, \tag{6.3}$$

$$v_{ij}^0 = \phi_{ij} \qquad on \;\; \Omega_h, \quad 1 \le k \le J, \tag{6.4}$$

$$v_{ij}^k = 0 \qquad on \;\; \partial\Omega_h, \quad 0 \le k \le J. \tag{6.5}$$

Then the inequalities

$$E_1^k \le \frac{3}{2} e^{\frac{3}{2} k\tau} \left[E_1^0 + \tau \sum_{l=1}^{k} \left(\|g^l\|^2 + |||g^l|||_{I,1}^2 \right) \right], \tag{6.6}$$

$$E_2^k \le \frac{3}{2}(J+1)\exp\left(\frac{3}{2}(1+T)k\tau\right)E_2^0 + \frac{3T}{2}\exp\left(\frac{3}{2}(1+T)k\tau\right)\left[4\tau\sum_{l=0}^{k}E_1^l + \tau\sum_{l=1}^{k}(\|g^l\|^2 + |||g^l|||_{I,2}^2)\right] \tag{6.7}$$

hold for $0<\tau\le 1/3$, $1\le k\le J$, *where* E_1^k, E_2^k, $|||g^k|||_{I,1}$, *and* $|||g^k|||_{I,2}$ *are defined by* (A.1)–(A.4), *respectively, in the case of* $p_2=1$.

Proof. (Estimation on E_1^k.) For the derivation of the H^1 estimation, as was done for Lemma 4.1, this proof is written out only for two subdomains, $\Omega_{\xi,1}$ ($\xi=1,2$), with a zigzag interface $\Gamma_{h,1}^1$ defined by (5.1) for $1\le m_1\le N-3$. For completeness, two cases of m_1 are discussed: one is $2\le m_1\le N-3$, the other is $m_1=1$. Denoting $w_{i_1,j}^k=(\tilde v_{i_1j}^k-v_{i_1j}^k)/h^2$, we derive that

$$h^2 w_{i_1j}^k=(\tilde v_{i_1j}^k-v_{i_1j}^k)=-\tau^2(\partial_t\Delta_h v_{i_1j}^k)-\tau(g_{i_1j}^k-\tilde g_{i_1j}^{k-1}),\qquad 1\le j\le N-1,$$

from the interface schemes (6.1) and (6.3). For convenience, the notation $w_{i_1,0}^k$ and $w_{i_1,N}^k$ is also used with complementary definitions $w_{i_1,0}^k=w_{i_1,N}^k=0$. For $j=1,2,\dots,N-1$, we denote $j\in$ odd to indicate integer j is odd and $j\in$ even to indicate j is even.

(a) If $2\le m_1\le N-3$, the schemes (6.2)–(6.3) would read

$$\partial_t v_{ij}^k=\Delta_h v_{ij}^k+g_{ij}^k,\qquad 0<i<m_1-1,\ 0<j<N, \tag{6.8}$$

$$\partial_t v_{m_1-1,j}^k=\Delta_h v_{m_1-1,j}^k+g_{m_1-1,j}^k,\qquad j\in\text{odd}, \tag{6.9}$$

$$\partial_t v_{m_1-1,j}^k=\Delta_h v_{m_1-1,j}^k+g_{m_1-1,j}^k+a_{m_1-\frac12,j}^1 w_{m_1,j}^k,\qquad j\in\text{even}, \tag{6.10}$$

$$\partial_t v_{m_1j}^k=\Delta_h v_{m_1j}^k+g_{m_1j}^k+a_{(m_1+1)-\frac12,j}^1 w_{m_1+1,j}^k + a_{m_1,j-\frac12}^2 w_{m_1,j-1}^k+a_{m_1,j+\frac12}^2 w_{m_1,j+1}^k,\qquad j\in\text{odd}, \tag{6.11}$$

$$\partial_t v_{m_1j}^k=\Delta_h v_{m_1j}^k+g_{m_1j}^k,\qquad j\in\text{even}, \tag{6.12}$$

$$\partial_t v_{m_1+1,j}^k=\Delta_h v_{m_1+1,j}^k+g_{m_1+1,j}^k,\qquad j\in\text{odd}, \tag{6.13}$$

$$\partial_t v_{m_1+1,j}^k=\Delta_h v_{m_1+1,j}^k+g_{m_1+1,j}^k+a_{m_1+\frac12,j}^1 w_{m_1j}^k + a_{m_1+1,j-\frac12}^2 w_{m_1+1,j-1}^k+a_{m_1+1,j+\frac12}^2 w_{m_1+1,j+1}^k,\qquad j\in\text{even}, \tag{6.14}$$

$$\partial_t v_{m_1+2,j}^k=\Delta_h v_{m_1+2,j}^k+g_{m_1+2,j}^k+a_{(m_1+1)+\frac12,j}^1 w_{m_1+1,j}^k,\qquad j\in\text{odd}, \tag{6.15}$$

$$\partial_t v_{m_1+2,j}^k=\Delta_h v_{m_1+2,j}^k+g_{m_1+2,j}^k,\qquad j\in\text{even}, \tag{6.16}$$

$$\partial_t v_{ij}^k=\Delta_h v_{ij}^k+g_{ij}^k,\qquad m_1+2<i<N,\ 0<j<N. \tag{6.17}$$

From (6.8)–(6.17), it is easy to obtain the following equality with regard to the inner product $\langle\cdot,\cdot\rangle$:

$$2\langle\partial_t v^k,\partial_t v^k\rangle=2\langle\partial_t v^k,\Delta_h v^k\rangle+2\langle\partial_t v^k,g^k\rangle+\sum_{j=1}^{N-1}Q_{1i_1j}^k, \tag{6.18}$$

with $\sum_{j=1}^{N-1}Q_{1i_1j}^k=S_1+S_2$, where

$$S_1=2h^2\sum_{j\in even}w_{m_1,j}^k\big(a_{m_1-\frac12,j}^1\partial_t v_{m_1-1,j}^k+a_{m_1+\frac12,j}^1\partial_t v_{m_1+1,j}^k\big) + 2h^2\sum_{j\in odd}w_{m_1+1,j}^k\big(a_{(m_1+1)-\frac12,j}^1\partial_t v_{m_1,j}^k+a_{(m_1+1)+\frac12,j}^1\partial_t v_{m_1+2,j}^k\big),$$

$$S_2 = 2h^2 \sum_{j\in odd} \big(a^2_{m_1,j-\frac{1}{2}} w^k_{m_1,j-1} + a^2_{m_1,j+\frac{1}{2}} w^k_{m_1,j+1}\big)\partial_t v^k_{m_1 j} + 2h^2 \sum_{j\in even} \big(a^2_{m_1+1,j-\frac{1}{2}} w^k_{m_1+1,j-1} + a^2_{m_1+1,j+\frac{1}{2}} w^k_{m_1+1,j+1}\big)\partial_t v^k_{m_1+1,j}\ .$$

Recalling that $v^k_{i_1,0} = v^k_{i_1,N} = 0$ and $w^k_{i_1,0} = w^k_{i_1,N} = 0$, we transform S_2 into

$$S_2 = 2h^2 \sum_{j\in even} w^k_{m_1,j}\big(a^2_{m_1,j-\frac{1}{2}} \partial_t v^k_{m_1,j-1} + a^2_{m_1,j+\frac{1}{2}} \partial_t v^k_{m_1,j+1}\big) + 2h^2 \sum_{j\in odd} w^k_{m_1+1,j}\big(a^2_{m_1+1,j-\frac{1}{2}} \partial_t v^k_{m_1+1,j-1} + a^2_{m_1+1,j+\frac{1}{2}} \partial_t v^k_{m_1+1,j+1}\big).$$

Due to the definitions of $\Gamma^1_{h,1}$, $\Delta_h v^k_{ij}$, and $\bar a^{12}_{i_1,j}$, it holds that

$$\sum_{j=1}^{N-1} Q^k_{1i_1 j} = 2h^2 \sum_{j=1}^{N-1} w^k_{i_1,j}\big(a^1_{i_1-\frac{1}{2},j} \partial_t v^k_{i_1-1,j} + a^1_{i_1+\frac{1}{2},j} \partial_t v^k_{i_1+1,j}\big) + 2h^2 \sum_{j=1}^{N-1} w^k_{i_1,j}\big(a^2_{i_1,j-\frac{1}{2}} \partial_t v^k_{i_1,j-1} + a^2_{i_1,j+\frac{1}{2}} \partial_t v^k_{i_1,j+1}\big)$$

or, in a simpler form,

$$\sum_{j=1}^{N-1} Q^k_{1i_1 j} = 2\sum_{j=1}^{N-1} (\tilde v^k_{i_1 j} - v^k_{i_1 j})\left[h^2 \partial_t(\Delta_h v^k_{i_1 j}) + 2\bar a^{12}_{i_1 j} \partial_t v^k_{i_1 j}\right].$$

We observe that the interface term $\sum_{j=1}^N Q^k_{1i_1 j}$ is quite similar to $Q^k_{1i_1}$ defined in the proof of Lemma 4.1 by overlooking the summation for j. Therefore, we would deal with the term $\sum_{j=1}^N Q^k_{1i_1 j}$ in the same way and obtain

$$\sum_{j=1}^{N-1} Q^k_{1i_1 j} \le -2\tau^2 \sum_{j=1}^{N-1} \bar a^{12}_{i_1 j} \partial_t[(\Delta_h v^k_{i_1 j})^2] + 2\tau^2 \sum_{j=1}^{N-1} \bar a^{12}_{i_1 j} (\Delta_h v^k_{i_1 j})^2 + |||g^k|||^2_{I,1}. \tag{6.19}$$

Now inserting (6.19) into (6.18), we have

$$\partial_t E^k_1 \le E^k_1 + \|g^k\|^2 + |||g^k|||^2_{I,1}, \qquad 1 \le k \le J,$$

which implies (6.6) for $2 \le m_1 \le N-3$ owing to Lemma 2.3.

(b) If $m_1 = 1$, the schemes (6.8)–(6.10) are not needed and could be cleared away; however, the interface term $\sum_{j=1}^N Q^k_{1i_1 j}$ will have its own form because the term $a^1_{\frac{1}{2},j} \partial_t v^k_{0,j}$ of S_1 is a zero-valued term. Thus we can treat the interface term in the same way, and then (6.6) is also true of $m_1 = 1$.

(Estimation on E^k_2.) Similarly, for the derivation of the L^2 estimation, we could get the following equality:

$$2\langle v^k, \partial_t v^k\rangle = 2\langle v^k, \Delta_h v^k\rangle + 2\langle v^k, g^k\rangle + \sum_{j=1}^{N-1} Q^k_{2i_1 j},$$

where the interface term reads

$$\sum_{j=1}^{N-1} Q_{2i_1j}^k = 2\sum_{j=1}^{N-1}(\tilde{v}_{i_1j}^k - v_{i_1j}^k)\left(h^2\Delta_h v_{i_1,j}^k + 2\bar{a}_{i_1j}^{12} v_{i_1j}^k\right),$$

which is similar to $Q_{2i_1}^k$ defined in Lemma 4.2. Thus, we could obtain that

$$\begin{aligned}\sum_{j=1}^{N-1} Q_{2i_1j}^1 \le & -\tau^2h^2\sum_{j=1}^{N-1}\partial_t(\Delta_h v_{i_1j}^1)^2 - 2\tau\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}\partial_t(v_{i_1j}^1)^2\\ &+\tau^2h^2\sum_{j=1}^{N-1}(\Delta_h v_{i_1j}^1)^2 + 2\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(v_{i_1}^0)^2 + 2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(v_{i_1j}^0)^2\\ &+2\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(\Delta_h v_{i_1j}^1)^2 + 4\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(\Delta_h v_{i_1j}^0)^2 + |||g^1|||_{I,2}^2\end{aligned}$$

and, for $k \ge 2$,

$$\begin{aligned}\sum_{j=1}^{N-1} Q_{2i_1j}^k \le & -\tau^2h^2\sum_{j=1}^{N-1}\partial_t(\Delta_h v_{i_1j}^k)^2 - 2\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}\partial_t\left[\partial_t(v_{i_1j}^k)^2\right]\\ &+\tau^2h^2\sum_{j=1}^{N-1}(\Delta_h v_{i_1j}^k)^2 + 2\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(v_{i_1j}^{k-1})^2\\ &+2\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(\Delta_h v_{i_1j}^k)^2 + 6\tau^2\sum_{j=1}^{N-1}\bar{a}_{i_1j}^{12}(\Delta_h v_{i_1j}^{k-1})^2 + |||g^k|||_{I,2}^2.\end{aligned}$$

To complete the proof, one should present the similar arguments for m_1 described above and follow the proof of Lemma 4.2. ∎

LEMMA 6.2 (prior estimation). *Let $v_h^k = \{v_{ij}^k | 0 \le i,j \le N, 0 \le k \le J\}$ satisfy the CEIDD-ZI2 algorithm on $p_1 \times p_2$ subdomains decoupled by the zigzag-line interfaces $\Gamma_h = \Gamma_h^1 \cup \Gamma_h^2$, viz.,*

$$\frac{\tilde{v}_{ij}^k - v_{ij}^{k-1}}{\tau} = \Delta_h v_{ij}^{k-1} + \tilde{g}_{ij}^{k-1} \quad on\ \ \Gamma_h, \quad 1 \le k \le J, \tag{6.20}$$

$$\begin{cases}\partial_t v_{ij}^k = \Delta_h v_{ij}^k + g_{ij}^k & on\ \ \Omega_{\xi,\eta}\ \begin{pmatrix}\xi = 1,2,\ldots,p_1\\ \eta = 1,2,\ldots,p_2\end{pmatrix},\\ v_{ij}^k = \tilde{v}_{ij}^k & on\ \ \Gamma_h\end{cases} \quad 1 \le k \le J, \tag{6.21}$$

$$\partial_t v_{ij}^k = \Delta_h v_{ij}^k + g_{ij}^k \quad on\ \ \Gamma_h, \quad 0 \le k \le J, \tag{6.22}$$

$$v_{ij}^0 = \phi_{ij} \quad on\ \ \Omega_h, \quad 1 \le k \le J, \tag{6.23}$$

$$v_{ij}^k = 0 \quad on\ \ \partial\Omega_h, \quad 0 \le k \le J. \tag{6.24}$$

Then the inequalities

$$E_1^k \le \frac{3}{2}e^{\frac{3}{2}k\tau}\left[E_1^0 + \tau\sum_{l=1}^{k}\left(\|g^l\|^2 + |||g^l|||_{I,1}^2\right)\right], \tag{6.25}$$

$$\begin{aligned}E_2^k \le & \frac{3}{2}(J+1)\exp\left(\frac{3}{2}(1+T)k\tau\right)E_2^0\\ &+\frac{3T}{2}\exp\left(\frac{3}{2}(1+T)k\tau\right)\left[4\tau\sum_{l=0}^{k}E_1^l + \tau\sum_{l=1}^{k}(\|g^l\|^2 + |||g^l|||_{I,2}^2)\right]\end{aligned} \tag{6.26}$$

hold for $0 < \tau \le 1/3$, $1 \le k \le J$, *where* E_1^k, E_2^k, $|||g^k|||_{I,1}$, *and* $|||g^k|||_{I,2}$ *are defined by* (A.1)–(A.4), *respectively.*

Proof. As far as the general case of $p_1 \times p_2$ subdomains, we can discuss the 2×2 subdomains case shown in Figure 5.2. Without any essential difficulty, the schemes (6.21)–(6.22) can read as some similar point-related schemes such as (6.8)–(6.17); however, the lengthy descriptions are messier so they are omitted here. For the H^1 estimation, we get the interface term in the form

$$\sum_{j=1}^{N-1} Q_{1i_1j}^k + \sum_{\substack{i=1\\ i\neq i_1}}^{N-1} Q_{1ij_1}^k = 2\sum_{j=1}^{N-1} (\tilde{v}_{i_1j}^k - v_{i_1j}^k)\left[h^2\partial_t(\Delta_h v_{i_1j}^k) + 2\bar{a}_{i_1j}^{12}\partial_t v_{i_1j}^k\right] + 2\sum_{\substack{i=1\\ i\neq i_1}}^{N-1} (\tilde{v}_{ij_1}^k - v_{ij_1}^k)\left[h^2\partial_t(\Delta_h v_{ij_1}^k) + 2\bar{a}_{ij_1}^{12}\partial_t v_{ij_1}^k\right].$$

Similarly, for the L^2 estimation, the interface term reads

$$\sum_{j=1}^{N-1} Q_{2i_1j}^k + \sum_{\substack{i=1\\ i\neq i_1}}^{N-1} Q_{2ij_1}^k = 2\sum_{j=1}^{N-1} (\tilde{v}_{i_1j}^k - v_{i_1j}^k)\left(h^2\Delta_h v_{i_1j}^k + 2\bar{a}_{i_1j}^{12} v_{i_1j}^k\right) + 2\sum_{\substack{i=1\\ i\neq i_1}}^{N-1} (\tilde{v}_{ij_1}^k - v_{ij_1}^k)\left(h^2\Delta_h v_{ij_1}^k + 2\bar{a}_{ij_1}^{12} v_{ij_1}^k\right).$$

To complete the proof, one could follow the proof of Lemma 6.1. ∎

We note that Lemma 6.2 is true of the CEH scheme with $p_1 = p_2 = N/2$. The tiny difference is that the scheme (6.21), applied at the interior points of each subdomain, always reads as

$$\partial_t v_{ij}^k = \Delta_h v_{ij}^k + g_{ij}^k + a_{i-\frac{1}{2},j}^1 w_{i-1,j}^k + a_{i+\frac{1}{2},j}^1 w_{i+1,j}^k + a_{i,j-\frac{1}{2}}^2 w_{i,j-1}^k + a_{i,j+\frac{1}{2}}^2 w_{i,j+1}^k,$$

in which the four adjoining points of (ih, jh) belong to $\Gamma_h \cup \partial\Omega_h$, and the mesh function $w_{ij}^k = (\tilde{v}_{ij}^k - v_{ij}^k)/h^2$ is defined at interface points together with $w_h^k(\partial\Omega_h) \equiv 0$. Correspondingly, in the proof of Lemma 6.2, the interface terms will be replaced by

$$\sum_{\alpha=1}^{N/2-1}\sum_{j=1}^{N-1} Q_{1i_\alpha j}^k + \sum_{\beta=1}^{N/2-1}\sum_{\substack{i=1\\ i\neq i_\alpha}}^{N-1} Q_{1ij_\beta}^k \quad \text{and} \quad \sum_{\alpha=1}^{N/2-1}\sum_{j=1}^{N-1} Q_{2i_\alpha j}^k + \sum_{\beta=1}^{N/2-1}\sum_{\substack{i=1\\ i\neq i_\alpha}}^{N-1} Q_{2ij_\beta}^k,$$

but Q_{1ij}^k and Q_{2ij}^k have their own forms.

Analogous to the analysis for the 1-D CEIDD algorithm in section 4, it is easy to obtain the results on stability and convergence from Lemmas 6.1 and 6.2.

THEOREM 6.3 (stability). *Under the reasonable assumption* (1.4), *the CEIDD-ZI algorithm* (5.3)–(5.7) *on* $p_1 \times p_2$ *subdomains is unconditionally stable with respect to the* H^1 *seminorm and the* L^2 *norm.*

THEOREM 6.4 (convergence). *If the solution of problem* (1.1)–(1.3) *is sufficiently smooth, spacing* h *is sufficiently small, and time step size* $\tau = \mathcal{O}\left((p_1+p_2)^{-1/2}h^{1/2+\epsilon}\right)$ *for* $\epsilon > 0$, *the numerical solution of the CEIDD-ZI algorithm* (5.3)–(5.7) *converges to the exact solution of* (1.1)–(1.3) *with an order of* $\mathcal{O}(\sqrt{1+(p_1+p_2-2)h^{-1}}(\tau+h^2))$ *in the* H^1 *seminorm and the* L^2 *norm in the sense that*

$$|e^k|_{a,1} \le c_7\sqrt{1+(p_1+p_2-2)h^{-1}}(\tau+h^2), \quad 1 \le k \le J,$$

and

$$\sqrt{\tau\sum_{l=1}^{k}\|e^l\|^2}\le c_8\sqrt{1+(p_1+p_2-2)h^{-1}}(\tau+h^2),\quad 1\le k\le J.$$

Again, for the coarsely granular parallel algorithm in the case of $2<p\ll N^2$, the accuracy of the above CEIDD-ZI algorithms is $\mathcal{O}(h^{-1/2}(\tau+h^2))$ with $\tau=\mathcal{O}(h^{1/2+\epsilon})$, just about the accuracy of the 1-D CEIDD method. However, for the CEH scheme with $p_1=p_2=N/2$, the accuracy decreases to $\mathcal{O}(h^{-1}(\tau+h^2))$ with time step size $\tau=\mathcal{O}(h^{1+\epsilon})$. From the formulations of $|||g^k|||_{I,1}$ and $|||g^k|||_{I,2}$ defined in the appendix, we could conclude that the accumulated truncation errors on the zigzag-line interior boundaries affect the global accuracy of our CEIDD-ZI algorithms.

On the other hand, if the truncation errors on Γ_h satisfy $|r_{ij}^k-\tilde r_{ij}^{k-1}|\le c_1\tau h^2$ and $|r_{ij}^{k-1}-\tilde r_{ij}^{k-1}|\le c_1\tau h^2$, an improved accuracy of CEIDD-ZI algorithms would be obtained by our prior estimations (6.25) and (6.26), and the convergence rate reaches $\mathcal{O}(\tau+h^2)$ even for the CEH scheme (cf. Example 1).

In Theorem 6.4, we notice that, for given $p=p_1\times p_2$ processors, the error bounds arrive at the minimum value with $p_1=p_2$. It means that solutions of different CEIDD-ZI algorithms have some tiny differences in numerical precision although they have the same rate of convergence. Therefore, in point of numerical precision, we may also prefer the CEIDD-ZI2 method to the CEIDD-ZI1 method for large-scale parallel computations, as the following remark states.

Remark 6.5. For given $p=p_1\times p_2$ processors, the numerical solutions of the CEIDD-ZI2 algorithm may be more precise than those of the CEIDD-ZI1 procedure, and for $1<p_2\le\sqrt{p}$, the larger the value of p_2, the better the precision may be.

7. Numerical experiments. In this section, we present some experimental results of the CEIDD-ZI procedures for four 2-D model problems. With the different choices of the initial and boundary conditions (1.2)–(1.3), the first three examples are heat problems,

$$u_t=\Delta u+f(x,y,t),\quad (x,y)\in(0,1)^2,\ t\in(0,T], \tag{7.1}$$

where three different outer-forced terms f are chosen so that we have the following.

Example 1. $u(x,y,t)\equiv u_1(x,y,t)=100tx^3(1-x)^2\cos(2\pi y)$.

Example 2. $u(x,y,t)\equiv u_2(x,y,t)=e^{-2t}\sin(x+y)$.

Example 3. $u(x,y,t)\equiv u_3(x,y,t)=te^{2t}\cos(x+y)$.

The fourth example is a convection-diffusion problem,

$$u_t=\Delta u+30u_x-20u_y+f(x,y,t),\quad (x,y)\in(0,1)^2,\ t\in(0,T], \tag{7.2}$$

where f is chosen so that we have the following.

Example 4. $u(x,y,t)\equiv u_4(x,y,t)=te^{2t}\cos(x+y)$.

For the numerical approximations of each model problem, five different scenarios were considered:

(i) the backward Euler scheme (listed as "BEuler" in the tables) inside the entire nonpartitioned domain; the CEIDD-ZI procedure on

(ii) 2×1 subdomains with 1 zigzag-line interface (5.1) for $m_1=h^{-1}/2$;

(iii) 4×1 subdomains with 3 zigzag-line interfaces (5.1) for $m_\alpha=\alpha h^{-1}/4\,(\alpha=1,2,3)$;

TABLE 7.1
Stability: $u(x,y,t) = u_1(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 1$ with spatial step $h = 1/64$.

τ	1/400	1/200	1/100	1/50	1/25
BEuler	8.01e-04	8.03e-04	8.06e-04	8.12e-04	8.25e-04
2×1 subdomains	8.01e-04	8.02e-04	8.05e-04	8.12e-04	7.94e-04
4×1 subdomains	8.01e-04	8.02e-04	8.05e-04	8.20e-04	7.21e-04
2×2 subdomains	8.01e-04	8.02e-04	8.05e-04	8.25e-04	6.97e-04
CEH	8.01e-04	8.01e-04	8.10e-04	5.71e-04	2.95e-04

TABLE 7.2
Stability: $u(x,y,t) = u_2(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 1$ with spatial step $h = 1/64$.

τ	1/400	1/200	1/100	1/50	1/25
BEuler	7.87e-05	1.56e-04	3.11e-04	6.17e-04	1.22e-03
2×1 subdomains	9.30e-05	5.22e-04	2.50e-03	1.17e-02	5.37e-02
4×1 subdomains	3.05e-04	1.40e-03	6.26e-03	2.91e-02	1.22e-01
2×2 subdomains	2.40e-04	1.24e-03	5.16e-03	2.42e-02	9.83e-02
CEH	3.25e-03	1.45e-02	6.78e-02	2.18e-01	3.42e-01

TABLE 7.3
Stability: $u(x,y,t) = u_3(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 1$ with spatial step $h = 1/64$.

τ	1/400	1/200	1/100	1/50	1/25
BEuler	6.33e-04	1.27e-03	2.54e-03	5.10e-03	1.03e-02
2×1 subdomains	7.36e-04	3.99e-03	1.78e-02	7.06e-02	2.44e-01
4×1 subdomains	2.39e-03	1.07e-02	4.42e-02	1.66e-01	5.01e-01
2×2 subdomains	1.84e-03	8.48e-03	3.53e-02	1.33e-01	3.92e-01
CEH	2.47e-02	9.60e-02	3.44e-01	7.92e-01	1.14e+00

(iv) 2×2 subdomains with 2 intersectant zigzag-line interfaces (5.1) and (5.2) for $m_1 = n_1 = h^{-1}/2$; and
(v) the CEH algorithm on $\frac{N}{2} \times \frac{N}{2}$ subdomains.

We consider the backward Euler scheme on the entire nonpartitioned domain as the benchmark for our comparisons since it is the most stable method, and we consider the CEH scheme as a negative example since the scheme has the worst accuracy among our parallel CEIDD-ZI algorithms. In these runs, stability and convergence are carefully examined in the sense of the time-averaged L^2 error, such as the form in Theorem 6.4,

$$\mathcal{E}_h = \sqrt{\tau \sum_{k=1}^{J} \|u(\cdot, k\tau) - u_h^k\|^2} \quad . \tag{7.3}$$

In Tables 7.1–7.4, the discrete solutions are obtained by using the doubling temporal steps with the minimal size $\tau = 1/400$. The errors of the coarsely granular CEIDD-ZI algorithms remain relatively small; even τ is large relative to the spacing h, although they are relatively large to those of the fully implicit Euler scheme. Experimentally, the data in Tables 7.1–7.3 support the stability results (Theorem 6.3) of the CEIDD-ZI algorithms on mesh ratio r and subdomain partition; and the data in Table 7.4 suggest that our parallel procedures may be stable for convection-diffusion problems.

TABLE 7.4

Stability: $u(x,y,t) = u_4(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 1$ with spatial step $h = 1/64$.

τ	1/400	1/200	1/100	1/50	1/25
BEuler	2.25e-04	4.40e-04	8.70e-04	1.74e-03	3.49e-03
2×1 subdomains	2.37e-04	1.31e-03	5.96e-03	2.48e-02	9.52e-02
4×1 subdomains	7.88e-04	3.59e-03	1.51e-02	6.04e-02	2.15e-01
2×2 subdomains	5.37e-04	2.54e-03	1.09e-02	4.38e-02	1.58e-01
CEH	1.08e-01	3.65e+00	2.62e+00	5.24e-01	1.07e+00

TABLE 7.5

Convergence in h: $u(x,y,t) = u_1(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 0.5$ for fixed $r = 8$ with $\tau = 8h^2$, and variable $r = h^{-1}$ with $\tau = h$.

τ	h	BEuler	2×1 subdomains	4×1 subdomains	2×2 subdomains	CEH
1/32	1/16	4.49e-03	4.43e-03	4.28e-03	4.30e-03	3.82e-03
1/128	1/32	1.08e-03	1.07e-03	1.07e-03	1.07e-03	1.07e-03
1/512	1/64	2.66e-04	2.66e-04	2.66e-04	2.66e-04	2.65e-04
1/2048	1/128	6.63e-05	6.63e-05	6.63e-05	6.63e-05	6.63e-05
Rate		2.03e+00	2.02e+00	2.01e+00	2.01e+00	1.96e+00
1/16	1/16	4.71e-03	3.99e-03	3.38e-03	3.33e-03	2.61e-03
1/32	1/32	1.12e-03	1.04e-03	9.40e-04	9.15e-04	5.75e-04
1/64	1/64	2.72e-04	2.69e-04	2.58e-04	2.55e-04	1.34e-04
1/128	1/128	6.71e-05	6.70e-05	6.68e-05	6.69e-05	3.25e-05
Rate		2.04e+00	1.96e+00	1.88e+00	1.88e+00	2.11e+00

TABLE 7.6

Convergence in h: $u(x,y,t) = u_2(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 0.5$ for fixed mesh ratio $r = 8$ with time step size $\tau = 8h^2$, and variable mesh ratio $r = h^{-1}$ with time step size $\tau = h$.

τ	h	BEuler	2×1 subdomains	4×1 subdomains	2×2 subdomains	CEH
1/32	1/16	8.80e-04	5.44e-03	1.41e-02	1.06e-02	2.97e-02
1/128	1/32	2.24e-04	5.36e-04	1.52e-03	1.19e-03	7.72e-03
1/512	1/64	5.64e-05	4.19e-05	1.57e-04	1.22e-04	1.79e-03
1/2048	1/128	1.41e-05	4.60e-06	1.26e-05	9.01e-06	4.37e-04
Rate		1.99e+00	3.43e+00	3.37e+00	3.39e+00	2.04e+00
1/16	1/16	1.72e-03	2.12e-02	4.31e-02	3.35e-02	6.38e-02
1/32	1/32	8.75e-04	1.29e-02	2.88e-02	2.29e-02	6.54e-02
1/64	1/64	4.42e-04	6.50e-03	1.59e-02	1.32e-02	6.59e-02
1/128	1/128	2.22e-04	3.02e-03	7.56e-03	6.39e-03	6.61e-02
Rate		9.84e-01	9.42e-01	8.39e-01	7.96e-01	−1.65e-02

In Tables 7.5–7.8, the solution u is approximated on the halving grids with the coarsest 16×16 grids. Setting time step size $\tau = \mathcal{O}(h^2)$ or $\tau = \mathcal{O}(h)$, the experimental rate (listed as "Rate" in the tables) of convergence, in h, is computed by observing that $\mathcal{E}_h \approx c_9 h^q$ and doing a least squares fit to determine q.

We observe that the solution error of Example 1 presented in Table 7.5 behaves in the manner of the backward Euler scheme as the grid is refined. The improvement accuracy in h is the result of the improved truncation errors on interfaces and subdomains owing to $u_{2,tt}(x,y,t) \equiv 0$, as mentioned in the previous section.

In other cases, the errors of coarsely granular CEIDD-ZI algorithms are large compared to those of the backward Euler scheme for relatively coarse time-spatial meshes; however, as seen in Tables 7.6–7.8, they compare favorably with the error of the backward Euler scheme as the mesh is further refined. For those test problems,

TABLE 7.7

Convergence in h: $u(x,y,t) = u_3(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 0.5$ for fixed $r = 8$ with $\tau = 8h^2$, and variable $r = h^{-1}$ with $\tau = h$.

τ	h	BEuler	2×1 subdomains	4×1 subdomains	2×2 subdomains	CEH
1/32	1/16	2.03e-03	9.44e-03	2.43e-02	1.76e-02	4.92e-02
1/128	1/32	5.01e-04	1.10e-03	3.14e-03	2.37e-03	1.50e-02
1/512	1/64	1.25e-04	9.47e-05	3.42e-04	2.59e-04	3.82e-03
1/2048	1/128	3.12e-05	1.18e-05	2.79e-05	1.99e-05	9.62e-04
Rate		2.01e+00	3.25e+00	3.25e+00	3.25e+00	1.90e+00
1/16	1/16	4.11e-03	3.17e-02	6.51e-02	4.86e-02	9.90e-02
1/32	1/32	2.03e-03	2.08e-02	4.57e-02	3.51e-02	9.97e-02
1/64	1/64	1.00e-03	1.16e-02	2.76e-02	2.20e-02	9.95e-02
1/128	1/128	5.00e-04	5.83e-03	1.43e-02	1.17e-02	9.92e-02
Rate		1.01e+00	8.17e-01	7.30e-01	6.82e-01	−7.32e-04

TABLE 7.8

Convergence in h: $u(x,y,t) = u_4(x,y,t)$. The table lists the time-averaged L^2 error of the solution at $T = 0.5$ for fixed $r = 8$ with $\tau = 8h^2$, and variable $r = h^{-1}$ with $\tau = h$.

τ	h	BEuler	2×1 subdomains	4×1 subdomains	2×2 subdomains	CEH
1/32	1/16	7.43e-04	3.24e-03	8.59e-03	5.47e-03	2.12e-02
1/128	1/32	1.80e-04	3.67e-04	1.07e-03	7.20e-04	2.66e-02
1/512	1/64	4.47e-05	3.16e-05	1.15e-04	7.68e-05	2.27e-02
1/2048	1/128	1.12e-05	4.76e-06	9.12e-06	6.17e-06	8.56e-04
Rate		2.02e+00	3.18e+00	3.29e+00	3.26e+00	1.41e+00
1/16	1/16	1.48e-03	1.34e-02	3.41e-02	2.19e-02	7.60e-02
1/32	1/32	7.08e-04	7.39e-03	1.80e-02	1.26e-02	8.31e-02
1/64	1/64	3.48e-04	3.91e-03	9.88e-03	7.09e-03	8.63e-02
1/128	1/128	1.73e-04	2.00e-03	4.91e-03	3.59e-03	8.78e-02
Rate		1.03e+00	9.17e-01	9.26e-01	8.65e-01	−6.78e-02

a better rate of convergence than that predicted by Theorem 6.4 is observed; nevertheless, it is mysterious to us.

As far as the numerical precision of solutions, in those experiments, we notice that the 2×1-subdomain approach is always better than the 4-subdomain approaches (4×1 subdomains and 2×2 subdomains), and the 2×2-subdomain is better than the 4×1-subdomain. The former can be explained by Theorem 6.4, and the latter supports Remark 6.5 experimentally.

8. Concluding remarks. The EIDD methods are globally noniterative, nonoverlapping domain decomposition methods, which are computationally and communicationally efficient for each time step; however, they always suffer from either stability- or consistency-related temporal step size restrictions. For 1-D and 2-D heat equations, we developed a class of two-level CEIDD algorithms by adding an implicit correction technique and special zigzag-shaped interfaces to EIDD methods. The proposed parallel CEIDD-ZI algorithms, including their degenerate cases, the CEH schemes, are unconditionally stable (Theorems 4.3 and 6.3) without discarding the advantages of EIDD methods, including good parallelism, the localization of communication, the flexibility of domain partitioning, and the total quantity of computation (Propositions 3.1 and 5.1).

Compared with the SEIDD methods proposed in [11], the CEIDD-ZI methods not only are free from the parallel-time overload of recomputing interface solutions at correction step (Propositions 3.1 and 5.1), but they also increase the flexibility

in domain partitioning by using the noncrossover and crossover interfaces (see, e.g., Figures 5.1 and 5.2). More importantly for parallel simulations on given processors, the flexibility of domain decomposition has another advantage, that is, it reduces the cost of data communication without degrading the order of convergence and the numerical precision of solutions (Proposition 5.2 and Remark 6.5). Compared with the sequential backward Euler method, the coarsely granular CEIDD-ZI methods have an accuracy loss of $h^{-1/2}$ in our analysis; however, they exhibit better accuracy for the model problems in our numerical experiments than that predicted by Theorems 4.4 and 6.4. The point to note is that the algorithms perform well and are promising parallel procedures for solving parabolic problems on coarse-grain parallel machines.

Moreover, the parallel CEIDD-ZI procedures also can be extended to 3-D problems. Given that $2 \le p_m \le N/2\,(m = 1, 2, 3)$, we construct three zigzag-plane interior boundaries,

$$\Gamma^1_{h,\alpha} = \{(i_\alpha h, jh, lh) |\ i_\alpha = m_\alpha + \mathrm{mod}(j + l, 2),\ 3 \le m_\alpha + 2 \le m_{\alpha+1} \le N - 3\},$$

$$\Gamma^2_{h,\beta} = \{(ih, j_\beta h, lh) |\ j_\beta = n_\beta + \mathrm{mod}(i + l, 2),\ 3 \le n_\beta + 2 \le n_{\beta+1} \le N - 3\},$$

$$\Gamma^3_{h,\gamma} = \{(ih, jh, l_\gamma h) |\ l_\gamma = q_\gamma + \mathrm{mod}(i + j, 2),\ 3 \le q_\gamma + 2 \le q_{\gamma+1} \le N - 3\},$$

where $1 \le \alpha \le p_1 - 1$, $1 \le \beta \le p_2 - 1$, and $1 \le \gamma \le p_3 - 1$. Further, we assume that the three zigzag-plane interfaces satisfy the following intersectant conditions:

$$\mathrm{mod}(m_\alpha, 2) = \mathrm{mod}(n_\beta, 2) = \mathrm{mod}(q_\gamma, 2).$$

Denoting that $\Gamma^m_h = \bigcup_{\nu=1}^{p_m - 1} \Gamma^m_{h,\nu}$ and $\Gamma_h = \bigcup_{m=1}^{3} \Gamma^m_h$, the domain Ω_h is decoupled into $p = \prod_{m=1}^{3} p_m$ subdomains by the zigzag-plane interior interfaces Γ_h. Then the 3-D generalizations of our parallel CEIDD-ZI algorithms and their theoretical results are straightforward for 3-D heat equations. In what follows, we can focus on the 2-D problem for further research since the higher-dimensional case can always be considered in a similar way.

Generally, the CEIDD algorithm given in section 1 allows many choices for the shape of interface at step 0, the explicit predictor at step 1, and the subdomain solver at step 2. Different purposes determine different choices at the three steps, and the different choices make different CEIDD algorithms.

To minimize the parallel time cost on interfaces, the zigzag-shaped interfaces are adopted and the corresponding CEIDD-ZI algorithms are developed in this paper. Moreover, to reduce computation cost on each subdomain, some notable operator-splitting techniques based on fractional-step methods [10] could be employed. With this factorization, the resulting factorized CEIDD-ZI procedure becomes completely noniterative, both globally and on each subdomain.

Besides those zigzag-shaped interfaces, a more direct choice is some straight-line interfaces. Provided with possible choices at steps 1 and 2, whether the resulting CEIDD-SI (CEIDD based on straight-line interfaces) procedures are unconditionally stable or not is an interesting problem and still open to us, although some of those algorithms exhibit excellent stability experimentally [11].

It is likely that generalizations can be made to the cases of linear and nonlinear parabolic problems, and with the different space steps on different subdomains for some special purposes. We will consider those generalizations in the future.

Appendix. Some notation for 2-D CEIDD-ZI algorithms. For the 2-D parallel CEIDD-ZI algorithms described in section 5, we introduce the H^1 seminorm energy

$$E_1^k = |v^k|_{a,1}^2 + 2\tau^2 \sum_{\alpha=1}^{p_1-1} \sum_{j=1}^{N-1} \bar{a}_{i_\alpha j}^{12} (\Delta_h v_{i_\alpha j}^k)^2 + 2\tau^2 \sum_{\beta=1}^{p_2-1} \sum_i^* \bar{a}_{ij_\beta}^{12} (\Delta_h v_{ij_\beta}^k)^2 \tag{A.1}$$

and the L^2 norm energy

$$\begin{aligned} E_2^k = \tau \sum_{l=0}^{k} \|v^l\|^2 &+ \sum_{\alpha=1}^{p_1-1} \sum_{j=1}^{N-1} \left[\tau^3 h^2 \sum_{l=0}^{k} (\Delta_h v_{i_\alpha j}^l)^2 + 2\tau^2 \bar{a}_{i_\alpha j}^{12} (v_{i_\alpha j}^k)^2 \right] \\ &+ \sum_{\beta=1}^{p_2-1} \sum_i^* \left[\tau^3 h^2 \sum_{l=0}^{k} (\Delta_h v_{ij_\beta}^l)^2 + 2\tau^2 \bar{a}_{ij_\beta}^{12} (v_{ij_\beta}^k)^2 \right], \end{aligned} \tag{A.2}$$

where the summation $\sum_i^*$ is defined by

$$\sum_i^* = \sum_{\substack{i=1 \\ i \neq i_\alpha}}^{N-1}$$

for $1 \le \alpha \le p_1 - 1$. As for the interior interfaces, we define $|||g^k|||_{I,1}$ as

$$|||g^k|||_{I,1}^2 = \sum_{\alpha=1}^{p_1-1} \sum_{j=1}^{N-1} \mathcal{G}_{1i_\alpha j}^k + \sum_{\beta=1}^{p_2-1} \sum_i^* \mathcal{G}_{1ij_\beta}^k, \tag{A.3}$$

where

$$\mathcal{G}_{1ij}^k = \frac{1}{2}(h^2 + 4\bar{a}_{ij}^{12})(g_{ij}^k - \tilde{g}_{ij}^{k-1})^2 + 2\bar{a}_{ij}^{12}\tau(\tilde{g}_{ij}^{k-1})^2.$$

Similarly, we define $|||g^k|||_{I,2}$ as

$$|||g^k|||_{I,2}^2 = \sum_{\alpha=1}^{p_1-1} \sum_{j=1}^{N-1} \mathcal{G}_{2i_\alpha j}^k + \sum_{\beta=1}^{p_2-1} \sum_i^* \mathcal{G}_{2ij_\beta}^k, \tag{A.4}$$

where

$$\mathcal{G}_{2ij}^1 = 2\bar{a}_{ij}^{12}\tau^2 \left[(g_{ij}^1)^2 + 2(\tilde{g}_{ij}^0)^2 \right] + h^2 (g_{ij}^1 - \tilde{g}_{ij}^0)^2 + 2\bar{a}_{ij}^{12} (\tilde{g}_{ij}^0)^2$$

and, for $k \ge 2$,

$$\begin{aligned} \mathcal{G}_{2ij}^k = {} & 2\bar{a}_{ij}^{12}\tau^2 \left[(g_{ij}^k)^2 + 2(g_{ij}^{k-1})^2 + 2(\tilde{g}_{ij}^{k-1})^2 \right] \\ & + h^2 (g_{ij}^k - \tilde{g}_{ij}^{k-1})^2 + 2\bar{a}_{ij}^{12} (g_{ij}^{k-1} - \tilde{g}_{ij}^{k-1})^2. \end{aligned}$$

Acknowledgment. We would like to express our thanks to two anonymous reviewers whose invaluable critical comments and suggestions helped greatly to improve this article.

REFERENCES

[1] K. Black, *Polynomial collocation using a domain decomposition solution to parabolic PDE's via the penalty method and explicit/implicit time marching*, J. Sci. Comput., 7 (1992), pp. 313–338.

[2] X.-C. Cai, *Additive Schwarz algorithms for parabolic convention-diffusion equations*, Numer. Math., 50 (1991), pp. 41–52.

[3] X.-C. Cai, *Multiplicative Schwarz methods for parabolic problems*, SIAM J. Sci Comput., 15 (1994), pp. 587–603.

[4] C. N. Dawson, Q. Du, and T. F. Dupont, *A finite difference domain decomposition algorithm for numerical solution of the heat equation*, Math. Comp., 57 (1991), pp. 63–71.

[5] C. N. Dawson and T. F. Dupont, *Explicit/implicit, conservative domain decomposition procedures for parabolic problems based on block-centered finite differences*, SIAM J. Numer. Anal., 31 (1994), pp. 1045–1061.

[6] Q. Du, M. Mu, and Z. N. Wu, *Efficient parallel algorithms for parabolic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1469–1487.

[7] A. R. Gourlay, *Hopscotch: A fast second order partial differential equation solver*, J. Inst. Math. Appl., 6 (1970), pp. 375–390.

[8] Y. Kuznetsov, *New algorithms for approximate realization of implicit difference scheme*, Soviet J. Numer. Anal. Math. Modelling, 3 (1988), pp. 99–114.

[9] H. Qian and J. Zhu, *On the efficient parallel algorithm for solving time dependent partial differential equations*, in Proceedings of the 1998 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'1998), CSREA Press, Las Vegas, NV, 1998, pp. 394–401.

[10] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, New York, 1997, pp. 14–15.

[11] Y. Zhuang and X.-H. Sun, *Stabilized explicit-implicit domain decomposition methods for the numerical solution of parabolic equations*, SIAM J. Sci. Comput., 24 (2002), pp. 335–358.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1612–1638

A DOMAIN DECOMPOSITION METHOD BASED ON WEIGHTED INTERIOR PENALTIES FOR ADVECTION-DIFFUSION-REACTION PROBLEMS*

ERIK BURMAN† AND PAOLO ZUNINO‡

Abstract. We propose a domain decomposition method for advection-diffusion-reaction equations based on Nitsche's transmission conditions. The advection-dominated case is stabilized using a continuous interior penalty approach based on the jumps in the gradient over element boundaries. We prove the convergence of the finite element solutions of the discrete problem to the exact solution and propose a parallelizable iterative method. The convergence of the resulting domain decomposition method is proved, and this result holds true uniformly with respect to the diffusion parameter. The numerical scheme that we propose here can thus be applied straightforwardly to diffusion-dominated, advection-dominated, and hyperbolic problems. Some numerical examples are presented in different flow regimes showing the influence of the stabilization parameter on the performance of the iterative method, and we compare our method with some other domain decomposition techniques for advection-diffusion equations.

Key words. advection-diffusion problem, interior penalty, finite element approximation, domain decomposition, iterative methods, discontinuous coefficients

AMS subject classifications. 65N30, 65N12, 35L50, 65N55

DOI. 10.1137/050634736

1. Introduction.

The solution of large computational problems calls for efficient linear solvers. Domain decomposition has proved to be an attractive way to allow for parallel solving of large problems. A formulation for domain decomposition using a generalization of Nitsche's method for weak boundary conditions has been considered, for instance, by Becker, Hansbo, and Stenberg [2, 24] and by Heinrich and Pietsch [16] for the Poisson problem. This formulation was then generalized to the case of advection-diffusion problems by Toselli [26] using SUPG-type stabilization and more recently by Burman [5]. In this last case, continuous interior penalty stabilization was used to make the method stable in all flow regimes. The interior penalty finite element method for continuous approximation spaces was introduced by Douglas and Dupont [12] and analyzed by Burman and Hansbo in [7] and by Burman in [5].

In this paper we will give a detailed analysis of the domain decomposition method using Nitsche's method. In particular we consider a fully parallel iterative splitting method for advection-diffusion-reaction problems, and we prove its convergence. The present result also automatically carries over to discontinuous Galerkin interior penalty formulations of advection-diffusion problems. Overlapping domain decomposition methods for discontinuous Galerkin methods was considered by Lasser and Toselli [19] and substructuring iterative methods for domain decomposition using SUPG-type stabilized continuous approximation was considered by Rapin and Lube [23]. For an overview of results on domain decomposition for nonsymmetric problems, see Quarteroni and Valli [22] or Toselli and Widlund [27] and the references therein.

*Received by the editors June 29, 2005; accepted for publication (in revised form) March 23, 2006; published electronically August 7, 2006.

http://www.siam.org/journals/sinum/44-4/63473.html

†Institut d'analyse et calcul scientifique, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (Erik.Burman@epfl.ch).

‡MOX, Dipartimento di Matematica, Politecnico di Milano, P.zza L. Da Vinci 32, Milano, Italy (paolo.zunino@polimi.it).

The advantages of the method proposed in this paper are to allow for continuous and discontinuous approximation with uniform stability properties with respect to the Péclet number. The discontinuous formulation naturally leads to an iterative method and allows for conservation locally in each subdomain. The continuous approximation, on the other hand, is better suited to handle different diffusive regimes since the interior penalty stabilization parameter is independent of the diffusion parameter. Numerical tests show that the proposed method is robust with respect to varying coefficients. As a model problem we propose the advection-diffusion-reaction equation

$$\begin{cases} \beta \cdot \nabla u + \sigma u - \nabla \cdot \varepsilon \nabla u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded open connected subset of $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$, $d = 2$ or 3 is the space dimension, $\beta \in [W^{1,\infty}(\Omega)]^d$ is a velocity field, $\varepsilon \in L^\infty(\Omega)$, $\varepsilon > 0$, is a diffusion coefficient, and $\sigma > 0$ is the reaction coefficient, $f \in L^2(\Omega)$. The analysis extends to the case $\varepsilon = 0$ in the obvious way if the boundary conditions of the continuous problem are modified and β is such that the problem remains well-posed. We assume that the following coercivity condition holds:

$$\sigma - \frac{1}{2}\nabla \cdot \beta \geq \sigma_0 > 0. \tag{1.2}$$

We define the associated parameter σ_1 by

$$\sigma_1 := \operatorname{ess\,sup}_{x\in\Omega} \frac{|\sigma - \nabla \cdot \beta|^2}{\sigma_0}.$$

Consider a decomposition of the domain Ω into the disjoint subdomains Ω_i, $i = 1, \ldots, N$, with boundaries $\partial\Omega_i$ and with corresponding shape regular disjoint triangulations $\mathcal{T}_{h,i}$, such that $\mathcal{T}_h = \cup_{i=1}^N \mathcal{T}_{h,i} = \cup_{i=1}^N \bar{\Omega}_i = \bar{\Omega}$. Note that we do not suppose that neighboring meshes are conforming over the intersubdomain boundary. The set of interior faces of each triangulation $\mathcal{T}_{h,i}$ will be denoted by $\mathcal{F}_i$. On each triangulation we define a finite element space $V_{h,k,i}$ associated with the subdomain Ω_i,

$$V_{h,k,i} := \{v_h : v_h \in H^1(\Omega_i); v_h|_K \in P_k(K)\ \forall K \in \mathcal{T}_{h,i}\},$$

where $P_k(K)$ denotes the space of polynomials of degree $\leq k$ on K and we let $V_h = \sum_{i=1}^N V_{h,k,i}$. For every function $v_h \in V_h$ we introduce the restriction to subdomain Ω_i, $v_{h,i} = v_h|_{\Omega_i}$. To each subdomain boundary we associate the outward-oriented normal n_i. We will always assume that the solution is sufficiently smooth, i.e., $u \in H^1(\Omega) \cap (\cup_{i=1}^N H^2(\Omega_i))$, and we will assume (weak) continuity of fluxes between subdomains. Typically the diffusion parameter ε may be discontinuous over some subdomain interface, provided the interface is smooth. Let h_K denote the diameter of an element K, and ϱ_K the radius of the largest inscribed ball in K. We henceforth assume that for all meshes $\mathcal{T}_{h,i}$ there holds

$$c_{\mathcal{T}} \leq \max_{K\in\mathcal{T}_{h,i}} \frac{h_K}{\varrho_K} \tag{1.3}$$

with the same positive parameter $c_{\mathcal{T}}$. We introduce a mesh parameter function $\tilde{h}(x)|_K = h_K$ and let $h = \max_{K\in\mathcal{T}_{h,i}} h_K$. Moreover we shall assume that there exists

a constant $\rho > 1$ such that for all elements K in $\mathcal{T}_{h,i}$, $i = 1, \ldots, N$, we have

$$\max_{K' \in \mathcal{N}(K)} h_{K'} \leq \rho \min_{K' \in \mathcal{N}(K)} h_{K'}, \tag{1.4}$$

where $\mathcal{N}(K)$ is the set of elements K' such that $\bar{K} \cap \bar{K}' \neq \emptyset$. Property (1.4) is a local quasi-uniformity property of the mesh. The jump $[x]|_E$ of a quantity x over a face E will be defined by $[x(\xi)]|_E = \lim_{\delta \to 0}(x(\xi - n_E\delta) - x(\xi + n_E\delta))$, where $\xi \in E$ and n_E denotes a normal vector to the face E for interior faces where the normal is fixed but arbitrary, while for faces on a subdomain boundary $E \in \partial\Omega_i$ the normal is outward oriented with respect to the subdomain Ω_i and denoted n_i. Subscripts will be omitted when there is no ambiguity. For faces such that $E \cap \partial\Omega \neq \emptyset$ we set $[x(\xi)]|_E \equiv \lim_{\delta \to 0} x(\xi - n_E\delta)$. By $\{x(\xi)\}|_E$ we denote the average value of x over face E, $\{x(\xi)\}|_E = \lim_{\delta \to 0} \frac{1}{2}(x(\xi - n_E\delta) + x(\xi + n_E\delta))$. We will also use the weighted average $\{x(\xi)\}_w|_E = \lim_{\delta \to 0}(w^- x(\xi - n_E\delta) + w^+ x(\xi + n_E\delta))$, where w^- and w^+ are two positive weights such that $w_- + w_+ = 1$, and for faces on the boundary $\partial\Omega$ we define $\{x(\xi)\}|_E = \{x(\xi)\}_w|_E = \lim_{\delta \to 0} 2x(\xi - n_E\delta)$. Furthermore we will use the notation $(x, y)_X = \int_X x \cdot y \, \mathrm{dx}$, $\langle x, y \rangle_{\partial X} = \int_{\partial X} x \cdot y \, \mathrm{ds}$ with the elementwise counterparts $(x, y)_{X,h} = \sum_{K \in X} \int_K x \cdot y \, \mathrm{dx}$ and $\langle x, y \rangle_{\partial X,h} = \sum_{E \in \partial X} \int_E x \cdot y \, \mathrm{ds}$. Let $\|x\|_X = (x, x)_X^{\frac{1}{2}}$ denote the L^2-norm over X with the elementwise counterpart $\|x\|_{X,h} = (x, x)_{X,h}^{\frac{1}{2}}$. The norm of the space $H^i(X)$ will be denoted $\|x\|_{i,X}$ with $i = 0, 1, 2, \ldots$. The notations $\|x\|_X$ and $\|x\|_{0,X}$ are equivalent. The latter will be used only where it is more appropriate. For other functional spaces the notation will be made completely explicit. We will use c and C to denote generic positive constants independent of h_K but not necessarily of the local mesh geometry.

2. A domain decomposition method based on interior penalties. In this section we will show how domain decomposition using Nitsche's method leads to a continuous/discontinuous Galerkin-type penalty method in a natural way. The approximation is chosen to be continuous on each subdomain. We consider problem (1.1) on Ω and by taking V_h as trial and test space we propose the finite element method: find $u_h \in V_h$ such that

$$A(u_h, v_h) + J(u_h, v_h) + B(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{2.1}$$

where

$$A(u_h, v_h) := \sum_{i=1}^{N} \Big(((\sigma - \nabla \cdot \beta) u_h, v_h)_{\Omega_i} + (\varepsilon \nabla u_h, \nabla v_h)_{\Omega_i} - (u_h, \beta \cdot \nabla v_h)_{\Omega_i} \Big),$$

$$J(u_h, v_h) := \sum_{i=1}^{N} \sum_{E \in \mathcal{F}_i} \left\langle \tilde{\gamma}_{1,i}(h_E) \|\beta \cdot n\|_{L^\infty(E)} [\nabla u_h \cdot n], [\nabla v_h \cdot n] \right\rangle_E,$$

$$B(u_h, v_h) := \sum_{i=1}^{N} \Bigg(\left\langle \beta \cdot n_i^+ u_h, [v_h] \right\rangle_{\partial\Omega_i} - \frac{1}{2} \left\langle \{\varepsilon \nabla u_h \cdot n_i\}_w, [v_h] \right\rangle_{\partial\Omega_i}$$
$$- \frac{1}{2} \left\langle \{\varepsilon \nabla v_h \cdot n_i\}_w, [u_h] \right\rangle_{\partial\Omega_i} + \left\langle \frac{\gamma_{bc} \{\varepsilon\}_w}{\tilde{h}} [u_h], [v_h] \right\rangle_{\partial\Omega_i} \Bigg),$$

and $\beta \cdot n_i^{\pm} := \frac{1}{2}(|\beta \cdot n_i| \pm \beta \cdot n_i)$. The discretization of the advection term corresponds to the standard upwind flux after integration by parts. Note that the bilinear form A corresponds to a standard Galerkin formulation in each subdomain, supplemented with boundary terms on the inner and outer boundaries that appear naturally in the formulation to assure coercivity or consistency. We observe that terms associated with nonhomogeneous boundary data do not appear since we consider $u = 0$ on $\partial\Omega$. The interior penalty stabilization term has been decomposed into one term controlling the jumps in the gradient over *interior* faces of each subdomain Ω_i, that is, $J(u_h, v_h)$, and the terms controlling the jump of the solution over interior boundaries of neighboring subdomains, the upwind flux term and the penalty term $\langle(\gamma_{bc}\{\varepsilon\}_w/\tilde{h})[u_h], [v_h]\rangle_{\partial\Omega_i}$. The stabilization parameter $\tilde{\gamma}_{1,i}(h_E) = \gamma_{ip,i}h_E^2$ depends only on the mesh geometry of the subdomain triangulation $\mathcal{T}_{h,i}$.

Remark 2.1. If the triangulation of each subdomain consists of a single triangle, then the formulation (2.1) is equivalent to a standard interior penalty discontinuous Galerkin method for (1.1). This follows immediately by noting that the interior penalty term on the gradient jumps vanishes since there are no interior faces in the subdomains.

Remark 2.2. Recalling the framework for discontinuous Galerkin methods based on interior penalties by Arnold et al. [1], we observe that the definition of the coupling term $B(u_h, v_h)$ can be made more general by introducing a parameter s that allows us to switch between a symmetric and a nonsymmetric version. Precisely, we consider

$$B(u_h, v_h) := \sum_{i=1}^{N} \Bigg(\left\langle \beta \cdot n_i^{+} u_h, [v_h] \right\rangle_{\partial\Omega_i} - \frac{1}{2} \left\langle \{\varepsilon \nabla u_h \cdot n_i\}_w, [v_h] \right\rangle_{\partial\Omega_i} - \frac{s}{2} \left\langle \{\varepsilon \nabla v_h \cdot n_i\}_w, [u_h] \right\rangle_{\partial\Omega_i} + \left\langle \frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} [u_h], [v_h] \right\rangle_{\partial\Omega_i} \Bigg),$$

where the symmetric and the nonsymmetric cases are obtained by $s = 1$ and $s = -1$, respectively. In this work we mainly consider $s = 1$, but for comparison the nonsymmetric case will be addressed in section 4.

2.1. A priori error estimate. In this section we will prove that the finite element solution obtained from formulation (2.1) converges to the exact solution of (1.1). The a priori error estimate is proved using the techniques from [2] for the Nitsche matching conditions combined with the technique of [5] for the interior penalty stabilization. The main idea behind the stabilization based on the jump in the gradient between adjacent elements is to introduce a least squares control over the part of the convective derivative that is not in the finite element space. A key result is the following lemma. For a proof of the underlying approximation result between discrete spaces we refer to [18], and for a proof in the context of interior penalty stabilization we refer to [6]. First we define the Oswald quasi-interpolant π_h^* (see [17]).

DEFINITION 2.3. *For each node x_i, let n_i be the number of elements containing x_i as a node. We define a quasi-interpolant π_h^* of degree k by*

$$\pi_h^* v(x_i) := \frac{1}{n_i} \sum_{\{K\,:\,x_i \in K\}} v_{|K}(x_i) \quad \forall v \in \{v : v|_K \in P_k(K)\}.$$

THEOREM 2.4 (stability). *Let $\beta_h \in [V_{h,1,i}]^d$ be the Lagrange interpolant of β and let $u_h \in V_{h,k,i}$. Then there exists a constant $\gamma_{ip,i} \geq c_0 > 0$, depending only on the*

local mesh geometry, such that

$$\|\tilde{h}^{\frac{1}{2}}(\beta_h \cdot \nabla u_h - \pi_h^*(\beta_h \cdot \nabla u_h))\|_{\Omega_i}^2 \leq J_i(u_h, u_h)$$

with

$$J_i(u_h, u_h) = \sum_{E \in \mathcal{F}_i} \int_E \gamma_{ip,i} h_E^2 \|\beta_h \cdot n\|_{L^\infty(E)} [\nabla u_h]^2 \; ds. \tag{2.2}$$

Remark 2.5. Clearly then $\|\tilde{h}^{\frac{1}{2}}(\beta_h \cdot \nabla u_h - \pi_h^*(\beta_h \cdot \nabla u_h))\|_{\Omega_i}^2 \leq J(u_h, u_h)$ since $\|\beta_h \cdot n\|_{L^\infty(E)} \leq \|\beta \cdot n\|_{L^\infty(E)}$.

We define a triple norm on each subdomain as

$$|||w_h|||_i^2 = \|\sigma_0^{\frac{1}{2}} w_h\|_{\Omega_i}^2 + \|\varepsilon^{\frac{1}{2}} \nabla w_h\|_{\Omega_i}^2 + J_i(w_h, w_h) \tag{2.3}$$

and the global triple norm, taking into account also the interface interaction terms, as

$$|||w_h|||^2 = \sum_{i=1}^N \Big(|||w_h|||_i^2 + \|\delta(\varepsilon, \beta)[w_h]\|_{\partial\Omega_i}^2 \Big), \tag{2.4}$$

where $\delta(\varepsilon, \beta) = \frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} + \frac{1}{2}|\beta \cdot n|$. In what follows, we will also make use of the quantity $\delta^+(\varepsilon, \beta) = \frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} + \frac{1}{2}\beta \cdot n^+$. The explicit dependence of δ and δ^+ from ε and β will be omitted later on when there is no ambiguity of notation. For the continuity of the bilinear form we will also use the modified norm

$$\begin{aligned}]w_h[^2 = \sum_{i=1}^N \Big(& \|\sigma_1^{\frac{1}{2}} w_h\|_{\Omega_i}^2 + \|\beta\|_{L^\infty(\Omega)} \|\tilde{h}^{-\frac{1}{2}} w_h\|_{\Omega_i}^2 + \|\varepsilon^{\frac{1}{2}} \nabla w_h\|_{\Omega_i}^2 \\ & + \|(\beta \cdot n)^{+\,\frac{1}{2}} w_h\|_{\partial\Omega_i \setminus \partial\Omega}^2 + J_i(w_h, w_h) \Big) + \|(\tilde{h}\varepsilon)^{\frac{1}{2}} \nabla w_h \cdot n\|_{\partial\Omega_i}^2 + \|\delta(\varepsilon, \beta) w_h\|_{\partial\Omega_i}^2. \end{aligned} \tag{2.5}$$

To prove convergence of the discrete solutions of formulation (2.1) to the exact solution of (1.1) we will first prove three preliminary lemmas giving Galerkin orthogonality, coercivity, and approximability. Existence of discrete solutions follows by the coercivity and convergence and is proved in Theorem 2.12.

We first recall a trace inequality and the standard inverse inequality that we will use repeatedly:

$$\|v\|_{0,\partial K}^2 \leq C \Big(h_K^{-1} \|v\|_K^2 + h_K \|v\|_{1,K}^2 \Big) \quad \forall v \in H^1(K), \tag{2.6}$$

$$\|\nabla v\|_K \leq C_{inv} h_K^{-1} \|v\|_K . \tag{2.7}$$

For a proof of (2.6) we refer to [25, p. 26], and for a proof of (2.7) we refer to [9].

LEMMA 2.6 (Galerkin orthogonality). *Let $u \in \cup_{i=1}^N H^2(\Omega_i)$ be the exact solution of* (1.1) *and u_h the solution to* (2.1). *Then there holds*

$$A(u - u_h, v_h) + B(u - u_h, v_h) + J(u - u_h, v_h) = 0 \quad \forall v_h \in V_h.$$

Proof. By assumption we have that $[-\varepsilon \nabla u \cdot n + \beta \cdot n u] = [u] = 0$ in the sense of traces, and since $u \in \cup_{i=1}^N H^2(\Omega_i)$ there holds $J(u, v_h) = 0$. Therefore using the

equality $[ab] = [a]\{b\} + \{a\}[b]$ and the fact that $\{\varepsilon\nabla u\cdot n\}_w - \beta\cdot n^+ u = \{\varepsilon\nabla u\cdot n - \beta\cdot nu\}$ we have

$$
\begin{aligned}
(2.8)\qquad & A(u,v_h) + B(u,v_h) + J(u,v_h) \\
&= A(u,v_h) - \frac{1}{2}\sum_{i=1}^{N} \langle\{\varepsilon\nabla u\cdot n\}_w, [v_h]\rangle_{\partial\Omega_i} + \sum_{i=1}^{N} \left\langle \beta\cdot n^+ u, [v_h]\right\rangle_{\partial\Omega_i} \\
&= A(u,v_h) - \frac{1}{2}\sum_{i=1}^{N} \langle\{\varepsilon\nabla u\cdot n - \beta\cdot nu\}, [v_h]\rangle_{\partial\Omega_i} \\
&= A(u,v_h) - \frac{1}{2}\sum_{i=1}^{N} \int_{\partial\Omega_i\setminus\partial\Omega} [(\varepsilon\nabla u\cdot n - \beta\cdot nu)v_h]\,\mathrm{d}s - \langle\varepsilon\nabla u\cdot n - \beta\cdot nu, v_h\rangle_{\partial\Omega}\,.
\end{aligned}
$$

By an integration by parts in each subdomain we obtain

$$
\begin{aligned}
A(u,v_h) &= \sum_{i=1}^{N} \{(\varepsilon\nabla u, \nabla v_h)_{\Omega_i} - (u, \beta\cdot\nabla v) + ((\sigma - \nabla\cdot\beta)u, v)_{\Omega_i}\} \\
&= \sum_{i=1}^{N} (-\varepsilon\Delta u + \beta\cdot\nabla u + \sigma u, v_h)_{\Omega_i} + \sum_{i=1}^{N} \langle\varepsilon\nabla u\cdot n - \beta\cdot nu, v_h\rangle_{\partial\Omega_i} \\
&= \sum_{i=1}^{N} (f, v_h)_{\Omega_i} + \frac{1}{2}\sum_{i=1}^{N} \int_{\partial\Omega_i\setminus\partial\Omega} [(\varepsilon\nabla u\cdot n - \beta\cdot nu)v_h]\,\mathrm{d}s + \langle\varepsilon\nabla u\cdot n - \beta\cdot nu, v_h\rangle_{\partial\Omega}\,.
\end{aligned}
$$

It then follows from (2.8) that

$$A(u,v_h) + B(u,v_h) + J(u,v_h) = (f, v_h);$$

combining this equality with (2.1) completes the proof. ∎

LEMMA 2.7 (coercivity). *For the formulation* (2.1) *there holds*

$$c|||z_h||| \le A(z_h,z_h) + B(z_h,z_h) + J(z_h,z_h) \quad \forall z_h \in V_h.$$

Proof. We essentially only need to show that the weakly imposed boundary and interface conditions do not destroy coercivity. We have

$$
\begin{aligned}
A(z_h,z_h) + B(z_h,z_h) = \sum_{i=1}^{N} \Bigg(& \int_{\Omega_i} (\sigma - \nabla\cdot\beta)\, z_h^2\,\mathrm{dx} + \|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2 - (z_h, \beta\cdot\nabla z_h)_{\Omega_i} \\
& + \left\langle \beta\cdot n^+ z_h, [z_h]\right\rangle_{\partial\Omega_i} - \langle\{\varepsilon\nabla z_h\cdot n\}_w, [z_h]\rangle_{\partial\Omega_i} + \left\langle \frac{\gamma_{bc}\varepsilon}{\tilde{h}}[z_h], [z_h]\right\rangle_{\partial\Omega_i}\Bigg).
\end{aligned}
\tag{2.9}
$$

Consider the third term on the right-hand side of (2.9). Integration by parts yields

$$
\begin{aligned}
(2.10)\qquad \sum_{i=1}^{N} (\beta\cdot\nabla z_h, z_h)_{\Omega_i} &= -\frac{1}{2}(\nabla\cdot\beta\; z_h, z_h)_{\Omega} + \sum_{i=1}^{N} \frac{1}{2}\left\langle \beta\cdot n\, z_h, z_h\right\rangle_{\partial\Omega_i} \\
&= -\frac{1}{2}(\nabla\cdot\beta\; z_h, z_h)_{\Omega} + \sum_{i=1}^{N} \frac{1}{4}\left\langle \beta\cdot n,\, [z_h^2]\right\rangle_{\partial\Omega_i\setminus\partial\Omega} + \frac{1}{2}\left\langle \beta\cdot n\; z_h,\; z_h\right\rangle_{\partial\Omega}.
\end{aligned}
$$

Applying (2.10) to the third term of (2.9) and using the equality $a(a-b) = \frac{1}{2}(a^2 - b^2 + (a-b)^2)$ we get

$$\begin{aligned}(2.11)\quad &\sum_{i=1}^{N}\Big(-(z_h, \beta\cdot\nabla z_h)_{\Omega_i} + \langle \beta\cdot n^+ z_h, [z_h]\rangle_{\partial\Omega_i}\Big)\\ &= \sum_{i=1}^{N}\left(\frac{1}{2}(\nabla\cdot\beta\ z_h, z_h)_{\Omega_i} - \frac{1}{4}\langle \beta\cdot n, [z_h^2]\rangle_{\partial\Omega_i\setminus\partial\Omega}\right.\\ &\quad\left. + \frac{1}{2}\langle |\beta\cdot n|\ z_h,\ z_h\rangle_{\partial\Omega} + \frac{1}{2}\langle \beta\cdot n^+, [z_h^2]\rangle_{\partial\Omega_i\setminus\partial\Omega} + \frac{1}{2}\langle \beta\cdot n^+[z_h], [z_h]\rangle_{\partial\Omega_i\setminus\partial\Omega}\right).\end{aligned}$$

By observing that $\sum_{i=1}^{N}\frac{1}{2}\langle \beta\cdot n^+, [z_h^2]\rangle_{\partial\Omega_i\setminus\partial\Omega} = \sum_{i=1}^{N}\frac{1}{4}\langle \beta\cdot n, [z_h^2]\rangle_{\partial\Omega_i\setminus\partial\Omega}$ we conclude that

$$\begin{aligned}(2.12)\quad &\sum_{i=1}^{N}\Big(-(z_h, \beta\cdot\nabla z_h)_{\Omega_i} + \langle \beta\cdot n^+ z_h, [z_h]\rangle_{\partial\Omega_i}\Big)\\ &= \sum_{i=1}^{N}\left(\frac{1}{2}(\nabla\cdot\beta\ z_h, z_h)_{\Omega_i} + \frac{1}{2}\langle |\beta\cdot n|\ z_h,\ z_h\rangle_{\partial\Omega} + \frac{1}{2}\langle \beta\cdot n^+[z_h], [z_h]\rangle_{\partial\Omega_i\setminus\partial\Omega}\right).\end{aligned}$$

We now consider the second, fifth, and sixth terms of (2.9). The nonsymmetric boundary integral is split using a Cauchy–Schwarz inequality followed by Young's inequality and controlled by the symmetric terms in the following fashion:

$$\begin{aligned}(2.13)\quad &\sum_{i=1}^{N}\left(\|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2 - \langle\{\varepsilon\nabla z_h\cdot n\}_w, [z_h]\rangle_{\partial\Omega_i} + \left\langle\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}[z_h], [z_h]\right\rangle_{\partial\Omega_i}\right)\\ &\geq \sum_{i=1}^{N}\left(\|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2 - 2\alpha\|(\tilde{h}\varepsilon)^{\frac{1}{2}}\nabla z_h\cdot n\|_{\partial\Omega_i}^2 + \left\langle\left(\gamma_{bc} - \frac{1}{4\alpha}\right)\frac{\{\varepsilon\}_w}{\tilde{h}}[z_h], [z_h]\right\rangle_{\partial\Omega_i}\right).\end{aligned}$$

As a consequence of the trace inequality (2.6) and inverse estimates we have

$$(2.14)\qquad \|(\tilde{h}\varepsilon)^{\frac{1}{2}}\nabla z_h\cdot n\|_{\partial\Omega_i}^2 \leq C_t\|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2,$$

and by choosing $\alpha = (4C_t)^{-1}$ and $\gamma_{bc} = 2C_t$ we conclude that

$$\begin{aligned}(2.15)\quad &\sum_{i=1}^{N}\left(\|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2 - \langle\{\varepsilon\nabla z_h\cdot n\}_w, [z_h]\rangle_{\partial\Omega_i} + \left\langle\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}[z_h], [z_h]\right\rangle_{\partial\Omega_i}\right)\\ &\qquad\geq \frac{1}{2}\sum_{i=1}^{N}\left(\|\varepsilon^{\frac{1}{2}}\nabla z_h\|_{\Omega_i}^2 + \left\langle\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}[z_h], [z_h]\right\rangle_{\partial\Omega_i}\right).\end{aligned}$$

Combining the results of (2.9), (2.12), (2.15) and applying once again (2.14) and recalling the condition (1.2), the lemma follows, with a constant $c = \frac{1}{2}$. ∎

Remark 2.8. The constant C_t depends only on the mesh regularity and can be given an explicit expression in the case of piecewise linear elements (see [2]); for high order elements it can be computed by solving a small local eigenvalue problem (see [15]).

We will now proceed and prove approximability properties of the triple norm. The L^2-projection of u onto V_h will be denoted $\pi_h u$ and the nodal interpolation will be denoted $i_h u$. To avoid globally quasi-uniform meshes we need a stability estimate for the L^2-projection in weighted norms. This problem was considered in [13] and more recently in [3]. In [3] the following weighted stability estimate was proven:

$$\|\phi^* \pi_h u\|_\Omega \leq C \|\phi^* u\|_\Omega, \tag{2.16}$$

where ϕ^* is a piecewise linear weighting function satisfying

$$|\nabla \phi^*|_K| \leq \eta h_K^{-1} \max_{x \in K} \phi^* \tag{2.17}$$

for all K. Stability holds for η sufficiently small. We will use this stability result to prove the following

LEMMA 2.9. *If the polynomial order of the finite element space is k and $u \in H^{k+1}(\Omega)$, then there holds, for ρ sufficiently small,*

$$\sum_{K \in \mathcal{T}_{h,i}} (h_K^{-1} \|(\pi_h u - u)\|_K^2 + h_K \|\nabla(\pi_h u - u)\|_K^2) \leq C \sum_{K \in \mathcal{T}_{h,i}} h_K^{2k+1} \|u\|_{k+1,K}^2. \tag{2.18}$$

Proof. First note that by adding and subtracting the nodal interpolant $i_h u$ in the H^1 contribution of (2.18) and applying a local inverse inequality we have

$$\begin{aligned} &\sum_{K \in \mathcal{T}_{i,h}} h_K \|\nabla(\pi_h u - u)\|_K^2 \\ &\quad \leq C \sum_{K \in \mathcal{T}_{i,h}} (C_{inv}^2 h_K^{-1} \|(\pi_h u - i_h u)\|_K^2 + h_K \|\nabla(i_h u - u)\|_K^2). \end{aligned} \tag{2.19}$$

Hence it is sufficient to consider the L^2-part: $\sum_{K \in \mathcal{T}_{i,h}} h_K^{-\frac{1}{2}} \|(\pi_h u - u)\|_K^2$.

Take $\phi^* = \pi_h^* h_K^{-\frac{1}{2}}$. We must prove that this function satisfies (2.17) and that η can be made as small as needed by diminishing ρ. By the definition of the Oswald interpolant and the local quasi-regularity (1.4) one readily verifies that for all $K \in \mathcal{T}_{h,i}$

$$\max_{x \in K} |\nabla \phi^*| \leq h_K^{-1} \left| \max_{K' \in \mathcal{N}(K)} h_{K'}^{-\frac{1}{2}} - \min_{K' \in \mathcal{N}(K)} h_{K'}^{-\frac{1}{2}} \right| \leq h_K^{-1} (\rho^{\frac{1}{2}} - 1) \min_{K' \in \mathcal{N}(K)} (h_{K'}^{-\frac{1}{2}}).$$

Hence, using the inequality $\min_{K' \in \mathcal{N}(K)} (h_{K'}^{-\frac{1}{2}}) \leq \min_{x \in K} \phi^*$ we have $|\nabla \phi^*|_K| \leq (\rho^{\frac{1}{2}} - 1) h_K^{-1} \min_{x \in K} \phi^*$ on K, and therefore $\eta(\rho) = (\rho^{\frac{1}{2}} - 1)$ can be made arbitrarily small by choosing ρ small. Applying now the weighted stability estimate we have

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} h_K^{-1} \|(\pi_h u - u)\|_K^2 &\leq \rho^{\frac{1}{2}} \|\phi^* (\pi_h u - u)\|_\Omega^2 \\ &\leq 2\rho^{\frac{1}{2}} (\|\phi^* (\pi_h u - i_h u)\|_\Omega^2 + \|\phi^* (i_h u - u)\|_\Omega^2) \\ &\leq C(\rho) \|\phi^* (i_h u - u)\|_\Omega^2 \leq C(\rho) \sum_K h_K^{2k+1} \|u\|_{k+1,\Omega}^2. \end{aligned}$$

∎

LEMMA 2.10 (approximability). *Assume that the family of meshes $\mathcal{T}_{h,i}$ is locally quasi uniform with ρ such that Lemma 2.9 holds. Let $u \in \cup_{i=1}^N H^s(\Omega_i)$ with $s \geq k+1 \geq 2$ and let $\pi_h u$ denote the standard L_2-projection of u onto V_h; then we have that*

$$|||\pi_h u - u||| \leq C(\varepsilon^{\frac{1}{2}} \mathcal{H}(0,u) + \|\beta\|_{L^\infty(\Omega)}^{\frac{1}{2}} \mathcal{H}(1,u) + \sigma_0^{\frac{1}{2}} \mathcal{H}(2,u)),$$

where C is independent of σ, ε, β, and h but depends on the mesh geometry and

$$\mathcal{H}(\alpha,u)=\left(\sum_{i=1}^{N}\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k+\alpha}\|u\|_{k+1,K}^2\right)^{\frac{1}{2}}.$$

Proof. It follows from the stability of the L^2-projection and standard interpolation results that $\|\sigma_0^{\frac{1}{2}}(\pi_h u-u)\|_{\Omega_i}\leq\sigma_0^{\frac{1}{2}}(\sum_{K\in\mathcal{T}_{h,i}}h_K^{2(k+1)}\|u\|_{k+1,K}^2)^{\frac{1}{2}}$. We then write $\xi_h=\pi_h u-i_h u$, where i_h denotes the nodal interpolant, and note that $\xi_h=\pi_h(u-i_h u)$. By the H^1-stability of the L^2-projection on locally quasi-uniform meshes [4, 11] we may write

$$\|\nabla\xi_h\|_{\Omega_i}\leq\|\nabla(u-i_h u)\|_{\Omega_i}\leq C\left(\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k}\|u\|_{k+1,K}^2\right)^{\frac{1}{2}}. \tag{2.20}$$

It immediately follows by means of the triangular inequality that

$$\|\varepsilon^{\frac{1}{2}}\nabla(u-\pi_h u)\|_{\Omega_i}^2\leq C\varepsilon\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k}\|u\|_{k+1,K}^2,$$

and by an application of the inverse inequality and Lemma 2.9 we have

$$\sum_{K\in\mathcal{T}_{h,i}}h_K^3\|\nabla\xi_h\|_{1,\Omega}^2\leq C\sum_{K\in\mathcal{T}_{h,i}}h_K\|\nabla\xi_h\|_{\Omega}^2\leq\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k+1}\|u\|_{k+1,K}^2. \tag{2.21}$$

Using the trace inequality (2.6) together with (2.20) and (2.21), it follows that

$$\begin{aligned}\|(\varepsilon\tilde{h})^{\frac{1}{2}}\nabla(\pi_h u-u)\cdot n\|_{\partial\Omega_i}^2&\leq C\sum_{K\in\mathcal{T}_{h,i}}\left(\varepsilon\|\nabla(\pi_h u-u)\|_K^2+\varepsilon h_K^2\|\nabla(\pi_h u-u)\|_{1,K}^2\right)\\&\leq C\varepsilon\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k}\|u\|_{k+1,\Omega_i}^2.\end{aligned}$$

Using once again (2.6), (2.20), and (2.21) we get in a similar fashion

$$\begin{aligned}&J_1(u-\pi_h u,u-\pi_h u)\\&\leq C\sum_{i=1}^{N}\gamma_{ip,i}\|\beta\|_{L^\infty(\Omega_i)}\sum_{K\in\mathcal{T}_{h,i}}\left(\|\tilde{h}^{\frac{1}{2}}\nabla(u-\pi_h u)\|_K^2+\|\tilde{h}^{\frac{3}{2}}\nabla(u-\pi_h u)\|_{1,K}^2\right)\\&\leq\sum_{i=1}^{N}\|\beta\|_{L^\infty(\Omega_i)}\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k+1}\|u\|_{k+1,\Omega_i}^2.\end{aligned}$$

Finally we note that for the boundary term we have, using (2.6) and (2.20),

$$\begin{aligned}\langle\pi_h u-u,\pi_h u-u\rangle_{\partial\Omega_i}&\leq\sum_{K:\partial K\cap\partial\Omega_i\neq\emptyset}h_K^{-1}\|\pi_h u-u\|_K^2+h_K\|\nabla(\pi_h u-u)\|_K^2\\&\leq\sum_{K\in\mathcal{T}_{h,i}}h_K^{2k+1}\|u\|_{k+1,\Omega_i}^2,\end{aligned}$$

which concludes the proof. ∎

As an immediate consequence of the above result and Lemma 2.9 we have the following.

COROLLARY 2.11. *Under the same assumptions as in Lemma* 2.10 *we have that*

$$|]\pi_h u - u[| \le C(\varepsilon^{\frac{1}{2}}\mathcal{H}(0,u) + \|\beta\|^{\frac{1}{2}}_{L^\infty(\Omega)}\mathcal{H}(1,u) + \sigma_1^{\frac{1}{2}}\mathcal{H}(2,u)),$$

where C is independent of σ, ε, β, and h but depends on the mesh geometry.

THEOREM 2.12 (convergence). *Let $u \in \cup_{i=1}^N H^s(\Omega_i)$ with $s \ge k+1 \ge 2$ be the solution of* (1.1) *and let $u_h \in V_h$ be the solution of* (2.1). *Then the following a priori error estimate holds:*

$$|||u - u_h||| \le C\left(\varepsilon^{\frac{1}{2}}\mathcal{H}(0,u) + \|\beta\|^{\frac{1}{2}}_{L^\infty(\Omega)}\mathcal{H}(1,u) + \left(\sigma_1^{\frac{1}{2}} + \frac{|\beta|_{W^{1,\infty}(\Omega)}}{\sigma_0}\right)\mathcal{H}(2,u)\right).$$

Proof. We decompose the error into two parts: $\eta = u - \pi_h u$ and $\xi_h = \pi_h u - u_h$. It follows that $u - u_h = \eta + \xi_h$. By Lemma 2.10 we know that

$$|||\eta||| \le C(\varepsilon^{\frac{1}{2}}\mathcal{H}(0,u) + \|\beta\|^{\frac{1}{2}}_{L^\infty(\Omega)}\mathcal{H}(1,u) + \sigma_0^{\frac{1}{2}}\mathcal{H}(2,u)),$$

and it is therefore sufficient to study $\xi_h = \pi_h u - u_h$. Using Lemma 2.7 we have

$$c|||\xi_h|||^2 \le A(\xi_h,\xi_h) + B(\xi_h,\xi_h) + J(\xi_h,\xi_h),$$

and by Galerkin orthogonality

$$c|||\xi_h|||^2 \le A(\eta,\xi_h) + B(\eta,\xi_h) + J(\eta,\xi_h).$$

After an integration by parts in the convective term and an application of the Cauchy–Schwarz inequality in all other terms we have

$$c|||\xi_h|||^2 \le |]\eta[|\,|||\xi_h||| + |(\eta, \beta\cdot\nabla\xi_h)|.$$

Using now the orthogonality of the L^2-projection and Lemma 2.4 we may write

$$\begin{aligned} c|||\xi_h|||^2 &\le |]\eta[|\,|||\xi_h||| + |(\eta, \beta_h\cdot\nabla\xi_h - \pi^*\beta_h\cdot\nabla\xi_h)| + |(\eta,(\beta-\beta_h)\cdot\nabla\xi_h) \\ &\le |]\eta[|\,|||\xi_h||| + \|\beta\|_{L^\infty(\Omega)}\|\tilde{h}^{-\frac{1}{2}}\eta\|J(\xi_h,\xi_h)^{\frac{1}{2}} + |\beta|_{W^{1,\infty}(\Omega)}\|\eta\|\ \|\tilde{h}\nabla\xi_h\| \\ &\le |]\eta[|\,|||\xi_h||| + C_i\frac{|\beta|_{W^{1,\infty}(\Omega)}}{\sigma_0}\|\eta\|\ |||\xi_h|||. \end{aligned}$$

The theorem now follows by the approximation Lemma 2.10 and Corollary 2.11. ∎

Remark 2.13. The a priori error analysis carried out in this section holds true for any admissible choice of the weights w^+, w^- (such that $w^+, w^- > 0$ and $w^+ + w^- = 1$) that appear in the definition of $\{\cdot\}_w$ as also proved in [16] and [24]. In the following section we propose a definition of these weights according to the specific characteristics of the problem at hand.

2.2. Optimal choice of the averaging weights. To make the notation simpler, let us assume that only two subdomains Ω_i are considered with corresponding diffusivities ε_i, $i = 1, 2$. In this case, let $\partial\Omega_1 \setminus \partial\Omega$ be the interface between the subdomains and let n_1 be the outer normal with respect to Ω_1. Then we define the weighted average on the interface as $\{x(\xi)\}_w = \lim_{\delta\to 0}(w_1 x(\xi - n_1\delta) + w_2 x(\xi + n_1\delta))$.

The regularity assumptions on the solution u can be expected to hold only as long as $\varepsilon_i \ge \varepsilon_0 > 0$ in all the subdomains and the intersubdomain boundaries are

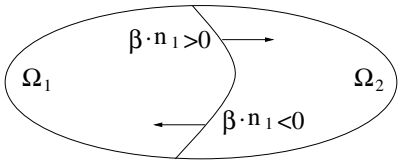


FIG. 2.1. *The model situation.*

smooth enough. In case ε_i vanishes in a subdomain, the weights w_i may be chosen so as to guarantee that the matching conditions automatically recover the physically correct behavior, relaxing the continuity of u but keeping the continuity of the fluxes. It turns out that balancing the diffusive fluxes yields a numerical scheme with the right asymptotic behavior if the diffusion coefficient vanishes in some subdomain. Let us exemplify this on a model case. We consider a domain Ω split into two neighboring subdomains Ω_1 and Ω_2 with a diffusion coefficient ε that is a regular function in each subdomain, but discontinuous across the interface $\partial\Omega_1 \cap \partial\Omega_2$. We choose the weights w_1 and w_2 such that

$$w_i(\xi) := \lim_{\delta \to 0} \frac{\varepsilon(\xi + \delta n_i)}{\varepsilon(\xi + \delta n_i) + \varepsilon(\xi - \delta n_i)} \quad \forall \xi \in \partial\Omega_1 \cap \partial\Omega_2,\ i = 1, 2, \tag{2.22}$$

where n_i is the outward unit normal with respect to Ω_i. We observe that such weights always satisfy $w_1(\xi) + w_2(\xi) = 1$ for all $\xi \in \partial\Omega_1 \cap \partial\Omega_2$. Moreover, in the case of smooth diffusivity across the interface, our choice coincides with the classical one, $w_1 = w_2 = \frac{1}{2}$. Furthermore, let us define $\omega(\xi) := w_1(\xi)\varepsilon_1(\xi) = w_2(\xi)\varepsilon_2(\xi)$. Our choice of the weights implies that $\{\varepsilon_i \nabla u_h \cdot n_i\}_w = 2\omega\{\nabla u_h \cdot n_i\}$, which shows that our method turns out to consider the arithmetic average of the gradients instead of the arithmetic average of the diffusion fluxes in order to construct the consistency term. Using these weights the coupling term between Ω_1 and Ω_2 becomes

$$\begin{aligned} B(u_h, v_h) = \sum_{i=1}^{2} \Bigg(& \langle \beta \cdot n^+ u_h, [v_h] \rangle_{\partial\Omega_i \setminus \partial\Omega} - \langle \omega\{\nabla u_h \cdot n\}, [v_h] \rangle_{\partial\Omega_i \setminus \partial\Omega} \\ & - \langle \omega\{\nabla v_h \cdot n\}, [u_h] \rangle_{\partial\Omega_i \setminus \partial\Omega} + \left\langle \frac{\gamma_{bc} 2\omega}{\tilde{h}} [u_h], [v_h] \right\rangle_{\partial\Omega_i \setminus \partial\Omega} \Bigg). \end{aligned}$$

Consider the case when ε_1 goes to zero; then only the upwind flux term remains. One may readily verify that the coupling term $B(u_h, v_h)$ corresponds to the weak formulation of the conditions

$$\begin{aligned} -\varepsilon_2 \nabla u_{2,h} \cdot n_1 + \beta \cdot n_1 u_{2,h} = \beta \cdot n_1 u_{1,h} \text{ on } \partial\Omega_1 \setminus \partial\Omega, \text{ where } \beta \cdot n_1 > 0, \\ u_{1,h} = u_{2,h} \text{ and } -\varepsilon_2 \nabla u_{2,h} \cdot n_1 = 0 \text{ on } \partial\Omega_1 \setminus \partial\Omega, \text{ where } \beta \cdot n_1 < 0, \end{aligned}$$

which were proposed for the hybrid elliptic-hyperbolic coupling in Gastaldi and Quarteroni [14] (see also [10]). By the symmetry of the weights the same holds in the case $\varepsilon_2 = 0$. The convergence analysis of the iterative method and numerical experience also indicates that this choice of w_1 and w_2 is the only viable one for the iterative algorithm.

3. An iterative splitting method. To introduce and analyze the iterative method we will restrict the discussion to the case of two subdomains Ω_i, $i = 1, 2$,

with interface $\partial\Omega_i \setminus \partial\Omega \neq \emptyset$. Nevertheless, the generalization to the multidomain case is straightforward and will be addressed later on. We denote with $u_{h,i} \in V_{h,k,i}$ the restriction on Ω_i of the global numerical solution. For the sake of simplicity, we also identify with $u_{h,i}$ the function on Ω that is obtained by extending $u_{h,i}$ to zero outside Ω_i. If we consider the formulation (2.1) and decouple the subdomains by using some approximation $u_{h,j}^k$ of $u_{h,j}$ with $j \neq i$ as boundary data from the neighboring subdomain with respect to Ω_i, we obtain the iterative scheme. Given $u_{h,1}^k, u_{h,2}^k$, for $k = 1, 2, \ldots$, find $u_{h,1}^{k+1} \in V_{h,1}$ such that

(3.1)
$$A(u_{h,1}^{k+1}, v_{h,1}) + \tilde{B}(u_{h,1}^{k+1}, u_{h,2}^k, v_{h,1}) + J(u_{h,1}^{k+1}, v_{h,1}) + S(u_{h,1}^{k+1}, u_{h,1}^k, v_{h,1}) = (f, v_{h,1})$$

and $u_{h,2}^{k+1} \in V_{h,2}$ such that

(3.2)
$$A(u_{h,2}^{k+1}, v_{h,2}) + \tilde{B}(u_{h,2}^{k+1}, u_{h,1}^k, v_{h,2}) + J(u_{h,2}^{k+1}, v_{h,2}) + S(u_{h,2}^{k+1}, u_{h,2}^k, v_{h,2}) = (f, v_{h,2}),$$

where

$$S(u_{h,i}^{k+1}, u_{h,i}^k, v_{h,i}) = \sum_{E \in G_h} \left\langle \frac{\gamma_{it}}{\tilde{h}} (u_{h,i}^{k+1} - u_{h,i}^k), v_{h,i} \right\rangle_E$$

are the terms that stabilize the iterations and the trace mesh is defined by

$$G_h = \{E \neq \emptyset : E = \partial K_i \cap \partial K_j;\ \forall K_i \in \mathcal{T}_{h,i};\ \forall K_j \in \mathcal{T}_{h,j};\ i \neq j\},$$

and we recall that $\tilde{h}(x)|_E = h_E$ for all $E \in G_h$.

The stabilization term $S(u_{h,i}^{k+1}, u_{h,i}^k, v_{h,i})$ corresponds to iteration relaxation and is mandatory to get good convergence properties. If S is omitted, we cannot prove convergence of the triple norm. In fact explicit control of the error in the jump over the interface is lost, and numerical experience shows very poor convergence as well for $S = 0$. Moreover, we note that the stabilization term is consistent in the sense that $S(u_{h,i}, u_{h,i}, v_{h,i}) = 0$. Finally, we have denoted with $\tilde{B}(u_{h,i}, u_{h,j}, v_{h,i})$, $i, j = 1, 2$, $j \neq i$, the interface/boundary penalty bilinear form after the iterative splitting, which is defined as follows:

$$\begin{aligned}
&\tilde{B}(u_{h,i}, u_{h,j}, v_{h,i}) = \langle \beta \cdot n_i^+ u_{h,i}, v_{h,i} \rangle_{\partial\Omega_i \setminus \partial\Omega} + \langle \beta \cdot n_i^- u_{h,j}, v_{h,i} \rangle_{\partial\Omega_i \setminus \partial\Omega} \\
&- \langle w_i \varepsilon_i \nabla u_{h,i} \cdot n_i + w_j \varepsilon_j \nabla u_{h,j} \cdot n_i, v_{h,i} \rangle_{\partial\Omega_i \setminus \partial\Omega} - \langle \varepsilon_i w_i \nabla v_{h,i} \cdot n_i, u_{h,i} - u_{h,j} \rangle_{\partial\Omega_i \setminus \partial\Omega} \\
&+ \left\langle 2 \frac{\gamma_{bc} \{\varepsilon\}_w}{\tilde{h}} (u_{h,i} - u_{h,j}), v_{h,i} \right\rangle_{\partial\Omega_i \setminus \partial\Omega} + \left\langle \frac{\gamma_{bc} \varepsilon}{\tilde{h}} u_{h,i}, v_{h,i} \right\rangle_{\partial\Omega_i \cap \partial\Omega} \\
&+ \langle \beta \cdot n^+ u_{h,i}, v_{h,i} \rangle_{\partial\Omega_i \cap \partial\Omega} - \langle \varepsilon_i \nabla u_{h,i} \cdot n, v_{h,i} \rangle_{\partial\Omega_i \cap \partial\Omega} - \langle \varepsilon_i \nabla v_{h,i} \cdot n, u_{h,i} \rangle_{\partial\Omega_i \cap \partial\Omega}.
\end{aligned}$$

Since the data on Ω_j are taken at the earlier iteration for both domains the two problems are decoupled and can be solved in parallel.

The present setting can easily be generalized to the case of several subdomains. Let $\bar{\Omega} = \cup_{i=1}^N \bar{\Omega}_i$ be the partition in N subdomains and let $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$ be the corresponding interfaces. Then, since the definition of A and J are already general with respect to N, problems (3.1) and (3.2) do not need to be modified in the multidomain case, provided that the definition of $\tilde{B}(u_{h,i}, u_{h,j}, v_{h,i})$ is adapted by replacing $\langle \cdot, \cdot \rangle_{\partial\Omega_i \setminus \partial\Omega}$ with $\sum_{i,j=1}^N \langle \cdot, \cdot \rangle_{\Gamma_{ij}}$. Thanks to the generality of the construction of G_h

the term $S(u_{h,i}, u_{h,i}, v_{h,i})$ remains unchanged. Moreover, in the multidomain case, the system of equations (3.1)–(3.2) should be complemented with one equation for each new subdomain. Although the formal generalization to the multidomain case is straightforward, we do not consider it here in order to reduce the notational complexity in the analysis of the iterative method.

LEMMA 3.1. *The subproblems* (3.1) *and* (3.2) *are well-posed in* $V_{h,i}$ *with respect to the norm* $|||\cdot|||_i$.

Proof. The proof is an immediate consequence of Lemma 2.7 restricted to one subdomain. ∎

We define the splitting error as $e_h^k = u_h - u_h^k$, where u_h is the solution to the finite element formulation of (2.1) and u_h^k is the solution after k iterations of (3.1) and (3.2). We will now state and prove the main result of this section.

THEOREM 3.2. *The iterative method defined by problems* (3.1) *and* (3.2) *converges when the relaxation parameter* γ_{it} *is chosen big enough. More precisely, there exists a positive constant* c (*that is, the coercivity constant of Theorem* 2.7) *such that*

(3.3)
$$c\sum_{k=1}^{\infty}\left|\left|\left|e_h^k\right|\right|\right|^2 \le \sum_{i=1,2}\left(\frac{c}{2}\left\|\varepsilon_i^{\frac{1}{2}}\nabla e_{h,i}^0\right\|_{\Omega_i}^2 + \left\|(\delta^+)^{\frac{1}{2}}e_{h,i}^0\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 + \left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^0\right\|_{\partial\Omega_i\setminus\partial\Omega}^2\right)$$

Proof. By subtracting the decoupled formulation given by (3.1) and (3.2) from the formulation (2.1) we have

$$A(e_{h,1}^{k+1}, v_{h,1}) + \tilde{B}(e_{h,1}^{k+1}, e_{h,2}^k, v_{h,1}) + J(e_{h,1}^{k+1}, v_{h,1}) + S(e_{h,1}^{k+1}, e_{h,1}^k, v_{h,1}) = 0 \tag{3.4}$$

and

$$A(e_{h,2}^{k+1}, v_{h,2}) + \tilde{B}(e_{h,2}^{k+1}, e_{h,1}^k, v_{h,2}) + J(e_h^{k+1}, v_{h,2}) + S(e_{h,2}^{k+1}, e_{h,2}^k, v_{h,2}) = 0. \tag{3.5}$$

We now choose $v_{h,i} = e_{h,i}^{k+1}$ to obtain

$$A(e_h^{k+1}, e_h^{k+1}) + \tilde{B}(e_{h,1}^{k+1}, e_{h,2}^k, e_{h,1}^{k+1}) + \tilde{B}(e_{h,2}^{k+1}, e_{h,1}^k, e_{h,2}^{k+1}) + J(e_h^{k+1}, e_h^{k+1}) + \sum_{i=1,2} S(e_{h,i}^{k+1}, e_{h,i}^k, e_{h,i}^{k+1}) = 0.$$

Proceeding now by adding and subtracting $B(e_h^{k+1}, e_h^{k+1})$ we may write

$$\begin{aligned} A(e_h^{k+1}, e_h^{k+1}) + B(e_h^{k+1}, e_h^{k+1}) + J(e_h^{k+1}, e_h^{k+1}) + \sum_{i=1,2} S(e_{h,i}^{k+1}, e_{h,i}^k, e_{h,i}^{k+1}) \\ = B(e_h^{k+1}, e_h^{k+1}) - \tilde{B}(e_{h,1}^{k+1}, e_{h,2}^k, e_{h,1}^{k+1}) - \tilde{B}(e_{h,2}^{k+1}, e_{h,1}^k, e_{h,2}^{k+1}). \end{aligned} \tag{3.6}$$

The first three terms on the left-hand side will be controlled by the coercivity Lemma 2.7, while the term that stabilizes the iterations can be rewritten as follows:

$$\begin{aligned} \sum_{i=1,2} S(e_{h,i}^{k+1}, e_{h,i}^k, e_{h,i}^{k+1}) &= \sum_{i=1,2}\sum_{E\in G_h}\left\langle\frac{\gamma_{it}}{\tilde{h}}(e_{h,i}^{k+1} - e_{h,i}^k), e_{h,i}^{k+1}\right\rangle_E \\ &= \frac{1}{2}\sum_{i=1,2}\left[\left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^{k+1}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 - \left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^k\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 \right. \\ &\qquad \left. + \left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}(e_{h,i}^{k+1} - e_{h,i}^k)\right\|_{\partial\Omega_i\setminus\partial\Omega}^2\right]. \end{aligned} \tag{3.7}$$

It remains to bound the interface residual of the right-hand side:

$$R(e_{h,1}^k, e_{h,1}^{k+1}, e_{h,2}^k, e_{h,2}^{k+1}) = B(e_h^{k+1}, e_h^{k+1}) - \tilde{B}(e_{h,1}^{k+1}, e_{h,2}^k, e_{h,1}^{k+1}) - \tilde{B}(e_{h,2}^{k+1}, e_{h,1}^k, e_{h,2}^{k+1}).$$

The residual R is different from zero only on the interface of the subdomains and consists of three parts:

(A) the advective interface flux term from the advection term;
(B) the symmetric interface flux term from the Laplacian;
(C) the interface penalization term.

We now rearrange the terms for the three above-mentioned cases.

(A) The advective interface fluxes:

$$\sum_{\substack{i,j=1,2\\ i\neq j}} \Big[\langle \beta\cdot n_i^+ e_{h,i}^{k+1}, e_{h,i}^{k+1} - e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} - \langle \beta\cdot n_i^+ e_{h,i}^{k+1}, e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} - \langle \beta\cdot n_i^+ e_{h,i}^{k}, -e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big]$$
$$= \sum_{\substack{i,j=1,2\\ i\neq j}} \Big[\langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,j}^{k+1} - e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} + \langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big].$$

We observe that

$$\langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,j}^{k+1} - e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} \leq \frac{1}{4\mu_i}\|(\beta\cdot n_i^+)^{\frac12}(e_{h,i}^k - e_{h,i}^{k+1})\|^2_{\partial\Omega_i\setminus\partial\Omega} + \mu_i\|(\beta\cdot n_i^+)^{\frac12}[e_h^{k+1}]\|^2_{\partial\Omega_i\setminus\partial\Omega}$$

and

$$\langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} = \frac12\|(\beta\cdot n_i^+)^{\frac12} e_{h,i}^k\|^2_{\partial\Omega_i\setminus\partial\Omega} - \frac12\|(\beta\cdot n_i^+)^{\frac12} e_{h,i}^{k+1}\|^2_{\partial\Omega_i\setminus\partial\Omega} - \frac12\|(\beta\cdot n_i^+)^{\frac12}(e_{h,i}^{k+1} - e_{h,i}^k)\|^2_{\partial\Omega_i\setminus\partial\Omega}.$$

By combining these results we obtain

(3.8)
$$\sum_{\substack{i,j=1,2\\ i\neq j}} \Big[\langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,j}^{k+1} - e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} + \langle \beta\cdot n_i^+ (e_{h,i}^{k} - e_{h,i}^{k+1}), e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big]$$
$$\leq \sum_{i=1,2}\Big[\mu_i\||\beta\cdot n_i|^{\frac12}[e_h^{k+1}]\|^2_{\partial\Omega_i\setminus\partial\Omega} + \frac{1-2\mu_i}{4\mu_i}\||\beta\cdot n_i|^{\frac12}(e_{h,i}^k - e_{h,i}^{k+1})\|^2_{\partial\Omega_i\setminus\partial\Omega} + \frac12\|(\beta\cdot n_i^+)^{\frac12} e_{h,i}^k\|^2_{\partial\Omega_i\setminus\partial\Omega} - \frac12\|(\beta\cdot n_i^+)^{\frac12} e_{h,i}^{k+1}\|^2_{\partial\Omega_i\setminus\partial\Omega}\Big].$$

(B) The boundary part of the Laplacian operator may then be written as

$$-\frac12\sum_{\substack{i,j=1,2\\ i\neq j}}\Big[2\langle\{\varepsilon\nabla e_h^{k+1}\cdot n_i\}_w, e_{h,i}^{k+1} - e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$-\langle \omega\nabla e_{h,i}^{k+1}\cdot n_i + \omega\nabla e_{h,j}^k\cdot n_i, e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} - \langle \omega\nabla e_{h,i}^{k+1}\cdot n_i, e_{h,i}^{k+1} - e_{h,j}^k\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$-\langle \omega\nabla e_{h,i}^{k}\cdot n_i + \omega\nabla e_{h,j}^{k+1}\cdot n_i, -e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega} - \langle \omega\nabla e_{h,j}^{k+1}\cdot n_i, e_{h,i}^{k} - e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big],$$

which can be rewritten as follows:

$$
\begin{aligned}
-\frac{1}{2}\sum_{\substack{i,j=1,2\\ i\neq j}}\Big[&\langle\omega\nabla e_{h,j}^{k+1}\cdot n_i, e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}-\langle\omega\nabla e_{h,j}^{k}\cdot n_i, e_{h,i}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&-\langle\omega\nabla e_{h,i}^{k+1}\cdot n_i, e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}+\langle\omega\nabla e_{h,i}^{k}\cdot n_i, e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&+\langle\omega\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i, e_{h,j}^{k+1}-e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&+\langle\omega\nabla(e_{h,j}^{k+1}-e_{h,j}^{k})\cdot n_i, e_{h,i}^{k+1}-e_{h,i}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big],
\end{aligned}
$$

where we recall that $w_1\varepsilon_1 = w_2\varepsilon_2 = \omega$. For this choice of the averaging weights the first four terms vanish, precisely:

$$
\begin{aligned}
\sum_{\substack{i,j=1,2\\ i\neq j}}\Big[&\langle\omega\nabla e_{h,j}^{k+1}\cdot n_i, e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}-\langle\omega\nabla e_{h,j}^{k}\cdot n_i, e_{h,i}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&-\langle\omega\nabla e_{h,i}^{k+1}\cdot n_i, e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}+\langle\omega\nabla e_{h,i}^{k}\cdot n_i, e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big]=0.
\end{aligned}
$$

By means of the Cauchy–Schwarz and Young inequalities, we have for the fifth term

$$
\begin{aligned}
\langle\omega\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i, e_{h,j}^{k+1}-e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}&=\sum_{E\in G_h}\langle\omega^{\frac12}\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i, \omega^{\frac12}(e_{h,j}^{k+1}-e_{h,j}^{k})\rangle_E\\
&\le\sum_{E\in G_h}2\left[h_E^{\frac12}\|\omega^{\frac12}\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i\|_E\cdot h_E^{-\frac12}\|\omega^{\frac12}(e_{h,j}^{k+1}-e_{h,j}^{k})\|_E\right]\\
&\le\sum_{E\in G_h}\left[\alpha_i h_E\|\omega^{\frac12}\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i\|_E^2+(\alpha_i h_E)^{-1}\|\omega^{\frac12}(e_{h,j}^{k+1}-e_{h,j}^{k})\|_E^2\right].
\end{aligned}
$$

Then, by virtue of trace and inverse inequalities (see Remark 2.8), there exists a positive constant C_t such that

$$
\begin{aligned}
\sum_{E\in G_h}h_E\|\omega^{\frac12}\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i\|_E^2&\le C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\|\varepsilon_i^{\frac12}\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\|_{\Omega_i}^2\\
&\le C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\left[\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k}\|_{\Omega_i}^2+\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k+1}\|_{\Omega_i}^2\right].
\end{aligned}
$$

We proceed analogously for the term $\langle\omega\nabla(e_{h,j}^{k+1}-e_{h,j}^{k})\cdot n_i, e_{h,i}^{k+1}-e_{h,i}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}$.

Summing up all the contributions we obtain that

$$
\begin{aligned}
-\frac{1}{2}\sum_{\substack{i,j=1,2\\ i\neq j}}\Big[&2\langle\omega\nabla(e_{h,i}^{k}-e_{h,i}^{k+1})\cdot n_i, e_{h,j}^{k+1}-e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&+\langle\omega\nabla e_{h,j}^{k+1}\cdot n_i, e_{h,i}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}-\langle\omega\nabla e_{h,j}^{k}\cdot n_i, e_{h,i}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&-\langle\omega\nabla e_{h,i}^{k+1}\cdot n_i, e_{h,j}^{k+1}\rangle_{\partial\Omega_i\setminus\partial\Omega}+\langle\omega\nabla e_{h,i}^{k}\cdot n_i, e_{h,j}^{k}\rangle_{\partial\Omega_i\setminus\partial\Omega}\Big]\\
\le\sum_{i=1,2}\Big[&\alpha_i C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\left(\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k+1}\|_{\Omega_i}^2+\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k}\|_{\Omega_i}^2\right)\\
&+\frac{\|w_i\varepsilon_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}}{\alpha_i}\|(\tilde h)^{-\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\|_{\partial\Omega_i\setminus\partial\Omega}\Big].
\end{aligned}
\tag{3.9}
$$

(C) For the interface penalization term we get

$$\sum_{\substack{i,j=1,2\\ i\neq j}} \left[\left\langle \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,j}^{k+1}), e_{h,i}^{k+1} - e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} \right.$$
$$\left. - \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,j}^{k}), e_{h,i}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} - \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,i}^{k}), e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} \right].$$

By means of algebraic manipulations we obtain

$$\left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,j}^{k}), e_{h,i}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} + \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,i}^{k}), e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$= \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,i}^{k}), e_{h,j}^{k+1} - e_{h,i}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} + \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,i}^{k}), e_{h,i}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$+ \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,j}^{k}), e_{h,i}^{k+1} - e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} + \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,j}^{k}), e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$+ \left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} [e_h^{k+1}] \right\|^2_{\partial\Omega_i\setminus\partial\Omega}.$$

By virtue of the particular choice of the weights that gives $2\omega = \{\varepsilon\}_w$ and by means of standard inequalities we observe that

$$\left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,i}^{k+1} - e_{h,i}^{k}), e_{h,j}^{k+1} - e_{h,i}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega} + \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,j}^{k}), e_{h,j}^{k+1} - e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$\leq \frac{1}{4\mu_i} \left[\left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} (e_{h,i}^{k+1} - e_{h,i}^{k}) \right\|^2_{\partial\Omega_i\setminus\partial\Omega} + \left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} (e_{h,j}^{k+1} - e_{h,j}^{k}) \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right]$$
$$\mu_i \left[\left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} [e_h^{k+1}] \right\|^2_{\partial\Omega_i\setminus\partial\Omega} + \left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} [e_h^{k+1}] \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right]$$
$$= \frac{1}{4\mu_i} \left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} (e_{h,i}^{k+1} - e_{h,i}^{k}) \right\|^2_{\partial\Omega_i\setminus\partial\Omega} + \mu_i \left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} [e_h^{k+1}] \right\|^2_{\partial\Omega_i\setminus\partial\Omega}$$

and that

$$\sum_{j=1,2} \left\langle \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right) (e_{h,j}^{k+1} - e_{h,j}^{k}), e_{h,j}^{k+1} \right\rangle_{\partial\Omega_i\setminus\partial\Omega}$$
$$= \frac{1}{2} \sum_{j=1,2} \left[\left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} e_{h,j}^{k+1} \right\|^2_{\partial\Omega_i\setminus\partial\Omega} - \left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} e_{h,j}^{k} \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right.$$
$$\left. + \left\| \left(\frac{\gamma_{bc}\omega}{\tilde{h}} \right)^{\frac{1}{2}} (e_{h,j}^{k+1} - e_{h,i}^{k}) \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right]$$
$$= \frac{1}{2} \left[\left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} e_{h,i}^{k+1} \right\|^2_{\partial\Omega_i\setminus\partial\Omega} - \left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} e_{h,i}^{k} \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right.$$
$$\left. + \left\| \left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}} \right)^{\frac{1}{2}} (e_{h,i}^{k+1} - e_{h,i}^{k}) \right\|^2_{\partial\Omega_i\setminus\partial\Omega} \right].$$

Summing up all the terms of the residual (C) we have

$$
\begin{aligned}
(3.10)\quad & \sum_{i=1,2}\Bigg[\left\langle\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)(e_{h,i}^{k+1}-e_{h,j}^{k+1}),e_{h,i}^{k+1}-e_{h,j}^{k+1}\right\rangle_{\partial\Omega_i\setminus\partial\Omega}\\
&-\left\langle\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)(e_{h,i}^{k+1}-e_{h,j}^{k}),e_{h,i}^{k+1}\right\rangle_{\partial\Omega_i\setminus\partial\Omega}-\left\langle\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)(e_{h,j}^{k+1}-e_{h,i}^{k}),e_{h,j}^{k+1}\right\rangle_{\partial\Omega_i\setminus\partial\Omega}\Bigg]\\
&\le\sum_{i=1,2}\Bigg[\frac{1-2\mu_i}{4\mu_i}\left\|\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)^{\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\right\|^2_{\partial\Omega_i\setminus\partial\Omega}+\mu_i\left\|\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)^{\frac12}[e_h^{k+1}]\right\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&\qquad+\frac12\left\|\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)^{\frac12}e_{h,i}^{k}\right\|^2_{\partial\Omega_i\setminus\partial\Omega}-\frac12\left\|\left(\frac{\gamma_{bc}\{\varepsilon\}_w}{\tilde{h}}\right)^{\frac12}e_{h,i}^{k+1}\right\|^2_{\partial\Omega_i\setminus\partial\Omega}\Bigg].
\end{aligned}
$$

By putting together (3.8), (3.9), (3.10) we obtain the following inequality:

$$
\begin{aligned}
(3.11)\quad & R(e_{h,1}^{k},e_{h,1}^{k+1},e_{h,2}^{k},e_{h,2}^{k+1})\\
&\le\sum_{i=1,2}\Bigg[\alpha_iC_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\left(\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k+1}\|^2_{\Omega_i}+\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k}\|^2_{\Omega_i}\right)+\mu_i\|\delta^{\frac12}[e_h^{k+1}]\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&\quad+\frac{1-2\mu_i}{4\mu_i}\|\delta^{\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\|^2_{\partial\Omega_i\setminus\partial\Omega}+\frac{\|w_i\varepsilon_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}}{\alpha_i}\|\tilde{h}^{-\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&\qquad+\frac12\|(\delta^+)^{\frac12}e_{h,i}^{k}\|^2_{\partial\Omega_i\setminus\partial\Omega}-\frac12\|(\delta^+)^{\frac12}e_{h,i}^{k+1}\|^2_{\partial\Omega_i\setminus\partial\Omega}\Bigg].
\end{aligned}
$$

It should be noted that the right-hand side of (3.11) consists of terms that are either telescoping or of one of the following forms:

- terms containing $\nabla e_{h,i}^{k+1}$;
- terms containing a part $[e_h^{k+1}]$;
- terms containing a part $e_{h,i}^{k+1}-e_{h,i}^{k}$.

The first and second contributions will be controlled by the triple norm, and the last type of contributions will be controlled by the relaxation terms of (3.7). More precisely, by replacing (3.11) and (3.7) in (3.6) we obtain

$$
\begin{aligned}
(3.12)\quad & \sum_{i=1,2}\Bigg[c\|\sigma_0^{\frac12}e_{h,i}^{k+1}\|^2_{\Omega_i}+cJ(e_{h,i}^{k+1},e_{h,i}^{k+1})\\
&+\left(\frac{\gamma_{it}}{2}-\frac{\|w_i\varepsilon_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}}{\alpha_i}\right)\|\tilde{h}^{-\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&\qquad-\frac{1-2\mu_i}{4\mu_i}\|\delta^{\frac12}(e_{h,i}^{k+1}-e_{h,i}^{k})\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&+(c-\alpha_iC_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)})\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k+1}\|^2_{\Omega_i}-\alpha_iC_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\|\varepsilon_i^{\frac12}\nabla e_{h,i}^{k}\|^2_{\Omega_i}\\
&+c\|\delta^{\frac12}[e_h^{k+1}]\|^2_{\partial\Omega_i\cap\partial\Omega}+(c-\mu_i)\|\delta^{\frac12}[e_h^{k+1}]\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&+\frac12\|(\delta^+)^{\frac12}e_{h,i}^{k+1}\|^2_{\partial\Omega_i\setminus\partial\Omega}-\frac12\|(\delta^+)^{\frac12}e_{h,i}^{k}\|^2_{\partial\Omega_i\setminus\partial\Omega}\\
&+\frac12\left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac12}e_{h,i}^{k+1}\right\|^2_{\partial\Omega_i\setminus\partial\Omega}-\frac12\left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac12}e_{h,i}^{k}\right\|^2_{\partial\Omega_i\setminus\partial\Omega}\Bigg]\le0.
\end{aligned}
$$

Then we choose the coefficients of Young's inequality, α_i and μ_i, as follows:

$$\alpha_i < \frac{c}{2C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}},\ \text{e.g., } \alpha_i = \frac{c}{4C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}};\quad \mu_i < c,\ \text{e.g., } \mu_i = \frac{c}{2};$$

and as a consequence of that, the relaxation parameter γ_{it} becomes

$$\gamma_{it} \geq \frac{8C_t\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\|w_i\varepsilon_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}}{c} + \frac{\max[1-c,0]}{2c}\left(\gamma_{bc}\|\{\varepsilon\}_w\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)} + \|\beta\cdot n\tilde{h}\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\right),\ i=1,2. \tag{3.13}$$

This allows us to rewrite (3.12) as follows:

$$\frac{c}{2}\left|\left|\left|e_h^{k+1}\right|\right|\right|^2 + \sum_{i=1,2}\left[\frac{c}{4}\left\|\varepsilon_i^{\frac{1}{2}}\nabla e_{h,i}^{k+1}\right\|_{\Omega_i}^2 - \frac{c}{4}\left\|\varepsilon_i^{\frac{1}{2}}\nabla e_{h,i}^{k}\right\|_{\Omega_i}^2 + \frac{1}{2}\left\|(\delta^+)^{\frac{1}{2}}e_{h,i}^{k+1}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 - \frac{1}{2}\left\|(\delta^+)^{\frac{1}{2}}e_{h,i}^{k}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 + \frac{1}{2}\left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^{k+1}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 - \frac{1}{2}\left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^{k}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2\right] \leq 0.$$

Finally, summing up from $k=0$ to $k=M-1$, we obtain

$$c\sum_{k=0}^{M-1}\left|\left|\left|e_h^{k+1}\right|\right|\right|^2 + \sum_{i=1,2}\left(\left\|(\delta^+)^{\frac{1}{2}}e_{h,i}^{M}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 + \left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^{M}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2\right) \leq \sum_{i=1,2}\left(\frac{c}{2}\left\|\varepsilon_i^{\frac{1}{2}}\nabla e_{h,i}^{0}\right\|_{\Omega_i}^2 + \left\|(\delta^+)^{\frac{1}{2}}e_{h,i}^{0}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2 + \left\|\left(\frac{\gamma_{it}}{\tilde{h}}\right)^{\frac{1}{2}}e_{h,i}^{0}\right\|_{\partial\Omega_i\setminus\partial\Omega}^2\right),$$

which implies (3.3). ∎

Remark 3.3. The general statement (3.13) implies the following choices of γ_{it}.

When $\varepsilon_i > 0$ for $i=1,2$ we have $c=\frac{1}{2}$ and $\gamma_{bc} \geq 2C_t$ in order to ensure coercivity. Then we insert $\gamma_{bc} = 2C_t$ into (3.13) and obtain

$$\gamma_{it} \geq 2C_t\left(1+8\|w_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}\right)\|w_i\varepsilon_i\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)} + \|\beta\cdot n\tilde{h}\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)},\ i=1,2.$$

For sufficiently small $\tilde{h}$ this expression can be summarized as $\gamma_{it} \simeq \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$. When $\varepsilon_1 = 0$ and $\varepsilon_2 > 0$ (or vice versa) we have $w_1 > 0$ and $w_2 = 0$. As a result of that the formula above becomes

$$\gamma_{it} \geq \frac{1}{2}\|\beta\cdot n\tilde{h}\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)},\ \text{i.e., } \frac{\gamma_{it}}{h} \geq \frac{1}{2}\|\beta\cdot n\|_{L^\infty(\partial\Omega_i\setminus\partial\Omega)}.$$

When $\varepsilon_1 = \varepsilon_2 = 0$ the coercivity constant becomes $c=1$. As a result of that (3.13) requires $\gamma_{it} \geq 0$.

4. Numerical results.

All the numerical experiments presented in this section were obtained using the `FreeFem++` library (http://www.freefem.org/ff++/index.htm).

4.1. Approximation and convergence properties of the iterative splitting method.

In this section we analyze the convergence of the iterative splitting method with respect to the mesh size $h = \max_{i=1,2}\max_{K\in\mathcal{T}_{h,i}} h_K$, the number of

TABLE 4.1
Convergence study with respect to h for conforming meshes.

Two subdomains, $h = 0.1$.

	P_1 FEM		P_2 FEM	
$\varepsilon = 1$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$
$2h$	$2.44\ 10^{-2}$	$5.82\ 10^{-1}$	$3.37\ 10^{-4}$	$5.05\ 10^{-2}$
h	$5.59\ 10^{-3}$	$2.65\ 10^{-1}$	4.62E-005	$1.26\ 10^{-2}$
Order	2.19	1.17	2.95	2.07
$\varepsilon = 10^{-3}$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$
$2h$	$1.65\ 10^{-2}$	$5.97\ 10^{-1}$	$8.88\ 10^{-4}$	$6.04\ 10^{-2}$
h	$3.64\ 10^{-3}$	$2.73\ 10^{-1}$	$1.02\ 10^{-4}$	$1.47\ 10^{-2}$
Order	2.24	1.16	3.21	2.10
$\varepsilon = 0$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$
$2h$	$1.69\ 10^{-2}$	$6.13\ 10^{-1}$	$9.95\ 10^{-4}$	$6.32\ 10^{-2}$
h	$3.80\ 10^{-3}$	$2.82\ 10^{-1}$	$1.23\ 10^{-4}$	$1.57\ 10^{-2}$
Order	2.22	1.16	3.10	2.07

Four subdomains $h = 0.08$.

	P_1 FEM		P_2 FEM	
$\varepsilon = 1$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_{1,\Omega}$
$2h$	$1.50\ 10^{-2}$	$4.48\ 10^{-1}$	$1.76\ 10^{-4}$	$3.25\ 10^{-2}$
h	$3.38\ 10^{-3}$	$2.06\ 10^{-1}$	$1.65\ 10^{-5}$	$7.68\ 10^{-3}$
Order	2.15	1.12	3.42	2.08

subdomains N, and the value of the penalty parameters γ_{it}, γ_{bc}, and $\gamma_{ip,i}$, $i = 1, 2$, for different values of the diffusion parameters ε_i and of the transport field β. To this aim, we consider problem (1.1), where $\sigma = 1$ is fixed and f is chosen so that the exact solution is

$$u(x, y) = \exp(xy)\sin(\pi x)\sin(\pi y), \tag{4.1}$$

on a domain $\Omega =]0, 1[\times]0, 1[$ that has been split into $N = n^2$ subdomains such that $\bar{\Omega} = \cup_{i=1}^{N} \bar{\Omega}_i = \cup_{i_1,i_2=1}^{n} [(i_1 - 1)/n, i_1/n] \times [(i_2 - 1)/n, i_2/n]$, obtaining a checkerboard partition of size $H = 1/n$. The simplest case of two subregions $\bar{\Omega}_1 = [0, \frac{1}{2}] \times [0, 1]$ and $\bar{\Omega}_2 = [\frac{1}{2}, 1] \times [0, 1]$ is also addressed. For each subdomain, we introduce N quasi-uniform meshes $\mathcal{T}_{h,i}$ that can be either conforming or nonconforming on their interfaces, but for the tests presented here we consider conforming discretizations. For the comparison of different cases we choose $u_{h,i}^0 = 0$ for $i = 1, \ldots, N$ and consider a convergence test on the triple norm of the incremental error, namely, the iterations are stopped if $|||u_h^{k+1} - u_h^k|||/|||u_h^{k+1}||| \leq tol$.

First of all, we aim to verify with numerical experiments the infinitesimal order with respect to h provided by Theorem 2.12. Table 4.1 shows that the optimal order of convergence is preserved for both linear and quadratic conforming elements. From now on, we will denote for simplicity $\|\cdot\|_{1,\Omega} \equiv (\sum_{i=1}^{N} \|\cdot\|_{1,\Omega_i})^{\frac{1}{2}}$.

Second, we aim to investigate the influence on the convergence rate of the iterative method of the parameters γ_{bc} and γ_{it} that appear in (3.1), (3.2). We study the number of iterations that the method needs to satisfy a tolerance $tol = 10^{-6}$ on the relative incremental error for several combinations of γ_{bc} and γ_{it}. Table 4.2 suggests

TABLE 4.2

The number of iterations necessary to converge with respect to a tolerance tol $= 10^{-6}$ *and on a quasi-uniform mesh of size* $h = 0.1$ *and a partition in two subdomains. Several combinations of the parameters* γ_{bc} *and* γ_{it} *in the case of the symmetric (right) and skewsymmetric coupling term (left) are addressed. In this case* $\varepsilon = 1$ *and* $\beta = [1, 1]$.

γ_{it}/γ_{bc}	$2\ 10^{0}$	$2\ 10^{1}$	$2\ 10^{2}$	$2\ 10^{-2}$	$2\ 10^{-1}$	$2\ 10^{0}$	$2\ 10^{1}$	$2\ 10^{2}$
$2\ 10^{-3}$	100	802	>1000	36	25	107	809	>1000
$2\ 10^{-2}$	101	803	>1000	34	25	107	810	>1000
$2\ 10^{-1}$	109	809	>1000	25	23	116	816	>1000
$2\ 10^{0}$	188	873	>1000	107	116	195	880	>1000
$2\ 10^{1}$	874	>1000	>1000	809	815	880	>1000	>1000
$2\ 10^{2}$	>1000	>1000	>1000	>1000	>1000	>1000	>1000	>1000

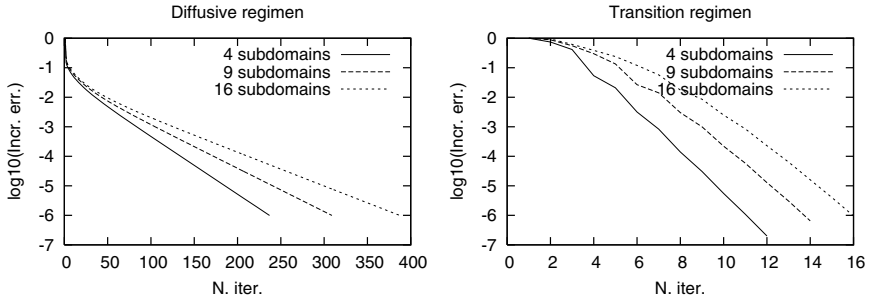


FIG. 4.1. *Convergence history of the iterative method for* $\varepsilon = 1.0$ *(right) and* $\varepsilon = 10^{-3}$ *(left) for the numerical tests on the coarse grids of Table* 4.3.

that an effective choice is to consider the small values of γ_{bc}, provided that the discrete problems (3.1), (3.2) remain well-posed according to Lemma 2.7. Recalling Remark 2.2, we analyze separately the symmetric and the nonsymmetric versions of the coupling term $B(u_h, v_h)$. In the symmetric case, Lemma 2.7 requires that $\gamma_{bc} = 2C_t$. In this case, Table 4.2 shows that the theoretical estimate obtained in Remark 3.3 is too restrictive for diffusion-dominated problems. Indeed, much smaller values of the estimated ones ensure better convergence properties. On the contrary, the numerical experiments presented in Table 4.5 suggest that the estimate of Remark 3.3 is effective for advection dominated problems. For the nonsymmetric case the limitations on γ_{bc} necessary for obtaining positivity of the discrete bilinear form change completely, in agreement with the analysis of interior penalty discontinuous Galerkin methods; see [1]. Indeed, only the restriction $\gamma_{bc} > 0$ is necessary. In this setting, the convergence properties of the iterative algorithm are much improved. Conversely, the approximation properties of the scheme are compromised since the discrete problem (2.1) is not adjoint consistent, and thus it does not enjoy optimal approximation properties in the L^2-norm (see [1] for a complete discussion). For the relaxation parameter γ_{it} we observe that in this case the choice $\gamma_{it} \simeq \gamma_{bc} \simeq 2\ 10^{-1}$ is effective.

The key point of this section is the characterization of the dependence of the convergence properties from the maximal mesh element size h and the number of subdomains N for different values of ε and β. More precisely, we analyze the diffusion dominated regimen ($\varepsilon = 1$), the transition regimen ($\varepsilon = 10^{-3}$), and the hyperbolic regimen ($\varepsilon = 0$). Indeed, Figure 4.1 and Table 4.3 show that the behavior of the method differs form one regimen to another. First of all, although Theorem 3.2 does

TABLE 4.3

The number of iterations necessary to converge with respect to a tolerance tol $= 10^{-6}$ *for several configurations of the partition in subdomains and several values of* ε. $\gamma_{bc} = 2$, $\gamma_{it} = \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$, *and* $\beta = [1, 1]$ *are fixed.*

H, N	1/2, 4		1/3, 9		1/4, 16	
h	0.13	0.06	0.12	0.06	0.12	0.07
$\varepsilon = 1$	237	445	309	579	388	723
Order h	-0.85		-0.91		-1.08	
Order H	–		-0.65		-0.77	
$\varepsilon = 10^{-3}$	12	14	14	17	16	20
Order h	-0.21		-0.28		-0.39	
Order H	–		-0.48		-0.56	
$\varepsilon = 0$	4	4	6	6	8	8
Order h	0		0		0	
Order H	–		-1		-1	

TABLE 4.4

The number of iterations necessary to converge with respect to a tolerance tol $= 10^{-6}$ *for different combinations of* β *and* ε *and for the case of* 16 *subdomains and* $h = 0.12$. $\gamma_{bc} = 2$, $\gamma_{it} = \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$ *are fixed.*

	$\varepsilon = 1$	$\varepsilon = 10^{-3}$	$\varepsilon = 0$
$\beta = [1, 1]$	388	16	8
$\beta = [1, 0]$	389	16	5
$\beta = [0, 1]$	389	16	5

not characterize the convergence behavior of the iterative method (3.1)–(3.2), Figure 4.1 puts into evidence that the incremental error is reduced according to the law C^k, where k is the iteration index and the constant $0 < C < 1$ is the convergence rate. Following this assumption, the number of iterations needed to satisfy a suitable tolerance on the incremental error is directly proportional to the convergence rate. As a consequence of that, Table 4.3 shows that in the diffusion-dominated regimen the convergence rate is inversely proportional to h and H. Following the heuristic motivations that are presented in [22] and [27], the inverse dependence on H can be explained observing that an iterative method that only exchanges information between neighboring subregions necessarily requires a number of steps to converge that is at least equal to the diameter of the dual graph corresponding to the subdomain partition, which is equivalent to $\mathcal{O}(H^{-1})$ when the diameter of Ω is unitary. The dual graph is constructed by introducing a vertex for each subregion and an edge between two subregions that share an interface. The inverse dependence on h is a consequence of (3.13) (see also Remark 3.3) which states that the relaxation term must be proportional to $\|\varepsilon\|_{L^\infty(\Omega)}/h$. Accordingly, by refining the mesh by a factor two, the number of iterations is doubled. Always in agreement with Remark 3.3 and with the fact that the relaxation term is allowed to vanish together with ε, the convergence rate of the method is less sensitive with respect to h for the transition case and completely insensitive with respect to the mesh size in the hyperbolic case. Indeed, when $\varepsilon = 0$ the number of iterations is only inversely dependent on H, and it is exactly equivalent to the number of steps that are needed to propagate the information along the diagonal of the checkerboard mesh defined by the subdomains, since the transport field is oriented along the diagonal. Furthermore, Table 4.4 suggests that

these results do not deteriorate if the orientation of the transport field β is modified. Indeed, this is an advantage of the method proposed here with respect to the family of nonoverlapping domain decomposition methods arising from transmission conditions of Robin type, whose convergence may turn out to be slow when the transport field is tangential to the interface [21]. This benefit is due to the use of the upwind flux for the advection term. As a consequence of that, the corresponding transmission conditions are not symmetric with respect to β, in contrast to what happens for the family of methods inspired by transmission conditions of Robin type. Finally, we observe that in the hyperbolic case a multiplicative (Gauss–Seidel) iterative scheme is more preforming than the additive (Jacobi) method. For instance, since the subdomains in the checkerboard partition have been numbered by rows, when the transport field β is oriented in the vertical direction the multiplicative algorithm converges in 2 iterations, irrespectively of h and H.

4.2. Comparison of iterative methods. In order to assess the performance of the iterative method based on Nitsche's transmission conditions (denoted with a in Table 4.5 and defined by problems (3.1) and (3.2)) we compare it with the nonoverlapping Schwarz method proposed in [20] (denoted with b) and with the overlapping Schwarz method (denoted with c). For this comparison, we consider the test case proposed in the previous section where the domain Ω has been split into two subdomains, $\Omega_1 = [0, \frac{1}{2}] \times [0, 1]$ and $\Omega_2 = [\frac{1}{2}, 1] \times [0, 1]$. In the case of the overlapping Schwarz method we also introduce two overlapping domains, $\Omega_1^* = [0, \frac{1}{2} + \frac{1}{2}\delta] \times [0, 1]$, $\Omega_2^* = [\frac{1}{2} - \frac{1}{2}\delta, 1] \times [0, 1]$, and corresponding discretizations $\mathcal{T}_{h,i}^*$, $i = 1, 2$. Let $V_{h,i}^*$ be the finite element spaces defined on these meshes. Then, given $u_{h,i}^0$, for $k = 1, 2, \ldots$ we look for $u_{h,i}^k \in V_{h,i}^*$, $i = 1, 2$, such that

$$
\begin{aligned}
A(u_{h,1}^{k+1}, v_{h,1}) + J(u_{h,1}^{k+1}, v_{h,1}) &= (f_1, v_{h,1}) \ \forall v_{h,1} \in V_{h,1}^*, \quad \hat{u}_{h,1}^{k+1} = u_{h,2}^k \text{ on } \partial\Omega_1^* \cap \Omega_2^*, \\
A(u_{h,2}^{k+1}, v_{h,2}) + J(u_{h,2}^{k+1}, v_{h,2}) &= (f_2, v_{h,2}) \ \forall v_{h,2} \in V_{h,2}^*, \quad \hat{u}_{h,2}^{k+1} = u_{h,1}^k \text{ on } \partial\Omega_2^* \cap \Omega_1^*, \\
u_{h,i}^{k+1} &= \frac{1}{2}\hat{u}_{h,i}^{k+1} + \frac{1}{2}u_{h,i}^k \ i = 1, 2.
\end{aligned}
$$

Recalling that the convergence of the overlapping Schwarz method can be accelerated by increasing the thickness of the overlapping region, that is, δ, we consider three cases, $\delta = \bar{h}$, $\delta = 2\bar{h}$, and $\delta = 4\bar{h}$, where $\bar{h}$ is the characteristic size of the quasi-uniform discretizations of Ω_1^* and Ω_2^*. The comparison with these cases will give a measure of the convergence performance of our method.

In Table 4.5, we compare the convergence and the approximation properties of these methods for the diffusion-dominated, the transition, and the hyperbolic regimens. The analysis of this table immediately shows that the method that we propose here is effective for the advection dominated and the hyperbolic regimens. In this case Nitsche's method a provides in general the best performances both for the convergence and the approximation properties for a fixed tolerance on the incremental error $tol = 10^{-6}$ and a given quasi-uniform mesh with $h = 0.05$.

In the diffusion-dominated case, the convergence of method a in the symmetric case is partially slowed down by the relaxation term. We have already observed that the choice $\gamma_{it} = \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$, motivated by the theoretical estimate derived in Remark 3.3, is not optimal. Indeed, the number of iterations needed to fulfill a tolerance of 10^{-6} on the incremental error is reduced from 354 to 190 if the parameter γ_{it} is divided by a factor of 100. In any case, this correction does not make method a with $s = 1$ (see Remark 2.2) competitive with method b in the diffusion-dominated

TABLE 4.5

The number of iterations necessary to converge with respect to a tolerance tol $= 10^{-6}$ and the approximation error on a given quasi-uniform mesh characterized by $h = 0.05$ and a partition in 2 subdomains. Several instances of the iterative algorithms a, b, and c are considered. The instance of algorithm a with symmetric coupling terms is denoted with $s = 1$, while the nonsymmetric version is denoted with $s = -1$.

Diffusion dominated regimen $\varepsilon = 1$, $\beta = [1, 1]$.

Method	N. iter.	$\|u - u_h\|_{0,\Omega}$	$\|u - u_h\|_{1,\Omega}$
a, $s = 1$, $\gamma_{bc} = 2$, $\gamma_{it} = \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$	354	$1.38\ 10^{-3}$	$1.32\ 10^{-1}$
a, $s = 1$, $\gamma_{bc} = 2$, $\gamma_{it} = 10^{-2}\gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$	190	$1.38\ 10^{-3}$	$1.32\ 10^{-1}$
a, $s = -1$, $\gamma_{bc} = 2\ 10^{-1}$, $\gamma_{it} = 2\ 10^{-1}$	43	$1.93\ 10^{-3}$	$1.25\ 10^{-1}$
a-hybrid	108	$1.37\ 10^{-3}$	$1.32\ 10^{-1}$
b	96	$1.37\ 10^{-3}$	$1.32\ 10^{-1}$
c, $\delta = \bar{h}$	210	$3.25\ 10^{-3}$	$1.27\ 10^{-1}$
c, $\delta = 2\bar{h}$	115	$2.35\ 10^{-3}$	$1.28\ 10^{-1}$
c, $\delta = 4\bar{h}$	65	$2.02\ 10^{-3}$	$1.36\ 10^{-1}$

Transition regimen $\varepsilon = 10^{-3}$, $\beta = [1, 1]$, $\gamma_{bc} = 2$, and $\gamma_{it} = \gamma_{bc}\|\varepsilon\|_{L^\infty(\Omega)}$.

Method	N. iter.	$\|u - u_h\|_{0,\Omega}$	$\|u - u_h\|_{1,\Omega}$
a, $s = 1$	12	$8.76\ 10^{-4}$	$1.33\ 10^{-1}$
a, $s = -1$	13	$8.75\ 10^{-4}$	$1.33\ 10^{-1}$
b	17	$1.03\ 10^{-3}$	$1.47\ 10^{-1}$
c, $\delta = \bar{h}$	46	$1.00\ 10^{-3}$	$1.37\ 10^{-1}$
c, $\delta = 2\bar{h}$	56	$1.27\ 10^{-3}$	$1.41\ 10^{-1}$
c, $\delta = 4\bar{h}$	42	$1.21\ 10^{-3}$	$1.51\ 10^{-1}$

Hyperbolic regimen $\varepsilon = 0$, $\beta = [1, 1]$, $\gamma_{bc} = 2$, and $\gamma_{it} = 0$.

Method	N. iter.	$\|u - u_h\|_{0,\Omega}$	$\|u - u_h\|_{1,\Omega}$
a, $s = \pm 1$	2	$9.48\ 10^{-4}$	$1.40\ 10^{-1}$
b	57	$2.44\ 10^{-3}$	$2.96\ 10^{-1}$
c, $\delta = \bar{h}$	52	$1.10\ 10^{-3}$	$1.45\ 10^{-1}$
c, $\delta = 2\bar{h}$	59	$1.48\ 10^{-3}$	$1.52\ 10^{-1}$
c, $\delta = 4\bar{h}$	45	$1.39\ 10^{-3}$	$1.63\ 10^{-1}$

case. Conversely, we observe that the convergence properties of the nonsymmetric version of method a is very satisfactory, while the approximation error in the L^2-norm reflects the suboptimality of this method. By comparing the properties of the symmetric and the nonsymmetric versions of method a, we observe that it may be possible to blend the benefits of the two methods by setting up a hybrid strategy (see Table 4.5, method a-hybrid). This consists in applying method a with $s = -1$, $\gamma_{bc} = \gamma_{it} = 2\ 10^{-1}$ until the tolerance equal to 10^{-6} is satisfied on the relative incremental error. As reported in Table 4.5, this procedure requires 43 iterations. Then, starting from the discrete solution computed in this way, we apply method a with $s = 1$, $\gamma_{bc} = 2$, $\gamma_{it} = 2\ 10^{-1}$ in order to improve the approximation error. This method requires 65 additional iterations to converge, and it reduces the L^2 approximation error of the nonsymmetric case from $1.93\ 10^{-3}$ to $1.37\ 10^{-3}$, which is equivalent to the error of the symmetric case. Since it is accurate and converges rapidly, the hybrid method outperforms both the symmetric and the nonsymmetric versions of method a. In the diffusive case, the hybrid method turns out to be almost

equivalent to method b. These considerations promote further studies of the hybrid method and suggest investigating in detail whether the nonsymmetric formulation might be applied as a preconditioner for the symmetric case. Finally, a heuristic comparison with the overlapping Schwarz methods c suggests that method b behaves as an additive overlapping Schwarz algorithm with a relatively generous overlap of magnitude $\delta = 2\bar{h} \equiv 6\%$ of the diameter of Ω. On the other hand, the symmetric Nitsche's method a is almost equivalent to the overlapping method with small overlap $\delta = \bar{h} \equiv 3\%$.

From the point of view of computational cost we observe that the scheme (3.1)–(3.2) requires more effort for the construction of the finite element matrix corresponding to the coupling terms $B(u_h, v_h)$ than the family of Robin–Robin methods. Indeed, for the Robin–Robin methods the coupling matrix is easily constructed since it corresponds to a mass matrix on the degrees of freedom at the interface. Moreover in our case the bandwidth of the coupling matrix is increased because of the presence of first order derivatives in the coupling terms. This drawback is balanced by the fact that basic Robin–Robin iterative splitting methods preserve the optimal approximation properties of Lagrangian finite elements only if a superpenalty technique is applied; see [8]. This technique, however, compromises the convergence properties of the iterative algorithm.

4.3. Approximation of problems with discontinuous coefficients. In this section, we apply the numerical scheme (3.1)–(3.2) for the approximation of advection diffusion problems with discontinuous coefficients. To this purpose, the domain Ω has been split into two subdomains, $\Omega_1 = [0, \frac{1}{2}] \times [0, 1]$ and $\Omega_2 = [\frac{1}{2}, 1] \times [0, 1]$ with $\varepsilon(x) = 1.0$ for $x \in \Omega_1$ and $\varepsilon(x) = 2\ 10^{-2}$ for $x \in \Omega_2$. In the case $\sigma = 0$ and $f = 0$, the exact solution on each subregion Ω_1, Ω_2 can be easily expressed as an exponential function with respect to the x coordinate independently from the y coordinate. The global solution $u(x, y)$ is provided by choosing the value at the interface $x = \frac{1}{2}$ in order to ensure the following matching conditions:

$$\lim_{x \to \frac{1}{2}^-} u(x, y) = \lim_{x \to \frac{1}{2}^+} u(x, y) \quad \text{and} \quad \lim_{x \to \frac{1}{2}^-} -\varepsilon(x)\partial_x u(x, y) = \lim_{x \to \frac{1}{2}^+} -\varepsilon(x)\partial_x u(x, y).$$

More precisely, we set $u(0, y) = 1$, $u(1, y) = 0$, and by consequence of the matching conditions, we obtain

$$u\left(\frac{1}{2}, y\right) = \left[\frac{u(0, y)\exp(\frac{\beta}{2\varepsilon_1})}{1 - \exp(\frac{\beta}{2\varepsilon_1})} + \frac{u(1, y)}{1 - \exp(\frac{\beta}{2\varepsilon_2})}\right]\left[\frac{\exp(\frac{\beta}{2\varepsilon_1})}{1 - \exp(\frac{\beta}{2\varepsilon_1})} + \frac{1}{1 - \exp(\frac{\beta}{2\varepsilon_2})}\right]^{-1}.$$

As a result of that, the exact solution in each subdomain can be expressed as

$$u_1(x, y) = \frac{u(\frac{1}{2}, y) - \exp(\frac{\beta}{2\varepsilon_1})u(0, y) + [u(0, y) - u(\frac{1}{2}, y)]\exp(\frac{\beta x}{\varepsilon_1})}{1 - \exp(\frac{\beta}{2\varepsilon_1})},$$

$$u_2(x, y) = \frac{u(1, y) - \exp(\frac{\beta}{2\varepsilon_2})u(\frac{1}{2}, y) + [u(\frac{1}{2}, y) - u(1, y)]\exp(\frac{\beta(x - \frac{1}{2})}{\varepsilon_2})}{1 - \exp(\frac{\beta}{2\varepsilon_2})}.$$

The resulting function is represented in Figure 4.2. We aim to compare on the test problem defined above the accuracy of the scheme (3.1)–(3.2) with linear elements, precisely $V_h = \sum_{i=1}^{2} V_{h,1,i}$ (denoted by A) with the classical lagrangian linear elements over the whole domain Ω (denoted by B). We point out that in both cases the

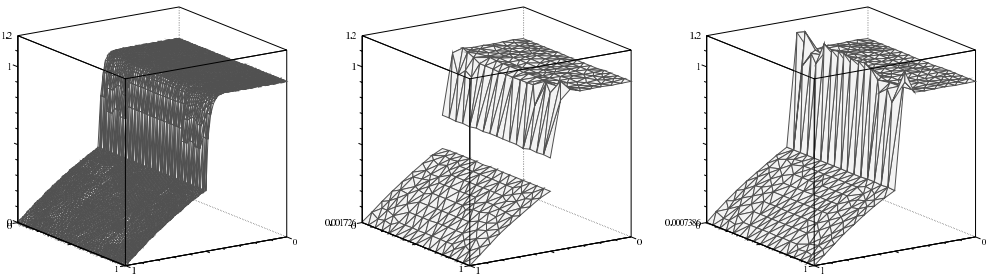


FIG. 4.2. *The nodal interpolant on a very refined mesh of the exact solution u of the test problem at hand (left). The numerical approximation u_h obtained with method A (middle) and method B (right) in the case of the discretization characterized by $h_1 = 0.1$.*

TABLE 4.6

The quantitative comparison of the accuracy of methods A and B. The L^2-norm, $\|u_h - u\|_{0,\Omega}$, the H^1-norm, $\|u_h - u\|_{1,\Omega}$, and the maximum norm, $\|u_h - u\|_{L^\infty(\Omega)}$, are displayed.

	$\|u - u_h\|_{0,\Omega}$		$\|u - u_h\|_{1,\Omega}$		$\|u - u_h\|_{L^\infty(\Omega)}$	
h	Method A	Method B	Method A	Method B	Method A	Method B
0.1	$1.81\ 10^{-2}$	$2.78\ 10^{-2}$	2.07	1.78	$2.13\ 10^{-1}$	$1.71\ 10^{-1}$
0.05	$6.98\ 10^{-3}$	$8.95\ 10^{-3}$	1.31	1.06	$1.07\ 10^{-1}$	$6.32\ 10^{-2}$
0.026	$2.56\ 10^{-3}$	$2.66\ 10^{-3}$	$7.49\ 10^{-1}$	$5.82\ 10^{-1}$	$1.40\ 10^{-1}$	$2.34\ 10^{-2}$

continuous interior penalty stabilization method with $\gamma_{ip,i} = 2\ 10^{-2}$ has been applied to cure the instability of finite elements in the case of advection-dominated problems. We compare the two schemes on a family of quasi-uniform triangulations on Ω_1 and Ω_2 that are conforming at the interface of the subdomains and are characterized by a decreasing maximal element size $h_1 = 0.1$, $h_2 = 0.05$, and $h_3 = 0.026$. The quantitative analysis of the accuracy is based on the following indicators: the L^2-norm of the error, $\|u_h - u\|_{0,\Omega}$; the H^1-norm, $\|u_h - u\|_{1,\Omega}$, which is well defined since $u \in H^1(\Omega)$; and the maximum norm, $\|u_h - u\|_{L^\infty(\Omega)}$. The quantitative data are reported in Table 4.6, while a visual comparison is given in Figure 4.2. The analysis of the results suggests that the scheme (3.1)–(3.2) performs well for the approximation of problems with discontinuous coefficients when the mesh size is not small enough to fully resolve the boundary layers arising in the neighborhood of the region of discontinuity. The benefit of the scheme presented here with respect to the application of classical Lagrangian elements over Ω emerges if we consider the L^2-norm. For the mesh size h_1 method A provides numerical solutions that are smoother than method B (see Figure 4.2), where spurious oscillations appear in the neighborhood of the boundary layer that arise because of the discontinuity of ε. However, we observe that the L^∞ error of method B is smaller than in the case of method A, since for this method L^∞ errors arise when the very steep boundary layer across the discontinuity of ε is approximated with a jump. Finally, the analysis of the H^1-norm of the errors suggests that method B seems to be more prone to approximate the gradients of the solution in the boundary layer, although this benefit is effective when the computational mesh becomes fine enough to reasonably approximate the boundary layer.

5. Concluding remarks. In conclusion, the discretization scheme and the associated iterative method that we have proposed here turn out to be appealing for advection-dominated problems and in the case of discontinuous coefficients. Indeed,

in these cases the method is competitive from the point of view of both computational effort and accuracy. A key role for the good properties when treating such problems is played by the average weights and the upwind treatment of the advection term in the interior penalty strategy applied for the coupling of the subdomains.

Acknowledgment. The authors are grateful for the referee's detailed and constructive criticisms and for the timely management of the manuscript.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] R. Becker, P. Hansbo, and R. Stenberg, *A finite element method for domain decomposition with non-matching grids*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 209–225.

[3] M. Boman, *Estimates for the L_2 projection onto continuous finite element spaces in a weighted L_p-norm*, BIT, 46 (2006), to appear.

[4] J. Bramble, J. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 71 (2002), pp. 147–156.

[5] E. Burman, *A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty*, SIAM J. Numer. Anal., 43 (2005), pp. 2012–2033.

[6] E. Burman, M. Fernández, and P. Hansbo, *Edge Stabilization: An Interior Penalty Method for the Incompressible Navier-Stokes Equation*, Tech. Rep. 23.2004, Ecole Polytechnique Federale de Lausanne, 2004.

[7] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–2453.

[8] T. Chacón-Rebollo and E. Chacón Vera, *Study of a non-overlapping domain decomposition method: Poisson and Stokes problems*, Appl. Numer. Math., 48 (2004), pp. 169–194.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, PA, 2002.

[10] J.-P. Croisille, A. Ern, T. Lelievre, and J. Proft, *Analysis and simulation of a coupled hyperbolic/parabolic model problem*, J. Numer. Math., 13 (2005), pp. 81–103.

[11] M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[12] J. Douglas, Jr. and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences (Second International Symposium, Versailles, 1975), Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216.

[13] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* II. *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.

[14] F. Gastaldi and A. Quarteroni, *On the coupling of hyperbolic and parabolic systems: Analytical and numerical approach*, Appl. Numer. Math., 6 (1989/90), pp. 3–31.

[15] P. Hansbo and M. G. Larson, *Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1895–1908.

[16] B. Heinrich and K. Pietsch, *Nitsche type mortaring for some elliptic problem with corner singularities*, Computing, 68 (2002), pp. 217–238.

[17] R. Hoppe and B. Wohlmuth, *Element-oriented and edge-oriented local error estimators for nonconforming finite element methods*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 237–263.

[18] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[19] C. Lasser and A. Toselli, *An overlapping domain decomposition preconditioner for a class of discontinuous Galerkin approximations of advection-diffusion problems*, Math. Comp., 72 (2003), pp. 1215–1238.

[20] G. Lube, L. Müller, and F. C. Otto, *A non-overlapping domain decomposition method for the advection-diffusion problem*, Computing, 64 (2000), pp. 49–68.

[21] F. Nataf and F. Rogier, *Factorization of the convection-diffusion operator and the Schwarz algorithm*, Math. Models Methods Appl. Sci., 5 (1995), pp. 67–93.

[22] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Numer. Math. Sci. Comput., Clarendon Press, Oxford University Press, New York, 1999.

[23] G. Rapin and G. Lube, *A stabilized three-field formulation for advection-diffusion equations*, Computing, 73 (2004), pp. 155–178.
[24] R. Stenberg, *Mortaring by a method of J. A. Nitsche*, in Computational Mechanics (Buenos Aires, 1998), CD-ROM, Centro Internac. Métodos Numér. Ing., Barcelona, 1998.
[25] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 1997.
[26] A. Toselli, *hp-finite element approximations on non-matching grids for partial differential equations with non-negative characteristic form*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 91–115.
[27] A. Toselli and O. Widlund, *Domain decomposition methods—algorithms and theory*, Springer Ser. Comput. Math. 34, Springer-Verlag, Berlin, 2005.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1639–1654

SUBGRID STABILIZED DEFECT CORRECTION METHODS FOR THE NAVIER–STOKES EQUATIONS*

SONGUL KAYA†, WILLIAM LAYTON‡, AND BÉATRICE RIVIÈRE‡

Abstract. We consider the synthesis of a recent subgrid stabilization method with defect correction methods. The combination is particularly efficient and combines the best algorithmic features of each. We prove convergence of the method for a fixed number of corrections as the mesh size goes to zero and derive parameter scalings from the analysis. We also present some numerical tests which both verify the theoretical predictions and illustrate the method's promise.

Key words. eddy viscosity, variational multiscale method, high Reynolds numbers, correction steps

AMS subject classifications. 65N30, 65N15, 76D05

DOI. 10.1137/050623942

1. Introduction.

This report studies the synthesis of defect correction methods and subgrid stabilization. Our proposed method adds an eddy viscosity stabilization on only the last few resolved scales on arbitrary, unstructured meshes. Computational considerations for total algorithmic efficiency suggest combining the stabilization method with defect correction when solving *underresolved, equilibrium flow problems*. In this work, we study precisely this combination in that context. We analyze convergence of the combination for the (nonlinear) Navier–Stokes equations. This analysis gives mathematical guidance on the selection of the method's algorithmic parameters. In our accompanying tests, we observe that the subgrid stabilized defect correction method has greater accuracy than the artificial viscosity method without the oscillations reported in the usual (centered) Galerkin finite element method or the unmodified defect correction finite element method.

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ *or* 3) denote a bounded, regular flow domain. We consider the approximate solution of the Navier–Stokes equations for internal flow on Ω: find $u : \Omega \to \mathbb{R}^d$, $p : \Omega \to \mathbb{R}$ satisfying

$$\begin{aligned} -\nu\Delta u + (u\cdot\nabla)u + \nabla p &= f \quad \text{in } \Omega, \\ \nabla\cdot u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \\ \textstyle\int_\Omega p\, dx &= 0. \end{aligned} \tag{1.1}$$

In (1.1), the coefficient ν is the kinematic viscosity of the fluid, and $f \in L^2(\Omega)^d$ is the body force driving the flow.

Let $(\cdot,\cdot)$, $\|\cdot\|$ denote the usual L^2 inner product and norm, respectively. Define $X := H_0^1(\Omega) := \{v \in L^2(\Omega)^d : \nabla v \in L^2(\Omega)^{d\times d}$ and $v = 0$ on $\partial\Omega\}$, $Q := L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q\, dx = 0\}$, and define $b^*(u,v,w) := \frac{1}{2}(u\cdot\nabla v, w) - \frac{1}{2}(u\cdot\nabla w, v)$ for all

*Received by the editors February 7, 2005; accepted for publication (in revised form) January 11, 2006; published electronically August 16, 2006.
http://www.siam.org/journals/sinum/44-4/62394.html

†Department of Mathematics, Middle East Technical University, Ankara 06531, Turkey (songul@math.metu.edu.tr). This author was partially supported by NSF grant 0207627.

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (wjl@pitt.edu, riviere@math.pitt.edu). The second author was partially supported by NSF grant DMS 0207627 and DMS 0508260. The third author was partially supported by NSF grant DMS 0506039.

$u, v, w \in X$. Integrating by parts gives the following standard variational formulation of (1.1): find $u \in X$, $p \in Q$ satisfying

$$\begin{aligned} \nu(\nabla u, \nabla v) + b^*(u,u,v) - (p, \nabla \cdot v) &= (f,v) \quad \forall v \in X, \\ (\nabla \cdot u, q) &= 0 \qquad \forall q \in Q. \end{aligned} \tag{1.2}$$

For the finite element discretization, we choose the conforming velocity-pressure finite element spaces, $X_h \subset X$ and $Q_h \subset Q$, satisfying the discrete inf-sup condition

$$\inf_{q_h \in Q_h} \sup_{v_h \in X_h} \frac{(q_h, \nabla \cdot v_h)}{\|q_h\| \|\nabla v_h\|} \geq \beta > 0, \tag{1.3}$$

where β is independent of h.

The stabilization we consider requires a coarser finite element velocity space, $X_H \subset X$, corresponding to the large scales of the fluid velocity. Since finite element spaces are constructed based upon triangulations of the domain Ω, typically (although not necessarily for our analysis) $X_H \subset X_h \subset X$. We then define the following space:

$$L_H = \nabla X_H \subset L = L^2(\Omega)^{d \times d}. \tag{1.4}$$

The stabilized finite element method we consider herein is: find $u_h \in X_h$, $p_h \in Q_h$, and $g_H \in L_H$ satisfying

$$\begin{aligned} (\nu + \alpha)(\nabla u_h, \nabla v_h) + b^*(u_h, u_h, v_h) & \\ -\alpha(g_H, \nabla v_h) - (p_h, \nabla \cdot v_h) &= (f, v_h) \quad \forall v_h \in X_h, \\ (\nabla \cdot u_h, q_h) &= 0 \qquad \forall q_h \in Q_h, \\ (g_H - \nabla u_h, l_H) &= 0 \qquad \forall l_H \in L_H, \end{aligned} \tag{1.5}$$

where α is the user-selected stabilization parameter and typically, $\alpha = \mathcal{O}(h)$. It is easy to verify that the last equality in (1.5) implies that g_H is the L^2 projection of ∇u^h, denoted by $\overline{\nabla u_h}$.

In a typical implementation of (1.5), the variables g_H, l_H in L_H are defined on macroelements, i.e., elements of the coarse mesh. Thus, solving (1.5) involves coupling of microvariables (functions in X_h, Q_h) across macroelements. Thus, although these terms are cheap to evaluate in a residual calculation, the bandwidth of the linearized system arising from (1.5) increases substantially, and the solution of the linear system containing these terms is not attractive. This issue is discussed briefly in Layton [31] and at some length in John, Kaya, and Layton [27] and Anitescu, Layton, and Pahlevani [2] for the evolutionary problem.

For this reason, we consider a further defect correction discretization of (1.2) herein. This combination greatly increases efficiency by shifting the macro-micro coupling to a residual calculation. The method consists of an initialization step followed by k-correction steps, where k is the local polynomial degree of X_h.

Initialization step. Solve for $(u_h^1, p_h^1) \in (X_h, Q_h)$ such that

$$\begin{aligned} (\nu + \alpha)(\nabla u_h^1, \nabla v_h) + b^*(u_h^1, u_h^1, v_h) - (p_h^1, \nabla \cdot v_h) &= (f, v_h) \quad \forall v_h \in X_h, \\ (\nabla \cdot u_h^1, q_h) &= 0 \qquad \forall q_h \in Q_h. \end{aligned} \tag{1.6}$$

k-correction steps. Given $(u_h^j, p_h^j) \in (X_h, Q_h)$ for $j = 1, 2 \ldots, k$, solve for $(u_h^{j+1},$

$p_h^{j+1}) \in (X_h, Q_h)$ satisfying

$$
\begin{aligned}
&(\nu+\alpha)(\nabla(u_h^{j+1}-u_h^j),\nabla v_h)+b^*(u_h^{j+1},u_h^{j+1},v_h)-b^*(u_h^j,u_h^j,v_h)-(p_h^{j+1}-p_h^j,\nabla\cdot v_h)\\
&=(f,v_h)-[(\nu+\alpha)(\nabla u_h^j,\nabla v_h)+b^*(u_h^j,u_h^j,v_h)-(p_h^j,\nabla\cdot v_h)-\alpha(g_H^j,\nabla v_h)]\quad\forall v_h\in X_h,\\
&\qquad\qquad(\nabla\cdot u_h^{j+1},q_h)=0\quad\forall q_h\in Q_h,\\
&\qquad\qquad(g_H^j-\nabla u_h^j,l_H)=0\quad\forall l_H\in L_H.
\end{aligned}
\tag{1.7}
$$

Remark 1.1. It is typical that while defect correction methods are algorithmically simple to implement, they are cumbersome to write (as above) and can resist analysis. There are also several equivalent formulations of (1.7) and several algorithmic options within the defect correction idea. We stress that this is not an iteration: only k steps are performed, where k is the local polynomial degree of X_h. Thus, an asymptotic analysis as $j\to\infty$ is irrelevant; we analyze herein the method as $h\to 0$ for fixed j. The algorithmic efficiency of the defect correction method (1.6), (1.7) can be seen by rewriting (1.7) as follows: find $(u_h^{j+1},p_h^{j+1})\in(X_h,Q_h)$ satisfying

$$
\begin{aligned}
&(\nu+\alpha)(\nabla u_h^{j+1},\nabla v_h)+b^*(u_h^{j+1},u_h^{j+1},v_h)\\
&\qquad-(p_h^{j+1},\nabla\cdot v_h)=(f,v_h)+\alpha(g_H^j,\nabla v_h)\quad\forall v_h\in X_h,\\
&\qquad(\nabla\cdot u_h^{j+1},q_h)=0\quad\forall q_h\in Q_h,\\
&\qquad(g_H^j-\nabla u_h^j,l_H)=0\quad\forall l_H\in L_H.
\end{aligned}
\tag{1.8}
$$

Since u_h^j is known in (1.8), g_H^j is explicitly calculable by computing the L^2 projection operator of ∇u_h^j into L_H. Because the natural choice for L_H is $L_H=\nabla X_H$, L_H is typically a space of discontinuous piecewise polynomials of degree $k-1$ on a coarse mesh. Therefore, this projection calculation uncouples into one well-conditioned small linear system per coarse mesh element. Given g_H^j, the solution u_h^{j+1} then only involves solving an artificial viscosity discretization of the Navier–Stokes equations. If $\alpha=\mathcal{O}(h)$, this is known to lead to linearized systems which can be solved efficiently.

An alternative formulation of a defect correction method is to begin with nonlinear, stabilized artificial viscosity approximation (1.6) for (u_h^1,p_h^1) and then correct by solving the linearized problem instead of the nonlinear one. This has the advantage that only one linear solution is needed per correction step. This variation reads: given (u_h^j,p_h^j), find (u_h^{j+1},p_h^{j+1}) satisfying

$$
\begin{aligned}
&(\nu+\alpha)(\nabla(u_h^{j+1}-u_h^j),\nabla v_h)+b^*(u_h^j,u_h^{j+1}-u_h^j,v_h)+b^*(u_h^{j+1}-u_h^j,u_h^j,v_h)\\
&\qquad-(p_h^{j+1}-p_h^j,\nabla\cdot v_h)=(f,v_h)-[(\nu+\alpha)(\nabla u_h^j,\nabla v_h)+b^*(u_h^j,u_h^j,v_h)\\
&\qquad\qquad-(p_h^j,\nabla\cdot v_h)-\alpha(g_H^j,\nabla v_h)]\quad\forall v_h\in X_h,\\
&\qquad\qquad(\nabla\cdot u_h^{j+1},q_h)=0\quad\forall q_h\in Q_h,\\
&\qquad\qquad(g_H^j-\nabla u_h^j,l_H)=0\quad\forall l_H\in L_H.
\end{aligned}
\tag{1.9}
$$

The correction (1.9) is in the familiar residual-update form. It can be simplified to read: find (u_h^{j+1},p_h^{j+1}) satisfying

$$
\begin{aligned}
&(\nu+\alpha)(\nabla u_h^{j+1},\nabla v_h)+b^*(u_h^j,u_h^{j+1},v_h)+b^*(u_h^{j+1},u_h^j,v_h)\\
&-(p_h^{j+1},\nabla\cdot v_h)=(f,v_h)+b^*(u_h^j,u_h^j,v_h)+\alpha(g_H^j,\nabla v_h)\quad\forall v_h\in X_h,\\
&\qquad\qquad(\nabla\cdot u_h^{j+1},q_h)=0\quad\forall q_h\in Q_h,\\
&\qquad\qquad(g_H^j-\nabla u_h^j,l_H)=0\quad\forall l_H\in L_H.
\end{aligned}
\tag{1.10}
$$

1.1. Literature review for the defect correction method. The idea of defect correction is simple and universal. In its initial form, it was considered an algorithmically efficient way to perform Richardson's extrapolation, e.g., Stetter [38]. Since most practical problems do not have enough regularity, the practical importance was not recognized until the work of Hemker [21, 20] and Hemker and Koren [23, 22]. One current view of the defect correction method is that it allows for a solution that is nearly nonsingular for ill-conditioned problems through stabilization and correction; for a sample of recent works, see, e.g., Altase and Burrage [1], Axelsson and Nikolova [4], Juncu [28], Graziadei, Mattheij, and Boonkkamp [16], Heinrichs [19, 18], Desideri and Hemker [6], Nefedov and Mattheij [34], Shaw and Crumpton [37]. For example, when applied to viscoelastic fluid flow (Lee [32]), the defect correction method proved to be the key algorithmic idea for computing with a Weissenberg number beyond which other algorithms failed.

Much analytical insight into defect correction methods was obtained early for periodic constant coefficient problems by local model analysis. The first complete convergence theory for defect correction methods for convection dominated problems in one dimension was performed in Ervin and Layton [8] in which uniform epsilon convergence was proven away from layers. This result was extended to higher dimensions, higher order methods, and unstructured meshes in Axelsson and Layton [3]. Recently, global uniform in epsilon convergence on Shishkin meshes has been proven in one dimension (Frohner, Linss, and Roos [11], Frohner and Roos [12], Hemker, Shishkin, and Shishkina [24]).

It was noticed early by Hemker [21] that the defect correction method overcorrects near layers and should be modified. Various proposals have been advanced, e.g., Hemker [21, 20], Hemker and Koren [22, 23], Ervin and Layton [7]. The one considered herein is to correct the large scales only and leave a small amount of stabilization in the small scales. This is a discretization idea of Layton [31] which is related to ideas of Maday and Tadmor [33], Guermond [17], and Hughes, Mazzei, and Jansen [25]. For current work on this discretization for flow problems, see, e.g., Kaya and Rivière [29], John and Kaya [26], and John, Kaya, and Layton [27].

Because of the attractive form of the defect correction method, it is particularly efficient when used in conjunction with adaptivity. The first theoretical study and computational testing of defect correction plus adaptivity was in Ervin, Layton, and Maubach [9, 10] and Cawood et al. [5]. Interesting recent work in this direction has been done by Nikolova [35] and Axelsson and Nikolova [4]. In particular, [10] considers the problems of stationary turbulence with the Smagorinsky model. It was noted there that the estimators decompose into residuals associated with the base discretization's numerical error, the defect correction method's update error, and the turbulence model's modelling error—an interesting feature of both the defect correction method and adaptive solution of various turbulence models.

2. Mathematical preliminaries. The error analysis we shall perform for the method (1.6), (1.7) will be for nonsingular solutions of the Navier–Stokes equations (1.1), (1.2). We thus collect a few useful facts about nonsingular solutions.

DEFINITION 2.1. *Let V and V_h denote, respectively, the divergence free subspaces of X and X_h:*

$$V := \{v \in X : (q, \nabla \cdot v) = 0 \quad \forall q \in Q\},$$
$$V_h := \{v_h \in X_h : (q_h, \nabla \cdot v_h) = 0 \quad \forall q \in Q_h\}.$$

Although typically $V_h \subsetneq V$, it is known that under the discrete inf-sup condition (1.3), functions in V are well approximated by ones in V_h (Girault and Raviart [14]).

LEMMA 2.2. *Let the discrete inf-sup condition* (1.3) *hold. Then for any* $v \in V$

$$\inf_{v_h \in V_h} \|\nabla(v - v_h)\| \le C \left(1 + \frac{1}{\beta}\right) \inf_{v_h \in X_h} \|\nabla(v - v_h)\|. \tag{2.1}$$

Proof. We refer to [14, p. 60, (1.12)] for the proof of this lemma. ∎

We shall define by M as a finite constant with

$$M = \sup_{u,v,w \in X} \frac{|b^*(u,v,w)|}{\|\nabla u\| \|\nabla v\| \|\nabla w\|}.$$

DEFINITION 2.3. *u is a nonsingular solution of* (1.1) *if there is a $\mu(u,\nu) > 0$ such that*

$$\inf_{v \in V} \sup_{w \in V} \frac{\nu(\nabla v, \nabla w) + b^*(u,v,w) + b^*(v,u,w)}{\|\nabla v\| \|\nabla w\|} \ge \mu(u,\nu) > 0. \tag{2.2}$$

DEFINITION 2.4. *u is an isolated solution of* (1.1) *if there is a $\delta > 0$ such that there exists no other solution u' of* (1.1) *with $\|\nabla(u - u')\| < \delta$.*

The following basic facts are known concerning the equilibrium Navier–Stokes equations (1.1).

PROPOSITION 2.5. (a) *Given $f \in H^{-1}(\Omega)^d$, there exists at least one $(u,p) \in (X,Q)$ satisfying* (1.1).

(b) *For $\|f\|$ small enough, that solution is unique and nonsingular.*

(c) *There is an open dense subset $\mathcal{D} \subset H^{-1}(\Omega)^d$ such that for all $f \in \mathcal{D}$, all solutions of* (1.1) *are nonsingular and the number of solutions for each $f \in \mathcal{D}$ is finite and odd.*

(d) *A nonsingular solution is isolated.*

(e) *Let u be a nonsingular solution of* (1.1) *with data f, and $\tilde{u}$ another solution with data $\tilde{f}$. If $\|\nabla(u - \tilde{u})\| \le \mu(u,\nu)/(2M)$, then*

$$\|\nabla(u - \tilde{u})\| \le \frac{2}{\mu(u,\nu)} \|f - \tilde{f}\|_{-1}.$$

Proof. (a), (b), (c) are well known in the Navier–Stokes equations literature; see, e.g., [14, Theorems 2.4, p. 115, and 2.5, p. 118] for (a), [14, Theorem 1.3, p. 108] for (b), and Temam [39] for (c). Nonsingularity of the (unique) solution under the small data condition is proven in Remark 2.3, p. 119 of [14]. Part (d) is a standard result about nonsingular solutions of nonlinear problems; see, e.g., the remark on page 116 of [14]. Part (e) was proven in [30] and follows from nonsingularity and the mean value theorem. Indeed, subtraction and the usual type arguments imply that

$$\mu(u,\nu)\|\nabla(u - \tilde{u})\| \le \|f - \tilde{f}\|_{-1} + 2M\|\nabla(u - \tilde{u})\|^2,$$

from which (e) follows. ∎

Since the set of invertible operators is open and $b^*(\cdot,\cdot,\cdot)$ is continuous in X, it is known that the point of linearization in various terms of (2.2) can be shifted slightly without changing the essential conclusions.

LEMMA 2.6. *Let u be a nonsingular solution of* (1.1). *Then there is a $\delta > 0$ such that for any $\alpha < \delta$, u' and $u'' \in V$ with $\|\nabla(u - u')\| < \delta$, $\|\nabla(u - u'')\| < \delta$ satisfying*

$$\inf_{v \in V} \sup_{w \in V} \frac{(\nu + \alpha)(\nabla v, \nabla w) + b^*(u',v,w) + b^*(v,u'',w)}{\|\nabla v\| \|\nabla w\|} \ge \frac{1}{2}\mu(u,\nu).$$

Proof. This is a standard result for $\mathcal{C}^1$ operators in nonlinear analysis (Schwartz [36]). It is used in error analysis in [14, Lemma 3.3, p. 130]. ∎

It will be important to note that if (1.3) holds, the infimum and supremum in Lemma 2.6 can also be taken over V_h.

LEMMA 2.7. *Let u be a nonsingular solution of* (1.1). *Then there is a $\delta > 0$ such that for any $\alpha < \delta$, u' and $u'' \in V$ or V_h with $\|\nabla(u - u')\| < \delta$, $\|\nabla(u - u'')\| < \delta$ satisfying*

$$\inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{(\nu + \alpha)(\nabla v_h, \nabla w_h) + b^*(u', v_h, w_h) + b^*(v_h, u'', w_h)}{\|\nabla v_h\| \|\nabla w_h\|} \geq \frac{1}{2}\mu(u, \nu).$$

Proof. For the proof, see Girault and Raviart [14, Lemma 3.3, p. 130] and especially [14, Remark 3.1, p. 130] that follows. The analysis in Remark 3.1 is a continuity argument and applies with $u'' \neq u'$ (both close enough to u). ∎

3. Error analysis. The basic principle of the defect correction method in this context is that each step attempts to increase the rate of convergence by one power of h up to the order of the basic method. To begin, note that (u_h^1, p_h^1) is just the usual artificial viscosity approximation to (u, p). Since the error analysis for this step is standard (and a special case of the general step in which $(u_h^0, p_h^0) = (0, 0)$), we present the result only. The error analysis uses basic tools from [14, 15] and requires a few basic assumptions on (X_h, Q_h) that assume that (1.3) holds and that (X_h, Q_h) become dense in (X, Q) as $h \to 0$. Specifically, we assume the following proposition.

PROPOSITION 3.1. *Let* (1.3) *hold. For any $\alpha \geq 0$ and $f \in H^{-1}(\Omega)^d$, the algorithm* (1.6), (1.7) *is well defined: there exist approximate solutions (u_h^j, p_h^j) for $j = 1, 2, \ldots$.*

Proof. Existence of (u_h^1, p_h^1) follows from the fact that $(X_h,\ Q_h)$ is finite dimensional, the fixed point theory, and the following a priori bounds:

$$(\alpha + \nu)\|\nabla u_h^1\| \leq \|f\|_{-1}, \tag{3.1}$$

$$\|p_h^1\| \leq \beta^{-1}(2 + M(\alpha + \nu)^{-2})\|f\|_{-1}. \tag{3.2}$$

The first result (3.1) is obtained by setting $v_h = u_h^1$ in (1.1). The second (3.2) follows from (1.3), exactly as for the usual Galerkin approximation.

Given $(u_h^j, p_h^j) \in (X_h, Q_h)$, the same argument can be used only in the formulation (1.8) to prove existence of (u_h^{j+1}, p_h^{j+1}), provided only that $\|g_H^j\|$ is bounded. To see this, note that in the second equation of (1.8), g_H^j is the L^2 projection into L_H of ∇u_h^j. Thus, $\|g_H^j\| \leq \|\nabla u_h^j\| < \infty$, which is the required bound. ∎

A similar result is true for the defect correction using the linearization (1.6), (1.9), provided that h is small enough.

PROPOSITION 3.2. *Let* (1.3) *hold. Consider the algorithm* (1.6), (1.9) *(or equivalently,* (1.6), (1.10)*). Assume that u is a nonsingular solution of* (1.1). *Fix a function $f \in H^{-1}(\Omega)$. Let $\alpha \geq 0$ tend to zero as h tends to zero. Then, there is $h_0 > 0$ such that for $h \leq h_0$, (u_h^2, p_h^2) exists and is unique. More generally, if u_h^j is close enough to u in X, then (u_h^{j+1}, p_h^{j+1}) exists and is unique.*

Proof. This is a linearization argument. First we note that since (u_h^1, p_h^1) is the artificial viscosity approximation to a nonsingular solution, standard error analysis of branches of nonsingular solutions, e.g., [14, 15], shows that $\lim u_h^1 = u$ as h tends to 0 (see Proposition 3.3 for more detail). Now consider the correction step (1.10). It can be rewritten as: find $u_h^2 \in V_h$ satisfying

$$(\nu + \alpha)(\nabla u_h^2, \nabla v_h) + b^*(u_h^1, u_h^2, v_h) + b^*(u_h^2, u_h^1, v_h) = (G, v_h) \quad \forall v_h \in V_h,$$

where $G = G(u_h^1, p_h^1, g_H^1, \dots)$ is known in terms of problem data and the solution of the first step (u_h^1, p_h^1). Since $u_h^1 \to u$ and u is a nonsingular solution, Lemma 2.7 implies that this linear problem has a unique solution. Thus, u_h^2 exists.

The remainder of the proof is an induction argument which follows similarly: once u_h^2 exists uniquely and the linearization (1.10) at u_h^1 is invertible, it will follow that u_h^2 converges to u as h tends to 0 (with appropriate error estimates). This implies that u_h^3 exists uniquely and the argument is repeated. ∎

Remark 3.1. This argument fails if (1.10) is an *iteration* since the constants involved depend on j, but it is correct since it is a *correction* only performed a fixed number of times. Concerning the error in u_h^1, we have the following proposition.

PROPOSITION 3.3. *Assume the spaces* (X_h, Q_h) *satisfy* (1.3). *Suppose* u *is a nonsingular solution of the* (1.1). *Suppose* α *tends to* 0 *as* h *tends to* 0. *Then, there is* $h_0 > 0$ *such that for* $h \le h_0$ *the error in* (u_h^1, p_h^1) *satisfies*

$$\|\nabla(u - u_h^1)\| \le C(\beta)\left[\frac{2}{\mu(u,\nu)}\left(\alpha + \nu + \frac{2M}{\nu}\|f\|_{-1}\right) + 1\right] \inf_{v_h \in X_h} \|\nabla(u - v_h)\| + \frac{2}{\mu(u,\nu)}\left[\inf_{\lambda_h \in Q_h} \|p - \lambda_h\| + \alpha\|\nabla u\|\right],$$

$$\|p - p_h^1\| \le \left(1 + \frac{1}{\beta}\right) \inf_{\lambda_h \in Q_h} \|p - \lambda_h\| + \frac{1}{\beta}\left(\nu + \alpha + \frac{2M}{\nu}\|f\|_{-1}\right)\|\nabla(u - u_h^1)\| + \frac{\alpha}{\beta}\|\nabla u\|.$$

Proof. The proof that $u_h^1 \to u$ is standard, since u_h^1 is just the usual artificial viscosity approximation, following, e.g., [14, 15]. We shall thus give the proof of only the error bound since it gives the ideas of the proof of the general case in a simpler context. The true solution (u, p) satisfies for any $v_h \in V_h$, $\lambda_h \in Q_h$

$$(\nu + \alpha)(\nabla u, \nabla v_h) + b^*(u, u, v_h) - (p - \lambda_h, \nabla \cdot v_h) = (f, v_h) + \alpha(\nabla u, \nabla v_h). \tag{3.3}$$

Let $\tilde{u} \in V_h$ be an approximation to u and write $e^1 = u - u_h^1 = \eta - \phi_h$, where $\phi_h = u_h^1 - \tilde{u}$ and $\eta = u - \tilde{u}$. Subtracting from (3.3), equation (1.6) for (u_h^1, p_h^1) gives

$$(\nu + \alpha)(\nabla e^1, \nabla v_h) + b^*(u, u, v_h) - b^*(u_h^1, u_h^1, v_h) = (p - \lambda_h, \nabla \cdot v_h) + \alpha(\nabla u, \nabla v_h). \tag{3.4}$$

The nonlinear term can be rewritten as

$$\begin{aligned} b^*(u, u, v_h) - b^*(u_h^1, u_h^1, v_h) &= b^*(e^1, u, v_h) + b^*(u_h^1, e^1, v_h) \\ &= b^*(\eta, u, v_h) - b^*(\phi_h, u, v_h) + b^*(u_h^1, \eta, v_h) - b^*(u_h^1, \phi_h, v_h). \end{aligned}$$

Using this decomposition of $b^*(\cdot,\cdot,\cdot)$ and splitting $e^1 = \eta - \phi_h$ gives

$$\begin{aligned} &(\nu + \alpha)(\nabla \phi_h, \nabla v_h) + b^*(\phi_h, u, v_h) + b^*(u_h^1, \phi_h, v_h) \\ &\quad = (\nu + \alpha)(\nabla \eta, \nabla v_h) + b^*(\eta, u, v_h) \\ &\qquad + b^*(u_h^1, \eta, v_h) - (p - \lambda_h, \nabla \cdot v_h) - \alpha(\nabla u, \nabla v_h) \quad \forall (v_h, \lambda_h) \in (V_h, Q_h). \end{aligned} \tag{3.5}$$

Applying standard bounds to the right-hand side of (3.5) gives

$$\begin{aligned} &\frac{1}{\|\nabla v_h\|}[(\nu + \alpha)(\nabla \phi_h, \nabla v_h) + b^*(\phi_h, u, v_h) + b^*(u_h^1, \phi_h, v_h)] \\ &\le (\nu + \alpha)\|\nabla \eta\| + M(\|\nabla u\| + \|\nabla u_h^1\|)\|\nabla \eta\| + \|p - \lambda_h\| + \alpha\|\nabla u\|. \end{aligned}$$

Taking the supremum over $v_h \in V_h$, using Lemma 2.7 and a priori bounds on $\|\nabla u\|$ and $\|\nabla u_h^1\|$ yield

$$\frac{1}{2}\mu(u,\nu)\|\nabla\phi_h\| \le \left[\alpha+\nu+M\|f\|_{-1}\left(\frac{1}{\nu}+\frac{1}{\nu+\alpha}\right)\right]\|\nabla\eta\| + \|p-\lambda_h\| + \alpha\|\nabla u\|.$$

By using the triangle inequality, taking the infimum over $v^h \in V^h$, $\lambda^h \in Q^h$, and using Lemma 2.2, one obtains the required result.

For the pressure estimate (just as for the Stokes problem) we begin with the error equation for $v_h \in X_h$ (rather than V_h):

$$(p-p_h^1, \nabla\cdot v_h) = (\nu+\alpha)(\nabla e^1, \nabla v_h) + b^*(e^1, u, v_h) - b^*(u_h^1, e^1, v_h) - \alpha(\nabla u, \nabla v_h).$$

Write $p - p_h^1 = p - \lambda_h - (p_h^1 - \lambda_h)$, where $\lambda_h \in Q_h$ approximates p well. Then

$$\begin{aligned}(p_h^1-\lambda_h, \nabla\cdot v_h) &= (p-\lambda_h, \nabla\cdot v_h) - (\nu+\alpha)(\nabla e^1, \nabla v_h)\\ &\quad - b^*(e^1, u, v_h) + b^*(u_h^1, e^1, v_h) + \alpha(\nabla u, \nabla v_h)\\ &\le [\|p-\lambda_h\| + (\nu+\alpha)\|\nabla e^1\| + M(\|\nabla u\| + \|\nabla u_h^1\|)\|\nabla e^1\|\\ &\quad + \alpha\|\nabla u\|]\|\nabla v_h\|.\end{aligned}$$

Dividing by $\|\nabla v_h\|$, taking the supremum over $v_h \in X_h$, using the inf-sup condition (1.3), and the triangle inequality, we have

$$\begin{aligned}\|p-p_h^1\| &\le \left(1+\frac{1}{\beta}\right)\|p-\lambda_h\| + \frac{1}{\beta}(\nu+\alpha)\|\nabla e^1\|\\ &\quad + \frac{M}{\beta}(\|\nabla u\| + \|\nabla u_h^1\|)\|\nabla e^1\| + \frac{\alpha}{\beta}\|\nabla u\|.\end{aligned}$$

Finally, using a priori bounds (3.1), $\nu\|\nabla u\| \le \|f\|_{-1}$ gives the required result. ∎

Concerning the error in the method we consider the variant (1.6), (1.9) in which one linearized problem is solved per correction step. Intuitively, one would expect that the defect correction method (1.6), (1.7) would be more robust and more accurate. On the other hand, complete error analysis of the defect correction method with nonlinear correction (1.6), (1.7) is an open problem in the case of large data and nonsingular solutions.

PROPOSITION 3.4. *Consider* (1.6), (1.9). *Let u be a nonsingular solution of the* (1.1) *and suppose* (1.3) *holds. Then, there is a $\delta > 0$ such that if $\|\nabla(u-u_h^j)\| < \delta$, for $j = 1, 2\ldots,$*

$$\begin{aligned}\frac{1}{2}\mu(u,\nu)\|\nabla(u-u_h^{j+1})\| &\le C(\beta)\left(\nu+\alpha+\frac{1}{2}\mu(u,\nu)+2M(\delta+\|\nabla u\|)\right)\inf_{v_h\in X_h}\|\nabla(u-v_h)\|\\ &\quad + \sqrt{2}\inf_{\lambda_h\in Q_h}\|p-\lambda_h\| + \alpha\|\nabla u - \overline{\nabla u}\|\\ &\quad + M\|\nabla(u-u_h^j)\|^2 + \alpha\|\overline{\nabla(u-u_h^j)}\|,\\ \beta\|p-p_h^{j+1}\| &\le C\inf_{\lambda_h\in Q_h}\|p-\lambda_h\| + [\nu+\alpha+2M\|\nabla u\|]\|\nabla(u-u_h^{j+1})\|\\ &\quad + M\|\nabla(u-u_h^j)\|^2 + \alpha\|\overline{\nabla(u-u_h^j)}\| + \alpha\|\nabla u - \overline{\nabla u}\|.\end{aligned}$$

Proof. The variational formulation of (1.1) can be rewritten as follows: for any $v_h \in V_h$ and $\lambda_h \in Q_h$,

$$\begin{aligned}&(\nu+\alpha)(\nabla u,\nabla v_h)+b^*(u_h^j,u,v_h)+b^*(u,u_h^j,v_h)-(p-\lambda_h,\nabla\cdot v_h)\\&=(f,v_h)+[b^*(u_h^j,u,v_h)+b^*(u,u_h^j,v_h)-b^*(u,u,v_h)]\\&\quad+\alpha(\overline{\nabla u},\nabla v_h)+\alpha(\nabla u-\overline{\nabla u},\nabla v_h).\end{aligned}\tag{3.6}$$

The square bracketed term on the right-hand side of (3.6) becomes

$$b^*(u_h^j,u,v_h)+b^*(u,u_h^j,v_h)-b^*(u,u,v_h)=-b^*(u-u_h^j,u-u_h^j,v_h)+b^*(u_h^j,u_h^j,v_h).\tag{3.7}$$

Let $e^{j+1}=u-u_h^{j+1}$, $e^j=u-u_h^j$ and note that $g_H=\overline{\nabla u_h^j}$ (by (1.10)). With this notation subtract (1.10) from (3.6) and use (3.7) for the nonlinear terms on the right-hand side. This gives

$$\begin{aligned}&(\nu+\alpha)(\nabla e^{j+1},\nabla v_h)+b^*(u_h^j,e^{j+1},v_h)+b^*(e^{j+1},u_h^j,v_h)\\&=(p-\lambda_h,\nabla\cdot v_h)-b^*(e^j,e^j,v_h)+\alpha(\overline{\nabla e^j},\nabla v_h)\\&+\alpha(\nabla u-\overline{\nabla u},\nabla v_h)\quad\forall(v_h,\lambda_h)\in(V_h,Q_h).\end{aligned}$$

The remainder of the proof follows that of Proposition 3.3: we first split $e^{j+1}=\eta-\phi_h$, $\eta=u-\tilde{u}$, and $\phi_h=u_h^{j+1}-\tilde{u}$, where $\tilde{u}\in V_h$ approximates u well. By using this decomposition, nonlinear terms can be written as

$$\begin{aligned}&b^*(u_h^j,e^{j+1},v_h)+b^*(e^{j+1},u_h^j,v_h)\\&=b(u_h^j,\eta,v_h)-b(u_h^j,\phi_h,v_h)+b(\eta,u_h^j,v_h)-b(\phi_h,u_h^j,v^h).\end{aligned}$$

Then, the use of splitting error and Lemma 2.7 give that for δ small enough (i.e., for u_h^j close enough to u) the linear problem for e^{j+1} satisfies the inf-sup stability condition. Thus, the following inequality for ϕ_h holds:

$$\begin{aligned}\frac{1}{2}\mu(u,\nu)\|\nabla\phi_h\|&\le(\nu+\alpha+2M\|\nabla u_h^j\|)\|\nabla\eta\|\\&\quad+\sqrt{2}\|p-\lambda_h\|+M\|\nabla e^j\|^2+\alpha\|\overline{\nabla e^j}\|+\alpha\|\nabla u-\overline{\nabla u}\|.\end{aligned}$$

Since u_h^j is close enough to $u\in X$, we have $\|\nabla u_h^j\|\le\delta+2\|\nabla u\|$. The triangle inequality then implies that

$$\begin{aligned}\frac{1}{2}\mu(u,\nu)\|\nabla e^{j+1}\|&\le\left(\nu+\alpha+\frac{1}{2}\mu(u,\nu)+2M(\delta+\|\nabla u\|)\right)\|\nabla\eta\|\\&\quad+\sqrt{2}\|p-\lambda_h\|+M\|\nabla e^j\|^2+\alpha\|\overline{\nabla e^j}\|+\alpha\|\nabla u-\overline{\nabla u}\|.\end{aligned}$$

The pressure bound also follows from the case $j=1$. ∎

The error estimate Proposition 3.4 has four terms. The first term $\|\nabla(u-v_h)\|$ is the error in the best approximation to $(u,p)\in(X_h,Q_h)$. The second term $\|\nabla e^j\|^2$ is the linearization error. Since this is quadratic, it is typically a higher order term. The third, $\alpha\|\overline{\nabla e^j}\|$, shows that each step of the defect correction method improves the error in the previous step by one power of h (recall that typically $\alpha=\mathcal{O}(h)$). The last term $\alpha\|\nabla u-\overline{\nabla u}\|$ arises from the error of the stabilized discretization.

As a result of Proposition 3.4, we can give the following corollaries.

COROLLARY 3.5. *In addition to the assumptions of Proposition* 3.4*, suppose* $h \le H \to 0$ *as* $h \to 0$ *and* $\alpha \to 0$ *as* $h \to 0$*. Suppose also* X_h, Q_h*,* L_H *become dense in* X, Q*, and* L*, respectively, as* $h \to 0$*. Then, there is an* $h_0 > 0$ *such that for* $h \le h_0$ *and* $j = 1, 2, \dots k$*,* $u_h^j \to u$ *as* $h \to 0$*.*

COROLLARY 3.6. *Suppose* X_h*,* Q_h*,* L_H *consist of piecewise polynomials of degree* $k, k-1$*, and* $k-1$*, respectively. Suppose also that* $u \in H^{k+1}(\Omega) \cap X$*,* $p \in H^k(\Omega) \cap Q$*. Then,*

$$\|\nabla(u - u_h^1)\| \le C(u,p)[h^k + \alpha],$$

and in general,

$$\|\nabla(u - u_h^j)\| \le C(u,p,j)[h^k + \alpha H^k + \alpha^j].$$

This follows by inserting the approximation theoretical orders of convergence into the right-hand side of Proposition 3.4 and keeping only dominant terms. For example,

$$\begin{aligned}\|\nabla(u - u_h^2)\| &\le C(u,p)[h^k + (h^k + \alpha)^2 + \alpha(h^k + \alpha) + \alpha H^k] \\ &\le C(u,p)[h^k + \alpha H^k + \alpha^2].\end{aligned}$$

The error estimate in the corollary explains the typical algorithmic choices:

$$\begin{aligned} j &\ge k : \text{correction step}, \\ \alpha &= \alpha_0 h : \text{regularization parameter}, \\ H &\le C h^{1-\frac{1}{k}} : \text{length scale of large structures}. \end{aligned}$$

4. Numerical studies. In this section, we consider some numerical experiments for the implementation of defect correction algorithms proposed with (1.6), (1.10). In particular, we present two numerical examples: one is a known analytical solution; and the other is the driven cavity problem.

All computations are carried out in the domain $\Omega = [0,1] \times [0,1]$. We divide our domain into triangles. We use Taylor–Hood elements, i.e., continuous piecewise quadratic functions for the velocity space X_h and continuous piecewise linear functions for the pressure space Q_h. It is well known that this conforming pair of finite element spaces satisfies the inf-sup condition (1.3). The coarse space L_H is chosen to be ∇X_H, where X_H is the space of continuous piecewise quadratics on the coarse mesh. For every grid, the first artificial viscosity system (1.6) is solved with $\alpha = h$. Then, k (the polynomial degree of the velocity approximation) defect correction steps are performed. Hence, two correction steps are required for the Taylor–Hood element. All the nonlinear systems are solved by using the Newton method with stopping criterion 10^{-6}. Corollary 3.6 suggests that the algorithmic choices for h and H should be $h \sim H^2$ or, equivalently, $H \sim h^{1/2}$ in order to obtain optimal error rates.

As a first numerical illustration, we study a numerical convergence to confirm the error estimate given in Proposition 3.4. The prescribed solution is given by

$$u = -4y^3x^2, \quad v = 2xy^4, \quad p = 2x + 3y - 2.$$

Dirichlet boundary conditions are chosen, and the right-hand side f is such that (u, v, p) is the solution of (1.1). In this example, our numerical results are performed for $\nu = 1$. For the Taylor–Hood finite element spaces, the theory predicts a convergence rate of $\mathcal{O}(h^2)$ in the energy norm, $\mathcal{O}(h^3)$ in the L^2 norm for the velocity, and $\mathcal{O}(h^2)$ for the pressure.

TABLE 1
Convergence rates by using the Galerkin finite element method.

h	L^2 *error*	Rate	H_0^1 *error*	Rate	L^2 pressure	Rate
h=1/2	0.0093		0.2894		0.2934	
h=1/4	0.0011	3.07	0.0710	2.02	0.0786	1.90
h=1/8	1.3181e-004	3.06	0.0177	2.00	0.0200	1.97
h=1/16	1.6321e-005	3.01	0.0044	2.00	0.0051	1.97

TABLE 2
Convergence rates by using the artificial viscosity method.

h	L^2 *error*	Rate	H_0^1 *error*	Rate	L^2 pressure	Rate
h=1/4	0.0061		0.0844		0.4671	
h=1/8	0.0032	0.93	0.0306	1.46	0.2146	1.12
h=1/16	0.0016	1.00	0.0137	1.15	0.1023	1.06

TABLE 3
Convergence rates of the subgrid stabilized defect correction method.

H	h	L^2 *error*	Rate	H_0^1 *error*	Rate	L^2 pressure	Rate
1/2	1/4	0.0012		0.0701		0.0824	
1/4	1/8	1.4146e-004	3.08	0.0174	2.01	0.0224	1.87
1/8	1/16	1.7565e-005	3.01	0.0043	2.01	0.0059	1.92
1/16	1/32	2.1926e-006	3.00	0.0011	1.96	0.0015	1.97

Note that, since we try to verify the theory in this simplest setting, the first numerical test problem does not require either a subgrid eddy viscosity method or defect correction method for an accurate solution. However, the method (1.6), (1.10) is fully comparable to the standard finite approach in this laminar case. In Table 1, the error in the usual Galerkin discretization of the Navier–Stokes equations is presented. In particular, we give L^2 and H_0^1 errors and the corresponding convergence rates. As theory predicts, the optimal convergence rates are obtained. In Table 2, we present convergence rates by using the artificial viscosity (AV) method where we perform only initialization step (1.6) to solve Navier–Stokes equations. Since we choose $\alpha = h$, it is expected and observed that the convergence rates for this method are suboptimal. In Table 3, the experimental rates of convergence for the subgrid stabilized defect correction method are presented. The scalings between coarse and fine mesh are chosen such that $H \leq h^{1/2}$ is satisfied. These numerical results demonstrate that the rates of convergence are optimal, as the theory predicts. Hence, the stabilization used in the method (1.6), (1.10) does not degrade rates of convergence in laminar flows.

Our second example is the benchmark problem of the underresolved driven cavity at high Reynolds numbers. This problem is chosen because there is benchmark data available for comparison. Even though the zero boundary condition assumed in the theory is not valid here, the convergence theory can be extended to nonzero boundary conditions smooth enough to be the trace of an H^1 function. In this benchmark problem, flow is driven by the tangential velocity field applied to the top boundary in the absence of other body forces. On the segment $\{(x, 1) : 0 < x < 1\}$, the velocity is equal to $u = (1, 0)$. On the rest of the boundary, zero Dirichlet conditions are imposed.

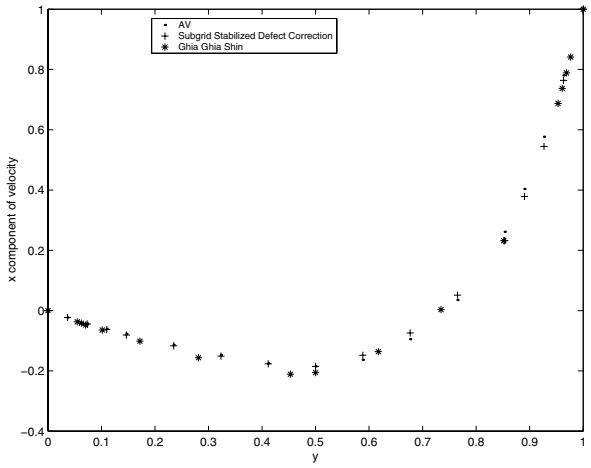


FIG. 1. *Vertical midlines for* $\nu = 10^{-2}$ *for* $H = 1/4, h = 1/8$.

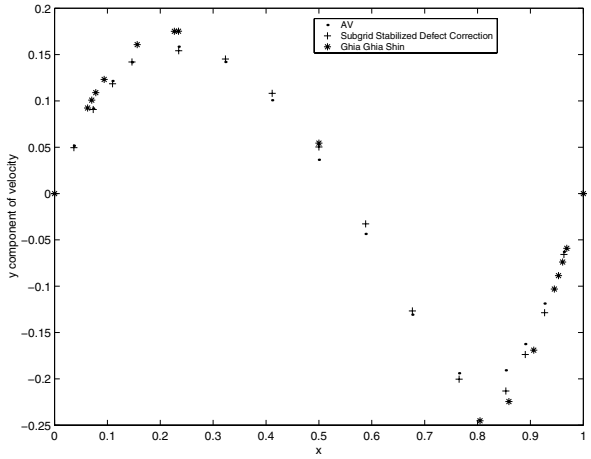


FIG. 2. *Horizontal midlines for* $\nu = 10^{-2}$ *for* $H = 1/4, h = 1/8$.

The drawbacks of usual, centered Galerkin methods for convection dominated problems are well known and well documented. Also, the drawbacks of the unmodified defect correction method, although less well known, are very well documented since the 1982 work of Hemker [21, 20]. Thus, the focus of our experiments on the driven cavity problem is to (i) show that the subgrid stabilized defect correction method gives high quality, coarse mesh solutions (comparable to the benchmark, fine mesh results of Ghia, Ghia, and Shin [13]), (ii) illustrate that stabilization of the finest resolved scales is effective in suppressing the oscillations on the scales typical of the unstabilized defect correction method, and (iii) illustrate the very substantial improvement in the results produced by the relatively inexpensive correction steps.

We compute an approximate solution for $\nu = 10^{-2}$, $\nu = 25 \times 10^{-4}$, and $\nu = 3125 \times 10^{-7}$ for the driven cavity flow with regularization parameter $\alpha_0 = 0.1$. In particular, we draw the x component of velocity along the vertical centerline and

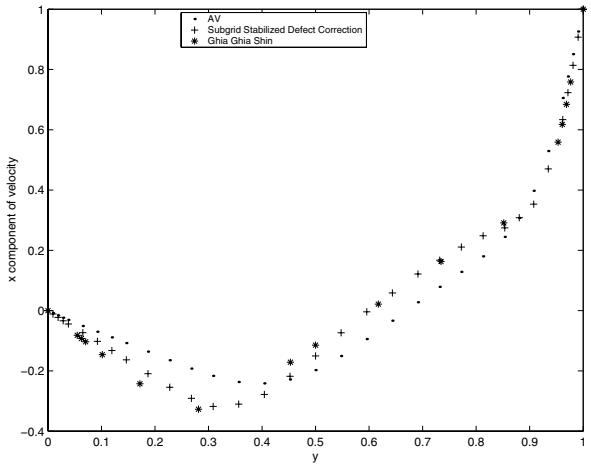


FIG. 3. *Vertical midlines for* $\nu = 25 \times 10^{-4}$ *for* $H = 1/8, h = 1/16$.

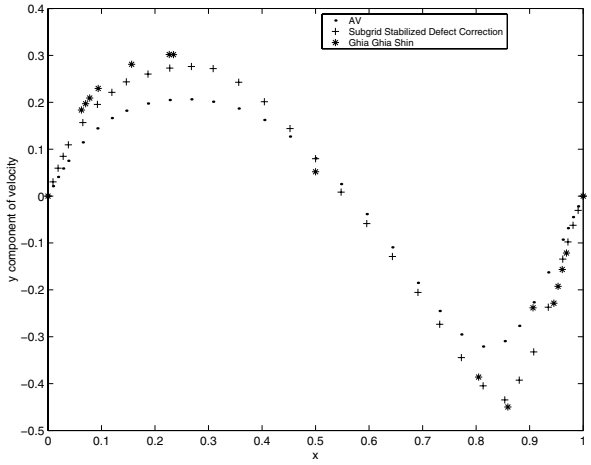


FIG. 4. *Horizontal midlines for* $\nu = 25 \times 10^{-4}$ *for* $H = 1/8, h = 1/16$.

y component of velocity along the horizontal centerlines. We compare our results to those obtained by Ghia, Ghia, and Shin [13]. The present numerical simulations are considered on a very coarse mesh ($h = 1/8, h = 1/16, h = 1/32$) and they are compared to the very fine mesh ($h = 1/129$) of [13]. Ghia's algorithm is based on the time dependent streamfunction using the coupled implicit and multigrid methods. Their results are used as benchmark data as basis for comparison.

In Figures 1–6, we compare the results obtained by the AV method, the subgrid stabilized defect correction method (1.6), (1.10), and the results of [13]. In the case $\nu = 10^{-2}$, there is very little difference between the vertical midlines for all three methods (Figure 1). For the horizontal midlines, the subgrid stabilized defect correction method is closer to Ghia, Ghia, and Shin's results than the artificial viscosity (see Figure 2).

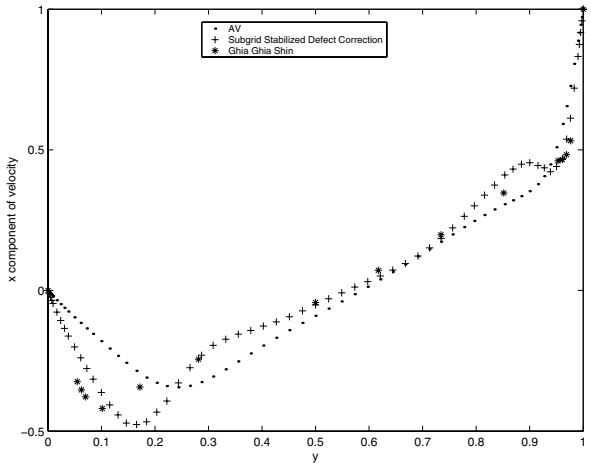


FIG. 5. *Vertical midlines for* $\nu = 3125 \times 10^{-7}$ *for* $H = 1/16, h = 1/32$.

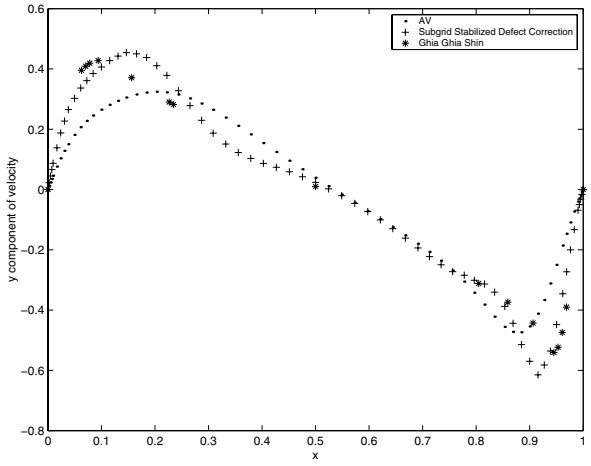


FIG. 6. *Horizontal midlines for* $\nu = 3125 \times 10^{-7}$ *for* $H = 1/16, h = 1/32$.

In the cases of $\nu = 25 \times 10^{-4}$ and $\nu = 3125 \times 10^{-7}$, namely for higher Reynolds number, Figures 3, 4, 5, and 6 clearly show that the subgrid stabilized defect correction method performs much better than the artificial viscosity method, and is comparable to the results obtained by Ghia, Ghia, and Shin on a more refined mesh.

5. Conclusion. The natural combination of defect correction with multiscale stabilization retains the best features of both methods and overcomes many of their deficits. The combination is accurate and efficient, and a convergence theory of the combination is developed. This latter theory shows that the good accuracy and stability properties are no accident—they are general features of the method.

This combination has strong promise, but many open questions remain including the correct extension of the method to time dependent problems, further numerical analysis (other norms, error functionals, ...), and more extensive testing.

REFERENCES

[1] I. Altas and K. Burrage, *A high-accuracy defect-correction multigrid method for the steady incompressible Navier–Stokes equations*, J. Comput. Phys., 114 (1994), pp. 227–233.

[2] M. Anitescu, W. J. Layton, and F. Pahlevani, *Implicit for local effects and explicit for nonlocal effects is unconditionally stable*, Electron. Trans. Numer. Anal., 18 (2004), pp. 174–187.

[3] O. Axelsson and W. Layton, *Defect correction methods for convection dominated convection-diffusion equation*, RAIRO J. Numer. Anal. (Now M2AN), 24 (1990), pp. 423–455.

[4] O. Axelsson and M. Nikolova, *Adaptive refinement for convection-diffusion problems based on a defect-correction technique and finite difference method*, Computing, 58 (1997), pp. 1–30.

[5] M. E. Cawood, V. J. Ervin, W. J. Layton, and J. M. Maubach, *Adaptive defect correction methods for convection dominated, convection diffusion problems*, J. Comput. Appl. Math., 116 (2000), pp. 1–21.

[6] J.-A. Desideri and P. W. Hemker, *Convergence analysis of the defect-correction iteration for hyperbolic problems*, SIAM J. Sci. Comput., 16 (1995), pp. 88–118.

[7] V. J. Ervin and W. J. Layton, *High resolution, minimal storage algorithms for convection dominated, convection-diffusion equations*, in Transactions of the Fourth Army Conference on Applied Mathematics and Computing, New York, U.S. Army Research Office, Research Triangle Park, NC, 1987, pp. 1173–1201.

[8] V. J. Ervin and W. J. Layton, *A study of defect correction, finite difference methods for convection diffusion equations*, SIAM J. Numer. Anal., 26 (1989), pp. 169–179.

[9] V. J. Ervin, W. J. Layton, and J. M. Maubach, *An adaptive defect correction approach for convection, dominated convection diffusion problems*, in Computational Techniques and Applications: CTAC95, R. L. May and A. K. Easton, eds., World Scientific, River Edge, NJ, 1996, pp. 287–294.

[10] V. J. Ervin, W. J. Layton, and J. M. Maubach, *Adaptive defect-correction methods for viscous incompressible flow problems*, SIAM J. Numer. Anal., 37 (2000), pp. 1165–1185.

[11] A. Frohner, T. Linss, and H.-G. Roos, *Defect correction on Shishkin-type meshes*, Numer. Algorithms, 26 (2001), pp. 281–299.

[12] A. Frohner and H. G. Roos, *The epsilon-uniform convergence of a defect correction method on a Shishkin mesh*, Appl. Numer. Math., 37 (2001), pp. 79–94.

[13] U. Ghia, K. N. Ghia, and C. T. Shin, *High-resolutions for incompressible flow using the Navier–Stokes equations and a multigrid method*, J. Comput. Phys., 48 (1982), pp. 387–411.

[14] V. Girault and P.-A. Raviart, *Finite Methods Element Approximation of the Navier–Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.

[15] V. Girault and P.-A. Raviart, *Finite Element Methods for the Navier–Stokes Equations: Theory and Algorithms*, Springer Ser. Comput. Math., 5, Springer, Berlin, 1986.

[16] M. Graziadei, R. M. M. Mattheij, and J. H. M. T. Boonkkamp, *Local defect correction with slanting grids*, Numer. Methods Partial Differential Equations, 20 (2004), pp. 1–17.

[17] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modelling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.

[18] W. Heinrichs, *Defect correction for the advection-diffusion equation*, Comput. Method Appl. Mech. Engrg., 119 (1994), pp. 191–197.

[19] W. Heinrichs, *Defect correction for convection-dominated flow*, SIAM J. Sci. Comput., 17 (1996), pp. 1082–1091.

[20] P. W. Hemker, *An accurate method without directional bias for the numerical solution of a 2-D elliptic singular perturbation problem*, in Theory and Applications of Singular Pertubations, Lecture Notes in Math. 942, W. Eckhaus and E. M. Jaeger, eds., Springer, Berlin, 1982.

[21] P. W. Hemker, *Mixed defect correction iteration for the accurate solution of the convection diffusion equation*, in Multigrid Methods, Lecture Notes in Math. 960, W. Hackbusch and V. Trottenberg, eds., Springer, Berlin, 1982, pp. 485–501.

[22] P. W. Hemker and B. Koren, *Multigrid, defect correction and upwind schemes for the steady Navier–Stokes equations*, in Numerical Methods for Fluid Dynamics, III (Oxford, 1988), K. W. Morton and M. J. Baines, eds., Oxford University Press, New York, 1988, pp. 153–170.

[23] P. W. Hemker and B. Koren, *Defect correction and nonlinear multigrid for the steady Euler equations*, in Advances in Computational Fluid Dynamics, W. G. Habashi and M. M. Hafez, eds., Cambridge University Press, Cambridge, UK, 1992, pp. 273–291.

[24] P. W. Hemker, G. I. Shishkin, and L. P. Shishkina, *The use of defect correction for the solution of parabolic singular perturbation problems*, Z. Angew Math. Mech., 77 (1997), pp. 59–74.
[25] T. J. R. Hughes, L. Mazzei, and K. E. Jansen, *Large eddy simulation and the variational multiscale method*, Comput. Visual Sci., 3 (2000), pp. 47–59.
[26] V. John and S. Kaya, *A finite element variational multiscale method for the Navier–Stokes equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1485–1503.
[27] V. John, S. Kaya, and W. J. Layton, *A two-level variational multiscale method for convection-diffusion equations*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 4594–4603.
[28] G. Juncu, *A numerical study of steady viscous flow past a fluid sphere*, Int. J. Heat Fluid Flow, 20 (1999), pp. 414–421.
[29] S. Kaya and B. Rivière, *A discontinuous subgrid eddy viscosity method for the time-dependent Navier–Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1572–1595.
[30] W. J. Layton, *Solution algorithm for incompressible viscous flows at high Reynolds number*, Vestnik Mosk. Gos. Univ. Ser., 15 (1996), pp. 25–35.
[31] W. J. Layton, *A connection between subgrid scale eddy viscosity and mixed methods*, Appl. Math. Comput., 133 (2002), pp. 147–157.
[32] H. K. Lee, *Analysis of a defect correction method for viscoelastic fluid flow*, Comput. Math. Appl., 48 (2004), pp. 1213–1229.
[33] Y. Maday and E. Tadmor, *Analysis of the spectral vanishing viscosity method for periodic conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 854–870.
[34] V. Nefedov and R. M. M. Mattheij, *Local defect correction with different grid types*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 454–468.
[35] M. Nikolova, *Adaptive Refinement Methods for Singulary Perturbed Convection-Diffusion Problems*, Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands, 1999.
[36] J. T. Schwartz, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.
[37] G. J. Shaw and P. I. Crumpton, *An assessment of the finite termination property of the defect correction method*, Internat. J. Numer. Methods Fluids, 16 (1993), pp. 199–215.
[38] H. J. Stetter, *The defect correction principle and discretization methods*, Numer. Math., 29 (1978), pp. 425–443.
[39] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 41, SIAM, Philadelphia, 1983.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1655–1680

CONVERGENCE OF A FINITE DIFFERENCE SCHEME FOR THE CAMASSA–HOLM EQUATION*

HELGE HOLDEN† AND XAVIER RAYNAUD‡

Abstract. We prove that a certain finite difference scheme converges to the weak solution of the Cauchy problem on a finite interval with periodic boundary conditions for the Camassa–Holm equation $u_t - u_{xxt} + 3uu_x - 2u_xu_{xx} - uu_{xxx} = 0$ with initial data $u|_{t=0} = u_0 \in H^1([0,1])$. Here it is assumed that $u_0 - u_0'' \geq 0$, and in this case the solution is unique, globally defined, and energy preserving.

Key words. Camassa–Holm equation, convergence of finite difference schemes

AMS subject classifications. Primary, 65M06, 65M12; Secondary, 35B10, 35Q53

DOI. 10.1137/040611975

1. Introduction.

In the past decade, the Camassa–Holm equation [3]

$$u_t - u_{xxt} + 2\kappa u_x + 3uu_x - 2u_xu_{xx} - uu_{xxx} = 0 \tag{1.1}$$

has received considerable attention. With κ positive it models (see [4, 16, 12]) propagation of unidirectional gravitational waves in a shallow water approximation, with u representing the fluid velocity. The Camassa–Holm equation possesses many intriguing properties: It is, for instance, completely integrable and experiences wave breaking in finite time for a large class of initial data. Most attention has been given to the case with $\kappa = 0$ on the full line, that is,

$$u_t - u_{xxt} + 3uu_x - 2u_xu_{xx} - uu_{xxx} = 0, \tag{1.2}$$

which has so-called peakon solutions, i.e., solutions of the form $u(x,t) = ce^{-|x-ct|}$ for real constants c. Local and global well-posedness results as well as results concerning breakdown are proved in [9, 14, 17, 20].

In this paper we study the Camassa–Holm equation (1.1) on a finite interval with periodic boundary conditions. It is known that certain initial data give global solutions, while other classes of initial data experience wave breaking in the sense that u_x becomes unbounded while the solution itself remains bounded. It suffices to treat the case $\kappa = 0$, since solutions with nonzero κ are obtained from solutions with zero κ by the transformation $v(x,t) = u(x + \kappa t, t) - \kappa$. More precisely, the fundamental existence theorem, due to Constantin and Escher [10], reads as follows: If $u_0 \in H^3([0,1])$ and $m_0 := u_0 - u_0'' \in H^1([0,1])$ is nonnegative, then (1.2) has a unique global solution $u \in C([0,T), H^3([0,1])) \cap C^1([0,T), H^2([0,1]))$ for any T positive. However, if $m_0 \in H^1([0,1])$, with u_0 not identically zero but $\int m_0\,dx = 0$, then the maximal time interval of existence is finite. Furthermore, if $u_0 \in H^1([0,1])$

*Received by the editors July 20, 2004; accepted for publication (in revised form) March 2, 2006; published electronically August 16, 2006. This work was supported in part by the Research Council of Norway.

http://www.siam.org/journals/sinum/44-4/61197.html

†Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, and Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053, Blindern, NO-0316 Oslo, Norway (holden@math.ntnu.no).

‡Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway (raynaud@math.ntnu.no).

and $m_0 = u_0 - u_0''$ is a positive Radon measure on $[0,1]$, then (1.2) has a unique global weak solution. Additional results in the periodic case can be found in [7, 10, 8, 11, 18]. Numerical results can be found in [4], where Camassa, Holm, and Hyman study (1.2) using a pseudospectral method. Numerical schemes based on multipeakons are examined in [2, 6, 5, 15].

In this paper, we prove convergence of a particular finite difference scheme for the equation, thereby giving a constructive approach to the actual determination of the solution. We work in the case where one has global solutions, that is, when $m_0 \geq 0$. The scheme is semidiscrete: Time is not discretized, and we have to solve a system of ordinary differential equations. We reformulate (1.1) to give meaning in $C([0,T]; H^1[0,1])$ to solutions such as peakons, and we prove that our scheme converges in $C([0,T]; H^1[0,1])$.

More precisely, we prove the following: Assume that v^n is a sequence of continuous, periodic, and piecewise linear functions on intervals $[(i-1)/n, i/n]$, $i = 1, \dots, n$, that converges to the initial data v in $H^1([0,1])$ as $n \to \infty$. Let $u^n = u^n(x,t)$ be the solution of the following system of equations:

$$\begin{aligned} m_t^n &= -D_-(m^n u^n) - m^n D u^n, \\ m^n &= u^n - D_- D_+ u^n, \end{aligned} \tag{1.3}$$

with initial condition $u^n|_{t=0} = v^n$. Here $D_\pm$ denotes forward and backward difference operators relative to the lattice with spacing $1/n$, and $D = (D_+ + D_-)/2$. Extrapolate u^n from its lattice values at points i/n to obtain a continuous, periodic, and piecewise linear function also denoted u^n. Assume that $v^n - D_- D_+ v^n \geq 0$. Then u^n converges in $C([0,T]; H^1([0,1]))$ as $n \to \infty$ to the solution u of the Camassa–Holm equation with initial condition $u|_{t=0} = v$. The result includes the case when the initial data $v \in H^1$ is such that $v - v_{xx}$ is a positive Radon measure; see Corollary 2.5. For the actual computations we discretize (1.3) using the forward Euler method. We prove convergence of that method; see Theorem 3.1.

The numerical scheme (1.3) is tested on various initial data. In addition, we study experimentally the convergence of other numerical schemes for the Camassa–Holm equation. The numerical results are surprisingly sensitive in the explicit form of the scheme, and, among the various schemes we have implemented, only the scheme (1.3) converges to the unique solution.

2. Convergence of the numerical scheme. We consider periodic boundary conditions and solve the equation on the interval $[0,1]$. We are looking for solutions that belong to $H^1([0,1])$, which is the natural space for the equation. Introduce the partition of $[0,1]$ in points separated by a distance $h = 1/n$ denoted $x_i = hi$ for $i = 0, \dots, n-1$. For any $(u_0, \dots, u_{n-1})$ in $\mathbb{R}^n$, we can define a continuous, periodic, piecewise linear function u by

$$u(x_i) = u_i, \tag{2.1}$$

in other words, the periodic polygon that passes through the points (x_i, u_i) for $i = 0, \dots, n-1$. It defines a bijection between $\mathbb{R}^n$ and the set of continuous, periodic, piecewise linear functions with possible break points at x_i, and we will use this bijection throughout this paper.

Given $u = (u_0, \dots, u_{n-1})$, the quantity $D_\pm u$ given by

$$(D_\pm u)_i = \frac{\pm 1}{h}(u_{i\pm 1} - u_i)$$

gives the right and left derivatives, respectively, of u at x_i. In these expressions, u_{-1} and u_n are derived from the periodicity conditions: $u_{-1} = u_{n-1}$ and $u_n = u_0$. The average Du between the left and right derivatives is given by

$$(Du)_i = \frac{1}{2}\big((D_+u)_i + (D_-u)_i\big) = \frac{1}{2h}(u_{i+1} - u_{i-1}).$$

The Camassa–Holm equation preserves the H^1-norm. In order to see that, we rewrite (1.2) in its Hamiltonian form (see [3]):

$$m_t = -(mu)_x - mu_x, \tag{2.2}$$

with

$$m = u - u_{xx}. \tag{2.3}$$

Assuming that u is smooth enough so that integration by parts can be carried out, we get

$$\begin{aligned}\frac{d}{dt}\|u\|_{H^1}^2 &= 2\int_0^1 (u_t - u_{xxt})u\,dx = 2\int_0^1 um_t\,dx\\ &= -2\int_0^1 u(mu)_x\,dx - 2\int_0^1 umu_x\,dx\\ &= 2\int_0^1 u_xmu\,dx - 2\int_0^1 umu_x\,dx = 0,\end{aligned}$$

and the H^1-norm of u is preserved.

From (2.3) and (2.2), we derive a finite difference approximation scheme for the Camassa–Holm equation and prove that it converges to the right solution. This is our main result.

THEOREM 2.1. *Let v^n be a sequence of continuous, periodic, and piecewise linear functions on $[0,1]$ that converges to v in $H^1([0,1])$ as $n \to \infty$ and such that $v^n - D_-D_+v^n \geq 0$. Then, for any given $T > 0$, the sequence $u^n = u^n(x,t)$ of continuous, periodic, and piecewise linear functions determined by the system of ordinary differential equations*

$$\begin{aligned}m_t^n &= -D_-(m^nu^n) - m^nDu^n,\\ m^n &= u^n - D_-D_+u^n,\end{aligned} \tag{2.4}$$

with initial condition $u^n|_{t=0} = v^n$, converges in $C([0,T];H^1([0,1]))$ as $n \to \infty$ to the solution u of the Camassa–Holm equation (1.2) *with initial condition $u|_{t=0} = v$.*

If we interpret the functions as vectors in (2.4) (cf. (2.1)), the multiplications are term-by-term multiplications of vectors. We also have to rewrite (1.2) in order to make it well defined in the sense of distributions for functions that at least belong to $C([0,T];H^1([0,1]))$; more precisely,

$$u_t - u_{xxt} = -\frac{3}{2}(u^2)_x - \frac{1}{2}(u_x^2)_x + \frac{1}{2}(u^2)_{xxx}. \tag{2.5}$$

A function u in $L^\infty([0,T];H^1)$ is said to be a solution of the periodic Camassa–Holm equation if it is periodic and satisfies (2.5) in the sense of distributions. In [11], a

different definition of weak solutions for the Camassa–Holm equation is presented. After proving our main theorem at the end of this section, we also prove that these two definitions are equivalent.

In order to solve (2.4), we need to compute u^n from m^n. It is simpler first to consider sequences that are defined in $\mathbb{R}^{\mathbb{Z}}$, the set of all sequences, and then discuss the periodic case. Let L denote the linear operator from $\mathbb{R}^{\mathbb{Z}}$ to $\mathbb{R}^{\mathbb{Z}}$ given, for all $u \in \mathbb{R}^{\mathbb{Z}}$, by

$$Lu = u - D_- D_+ u.$$

We want to find an expression for L^{-1}. Introduce the Kronecker delta by $\delta_i = 1$ if $i = 0$, and zero otherwise. It is enough to find a solution g of

$$Lg = \delta$$

which decays sufficiently fast at infinity because $L^{-1}m$ is then given, for any bounded $m \in \mathbb{R}^{\mathbb{Z}}$, by the discrete convolution product of g and m:

$$L^{-1}m_i = \sum_{j \in \mathbb{Z}} g_{i-j} m_j.$$

For i nonzero the function g satisfies

$$g_i - n^2(g_{i+1} - 2g_i + g_{i-1}) = 0. \tag{2.6}$$

The general solution of (2.6) for all $i \in \mathbb{Z}$ is given by

$$g_i = Ae^{\kappa_1 i} + Be^{\kappa_2 i},$$

where A, B are constants, $\kappa_1 = \ln x_1$, $\kappa_2 = \ln x_2$, and x_1 and x_2 are the solutions of

$$-n^2x^2 + (1 + 2n^2)x - n^2 = 0.$$

Here x_1 and x_2 are real and positive, and $x_1x_2 = 1$ implies that $\kappa_2 = -\kappa_1$. We set $\kappa = \kappa_1 = -\kappa_2$. After some calculations, we get

$$\kappa = \ln\left(\frac{1 + 2n^2 + \sqrt{1 + 4n^2}}{2n^2}\right). \tag{2.7}$$

We take g of the form

$$g_i = c\, e^{-\kappa|i|}$$

so that g satisfies (2.6) for all $i \neq 0$ and decays at infinity. The constant c is determined by the condition that $(Lg)_0 = 1$, which yields

$$c = \frac{1}{1 + 2n^2(1 - e^{-\kappa})}.$$

We periodize g in the following manner:

$$g_i^p \equiv \sum_{k \in \mathbb{Z}} g_{i+kn} = c\frac{e^{-\kappa i} + e^{\kappa(i-n)}}{1 - e^{-\kappa n}}$$

for $i \in \{0, \dots, n-1\}$. The inverse of L on the set of periodic sequences is then given by

$$u_i = L^{-1} m_i = \sum_{j=0}^{n-1} g_{i-j}^p m_i = \frac{c}{1 - e^{-\kappa n}} \sum_{j=0}^{n-1} (e^{-\kappa(i-j)} + e^{\kappa(i-j-n)}) m_j. \tag{2.8}$$

Hence,

$$L\left(\sum_{j=0}^{n-1} g_{i-j}^p m_j \right)_i = L\left(\sum_{l \in \mathbb{Z}} g_{i-l} m_l \right)_i = m_i.$$

For sufficiently smooth initial data ($u_0 \in H^3$ and $m_0 \in H^1$) which satisfies $m_0 \geq 0$, Constantin and Escher [9] proved that there exists a unique global solution of the Camassa–Holm equation belonging to $C(\mathbb{R}_+; H^3) \cap C^1(\mathbb{R}_+; H^2)$. The proof of this result relies heavily on the fact that if m is nonnegative at $t = 0$, then m remains nonnegative for all $t > 0$. An important feature of our scheme is that it preserves this property. (For simplicity we have here dropped the superscript n appearing on u and m.)

LEMMA 2.2. *Assume that $m_i(0) \geq 0$ for all $i = 0, \dots, n-1$. For any solution $u(t)$ of the system (2.4), we have that $m_i(t) \geq 0$ for all $t \geq 0$ and for all $i = 0, \dots, n-1$.*

Proof. Let us assume that there exist $t > 0$ and $i \in \{0, \dots, n-1\}$ such that

$$m_i(t) < 0. \tag{2.9}$$

We consider the time interval F in which m remains positive:

$$F = \{t \geq 0 \mid m_i(\tilde{t}) \geq 0 \text{ for all } \tilde{t} \leq t \text{ and } i \in \{0, \dots, n-1\}\}.$$

Because of assumption (2.9), F is bounded and we define

$$T = \sup F.$$

By definition of T, for any integer $j > 0$, there exists a $\tilde{t}_j$ and an i_j such that $T < \tilde{t}_j < T + \frac{1}{j}$ and $m_{i_j}(\tilde{t}_j) < 0$. The function $m_{i_j}(t)$ is a continuously differentiable function of t. Hence, $m_{i_j}(T) \geq 0$ and there exists a t_j such that

$$m_{i_j}(t_j) = 0,$$

with $T \leq t_j < T + \frac{1}{j}$.

Since i_j can only take a finite number of values ($i_j \in \{0, \dots, n-1\}$), there exists a $p \in \{0, \dots, n-1\}$ and a subsequence j_k such that $i_{j_k} = p$. The function $m_p(t)$ belongs to C^1 and, since $t_{j_k} \to T$, we have

$$m_p(T) = 0. \tag{2.10}$$

We denote by G the set of indices for which (2.10) holds:

$$G = \{k \in \{0, \dots, n-1\} \mid m_k(T) = 0\}.$$

G is nonempty because it contains p. If $G = \{0, \dots, n-1\}$, then $m_k(T) = 0$ for all k and m must be the zero solution, because we know from Picard's theorem that the solution of (2.4) is unique.

If $G \neq \{0, \dots, n-1\}$, then there exists an $l \in \{0, \dots, n-1\}$ such that

$$m_{l-1}(T) > 0, \quad m_l(T) = 0, \quad \frac{dm_l}{dt}(T) \leq 0. \tag{2.11}$$

The last condition, $\frac{dm_l}{dt}(T) \leq 0$, comes from the definition of T that would be contradicted if we had $\frac{dm_l}{dt}(T) > 0$. Note that we also use the periodicity of m, which in particular means that if $l = 0$, then $m_{l-1}(T) = m_{-1}(T) = m_{n-1}(T)$.

In (2.4), for $i = l$ and $t = T$, the terms involving $m_l(T)$ cancel and

$$\frac{dm_l}{dt}(T) = \frac{m_{l-1}(T)u_{l-1}(T)}{h}.$$

The fact that all the $m_i(T)$ are positive, with one of them, $m_{l-1}(T)$, strictly positive, implies that u_i is strictly positive for all indices i; see (2.8). Since, in addition, $m_{l-1}(T) > 0$, we get

$$\frac{dm_l}{dt}(T) > 0,$$

which contradicts the last inequality in (2.11), and therefore our primary assumption (2.9) does not hold. The lemma is proved. □

We want to establish a uniform bound on the H^1-norm of the sequence u^n. Recall that u^n is a continuous piecewise linear function (with respect to the space variable), and its L^2-norm can be computed exactly. We find

$$\|u^n\|_{L^2}^2 = \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{3}((u_{i+1}^n)^2 + u_i^n u_{i+1}^n + (u_i^n)^2). \tag{2.12}$$

The derivative u_x^n of u^n is piecewise constant, and therefore we have

$$\|u_x^n\|_{L^2}^2 = \frac{1}{n}\sum_{i=0}^{n-1}(D_+u^n)_i^2. \tag{2.13}$$

We define a renormalized norm $\|\cdot\|_{l^2}$ and the corresponding scalar product on $\mathbb{R}^n$ by

$$\|u^n\|_{l^2} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(u_i^n)^2}, \quad \langle u^n, v^n\rangle_{l^2} = \frac{1}{n}\sum_{i=0}^{n-1}u_i^n v_i^n.$$

The following inequalities hold:

$$\frac{1}{2}\|u^n\|_{l^2} \leq \|u^n\|_{L^2} \leq \|u^n\|_{l^2}, \tag{2.14}$$

which make the two norms $\|\cdot\|_{l^2}$ and $\|\cdot\|_{L^2}$ uniformly equivalent independently of n. In (2.14), u^n either denotes an element of $\mathbb{R}^n$ or the corresponding continuous piecewise linear function as defined previously. By using the Cauchy–Schwarz inequality and the periodicity of u^n, it is not hard to prove that

$$\|u^n\|_{L^2} \leq \|u^n\|_{l^2}.$$

For the other equality, it suffices to see that (2.12) can be rewritten as

$$\|u^n\|_{L^2}^2 = \frac{1}{3n}\sum_{i=0}^{n-1}\left[\left(u_{i+1}^n + \frac{1}{2}u_i^n\right)^2 + \frac{3}{4}(u_i^n)^2\right],$$

which implies

$$\frac{1}{2}\|u^n\|_{l^2} \le \|u^n\|_{L^2}.$$

We are now in a position to establish a uniform bound on the H^1-norm of u^n. Let $E_n(t)$ denote

$$E_n(t) = \left(\|u^n(t)\|_{l^2}^2 + \|D_+u^n(t)\|_{l^2}^2\right)^{\frac{1}{2}}, \tag{2.15}$$

which provides an approximation of the H^1-norm of $u^n(t)$. We have, from (2.14) and (2.13),

$$\frac{1}{2}\|u^n(t)\|_{H^1} \le E_n(t) \le \|u^n(t)\|_{H^1}. \tag{2.16}$$

The derivative of $E_n(t)^2$ reads

$$\begin{aligned}
\frac{dE_n(t)^2}{dt} &= \frac{2}{n}\sum_{i=0}^{n-1}\left[u_i^n u_{i,t}^n + D_+u_i^n D_+u_{i,t}^n\right] \\
&= \frac{2}{n}\sum_{i=0}^{n-1}(u_i^n - D_-D_+u_i^n)_t u_i^n \quad \text{(summation by parts)} \\
&= -\frac{2}{n}\sum_{i=0}^{n-1}\left[D_-(m^nu^n)_i u_i^n + m_i^n Du_i^n u_i^n\right] \quad \text{by (2.4)} \\
&= \frac{2}{n}\sum_{i=0}^{n-1}\left[m_i^n u_i^n (D_+u_i^n - Du_i^n)\right].
\end{aligned}$$

Since

$$D_+u_i^n - Du_i^n = \frac{1}{2}\left[D_+u_i^n - D_+u_{i-1}^n\right] = \frac{1}{2n}D_-D_+u_i^n,$$

we get

$$\frac{dE_n(t)^2}{dt} = \frac{1}{n}\sum_{i=0}^{n-1}\left[m_i^n u_i^n \frac{1}{n}D_-D_+u_i^n\right] = \frac{1}{n^2}\sum_{i=0}^{n-1}\left[m_i^n u_i^n(-m_i^n + u_i^n)\right], \tag{2.17}$$

and, because u_i^n is positive (see (2.8)),

$$\frac{dE_n^2(t)}{dt} \le \frac{1}{n^2}\sum_{i=0}^{n-1} m_i^n (u_i^n)^2. \tag{2.18}$$

A summation by parts gives us that

$$\frac{1}{n}\sum_{i=0}^{n-1} m_i^n u_i^n = E_n(t)^2.$$

Since L^∞ is continuously embedded in H^1, there exists a constant $\mathcal{O}(1)$, independent of n, such that

$$\max_i u_i^n \le \mathcal{O}(1)\,\|u^n\|_{H^1} \le \mathcal{O}(1)E_n(t).$$

Hence, (2.18) implies

$$E_n'(t) \le \frac{\mathcal{O}(1)}{n} E_n(t)^2$$

and, after integration,

$$\frac{1}{E_n(t)} \ge \frac{1}{E_n(0)} - \frac{\mathcal{O}(1)}{n} t.$$

Since $u^n(0) = v^n$ tends to v in H^1, $\|u^n(0)\|_{H^1}$ and therefore $E_n(0)$ are bounded. It implies that $E_n(0)^{-1}$ is bounded from below by a strictly positive constant and, for any given $T > 0$, there exists $N \ge 0$ and constant $C' > 0$ such that for all $n \ge N$ and all $t \in [0, T]$, we have $E_n(0)^{-1} - \mathcal{O}(1)t/n \ge 1/C'$. Hence,

$$\|u^n\|_{H^1} \le 2E_n(t) \le 2C' \tag{2.19}$$

and, by (2.16), the H^1-norm of $u^n(t)$ is uniformly bounded in $[0, T]$. This result also guarantees the existence of solutions to (2.4) in $[0, T]$ (at least, for n big enough) because, on $[0, T]$, we have that $\max_i |u_i^n(t)| = \|u^n(\cdot\,, t)\|_{L^\infty} \le \mathcal{O}(1)\,\|u^n(t)\|_{H^1}$ remains bounded.

To prove that we can extract a converging subsequence of u^n, we need some estimates on the derivative of u^n.

LEMMA 2.3. *We have the following properties:*

(i) u_x^n *is uniformly bounded in* $L^\infty([0, 1])$.
(ii) u_x^n *has a uniformly bounded total variation.*
(iii) u_t^n *is uniformly bounded in* $L^2([0, 1])$.

Proof. (i) From (2.8), we get

$$D_+ u_i^n = \frac{c}{1 - e^{-\kappa n}} \sum_{j=0}^{n-1} \left[m_j^n e^{-\kappa(i-j)} \left(\frac{e^{-\kappa} - 1}{h} \right) + m_j^n e^{\kappa(i-j-n)} \left(\frac{e^{\kappa} - 1}{h} \right) \right],$$

where κ is given by (2.7).

One easily gets the following expansion for κ as h tends to 0:

$$\kappa = h + o(h^2),$$

which implies that for all $i \in \{0, \dots, n-1\}$,

$$\begin{aligned} |D_+ u_i^n| &\le (1 + \mathcal{O}(h)) \frac{c}{1 - e^{-\kappa n}} \sum_{j=0}^{n-1} \left(\left| m_j^n \right| e^{-\kappa(i-j)} + \left| m_j^n \right| e^{\kappa(i-j-n)} \right) \\ &\le (1 + \mathcal{O}(h)) \frac{c}{1 - e^{-\kappa n}} \sum_{j=0}^{n-1} \left(m_j^n e^{-\kappa(i-j)} + m_j^n e^{\kappa(i-j-n)} \right) \\ &\le (1 + \mathcal{O}(h))\, u_i^n, \end{aligned} \tag{2.20}$$

where we have used the positivity of m^n and relation (2.8). Hence, since $\|u^n\|_{L^\infty}$ is uniformly bounded, we get a uniform bound on $\|u_x^n\|_{L^\infty}$.

(ii) For each t the total variation of $u_x^n(\,\cdot\,,t)$ is given by

$$\mathrm{TV}(u_x^n) = \sup_{\phi\in C^1, \|\phi\|_{L^\infty}\le 1} \int_0^1 u_x^n(x)\phi_x(x)\,dx.$$

On the interval (x_i, x_{i+1}), the function u_x^n is constant and equal to $D_+u_i^n$. Therefore,

$$\begin{aligned}\int_0^1 u_x^n(x)\phi_x(x)\,dx &= \sum_{i=0}^{n-1} D_+u_i^n \int_{x_i}^{x_{i+1}} \phi_x(x)\,dx = \sum_{i=0}^{n-1} D_+u_i^n(\phi(x_{i+1})-\phi(x_i))\\ &= \sum_{i=0}^{n-1} \frac{1}{n} D_+u_i^n D_+\phi(x_i) = -\sum_{i=0}^{n-1} \frac{1}{n}(D_-D_+u_i^n)\phi(x_i)\end{aligned}$$

and

$$\mathrm{TV}(u_x^n) \le \frac{1}{n}\sum_{i=0}^{n-1} |D_-D_+u_i^n|\,.$$

Since m_i^n and u_i^n are positive for all i,

$$|D_-D_+u_i^n| = |m_i^n - u_i^n| \le m_i^n + u_i^n \le 2u_i^n - D_-D_+u_i^n.$$

When summing over i on the right-hand side of the last inequality, we see that the term $D_-D_+u_i^n$ disappears, and we get

$$\mathrm{TV}(u_x^n) \le 2\max_i u_i^n \le \mathcal{O}(1)\,\|u^n\|_{H^1} \le \mathcal{O}(1)$$

for all t.

(iii) In order to make the ideas clearer, we first sketch the proof directly on (2.2). Assuming that m is positive and u is in H^1, we see how, from (2.2), u_t can be defined as an element of $L^2([0,1])$. This will be useful when we afterwards derive a uniform bound for u_t^n in $L^2([0,1])$.

For all smooth v, we have

$$\int_0^1 u_t\, v\,dx = \int_0^1 (\mathcal{L}^{-1}m_t)\, v\,dx,$$

where $\mathcal{L}$ denotes the operator $\mathcal{L}u = u - u_{xx}$, which is a self-adjoint homeomorphism from H^2 to L^2. If we let $w = \mathcal{L}^{-1}v$, the continuity of $\mathcal{L}^{-1}$ implies

$$\|w\|_{H^2} \le \mathcal{O}(1)\,\|v\|_{L^2} \tag{2.21}$$

for some constant $\mathcal{O}(1)$ independent of v.

We find

$$\begin{aligned}\int_0^1 u_t\, v\,dx &= \int_0^1 \left(\mathcal{L}^{-1}m_t\right) v\,dx = \int_0^1 m_t\,\mathcal{L}^{-1}v\,dx \quad (\mathcal{L}^{-1}\text{ is self-adjoint})\\ &= -\int_0^1 ((mu)_x + mu_x)w\,dx = \int_0^1 (muw_x - mu_xw)\,dx.\end{aligned}$$

The integrals here must be understood as distributions. Even so, some terms (like mu_x) are not well defined as distributions. However, we get the same results rigorously by considering the equation written as a distribution (2.5). We have

$$\left|\int_0^1 u_t\, v\, dx\right| \le \int_0^1 (|muw_x| + |mu_x w|)\, dx \\ \le (\|u\|_{L^\infty}\, \|w_x\|_{L^\infty} + \|u_x\|_{L^\infty}\, \|w\|_{L^\infty}) \int_0^1 |m|\ dx.$$

Recall that $\|u\|_{L^\infty}$ and $\|u_x\|_{L^\infty}$ are uniformly bounded. Furthermore, m positive implies $\int_0^1 |m| = \int_0^1 m = \int_0^1 u \le \|u\|_{L^\infty}$, and therefore m is also uniformly bounded. From (2.21) and the fact that H^1 is continuously embedded in L^∞, we get

$$\|w_x\|_{L^\infty} \le \mathcal{O}(1)\, \|w_x\|_{H^1} \le \mathcal{O}(1)\, \|w\|_{H^2} \le \mathcal{O}(1)\, \|v\|_{L^2}\,,$$

and similarly

$$\|w\|_{L^\infty} \le \mathcal{O}(1)\, \|v\|_{L^2}\,.$$

Finally,

$$\left|\int_0^1 u_t\, v\, dx\right| \le \mathcal{O}(1)\, \|v\|_{L^2}\,,$$

which implies, by Riesz's representation theorem, that u_t is in L^2 and

$$\|u_t\|_{L^2} \le \mathcal{O}(1).$$

We now turn to the analogous derivations in the discrete case. Consider the sequence u^n. The aim is to derive a uniform bound for u_t^n in L^2. We take a continuous piecewise linear function v^n,

$$\langle u_t^n, v^n\rangle_{l^2} = \left\langle L^{-1} m_t^n, v^n\right\rangle_{l^2} = \left\langle m_t^n, L^{-1} v^n\right\rangle_{l^2}\,, \tag{2.22}$$

because L and therefore L^{-1} are self-adjoint.

Let w^n denote

$$w^n = L^{-1} v^n.$$

We have

$$\langle v^n, w^n\rangle_{l^2} = \langle Lw^n, w^n\rangle_{l^2} = \frac{1}{n}\sum_{i=0}^{n-1} (w_i^n - D_- D_+ w_i^n)\, w_i^n = \frac{1}{n}\sum_{i=0}^{n-1} \left[(w_i^n)^2 + (D_+ w_i^n)^2\right].$$

Then, after using (2.16) and Cauchy–Schwarz, we get

$$\|w^n\|_{H^1}^2 \le 4\, \|v^n\|_{l^2}\, \|w^n\|_{l^2}\,.$$

By (2.14), (2.16) we find

$$\|w^n\|_{H^1}^2 \le \mathcal{O}(1)\, \|v^n\|_{l^2}\, \|w^n\|_{H^1}$$

and

$$\|w^n\|_{H^1} \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,, \tag{2.23}$$

where $\mathcal{O}(1)$ is a constant independent of n. Since H^1 is continuously embedded in L^∞, we get

$$\max_i |w_i^n| \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,. \tag{2.24}$$

Let us define y^n as follows:

$$y_i^n = (D_+ w^n)_{i-1}.$$

We want to find a bound on y^n. From (2.14) and (2.23), we get

$$\|y^n\|_{l^2} \leq \|w^n\|_{H^1} \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,. \tag{2.25}$$

We also have, using the definition of y^n and w^n,

$$D_+ y^n = D_- D_+ w^n = w^n - v^n,$$

which gives

$$\|D_+ y^n\|_{l^2} \leq \mathcal{O}(1)\,\|v^n\|_{l^2} \tag{2.26}$$

because, by (2.23),

$$\|w^n\|_{l^2} \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,.$$

Equations (2.25), (2.26), and (2.16) give us a uniform bound on the H^1-norm of y^n:

$$\|y^n\|_{H^1} \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,.$$

Since H^1 is continuously embedded in L^∞, we get

$$\max_i |D_+ w_i^n| = \max_i |y_i^n| = \|y^n\|_{L^\infty} \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,. \tag{2.27}$$

Going back to (2.22), we have

$$\begin{aligned}\langle u_t^n, v^n\rangle_{l^2} &= \langle m_t^n, w^n\rangle_{l^2} = \langle -D_-(m^n u^n) - m^n D u^n, w^n\rangle_{l^2} \\ &= \langle m^n u^n, D_+ w^n\rangle_{l^2} - \langle m^n D u^n, w^n\rangle_{l^2}\,.\end{aligned}$$

Hence,

$$|\langle u_t^n, v^n\rangle_{l^2}| \leq \frac{1}{n}\left(\max_i |u_i^n| \max_i |D_+ w_i^n| + \max_i |D_+ u_i^n| \max_i |w_i^n|\right) \sum_{i=0}^{n-1} |m_i^n|\,.$$

The functions u_i^n and $D_+ u_i^n$ are uniformly bounded with respect to n and

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1} |m_i^n| &= \frac{1}{n}\sum_{i=0}^{n-1} m_i^n \quad (m^n \text{ is positive}) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} u_i^n \quad \left(\text{cancellation of } \textstyle\sum_{i=0}^{n-1} D_- D_+ u_i^n\right) \\ &\leq \mathcal{O}(1) \qquad (u_i^n \text{ is bounded}).\end{aligned}$$

Finally, using the bounds we have derived on w^n (see (2.24)) and D_+w^n (see (2.27)), we get

$$|\langle u_t^n, v^n\rangle_{l^2}| \leq \mathcal{O}(1)\,\|v^n\|_{l^2}\,.$$

Taking $v^n = u_t^n$ yields

$$\|u_t^n\|_{l^2} \leq \mathcal{O}(1),$$

which, since the l^2- and L^2-norms are uniformly equivalent, gives us a uniform bound on $\|u_t^n\|_{L^2}$. □

To prove the existence of a converging subsequence of u^n in $C([0,T],H^1)$ we recall the following compactness theorem given by Simon [21, Corollary 4].

THEOREM 2.4 (Simon [21]). *Let X, B, Y be three continuously embedded Banach spaces*

$$X \subset B \subset Y,$$

with the first inclusion, $X \subset B$, compact. We consider a set $\mathcal{F}$ of functions mapping $[0,T]$ into X. If $\mathcal{F}$ is bounded in $L^\infty([0,T],X)$ and $\frac{\partial \mathcal{F}}{\partial t} = \{\frac{\partial f}{\partial t} \mid f \in \mathcal{F}\}$ is bounded in $L^r([0,T],Y)$, where $r > 1$, then $\mathcal{F}$ is relatively compact in $C([0,T],B)$.

We now turn to the proof of our main theorem.

Proof of Theorem 2.1. (i) First we establish that there exists a subsequence of u^n that converges in $C([0,T],H^1)$ to an element $u \in H^1$. To apply Theorem 2.4, we have to determine the Banach spaces with the required properties. In our case, we take X as the set of functions of H^1 which have derivatives of bounded variation:

$$X = \left\{v \in H^1 \mid v_x \in \mathrm{B}V\right\}.$$

X endowed with the norm

$$\|v\|_X = \|v\|_{H^1} + \|v_x\|_{BV} = \|v\|_{H^1} + \|v_x\|_{L^\infty} + \mathrm{TV}(v_x)$$

is a Banach space. Let us prove that the injection $X \subset H^1$ is compact. We consider a sequence v_n which is bounded in X. Since $\|v_n\|_{L^\infty}$ is bounded ($H^1 \subset L^\infty$ continuously), there exists a point x_0 such that $v_n(x_0)$ is bounded and we can extract a subsequence (that we still denote v_n) such that $v_n(x_0)$ converges to some $l \in \mathbb{R}$. By Helly's theorem, we can also extract a subsequence such that

$$v_{n,x} \to w \quad \text{a.e.} \tag{2.28}$$

for some $w \in L^\infty$. By Lebesgue's dominated convergence theorem, it implies that $v_{n,x} \to w$ in L^2. We set

$$v(x) = l + \int_{x_0}^{x} w(s)\,ds.$$

We have that $v_x = w$ a.e. We also have

$$v_n(x) = v_n(x_0) + \int_{x_0}^{x} v_{n,x}(s)\,ds,$$

which together with (2.28) implies that v_n converges to v in L^∞. Therefore v_n converges to v in H^1 and X is compactly embedded in H^1.

The estimates we have derived previously give us that u^n and u_t^n are uniformly bounded in $L^\infty([0,T],X)$ and $L^\infty([0,T],L^2)$, respectively. Since $X \subset H^1 \subset L^2$ with the first inclusion compact, Simon's theorem gives us the existence of a subsequence of u^n that converges in $C([0,T],H^1)$ to some $u \in H^1$.

(ii) Next we show that the limit we get is a solution of the Camassa–Holm equation (1.2).

Let us now take φ in $C^\infty([0,1]\times[0,T])$ and multiply, for each i, the first equation in (2.4) by $h\varphi(x_i,t)$. We denote by φ^n the continuous piecewise linear function given by $\varphi^n(x_i,t)=\varphi(x_i,t)$. We sum over i and get, after one summation by parts,

$$\begin{aligned}\sum_{i=0}^{n-1} h\left(u_{i,t}^n - (D_-D_+u_i^n)_t\right)\varphi_i^n = &\underbrace{\sum_{i=0}^{n-1} h(u_i^n)^2 D_+\varphi_i}_{\text{A}} - \underbrace{\sum_{i=0}^{n-1} hu_i^n D_-D_+u_i^n D_+\varphi_i^n}_{\text{B}}\\ &\underbrace{-\sum_{i=0}^{n-1} hu_i^n Du_i^n\varphi_i^n}_{\text{C}} + \underbrace{\sum_{i=0}^{n-1} hD_-D_+u_i^n Du_i^n\varphi_i^n}_{\text{D}}.\end{aligned} \tag{2.29}$$

We are now going to prove that each term in this equality converges to the corresponding terms in (2.5).

Term A. We want to prove that

$$\left\langle (u^n)^2 D_+\varphi^n\right\rangle \to \int_0^1 u^2\varphi_x\,dx, \tag{2.30}$$

where we have introduced the notation

$$\langle u\rangle = h\sum_{i=0}^{n-1} u_i$$

to denote the average of a quantity u. We have

$$\begin{aligned}\left|\int_0^1 u^2\varphi_x\,dx - \left\langle (u^n)^2 D_+\varphi^n\right\rangle\right| \le &\left|\int_0^1 (u^2-(u^n)^2)\varphi_x\,dx\right|\\ &+\left|\int_0^1 (u^n)^2(\varphi_x - D_+\varphi^n)\,dx\right|\\ &+\left|\int_0^1 (u^n)^2 D_+\varphi^n\,dx - \left\langle (u^n)^2 D_+\varphi^n\right\rangle\right|.\end{aligned}$$

The first term tends to zero because $u^n \to u$ in L^2 for all $t\in[0,T]$. The second tends to zero by Lebesgue's dominated convergence theorem. It remains to prove that the last term tends to zero.

The integral of a product between two continuous piecewise linear function, v and w, and a piecewise constant function z can be computed explicitly. We skip the details of the calculation and directly give the result:

$$\int_0^1 zvw\,dx = \frac{1}{3}\langle zS_+vS_+w\rangle + \frac{1}{6}\langle zS_+vw\rangle + \frac{1}{6}\langle zvS_+w\rangle + \frac{1}{3}\langle zvw\rangle. \tag{2.31}$$

Here S_+ and S_- denote shift operators

$$(S_\pm u)_i = u_{i\pm 1}.$$

After using (2.31) with $v = w = u^n$ and $z = D_+\varphi^n$, we get

$$\int_0^1 (u^n)^2 D_+\varphi^n - \left\langle (u^n)^2 D_+\varphi^n \right\rangle = \frac{1}{3}\left\langle (S_+u^n - u^n) D_+\varphi^n u^n \right\rangle + \frac{1}{3}\left\langle (u^n)^2 D_+(S_-\varphi^n - \varphi^n)\right\rangle.$$

We use the uniform equivalence of the l^2- and L^2-norms to get the following estimate:

$$\begin{aligned}\left\langle (S_+u^n - u^n) D_+\varphi^n u^n \right\rangle &\le \|S_+u^n - u^n\|_{l^2}\, \|D_+\varphi^n u^n\|_{l^2} \quad \text{(Cauchy–Schwarz)}\\ &\le \mathcal{O}(1)\, \|u^n(\,\cdot + h) - u^n(\,\cdot\,)\|_{L^2}. \end{aligned} \tag{2.32}$$

Since $u_n \in H^1$, we have (see, for example, [1])

$$\|u^n(\,\cdot + h) - u^n(\,\cdot\,)\|_{L^2} \le h\, \|u_x^n\|_{L^2} \le \mathcal{O}(1)h$$

because $\|u_x^n\|_{L^\infty}$ is uniformly bounded. Hence $|\langle (S_+u^n - u^n) D_+\varphi^n u^n\rangle|$ tends to zero. The quantity $\left\langle (u^n)^2 D_+(S_-\varphi^n - \varphi^n)\right\rangle$ tends to zero because φ is C^∞ and u^n uniformly bounded. We have proved (2.30).

Term B. We want to prove

$$\langle u^n D_-D_+u^n D_+\varphi^n\rangle \to \frac{1}{2}\int_0^1 u^2\varphi_{xxx}\,dx - \int_0^1 u_x^2\varphi_x. \tag{2.33}$$

We rewrite $u^nD_-D_+u^n$ in such a way that the discrete double derivative D_-D_+ does not appear in a product (so that we can later sum by parts). We have

$$u^nD_-D_+u^n = \frac{1}{2}(D_-D_+((u^n)^2) - D_+u^nD_+u^n - D_-u^nD_-u^n).$$

We can prove in the same way as we did for term A that

$$\begin{aligned}\left\langle D_-D_+((u^n)^2) D_+\varphi^n\right\rangle &= \left\langle (u^n)^2 D_-D_+D_+\varphi^n\right\rangle \quad \text{(summation by parts)}\\ &\to \int_0^1 u^2\varphi_{xxx}\,dx.\end{aligned}$$

The quantity $(u_x^n)^2\varphi_x^n$ is a piecewise constant function. Therefore,

$$\int_0^1 (u_x^n)^2\varphi_x^n\,dx = \langle D_+u^nD_+u^nD_+\varphi^n\rangle.$$

Since $u_x^n \to$ in L^2 for all $t \in [0,T]$ and

$$\int_0^1 u_x^2\varphi_x\,dx - \langle D_+u^nD_+u^nD_+\varphi^n\rangle = \int_0^1 (u_x^2 - (u_x^n)^2)\varphi_x\,dx + \int_0^1 (u_x^n)^2(\varphi_x - \varphi_x^n)\,dx,$$

we have

$$\langle D_+u^nD_+u^nD_+\varphi^n\rangle \to \int_0^1 u_x^2\varphi_x\,dx.$$

In the same way, we get

$$\langle D_- u_i^n D_- u_i^n D_+ \varphi^n \rangle \to \int_0^1 u_x^2 \varphi_x$$

and (2.33) is proved.

Term C. We want to prove

$$\langle u^n D u^n \varphi^n \rangle \to \int_0^1 u u_x \varphi \, dx. \tag{2.34}$$

We have

$$\begin{aligned}\int_0^1 u u_x \varphi \, dx - \langle u^n D_+ u^n \varphi^n \rangle = &\int_0^1 (u - u^n) u_x \varphi \, dx + \int_0^1 u^n (u_x - u_x^n) \varphi \, dx \\ &+ \int_0^1 u^n u_x^n (\varphi - \varphi^n) \, dx + \int_0^1 u^n u_x^n \varphi^n \, dx \\ &- \langle u^n D_+ u^n \varphi^n \rangle .\end{aligned}$$

The first two terms converge to zero because $u^n \to u$ in H^1 for all $t \in [0, T]$. The third term converges to zero by Lebesgue's dominated convergence theorem. We use (2.31) to evaluate the last integral:

$$\begin{aligned}\int_0^1 u^n u_x^n \varphi^n \, dx = &\frac{1}{3} \langle D_+ u^n S_+ u^n S_+ \varphi^n \rangle + \frac{1}{6} \langle D_+ u^n S_+ u^n \varphi^n \rangle \\ &+ \frac{1}{6} \langle D_+ u^n u^n S_+ \varphi^n \rangle + \frac{1}{3} \langle D_+ u^n u^n \varphi^n \rangle .\end{aligned}$$

Using the same type of arguments as those we have just used for term A, we can show that

$$\int_0^1 u^n u_x^n \varphi^n \, dx \to \langle D_+ u^n u^n \varphi^n \rangle .$$

Thus, in order to prove (2.34), it remains to prove that

$$\langle D_+ u^n u^n \varphi^n \rangle - \langle D u^n u^n \varphi^n \rangle \to 0. \tag{2.35}$$

Since $D = \frac{1}{2}(D_+ + D_-)$, we have

$$\langle D_+ u^n u^n \varphi^n \rangle - \langle D u^n u^n \varphi^n \rangle = \frac{1}{2} \langle (D_+ u^n - D_- u^n) u^n \varphi^n \rangle$$

and

$$\begin{aligned}|\langle (D_+ u^n - D_- u^n) u^n \varphi^n \rangle| &\le C \sum_{i=0}^{n-1} h \left| D_+ u_i^n - D_+ u_{i-1}^n \right| \\ &\le \mathcal{O}(1) \int_0^1 |u_x^n(x) - u_x^n(x-h)| \, dx \\ &\le \mathcal{O}(1) h \ \mathrm{TV}(u_x^n).\end{aligned}$$

Since $\mathrm{TV}(u_x^n)$ is uniformly bounded, (2.35) holds and we have proved (2.34).

Term D. We want to prove that

$$\langle D_- D_+ u^n D u^n \varphi^n \rangle \to -\frac{1}{2} \int_0^1 u_x^2 \varphi_x \, dx. \tag{2.36}$$

We have

$$\frac{1}{2} \int_0^1 u_x^2 \varphi_x \, dx + \langle D_- D_+ u^n D u^n \varphi^n \rangle \tag{2.37}$$

$$= \frac{1}{2} \int_0^1 (u_x^2 - (u_x^n)^2) \varphi_x \, dx + \frac{1}{2} \int_0^1 (u_x^n)^2 (\varphi_x - D_- \varphi^n) \, dx \tag{2.38}$$

$$- \frac{1}{2} \langle D_+ (D_+ u^n D_+ u^n) \varphi^n \rangle + \langle D_- D_+ u^n D u^n \varphi^n \rangle . \tag{2.39}$$

The two first terms on the right-hand side tend to zero. After using the identity

$$D_+ (D_+ u^n D_+ u^n) = D_+ D_+ u^n D_+ u^n + D_+ D_+ u^n D_+ S_+ u^n,$$

we can rewrite the two last terms in (2.37) as

$$\begin{aligned}
&-\frac{1}{2} \langle D_+ (D_+ u^n D_+ u^n) \varphi^n \rangle + \langle D_- D_+ u^n D u^n \varphi^n \rangle \\
&\quad = -\frac{1}{2} \langle D_- D_+ S_+ u^n D_+ u^n \varphi^n \rangle - \frac{1}{2} \langle D_- D_+ S_+ u^n D_+ S_+ u^n \varphi^n \rangle \\
&\qquad + \frac{1}{2} \langle D_- D_+ u^n D_+ S_- u^n \varphi^n \rangle + \frac{1}{2} \langle D_- D_+ u^n D_+ u^n \varphi^n \rangle \\
&\quad = \frac{1}{2} \langle D_- D_+ u^n D_- u^n (\varphi^n - S_- \varphi^n) \rangle + \frac{1}{2} \langle D_- D_+ u^n D_+ u^n (\varphi^n - S_- \varphi^n) \rangle ,
\end{aligned}$$

which tends to zero because, as we have seen before, due to the positivity of m, $\langle |D_- D_+ u_i^n D_+ u_i^n| \rangle$ is uniformly bounded. We have proved (2.36).

Up to now we have not really considered the time variable. We integrate (2.29) with respect to time and integrate by parts the left-hand side:

$$\begin{aligned}
\int_0^T \sum_{i=0}^{n-1} h \left(u_{i,t}^n - D_- D_+ u_{i,t}^n \right) \varphi(x_i, t) \, dt = &- \int_0^T \sum_{i=0}^{n-1} h \left(u_i^n - D_- D_+ u_i^n \right) \varphi_t(x_i, t) \, dt \\
&+ \left[\sum_{i=0}^{n-1} h \left(u_i^n - D_- D_+ u_i^n \right) \varphi(x_i, t) \right]_{t=0}^{t=T};
\end{aligned}$$

after summing by parts, the limit of this expression is (we use Lebesgue's dominated convergence theorem with respect to x and t)

$$- \int_0^T \int_0^1 u(\varphi_t - \varphi_{txx}) \, dx dt + \left[\int_0^1 u(\varphi - \varphi_{xx}) \, dx \right]_{t=0}^{t=T}.$$

It is not hard to see that the right-hand side of (2.29) is uniformly bounded by a constant, and we can integrate over time and use the Lebesgue dominated convergence theorem to conclude that u is indeed a solution of (2.5) in the sense of distribution.

The analysis in [11] shows that the weak solution of the Camassa–Holm equation with initial conditions satisfying $m(x, 0) \geq 0$ is unique. This implies that in our

algorithm not only a subsequence but the whole sequence u^n converges to the solution. However, in [11], a solution of the Camassa–Holm equation is defined as an element u of H^1 satisfying

$$u_t + uu_x + \left[\int_{-\infty}^{\infty} p(x-y)[u^2(y,t) + \frac{1}{2}u_x^2(y,t)]\,dy\right]_x = 0, \tag{2.40}$$

where p is the solution of

$$\mathcal{A}p \equiv (I - \partial_x^2)p = \delta.$$

We want to prove that weak solutions of (2.40) and (2.5) are the same. Periodic distributions belong to the class of tempered distribution $\mathcal{S}'$ (see, for example, [13]). The operator $\mathcal{A}$ defines a homeomorphism on the Schwartz class $\mathcal{S}$ (or class of rapidly decreasing function): The Fourier transform is a homeomorphism on $\mathcal{S}$, and $\mathcal{A}$ restricted to $\mathcal{S}$ can be written as

$$\mathcal{A} = \mathcal{F}^{-1}(1+\xi^2)\mathcal{F}, \tag{2.41}$$

where ξ denotes the frequency variable. It is clear from (2.41) that the inverse of $\mathcal{A}$ in $\mathcal{S}$ is

$$\mathcal{A}^{-1} = \mathcal{F}^{-1}\frac{1}{1+\xi^2}\mathcal{F}.$$

Hence $\mathcal{A}$ is a homeomorphism on $\mathcal{S}$.

We can now define the inverse $\mathcal{A}^{-1}$ of $\mathcal{A}$ in $\mathcal{S}'$. Given T in $\mathcal{S}'$, $\mathcal{A}^{-1}T$ is given by

$$\left\langle \mathcal{A}^{-1}T, \phi \right\rangle = \left\langle T, \mathcal{A}^{-1}\phi \right\rangle, \quad \phi \in \mathcal{S}.$$

It is easy to check that $\mathcal{A}^{-1}$ indeed satisfies

$$\mathcal{A}^{-1}\mathcal{A} = \mathcal{A}\mathcal{A}^{-1} = \mathrm{Id},$$

and that $\mathcal{A}^{-1}$ is continuous on $\mathcal{S}'$. The operator $\mathcal{A}$ is therefore a homeomorphism on $\mathcal{S}'$.

Let u be a solution of (2.40). Then we have

$$u_t + \partial_x\left(\frac{u^2}{2}\right) + \partial_x\mathcal{A}^{-1}\left[u^2 + \frac{1}{2}u_x^2\right] = 0. \tag{2.42}$$

The operators ∂_x and $\mathcal{A}^{-1}$ commute because ∂_x and $\mathcal{A}$ commute. We apply $\mathcal{A}$ on both sides of (2.42) and get

$$u_t - u_{xxt} + \mathcal{A}\partial_x\left(\frac{1}{2}u^2\right) + \partial_x\left[u^2 + \frac{1}{2}u_x^2\right] = 0, \tag{2.43}$$

which is exactly (2.5). Since $\mathcal{A}$ is a bijection, (2.43) also implies (2.42), and we have proved that the weak solutions of (2.5) are the same as the weak solutions given by (2.40). □

In Theorem 2.1, some restrictions on the initial data v are implicitly imposed by the condition $v^n - D_-D_+v^n \geq 0$. We are going to prove that if $v \in H^1([0,1])$ is periodic with $v - v_{xx} \in \mathcal{M}^+$, where $\mathcal{M}^+$ denotes the space of positive Radon measures, then there exists a sequence of piecewise linear, continuous, periodic functions v^n that converges to v in H^1 and satisfies $v^n - D_-D_+v^n \geq 0$ for all n.

We can then apply Theorem 2.1 and get the existence result contained in the following corollary, which coincides with results obtained in [11] by a different method.

COROLLARY 2.5. *If $u_0 \in H^1$ is such that $u_0 - u_{0,xx} \in \mathcal{M}^+$, then the Camassa–Holm equation has a global solution in $C(\mathbb{R}_+, H^1)$. The solution is obtained as a limit of the numerical scheme defined by* (2.4).

To apply Theorem 2.1, we need to prove that, given $u \in H^1([0,1])$ such that $u - u_{xx} \in \mathcal{M}^+$, there exists a sequence u^n of piecewise linear, continuous, and periodic functions such that

$$u^n \to u \text{ in } H^1,$$
$$u^n - D_- D_+ u^n \geq 0.$$

Let $\{\psi_i^n\}$ be a partition of unity associated with the covering $\cup_{i=0}^{n-1}(x_{i-1}, x_{i+1})$. For all $i \in \{0, \dots, n-1\}$, the functions ψ_i^n are nonnegative with supp $\psi_i^n \subset (x_{i-1}, x_{i+1})$, and $\sum_{i=0}^{n-1} \psi_i^n = 1$. Define

$$v_i^n = \frac{1}{h} \langle u - u_{xx}, \psi_i^n \rangle$$

and

$$u_i^n - D_- D_+ u_i^n = v_i^n. \tag{2.44}$$

Recall that the operator $u^n - D_- D_+ u^n$ is invertible (see (2.8)), so that u^n is well defined by (2.44). Since $u - u_{xx}$ belongs to $\mathcal{M}^+$ and $\psi_i^n \geq 0$, we have $v_i^n = u_i^n - D_- D_+ u_i^n \geq 0$, and it only remains to prove that u^n converges to u in H^1. Since the application $\mathcal{L} : H^1 \to H^{-1}$ given by $\mathcal{L}u = u - u_{xx}$ is an homeomorphism, it is equivalent to prove that

$$u^n - u_{xx}^n \to u - u_{xx} \text{ in } H^{-1}.$$

The homeomorphism $\mathcal{L}$ is also an isometry, so that

$$\|\mathcal{L}u\|_{H^{-1}} = \|u\|_{H^1}.$$

We can find a bound on $\|u^n\|_{H^1}$. Let E_n be defined, as before, by

$$E_n = \left(h \sum_{i=0}^{n-1} \left[(u_i^n)^2 + (D_+ u^n)_i^2 \right] \right)^{\frac{1}{2}}.$$

Inequality (2.16) still holds. We have

$$\begin{aligned}
E_n^2 &= h \sum_{i=0}^{n-1} (u_i^n - D_- D_+ u_i^n) u_i^n \\
&= h \sum_{i=0}^{n-1} v_i^n u_i^n \\
&\leq \|u^n\|_{L^\infty} \sum_{i=0}^{n-1} h v_i^n \\
&\leq \|u^n\|_{L^\infty} \left\langle u - u_{xx}, \sum_{i=0}^{n-1} \psi_i^n \right\rangle \\
&\leq \|u^n\|_{L^\infty} \|u - u_{xx}\|_{\mathcal{M}^+} \quad \left(\text{since } \textstyle\sum_{i=0}^{n-1} \psi_i^n = 1\right).
\end{aligned}$$

Hence, since L^∞ is continuously embedded in H^1, there exists a constant C (independent of n) such that

$$E_n^2 \le C \|u^n\|_{H^1} \|u - u_{xx}\|_{\mathcal{M}^+}.$$

We use (2.16) to get the bound on $\|u^n\|_{H^1}$ we were looking for:

$$\|u^n\|_{H^1} \le 4C \|u - u_{xx}\|_{\mathcal{M}^+}.$$

To prove that $u^n - u^n_{xx} \to u - u_{xx}$ in H^{-1}, since $\|u^n - u^n_{xx}\|_{H^{-1}} = \|u^n\|_{H^1}$ is uniformly bounded, we just need to prove that

$$\langle u^n - u^n_{xx}, \varphi \rangle \to \langle u - u_{xx}, \varphi \rangle$$

for all φ belonging to a dense subset of H^1 (for example, C^∞).

The function u^n is continuous and piecewise linear. Its second derivative u^n_{xx} is therefore a sum of Dirac functions,

$$u^n_{xx} = \sum_{i=0}^{n-1} h D_- D_+ u_i^n \delta_{x_i},$$

and, for any φ in C^∞, we have

$$\begin{aligned}
\langle u^n - u^n_{xx}, \varphi \rangle &= \int_0^1 u^n(x)\varphi(x)\,dx - h\sum_{i=0}^{n-1} D_- D_+ u_i^n \varphi(x_i) \\
&= \int_0^1 u^n(x)(\varphi(x) - \varphi^n(x))\,dx + \int_0^1 u^n(x)\varphi^n(x)\,dx \\
&\quad - h\sum_{i=0}^{n-1} u_i^n \varphi_i^n + h\sum_{i=0}^{n-1} v_i \varphi_i^n,
\end{aligned} \tag{2.45}$$

where φ^n denotes the piecewise linear, continuous function that coincides with φ on x_i, $i = 0, \dots, n-1$.

The first integral in (2.45) tends to zero by the Lebesgue dominated convergence theorem. We use (2.31) to compute the second integral:

$$\int_0^1 u^n(x)\varphi^n(x)\,dx = \frac{2}{3}\langle u^n \varphi^n \rangle + \frac{1}{6}\langle S_+ u^n \varphi^n \rangle + \frac{1}{6}\langle u^n S^+ \varphi^n \rangle.$$

One can prove that this term tends to $\langle u^n \varphi^n \rangle$ (see the proof of the convergence of term A in the proof of Theorem 2.1). The last sum equals

$$\sum_{i=0}^{n-1} h v_i^n \varphi(x_i) = \left\langle u - u_{xx}, \sum_{i=0}^{n-1} \varphi_i^n \psi_i^n(x) \right\rangle.$$

For all $x \in [0,1]$, there exists a k such that $x \in [x_k, x_{k+1}]$. Then

$$\begin{aligned}
\left| \varphi(x) - \sum_{i=0}^{n-1} \varphi_i^n \psi_i^n(x) \right| &= \left| \sum_{i=0}^{n-1} (\varphi(x) - \varphi(x_i)) \psi_i^n(x) \right| \\
&\le |\varphi(x) - \varphi(x_k)| + |\varphi(x) - \varphi(x_{k+1})| \\
&\le 2 \sup_{|z-y| \le h} |\varphi(y) - \varphi(z)|
\end{aligned}$$

and therefore, by the uniform continuity of φ,

$$\sum_{i=0}^{n-1} \varphi(x_i)\psi_i^n(x) \to \varphi(x) \text{ in } L^\infty.$$

Thus,

$$\sum_{i=0}^{n-1} hv_i^n \varphi(x_i) = \left\langle u - u_{xx}, \sum_{i=0}^{n-1} \varphi(x_i)\psi_i^n \right\rangle \to \langle u - u_{xx}, \varphi \rangle$$

and, from (2.45), we get

$$\langle u^n - u_{xx}^n, \varphi \rangle \to \langle u - u_{xx}, \varphi \rangle .$$

As already explained, this implies that

$$u^n \to u \text{ in } H^1.$$

3. Numerical results. The numerical scheme (2.4) is semidiscrete: The time derivative has not been discretized, and hence we work with a system of ordinary differential equations. However, for numerical computations we integrate in time by using an explicit Euler method. Given a positive time T and $l \in \mathbb{N}$, we consider the time step $\Delta t = T/l$. We compute $m_j^{n,l}$, the approximate value of m^n at time $t_j = j\Delta t$, by taking

$$m_{j+1}^{n,l} = m_j^{n,l} + \Delta t \left(-D_-(m_j^{n,l} u_j^{n,l}) - m_j^{n,l} D u_j^{n,l} \right), \tag{3.1}$$

where

$$m_j^{n,l} = u_j^{n,l} - D_- D_+ u_j^{n,l}. \tag{3.2}$$

Here $m_j^{n,l} = (m_{0,j}^{n,l}, \dots, m_{n-1,j}^{n,l})$ and $u_j^{n,l} = (u_{0,j}^{n,l}, \dots, u_{n-1,j}^{n,l})$. Given $m_j^{n,l}$, one can still recompute $u_j^{n,l}$ using (2.8), that is,

$$u_{i,j}^{n,l} = L^{-1} m_{i,j}^{n,l} = \frac{c}{1 - e^{-\kappa n}} \sum_{k=0}^{n-1} (e^{-\kappa(i-k)} + e^{\kappa(i-k-n)}) m_{k,j}^{n,l}. \tag{3.3}$$

Lemma 2.2 does not apply in this setting, and the proof of convergence for the fully discrete scheme proceeds differently. Writing (2.4) as

$$m_t^n = f(m^n),$$

where $f\colon \mathbb{R}^n \to \mathbb{R}^n$, we observe (cf. (2.4) and (2.8)) that each component of $f(x)$ is a polynomial in the components $x_0, \dots, x_{n-1}$ of x. Hence, f is continuously differentiable. From (2.19) and (2.4), we obtain that when n is large enough, there exists a constant C which is independent of n such that

$$|m_i^n(t)| \le 5n^2 \max_i |u_i^n(t)| \le Cn^2$$

for all $t \in [0, T]$. Hence, $m^n(t)$ is bounded in $[0, T]$ and therefore the Euler method converges (see, for example, [19]), that is,

$$\lim_{l \to \infty} \max_{j=1,\dots,l} \left\| m_j^{n,l} - m^n(t_j) \right\| = 0. \tag{3.4}$$

All norms are equivalent in finite-dimensional vector spaces, and therefore (3.4) holds for any norm in $\mathbb{R}^n$. We denote by $m^{n,l}(t)$ the piecewise linear function in $C([0,T],\mathbb{R}^n)$ satisfying $m^{n,l}(t_j) = m_j^{n,l}$. It is given by

$$m^{n,l}(t) = \frac{1}{\Delta t}(t_{j+1} - t)m_j^{n,l} + \frac{1}{\Delta t}(t - t_j)m_{j+1}^{n,l}$$

for $t \in [t_j, t_{j+1}]$. Let us prove that

$$\lim_{l\to\infty} \left\|m^{n,l} - m^n\right\|_{C([0,T],\mathbb{R}^n)} = 0. \tag{3.5}$$

We have, for $t \in [t_j, t_{j+1}]$,

(3.6)

$$\begin{aligned} m^{n,l}(t) - m^n(t) = {} & \frac{1}{\Delta t}(t_{j+1} - t)(m_j^{n,l} - m^n(t_j)) + \frac{1}{\Delta t}(t - t_j)(m_{j+1}^{n,l} - m^n(t_{j+1})) \\ & + \frac{1}{\Delta t}(t_{j+1} - t)(m^n(t_j) - m^n(t)) + \frac{1}{\Delta t}(t - t_j)(m^n(t_{j+1}) - m^n(t)). \end{aligned}$$

Let $\varepsilon > 0$. Since $m^n \in C([0,T],\mathbb{R}^n)$, m^n is uniformly continuous, and there exists $\delta > 0$ such that $\|m^n(t_1) - m^n(t_2)\| < \varepsilon/2$ for all $t_1, t_2 \in [0,T]$ with $|t_2 - t_1| < \delta$. We can choose l large enough so that $\Delta t = T/l < \delta$. Then, for $t \in [t_j, t_{j+1}]$, we have $t - t_j < \delta$ and $t_{j+1} - t < \delta$, and

$$\begin{aligned} & \left\| \frac{1}{\Delta t}(t_{j+1} - t)(m^n(t_j) - m^n(t)) + \frac{1}{\Delta t}(t - t_j)(m^n(t_{j+1}) - m^n(t_{j+1})) \right\| \\ & \qquad < \frac{1}{\Delta t}(t_{j+1} - t)\frac{\varepsilon}{2} + \frac{1}{\Delta t}(t - t_j)\frac{\varepsilon}{2} \\ & \qquad < \frac{\varepsilon}{2}. \end{aligned} \tag{3.7}$$

By (3.4), we can choose l large enough so that $\max_{j=1,\dots,l} \|m_j^{n,l} - m^n(t_j)\| < \varepsilon/2$. Hence,

$$\left\| \frac{1}{\Delta t}(t_{j+1} - t)(m_j^{n,l} - m^n(t_j)) + \frac{1}{\Delta t}(t - t_j)(m_{j+1}^{n,l} - m^n(t_{j+1})) \right\| < \frac{\varepsilon}{2}. \tag{3.8}$$

Comparing (3.6), (3.7), and (3.8), we obtain

$$\left\|m^{n,l}(t) - m^n(t)\right\| < \varepsilon$$

for l large enough and any $t \in [0,T]$. Hence, (3.5) is proved. The mapping $L^{-1}\colon \mathbb{R}^n \to \mathbb{R}^n$, $L^{-1}m^n = u^n$, is continuous, and therefore we have $\lim_{l\to\infty} \left\|u^{n,l} - u^n\right\|_{C([0,T],\mathbb{R}^n)} = 0$. Finally, after using the identification of $\mathbb{R}^n$ with the set of continuous, periodic, piecewise linear functions, we get that

$$\lim_{l\to\infty} u^{n,l} = u^n,$$

and, from Theorem 2.1,

$$\lim_{n\to\infty} \lim_{l\to\infty} u^{n,l} = u$$

in $C([0,T],H^1)$. We summarize the result in the following theorem.

THEOREM 3.1. *Let $\Delta t = T/l$, and define the function $u_{i,j}^{n,l}$ by (3.1)–(3.3). Define the corresponding interpolating function $u^{n,l}$ in $C([0,T],H^1)$ by*

$$u^{n,l}(x,t) = \frac{n}{\Delta t}\Big((t_{j+1}-t)\big[(x_{i+1}-x)u_{i,j}^{n,l} + (x-x_i)u_{i+1,j}^{n,l}\big] + (t-t_j)\big[(x_{i+1}-x)u_{i,j+1}^{n,l} + (x-x_i)u_{i+1,j+1}^{n,l}\big]\Big)$$

for $x \in [x_i, x_{i+1}]$ and $t \in [t_j, t_{j+1}]$. Then

$$\lim_{n\to\infty}\lim_{l\to\infty} u^{n,l} = u \tag{3.9}$$

in $C([0,T],H^1)$, where u is the solution of the Camassa–Holm equation (1.2).

To compute the discrete spatial derivative, we need at each step to compute u from m. The function u is given by a discrete convolution product,

$$u_i = h\sum_{j=0}^{n-1} g_{i-j}^p m_j.$$

It is advantageous to apply the fast Fourier transform (FFT); see [13]. In the frequency space, a convolution product becomes a multiplication which is cheap to evaluate. Going back and forth to the frequency space is not very expensive due to the efficiency of the FFT. We use a formula of the form (see [13] for more details):

$$u = \mathcal{F}_N^{-1}(\mathcal{F}_N[g]\cdot\mathcal{F}_N[m]),$$

where $\mathcal{F}_N$ denotes the FFT.

We have tested algorithm (3.1) with single and double peakons. In the single peakon case, the initial condition is given by

$$u(x,0) = c\frac{\cosh(d-\frac{a}{2})}{\sinh\frac{a}{2}}, \tag{3.10}$$

which is the periodized version of $u(x,0) = ce^{-|x|}$. The period is denoted by a, and $d = \min(x, a-x)$ is the distance from x to the boundary of the interval $[0,a]$. The peakons travel at a speed equal to their height, that is,

$$u(x,t) = ce^{-|x-ct|}.$$

If u satisfies the initial condition $u(x,0) = e^{-|x|}$, then $m = 2\delta$ at $t = 0$ and we take

$$m_i(0) = \begin{cases} \frac{2}{h} & \text{if } i = 0, \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

as the initial discrete condition. The function m_i gives a discrete approximation of 2δ. Figure 1 shows the result of the computation for different refinements. Figure 2 indicates that the computed solution converges to the exact solution.

The sharp increase of the error $\|u(t) - u^n(t)\|_{H^1}$ at time $t = 0$ can be predicted by looking at (2.17), which gives a first-order approximation of the time derivative of $\|u(t)\|_{H^1}^2$:

$$\frac{dE_n(t)^2}{dt} = -\sum_{i=0}^{n-1} u_i(hm_i)^2 + \mathcal{O}(h).$$

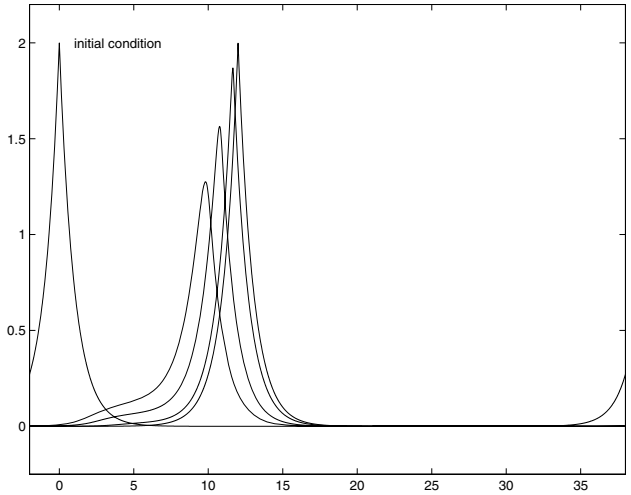


FIG. 1. *Periodic single peakon. The initial condition is given by* $u(x,0) = 2e^{-|x|}$ *and period* $a = 40$. *The computed solutions are shown at time* $t = 6$ *for* (*from left to right*) $n = 2^{10}$, $n = 2^{12}$, $n = 2^{14}$ *together with the exact solution* (*at the far right*).

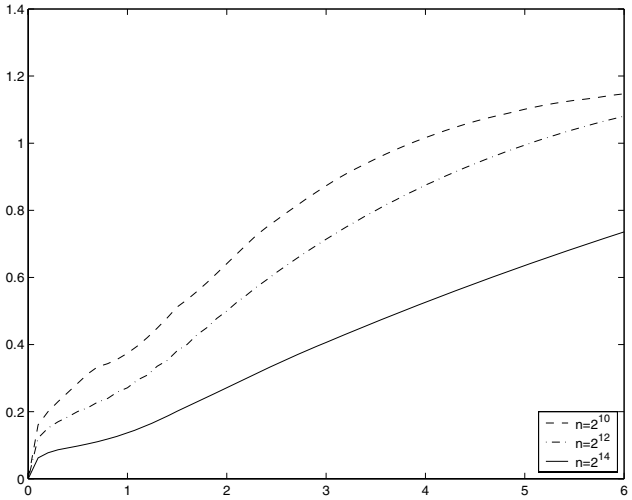


FIG. 2. *Plot of* $\|u(t) - u^n(t)\|_{H^1} / \|u(t)\|_{H^1}$ *in the one peakon case of Figure* 1.

Hence,

$$\frac{d\,\|u\|_{H^1}^2}{dt} \approx \frac{dE_n(t)^2}{dt} \approx -4 \quad \text{at } t = 0.$$

At the beginning of the computation, we can therefore expect a sharp decrease of the H^1-norm. To get convergence in H^1, it is therefore necessary that the solution becomes smooth enough so that $\frac{d\|u\|_{H^1}^2}{dt} \to 0$. In any case, we cannot hope for high accuracy and convergence rate in this case. Figures 3 and 4 show the same plots in the two-peakon case.

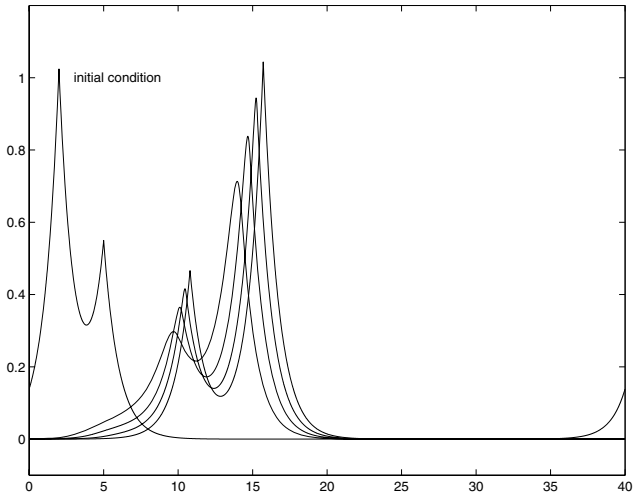


FIG. 3. *Two peakon case. The initial condition is the periodized version of* $2e^{-|x-2|} + e^{-|x-5|}$. *The computed solutions are shown at time* $t = 12$ *for* (*from left to right*) $n = 2^{10},\ n = 2^{12},\ n = 2^{14}$ *together with the exact solution* (*at the far right*).

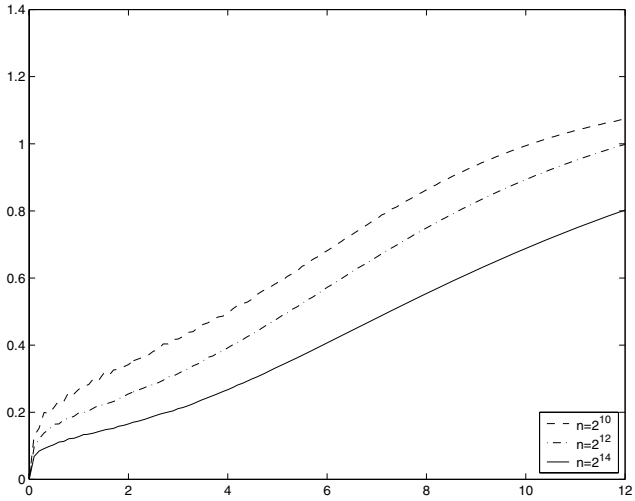


FIG. 4. *Plot of* $\|u(t) - u^n(t)\|_{H^1} / \|u(t)\|_{H^1}$ *in the two peakon case of Figure* 3.

We have tested our algorithm with smooth initial conditions. In this case, the H^1-norm remains constant in a much more accurate manner. The convergence is probably much better than for nonsmooth solutions, but we have no analytical solution with which to compare.

Other time integration methods (second-order Runge–Kutta method, variable-order Adams–Bashforth–Moulton) have also been tried and the results do not differ significantly from those given by (3.1). It follows that the Camassa–Holm equation is not very sensitive to the way time is discretized. But the situation is completely

different when we consider different space discretizations. The schemes

$$m_t = -D_-(mu)_i - m_i D_+ u_i, \tag{3.12}$$

$$m_t = -D(mu)_i - m_i D u_i, \tag{3.13}$$

$$m_t = -D_+(mu)_i - m_i D_- u_i \tag{3.14}$$

are all at first glance good candidates for solving the Camassa–Holm equation. They preserve the H^1-norm, are finite difference approximations of (2.2), and finally look very similar to (2.4). But, tested on a single peakon, (3.12) produces a peakon that grows, (3.13) produces oscillations, and (3.14) behaves in a completely unexpected manner (at the first time step, m becomes a negative Dirac function and starts traveling backward!).

Let us have a closer look at the scheme (3.12). We compute $\frac{dE_n^2}{dt}$:

$$\frac{1}{2}\frac{dE_n^2}{dt} = \sum_{i=0}^{n-1} m_{i,t}^n u_i^n = \sum_{i=0}^{n-1} \big(- D_-(m^n u^n)_i u_i - m_i^n D_+ u_i u_i \big) = 0.$$

Thus, E_n is exactly preserved. Lemma 2.2 still holds since the same proof applies to (3.12). It allows us to derive the bounds of Lemma 2.3 and, after applying Simon's theorem, we get the existence of a converging subsequence. The problem is that, in general, this subsequence *does not* converge to the solution of the Camassa–Holm equation. In order to see that, we compare how our original algorithm (3.12) and algorithm (3.13) handle a peakon solution $u = ce^{-|x-ct|}$. The only terms that differ are $m^n Du^n$ and $m^n D_+ u^n$. We have proved earlier that for any smooth function φ,

$$\sum_{i=0}^{n-1} m_i^n Du_i^n \varphi(x_i) \to \frac{1}{2}\int_0^1 (u^2 - u_x^2)\varphi(x)\,dx$$

as $n \to \infty$. In the peakon case, $u^2 = u_x^2$ and this term tends to zero. Roughly speaking, we can say that m^n converges to a Dirac function (see (3.11)), but at the same time it is multiplied by Du^n, which is the average of the left and right derivatives and which tends to zero at the top of the peak. Eventually the whole product $m^n Du^n$ tends to zero. We follow the same heuristic approach with the term $m^n D_+ u^n$ in (3.13). This time, m^n is multiplied by the right derivative $D_+ u^n$ of u^n, which tends, at the top of the peak, to $-c$. Hence, $-m^n D_+ u^n$ tends to $c\delta$ and not zero, as it would if (3.13) converged to the correct solution. This example shows how sensitive the numerical approximation is, regarding the explicit form of the finite difference scheme, for the Camassa–Holm equation.

Acknowledgments. H. H. acknowledges helpful discussions with Nils Henrik Risebro and Kenneth H. Karlsen on discretizations of the Camassa–Holm equation.

REFERENCES

[1] H. Brezis, *Analyse Fonctionnelle. Théorie et applications*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1983.

[2] R. Camassa, *Characteristics and the initial value problem of a completely integrable shallow water equation*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 115–139.

[3] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[4] R. Camassa, D. D. Holm, and J. Hyman, *A new integrable shallow water equation*, Adv. Appl. Mech., 31 (1994), pp. 1–33.

[5] R. Camassa, J. Huang, and L. Lee, *Integral and integrable algorithms for a nonlinear shallow-water wave equation*, J. Comput. Phys., 216 (2006), pp. 547–572.
[6] R. Camassa, J. Huang, and L. Lee, *On a completely integrable numerical scheme for a nonlinear shallow-water wave equation*, J. Nonlinear Math. Phys., 12 (2005), pp. 146–162.
[7] A. Constantin, *On the Cauchy problem for the periodic Camassa–Holm equation*, J. Differential Equations, 141 (1997), pp. 218–235.
[8] A. Constantin, *Existence of permanent and breaking waves for a shallow water equation: A geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.
[9] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 26 (1998), pp. 303–328.
[10] A. Constantin and J. Escher, *Well-posedness, global existence, and blowup phenomena for a periodic quasi-linear hyperbolic equation*, Comm. Pure Appl. Math., 51 (1998), pp. 475–504.
[11] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.
[12] H. R. Dullin, G. A. Gottwald, and D. D. Holm, *Camassa–Holm, Korteweg–de Vries-5 and other asymptotically equivalent equations for shallow water waves*, Fluid Dynam. Res., 33 (2003), pp. 73–95.
[13] C. Gasquet and P. Witomski, *Fourier Analysis and Applications. Filtering, Numerical Computation, Wavelets*, Texts Appl. Math. 30, Springer-Verlag, New York, 1999.
[14] A. A. Himonas and G. Misiołek, *The Cauchy problem for an integrable shallow-water equation*, Differential Integral Equations, 14 (2001), pp. 821–831.
[15] H. Holden and X. Raynaud, *A convergent numerical scheme for the Camassa–Holm equation based on multipeakons*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 505–523.
[16] R. S. Johnson, *Camassa–Holm, Korteweg–de Vries, and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[17] Y. A. Li and P. J. Olver, *Well-posedness and blow-up solutions for an integrable nonlinearly dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.
[18] G. Misiołek, *Classical solutions of the periodic Camassa–Holm equation*, Geom. Funct. Anal., 12 (2002), pp. 1080–1104.
[19] R. Plato, *Concise Numerical Mathematics*, Grad. Stud. Math. 57, AMS, Providence, RI, 2003.
[20] G. Rodríguez-Blanco, *On the Cauchy problem for the Camassa–Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.
[21] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1681–1702

POSTPROCESSING THE FINITE ELEMENT METHOD FOR SEMILINEAR PARABOLIC PROBLEMS*

YUBIN YAN†

Abstract. In this paper we consider postprocessing of the finite element method for semilinear parabolic problems. The postprocessing amounts to solving a linear elliptic problem on a finer grid (or higher-order space) once the time integration on the coarser mesh is completed. The convergence rate is increased at almost no additional computational cost. This procedure was introduced and analyzed in García-Archilla and Titi [*SIAM J. Numer. Anal.*, 37 (2000), pp. 470–499]. We extend the analysis to the fully discrete case and prove error estimates for both space and time discretization. The analysis is based on error estimates for the approximation of time derivatives by difference quotients.

Key words. time derivative, postprocessing, finite element method, backward Euler method, error estimates, semilinear parabolic problem

AMS subject classifications. 65M60, 65M15, 65M20

DOI. 10.1137/S0036142903430931

1. Introduction.

In this paper we shall consider postprocessing of the finite element method for the semilinear parabolic problem

$$u_t - \Delta u = F(u) \quad \text{in}\,\Omega \quad \text{for}\ t \in (0,T], \tag{1.1}$$

$$u = 0 \quad \text{on}\ \partial\Omega \quad \text{for}\ t \in (0,T], \quad \text{with}\, u(0) = v,$$

where Ω is a bounded domain in $\mathbf{R}^d$, $d = 1, 2, 3$, with a sufficiently smooth boundary $\partial\Omega$, $u_t = \partial u/\partial t$, Δ is the Laplacian, and $F : \mathbf{R} \to \mathbf{R}$ is a smooth function.

Let $H = L_2(\Omega)$. We define the unbounded operator $A = -\Delta$ on H with domain of definition $\mathcal{D}(A) = H^2 \cap H_0^1$, where, for integer $m \geq 1$, $H^m = H^m(\Omega)$ denotes the standard Sobolev space $W_2^m(\Omega)$, and $H_0^1 = H_0^1(\Omega) = \{v \in H^1 : v|_{\partial\Omega} = 0\}$. Then A is a closed, densely defined, and self-adjoint positive definite operator in H with compact inverse. The initial-boundary value problem (1.1) may then be formulated as the following initial value problem:

$$u_t + Au = F(u) \quad \text{for}\ 0 < t \leq T, \quad \text{with}\, u(0) = v, \tag{1.2}$$

in the Hilbert space H, where $F : H \to H$ is a nonlinear operator, and $v \in H$.

Recently, a postprocessing technique was introduced to increase the efficiency of the Galerkin method of spectral type; see Canuto et al. [5], de Frutos, García-Archilla, and Novo [7], and de Frutos and Novo [8], [10]. Postprocessed methods yield greater accuracy than standard Galerkin schemes at nearly the same computational cost. In García-Archilla and Titi [14], the postprocessing technique was extended to the h-version of the finite element method for dissipative partial differential equations. There, the authors prove that the postprocessed method has a higher rate of convergence than the standard finite element method when higher-order finite elements,

*Received by the editors June 29, 2003; accepted for publication (in revised form) March 23, 2006; published electronically September 15, 2006.
http://www.siam.org/journals/sinum/44-4/43093.html
†Department of Automatic Control and Systems Engineering, The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK (Y.Yan@sheffield.ac.uk).

rather than linear finite elements, are used. Error estimates in L_2 and H^1 norms in the spatially semidiscrete case are obtained. More recently, in de Frutos and Novo [9], the authors show that the postprocessing technique can also be applied to linear finite elements and the convergence rate can be improved in the H^1 norm, but not in the L_2 norm. The analysis is restricted to the spatially semidiscrete case.

The purpose of the present paper is to derive the error estimates in the fully discrete case for the postprocessed finite element method applied to (1.2). To do this, we introduce the time-stepping method to compute the discrete solution of (1.2) and define a difference quotient approximation to the time derivative. We then define the postprocessing step in the fully discrete case and show the error estimates for the postprocessing method by using the error estimates for time derivatives. For simplicity we consider error estimates only in the L_2 norm. Our technique of proof is related to, but different from, the one employed by García-Archilla and Titi [14]. Our technique is applicable to both semidiscrete and fully discrete cases. However, we should point out that García-Archilla and Titi [14] also treat a nonlinear convection term. It is not quite clear how it is possible to generalize our method to deal with a nonlinear convection term.

The paper is organized as follows. In section 2, we introduce some basic notation and lemmas. In section 3 we consider error estimates for the postprocessed finite element method in the semidiscrete case. In section 4, we consider error estimates in the fully discrete case. In section 5, we consider the starting approximation of time derivatives. Finally, in section 6, we consider higher-order time-stepping in the context of the linear homogeneous problem.

By C_0 we denote positive constant independent of the functions and parameters concerned, but not necessarily the same at different occurrences.

2. Preliminaries. Let $\mathcal{T}$ denote a partition of Ω into disjoint triangles τ such that no vertex of any triangle lies on the interior of a side of another triangle and such that the union of the triangles determines a polygonal domain $\Omega_h \subset \Omega$ with boundary vertices on $\partial\Omega$. Let h denote the maximal length of the sides of the triangulation $\mathcal{T}_h$. We assume that the triangulations are quasi-uniform in the sense that the triangles of $\mathcal{T}_h$ are of essentially the same size.

Let r be any nonnegative integer. We denote by $\|\cdot\|_r$ the norm in H^r. Let $\{S_h\} = \{S_{h,r}\} \subset H_0^1$ be a family of finite element spaces with accuracy of order $r \geq 2$, i.e., S_h consists of continuous functions on the closure $\bar{\Omega}$ of Ω which are polynomials of degree at most $r-1$ in each triangle of $\mathcal{T}_h$ and which vanish outside Ω_h, such that, for small h,

$$\inf_{\chi \in S_h} \{\|v-\chi\| + h\|\nabla(v-\chi)\|\} \leq Ch^s\|v\|_s \quad \text{for } 1 \leq s \leq r,$$

when $v \in H^s \cap H_0^1$.

The semidiscrete problem of (1.2) is to find the approximate solution $u_h(t) = u_h(\cdot,t) \in S_h$ for each t, such that

$$u_{h,t} + A_h u_h = P_h F(u_h), \quad \text{with } u_h(0) = v_h, \tag{2.1}$$

where $v_h \in S_h$, $P_h : L_2 \to S_h$ is the L_2 projection onto S_h, and $A_h : S_h \to S_h$ is the discrete analogue of A, defined by

$$(A_h\psi, \chi) = A(\psi, \chi) \quad \forall\, \psi, \chi \in S_h. \tag{2.2}$$

Here $A(\cdot,\cdot) = (\nabla\cdot, \nabla\cdot)$ is the bilinear form on H_0^1 obtained from A.

Error estimates for finite element methods for semilinear parabolic problems with various conditions on the nonlinearity have been considered in many papers; see, e.g., Akrivis, Crouzeix, and Makridakis [1], [2], Crouzeix, Thomée, and Wahlbin [6], Elliott and Larsson [11], [12], Helfrich [16], Johnson et al. [17], Thomée [27], Thomée and Wahlbin [28], and Wheeler [29]. The long time behavior of finite element solutions was studied by Elliott and Stuart [13], Larsson [18], [19], and Larsson and Sanz-Serna [20], [21].

Let us now describe the idea of the postprocessed finite element method proposed by García-Archilla and Titi [14]. Suppose that we want to obtain high-order approximation, for instance, $O(h^{r+2})$. Then we can use, in every time step, either a family of high-order finite element spaces $\tilde{S}_h := S_{h,r+2}$ with accuracy of order $r+2$, or a family of finite element space $\tilde{S}_h := S_{\tilde{h},r}$ with accuracy of order r, but with finer partition $\mathcal{T}_{\tilde{h}}$ of the domain Ω, such that $h^{r+2} = \tilde{h}^r$. In [14], another technique, called the *postprocessed finite element method*, is presented, which improves the convergence rate without using a high-order finite element space $\tilde{S}_h$ in every time step. Suppose that we are interested in the solution of (1.2) at a given time T. At time T, rewriting (1.2), we have

$$Au(T) = -u_t(T) + F(u(T)). \tag{2.3}$$

Thus, $u(T)$ can be seen as the solution of an elliptic problem whose right-hand side is not known but can be approximated. García-Archilla and Titi first compute $u_h(T)$ by (2.1) in the finite element space S_h, then replace $u_t(T)$ by $u_{h,t}(T)$ and solve (or, in practice, approximate) the following linear elliptic problem: find $\tilde{u}(T) \in \mathcal{D}(A)$, such that

$$A\tilde{u}(T) = -u_{h,t}(T) + F(u_h(T)), \tag{2.4}$$

which is the postprocessing step.

They obtained the following error estimate, with $\ell_h = 1 + \log(T/h^2)$:

$$\|\tilde{u}(T) - u(T)\| \le C(u)\ell_h h^{r+2} \quad \text{for } r \ge 4, \tag{2.5}$$

where $C(u)$ is some constant depending on u. A similar result holds for $r \ge 3$ with order $O(h^{r+1})$, but without the factor ℓ_h.

The proof is based on superconvergence for elliptic finite element methods in norms of negative order, which is the reason for the restriction $r \ge 3$.

We note that the bound (2.5) is an improvement over the error estimates for the standard Galerkin method, which is $O(h^r)$. In practice $\tilde{u}$ cannot be computed exactly, since in general it does not belong to a finite element space. However, one can approximate the solution $\tilde{u}$ of (2.4) by some $\tilde{u}_h$ belonging to a finite element space $\tilde{S}_h$ of approximation order $r+2$ as described above. More precisely, we pose the following semidiscrete problem corresponding to (2.4): find $\tilde{u}_h \in \tilde{S}_h$, such that

$$\tilde{A}_h\tilde{u}_h(T) = \tilde{P}_h\big(-u_{h,t}(T) + F(u_h(T))\big), \tag{2.6}$$

where $\tilde{P}_h : L_2 \to \tilde{S}_h$ is the L_2 projection onto $\tilde{S}_h$ and $\tilde{A}_h$ is the discrete analogue of A with respect to $\tilde{S}_h$. The standard error estimate reads (see, e.g., Brenner and Scott [4])

$$\|\tilde{u}_h(T) - \tilde{u}(T)\| \le C(u)h^{r+2}. \tag{2.7}$$

Combining (2.5) and (2.7), we have

$$\|\tilde{u}_h(T) - u(T)\| \le \|\tilde{u}_h(T) - \tilde{u}(T)\| + \|\tilde{u}(T) - u(T)\| \le C(u)\ell_h h^{r+2} \quad \text{for } r \ge 4.$$

Let us now introduce norms of negative order. Consider the stationary problem

$$Au = f. \tag{2.8}$$

The variational form of this problem is to find $u \in H_0^1 = H_0^1(\Omega)$, such that

$$A(u, \varphi) = (f, \varphi) \quad \forall\, \varphi \in H_0^1.$$

The standard Galerkin finite element problem is to find $u_h \in S_h$, such that

$$A(u_h, \chi) = (f, \chi) \quad \forall\, \chi \in S_h. \tag{2.9}$$

Let $G : L_2 \to H_0^1$ be the exact solution operator of (2.8) and define the approximate solution operator $G_h : L_2 \to S_h$ by $G_h f = u_h$ so that $u_h = G_h f \in S_h$ is the solution of (2.9). We recall that G_h is the self-adjoint, positive semidefinite on L_2 and positive definite on S_h. We note that $G : L_2 \to H_0^1 \cap H^2$ is the inverse operator of $A : H_0^1 \cap H^2 \to L_2$, i.e., $G = A^{-1}$, and similarly $G_h = A_h^{-1}$ on S_h, where A_h is the discrete Laplacian of A defined by (2.2). Moreover, we will use the following properties (see, Thomée [27, Chapter 2]):

$$G_h P_h = G_h \quad \text{and} \quad G_h = R_h G, \tag{2.10}$$

where $R_h : H_0^1 \to S_h$ is the elliptic projection, or Ritz projection, defined by

$$A(R_h u, \chi) = A(u, \chi) \quad \forall\, \chi \in S_h. \tag{2.11}$$

The negative order norm is defined by

$$|v|_{-s} = \|G^{s/2} v\| = (G^s v, v)^{1/2} \quad \text{for } s \ge 0.$$

We have (see Thomée [27, Chapter 6])

$$\|(G_h - G)f\| \le Ch^r \|f\|_{r-2} \quad \text{for } f \in H^{r-2},\ r \ge 2, \tag{2.12}$$

and

$$|(G_h - G)f|_{-2} \le Ch^{r+2} \|f\|_{r-2} \quad \text{for } f \in H^{r-2}, \quad r \ge 4. \tag{2.13}$$

We will use (2.12) and (2.13) in section 6 for the homogeneous parabolic problem by using a higher-order time-stepping method.

Remark 2.1. The estimate (2.13) and its like require higher elliptic regularity of the A operator. To see this, let us mention some related results here; see Schatz, Sloan, and Wahlbin [22]. We denote by $\|\cdot\|_r$ the standard Sobolev norm in H^r for $r > 0$.

Case 1. A Dirichlet problem in a plane polygonal domain. Consider the problem

$$-\Delta u = f \quad \text{in}\, \Omega, \qquad u = 0 \quad \text{on}\, \partial\Omega,$$

where Ω is a plane polygonal domain. It is well known how the solution behaves near the corners of the domain (see Grisvard [15]), and using suitable mesh refinements if necessary (see Babuška [3]) we shall assume that

$$\min_{\chi \in S_h} \|u - \chi\|_1 \le Ch^{r-1} \|u\|_r, \quad r \ge 2.$$

A standard duality argument gives

$$\|u_h - u\|_{-(r-2)} \le Ch^{2r-2}, \quad r \ge 2.$$

Case 2. A Dirichlet problem in smooth plane domain. Consider the problem

$$Au = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega,$$

where Ω is a smooth domain and A is an elliptic operator. As in Scott [25], assume that the boundary interpolation nodes satisfy some special conditions; if A is properly elliptic, then we can show

$$\|u_h - u\|_{-s} \le Ch^{r+s}\big(\|u\|_r + \|g\|_{r+s}\big), \quad s \ge 0.$$

Case 3. A homogeneous Dirichlet problem in smooth domain in $\mathbf{R}^d$. Consider the problem

$$Au = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

where Ω is a smooth domain and A is an elliptic operator. For our finite element spaces we take isoparametric elements which approximate the boundary to order h^r. For u smooth enough it was proved in Schatz and Wahlbin [24], with $\ell_h = \ln 1/h$, that

$$\|u_h - u\|_{W^1_\infty} \le C\ell_h h^{r-1}.$$

By the duality argument, we then get

$$\|u_h - u\|_{-(r-2)} \le C\ell_h^2 h^{2r-2}.$$

Case 4. A homogeneous Neumann problem in smooth domain in $\mathbf{R}^d$. Consider the problem

$$Au = f \quad \text{in } \Omega, \qquad \frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega,$$

where Ω is a smooth domain and A is an elliptic operator. Let $\partial\Omega$ be smooth and let the finite elements at $\partial\Omega$ be curved, exactly fitting $\partial\Omega$. It was proved in Scott [26] that, with $\ell_h = \ln 1/h$,

$$\|u_h - u\|_{W^1_\infty} \le C\ell_h h^{r-1}.$$

By the duality argument, we then get

$$\|u_h - u\|_{-(r-2)} \le C\ell_h^2 h^{2r-2}.$$

We also introduce a discrete negative order seminorm on L_2 by

$$|v|_{-s,h} = \|G_h^{s/2} v\| = (G_h^s v, v)^{1/2} \quad \text{for } s \ge 0;$$

it corresponds to the discrete semi-inner product $(v, w)_{-s,h} = (G_h^s v, w)\ \forall\ v, w \in L_2$. Since G_h is positive definite on S_h, $|v|_{-s,h}$ and $(v, w)_{-s,h}$ define a norm and an inner product there. We also find that the discrete negative-order seminorm is equivalent

to the corresponding continuous norm, modulo a small error. More precisely, we have the following bounds; see, e.g., Thomée [27, Lemma 6.3].

LEMMA 2.1. *We have, for* $0 \le s \le r$,

$$|v|_{-s,h} \le C_0(|v|_{-s} + h^s\|v\|) \quad \text{and} \quad |v|_{-s} \le C_0(|v|_{-s,h} + h^s\|v\|).$$

We also need Gronwall's lemma.

LEMMA 2.2. *If* a, b *are nonnegative constants and*

$$0 \le u(t) \le a + b\int_0^t u(s)\,ds \quad \text{for } 0 \le t \le T,$$

then we have

$$u(t) \le ae^{bt} \quad \text{for } 0 \le t \le T.$$

For the nonlinear operator F, we have the following bounds; see García-Archilla and Titi [14, Lemma 3].

LEMMA 2.3. *Let* $u \in H^r(\Omega) \cap H_0^1(\Omega)$, $r \ge 4$, *and* $\chi \in H_0^1(\Omega) \cap L^\infty(\Omega)$. *Assume that* F *is a smooth function. Further assume that* $d \le 3$ *and* $\|u-\chi\|_{L_\infty} \le K$ *for some positive number* K. *Then there is a constant* $C = C(\|u\|_r, K)$ *such that*

$$\|F(u) - F(\chi)\| \le C\|u - \chi\| \tag{2.14}$$

and

$$|F(u) - F(\chi)|_{-2} \le C\big(|u - \chi|_{-2} + \|u - \chi\|^2\big). \tag{2.15}$$

Remark 2.2. In our application of Lemma 2.3, we will choose u to be the solution of (1.2) and χ to be the corresponding finite element approximation solution u_h. It is obvious that u_h and u satisfy the assumptions of Lemma 2.3. For instance, $\|u_h - u\|_\infty \le K$ can be achieved by using the inverse inequality, provided we know that the L_2 error estimate for $u_h - u$ is $O(h^r)$; see Thomée [27, Chapter 14].

3. Semidiscrete approximation. In this section we will consider the error estimates for the *postprocessed finite element method* for the semilinear parabolic problem (1.2) in the semidiscrete case. The main theorem in this section is the following.

THEOREM 3.1. *Let* $r \ge 4$ *and let* S_h *and* $\tilde{S}_h$ *be the finite element spaces of orders* r *and* $r + 2$, *respectively, as described in section* 2. *Let* $\tilde{u}_h$ *and* u *be the solutions of* (2.6) *and* (1.2), *respectively. Assume that* F *satisfies* $\|F(u)\|_r \le C_0$ *in addition to the assumptions in Lemma* 2.3. *Let* u_h *be the solution of* (2.1). *Assume that* $v_h = R_h v, v \in H_0^1$, *and*

$$\sup_{s\in[0,T]} \|u_h(s) - u(s)\|_{L_\infty} \le K$$

and

$$\sup_{s\in[0,T]} \big(\|u(s)\|_r + \|u_t(s)\|_r + \|u_{tt}(s)\|_r\big) \le M \tag{3.1}$$

for some positive numbers K, M, T. *Then there is a constant* $C = C(K, M, T)$ *such that, with* $\ell_h = 1 + \log(T/h^2)$,

$$\|\tilde{u}_h(T) - u(T)\| \le C\ell_h h^{r+2}. \tag{3.2}$$

To prove Theorem 3.1, it suffices to show the bounds of $|u_h - u|_{-l}$ and $|u_{h,t} - u_t|_{-l}$ for $l = 0, 2$. We first split

$$u_h - u = (u_h - \hat{u}_h) + (\hat{u}_h - u) = \eta + e, \tag{3.3}$$

where $\hat{u}_h$ satisfies

$$\hat{u}_{h,t} + A_h \hat{u}_h = P_h F(u), \quad \hat{u}_h(0) = v_h. \tag{3.4}$$

Since u satisfies

$$u_t + Au = F(u), \quad u(0) = v, \tag{3.5}$$

the desired bounds of $e = \hat{u}_h - u$ and e_t follow from the error estimates for the linear parabolic problem because the right-hand side of (3.4) is independent of $\hat{u}_h$. In other words we only need to consider the nonlinear term F when we show the bounds of $\eta = u_h - \hat{u}_h$ and η_t. Note that η satisfies

$$\eta_t + A_h \eta = P_h(F(u_h) - F(u)), \quad \eta(0) = 0. \tag{3.6}$$

By Duhamel's principle, we have

$$\eta(t) = \int_0^t E_h(t-s) P_h \big(F(u_h(s)) - F(u(s))\big)\, ds. \tag{3.7}$$

Our main task is to consider the bounds for $|\eta|_{-l}$, $|\eta_t|_{-l}$, $l = 0, 2$.

We remark that since $\eta(0) = 0$, we don't need to consider the term $E_h(T)\eta(0)$ in (3.7). This observation is very useful in the fully discrete case.

Our first lemma in this section is the error estimate for the solution of (1.2).

LEMMA 3.2. *Let u_h and u be the solutions of* (2.1) *and* (1.2), *respectively. Assume that F satisfies the assumptions in Lemma* 2.3. *Further assume that $v_h = R_h v$ and*

$$\sup_{0 \le s \le T} \|u_h(s) - u(s)\|_{L_\infty} \le K \tag{3.8}$$

and

$$\sup_{0 \le s \le T} \big(\|u(s)\|_r + \|u_t(s)\|_r\big) \le M_1 \tag{3.9}$$

for some positive numbers K, M_1, T. Then there is a constant $C = C(K, M_1, T)$ such that

$$\sup_{0 \le t \le T} \|u_h(t) - u(t)\| \le Ch^r \quad \textit{for } r \ge 2 \tag{3.10}$$

and

$$\sup_{0 \le t \le T} |u_h(t) - u(t)|_{-2} \le Ch^{r+2} \quad \textit{for } r \ge 4. \tag{3.11}$$

The proof of Lemma 3.2 is similar to the proof of Lemma 3.3, so we omit it here.

Our next lemma is the error estimates for the time derivative of the solution of (1.2).

LEMMA 3.3. *Let u_h and u be the solutions of (2.1) and (1.2), respectively. Assume that F satisfies the assumptions in Lemma 2.3. Further assume that $v_h = R_h v$, $v \in H_0^1$, and*

$$\sup_{0\le s\le T} \|u_h(s) - u(s)\|_{L_\infty} \le K \tag{3.12}$$

and

$$\sup_{0\le s\le T} \big(\|u(s)\|_r + \|u_t(s)\|_r + \|u_{tt}(s)\|_r\big) \le M_2 \tag{3.13}$$

for some positive numbers K, M_2, T. Then there is a constant $C = C(K, M_2, T)$ such that, with $\ell_h = 1 + \log(T/h^2)$,

$$\sup_{0\le t\le T} \|u_{h,t}(t) - u_t(t)\| \le C\ell_h h^r, \quad r \ge 2, \tag{3.14}$$

and

$$\sup_{0\le t\le T} |u_{h,t}(t) - u_t(t)|_{-2} \le C\ell_h h^{r+2}, \quad r \ge 4. \tag{3.15}$$

Proof. We write

$$u_{h,t} - u_t = (u_{h,t} - \hat{u}_{h,t}) + (\hat{u}_{h,t} - u_t) = \eta_t + e_t.$$

Following the proofs of Theorems 1.3 and 6.2 in Thomée [27] for the error estimate $|e|_{-l}$, $l = 0, 2$, we can show the following error estimates for $|e_t|_{-l}$, $l = 0, 2$:

$$\|e_t(t)\| \le \|\hat{u}_{h,t}(0) - u_t(0)\| + C_0 h^r \Big(\|u_t(0)\|_r + \int_0^t \|u_{tt}\|_r \, ds\Big)$$

and

$$|e_t(t)|_{-2} \le |\hat{u}_{h,t}(0) - u_t(0)|_{-2} + C_0 h^{r+2} \Big(\|u_t(0)\|_r + \int_0^t \|u_{tt}\|_r \, ds\Big).$$

We observe that, by (3.4), and by noting that $\hat{u}_h(0) = R_h u(0)$,

$$\begin{aligned} \hat{u}_{h,t}(0) &= -A_h \hat{u}_h(0) + P_h F(u(0)) = -A_h R_h u(0) + P_h F(u(0)) \\ &= P_h\big(Au(0) + F(u(0))\big) = P_h u_t(0). \end{aligned}$$

We therefore have, by the error bounds for the L_2 projection,

$$\|\hat{u}_{h,t}(0) - u_t(0)\| = \|(P_h - I)u_t(0)\| \le C_0 h^r \|u_t(0)\|_r$$

and

$$|\hat{u}_{h,t}(0) - u_t(0)|_{-2} \le C_0 h^{r+2} \|u_t(0)\|_r.$$

Thus, we get

$$\|e_t(t)\| \le C_0 h^r \Big(\|u_t(0)\|_r + \int_0^t \|u_{tt}\|_r \, ds\Big) \le C(M_2, T) h^r \tag{3.16}$$

and, similarly,

$$|e_t(t)|_{-2} \le C(M_2,T)h^{r+2}. \tag{3.17}$$

We now turn to $|\eta_t|_{-l},\ l=0,2$. Using the fact $\|A_hE_h(t)\| \le C_0(t+h^2)^{-1}$ (see Schatz, Thomée, and Wahlbin [23]), we have

$$\int_0^t \|A_hE_h(t-s)\|\,ds \le C_0(1+\log(T/h^2)) \le C_0\ell_h. \tag{3.18}$$

By (3.7), we have

$$\begin{aligned}\eta_t(t) &= P_h(F(u_h(t))-F(u(t))) \\ &\quad - \int_0^t A_hE_h(t-s)P_h\big(F(u_h(s))-F(u(s))\big)\,ds.\end{aligned} \tag{3.19}$$

Thus, by (3.10), (3.18), and Lemma 2.3,

$$\begin{aligned}\|\eta_t(t)\| &\le \|P_h(F(u_h(t))-F(u(t)))\| \\ &\quad + \int_0^t \big\|A_hE_h(t-s)P_h\big(F(u_h(s))-F(u(s))\big)\big\|\,ds \\ &\le C(K,M_2,T)(1+\ell_h)\sup_{0\le s\le T}\|u_h(s)-u(s)\| \le C(K,M_2,T)\ell_h h^r.\end{aligned} \tag{3.20}$$

For $|\eta_t(t)|_{-2}$, we have, by (3.19),

$$\begin{aligned}|\eta_t(t)|_{-2} &\le |P_h(F(u_h(t))-F(u(t)))|_{-2} \\ &\quad + \int_0^t \big|A_hE_h(t-s)P_h\big(F(u_h(t))-F(u(t))\big)\big|_{-2}\,ds.\end{aligned}$$

Here, by Lemmas 2.1 and 2.3 and by (3.10), (3.11),

$$\begin{aligned}|P_h(F(u_h)-F(u))|_{-2} &\le C_0\big(h^2\|P_h(F(u_h)-F(u))\| + \|G_hP_h(F(u_h)-F(u))\|\big) \\ &\le C(\|u\|_r,K)\big(h^2\|u_h-u\| + \|u_h-u\|^2 + |u_h-u|_{-2}\big) \\ &\le C(K,M_2,T)h^{r+2}.\end{aligned}$$

Thus, by (3.18),

$$|\eta_t(t)|_{-2} \le C(K,M_2,T)\ell_h h^{r+2}.$$

Together these estimates complete the proof. ∎

Proof of Theorem 3.1. Combining (2.3) and (2.4), we have, with $\tilde{G}_h = \tilde{A}_h^{-1}$,

$$\begin{aligned}\tilde{u}_h(T)-u(T) &= \tilde{G}_h\tilde{P}_h(-u_{h,t}+F(u_h)) - G(-u_t+F(u)) \\ &= (\tilde{G}_h\tilde{P}_h - G)\big(-u_{h,t}+F(u_h)+u_t-F(u)\big) \\ &\quad - (\tilde{G}_h\tilde{P}_h - G)(u_t-F(u)) \\ &\quad + G(-u_{h,t}+F(u_h)+u_t-F(u)).\end{aligned}$$

Thus, by Lemmas 2.3, 3.2, and 3.3, we get, noting that $\|(\tilde{G}_h\tilde{P}_h - G)f\| \le Ch^s\|f\|_{s-2}$ for $0 \le s \le r+2$,

$$\begin{aligned}\|\tilde{u}_h(T) - u(T)\| &\le C_0h^2\big(\|u_{h,t} - u_t\| + \|F(u_h) - F(u)\|\big)\\ &\quad + C_0h^{r+2}\big(\|u_t\|_r + \|F(u)\|_r\big)\\ &\quad + |u_{h,t} - u_t|_{-2} + |F(u_h) - F(u)|_{-2}\\ &\le C(K, M, T)\ell_h h^{r+2}.\end{aligned}$$

The proof is complete. ∎

Remark 3.1. We remark that the "closeness assumption" of Theorem 3.1, i.e., $\sup_{s\in[0,T]}\|u_h(s) - u(s)\|_{L_\infty} \le K$, as well as similar assumptions in Lemmas 3.2 and 3.3 (inequalities (3.8), (3.12), respectively), require some restrictions on the semilinearity. For example, as stated in Thomée [27, Chapter 14, pp. 224], $F'(u)$ needs to grow "mildly" in order to obtain the "closeness assumption," without using inverse inequalities. For $d = 3$, the semilinearity assumption reads $|F'(u)| \le C(1+|u|^p)$, $p \le 2$, while for $d = 2$, p can be chosen arbitrarily.

Otherwise, we may need the inverse inequalities. However, as indicated in Thomée [27, Chapter 14, pp. 224] such properties are valid for $r > d/2$ for quasi-uniform partitions in the d-dimensional case. Similar observations are also needed for the fully discrete case.

4. Completely discrete approximation. In this section we will consider the postprocessed finite element method for (1.2) in the fully discrete case.

We use the similar technique developed in section 3 to derive the error estimates in the fully discrete case. Let $t_n = nk$, k being the time step. We will use the notation $u^j = u(t_j)$ and $u_t^j = u_t(t_j)$ below. We define the following backward Euler method, with $\bar{\partial}U^n = (U^n - U^{n-1})/k$:

$$\bar{\partial}U^n + A_hU^n = P_hF(U^n), \quad n \ge 1, \quad \text{with } U^0 = v_h. \tag{4.1}$$

It is natural to approximate $u_{h,t}(T)$, $T = t_n$, in (2.4) by $\bar{\partial}U^n$ for fixed n. The postprocessing step in the fully discrete case is to find $\tilde{u}(T) \in \mathcal{D}(A)$, such that

$$A\tilde{u}(T) = -\bar{\partial}U^n + F(U^n). \tag{4.2}$$

The semidiscrete problem of (4.2) is to find $\tilde{u}_h(T) \in \tilde{S}_h$, such that

$$\tilde{A}_h\tilde{u}_h(T) = \tilde{P}_h(-\bar{\partial}U^n + F(U^n)). \tag{4.3}$$

Let $\hat{U}^n$ be the solution of

$$\bar{\partial}\hat{U}^n + A_h\hat{U}^n = P_hF(u^n), \quad n \ge 1, \quad \text{with } \hat{U}^0 = v_h. \tag{4.4}$$

We have the following theorem.

THEOREM 4.1. *Let $r \ge 4$ and let S_h and $\tilde{S}_h$ be the finite element spaces of orders r and $r+2$, respectively, as described in section 2. Let $\tilde{u}_h$ and u be the solutions of (4.3) and (1.2), respectively. Assume that F satisfies $\|F(u^n)\|_r \le C_0$ in addition to the assumptions in Lemma 2.3. Let $T = t_n$ be a fixed time. Let U^n be the solution of (4.1). Assume that $v_h = R_hv, v \in H_0^1$, and*

$$\sup_{0\le t_n\le T}\|U^n - u(t_n)\|_{L_\infty} \le K$$

and

$$\sup_{0\le s\le T}\big(\|u(s)\|_r+\|u_t(s)\|_r+\|u_{tt}(s)\|+|u_{tt}(s)|_{-2}+\|Au_{tt}(s)\|\big)\le M \tag{4.5}$$

for some positive numbers K, M, T. Then there is a constant $C = C(K, M, T)$ such that, with $\ell_k = 1 + \log(T/k)$, k being the time step,

$$\|\tilde u_h(T)-u(T)\|\le C_0\big(\|\bar\partial\hat U^1-u_t(t_1)\|+|\bar\partial\hat U^1-u_t(t_1)|_{-2}\big)+C\ell_k(h^{r+2}+k).$$

We now state a lemma for the error estimate of the approximation U^n of $u(t_n)$ in the L_2 norm.

LEMMA 4.2. *Let U^n and u be the solutions of* (4.1) *and* (1.2)*, respectively. Assume that F satisfies the assumptions in Lemma* 2.3*. Further assume that $v_h = R_h v$, and*

$$\sup_{0\le t_n\le T}\|U^n-u(t_n)\|_{L_\infty}\le K \tag{4.6}$$

and

$$\sup_{0\le s\le T}\big(\|u(s)\|_r+\|u_t(s)\|_r+\|u_{tt}(s)\|+|u_{tt}(s)|_{-2}\big)\le M_3 \tag{4.7}$$

for some positive numbers K, M_3, T. Then there is a constant $C = C(K, M_3, T)$ such that

$$\sup_{0\le t_n\le T}\|U^n-u(t_n)\|\le C(h^r+k),\quad r\ge 2, \tag{4.8}$$

and

$$\sup_{0\le t_n\le T}|U^n-u(t_n)|_{-2}\le C(h^{r+2}+k),\quad r\ge 4. \tag{4.9}$$

Proof. We split

$$U^n-u(t_n)=(U^n-\hat U^n)-(\hat U^n-u(t_n))=\eta^n+e^n,$$

where $\hat U^n$ is defined by (4.4).

For $e^n=\hat U^n-u(t_n)$, we have, by the standard error estimates for linear parabolic problems (see, e.g., Thomée [27, Theorem 1.5]),

$$\begin{aligned}\|e^n\|&\le C_0\|R_hv-v\|+C_0h^r\Big(\|v\|_r+\int_0^{t_n}\|u_t\|_r\,ds\Big)+C_0k\int_0^{t_n}\|u_{tt}(s)\|\,ds\\&\le C(M_3,T)(h^r+k).\end{aligned} \tag{4.10}$$

For $\eta^n=U^n-\hat U^n$, noting that, by (4.4) and (4.1),

$$\begin{cases}\bar\partial\eta^n+A_h\eta^n=P_h(F(U^n)-F(u^n))\quad\text{for } n\ge 1,\\ \eta^0=0,\end{cases} \tag{4.11}$$

we have, by Lemma 2.3, with $r(\lambda)=1/(1+\lambda)$,

$$\begin{aligned}\|\eta^n\|&\le k\sum_{j=1}^n\|r(kA_h)^{n-j+1}\|\,\|P_h(F(U^j)-F(u^j))\|\\&\le C_0k\sum_{j=1}^n\|F(U^j)-F(u^j)\|\le C(K,M_3)\Big(k\sum_{j=1}^n\|\eta^j\|+k\sum_{j=1}^n\|e^j\|\Big).\end{aligned}$$

Further, by the discrete Gronwall lemma and (4.10), we have

$$\|\eta^n\| \le C(K, M_3, T)(h^r + k),$$

which shows (4.8).

Now we turn to (4.9). Following the proof of (4.10), we can show that

$$\begin{aligned} |e^n|_{-2} &\le C_0|R_h v - v|_{-2} + C_0 h^{r+2}\Big(\|v\|_r + \int_0^t \|u_t\|_r\, ds\Big) \\ &\quad + C_0 k \int_0^{t_n} |u_{tt}(s)|_{-2}\, ds \\ &\le C(M_3, T)(h^{r+2} + k). \end{aligned} \tag{4.12}$$

To estimate $|\eta^n|_{-2}$, we first note that, by Lemma 2.1,

$$|\eta^n|_{-2} \le C_0(h^2\|\eta^n\| + \|G_h\eta^n\|). \tag{4.13}$$

Here, by (4.11), $G_h\eta^n$ satisfies

$$\begin{cases} \bar\partial(G_h\eta^n) + A_h(G_h\eta^n) = G_h P_h(F(U^n) - F(u^n)) \quad \text{for } n \ge 1, \\ \eta^0 = 0, \end{cases} \tag{4.14}$$

which implies

$$G_h\eta^n = k\sum_{j=1}^{n} r(kA_h)^{n-j+1} G_h P_h(F(U^j) - F(u^j)).$$

Note that, by Lemmas 2.1 and 2.3,

$$\begin{aligned} \|G_h P_h(F(U^j) - F(u^j))\| &= |F(U^j) - F(u^j)|_{-2,h} \\ &\le C(\|u\|_r, K)\big(h^2\|U^j - u^j\| + \|U^j - u^j\|^2 + |U^j - u^j|_{-2}\big). \end{aligned}$$

Hence, by the stability of $r(\lambda)$,

$$\begin{aligned} \|G_h\eta^n\| \le C(K, M_3)\Big(& k\sum_{j=1}^{n} |\eta^j|_{-2} + h^2 k\sum_{j=1}^{n} \|U^j - u^j\| \\ & + k\sum_{j=1}^{n}(\|U^j - u^j\|^2 + |e^j|_{-2})\Big). \end{aligned}$$

Combining this with (4.13) and using the discrete Gronwall lemma, we get, by (4.8) and (4.12),

$$|\eta^n|_{-2} \le C(K, M_3, T)(h^{r+2} + k). \tag{4.15}$$

Together these estimates complete the proof. ∎

We also need the following lemma for the error estimate of the approximation $\bar\partial U^n$ of $u_t(t_n)$.

LEMMA 4.3. *Let U^n and u be the solutions of* (4.1) *and* (1.2)*, respectively. Assume that F satisfies the assumptions in Lemma* 2.3*. Further assume that $v_h = R_h v, v \in H_0^1$, and*

$$\sup_{0\le t_n\le T} \|U^n - u(t_n)\|_{L_\infty} \le K \tag{4.16}$$

and

$$\sup_{0\le s\le T}\big(\|u(s)\|_r+\|u_t(s)\|_r+\|u_{tt}(s)\|_r+\|u_{tt}(s)\|+\|Au_{tt}(s)\|\big)\le M_4 \tag{4.17}$$

for some positive numbers K, M_4, T. Then there is a constant $C=C(K,M_4,T)$ such that, with $\ell_k=1+\log(T/k)$, k being the time step,

$$\sup_{k\le t_n\le T}\|\bar\partial U^n-u_t(t_n)\|\le C_0\|\bar\partial\hat U^1-u_t(t_1)\|+C\ell_k(h^r+k) \tag{4.18}$$

and

$$\sup_{k\le t_n\le T}|\bar\partial U^n-u_t(t_n)|_{-2}\le C_0|\bar\partial\hat U^1-u_t(t_1)|_{-2}+C\ell_k(h^{r+2}+k). \tag{4.19}$$

Proof. We use the same notation as in Lemma 4.2 and write

$$\begin{aligned}\bar\partial U^n-u_t(t_n)&=(\bar\partial U^n-\bar\partial\hat U^n)+(\bar\partial\hat U^n-u_t(t_n))\\&=\bar\partial\eta^n+\big(\bar\partial\hat U^n-u_t(t_n)\big).\end{aligned}$$

We first show

$$\begin{aligned}\|\bar\partial\hat U^n-u_t(t_n)\|&\le C_0\|\bar\partial\hat U^1-u_t(t_1)\|+C_0h^r\Big(\|u_t(0)\|_r+\int_0^{t_n}\|u_{tt}\|_r\,ds\Big)\\&\quad+C_0k\int_0^{t_n}\|Au_{tt}(s)\|\,ds\\&\le C_0\|\bar\partial\hat U^1-u_t(t_1)\|+C(M_4,T)(h^r+k).\end{aligned} \tag{4.20}$$

To show (4.20), we write

$$\bar\partial\hat U^n-u_t(t_n)=(\bar\partial\hat U^n-R_hu_t(t_n))+(R_hu_t(t_n)-u_t(t_n))=\theta^n+\rho^n.$$

In the standard way ρ^n is bounded as desired, and it remains to consider $\theta^n\in S_h$. We have

$$\bar\partial\theta^n+A_h\theta^n=P_h\omega^n\quad\text{for } n\ge 2,$$

where

$$\omega^n=(R_h-I)\bar\partial u_t(t_n)+A(\bar\partial u^n-u_t^n)=\sigma^n+\tau^n.$$

By the stability estimate (see, e.g., Thomée [27, Theorem 10.2]),

$$\|\theta^n\|\le C_0\|\theta^1\|+C_0k\sum_{j=2}^n\|\sigma^j\|+C_0k\sum_{j=2}^n\|\tau^j\|\quad\text{for } n\ge 2. \tag{4.21}$$

We have

$$k\|\sigma^n\|\le C_0h^r\int_{t_{n-1}}^{t_n}\|u_{tt}\|_r\,ds$$

and

$$k\|\tau^n\|\le C_0k\|A(\bar\partial u^n-u_t^n)\|\le C_0k\int_{t_{n-1}}^{t_n}\|Au_{tt}(s)\|\,ds.$$

Together with $\|\theta^1\| \le \|\bar\partial U^1 - u_t^1\| + \|\rho^1\|$, with the obvious bounds for $\|\rho^1\|$, this completes the proof of (4.20).

For $\|\bar\partial\eta^n\|$, we have, by (4.11),

$$\bar\partial\eta^n = P_h(F(U^n) - F(u^n)) - k\sum_{j=1}^{n} A_h r(kA_h)^{n-j+1} P_h(F(U^n) - F(u^n)). \tag{4.22}$$

Using the smoothing property

$$k\sum_{j=1}^{n} \|A_h r(kA_h)^{n-j+1}\| \le C_0\ell_k, \tag{4.23}$$

which follows from

$$\begin{aligned} k\sum_{j=1}^{n} \|A_h r(kA_h)^{n-j+1}\| &\le C_0 k\sum_{j=1}^{n} t_{n-j+1}^{-1} = C_0\Big(1 + k\sum_{j=1}^{n-1} t_{n-j+1}^{-1}\Big) \\ &\le C_0\Big(1 + \int_{t_1}^{t_n} \frac{1}{s}\,ds\Big) \le C_0(1 + \log(t_n/k)) \le C_0\ell_k, \end{aligned}$$

we have, by Lemma 2.3 and (4.8),

$$\begin{aligned} \|\bar\partial\eta^n\| &\le C(K, M_4)\big(\|U^n - u^n\| + \ell_k \max_{1\le j\le n} \|U^j - u^j\|\big) \\ &\le C(K, M_4, T)\ell_k(h^r + k). \end{aligned} \tag{4.24}$$

Together these estimates complete the proof of (4.18).

Now we turn to estimate (4.19). Following the proof of (4.20), we can show

$$\begin{aligned} |\bar\partial\hat U^n - u_t(t_n)|_{-2} &\le C_0|\bar\partial\hat U^1 - u_t(t_1)|_{-2} + C_0h^{r+2}\Big(\|u_t(0)\|_r + \int_0^{t_n} \|u_{tt}\|_r\,ds\Big) \\ &\quad + C_0k\int_0^{t_n} \|u_{tt}(s)\|\,ds, \\ &\le C_0|\bar\partial\hat U^1 - u_t(t_1)|_{-2} + C(M_4, T)(h^{r+2} + k). \end{aligned} \tag{4.25}$$

For $|\bar\partial\eta^n|_{-2}$, we have, using (4.22), and by Lemmas 2.1 and 2.3,

$$|\bar\partial\eta^n|_{-2} \le C(K, M_4, T)\ell_k \max_{1\le j\le n} \big(h^2\|U^j - u^j\| + \|U^j - u^j\|^2 + |U^j - u^j|_{-2}\big).$$

Thus, by (4.8) and (4.9),

$$|\bar\partial\eta^n|_{-2} \le C(K, M_4, T)\ell_k(h^{r+2} + k). \tag{4.26}$$

Together these estimates complete the proof. □

Proof of Theorem 4.1. Combining (2.3) and (4.3), we have, with $\tilde G_h = \tilde A_h^{-1}$,

$$\begin{aligned} \tilde u_h(T) - u(T) &= \tilde G_h\tilde P_h(-\bar\partial U^n + F(U^n)) - G(-u_t(t_n) + F(u^n)) \\ &= (\tilde G_h\tilde P_h - G)\big(-\bar\partial U^n + F(U^n) + u_t(t_n) - F(u^n)\big) \\ &\quad - (\tilde G_h\tilde P_h - G)\big(u_t(t_n) - F(u^n)\big) \\ &\quad + G\big(-\bar\partial U^n + F(U^n) + u_t(t_n) - F(u^n)\big). \end{aligned}$$

Thus, noting that $\|(\tilde{G}_h\tilde{P}_h - G)f\| \le Ch^s\|f\|_{s-2}$ for $0 \le s \le r+2$, we get

$$\begin{aligned}\|\tilde{u}_h(T) - u(T)\| \le{}& C_0h^2\big(\|\bar{\partial}U^n - u_t(t_n)\| + \|F(U^n) - F(u^n)\|\big)\\ &+ C_0h^{r+2}\|u_t(t_n) - F(u^n)\|_r\\ &+ |\bar{\partial}U^n - u_t(t_n)|_{-2} + |F(U^n) - F(u^n)|_{-2}.\end{aligned}$$

Combining this with Lemmas 2.3, 4.2, and 4.3, we complete the proof. ∎

Remark 4.1. The algorithm works also for the Crank–Nicolson method. We can easily extend the proof of the backward Euler method to the Crank–Nicolson method.

The algorithm works also for the backward Euler method with variable time step if F is independent of solution u. It is not clear how to get the postprocessing error estimates if F depends on u. We will study this in future work.

5. Error estimate for the starting approximation. In this section we will consider the error estimate for the starting approximation of the time derivative $|\bar{\partial}\hat{U}^1 - u_t(t_1)|_{-s},\ s = 0, 2$, which appears in Theorem 4.1, where u and $\hat{U}^1$ satisfy

$$u_t + Au = F(u), \quad \text{with } u(0) = v, \tag{5.1}$$

and

$$\bar{\partial}\hat{U}^1 + A_h\hat{U}^1 = P_hF(u^1), \quad \text{with } \hat{U}^0 = v_h = R_hv, \tag{5.2}$$

respectively.

The semidiscrete problem of (5.1) is to find $\hat{u}_h \in S_h$ such that

$$\hat{u}_{h,t} + A_h\hat{u}_h = P_hF(u), \quad \text{with } \hat{u}_h(0) = R_hv. \tag{5.3}$$

We observe that we use $F(u^1)$ in (5.2); thus $|\bar{\partial}\hat{U}^1 - u_t(t_1)|_{-s},\ s = 0, 2$, can be bounded by the standard technique for nonhomogeneous linear parabolic problems. We have the following theorem.

THEOREM 5.1. *Let $\hat{U}^1$ and u be the solutions of* (5.2) *and* (5.1)*, respectively. Assume that F is continuously differentiable and*

$$\|Au_t(0)\| + \|u_t(0)\|_r + \max_{0\le\tau\le k}\big(\|F'(u(\tau))u_t(\tau)\| + \|u_{tt}(\tau)\|_r\big) \le M_0$$

for some positive number M_0. Then there is a constant $C = C(M_0)$ such that

$$\|\bar{\partial}\hat{U}^1 - u_t(t_1)\| \le C(h^r + k) \tag{5.4}$$

and

$$|\bar{\partial}\hat{U}^1 - u_t(t_1)|_{-2} \le C(h^{r+2} + k). \tag{5.5}$$

Proof. We first show (5.4). We write

$$\bar{\partial}\hat{U}^1 - u_t(t_1) = \big(\bar{\partial}\hat{U}^1 - \hat{u}_{h,t}(t_1)\big) + \big(\hat{u}_{h,t}(t_1) - u_t(t_1)\big).$$

By (3.16), we have

$$\|\hat{u}_{h,t}(t_1) - u_t(t_1)\| \le C_0h^r\Big(\|u_t(0)\|_r + \int_0^{t_1}\|u_{tt}(s)\|_r\,ds\Big). \tag{5.6}$$

For $\bar{\partial}\hat{U}^1 - \hat{u}_{h,t}(t_1)$, we have, by (5.2) and (5.3),

$$\bar{\partial}\hat{U}^1 - \hat{u}_{h,t}(t_1) = A_h(\hat{U}^1 - \hat{u}_h^1).$$

Here, by Taylor's formula, with $r(\lambda) = 1/(1+\lambda)$, $E_h(t) = e^{-tA_h}$,

$$\begin{aligned}\hat{U}^1 - \hat{u}_h^1 &= \big(r(kA_h) - E_h(t_1)\big)R_h v + kr(kA_h)P_h F(u^1) \\ &\quad - \int_0^{t_1} E_h(t_1 - s)P_h F(u(s))\,ds \\ &= \big(r(kA_h) - E_h(t_1)\big)R_h v + kb_0(kA_h)P_h F(u(0)) + kR(F),\end{aligned}$$

where

$$b_0(\lambda) = r(\lambda) - \int_0^1 e^{-(1-s)\lambda}\,ds$$

and

$$\begin{aligned}R(F) &= r(kA_h)\int_0^k P_h F'(u(\tau))u_t(\tau)\,d\tau \\ &\quad - \int_0^1 e^{-(1-s)kA_h}\int_0^{ks} P_h F'(u(\tau))u_t(\tau)\,d\tau\,ds.\end{aligned}$$

Thus, we have

$$\begin{aligned}\bar{\partial}\hat{U}^1 - \hat{u}_{h,t}(t_1) &= \big(r(kA_h) - E_h(t_1)\big)A_h R_h v \\ &\quad + kA_h b_0(kA_h)P_h F(u(0)) + kA_h R(F).\end{aligned}$$

Noting that $A_h R_h = P_h A$ and $\lambda b_0(\lambda) = -\big(r(\lambda) - e^{-\lambda}\big)$, we get

$$\begin{aligned}\bar{\partial}\hat{U}^1 - \hat{u}_{h,t}(t_1) &= \big(r(kA_h) - E_h(t_1)\big)P_h\big(Av - F(u(0))\big) + kA_h R(F) \\ &= \big(r(kA_h) - E_h(t_1)\big)P_h u_t(0) + kA_h R(F) \\ &= I + II.\end{aligned} \tag{5.7}$$

For I, we have, by the error estimate for homogeneous parabolic problems,

$$\begin{aligned}\|I\| &\le \big\|\big(r(kA_h) - E_h(t_1)\big)(P_h - R_h)u_t(0)\big\| + \big\|\big(r(kA_h) - E_h(t_1)\big)R_h u_t(0)\big\| \\ &\le \|(P_h - R_h)u_t(0)\| + C_0 k\|A_h R_h u_t(0)\| \\ &\le C_0 h^r\|u_t(0)\|_r + C_0 k\|Au_t(0)\|.\end{aligned}$$

For II, we write

$$\begin{aligned}II &= kA_h r(kA_h)\int_0^k P_h F'(u(\tau))u_t(\tau)\,d\tau \\ &\quad - kA_h\int_0^1 e^{-(1-s)kA_h}\int_0^{ks} P_h F'(u(\tau))u_t(\tau)\,d\tau\,ds \\ &= II_1 + II_2.\end{aligned}$$

We have, noting that $|\lambda r(\lambda)| \le 1,\ \|P_h\| \le 1$,

$$\begin{aligned}\|II_1\| &\le \|kA_h r(kA_h)\| \int_0^k \|P_h F'(u(\tau))u_t(\tau)\|\, d\tau \\ &\le k \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\|,\end{aligned}$$

and, by exchanging the integral order and noting that $\int_\epsilon^1 \lambda e^{-(1-s)\lambda}\, ds \le 1$ for $0 \le \epsilon \le 1$,

$$\begin{aligned}\|II_2\| &= \Big\| kA_h \int_0^k P_h F'(u(\tau))u_t(\tau) \int_{\tau/k}^1 e^{-(1-s)kA_h}\, ds d\tau \Big\| \\ &\le k \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\| \Big\| kA_h \int_{\tau/k}^1 e^{-(1-s)kA_h}\, ds \Big\| \\ &\le k \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\|.\end{aligned}$$

Together these estimates show

$$\|\bar\partial \hat U^1 - \hat u_{h,t}(t)\| \le C_0\big(h^r \|u_t(0)\|_r + k\|Au_t(0)\| + k \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\|\big). \tag{5.8}$$

Combining this with (5.6) shows (5.4).

We now turn to (5.5). We again write

$$\bar\partial \hat U^1 - u_t(t_1) = \big(\bar\partial \hat U^1 - \hat u_{h,t}(t_1)\big) + \big(\hat u_{h,t}(t_1) - u_t(t_1)\big).$$

The desired bound for $|\hat u_{h,t}(t_1) - u_t(t_1)|_{-2}$ follows from (3.17).

For $\bar\partial \hat U^1 - \hat u_{h,t}(t_1)$, we have, by Lemma 2.1,

$$|\bar\partial \hat U^1 - \hat u_{h,t}(t_1)|_{-2} \le C_0\big(h^2 \|\bar\partial \hat U^1 - \hat u_{h,t}(t_1)\| + |\bar\partial \hat U^1 - \hat u_{h,t}(t_1)|_{-2,h}\big).$$

Thus, by (5.7),

$$|\bar\partial \hat U^1 - \hat u_{h,t}(t_1)|_{-2,h} \le |I|_{-2,h} + |II|_{-2,h}.$$

For $|I|_{-2,h}$, we have, by the error estimate for homogeneous parabolic problems [30],

$$|I|_{-2,h} = |(r(kA_h) - E_h(t_1))P_h u_t(0)|_{-2,h} \le C_0\big(h^{r+2}\|u_t(0)\|_r + k\|Au_t(0)\|\big).$$

For $|II|_{-2,h}$, we have, noting that $|r(\lambda)| \le 1,\ \int_\epsilon^1 e^{-(1-s)\lambda}\, ds \le 1$ for $0 \le \epsilon \le 1$,

$$\begin{aligned}|II|_{-2,h} &\le \int_0^k \|kr(kA_h)\|\, \|P_h F'(u(\tau))u_t(\tau)\|\, d\tau \\ &\quad + \Big\| k \int_0^k P_h F'(u(\tau))u_t(\tau) \int_{\tau/k}^1 e^{-(1-s)kA_h}\, ds\, d\tau \Big\| \\ &\le k^2 \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\|.\end{aligned}$$

Hence we get

$$|\bar\partial \hat U^1 - \hat u_{h,t}(t_1)|_{-2,h} \le C_0\big(h^{r+2}\|u_t(0)\|_r + k\|Au_t(0)\| + k^2 \max_{0\le\tau\le k} \|F'(u(\tau))u_t(\tau)\|\big).$$

Combining this with (5.8) shows (5.5).

Together these estimates complete the proof of the theorem. ∎

6. High-order time-stepping. The postprocessing requires very accurate time-stepping in order to match the high-order spatial approximation. It would be natural then to use a time-stepping method of higher order than the backward Euler method of section 4. However, we have not been able to analyze such methods except in the case of linear homogeneous problems, where we can apply the analysis of time derivative approximation from [30].

In this section we consider the linear homogeneous parabolic problem

$$u_t + Au = 0 \quad \text{for } t > 0, \quad \text{with } u(0) = v. \tag{6.1}$$

We define the time-stepping method

$$U^n = r(kA_h)U^{n-1}, \quad U^0 = v_h, \tag{6.2}$$

where $r(\lambda)$ is a rational function and accurate of order $p \geq 1$, i.e.,

$$r(\lambda) - e^{-\lambda} = O(\lambda^{p+1}), \quad \lambda \to 0.$$

For example, if $r(\lambda) = 1/(1+\lambda)$, then we have $(1+kA_h)U^n = U^{n-1}$, which is the backward Euler method. If $r(\lambda) = (1 - \frac{\lambda}{2})/(1 + \frac{\lambda}{2})$, then we have $(1 + \frac{1}{2}kA_h)U^n = (1 - \frac{1}{2}kA_h)U^{n-1}$, which is the Crank–Nicolson method.

Further we define the quotient Q_kU^n to approximate the time derivative $u_{h,t}(t_n)$, with positive integers m_1, m_2 and real numbers c_ν,

$$Q_kU^n = k^{-1} \sum_{\nu=-m_1}^{m_2} c_\nu U^{n+\nu} \quad \text{for } n \geq m_1. \tag{6.3}$$

We assume that the operator Q_k satisfies, for any smooth function u,

$$Q_ku^n - u_t(t_n) = O(k^p), \quad k \to 0. \tag{6.4}$$

The postprocessing step in the fully discrete case is to find $\tilde{u}(T) \in S_h$, $T = t_n$, such that

$$A\tilde{u}(T) = -Q_kU^n. \tag{6.5}$$

The finite element solution of the elliptic problem (6.5) with respect to $\tilde{S}_h$ is to find $\tilde{u}_h(T) \in \tilde{S}_h$, such that

$$\tilde{A}_h\tilde{u}_h(T) = \tilde{P}_h(-Q_kU^n). \tag{6.6}$$

Our main theorem in this section is the following.

THEOREM 6.1. *Let $r \geq 4$ and let S_h and $\tilde{S}_h$ be the finite element spaces of orders r and $r+2$, respectively, as described in section 2. Let $\tilde{u}_h$ and u be the solutions of (6.6) and (6.1), respectively. Let $T = t_n$ be a fixed time. Then we have, if $v_h = R_hv$,*

$$\|\tilde{u}_h(T) - u(T)\| \leq C_0\big(h^{r+2}|v|_{r+2} + k^p|v|_{2(p+1)} + h^{r+2}\|u_t(T)\|_r\big) \quad \textit{for } r \geq 4.$$

Recalling the proof of Theorem 4.1, we note that Theorem 6.1 follows once we have proved appropriate estimates of $\|Q_kU^n - u_t(t_n)\|$ and $|Q_kU^n - u_t(t_n)|_{-2}$, which are given in the following two lemmas.

LEMMA 6.2. *Let U^n and u be the solutions of* (6.2) *and* (6.1), *respectively. Assume that $|r(\lambda)| < 1$ for $\lambda > 0$. Then we have, if $v_h = R_h v$,*

$$\|Q_k U^n - u_t(t_n)\| \le C_0(h^r |v|_{r+2} + k^p |v|_{2(p+1)}).$$

Lemma 6.2 was proved in [30].

LEMMA 6.3. *Let U^n and u be the solutions of* (6.2) *and* (6.1), *respectively. Assume that $|r(\lambda)| < 1$ for $\lambda > 0$. Then we have, if $v_h = R_h v$,*

$$|Q_k U^n - u_t(t_n)|_{-2} \le C_0\big(h^{r+2} |v|_{r+2} + k^p |v|_{2(p+1)}\big).$$

Proof. By Thomée [27, Theorem 6.4], we have

$$\big|u_{h,t}(t) - u_t(t)\big|_{-2} \le C h^{r+2} |v|_{r+2}.$$

Therefore it suffices to show

$$\big|Q_k U^n - u_{h,t}(t_n)\big|_{-2} \le C\big(h^{r+2} |v|_{r+2} + k^p |v|_{2(p+1)}\big), \tag{6.7}$$

which we will prove now.

We first estimate $\big|Q_k U^n - u_{h,t}(t_n)\big|_{-2,h}$. Noting that, with $v_h = R_h v = G_h A v$,

$$\begin{aligned} Q_k U^n - u_{h,t}(t_n) &= k^{-1}\Big(\sum_{\nu=-m_1}^{m_2} c_\nu U^{n+\nu} - (-A_h) e^{-nkA_h}\Big) G_h A v \\ &= k^{-1} g_n(kA_h) G_h A v, \end{aligned}$$

where $g_n(\lambda) = \sum_{\nu=-m_1}^{m_2} r(\lambda)^{n+\nu} - (-\lambda) e^{-n\lambda}$, we need to show

$$\big\|G_h\big(k^{-1} g_n(kA_h) G_h A v\big)\big\| \le C_0\big(h^{r+2} |v|_{r+2} + k^p |v|_{2(p+1)}\big).$$

To do this we set

$$v_k = \sum_{k\lambda_l \le 1} (v, \varphi_l)\varphi_l,$$

where φ_l and λ_l are the eigenfunctions and eigenvalues of the operator A. Then $v_k \in \dot{H}^s$ for each $s \ge 0$. Further, by the definition of the norm in $\dot{H}^s$, we easily find

$$\|A(v - v_k)\| \le k^p |v|_{2p+2}, \tag{6.8}$$

$$|v_k|_{2(p+1)} \le |v|_{2(p+1)}, \tag{6.9}$$

and

$$|v_k|_{r+2l+2} \le k^{-l} |v|_{r+2} \quad \text{for } 0 \le l \le p-1. \tag{6.10}$$

Applying now the identity

$$v = \sum_{j=0}^{p-1} G_h^j (G - G_h) A^{j+1} v + G_h^p A^p v \quad \text{for } v \in \dot{H}^{2p}, \quad \text{where } G_h^0 = I, \tag{6.11}$$

to $v = Av_k$, we get

$$G_h g_n(kA_h)G_hAv_k = \sum_{l=0}^{p-1} g_n(kA_h)G_h^{l+1}\big(G_h(G-G_h)A^{l+2}v_k\big) + G_h g_n(kA_h)G_h^{p+1}A^{p+1}v_k.$$

It is easy to show that (see, e.g., [30, Lemma 3.9])

$$\|g_n(kA_h)G_h^{l+1}\| \le C_0 k^{l+1} \quad \text{for } 0 \le l \le p,\ n \ge 0. \tag{6.12}$$

Thus, by (6.9) and noting the boundedness of G_h,

$$\begin{aligned}\|G_h g_n(kA_h)G_h^{p+1}A^{p+1}v_k\| &\le \|g_n(kA_h)G_h^{p+1}A^{p+1}v_k\| \\ &\le C_0 k^{p+1}\|A^{p+1}v_k\| \le C_0 k^{p+1}|v_k|_{2(p+1)} \le C_0 k^{p+1}|v|_{2(p+1)}.\end{aligned}$$

Further, by (6.10), (6.12), and using (2.13), we have, with $0 \le l \le p-1$,

$$\begin{aligned}&\big\|g_n(kA_h)G_h^{l+1}\big(G_h(G-G_h)A^{l+2}v_k\big)\big\| \le C_0 k^{l+1}\|G_h(G-G_h)A^{l+2}v_k\| \\ &\le C_0 k^{l+1}h^2\|(G-G_h)(A^{l+2}v_k)\| + C_0 k^{l+1}h^{r+2}|A^{l+2}v_k|_{r-2} \\ &\le C_0 k^{l+1}h^{r+2}\|A^{l+2}v_k\|_{r-2} \le C_0 k^{l+1}h^{r+2}|v_k|_{r+2l+2} \le C_0 k h^{r+2}|v|_{r+2}.\end{aligned}$$

Together these estimates imply

$$\|G_h g_n(kA_h)G_hAv_k\| \le C_0 k(h^{r+2}|v|_{r+2} + k^p|v|_{2(p+1)}).$$

Since obviously, by (6.8), the boundedness of G_h, and stability, we get

$$\begin{aligned}\|G_h g_n(kA_h)G_hA(v-v_k)\| &\le \|g_n(kA_h)G_hA(v-v_k)\| \\ &\le C_0 k\|A(v-v_k)\| \le C_0 k^{p+1}|v|_{2(p+1)},\end{aligned}$$

we conclude that

$$\begin{aligned}\|G_h(Q_kU^n - u_{h,t}(t_n))\| &= k^{-1}\|G_h g_n(kA_h)G_hAv\| \\ &\le C_0(h^{r+2}|v|_{r+2} + k^p|v|_{2(p+1)}).\end{aligned}$$

By [30, Theorem 3.8], we have

$$\|Q_kU^n - u_{h,t}(t_n)\| \le C_0(h^r|v|_{r+2} + k^p|v|_{2p}).$$

Thus

$$\begin{aligned}|Q_kU^n - u_{h,t}(t_n)|_{-2} &\le \|(G-G_h)(Q_kU^n - u_{h,t}(t_n))\| \\ &\quad + \|G_h(Q_kU^n - u_{h,t}(t_n))\| \\ &\le C_0(h^{r+2}|v|_{r+2} + k^p|v|_{2(p+1)}).\end{aligned}$$

Together these estimates complete the proof. □

After the preparations above we now come to the proof of Theorem 6.1.

Proof of Theorem 6.1. Combining (6.6) and (6.1), we get, with $\tilde{G}_h = \tilde{A}_h^{-1}$,

$$\begin{aligned}\tilde{u}_h(T) - u(T) &= \tilde{G}_h\tilde{P}_h(-Q_kU^n) - G(-u_t) \\ &= (\tilde{G}_h\tilde{P}_h - G)(-Q_kU^n + u_t(t_n)) \\ &\quad - (\tilde{G}_h\tilde{P}_h - G)u_t(t_n) + G(Q_kU^n - u_t).\end{aligned}$$

Thus, by Lemmas 6.2 and 6.3, and noting that $\|(\tilde{G}_h\tilde{P}_h - G)f\| \le Ch^s\|f\|_{s-2}$ for $0 \le s \le r+2$, we have

$$\begin{aligned}\|\tilde{u}_h(T) - u(T)\| &\le C_0h^2\|Q_kU^n - u_t(t_n)\| \\ &\quad + C_0h^{r+2}\|u_t(t_n)\|_r + |(Q_kU^n - u_t(t_n))|_{-2} \\ &\le C_0\big(h^{r+2}|v|_{r+2} + k^p|v|_{2(p+1)} + h^{r+2}\|u_t(t_n)\|_r\big).\end{aligned}$$

Together these estimates complete the proof. ∎

Acknowledgments. I wish to express my sincere gratitude to Prof. Stig Larsson, Chalmers University of Technology, for suggesting the topic of this paper, and for his support and valuable criticism. I thank Dr. Tony Shardlow for helpful discussions and also thank the referees for a number of valuable comments.

REFERENCES

[1] G. Akrivis, M. Crouzeix, and C. Makridakis, *Implicit-explicit multistep finite element methods for nonlinear parabolic problems*, Math. Comp., 67 (1998), pp. 457–477.
[2] G. Akrivis, M. Crouzeix, and C. Makridakis, *Implicit-explicit multistep methods for quasilinear parabolic equations*, Numer. Math., 82 (1999), pp. 521–541.
[3] I. Babuška, *Finite element method for domains with corners*, Computing, 6 (1970), pp. 264–273.
[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Ser. Comput. Phys., Springer-Verlag, New York, 1988.
[6] M. Crouzeix, V. Thomée, and L. B. Wahlbin, *Error estimates for spatially discrete approximations of semilinear parabolic equations with initial data of low regularity*, Math. Comp., 53 (1989), pp. 25–41.
[7] J. de Frutos, B. García-Archilla, and J. Novo, *A postprocessed Galerkin method with Chebyshev or Legendre polynomials*, Numer. Math., 86 (2000), pp. 419–442.
[8] J. de Frutos and J. Novo, *A postprocess based improvement of the spectral element method*, Appl. Numer. Math., 33 (2000), pp. 217–223.
[9] J. de Frutos and J. Novo, *Postprocessing the linear finite element method*, SIAM J. Numer. Anal., 40 (2002), pp. 805–819.
[10] J. de Frutos and J. Novo, *A spectral element method for the Navier–Stokes equations with improved accuracy*, SIAM J. Numer. Anal., 38 (2000), pp. 799–819.
[11] C. M. Elliott and S. Larsson, *Error estimates with smooth and nonsmooth data for a finite element method for the Cahn–Hilliard equation*, Math. Comp., 58 (1992), pp. 603–630.
[12] C. M. Elliott and S. Larsson, *A finite element model for the time-dependent Joule heating problem*, Math. Comp., 64 (1995), pp. 1433–1453.
[13] C. M. Elliott and A. M. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal., 30 (1993), pp. 1622–1663.
[14] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (2000), pp. 470–499.
[15] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.
[16] H. P. Helfrich, *Error estimates for semidiscrete Galerkin type approximations to semilinear evolution equations with nonsmooth initial data*, Numer. Math., 51 (1987), pp. 559–569.
[17] C. Johnson, S. Larsson, V. Thomée, and L. B. Wahlbin, *Error estimates for spatially discrete approximations of semilinear parabolic equations with nonsmooth initial data*, Math. Comp., 49 (1987), pp. 331–357.
[18] S. Larsson, *Nonsmooth Data Error Estimates with Applications to the Study of the Long-Time Behavior of Finite Element Solutions of Semilinear Parabolic Problems*, Department of Mathematics, Chalmers University of Technology, 1992–36, preprint.
[19] S. Larsson, *The long-time behavior of finite-element approximations of solutions to semilinear parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 348–365.

[20] S. Larsson and J.-M. Sanz-Serna, *The behavior of finite element solutions of semilinear parabolic problems near stationary points*, SIAM J. Numer. Anal., 31 (1994), pp. 1000–1018.

[21] S. Larsson and J.-M. Sanz-Serna, *A shadowing result with applications to finite element approximation of reaction-diffusion equations*, Math. Comp., 68 (1999), pp. 55–72.

[22] A. H. Schatz, I. H. Sloan, and L. B. Wahlbin, *Superconvergence in finite element methods and meshes that are locally symmetric with respect to a point*, SIAM J. Numer. Anal., 33 (1996), pp. 505–521.

[23] A. H. Schatz, V. Thomée, and L. B. Wahlbin, *Stability, analyticity, and almost best approximation in maximum norm for parabolic finite element equations*, Comm. Pure Appl. Math., 51 (1998), pp. 1349–1385.

[24] A. H. Schatz and L. B. Wahlbin, *On the quasi-optimality in L_∞ of the $\dot{H}^1$-projection into finite element spaces*, Math. Comp., 38 (1982), pp. 1–22.

[25] R. Scott, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal., 12 (1975), pp. 404–427.

[26] R. Scott, *Optimal L^∞ estimates for the finite element method on irregular meshes*, Math. Comp., 30 (1976), pp. 681–697.

[27] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

[28] V. Thomée and L. Wahlbin, *On Galerkin methods in semilinear parabolic problems*, SIAM J. Numer. Anal., 12 (1975), pp. 378–389.

[29] M. F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

[30] Y. Yan, *Smoothing properties and approximation of time derivatives for parabolic equations: Constant time steps*, IMA J. Numer. Anal., 23 (2003), pp. 465–487.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1703–1720

ERROR ESTIMATES TO SMOOTH SOLUTIONS OF RUNGE–KUTTA DISCONTINUOUS GALERKIN METHOD FOR SYMMETRIZABLE SYSTEMS OF CONSERVATION LAWS*

QIANG ZHANG† AND CHI-WANG SHU‡

Abstract. In this paper we study the error estimates to sufficiently smooth solutions of symmetrizable systems of conservation laws for the Runge–Kutta discontinuous Galerkin (RKDG) method. Time discretization is the second-order explicit TVD (total variation diminishing) Runge–Kutta method, and the $\mathbb{P}^k$ (piecewise polynomial) finite element is used. When $k = 1$ (piecewise linear finite element), the error estimate is obtained under the usual CFL condition $\tau \leq \beta h$ for nonlinear systems in one dimension and for linear systems in multiple space dimensions. Here, h is the maximum element length, τ is the time step, and β is a positive constant independent of h and τ. Error estimates for $\mathbb{P}^k$ finite elements with $k > 1$ are obtained under a more restrictive CFL condition.

Key words. discontinuous Galerkin method, finite element method, TVD Runge–Kutta method, error estimates, symmetrizable system, conservation laws

AMS subject classification. 65M15

DOI. 10.1137/040620382

1. Introduction.

In this paper, we continue our work in [24] and present the error estimates of the Runge–Kutta discontinuous Galerkin (RKDG) method for smooth solutions of symmetrizable systems of conservation laws,

$$\boldsymbol{u}_{,t} + \boldsymbol{f}^{(i)}_{,x_i}(\boldsymbol{u}) = 0, \qquad (x,t) \in \Omega \times (0,T], \tag{1.1a}$$

$$\boldsymbol{u}(x,0) = \boldsymbol{u}_0, \qquad x \in \Omega, \tag{1.1b}$$

in the spatial domain $\Omega \in \mathbb{R}^d$ and the time interval $[0,T]$. Here, $\boldsymbol{u}(x,t)\colon \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^m$ is the dependent solution variable, $\boldsymbol{f}(\boldsymbol{u}) = (\boldsymbol{f}^{(1)}(\boldsymbol{u}), \ldots, \boldsymbol{f}^{(d)}(\boldsymbol{u}))\colon \mathbb{R}^m \to \mathbb{R}^{m\times d}$ is the vector-valued flux function, and the implied summation on the index i is used in (1.1a), i.e., $\boldsymbol{f}^{(i)}_{,x_i} = \sum_{i=1}^d \partial \boldsymbol{f}^{(i)}/\partial x_i$. We do not pay attention to boundary conditions in this paper; hence the solution is considered to be either periodic or compactly supported. For simplicity of presentation, we will give detailed analysis only for the one-dimensional case where $d = 1$ and $\Omega = I = (0,1)$; herein we drop the index i in (1.1). We will, however, point out similarities and differences when the analysis is generalized to multiple space dimensions. We assume, in addition, that each component of the flux function $\boldsymbol{f}(\boldsymbol{u})$ is smooth enough in $\boldsymbol{u}$; for our purpose $C^3(\mathbb{R}^m)$ will suffice. The analysis in this paper is for *smooth* solutions of (1.1). Discontinuous solutions with shocks are not considered.

*Received by the editors December 7, 2004; accepted for publication (in revised form) April 28, 2006; published electronically September 15, 2006.
http://www.siam.org/journals/sinum/44-4/62038.html

†College of Mathematical Science, Nankai University, Tianjin, 300071, P.R. China (qzh@nankai.edu.cn). The research of this author was supported by CNNSF grant 10301016.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (shu@dam.brown.edu). The research of this author was supported by the Chinese Academy of Sciences while the author was in residence at the University of Science and Technology of China (grant 2004-1-8) and at the Institute of Computational Mathematics and Scientific / Engineering Computing. Additional support was provided by ARO grant W911NF-04-1-0291, NSF grants DMS-0207451 and DMS-0510345, and AFOSR grant FA9550-05-1-0123.

The RKDG method was introduced and developed by Cockburn and coworkers [4, 5, 3, 2, 6] for solving nonlinear hyperbolic conservation laws; the method uses discontinuous Galerkin (DG) discretization in space and combines it with an explicit total variation diminishing (TVD) Runge–Kutta time-marching algorithm [21]. This method has a good stability property, is flexible for h-p adaptivity, and has a high parallel efficiency. In recent years there has been a lot of activity in the design, analysis, and application of RKDG methods. For more details, we refer to the review article [8].

Although error estimates for linear equations and for the method of lines (continuous in time) version of the RKDG method have been available for a long time (e.g., [15, 14, 7]), error estimates for the fully discrete RKDG method for nonlinear conservation laws with smooth solutions have become available only recently. In [24] we obtained error estimates for scalar conservation laws with piecewise kth-degree polynomial DG spatial discretization coupled with second-order TVD Runge–Kutta time discretization. The analysis assumed the usual CFL condition $\tau \le \beta h$ for the piecewise linear $k = 1$ case, where h is the maximum element length, τ is the time step, and β is a suitable positive constant independent of h and τ. For the higher-order $k > 1$ case, the proof had to assume a much stronger CFL condition $\tau \le \beta h^{4/3}$ for an arbitrary positive constant β. In this paper, we extend these error estimates to symmetrizable systems (see Theorem 2.1).

In the symmetrization theory [20] for the first-order conservation laws, one seeks a mapping $\boldsymbol{u}(\boldsymbol{v})\colon \mathbb{R}^m \to \mathbb{R}^m$ applied to (1.1a) so that when transformed,

$$\boldsymbol{u}_{,\boldsymbol{v}}\boldsymbol{v}_{,t} + \boldsymbol{f}_{,\boldsymbol{v}}\boldsymbol{v}_{,x} = 0, \tag{1.2}$$

the matrix $\boldsymbol{u}_{,\boldsymbol{v}}$ is symmetric positive definite (SPD) and the matrix $\boldsymbol{f}_{,\boldsymbol{v}} = \boldsymbol{f}_{,\boldsymbol{u}}\boldsymbol{u}_{,\boldsymbol{v}}$ is also symmetric, where $\boldsymbol{f}_{,\boldsymbol{v}} = \{\boldsymbol{f}(\boldsymbol{u}(\boldsymbol{v}))\}_{,\boldsymbol{v}}$. We further assume that each component of $\boldsymbol{u}_{,\boldsymbol{v}}$ is Lipschitz continuous with respect to the variable $\boldsymbol{v}$. As is well known, a conservation law system (1.1a) is symmetrizable if and only if it has a convex entropy function [10]. Well-known systems such as the Euler equations of compressible gas dynamics are symmetrizable. If $\boldsymbol{f}_{,\boldsymbol{u}}$ is already symmetric, the system (1.1a) is symmetric. It is rather straightforward to generalize the error estimates in [24] from the scalar case to symmetric systems. However, there are not that many physical systems that are symmetric. On the other hand, as we will see later in this paper, it is significantly more difficult to generalize the error estimates in [24] from the scalar case to symmetrizable systems.

The line of analysis in this paper follows that of [24]. The main techniques are Taylor expansions and energy analysis. In generalizing the analysis from the scalar case to systems, we need to pay attention to the suitable norm in the analysis, to a careful classification of the necessary properties for the numerical fluxes, to the complication related to the fact that derivatives (Jacobians) and second derivatives of the flux functions are matrices and supermatrices, and to a suitable generalization of the a priori assumption about the numerical solution. We will present a series of lemmas, which mostly correspond to those in [24]. If the proofs have only minor differences, we will comment on such differences and will not repeat the details. We will concentrate our analysis on the piecewise $k = 1$ case for the DG method, since higher-order cases can be analyzed with the stronger CFL condition $\tau \le \beta h^{4/3}$, following the same lines as those in [24], once the $k = 1$ case is proved.

An outline of this paper is as follows. In section 2 we present, for (1.1), the RKDG method and the corresponding convergence theorem with the second-order

TVD Runge–Kutta time discretization, where we introduce a definition of the generalized E-flux property for symmetrizable systems of conservation laws. In section 3 we present a proposition for an important matrix which measures the amount of numerical viscosity on each element interface, and perform some elementary analysis to the error equations. We prove the convergence theorem in section 4. Section 5 is an appendix, in which we give the technical details of the estimates omitted in previous sections.

2. RKDG method and the convergence theorem. In this section we will follow [6] and define the RKDG method for the problem (1.1) in one space dimension, presenting the corresponding convergence theorem without proof. The multidimensional scheme can be similarly defined, and the analysis can be carried out in a similar fashion for linear as well as nonlinear symmetrizable systems with a few modifications; see Remark 4.2.

2.1. RKDG method. For each partition of the interval $I = (0,1)$, $\{x_{j+\frac{1}{2}}\}_{j=0}^{N}$, we set $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ and $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ for $j = 1, \ldots, N$; we denote the quantity $\max_{1\le j\le N} h_j$ by h. For a given time step $\tau \equiv \tau^n$ (which could actually change from step to step, but is taken as a constant with respect to the time level n for simplicity), the solution of the scheme is denoted by $\boldsymbol{u}_h^n(x) = \boldsymbol{u}_h(x, t^n) = \boldsymbol{u}_h(x, n\tau)$, which belongs to the finite element space

$$\mathbb{V}_h = \{\, \boldsymbol{v} \in [L^1(0,1)]^m : \boldsymbol{v}|_{I_j} \in [\mathbb{P}^k(I_j)]^m,\ j = 1, \ldots, N \,\}, \tag{2.1}$$

where $\mathbb{P}^k(I_j)$ denotes the space of polynomials in I_j of degree at most k. Note that each component of a vector-valued function in $\mathbb{V}_h$ is allowed to have discontinuities across element interfaces.

In what follows, we will consider the standard L^2-projection of a vector-valued function $\boldsymbol{p} \in [L^2(0,1)]^m$ into the finite element space $\mathbb{V}_h$, denoted by $\mathbb{P}_h \boldsymbol{p}$, which is defined as the unique vector-valued function in $\mathbb{V}_h$ such that

$$\int_0^1 \boldsymbol{z}_h^T(x)(\mathbb{P}_h \boldsymbol{p}(x) - \boldsymbol{p}(x))\, dx = 0 \quad \forall \boldsymbol{z}_h \in \mathbb{V}_h, \tag{2.2}$$

where $\boldsymbol{z}^T$ denotes the transpose of the vector $\boldsymbol{z}$.

As usual, at each element interface we will denote, for a vector-valued function $\boldsymbol{z}$, two limiting values from different directions by $\boldsymbol{z}^{\pm}_{j+1/2} = \boldsymbol{z}(x_{j+1/2} \pm 0)$, and denote the average and jump by $\bar{\boldsymbol{z}} = (\boldsymbol{z}^+ + \boldsymbol{z}^-)/2$ and $[\boldsymbol{z}] = \boldsymbol{z}^+ - \boldsymbol{z}^-$, respectively. We also define, for any vector-valued functions $\boldsymbol{p}$ and $\boldsymbol{z}$, the following functional corresponding to the DG spatial discretization:

$$\mathcal{H}_j(\boldsymbol{p}, \boldsymbol{z}) = \int_{I_j} \boldsymbol{z}_{,x}^T \boldsymbol{f}(\boldsymbol{p})\, dx - (\boldsymbol{z}^-_{j+\frac{1}{2}})^T \hat{\boldsymbol{h}}(\boldsymbol{p})_{j+\frac{1}{2}} + (\boldsymbol{z}^+_{j-\frac{1}{2}})^T \hat{\boldsymbol{h}}(\boldsymbol{p})_{j-\frac{1}{2}}, \tag{2.3}$$

where $\hat{\boldsymbol{h}}(\boldsymbol{p}) \equiv \hat{\boldsymbol{h}}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ is a given (locally) Lipschitz continuous numerical flux function consistent with the flux function $\boldsymbol{f}(\boldsymbol{p})$, that is, $\hat{\boldsymbol{h}}(\boldsymbol{p}, \boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{p})$. In this paper, we will also assume that $\hat{\boldsymbol{h}}(\boldsymbol{p})$ is a generalized E-flux function, to be defined in the next subsection 2.2.

The approximate solution in $\mathbb{V}_h$ from time $n\tau$ to $(n+1)\tau$ given by the RKDG method with second-order TVD time discretization can now be defined as follows:

find successively $\boldsymbol{w}_h^n \equiv \boldsymbol{w}_h^n(x) \in \mathbb{V}_h$ and $\boldsymbol{u}_h^{n+1} \equiv \boldsymbol{u}_h^{n+1}(x) \in \mathbb{V}_h$ such that, for any $\boldsymbol{z}_h \equiv \boldsymbol{z}_h(x) \in [\mathbb{P}^k(I_j)]^m$ and $1 \le j \le N$,

$$\int_{I_j} \boldsymbol{z}_h^T \boldsymbol{w}_h^n \, dx = \int_{I_j} \boldsymbol{z}_h^T \boldsymbol{u}_h^n \, dx + \tau \mathcal{H}_j(\boldsymbol{u}_h^n, \boldsymbol{z}_h), \tag{2.4a}$$

$$\int_{I_j} \boldsymbol{z}_h^T \boldsymbol{u}_h^{n+1} \, dx = \frac{1}{2} \int_{I_j} \boldsymbol{z}_h^T \boldsymbol{u}_h^n \, dx + \frac{1}{2} \int_{I_j} \boldsymbol{z}_h^T \boldsymbol{w}_h^n \, dx + \frac{\tau}{2} \mathcal{H}_j(\boldsymbol{w}_h^n, \boldsymbol{z}_h), \tag{2.4b}$$

with the initial value $\boldsymbol{u}_h^0 = \mathbb{P}_h \boldsymbol{u}_0(x)$. This is an explicit time marching method when a local orthogonal basis is chosen for polynomials on I_j or when a small local mass matrix on I_j is inverted. Numerical results and details of this scheme can be found in [9] and [8].

2.2. Generalized E-flux. In this subsection we will introduce an important assumption on the numerical fluxes used in the scheme (2.4). The symmetrizable theory implies that the Jacobian $\boldsymbol{f}_{,\boldsymbol{u}}$ is similar to a symmetric matrix, since

$$\boldsymbol{u}_{,\boldsymbol{v}}^{-1/2} \boldsymbol{f}_{,\boldsymbol{u}} \boldsymbol{u}_{,\boldsymbol{v}}^{1/2} = \boldsymbol{u}_{,\boldsymbol{v}}^{-1/2} \boldsymbol{f}_{,\boldsymbol{v}} \boldsymbol{u}_{,\boldsymbol{v}}^{-1/2} \tag{2.5}$$

and the matrix on the right-hand side is symmetric. It give us a motivation to preserve more properties of the numerical flux from the scalar and symmetric system cases to the symmetrizable system case.

We assume in this paper that the numerical flux function $\hat{\boldsymbol{h}}(\boldsymbol{p})$ is locally Lipschitz continuous and, for $\boldsymbol{r}_i = \boldsymbol{p}^-$, $\bar{\boldsymbol{p}}$, and $\boldsymbol{p}^+$, satisfies

$$(\boldsymbol{p}^+ - \boldsymbol{p}^-)^T \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_i)\{\boldsymbol{f}(\boldsymbol{r}_i) - \hat{\boldsymbol{h}}(\boldsymbol{p})\} \ge 0, \quad i = 1, 2, 3, \tag{2.6}$$

for some reference vectors $\boldsymbol{s}_i$, $i = 1, 2, 3$, which are inside the m-dimensional cube, with $\boldsymbol{p}^+$ and $\boldsymbol{p}^-$ being the endpoints of the longest diagonal line. These reference vectors may depend on the numerical fluxes $\hat{\boldsymbol{h}}$ and on $\boldsymbol{p}^{\pm}$.

The inequality (2.6), for a symmetric system of conservation laws (in which case $\boldsymbol{v}_{,\boldsymbol{u}} \equiv \mathbf{I}$), has been considered in [13] as an E-flux (see [17] for the definition of E-fluxes for scalar conservation laws). Therefore, we refer to a numerical flux satisfying (2.6) as a generalized E-flux. It is easy to verify that the property (2.6) holds for many numerical flux functions constructed from approximate Riemann solvers: for example, the Roe linearization flux function [19], with or without Harten's entropy fix [11], and the global (local) Lax–Friedrichs flux, where $\boldsymbol{s}_i$ may be chosen as the so-called Roe average [19] of $\boldsymbol{p}^+$ and $\boldsymbol{p}^-$.

2.3. The convergence theorem. We now present the main convergence theorem of the RKDG scheme (2.4). The proof will be given in the next two sections.

THEOREM 2.1. *For the symmetrizable system of conservation laws* (1.1)*, assume that the solution* $\boldsymbol{u}$ *and the flux function* $\boldsymbol{f}(\boldsymbol{u})$ *are sufficiently smooth with bounded derivatives. Let* $\boldsymbol{u}_h$ *be the numerical approximate solution of the RKDG scheme* (2.4) *with the second-order TVD Runge–Kutta time discretization, where the numerical flux* $\hat{\boldsymbol{h}}(\cdot, \cdot)$ *is assumed to be a generalized E-flux; namely,* (2.6) *is satisfied. For regular triangulations of* $I = (0, 1)$*, if the finite element space* $\mathbb{V}_h$ *is of piecewise polynomials of degree* $k \ge 1$*, then for small enough* h *there holds the following estimate:*

$$\max_{n\tau \le T} ||\boldsymbol{u}(t^n) - \boldsymbol{u}_h^n||_{L^2(0,1)} \le C(h^{k+1/2} + \tau^2), \tag{2.7}$$

where the positive constant C is independent of h, τ, and the approximate solution $\boldsymbol{u}_h$. This estimate holds for $k \geq 2$ under the restrictive time step condition $\tau \leq \beta h^{4/3}$ with any given positive constant β; meanwhile, it holds for $k = 1$ under the usual CFL condition $\tau \leq \beta h$ with a suitable positive CFL number β which is independent of τ and h.

We remark that the power $h^{k+1/2}$ is optimal for general triangulations [18] for the scalar case but is suboptimal for the one-dimensional case with scalar equations. The proof of the optimal order h^{k+1} for the scalar case requires special upwind fluxes [24], which can be done for the system case (1.1) as well in some special situations. See Remark 4.3 in section 4.

For the generalization of these results to multiple space dimensions, see Remark 4.2 in section 4.

In what follows, we would like to assume as in [24] that each component of the flux function $\boldsymbol{f}(\cdot)$ itself, and its derivatives up to third order, are bounded in the domain $\mathbb{R}^m$. This assumption is nonessential if we consider only smooth solutions of (1.1) to a finite time T. We could achieve the desired boundedness by redefining the flux function $\boldsymbol{f}(\boldsymbol{u})$ outside the range of the solution $\boldsymbol{u}$; cf. [24].

We denote the inverse mapping of $\boldsymbol{u}(\boldsymbol{v})$ by $\boldsymbol{v}(\boldsymbol{u})$. The symmetrizable theory provides that the Jacobians $\boldsymbol{u}_{,\boldsymbol{v}}(\boldsymbol{v})$ and $\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u})$ are both SPD and Lipschitz continuous. Similarly as above, we assume that these properties hold uniformly in the domain $\mathbb{R}^m$ and that the spectrum of the Jacobians is bounded.

We would also like to denote, by C, $C_\star$, M, or ε, a generic positive constant independent of n, h, and τ. Herein, M and ε are used to denote constants which are independent of the solution of (1.1). $C_\star$ is used to emphasize the nonlinearity of $\boldsymbol{f}(\boldsymbol{u})$; i.e., $C_\star = 0$ for a linear flux function $\boldsymbol{f}(\boldsymbol{u}) = \mathbb{C}\boldsymbol{u}$. These constants may have a different value in each occurrence.

3. Error equations, energy equality, and a few estimates. We follow the idea in [24] to obtain the error estimate to sufficiently smooth solutions for the RKDG scheme (2.4). In this section we will present some elementary development similar to that in [24], and we omit the detailed proof if it is similar to that in [24].

3.1. Error equations and energy equality. We denote the error at each stage of the considered RKDG scheme by $\boldsymbol{e}_{\boldsymbol{u}}^n = \boldsymbol{u}(x, t^n) - \boldsymbol{u}_h^n$ and $\boldsymbol{e}_{\boldsymbol{w}}^n = \boldsymbol{w}(x, t^n) - \boldsymbol{w}_h^n$, respectively, where $\boldsymbol{w}(x, t)$ is a vector-valued function in parallel to an Euler forward time marching, namely

$$\boldsymbol{w}(x, t) = \boldsymbol{u}(x, t) + \tau \boldsymbol{u}_{,t}(x, t). \tag{3.1}$$

As is customary in finite element error analysis, we define $\boldsymbol{\xi}_{\boldsymbol{p}} = \mathbb{P}_h \boldsymbol{p} - \boldsymbol{p}_h$ and $\boldsymbol{\eta}_{\boldsymbol{p}} = \mathbb{P}_h \boldsymbol{p} - \boldsymbol{p}$, where $\mathbb{P}_h$ is the local L^2-projection. Then the error is decomposed by $\boldsymbol{e}_{\boldsymbol{p}}^n = \boldsymbol{\xi}_{\boldsymbol{p}}^n - \boldsymbol{\eta}_{\boldsymbol{p}}^n$, where $\boldsymbol{p} = \boldsymbol{u}$ or $\boldsymbol{w}$. The estimates for $\boldsymbol{\eta}_{\boldsymbol{p}}$ will be discussed easily in subsection 3.2, while the estimates to $\boldsymbol{\xi}_{\boldsymbol{p}}$ contain the main difficulty in the analysis.

To perform this analysis, we need the error equations of the RKDG scheme (2.4). These can be obtained by algebraic manipulations similar to those in [23, 24] (cf. Lemma 4.1 in [24]). The error equations are given by

$$\int_{I_j} \boldsymbol{z}_h^T \boldsymbol{\xi}_{\boldsymbol{w}}^n \, dx = \int_{I_j} \boldsymbol{z}_h^T \boldsymbol{\xi}_{\boldsymbol{u}}^n \, dx + \mathcal{K}_j^n(\boldsymbol{z}_h), \tag{3.2a}$$

$$\int_{I_j} \boldsymbol{z}_h^T \boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} \, dx = \int_{I_j} \boldsymbol{z}_h^T \boldsymbol{\xi}_{\boldsymbol{u}}^n \, dx + \frac{1}{2}\mathcal{K}_j^n(\boldsymbol{z}_h) + \frac{1}{2}\mathcal{L}_j^n(\boldsymbol{z}_h), \tag{3.2b}$$

for any $\boldsymbol{z}_h(x) \in [\mathbb{P}^k(I_j)]^m$ and $1 \le j \le N$, where

(3.2c)
$$\mathcal{K}_j^n(\boldsymbol{z}_h) = \int_{I_j} \boldsymbol{z}_h^T(\boldsymbol{\eta}_{\boldsymbol{w}}^n - \boldsymbol{\eta}_{\boldsymbol{u}}^n)\,dx + \tau\mathcal{H}_j(\boldsymbol{u}(t^n), \boldsymbol{z}_h) - \tau\mathcal{H}_j(\boldsymbol{u}_h^n, \boldsymbol{z}_h),$$

(3.2d)
$$\mathcal{L}_j^n(\boldsymbol{z}_h) = \int_{I_j} \boldsymbol{z}_h^T(2\boldsymbol{\eta}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\eta}_{\boldsymbol{w}}^n - \boldsymbol{\eta}_{\boldsymbol{u}}^n + 2\boldsymbol{\zeta}^n)\,dx + \tau\mathcal{H}_j(\boldsymbol{w}(t^n), \boldsymbol{z}_h) - \tau\mathcal{H}_j(\boldsymbol{w}_h^n, \boldsymbol{z}_h).$$

Here $\boldsymbol{\zeta}^n$ is the truncation error in time, with the size $\mathcal{O}(\tau^3)$. Below we will use the short notation $\mathcal{K}^n(\boldsymbol{z}_h) = \sum_{1\le j\le N} \mathcal{K}_j^n(\boldsymbol{z}_h)$ and $\mathcal{L}^n(\boldsymbol{z}_h) = \sum_{1\le j\le N} \mathcal{L}_j^n(\boldsymbol{z}_h)$.

We will use energy estimates to analyze the error of the RKDG scheme (2.4). To this end, we define a norm which depends on the time level n, given by $||\boldsymbol{p}||_n = ||\boldsymbol{v}_{,\boldsymbol{u}}^{1/2}(\boldsymbol{u}_c^n)\boldsymbol{p}||$, for any vector-valued function $\boldsymbol{p}$, where $\boldsymbol{u}_c^n$ is the piecewise constant vector-valued function that is equal to the vector $\boldsymbol{u}(x_j, t^n)$ in each element I_j. The symmetrizable theory guarantees that the $||\cdot||_n$ norm is equivalent to the usual L^2-norm $||\cdot||$.

We first take the test function $\boldsymbol{z}_h = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{\xi}_{\boldsymbol{u}}^n$ in (3.2a) and $\boldsymbol{z}_h = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{w}_c^n)\boldsymbol{\xi}_{\boldsymbol{w}}^n$ in (3.2b), respectively, which belongs to the finite element space $\mathbb{V}_h$. By adding the two equalities together, we obtain the energy equation

$$||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}||_n^2 - ||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 = ||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^n||_n^2 + \mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{\xi}_{\boldsymbol{u}}^n) + \mathcal{L}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{w}_c^n)\boldsymbol{\xi}_{\boldsymbol{w}}^n) + \mathcal{E}^n, \tag{3.3a}$$

where

$$\mathcal{E}^n = \int_I (\boldsymbol{\xi}_{\boldsymbol{w}}^n)^T(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{w}_c^n))(2\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{u}}^n - \boldsymbol{\xi}_{\boldsymbol{w}}^n)\,dx \tag{3.3b}$$

and $\boldsymbol{w}_c^n$ is defined in the same way as $\boldsymbol{u}_c^n$; i.e., it is the piecewise constant vector-valued function which is equal to the vector $\boldsymbol{w}(x_j, t^n)$ in each element I_j. In order to obtain the error estimate, we shall analyze carefully each term on the right-hand side of this important energy equation (3.3) in the next section.

3.2. Properties of the finite element spaces. In this subsection we present some interpolation approximation and inverse inequalities of the finite element space $\mathbb{V}_h$, which consists of piecewise polynomials of degree at most k. The usual notation for norms and seminorms in Sobolev spaces will be used below.

The local L^2-projection is enough to prove Theorem 2.1 for general numerical flux functions with the suboptimal error bound $Ch^{k+1/2}$. By the standard scaling theory, it is easy to show (cf. [1]) that if $\boldsymbol{u}$ and $\boldsymbol{u}_{,t} \in L^\infty([0,T]; [H^{k+1}(I)]^m)$, then

$$||\boldsymbol{\eta}_{\boldsymbol{p}}^n|| + h||\boldsymbol{\eta}_{\boldsymbol{p}}^n||_\infty + h^{\frac{1}{2}}||\boldsymbol{\eta}_{\boldsymbol{p}}^n||_{\Gamma_h} \le Ch^{k+1} \quad (\boldsymbol{p} = \boldsymbol{u}, \boldsymbol{w},\ \forall n : n\tau \le T), \tag{3.4a}$$

where Γ_h is the set of boundary interfaces of all elements and $||\cdot||_{\Gamma_h}$ is the usual L^2-norm on Γ_h. Noticing the definition (3.1) of $\boldsymbol{w}$ and the linearity of the L^2-projection $\mathbb{P}_h$, we can conclude that if $\boldsymbol{u}_{,t} \in L^\infty([0,T]; [H^{k+1}(I)]^m)$, then

$$||\boldsymbol{\eta}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\eta}_{\boldsymbol{u}}^n|| + ||\boldsymbol{\eta}_{\boldsymbol{w}}^n - \boldsymbol{\eta}_{\boldsymbol{u}}^n|| \le Ch^{k+1}\tau \quad \forall n : n\tau < T. \tag{3.4b}$$

In these inequalities the positive constant C depends solely on $\boldsymbol{u}, \boldsymbol{w}$, and/or $\boldsymbol{u}_{,t}$ and is independent of n, h, and τ.

In the following analysis we will also use some inverse inequalities of the finite element space $\mathbb{V}_h$. For any vector-valued function $\boldsymbol{z}_h \in \mathbb{V}_h$, there is a positive constant C independent of $\boldsymbol{z}_h$ and h such that

$$\text{(i) } ||(\boldsymbol{z}_h)_{,x}|| \le Ch^{-1}||\boldsymbol{z}_h||, \quad \text{(ii) } ||\boldsymbol{z}_h||_{\Gamma_h} \le Ch^{-\frac{1}{2}}||\boldsymbol{z}_h||, \quad \text{(iii) } ||\boldsymbol{z}_h||_\infty \le Ch^{-\frac{1}{2}}||\boldsymbol{z}_h||.$$

For more details of these inverse inequalities, we refer to [1].

3.3. An important matrix related to the numerical flux. In this subsection we will introduce an important matrix $\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p})$ associated with a generalized E-flux function $\hat{\boldsymbol{h}}(\boldsymbol{p})$ satisfying (2.6), which measures the numerical viscosity of the flux on each element interface. It is a generalization of a similar quantity in [24] for the scalar case, although we make a minor modification because the definition of the generalized E-flux (2.6) is weaker than that of an E-flux, as we require the inequality (2.6) to hold only for the end points $\boldsymbol{p}^\pm$ and the midpoint $\bar{\boldsymbol{p}}$, rather than for all points between $\boldsymbol{p}^\pm$ for an E-flux. See also [11].

In the next proposition we will use the following notation. If there exists an invertible matrix $\mathbb{T}$ such that $\mathbb{A} = \mathbb{T}^{-1}\mathbb{B}\mathbb{T}$, where $\mathbb{B} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix, then we denote its absolute value matrix by $|\mathbb{A}| = \mathbb{T}^{-1}\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_m|)\mathbb{T}$. We also denote $\mathbb{J}(\boldsymbol{p}) = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{p})\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{p}) = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{p})\boldsymbol{f}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{p}))\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{p})$, which is a symmetric matrix in the symmetrizable theory.

PROPOSITION 3.1. *Assume that the generalized E-flux property* (2.6) *holds for the numerical flux* $\hat{\boldsymbol{h}}(\boldsymbol{p}) \equiv \hat{\boldsymbol{h}}(\boldsymbol{p}^-, \boldsymbol{p}^+)$*, which is consistent with the flux* $\boldsymbol{f}(\boldsymbol{p})$*. Define the matrix on each element interface*

$$\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p}) \equiv \mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p}^-,\boldsymbol{p}^+) := \begin{cases} \dfrac{1}{6}\mathcal{A}_1 + \dfrac{2}{3}\mathcal{A}_2 + \dfrac{1}{6}\mathcal{A}_3, & \text{if } [\boldsymbol{p}] \neq \boldsymbol{0}, \\ \big|\mathbb{J}(\bar{\boldsymbol{p}})\big|, & \text{if } [\boldsymbol{p}] = \boldsymbol{0}, \end{cases} \tag{3.5a}$$

where

$$\mathcal{A}_i = \frac{\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_i)\{\boldsymbol{f}(\boldsymbol{r}_i) - \hat{\boldsymbol{h}}(\boldsymbol{p})\}[\boldsymbol{p}]^T}{[\boldsymbol{p}]^T[\boldsymbol{p}]}, \qquad i = 1, 2, 3, \tag{3.5b}$$

and $\boldsymbol{r}_i$ *and* $\boldsymbol{s}_i$*,* $i = 1, 2, 3$*, are the vectors defined in subsection* 2.2*. Then for any vector* $\boldsymbol{p} \in \mathbb{R}^m$*, the spectrum of* $\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p})$ *is bounded and* $[\boldsymbol{p}]^T\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p})[\boldsymbol{p}] \ge 0$*; moreover,*

$$\frac{1}{3}[\boldsymbol{p}]^T\big|\mathbb{J}(\bar{\boldsymbol{p}})\big|[\boldsymbol{p}] \le [\boldsymbol{p}]^T\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p})[\boldsymbol{p}] + C_\star||[\boldsymbol{p}]||^3, \tag{3.6}$$

where the positive constant $C_\star$ *is determined solely by the nonlinearity of the flux* $\boldsymbol{f}(\boldsymbol{p})$*, and* $||[\boldsymbol{p}]||$ *is the length of the vector* $[\boldsymbol{p}]$*.*

Proof. The boundedness of spectrum is implied by the Lipschitz continuity of the numerical flux $\hat{\boldsymbol{h}}(\cdot,\cdot)$. The positive property $[\boldsymbol{p}]^T\mathcal{A}(\hat{\boldsymbol{h}};\boldsymbol{p})[\boldsymbol{p}] \ge 0$ is evident, since $[\boldsymbol{p}]^T\mathcal{A}_i[\boldsymbol{p}] \ge 0$ by the generalized E-flux property (2.6).

If $[\boldsymbol{p}] = \boldsymbol{0}$, the conclusion (3.6) is trivial. Otherwise, we start our proof from the equality

$$\begin{aligned}
[\boldsymbol{p}]^T\mathcal{A}_2[\boldsymbol{p}] &= [\boldsymbol{p}]^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_2)\{\boldsymbol{f}(\boldsymbol{r}_2) - \hat{\boldsymbol{h}}(\boldsymbol{p})\} \\
&= [\boldsymbol{p}]^T\boldsymbol{v}_{,\boldsymbol{u}}(\bar{\boldsymbol{p}})\{\boldsymbol{f}(\boldsymbol{r}_2) - \boldsymbol{f}(\boldsymbol{r}_j)\} + [\boldsymbol{p}]^T\{\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_2) - \boldsymbol{v}_{,\boldsymbol{u}}(\bar{\boldsymbol{p}})\}\{\boldsymbol{f}(\boldsymbol{r}_2) - \boldsymbol{f}(\boldsymbol{r}_j)\} \\
&\quad + [\boldsymbol{p}]^T\{\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_2) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_j)\}(\boldsymbol{f}(\boldsymbol{r}_j) - \hat{\boldsymbol{h}}(\boldsymbol{p})) + [\boldsymbol{p}]^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_j)(\boldsymbol{f}(\boldsymbol{r}_j) - \hat{\boldsymbol{h}}(\boldsymbol{p}))
\end{aligned}$$

and proceed to estimate each term on the right-hand side separately. In this equality $\boldsymbol{r}_2 = \bar{\boldsymbol{p}}$, and $\boldsymbol{r}_j$ will be taken as $j = 1$ and 3, namely $\boldsymbol{p}^+$ and $\boldsymbol{p}^-$.

It is easy to see that the absolute values of the middle two terms are bounded by $\mathcal{O}(|||[\boldsymbol{p}]|||^3)$, since $\boldsymbol{f}$, $\hat{\boldsymbol{h}}$, and $\boldsymbol{v}_{,\boldsymbol{u}}$ are all Lipschitz continuous, and that the last term is nonnegative by the generalized E-flux property (2.6).

We estimate the first term by simple Taylor expansions up to second order. We remark that the expansions are performed along the line with endpoints $\boldsymbol{p}^{\pm}$, i.e., for the single-variable function $\tilde{\boldsymbol{f}}(s) = \boldsymbol{f}(s\boldsymbol{p}^+ + (1-s)\boldsymbol{p}^-)$, where $s \in [0,1]$. It is obvious that $\boldsymbol{f}(\boldsymbol{p}^+) = \tilde{\boldsymbol{f}}(1)$, $\boldsymbol{f}(\bar{\boldsymbol{p}}) = \tilde{\boldsymbol{f}}(\frac{1}{2})$, and $\boldsymbol{f}(\boldsymbol{p}^-) = \tilde{\boldsymbol{f}}(0)$. Together with $[\boldsymbol{p}]^T \mathcal{A}_i [\boldsymbol{p}] \geq 0$ for $i = 1, 3$, the Taylor expansions at the point $\bar{\boldsymbol{p}}$ finally give that

$$[\boldsymbol{p}]^T \mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{p})[\boldsymbol{p}] \geq \pm \frac{1}{3} [\boldsymbol{p}]^T \boldsymbol{v}_{,\boldsymbol{u}}(\bar{\boldsymbol{p}}) \boldsymbol{f}_{,\boldsymbol{v}}(\boldsymbol{v}(\bar{\boldsymbol{p}})) \boldsymbol{v}_{,\boldsymbol{u}}(\bar{\boldsymbol{p}})[\boldsymbol{p}] - C_\star |||[\boldsymbol{p}]|||^3.$$

Due to the Lipschitz property of $\boldsymbol{f}$ and $\hat{\boldsymbol{h}}$, each component of $\mathcal{A}(\boldsymbol{h}; \bar{\boldsymbol{p}} - \frac{1}{2}\boldsymbol{r}, \bar{\boldsymbol{p}} + \frac{1}{2}\boldsymbol{r})$ is Lipschitz continuous with respect to the variable $\boldsymbol{p}$, on the m-dimensional sphere with diameter $|||[\boldsymbol{p}]|||$. Therefore we have

$$\frac{1}{3} \left| \boldsymbol{r}^T \mathbb{J}(\bar{\boldsymbol{p}}) \boldsymbol{r} \right| \leq \boldsymbol{r}^T \mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{p}) \boldsymbol{r} + C_\star |||[\boldsymbol{p}]|||^3 \quad \forall \boldsymbol{r} : ||\boldsymbol{r}|| = \frac{1}{2} |||[\boldsymbol{p}]|||.$$

Thus it is easy to conclude the inequality (3.6). This completes the proof of this proposition. ∎

Remark 3.1. The inequality (3.6) also holds for other numerical flux functions which may violate the generalized E-flux property (2.6) slightly, e.g., the Harten–Hyman flux function [12] and the local Lax–Friedrichs flux function with an entropy fix (cf. [16]). For these fluxes, the deviation to (2.6) is of the order $\mathcal{O}(|||[\boldsymbol{p}]|||_\infty^4)$, which does not affect the proof of inequality (3.6).

In what follows we will adopt some convenient notations about the matrix $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{p})$. If there is no confusion, we will, for any vector-valued function $\boldsymbol{p}$, denote

$$A(\boldsymbol{p}) = \sum_{1 \leq j \leq N} [\boldsymbol{p}]^T_{j+\frac{1}{2}} \mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{p})_{j+\frac{1}{2}} [\boldsymbol{p}]_{j+\frac{1}{2}}.$$

We will also denote by $\varrho(\boldsymbol{p})$ the maximum of the spectral radius of $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{p})$ over all element interfaces. As a result of Proposition 3.1, we have that each spectrum of $|\mathbb{J}(\boldsymbol{p})|$ is bounded by $3\varrho(\boldsymbol{p}) + C_\star |||[\boldsymbol{p}]|||$. Also, we have that each spectrum is "almost positive"; i.e., if the spectrum is less than zero, then its absolute value must be bounded by $C_\star |||[\boldsymbol{p}]|||$.

3.4. General estimates for the operators $\mathcal{L}$ and $\mathcal{K}$. In this subsection we present a few general inequalities with regard to the operators $\mathcal{L}$ and $\mathcal{K}$ for any test function. They will be used in the next section to estimate the error resulting from the second-order Runge–Kutta time discretization. We remark that all estimates given in this subsection hold for the finite element space $\mathbb{V}_h$ with any degree k.

By subtracting the error equation (3.2a) from (3.2b), for any $1 \leq j \leq N$ we have that

$$\int_{I_j} \boldsymbol{z}_h^T \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^n)\, dx = \frac{1}{2}(\mathcal{L}_j^n - \mathcal{K}_j^n)(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{z}_h) \quad \forall\, \boldsymbol{z}_h \in [\mathbb{P}^k(I_j)]^m, \tag{3.7}$$

and consequently $||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^n||_n^2 = \frac{1}{2}(\mathcal{L}^n - \mathcal{K}^n)(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^n))$. It is therefore natural to start the analysis with an estimate of the difference between $\mathcal{L}^n$ and $\mathcal{K}^n$ for

an arbitrary test function. The main ingredient in the analysis is to obtain a sharp bound for the errors occurring on the element interfaces.

For the following two lemmas, there are only minor modifications from the scalar case [24] to the system case in the analysis, so we will present here only the estimates without proof. The main modification is that the Taylor expansions are changed from single variable to multiple variables, where the derivative $f'(u)$ and its maximum magnitude are replaced by the Jacobian $\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u})$ and its maximum spectral radius, respectively. The definition (3.5) on each element interface also yields a few differences, resulting from a new choice of the reference vector for system case, i.e., $\frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^-) + \frac{2}{3}\boldsymbol{f}(\bar{\boldsymbol{u}}_h) + \frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^+)$, instead of $f(\bar{u}_h)$ in [24] for the scalar case.

LEMMA 3.1. *Assume that the interpolation property* (3.4) *and Proposition* 3.1 *hold. Given any small positive constant ε, we have for any $\boldsymbol{z}_h \in \mathbb{V}_h$ that*

$$\begin{aligned}(\mathcal{L}^n - \mathcal{K}^n)&(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{z}_h)\\ &\le \varepsilon||\boldsymbol{z}_h||_n^2 + M_\varepsilon\tau^2h^{-1}\varrho(\boldsymbol{u}_h^n)A(\boldsymbol{u}_h^n) + M_\varepsilon\tau^2h^{-1}\varrho(\boldsymbol{w}_h^n)A(\boldsymbol{w}_h^n)\\ &+ (C_\star\tau^2h^{-2}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + (C_\star\tau^2h^{-2}||\boldsymbol{e}_{\boldsymbol{w}}^n||_\infty^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{w}}^n||_n^2\\ &+ C(\Xi(n)h^{2k+2}\tau + \tau^6) - \tau\textstyle\sum_{1\le j\le N}\int_{I_j}\boldsymbol{z}_h^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}\,dx,\end{aligned}$$

where $\Xi(n) = 1 + C_\star h^{-1}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2 + C_\star h^{-1}||\boldsymbol{e}_{\boldsymbol{w}}^n||_\infty^2$, the positive constants C and $C_\star$ are independent of n, h, τ and the approximate solutions, and $M_\varepsilon = O(\varepsilon^{-1})$ depends on ε solely.

Similarly, we can get the following lemma to estimate $\mathcal{K}(\cdot)$ by using Taylor expansions of $\boldsymbol{f}(\boldsymbol{u})$ up to second- and third-order derivatives, respectively.

LEMMA 3.2. *Under the assumption of Lemma* 3.1*, we have, for any $\boldsymbol{z}_h \in \mathbb{V}_h$ and any small positive constant ε, the following estimates:*

$$\begin{aligned}\mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{z}_h) \le{}& \varepsilon||\boldsymbol{z}_h||_n^2 + M_\varepsilon\tau^2h^{-1}\varrho(\boldsymbol{u}_h)A(\boldsymbol{u}_h^n) + (C_\star h^{-2}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2\tau^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2\\ &+ (C + C_\star||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2)h^{2k+1}\tau - \tau\sum_{1\le j\le N}\int_{I_j}\boldsymbol{z}_h^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}\,dx,\\ \mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{z}_h) \le{}& \varepsilon||\boldsymbol{z}_h||_n^2 + C\tau^2h^{-2}||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + Ch^{2k}\tau^2,\end{aligned}$$

where the positive constants C and $C_\star$ are independent of n, h, τ and the approximate solutions and where $M_\varepsilon = O(\varepsilon^{-1})$ depends solely on ε.

From the error equation (3.2a) we get the identity $||\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 = \mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n))$. Thus we can take the test function $\boldsymbol{z}_h = \boldsymbol{\xi}_w^n - \boldsymbol{\xi}_u^n$ in the second inequality of Lemma 3.2 and choose the positive constant ε small enough to obtain the following corollary.

COROLLARY 3.1. *Under the assumption of Lemma* 3.1*, we have that*

$$||\boldsymbol{\xi}_{\boldsymbol{w}}^n|| \le C(||\boldsymbol{\xi}_{\boldsymbol{u}}^n|| + h^k\tau) \quad \forall n : n\tau < T \tag{3.8}$$

if the general time-step condition $\tau = \mathcal{O}(h)$ is satisfied, where the positive constant C is independent of n, h, τ and the approximate solutions.

4. Proof of the convergence theorem. In this section we are going to prove only the error estimate (2.7) of the RKDG method with finite element space of piecewise linear polynomials ($k = 1$). The generalization to high order ($k > 1$) with a more restrictive CFL condition is straightforward, along the same line as [24]. To this end, we will analyze each term on the right-hand side of the energy equation (3.3a) separately.

4.1. Estimates to each term on the right-hand side of the energy equation. We will estimate the first three terms in the next two lemmas. First, we look at the last term $\mathcal{E}^n$. By Young's inequality, it is easy to get that

$$\mathcal{E}^n \le \varepsilon ||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}||_n^2 \tau + C(||\boldsymbol{\xi}_{\boldsymbol{u}}^{n}||_n^2 + ||\boldsymbol{\xi}_{\boldsymbol{w}}^{n}||_n^2)\tau, \tag{4.1}$$

where ε is a suitably small positive constant, since each component of $\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{w}_c^n)$ is of order $\mathcal{O}(\tau)$ from the definition (3.1) of $\boldsymbol{w}(x,t)$.

LEMMA 4.1. *Let $\mathbb{V}_h$ be the space of piecewise linear polynomials ($k=1$). If the interpolation approximation property* (3.4) *and the time-step condition $\tau = \mathcal{O}(h)$ are satisfied, then we have the following estimate:*

$$\begin{aligned} ||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^{n}||_n^2 \le & C(\Xi(n)h^3\tau + \tau^6) + \delta_1(n)A(\boldsymbol{u}_h^n)\tau + \delta_2(n)A(\boldsymbol{w}_h^n)\tau \\ & + \left\{ \frac{C_\star \tau^2}{h^2}(||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2 + ||\boldsymbol{e}_{\boldsymbol{w}}^n||_\infty^2) + \frac{C_\star \tau^4}{h^4}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2 + C\tau^2 \right\} ||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2, \end{aligned} \tag{4.2}$$

where the positive constants C and $C_\star$ are independent of n, h, τ and the numerical solutions and where $\Xi(n)$ has been defined in Lemma 3.1. *Here*

$$\delta_1(n) = \frac{M_1\tau}{h}\varrho(\boldsymbol{u}_h^n) + \frac{M_2\tau^3}{h^3}\varrho(\boldsymbol{u}_h^n)\lambda^2(\boldsymbol{u}^n) \quad \textit{and} \quad \delta_2(n) = \frac{M_3\tau}{h}\varrho(\boldsymbol{w}_h^n), \tag{4.3}$$

where $\lambda(\boldsymbol{u}^n)$ is the maximum spectral radius of $\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}^n)$ on all element interfaces and where M_i, $i = 1, 2, 3$, are positive constants independent of the other parameters in (4.3).

Proof. We follow the analysis framework in [24] and sketch the two main steps in the proof.

First, we successively take two test functions both in (3.7) and in Lemma 3.1: the first one is $\boldsymbol{z}_h = \boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^{n}$, and the second is $\boldsymbol{z}_h = -\tau \boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}$. This is feasible since $\mathbb{V}_h$ is the discontinuous finite element space. By combining the two resulting inequalities and letting each parameter ε be small enough, we can obtain the following estimate:

$$\begin{aligned} ||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1} - \boldsymbol{\xi}_{\boldsymbol{w}}^{n}||_n^2 \le & C(\Xi(n)h^{2k+2}\tau + \tau^6) + M\tau^2 h^{-1}\varrho(\boldsymbol{w}_h^n)A(\boldsymbol{w}_h^n) + M\tau^2 h^{-1}\varrho(\boldsymbol{u}_h^n)A(\boldsymbol{u}_h^n) \\ & + (C_\star \tau^2 h^{-2}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + (C_\star \tau^2 h^{-2}||\boldsymbol{e}_{\boldsymbol{w}}^n||_\infty^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{w}}^n||_n^2 \\ & + M\tau^2 \sum_{1\le j\le N} \int_{I_j} (\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}^T \boldsymbol{v}_{,\boldsymbol{u}}^{1/2}(\boldsymbol{u}_c^n)\mathbb{S}^2(\boldsymbol{u}_c^n)\boldsymbol{v}_{,\boldsymbol{u}}^{1/2}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}\, dx, \end{aligned}$$

under the general time-step condition $\tau = \mathcal{O}(h)$, where $\mathbb{S}(\boldsymbol{u}_c^n) = \boldsymbol{v}_{,\boldsymbol{u}}^{1/2}(\boldsymbol{u}_c^n)\boldsymbol{f}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{u}_c^n))\boldsymbol{v}_{,\boldsymbol{u}}^{1/2}(\boldsymbol{u}_c^n)$, the positive constants C and $C_\star$ are independent of n, h, τ and the approximate solutions, and $M > 0$ is determined solely by the fixed constant ε. We remark that $\mathbb{S}(\boldsymbol{u}_c^n)$ is symmetric and has the same eigenvalues as $\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)$ due to the equality (2.5); therefore the last term in the estimate above is bounded by $M\tau^2\lambda^2(\boldsymbol{u}^n)||(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}||_n^2$.

Next, we need to obtain a sharp estimate to $||(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}||_n^2$ to complete the proof. To this end, we will use the discontinuity property of the finite element space, especially for the piecewise linear polynomials. Because $\boldsymbol{\xi}_{\boldsymbol{u}}^n \in \mathbb{V}_h$ under consideration is a piecewise linear vector-valued function, its derivative $(\boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}$ is a constant vector on each element I_j. Hence, for any vector-valued function $\boldsymbol{p}_h$ (even a function not belonging to $\mathbb{V}_h$) there holds

$$\int_{I_j} (\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h})^T \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}\, dx = 0 \quad (1 \le j \le N, \forall n : n\tau < T), \tag{4.4}$$

where $\widetilde{\boldsymbol{p}_h} = h_j^{-1}\int_{I_j}\boldsymbol{p}_h\,dx$ is the average of $\boldsymbol{p}_h$ on each element I_j; i.e., $\int_{I_j}(\boldsymbol{p}-\boldsymbol{p}_h)\,dx = 0$. This property plays an important role in the proof. Unfortunately it holds only for piecewise linear polynomials and not for higher-order piecewise polynomials. This is why we would need stronger time step restriction for the proof of the higher-order $k > 1$ cases.

It is also worthwhile to note, for any $\boldsymbol{p}_h \in \mathbb{V}_h$, that $\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h}$ and $(\boldsymbol{p}_h)_{,x} = (\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h})_{,x}$ both belong to the finite element space $\mathbb{V}_h$. This property holds for the DG method with any degree k, but not for the standard conforming finite element methods.

We take $\boldsymbol{p}_h = \boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n$ and get $||(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}||_n \le Ch^{-1}||\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h}||_n$ by the inverse inequality (i). After a simple manipulation, we can get from the error equation (3.2a) that

$$||\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h}||_n^2 = (\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n, \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h})) = \mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h})).$$

By taking the test function $\boldsymbol{z}_h = \boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h}$ in the first inequality of Lemma 3.2, we observe that the last integral term becomes 0, owing to the identity (4.4). By choosing ε small enough, we obtain the estimate $||\boldsymbol{p}_h - \widetilde{\boldsymbol{p}_h}||_n^2 \le \mathfrak{M}$, where

$$\mathfrak{M} = M\tau^2 h^{-1}\varrho(\boldsymbol{u}_h^n)A(\boldsymbol{u}_h^n) + (C_\star h^{-2}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2\tau^2 + C\tau^2)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + (C + C_\star||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2)h^3\tau.$$

As a result of this inequality, we have $||(\boldsymbol{\xi}_{\boldsymbol{w}}^n - \boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}||_n^2 \le Ch^{-2}\mathfrak{M}$. This sharp estimate completes the proof of this lemma. ∎

LEMMA 4.2. *Let $\mathbb{V}_h$ be the space of piecewise linear polynomials ($k = 1$). If the interpolation approximation property* (3.4) *is satisfied, then we have the following estimates:*

$$\mathcal{K}^n(\boldsymbol{\xi}_{\boldsymbol{u}}^n) \le \Phi(\boldsymbol{e}_{\boldsymbol{u}}^n)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2\tau - \frac{1}{2}A(\boldsymbol{u}_h^n)\tau + (C + C_\star||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2)h^3\tau, \tag{4.5a}$$

$$\mathcal{L}^n(\boldsymbol{\xi}_{\boldsymbol{w}}^n) \le \Phi(\boldsymbol{e}_{\boldsymbol{w}}^n)||\boldsymbol{\xi}_{\boldsymbol{w}}^n||_n^2\tau - \frac{1}{2}A(\boldsymbol{w}_h^n)\tau + (C + C_\star||\boldsymbol{e}_{\boldsymbol{w}}^n||_\infty^2)h^3\tau + C\tau^5, \tag{4.5b}$$

where the positive constants C and $C_\star$ are independent of n, h, τ and the numerical solutions and where $\Phi(\boldsymbol{e}_{\boldsymbol{p}}^n) = C + C_\star h^{-1}||\boldsymbol{e}_{\boldsymbol{p}}^n||_\infty^2$ for $\boldsymbol{p} = \boldsymbol{u}$ or $\boldsymbol{w}$.

Proof. We will prove only (4.5a) here, since the proof to (4.5b) is similar.

Noticing the periodic or zero (compactly supported) boundary conditions, after some elementary calculations we have an equivalent form of $\mathcal{K}^n(\boldsymbol{\xi}_{\boldsymbol{u}}^n)$. It reads

$$\begin{aligned}
\mathcal{K}^n(\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\boldsymbol{\xi}_{\boldsymbol{u}}^n) \equiv \sum_{j=1}^N \mathcal{K}_j^n(\boldsymbol{\xi}_{\boldsymbol{u}}^n) &:= \Pi_1 + \Pi_2 + \Pi_3 + \Pi_4 + \Pi_5 \\
&= \sum_{1\le j\le N}\int_{I_j}(\boldsymbol{\xi}_{\boldsymbol{u}}^n)^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)(\boldsymbol{\eta}_{\boldsymbol{w}}^n - \boldsymbol{\eta}_{\boldsymbol{u}}^n)\,dx \\
&+ \tau\sum_{1\le j\le N}\int_{I_j}(\boldsymbol{\xi}_{\boldsymbol{u}}^n)_{,x}^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)\big(\boldsymbol{f}(\boldsymbol{u}^n) - \boldsymbol{f}(\boldsymbol{u}_h^n)\big)\,dx \\
&+ \tau\sum_{1\le j\le N}\Big\{\big[\boldsymbol{\xi}_{\boldsymbol{u}}^n\big]^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b^n)\big(\boldsymbol{f}(\boldsymbol{u}^n) - \boldsymbol{f}_{Ref}\big)\Big\}_{j+\frac{1}{2}} \\
&+ \tau\sum_{1\le j\le N}\Big\{\big[\boldsymbol{\xi}_{\boldsymbol{u}}^n\big]^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b^n)\big(\boldsymbol{f}_{Ref} - \hat{\boldsymbol{h}}(\boldsymbol{u}_h^n)\big)\Big\}_{j+\frac{1}{2}} \\
&+ \tau\sum_{1\le j\le N}\Big\{\big((\boldsymbol{\xi}_{\boldsymbol{u}}^{n,+})^T\mathbb{E}_b^{n,+} - (\boldsymbol{\xi}_{\boldsymbol{u}}^{n,-})^T\mathbb{E}_b^{n,-}\big)\big(\boldsymbol{f}(\boldsymbol{u}^n) - \hat{\boldsymbol{h}}(\boldsymbol{u}_h^n)\big)\Big\}_{j+\frac{1}{2}},
\end{aligned}$$

where $\boldsymbol{f}_{Ref} = \frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^{-,n}) + \frac{2}{3}\boldsymbol{f}(\bar{\boldsymbol{u}}_h^n) + \frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^{+,n})$ is the reference vector defined on each element interface. As we have mentioned before, this is different from the original reference $f(\bar{u}_h^n)$ for the scalar case [24]. We will see in the appendix the usage of this reference vector. Here $\mathbb{E}_b^{n,\pm} = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_{b\pm1/2}^n) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b^n)$, where the subscript b is used to emphasize the evaluation on the element interface, namely, $\boldsymbol{u}_b^n = \boldsymbol{u}_{j+1/2}^n$. It follows from the smoothness assumption of $\boldsymbol{v}_{,\boldsymbol{u}}$ and $\boldsymbol{u}$ that each component of $\mathbb{E}_b^{n,\pm}$ is of the order $\mathcal{O}(h)$.

Below we will follow [24] and separately analyze each term of the above equality.

By the interpolation approximation property (3.4) of the finite element space $\mathbb{V}_h$ and Young's inequality, it is easy to estimate Π_1 and Π_5 in the form

$$\Pi_1 + \Pi_5 \le Ch^4\tau + C||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2\tau, \tag{4.6a}$$

where the inverse property (iii), the Lipschitz property of $\hat{\boldsymbol{h}}$, and the fact that $\mathbb{E}_b^{n,\pm}$ is of the order $\mathcal{O}(h)$ are also used.

We would like to use Taylor expansions up to third order and estimate Π_2 and Π_3 together. In this step, we have to overcome the difficulties resulting from the symmetrizable assumption of the system (1.1). We present only the result here and postpone the technical analysis to the appendix. The final estimate reads

$$\Pi_2 + \Pi_3 \le \Phi(\boldsymbol{e}_{\boldsymbol{u}}^n)||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2\tau + \{C + C_\star||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2\}h^3\tau. \tag{4.6b}$$

We can estimate the last term Π_4 by virtue of the matrix $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{u}_h^n)$ (see Proposition 3.1) and the following two properties: one is $[\boldsymbol{e}_{\boldsymbol{u}}^n] = -[\boldsymbol{u}_h^n]$ from the continuity of $\boldsymbol{u}^n$, and consequently $[\boldsymbol{\xi}_{\boldsymbol{u}}^n] = [\boldsymbol{\eta}_{\boldsymbol{u}}^n] - [\boldsymbol{u}_h^n]$; the other is the definition of $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{u}_h^n)$ and the identities $\boldsymbol{f}(\boldsymbol{r}_i^n) - \hat{\boldsymbol{h}}(\boldsymbol{u}_h^n) = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_i^n)^{-1}\mathcal{A}_i[\boldsymbol{u}_h^n]$, $i = 1, 2, 3$, where $\boldsymbol{r}_i^n$ is $\boldsymbol{u}_h^{n,-}, \bar{\boldsymbol{u}}_h^n$, and $\boldsymbol{u}_h^{n,+}$, respectively. We would also like to mention that $\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{s}_i^n) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b^n)$ is of the order $\mathcal{O}(||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty)$ by the smoothness of the mapping $\boldsymbol{v}(\boldsymbol{u})$ and the exact solution $\boldsymbol{u}$. After separating Π_4 into six parts as in [24], finally we obtain

$$\Pi_4 \le -\frac{1}{2}A(\boldsymbol{u}_h^n)\tau + C_\star h^{-1}||\boldsymbol{e}_{\boldsymbol{u}}^n||_\infty^2||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + Ch^3\tau, \tag{4.6c}$$

by Young's inequality, the properties of the finite element space $\mathbb{V}_h$, and the boundedness of $\boldsymbol{v}_{,\boldsymbol{u}}, \boldsymbol{u}_{,\boldsymbol{v}}$, and $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{u}_h^n)$.

Now we can get the estimate (4.5a) by summing up all of the estimates (4.6) and complete the proof of this lemma. ∎

4.2. Proof of the convergence theorem. In this subsection we will prove the convergence Theorem 2.1 for the $k = 1$ case using the estimates obtained in subsection 4.1. In addition, we would need to use the a priori technique below.

To deal with the nonlinearity of the flux function $\boldsymbol{f}(\boldsymbol{u})$, we assume a priori that for h small enough there holds

$$||\boldsymbol{u}^n - \boldsymbol{u}_h^n|| \le h. \tag{4.7}$$

This is obviously satisfied for $n = 0$ by $\boldsymbol{u}_h^0 = \mathbb{P}_h\boldsymbol{u}_0(x)$ and the interpolation approximation property (3.4a). We shall later verify the correctness of (4.7) and prove that it still holds true for $n + 1$ if it holds true for a given n. For a linear flux function $\boldsymbol{f} = \mathbb{C}\boldsymbol{u}$, where $\mathbb{C}$ is a constant matrix, this a priori assumption is unnecessary.

It follows that $||\boldsymbol{w}^n - \boldsymbol{w}_h^n|| \le Ch$ from the a priori assumption (4.7) and Corollary 3.1. Then the inverse inequality (iii), together with the approximation property (3.4a) of $\mathbb{V}_h$, implies that

$$||\boldsymbol{e}_{\boldsymbol{p}}^n||_\infty \le Ch^{1/2}, \quad \boldsymbol{p} = \boldsymbol{u}, \boldsymbol{w}. \tag{4.8}$$

By combining all the results in subsection 4.1, together with Corollary 3.1 and (4.8), we can get from the energy equation (3.3) that for h small enough there holds

$$\begin{aligned} &||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}||_n^2 - ||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2 + \frac{1}{2}A(\boldsymbol{u}_h^n)\tau + \frac{1}{2}A(\boldsymbol{w}_h^n)\tau \\ &\le \varepsilon||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}||_n^2\tau + C(||\boldsymbol{\xi}_{\boldsymbol{u}}^n||_n^2\tau + h^3\tau + \tau^5) + \delta_1(n)A(\boldsymbol{u}_h^n)\tau + \delta_2(n)A(\boldsymbol{w}_h^n)\tau, \end{aligned} \tag{4.9}$$

under a suitable CFL condition $\tau \le \beta h$, where the CFL number β will be determined later. Here, ε is an arbitrary positive constant, and C is a positive constant independent of n, h, τ and the approximate solutions.

The number β can be determined by, for example, $\delta_1(n) \le 1/4$ and $\delta_2(n) \le 1/4$. We would like to mention again that those positive constants that emerged in (4.3), namely M_1, M_2, and M_3, are independent of h and τ. Hence there exists a maximum positive constant r_0 also independent h and τ such that

$$M_1 r_0 \le \frac{1}{8}, \quad M_2 r_0^3 \le \frac{1}{8}, \quad \text{and} \quad M_3 r_0 \le \frac{1}{4}.$$

In the numerical simulation, we often determine each time step τ^n by $\tau^n \le \beta^n h$, where

$$\beta^n = r_0 \min\{\, \varrho(u_h^n)^{-1}, \varrho(w_h^n)^{-1}, (\varrho(u_h^n)\lambda^2(u^n))^{-1/3} \,\}. \tag{4.10}$$

However, in this paper we assume for convenience that the time step is constant τ; hence we write the CFL condition as $\tau \le \beta h$ instead of $\tau^n \le \beta^n h$, where $\beta = \min_{\forall n: n\tau \le T} \beta^n$.

Under the above CFL condition $\tau \le \beta h$, the summation of the inequality (4.9) over n, when ε is suitably small, yields that

$$||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}||^2 + \sum_{0\le m\le n} A(\boldsymbol{u}_h^m)\tau + \sum_{0\le m\le n} A(\boldsymbol{w}_h^m)\tau \le C\left(\sum_{0\le m\le n} ||\boldsymbol{\xi}_{\boldsymbol{u}}^m||^2\tau + h^3 + \tau^4\right),$$

where we use the equivalence of $||\cdot||_n$ and $||\cdot||$ and the fact that $\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^{n+1}) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c^n)$ is of order $\mathcal{O}(\tau)$. Thus by Gronwall's inequality we can get the following error estimate:

$$||\boldsymbol{\xi}_{\boldsymbol{u}}^{n+1}|| \le C(h^{3/2} + \tau^2) \quad \forall n:\ n\tau \le T, \tag{4.11}$$

where the positive constant C is independent of n, h, τ and the numerical solutions. Then we can easily get the estimate (2.7) for $k = 1$ by the triangle inequality and the interpolation approximation property (3.4a).

Finally let us verify the a priori assumption (4.7) before we complete the proof of Theorem 2.1 for $k = 1$. If assumption (4.7) is satisfied for a certain n, then it follows from (4.11) and the interpolation approximation property (3.4a) that it is also true for $n + 1$. Thus the given a priori assumption (4.7) is reasonable, and all of the estimates above hold for all $n:\ n\tau \le T$.

Remark 4.1. It is worth noting that the condition (4.10) is the usual CFL condition for systems of conservation laws. By Proposition 3.1 we know that for any numerical flux function $\hat{\boldsymbol{h}}$ the spectral radius of $\mathcal{A}(\hat{\boldsymbol{h}}; \boldsymbol{u}_h^n)$ is bounded by a constant times the maximum of the Lipschitz constant of $\hat{\boldsymbol{h}}$. For example, for the linear flux $f = \mathbb{C}u$ the CFL number β determined by (4.10) depends solely on the spectral radius of $\mathbb{C}$. This also explains why the CFL constant β is lower bounded away from zero during a mesh refinement.

Remark 4.2. We have carried out the error estimate and proof for the linear finite element $k = 1$ only in the one dimensional case with generalized E-flux functions. The estimate (2.7) also holds for the linear flux function $f = \mathbb{C}u$ in multiple dimensions, when the a priori assumption (4.7) is unnecessary.

For a higher-order finite element space $k > 1$ in d-dimensional space, the above analysis framework still works for symmetrizable nonlinear systems if we assume $d < 2k - 1$ to ensure an a priori assumption stronger than (4.7), e.g., $||\boldsymbol{u}^n - \boldsymbol{u}_h^n|| \le h^{1+\frac{d}{2}}$. In this case we use Taylor expansion only up to second order. Though Lemma 4.1 holds only for piecewise linear polynomials, we can similarly prove the estimate (2.7) under a more restrictive time-space condition, for example, $\tau = \mathcal{O}(h^{4/3})$. We remark that this stronger condition is necessary, since the method is linearly unstable under the usual CFL condition $\tau \le \beta h$ for $k \ge 2$. For more details, see [24].

Remark 4.3. For certain special numerical flux functions, we can upgrade the error estimate in Theorem 2.1 to be optimal, i.e., $\mathcal{O}(h^{k+1} + \tau^2)$.

These numerical flux functions include those constructed by the flux-vector splitting method, for example, the upwind numerical flux for a linear flux and the Steger–Warming flux [22] for Euler equations. Their common property is that the vector-valued physical flux function is homogeneous of degree 1; i.e., $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{u}$.

To obtain the optimal error estimate in this case, we would need to use a standard trick in the DG analysis as we have done in [24], which consists of two main ingredients. The first one is the Gauss–Radau projection instead of the local L^2-projection, and the other is the upwind setting of the reference vector on each element interface. All the analysis is carried out in projecting to each eigenvector direction. In this case we can strengthen the a priori assumption and consequently use only Taylor expansion up to second order. We omit the details of the proof, as they are similar to those in [24] for the scalar case.

5. Appendix. In this appendix we complete the convergence proof and give a detailed analysis of the inequality (4.6b), namely,

$$\Pi_2 \equiv \Pi_2(\boldsymbol{\xi_u}) = \tau \sum_{1\le j\le N} \int_{I_j} (\boldsymbol{\xi_u})_{,x}^T \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c)\big(\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{u}_h)\big)\, dx,$$
$$\Pi_3 \equiv \Pi_3(\boldsymbol{\xi_u}) = \tau \sum_{1\le j\le N} \Big\{ \big[\boldsymbol{\xi_u}\big]^T \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b)\big(\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}_{Ref}\big)\Big\}_{j+\frac{1}{2}},$$

where we suppress the subscripts n for clarity and $\boldsymbol{f}_{Ref} = \frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^-) + \frac{2}{3}\boldsymbol{f}(\bar{\boldsymbol{u}}_h) + \frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^+)$ is the reference vector (see Proposition 3.1). Since $\boldsymbol{e_u} = \boldsymbol{\xi_u} - \boldsymbol{\eta_u}$, we have $\Pi_i = \Pi_i(\boldsymbol{\eta_u}) + \Pi_i(\boldsymbol{e_u})$ for $i = 2, 3$. It is easy to estimate $\Pi_i(\boldsymbol{\eta_u})$, but more technical to estimate $\Pi_i(\boldsymbol{e_u})$.

We can estimate $\Pi_i(\boldsymbol{\eta_u})$, $i = 2, 3$, by the interpolation approximation property and the property of the local L^2-projection. The detailed analysis is very similar to

those in [24], and thus omitted. Finally, this result reads

$$\Pi_2(\boldsymbol{\eta_u}) + \Pi_3(\boldsymbol{\eta_u}) \le C(||\boldsymbol{\xi}||_n^2\tau + h^3\tau). \tag{5.1}$$

In what follows we will give the detailed analysis of $\Pi_i(\boldsymbol{e_u})$. For clarity we will suppress the subscripts $\boldsymbol{u}$ and use implied summation on the indices ι, κ, and σ. We would like to use the following Taylor expansions up to third order:

$$\begin{aligned}
\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{u}_h) &= \boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{e} - \frac{1}{2}\boldsymbol{e}^T\boldsymbol{f}_{,\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{e} + \frac{1}{6}\{\boldsymbol{f}(\boldsymbol{u}^\star)\}_{,u_\iota u_\kappa u_\sigma}\boldsymbol{e}_\iota\boldsymbol{e}_\kappa\boldsymbol{e}_\sigma \\
&= \boldsymbol{m}_1 + \boldsymbol{m}_2 + \boldsymbol{m}_3, \\
\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{r}_\natural) &= \boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u}_b)\boldsymbol{e}^\natural - \frac{1}{2}(\boldsymbol{e}^\natural)^T\boldsymbol{f}_{,\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{u}_b)\boldsymbol{e}^\natural + \frac{1}{6}\{\boldsymbol{f}(\boldsymbol{u}_b^\natural)\}_{,u_\iota u_\kappa u_\sigma}\boldsymbol{e}_\iota^\natural\boldsymbol{e}_\kappa^\natural\boldsymbol{e}_\sigma^\natural \\
&= \boldsymbol{n}_1^\natural + \boldsymbol{n}_2^\natural + \boldsymbol{n}_3^\natural
\end{aligned}$$

for x inside each element and x on each element interface, respectively. Here $\boldsymbol{u}^\star$ and $\boldsymbol{u}_b^\natural$ are some mean vectors in expansions; $\boldsymbol{r}_\natural$ points to $\boldsymbol{u}_h^-, \bar{\boldsymbol{u}}_h$, and $\boldsymbol{u}_h^+$; and consequently $\boldsymbol{e}^\natural$ points to $\boldsymbol{e}_h^-, \bar{\boldsymbol{e}}_h$, and $\boldsymbol{e}^+$, for $\natural = 1, 2, 3$, respectively.

Corresponding to the above expansions, we have the equality $\Pi_i(\boldsymbol{e_u}) = \sum_{\alpha=1}^3 \pi_{i\alpha}(\boldsymbol{e})$ for $i = 2$ and 3, where

$$\pi_{3\alpha}(\boldsymbol{e}) = \frac{1}{6}\pi_{3\alpha}^1(\boldsymbol{e}) + \frac{2}{3}\pi_{3\alpha}^2(\boldsymbol{e}) + \frac{1}{6}\pi_{3\alpha}^3(\boldsymbol{e}), \tag{5.2a}$$

and for $\alpha, \natural = 1, 2, 3$,

(5.2b)

$$\pi_{2\alpha}(\boldsymbol{e}) = \tau\sum_{1\le j\le N}\int_{I_j}\boldsymbol{e}_{,x}^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c)\boldsymbol{m}_\alpha\,dx, \qquad \pi_{3\alpha}^\natural(\boldsymbol{e}) = \tau\sum_{1\le j\le N}\left\{[\boldsymbol{e}]^T\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_b)\boldsymbol{n}_\alpha^\natural\right\}_{j+\frac{1}{2}}.$$

In what follows we analyze carefully the terms above in three groups, i.e., $\pi_{2\alpha}(\boldsymbol{e}) + \pi_{3\alpha}(\boldsymbol{e})$, where the symmetrizable property of the system (1.1) plays an important role.

As we have mentioned before, $\mathbb{J}(\boldsymbol{u}) = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u})$ is a symmetric matrix by the symmetrizable theory. A simple integration by parts reveals that

$$\pi_{21}(\boldsymbol{e}) + \pi_{31}(\boldsymbol{e}) = \tau\sum_{1\le j\le N}\left\{\int_{I_j}\boldsymbol{e}_{,x}^T\mathbb{E}^c(\boldsymbol{u})\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{e}\,dx - \frac{1}{2}\int_{I_j}\boldsymbol{e}^T\partial_x\mathbb{J}(\boldsymbol{u})\boldsymbol{e}\,dx\right\},$$

where $\mathbb{E}^c = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u}_c) - \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u})$ with each component of order $\mathcal{O}(h)$. The inverse inequality (i) together with the approximation property (3.4a) shows that $||\boldsymbol{e}_{,x}|| \le Ch^{-1}(||\boldsymbol{e}|| + h^2)$. Then it is easy to see from the formula above that

$$\pi_{21}(\boldsymbol{e}) + \pi_{31}(\boldsymbol{e}) \le C||\boldsymbol{e}||_n^2\tau + h^4\tau. \tag{5.3}$$

The estimate of the second group, i.e., $\pi_{22}(\boldsymbol{e}) + \pi_{32}(\boldsymbol{e})$, is one of the most difficult steps in generalizing the error analysis from the scalar equation to symmetrizable systems when the finite element space $\mathbb{V}_h$ is made up of piecewise linear polynomials. However, this inconvenience can be obviated for high-order piecewise polynomials, because the Taylor expansion would need to be carried out only to the second order if the a priori assumption is strengthened. See Remark 4.2.

Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)^T = \boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{e}$ and denote $\mathbb{G}(\boldsymbol{u}) = \boldsymbol{u}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{u}))\boldsymbol{f}_{,\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{u}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{u}))$. Here $\mathbb{G}$ is a supermatrix with the component $\mathbb{G}^{i_3}_{i_1,i_2} = \{\boldsymbol{u}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{u}))\boldsymbol{f}^{i_3}_{,\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{u}_{,\boldsymbol{v}}(\boldsymbol{v}(\boldsymbol{u}))\}_{i_1,i_2}$ at the (i_1, i_2, i_3)th position. Then it is easy to see that $\pi_{22}(\boldsymbol{e}) + \pi_{32}(\boldsymbol{e}) = \mathcal{R} + \mathcal{Q} + \mathcal{S}$, where

$$\mathcal{R} = -\frac{\tau}{2}\sum_{j=1}^{N}\int_{I_j} \boldsymbol{g}^T_{,x}\boldsymbol{g}^T\mathbb{G}(\boldsymbol{u})\boldsymbol{g}\,dx,$$

$$\mathcal{Q} = -\frac{\tau}{2}\sum_{j=1}^{N}[\boldsymbol{g}]^T_{j+\frac{1}{2}}\left(\frac{1}{6}(\boldsymbol{g}^-)^T\mathbb{G}(\boldsymbol{u}_b)\boldsymbol{g}^- + \frac{2}{3}\bar{\boldsymbol{g}}^T\mathbb{G}(\boldsymbol{u}_b)\bar{\boldsymbol{g}} + \frac{1}{6}(\boldsymbol{g}^+)^T\mathbb{G}(\boldsymbol{u}_b)\boldsymbol{g}^+\right)_{j+\frac{1}{2}},$$

$$\mathcal{S} = -\frac{\tau}{2}\sum_{j=1}^{N}\int_{I_j} \boldsymbol{e}^T_{,x}\mathbb{E}^c\boldsymbol{g}^T\mathbb{G}(\boldsymbol{u})\boldsymbol{g}\,dx + \frac{\tau}{2}\sum_{j=1}^{N}\int_{I_j} \boldsymbol{e}^T\partial_x\{\boldsymbol{v}_{,\boldsymbol{u}}(\boldsymbol{u})\}\boldsymbol{g}^T\mathbb{G}(\boldsymbol{u})\boldsymbol{g}\,dx.$$

Below we will separately analyze the above terms and will use the implied summation for the indices i_1, i_2, i_3, and j (or b, since $b = j + 1/2$) for clarity.

To this end, we would first like to point out that the component $\{\mathbb{G}(\boldsymbol{u})\}^{i_3}_{i_1i_2}$ is invariant for all rotations of the indices i_1, i_2, and i_3, which we also refer to as the symmetric property of the supermatrix $\mathbb{G}$. It is obvious that $\mathbb{G}^{i_3}_{i_1,i_2} = \mathbb{G}^{i_3}_{i_2,i_1}$, so we need only prove $\mathbb{G}^{i_3}_{i_1,i_2} = \mathbb{G}^{i_2}_{i_1,i_3}$ to verify this important property. We start with the definition of $\mathbb{G}$ and consequently the identity

$$\{\mathbb{G}(\boldsymbol{u})\}^{i_3}_{i_1i_2} = \frac{\partial^2 f_{i_3}}{\partial v_{i_1}\partial v_{i_2}} + \sum_{\kappa,\gamma,\sigma,\ell}\frac{\partial f_\kappa}{\partial u_\gamma}\frac{\partial u_\gamma}{\partial v_{i_3}}\frac{\partial u_\sigma}{\partial v_{i_1}}\frac{\partial u_\ell}{\partial v_{i_2}}\frac{\partial^2 v_\kappa}{\partial u_\sigma\partial u_\ell}, \tag{5.4}$$

where we have used the symmetric property of $\boldsymbol{u}_{,\boldsymbol{v}}$. We have mentioned before that in the symmetrizable theory $\mathbb{J} = \boldsymbol{v}_{,\boldsymbol{u}}\boldsymbol{f}_{,\boldsymbol{u}} = \boldsymbol{v}_{,\boldsymbol{u}}\boldsymbol{f}_{,\boldsymbol{v}}\boldsymbol{v}_{,\boldsymbol{u}}$ is a symmetric matrix. Clearly, a vector-valued function $\boldsymbol{c}(\boldsymbol{u})\colon \mathbb{R}^m \to \mathbb{R}^m$ can be found so that $\boldsymbol{c}_{,\boldsymbol{u}} = \boldsymbol{v}_{,\boldsymbol{u}}\boldsymbol{f}_{,\boldsymbol{u}}$, namely

$$\boldsymbol{c}_{\gamma,\sigma} = \boldsymbol{c}_{\sigma,\gamma} = \frac{\partial c_\sigma}{\partial u_\gamma} = \sum_\kappa \frac{\partial v_\sigma}{\partial u_\kappa}\frac{\partial f_\kappa}{\partial u_\gamma} = \sum_\kappa \frac{\partial v_\kappa}{\partial u_\sigma}\frac{\partial f_\kappa}{\partial u_\gamma}.$$

Its derivative with respect to u_ℓ is given by

$$\frac{\partial^2 c_\sigma}{\partial u_\gamma\partial u_\ell} = \frac{\partial \boldsymbol{c}_{\sigma,\gamma}}{\partial u_\ell} = \sum_\kappa \frac{\partial^2 v_\kappa}{\partial u_\sigma\partial u_\ell}\frac{\partial f_\kappa}{\partial u_\gamma} + \sum_\kappa \frac{\partial v_\kappa}{\partial u_\sigma}\frac{\partial^2 f_\kappa}{\partial u_\gamma\partial u_\ell},$$

which implies that $\sum_\kappa \frac{\partial^2 v_\kappa}{\partial u_\sigma\partial u_\ell}\frac{\partial f_\kappa}{\partial u_\gamma}$ is invariant under the exchange of indices γ and ℓ. As a result, the second term on the right-hand side of identity (5.4) is invariant under the exchange of indices i_2 and i_3. Since $\boldsymbol{f}_{,\boldsymbol{v},\boldsymbol{v}}$ is symmetric in the symmetrizable theory, we can assert from (5.4) that $\mathbb{G}^{i_3}_{i_1,i_2} = \mathbb{G}^{i_2}_{i_1,i_3}$. This verifies the symmetric property of the supermatrix $\mathbb{G}$.

By the symmetric property of $\mathbb{G}$, an integration by parts yields that

$$\mathcal{R} = \frac{\tau}{6}\{\mathbb{G}(\boldsymbol{u}_b)\}^{i_3}_{i_1i_2}[g_{i_1}g_{i_2}g_{i_3}]_b + \frac{\tau}{6}\int_I \partial_x\{\mathbb{G}(\boldsymbol{u})\}^{i_3}_{i_1i_2}g_{i_1}g_{i_2}g_{i_3}\,dx.$$

In order to estimate $\mathcal{R}$ and $\mathcal{Q}$ together, we would like to use the following equality:

$$\begin{aligned}\{\mathbb{G}(\boldsymbol{u}_b)\}^{i_3}_{i_1i_2}[g_{i_1}g_{i_2}g_{i_3}]_b &= \{\mathbb{G}(\boldsymbol{u}_b)\}^{i_3}_{i_1i_2}\left([g_{i_1}]g^+_{i_2}g^+_{i_3} + g^-_{i_1}[g_{i_2}]g^+_{i_3} + g^-_{i_1}g^-_{i_2}[g_{i_3}]\right)_b\\ &= \{\mathbb{G}(\boldsymbol{u}_b)\}^{i_3}_{i_1i_2}\left(g^+_{i_1}g^+_{i_2} + \frac{1}{2}g^-_{i_1}g^+_{i_2} + \frac{1}{2}g^-_{i_2}g^+_{i_1} + g^-_{i_1}g^-_{i_2}\right)_b[g^n_{i_3}]_b,\end{aligned}$$

thanks to the symmetric property of $\mathbb{G}$. Then a direct manipulation shows

$$\mathcal{R}+\mathcal{Q}=\frac{\tau}{6}\int_I \partial_x\{\mathbb{G}(\boldsymbol{u})\}_{i_1 i_2}^{i_3} g_{i_1} g_{i_2} g_{i_3}\,dx \le C||\boldsymbol{e}||_\infty||\boldsymbol{e}||^2. \tag{5.5}$$

This is the reason that the reference vector $\boldsymbol{f}_{Ref}$ is taken as $\frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^-)+\frac{2}{3}\boldsymbol{f}(\bar{\boldsymbol{u}}_h)+\frac{1}{6}\boldsymbol{f}(\boldsymbol{u}_h^+)$ in this paper. It avoids the estimate to the term corresponding to the second-order derivatives.

Since each component of $\mathbb{E}^c$ is of order $\mathcal{O}(h)$, it is easy to estimate the last term $\mathcal{S}$ in the second group. The estimate to the last group $\pi_{23}(\boldsymbol{e})+\pi_{33}(\boldsymbol{e})$ is easy to get in a similar way, where we use the inequality $||\boldsymbol{e}||_{\Gamma_h} \le Ch^{-\frac{1}{2}}(||\boldsymbol{e}||+h^2)$ implied by the inverse inequality (iii) and the approximation property (3.4a). Thus we give the final estimate together in the form

$$\mathcal{S}+\pi_{23}(\boldsymbol{e})+\pi_{33}(\boldsymbol{e}) \le C_\star(h^{-1}+1)||\boldsymbol{e}||_\infty^2(||\boldsymbol{e}||^2+h^4)\tau. \tag{5.6}$$

By collecting the above estimates and preprocessing the result using the simple inequality $||\boldsymbol{e}||_\infty \le \frac{1}{2}(h^{-1}||\boldsymbol{e}||_\infty^2+h)$, we can easily get the inequality (4.6b) since $h<1$. We have therefore completed the proof of the convergence theorem.

REFERENCES

[1] P. G. Ciarlet, *Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[2] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[3] B. Cockburn, S.-Y. Lin, and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[4] B. Cockburn and C.-W. Shu, *The Runge–Kutta local projection P^1–discontinuous Galerkin method for scalar conservation laws*, Math. Model. Numer. Anal., 25 (1991), pp. 337–361.

[5] B. Cockburn and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[6] B. Cockburn and C.-W. Shu, *The Runge–Kutta discontinuous Galerkin finite element method for conservation laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[7] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM. J. Numer. Anal., 35 (1998), pp. 2440–2463.

[8] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[9] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection–dominated problems*, in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, (A. Quarteroni, ed.), Lecture Notes in Math. 1697, Springer, New York, 1998, pp. 325–432.

[10] A. Harten, *On the symmetric form of systems of conservation laws with entropy*, J. Comput. Phys., 49 (1983), pp. 151–164.

[11] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[12] A. Harten and J. M. Hyman, *Self-adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comput. Phys., 50 (1983), pp. 235–269.

[13] S.-M. Hou and X.-D. Liu, *Solutions of multidimensional hyperbolic systems of conservation laws by square entropy condition satisfying discontinuous Galerkin method*, J. Sci. Comput., to appear.

[14] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[15] P. LeSaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations C. de Boor, ed., Academic Press, New York, 1974, pp. 89–145.
[16] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.
[17] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM. J. Numer. Anal., 21 (1984), pp. 217–235.
[18] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.
[19] P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.
[20] C. W. Schulz-Rinne, *Classification of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Math. Anal., 24 (1993), pp. 76–88.
[21] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[22] J. L. Steger and R. F. Warming, *Flux vector splitting of the inviscid gasdynamic equations with applications to finite-difference methods*, J. Comput. Phys., 40 (1981), pp. 263–293.
[23] L.-A. Ying, *A second order explicit finite element scheme to multi–dimensional conservation laws and its convergence*, Sci. China Ser. A, 43 (2000), pp. 945–957.
[24] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 641–666.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1721–1734

THE L^2 NORM ERROR ESTIMATES FOR THE DIV LEAST-SQUARES METHOD*

ZHIQIANG CAI† AND JAEUN KU†

Abstract. This paper studies L^2 norm error estimates for the div least-squares method for which the associated homogeneous least-squares functional is equivalent to the $H(\mathrm{div}) \times H^1$ norm for the respective dual and primal variables. Least-squares of this type for the second-order elliptic equations, elasticity, and the Stokes equations are an active area of research, and error estimates in the $H(\mathrm{div}) \times H^1$ norm were previously established. In this paper, we establish optimal L^2 norm error estimates for the primal variable under the minimum regularity requirement through a refined duality argument.

Key words. least-squares method, error estimate, elliptic equations, elasticity, Stokes, incompressible Newtonian flow

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/050636504

1. Introduction.

There is substantial interest in the use of least-squares principles for the approximate solution of partial differential equations with applications in both solid and fluid mechanics. One advantage of the least-squares approach is that the finite element spaces for the individual unknowns may be chosen independently, and thus based on simplicity, availability, and optimality, or may be chosen from the physics of the underlying problem. Moreover, the linear systems of algebraic equations resulting from well-posed least-squares discretizations are always self-adjoint and positive definite.

Many least-squares methods for scalar elliptic equations, elasticity, and the Stokes equations have been proposed and analyzed. The numerical properties depend on the form of the first-order system and the choice of the least-squares norm. Loosely speaking, there are three types of least-squares methods: the inverse approach, the div approach, and the div-curl approach. The inverse approach employs an inverse norm that is further replaced with either the weighted mesh-dependent norm (see [2]) or the discrete H^{-1} norm (see [4]) for computational feasibility. The div approach uses the L^2 norm, and the corresponding homogeneous least-squares functional is equivalent to the $H(\mathrm{div}) \times H^1$ norm. The homogeneous least-squares functional from the div-curl approach is equivalent to the $H(\mathrm{div}) \cap H(\mathrm{curl})$ norm for some variables.

The purpose of this paper is to study the L^2 norm error estimates for the div least-squares method. For the scalar elliptic equations, the div approach based on the flux-pressure formulation has been studied extensively (see, e.g., [3, 6, 9, 10, 12, 13, 14, 15, 16, 17, 18]). The pressure and the flux are referred to as the primal and dual variables, respectively. For elasticity and the Stokes equations, we recently proposed and analyzed the div least-squares approach in [7, 8] based on the stress-

*Received by the editors July 21, 2005; accepted for publication (in revised form) May 26, 2006; published electronically September 15, 2006. This work was supported in part by the U.S. Department of Energy University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48, by the National Science Foundation under grant DMS-0511430, and by the Korea Research Foundation under grant KRF-2002-070-C00014.

http://www.siam.org/journals/sinum/44-4/63650.html

†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (zcai@math.purdue.edu, jku@math.purdue.edu).

displacement and the stress-velocity formulations, respectively. The displacement and the velocity are referred to as the primal variables, and the stress as the dual variable. It was proved that the homogeneous least-squares functional is equivalent to the $H(\mathrm{div}) \times H^1$ norm. This immediately leads to the error estimate in such a norm. This estimate also yields an optimal L^2 norm error estimate for the dual variable when using appropriate approximation spaces. In this paper, we establish optimal error estimates for the primal variable in the L^2 norm under the minimum regularity requirement. As usual, this estimate is obtained through a duality argument (or the so-called Aubin–Nitche trick for the standard Galerkin finite element method). The key step of the duality argument presented in this paper is to express the L^2 norm error of the primal variable in terms of the bilinear form (see Lemma 5.1). While this is straightforward for the Galerkin method, it is less obvious for the div least-squares method applied to the elasticity and Stokes equations.

Previously, L^2 norm error estimates for the div least-squares method applied to the scalar elliptic problems were studied by several researchers. The optimal L^2 norm error estimate under the minimum regularity assumption was proved for the Poisson equation in [14] (see also [3]). By using extra regularity and special finite element spaces for the flux, an L^2 estimate was obtained in [20] for a divergence form of scalar elliptic problems.

The paper is organized as follows. Section 2 introduces the second-order scalar elliptic partial differential equations, elasticity, and the Stokes equations. The div least-squares formulation and a finite element approximation are described in sections 3 and 4, respectively. In Section 5, we establish optimal L^2 error estimates through the duality argument.

1.1. Notation. We use the standard notation and definitions for the Sobolev spaces $H^s(\Omega)^d$ and $H^s(\partial\Omega)^d$ for $s \geq 0$. The standard associated inner products are denoted by $(\cdot, \cdot)_{s,\Omega}$ and $(\cdot, \cdot)_{s,\partial\Omega}$, and the respective norms are denoted by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\partial\Omega}$. (We suppress the superscript d when the dimension is clear by context. We also omit the subscript Ω from the inner product and norm designation when there is no confusion.) For $s = 0$, $H^s(\Omega)^d$ coincides with $L^2(\Omega)^d$. In this case, the inner product and norm will be denoted by $\|\cdot\|$ and $(\cdot, \cdot)$, respectively. Set

$$H^1_D(\Omega) := \{q \in H^1(\Omega) \,:\, q = 0 \text{ on } \Gamma_D\}.$$

When $\Gamma = \partial\Omega$, denote $H^1_D(\Omega)$ by $H^1_0(\Omega)$. Finally, set

$$H(\mathrm{div}; \Omega) = \{\mathbf{v} \in L^2(\Omega)^d \,:\, \nabla \cdot \mathbf{v} \in L^2(\Omega)\},$$

which is a Hilbert space under the norm

$$\|\mathbf{v}\|_{H(\mathrm{div};\,\Omega)} = \left(\|\mathbf{v}\|^2 + \|\nabla \cdot \mathbf{v}\|^2\right)^{\frac{1}{2}},$$

and define its subspace

$$H_N(\mathrm{div}; \Omega) = \{\mathbf{v} \in H(\mathrm{div}; \Omega) \,:\, \mathbf{n} \cdot \mathbf{v} = 0 \text{ on } \Gamma_N\}.$$

2. Mathematical equations. Let Ω be a bounded, open, connected subset of $\Re^d$ ($d = 2$ or 3) with a Lipschitz continuous boundary $\partial\Omega$. Denote by $\mathbf{n} = (n_1, \ldots, n_d)$ the outward unit vector normal to the boundary. We partition the boundary of the domain Ω into two open subsets Γ_D and Γ_N such that $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. For simplicity, we assume that Γ_D is not empty (i.e., $\mathrm{mes}\,(\Gamma_D) \neq 0$). Otherwise, solutions of partial differential equations considered in this paper are unique up to an additive constant.

2.1. Second-order elliptic problems. Consider the second-order elliptic boundary value problem

$$-\nabla \cdot (A\nabla u) + Xu = f \quad \text{in } \Omega \tag{2.1}$$

with boundary conditions

$$u = 0 \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot A\nabla u = 0 \quad \text{on } \Gamma_N, \tag{2.2}$$

where the symbols $\nabla\cdot$ and ∇ stand for the divergence and gradient operators, respectively; A is a given $d \times d$ tensor function; X is an at most first-order linear differential operator; and f is a given scalar function. Assume that A is uniformly symmetric positive definite; then there exist positive constants $0 < \Lambda_0 \le \Lambda_1$ such that

$$\Lambda_0 \boldsymbol{\xi}^T \boldsymbol{\xi} \le \boldsymbol{\xi}^T A \boldsymbol{\xi} \le \Lambda_1 \boldsymbol{\xi}^T \boldsymbol{\xi}$$

for all $\boldsymbol{\xi} \in \Re^d$ and almost all $x \in \overline{\Omega}$. Here and hereafter, we assume homogeneous boundary conditions for simplicity. The corresponding variational form of system (2.1) is to find $u \in H_D^1(\Omega)$ such that

$$a(u, v) = (f, v) \quad \forall\, v \in H_D^1(\Omega), \tag{2.3}$$

where the bilinear form is defined by

$$a(u, v) = (A\nabla u, \nabla v) + (Xu, v).$$

The dual problem of (2.3) is to find $z \in H_D^1(\Omega)$ such that

$$a(v, z) = (f, v) \quad \forall\, v \in H_D^1(\Omega). \tag{2.4}$$

For simplicity, assume that both problems (2.3) and (2.4) satisfy the full H^2 regularity estimates

$$\|u\|_2 \le C\,\|f\| \quad \text{and} \quad \|z\|_2 \le C\,\|f\|. \tag{2.5}$$

(See section 5.3 for problems with low regularity.) Here and hereafter, we use C with or without subscripts to denote a generic positive constant, that is, a constant that is independent of the mesh size h and the parameter λ introduced in subsequent sections but that may depend on the domain Ω.

2.2. Elasticity and Stokes equations. For a vector field $\mathbf{v} = (v_1, \ldots, v_d)$, define its gradient as a $d \times d$ tensor,

$$\nabla \mathbf{v} = \begin{pmatrix} \dfrac{\partial v_1}{\partial x_1} & \cdots & \dfrac{\partial v_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial v_d}{\partial x_1} & \cdots & \dfrac{\partial v_d}{\partial x_d} \end{pmatrix} = \left(\frac{\partial v_i}{\partial x_j} \right)_{d\times d}.$$

Denote the symmetric part of $\nabla \mathbf{v}$ by

$$\boldsymbol{\epsilon}(\mathbf{v}) = \frac{1}{2}\left(\nabla \mathbf{v} + (\nabla \mathbf{v})^t\right) = \left(\epsilon_{ij}(\mathbf{v})\right)_{d\times d} \quad \text{with} \quad \epsilon_{ij}(\mathbf{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right).$$

For the displacement and the velocity fields $\mathbf{v}$, the symmetric part of the gradients $\boldsymbol{\epsilon}(\mathbf{v})$ is referred to as the strain tensor and the strain rate tensor, respectively. For a tensor function $\boldsymbol{\tau} = (\tau_{ij})_{d\times d}$, let $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{id})$ denote its ith-row for $i = 1, \ldots, d$ and define its divergence, normal, and trace by

$$\nabla \cdot \boldsymbol{\tau} = (\nabla \cdot \boldsymbol{\tau}_1, \ldots, \nabla \cdot \boldsymbol{\tau}_d), \quad \mathbf{n} \cdot \boldsymbol{\tau} = (\mathbf{n} \cdot \boldsymbol{\tau}_1, \ldots, \mathbf{n} \cdot \boldsymbol{\tau}_d), \quad \text{and} \quad \operatorname{tr} \boldsymbol{\tau} = \sum_{i=1}^{d} \tau_{ii},$$

respectively. For tensors $\boldsymbol{\tau} = (\tau_{ij})_{d\times d}$ and $\boldsymbol{\gamma} = (\gamma_{ij})_{d\times d}$, define

$$\boldsymbol{\tau} : \boldsymbol{\gamma} \equiv \sum_{i,j=1}^{d} \tau_{ij}\gamma_{ij}.$$

If $\boldsymbol{\tau}$ is symmetric and $\boldsymbol{\gamma}$ skew-symmetric, it is then easy to check that

$$\boldsymbol{\tau} : \boldsymbol{\gamma} = 0. \tag{2.6}$$

Elasticity and Stokes equations may be cast in the following pressure-perturbed form of the generalized Stokes equations

$$\begin{cases} -\nabla \cdot (2\mu\, \boldsymbol{\epsilon}(\mathbf{u}) - p\, \boldsymbol{\delta}) &= \mathbf{f} \quad \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} + \dfrac{1}{\lambda} p &= g \quad \text{in} \quad \Omega \end{cases} \tag{2.7}$$

with boundary conditions

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot (2\mu\, \boldsymbol{\epsilon}(\mathbf{u}) - p\, \boldsymbol{\delta}) = \mathbf{0} \quad \text{on } \Gamma_N, \tag{2.8}$$

where $\mathbf{u}$ represents the displacement and velocity vector field for solids and fluids, respectively; p represents the pressure; $\boldsymbol{\delta} = (\delta_{ij})_{d\times d}$ denotes the identity tensor; $\mathbf{f}$ is a given vector function; g is a given scalar function; and μ and λ are material constants such that $\mu \in [\mu_1, \mu_2]$ with $0 < \mu_1 < \mu_2$ and $\lambda \in (0, \infty]$. The λ and μ are the so-called Lamé and viscosity constants for solids and fluids, respectively. Materials are said to be incompressible when λ is infinite.

Equation (2.7) with $\lambda = \infty$ gives the Stokes equations

$$\begin{cases} -2\mu\, \nabla \cdot (\boldsymbol{\epsilon}(\mathbf{u})) + \nabla p &= \mathbf{f} \quad \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= g \quad \text{in} \quad \Omega \end{cases} \tag{2.9}$$

with boundary conditions (2.8). When $\Gamma_N = \emptyset$, p is unique up to an additive constant provided that the compatibility condition $\int_\Omega g\, dx = 0$ holds. For $\lambda < \infty$ and $g = 0$, eliminating p in (2.7) and (2.8) yields the elastic equations

$$-2\mu\, \nabla \cdot (\boldsymbol{\epsilon}(\mathbf{u})) - \lambda\, \nabla \nabla \cdot \mathbf{u} = \mathbf{f} \quad \text{in } \Omega \tag{2.10}$$

with boundary conditions

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot (2\mu\, \boldsymbol{\epsilon}(\mathbf{u}) + \lambda\, (\nabla \cdot \mathbf{u})\, \boldsymbol{\delta}) = \mathbf{0} \quad \text{on } \Gamma_N.$$

Again, we assume that both (2.9) and (2.10) satisfy the full H^2 regularity estimates:

$$\|\mathbf{u}\|_2 + \|p\|_1 \le C\, (\|\mathbf{f}\| + \|g\|_1) \quad \text{and} \quad \|\mathbf{u}\|_2 + \lambda\, \|\nabla \cdot \mathbf{u}\|_1 \le C\, \|\mathbf{f}\|, \tag{2.11}$$

respectively, where C is a positive constant independent of $\mathbf{u}$, p, and λ.

3. Least-squares formulations. Least-squares formulations for the scalar elliptic problems, elasticity, and Stokes equations have been extensively studied. They differ in such choices as the first-order system and the least-squares norm. In this paper, we study only the div least-squares method, which applies the simple L^2 norm to the natural first-order system arising from physical laws in both solid and fluid mechanics.

or the velocity for the

3.1. First-order systems. Introducing the flux (vector) variable

$$\sigma = -A\,\nabla u,$$

the scalar elliptic problems in (2.1) may be rewritten as the following first-order partial differential system

$$\left\{\begin{array}{rcll} A^{-1}\sigma + \nabla u &=& \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\sigma + Xu &=& f & \text{in} \quad \Omega \end{array}\right. \tag{3.1}$$

with boundary conditions

$$u = 0 \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n}\cdot\sigma = 0 \quad \text{on } \Gamma_N. \tag{3.2}$$

For the generalized Stokes equations, define the compliance tensor of fourth order, $\mathcal{A}_\lambda$, by

$$\mathcal{A}_\lambda\,\boldsymbol{\tau} = \left\{\begin{array}{ll} \dfrac{1}{2\mu}\left(\boldsymbol{\tau} - \dfrac{\lambda}{d\lambda + 2\mu}(\operatorname{tr}\boldsymbol{\tau})\,\boldsymbol{\delta}\right) & \text{for } \lambda \in (0,\,\infty), \\ \dfrac{1}{2\mu}\left(\boldsymbol{\tau} - \dfrac{1}{d}(\operatorname{tr}\boldsymbol{\tau})\,\boldsymbol{\delta}\right) & \text{for } \lambda = \infty. \end{array}\right.$$

Note that

$$\mathcal{A}_\infty\,\boldsymbol{\tau} = \lim_{\lambda\to\infty}\mathcal{A}_\lambda\,\boldsymbol{\tau}.$$

Without loss of generality, we take $\mu = 1/2$. Denote by $\boldsymbol{\sigma}$ the stress tensor and by $\mathbf{u}$ the displacement for linear elasticity or the velocity for the Stokes equations. Consider the following first-order system of partial differential equations studied in [8, 7]:

$$\left\{\begin{array}{rcll} \mathcal{A}_\lambda\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u}) &=& \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\boldsymbol{\sigma} &=& -\mathbf{f} & \text{in} \quad \Omega \end{array}\right. \tag{3.3}$$

with boundary conditions

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n}\cdot\boldsymbol{\sigma} = \mathbf{0} \quad \text{on } \Gamma_N. \tag{3.4}$$

Variables σ and u for the scalar elliptic problems in (3.1) are vector and scalar functions, respectively, while variables $\boldsymbol{\sigma}$ and $\mathbf{u}$ for the generalized Stokes equations in (3.3) are the respective tensor and vector functions. First-order system (3.3) may be considered the vector form of (3.1), even though they differ substantially in terms of analysis.

3.2. Least-squares variational problems. Define least-squares functionals as the sum of the L^2 norm of the residuals of the first-order systems in (3.1) and (3.3). Then least-squares variational problems are to minimize the least-squares functionals in appropriate solution spaces. To this end, let us first introduce solution spaces. Define the spaces

$$\mathbf{X}_\lambda \equiv \begin{cases} H_N(\mathrm{div};\,\Omega)^d & \text{for} \quad \lambda \in (0,\,\infty), \\ \{\boldsymbol{\tau} \in H_N(\mathrm{div};\,\Omega)^d \mid \int_\Omega \mathrm{tr}\,\boldsymbol{\tau}\,dx = 0\} & \text{for} \quad \lambda = \infty, \end{cases}$$

and let

$$\Sigma = \begin{cases} H_N(\mathrm{div};\Omega) & \text{for (3.1)}, \\ \mathbf{X}_\lambda & \text{for (3.3)}, \end{cases} \qquad U = \begin{cases} H_D^1(\Omega) & \text{for (3.1)}, \\ H_D^1(\Omega)^d & \text{for (3.3)}. \end{cases}$$

Now, for $f \in L^2(\Omega)$ and $\mathbf{f} \in L^2(\Omega)^d$, define the least-squares functional

$$G(\boldsymbol{\sigma},\,\mathbf{u};\,\mathbf{f}) = \begin{cases} \|A^{1/2}(A^{-1}\sigma + \nabla u)\|^2 + \|\nabla\cdot\sigma + Xu - f\|^2 & \text{for (3.1)}, \\ \|\mathcal{A}_\lambda\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2 & \text{for (3.3)} \end{cases} \tag{3.5}$$

for all $(\boldsymbol{\sigma},\,\mathbf{u}) \in \Sigma \times U$. Here and hereafter, we use boldface letters $(\boldsymbol{\sigma},\,\mathbf{u})$ to denote variables for both problems when there is no danger of confusion. Let

$$|||(\boldsymbol{\tau},\,\mathbf{v})||| = \left(\|\mathbf{v}\|_1^2 + \|\boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)}^2\right)^{\frac{1}{2}}.$$

The following theorem was proved in [6] for the scalar elliptic problems in (3.1) and in [8, 7] for the generalized Stokes equations in (3.3).

THEOREM 3.1. *The homogeneous functional $G(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0})$ is equivalent to $|||(\boldsymbol{\tau},\,\mathbf{v})|||^2$; i.e., there exists a positive constant C_1 such that*

$$\frac{1}{C_1}\,|||(\boldsymbol{\tau},\,\mathbf{v})|||^2 \le G(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0}) \le C_1\,|||(\boldsymbol{\tau},\,\mathbf{v})|||^2 \tag{3.6}$$

hold for all $(\boldsymbol{\tau},\,\mathbf{v}) \in \Sigma \times U$. Moreover, the constant C_1 is independent of λ for the generalized Stokes problems.

The variational problem corresponding to the L^2 norm least-squares functional is to minimize functional (3.5) over $\Sigma \times U$, that is, to find $(\boldsymbol{\sigma},\,\mathbf{u}) \in \Sigma \times U$ such that

$$G(\boldsymbol{\sigma},\,\mathbf{u}\,;\,\mathbf{f}) = \inf_{(\boldsymbol{\tau},\,\mathbf{v})\in\Sigma\times U} G(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{f}). \tag{3.7}$$

Let us define a bilinear form $b(\cdot;\,\cdot)$ on $(\Sigma \times U) \times (\Sigma \times U)$ by

$$b(\boldsymbol{\sigma},\mathbf{u};\,\boldsymbol{\tau},\mathbf{v}) = \begin{cases} (\sigma + A\nabla u,\,A^{-1}\tau + \nabla v) + (\nabla\cdot\sigma + Xu,\,\nabla\cdot\tau + Xv) & \text{for (3.1)}, \\ (\mathcal{A}_\lambda\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u}),\,\mathcal{A}_\lambda\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})) + (\nabla\cdot\boldsymbol{\sigma},\,\nabla\cdot\boldsymbol{\tau}) & \text{for (3.3)} \end{cases}$$

for all $(\boldsymbol{\sigma},\,\mathbf{u};\,\boldsymbol{\tau},\,\mathbf{v}) \in (\Sigma \times U) \times (\Sigma \times U)$, and define a linear form $F(\cdot)$ on $\Sigma \times U$ by

$$F(\boldsymbol{\tau},\mathbf{v}) = \begin{cases} (f,\,\nabla\cdot\tau + Xv) & \text{for (3.1)}, \\ (-\mathbf{f},\,\nabla\cdot\boldsymbol{\tau}) & \text{for (3.3)} \end{cases}$$

for all $(\boldsymbol{\tau},\,\mathbf{v}) \in \Sigma \times U$. Then (3.7) can be written in equivalent form as follows: Find $(\boldsymbol{\sigma},\,\mathbf{u}) \in \Sigma \times U$ such that

$$b(\boldsymbol{\sigma},\,\mathbf{u};\,\boldsymbol{\tau},\,\mathbf{v}) = F(\boldsymbol{\tau},\,\mathbf{v}) \quad \forall\ (\boldsymbol{\tau},\,\mathbf{v}) \in \Sigma \times U. \tag{3.8}$$

4. Least-squares finite element approximations. Theorem 3.1 guarantees that conforming finite element spaces of $\Sigma \times U$ for the dual and primal variables may be chosen independently. But it does not imply that any choice leads to an optimal approximation in terms of both the regularity and the approximation property. It is important to note that the basic error estimations of least-squares methods in the energy norm for the dual and primal variables cannot be obtained separately (see (4.6)) and that the smoothness of the dual variable is one order less than the primal variable in our applications. Because of these constraints, the only finite element spaces having optimal approximations in the above sense are the continuous piecewise polynomials for the primal variable and the Raviart–Thomas elements for the dual variable. Moreover, the system of algebraic equations resulting from these elements can be solved efficiently by fast multigrid methods (see, e.g., [1]). For the above reasons, only these elements are analyzed in this paper. However, it is clear that our subsequent analysis does apply to any other conforming finite element spaces [5, 11] with no essential modification.

For simplicity of presentation, we consider only triangular and tetrahedral elements for the respective two and three dimensions. Assuming that the domain Ω is polygonal, let $\mathcal{T}_h$ be a regular triangulation of Ω (see [11]) with triangular/tetrahedral elements of size $\mathcal{O}(h)$. Let $P_k(K)$ be the space of polynomials of degree k on triangle K and denote the local Raviart–Thomas space of order k on K:

$$RT_k(K) = P_k(K)^d + \mathbf{x}\, P_k(K)$$

with $\mathbf{x} = (x_1, \ldots, x_d)$. The standard $H(\mathrm{div};\, \Omega)$ conforming Raviart–Thomas space of index k [19] and the standard (conforming) continuous piecewise polynomials of degree $k+1$ are defined, respectively, by

$$\Sigma_h^k = \{\boldsymbol{\tau} \in \Sigma \,:\, \boldsymbol{\tau}|_K \in RT_k(K)^m \quad \forall\ K \in \mathcal{T}_h\},$$
$$V_h^{k+1} = \{\mathbf{v} \in U \,:\, \mathbf{v}|_K \in P_{k+1}(K)^m \quad \forall\ K \in \mathcal{T}_h\}$$

where $m = 1$ for the scalar elliptic problems and $m = d$ for the generalized Stokes equations. It is well known (see [11]) that V_h^{k+1} has the following approximation property: Let $k \geq 0$ be an integer and let $l \in [0,\, k+1]$,

$$\inf_{\mathbf{v} \in V_h^{k+1}} \|\mathbf{u} - \mathbf{v}\|_1 \leq C\, h^l\, \|\mathbf{u}\|_{l+1} \tag{4.1}$$

for $\mathbf{u} \in H^{l+1}(\Omega)^m \cap U$. It is also well known (see [19]) that Σ_h^k has the following approximation property: Let $k \geq 0$ be an integer and let $l \in [1,\, k+1]$,

$$\inf_{\boldsymbol{\tau} \in \Sigma_h^k} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\, \Omega)} \leq C\, h^l\, (\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l) \tag{4.2}$$

for $\boldsymbol{\sigma} \in H^l(\Omega)^{m \times m} \cap \Sigma$ with $\nabla \cdot \boldsymbol{\sigma} \in H^l(\Omega)^m$. Since the smoothness of $\boldsymbol{\sigma}$ and $\nabla \cdot \boldsymbol{\sigma}$ is one order less than $\mathbf{u}$, we choose k to be the smallest integer greater than or equal to $l - 1$.

The finite element discretization of our least-squares variational problem is the following: Find $(\boldsymbol{\sigma}_h,\, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$ such that

$$G(\boldsymbol{\sigma}_h,\, \mathbf{u}_h;\, \mathbf{f}) = \min_{(\boldsymbol{\tau},\, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}} G(\boldsymbol{\tau},\, \mathbf{v};\, \mathbf{f}). \tag{4.3}$$

Equivalently, to find $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$ such that

$$b(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\tau}, \mathbf{v}) = F(\boldsymbol{\tau}, \mathbf{v}) \quad \forall\, (\boldsymbol{\tau}, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}. \tag{4.4}$$

Note that we have the orthogonal property

$$b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h; \boldsymbol{\tau}, \mathbf{v}) = 0 \quad \forall\, (\boldsymbol{\tau}, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}. \tag{4.5}$$

By Theorem 3.1 and the fact that $\Sigma_h^k \times V_h^{k+1}$ is a subspace of $\Sigma \times U$, (4.3) has a unique solution. The following error estimations in the energy norm follow directly from Theorem 3.1, the orthogonal property in (4.5), the Cauchy–Schwarz inequality, and the approximation properties in (4.1) and (4.2) (see also [6, 7, 8]).

THEOREM 4.1. *Assume that the solution* $(\boldsymbol{\sigma}, \mathbf{u})$ *of* (3.7) *is in* $H^l(\Omega)^{m\times m} \times H^{l+1}(\Omega)^m$ *and that the divergence of the stress* $\nabla\cdot\boldsymbol{\sigma}$ *is in* $H^l(\Omega)^m$ $(m = 1$ *or* $d)$*. Let* k *be the smallest integer greater than or equal to* $l-1$*. Then with* $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$ *denoting the solution to* (4.3)*, the following error estimate holds:*

$$|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)||| \le C \left(\inf_{\boldsymbol{\tau}\in\Sigma_h^k} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} + \inf_{\mathbf{v}\in V_h^{k+1}} \|\mathbf{u} - \mathbf{v}\|_1 \right) \tag{4.6}$$

$$\le C\, h^l \left(\|\boldsymbol{\sigma}\|_l + \|\nabla\cdot\boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1} \right). \tag{4.7}$$

5. The L^2 norm error estimates. This section presents the main results of this paper on the optimal L^2 norm estimates of $\mathbf{u} - \mathbf{u}_h$. As usual, these estimates are established through the duality argument. The key step of this argument is to express $\|\mathbf{u} - \mathbf{u}_h\|^2$ in terms of the bilinear form (see Lemma 5.1). While this is straightforward for the Galerkin finite element method, it is nontrivial for the div least-squares method.

LEMMA 5.1. *Let* $(\boldsymbol{\sigma}, \mathbf{u})$ *and* $(\boldsymbol{\sigma}_h, \mathbf{u}_h)$ *be the solutions of* (3.7) *and* (4.3)*, respectively. Assume that the regularity estimates in* (2.5) *and* (2.11) *hold. Then there exists* $(\boldsymbol{\gamma}, \mathbf{w})$ *such that*

$$\|\mathbf{u} - \mathbf{u}_h\|^2 = b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h; \boldsymbol{\gamma}, \mathbf{w}) \tag{5.1}$$

and that

$$\|\boldsymbol{\gamma}\|_1 + \|\nabla\cdot\boldsymbol{\gamma}\|_1 + \|\mathbf{w}\|_2 \le C\, \|\mathbf{u} - \mathbf{u}_h\|. \tag{5.2}$$

For the Galerkin method, $(\boldsymbol{\gamma}, \mathbf{w})$ in Lemma 5.1 is simply the solution of the corresponding dual problem. As developed later in this section, for the least-squares method, $(\boldsymbol{\gamma}, \mathbf{w})$ is chosen to be the solution of another auxiliary problem whose right-hand side involves the solution of the corresponding dual problem. This lemma will be proved separately for the scalar elliptic problems and the generalized Stokes equations in the respective sections 5.1 and 5.2. With Lemma 5.1, the optimal L^2 norm error estimates may be proved easily.

THEOREM 5.1. *Under the assumptions of Theorem* 4.1 *and Lemma* 5.1*, the following error estimate holds:*

$$\|\mathbf{u} - \mathbf{u}_h\| \le C\, h\, |||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)||| \tag{5.3}$$

$$\le C\, h^{l+1} \left(\|\boldsymbol{\sigma}\|_l + \|\nabla\cdot\boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1} \right). \tag{5.4}$$

Proof. It follows from Lemma 5.1, the orthogonality property in (4.5), the Cauchy–Schwarz inequality, and the approximation properties in (4.1) and (4.2) with $l = 1$ that for any $(\boldsymbol{\gamma}_h, \mathbf{w}_h) \in \Sigma_h^k \times V_h^{k+1}$,

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}_h\|^2 &= b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h; \boldsymbol{\gamma}, \mathbf{w}) = b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h; \boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \mathbf{w} - \mathbf{w}_h)\\ &\le C\,|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)||| \inf_{(\boldsymbol{\gamma}_h, \mathbf{w}_h) \in \Sigma_h^k \times V_h^{k+1}} |||(\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \mathbf{w} - \mathbf{w}_h)|||\\ &\le C\,h\,|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)|||\,(\|\boldsymbol{\gamma}\|_1 + \|\nabla \cdot \boldsymbol{\gamma}\|_1 + \|\mathbf{w}\|_2)\\ &\le C\,h\,|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)||| \cdot \|\mathbf{u} - \mathbf{u}_h\|.\end{aligned}$$

Dividing $\|\mathbf{u} - \mathbf{u}_h\|$ on both sides of the above inequality gives (5.3) which, combined with Theorem 4.1, implies (5.4). This completes the proof of the theorem. ∎

5.1. The scalar elliptic problems.

Proof of Lemma 5.1. Let $z \in H_D^1(\Omega)$ be the solution of (2.4) with $f = u - u_h$. The regularity estimate in (2.5) for the dual problem implies

$$\|z\|_2 \le C\,\|u - u_h\|. \tag{5.5}$$

By taking $v = u - u_h$ in (2.4), adding and subtracting $(\sigma - \sigma_h, \nabla v)$, and integrating by parts, we have

$$\begin{aligned}\|u - u_h\|^2 &= a(u - u_h, z) = (A\nabla(u - u_h), \nabla z) + (X(u - u_h), z)\\ &= ((\sigma - \sigma_h) + A\nabla(u - u_h), \nabla z) + (\nabla \cdot (\sigma - \sigma_h) + X(u - u_h), z).\end{aligned}$$

Now, to show the validity of Lemma 5.1, it suffices to find $(\gamma, w) \in H_N(\mathrm{div}; \Omega) \times H_D^1(\Omega)$ such that

$$\begin{cases} A^{-1}\gamma + \nabla w &= \nabla z \quad \text{in} \quad \Omega,\\ \nabla \cdot \gamma + Xw &= z \quad \text{in} \quad \Omega \end{cases} \tag{5.6}$$

and that

$$\|w\|_2 + \|\gamma\|_1 + \|\nabla \cdot \gamma\|_1 \le C\,\|u - u_h\|. \tag{5.7}$$

To do so, let $w \in H_D^1(\Omega)$ be the solution of the scalar elliptic problems in (2.1) and (2.2) with the right-hand side $f = z - \nabla \cdot (A\nabla z)$. The regularity estimate in (2.5) for the dual problem and the triangle inequality implies

$$\|w\|_2 \le C\,\|z - \nabla \cdot (A\nabla z)\| \le C\,\|z\|_2 \le C\,\|u - u_h\|. \tag{5.8}$$

Let $\gamma = A\nabla(z - w)$. It is then easy to check that (γ, w) satisfies (5.6). Now, (5.7) follows from (5.5), (5.8), and the facts that

$$\|\gamma\|_1 = \|A\nabla(z - w)\|_1 \le C\,(\|z\|_2 + \|w\|_2)$$

and that

$$\|\nabla \cdot \gamma\|_1 = \|z - Xw\|_1 \le C\,(\|z\|_1 + \|w\|_2)\,.$$

This completes the proof of Lemma 5.1 for the scalar elliptic problems. ∎

5.2. The generalized Stokes equations. The regularity estimate of the generalized Stokes equation in (2.7) is a consequence (2.11).

LEMMA 5.2. *For* $\mathbf{f} \in L^2(\Omega)^d$ *and* $g \in H^1(\Omega)$, *solution* $(\mathbf{u}, p) \in \left(H_D^1(\Omega)^d \cap H^2(\Omega)^d\right) \times H^1(\Omega)$ *of the generalized Stokes equations in* (2.7) *and* (2.8) *satisfies the a priori estimate*

$$\|\mathbf{u}\|_2 + \|p\|_1 \leq C\left(\|\mathbf{f}\| + \|g\|_1\right), \tag{5.9}$$

where the positive constant C *is independent of* λ.

Proof. When $\lambda = \infty$, let $(\hat{\mathbf{u}}, \hat{p})$ be the solution of (2.9) and (2.8); then (5.9) is simply the first inequality in (2.11), that is,

$$\|\hat{\mathbf{u}}\|_2 + \|\hat{p}\|_1 \leq C\left(\|\mathbf{f}\| + \|g\|_1\right). \tag{5.10}$$

When $\lambda \neq \infty$, let $(\mathbf{u}, p)$ be the solution of (2.7) and (2.8), and set $(\mathbf{u}^*, p^*) = (\mathbf{u} - \hat{\mathbf{u}}, p - \hat{p})$. It is easy to check that $(\mathbf{u}^*, p^*)$ satisfies (2.7) with $\mathbf{f} = \mathbf{0}$, $g = -\frac{1}{\lambda}\hat{p}$ and boundary conditions (2.8), that is,

$$\left\{\begin{array}{rcll} -\nabla \cdot (\boldsymbol{\epsilon}(\mathbf{u}^*) - p^* \boldsymbol{\delta}) &=& \mathbf{0} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u}^* + \dfrac{1}{\lambda} p^* &=& -\dfrac{1}{\lambda}\hat{p} & \text{in } \Omega \end{array}\right. \tag{5.11}$$

with $\mathbf{u}^* = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot (\boldsymbol{\epsilon}(\mathbf{u}^*) - p^* \boldsymbol{\delta}) = \mathbf{0}$ on Γ_N. Substituting $p^* = -\lambda \nabla \cdot \mathbf{u}^* - \hat{p}$ into the first equation of (5.11) gives

$$-\nabla \cdot \boldsymbol{\epsilon}(\mathbf{u}^*) - \lambda \nabla \nabla \cdot \mathbf{u}^* = -\nabla \hat{p}$$

with $\mathbf{u}^* = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot (\boldsymbol{\epsilon}(\mathbf{u}^*) + \lambda\,(\nabla \cdot \mathbf{u}^*)\,\boldsymbol{\delta}) = \mathbf{0}$ on Γ_N. By the regularity estimate in (2.11) for the elastic equations, (5.10), and the triangle inequality, we have that

$$\|\mathbf{u}^*\|_2 + \lambda \|\nabla \cdot \mathbf{u}^*\|_1 \leq C \|\nabla \hat{p}\| \leq C\left(\|\mathbf{f}\| + \|g\|_1\right)$$

and that

$$\|p^*\|_1 = \|\lambda \nabla \cdot \mathbf{u}^* + \hat{p}\|_1 \leq \|\lambda \nabla \cdot \mathbf{u}^*\|_1 + \|\hat{p}\|_1 \leq C\left(\|\mathbf{f}\| + \|g\|_1\right).$$

Now, (5.9) is an immediate consequence of the fact that $(\mathbf{u}, p) = (\mathbf{u}^* + \hat{\mathbf{u}}, p^* + \hat{p})$, the triangle inequality, and (5.10). This completes the proof of the lemma. ∎

With the H^2 regularity estimate for the generalized Stokes equations, we are now ready to prove Lemma 5.1 for both the elastic and Stokes equations by the duality argument.

Proof of Lemma 5.1. Let $\mathbf{e}_h = \mathbf{u} - \mathbf{u}_h$ and $E_h = \boldsymbol{\sigma} - \boldsymbol{\sigma}_h$. It is easy to see that

$$\mathbf{e}_h = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot E_h = \mathbf{0} \quad \text{on } \Gamma_N. \tag{5.12}$$

Consider the following dual problem: Let $(\mathbf{z}, r) \in H_0^1(\Omega) \times L^2(\Omega)$ be the solution of the perturbed Stokes equation in (2.7) and (2.8) with the right-hand sides $\mathbf{f} = \mathbf{u} - \mathbf{u}_h$ and $g = 0$, that is,

$$\left\{\begin{array}{rcll} -\nabla \cdot (\boldsymbol{\epsilon}(\mathbf{z}) - r\,\boldsymbol{\delta}) &=& \mathbf{e}_h & \text{in } \Omega, \\ \nabla \cdot \mathbf{z} + \dfrac{1}{\lambda} r &=& 0 & \text{in } \Omega \end{array}\right. \tag{5.13}$$

with boundary conditions

$$\mathbf{z} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot (\boldsymbol{\epsilon}(\mathbf{z}) - r\,\boldsymbol{\delta}) = \mathbf{0} \quad \text{on } \Gamma_N. \tag{5.14}$$

It follows from Lemma 5.2 that

$$\|\mathbf{z}\|_2 + \|r\|_1 \le C\,\|\mathbf{u} - \mathbf{u}_h\|. \tag{5.15}$$

To establish Lemma 5.1, we first use the above dual problem to derive the following equality:

$$\|\mathbf{e}_h\|^2 = (\mathcal{A}_\lambda E_h - \boldsymbol{\epsilon}(\mathbf{e}_h),\, r\,\boldsymbol{\delta} - \nabla\,\mathbf{z}) + (\nabla \cdot E_h,\, -\mathbf{z}). \tag{5.16}$$

To this end, using the first equation in (5.13), integration by parts, boundary conditions (5.12) and (5.14), and (2.6), we have

$$\begin{aligned}
\|\mathbf{e}_h\|^2 &= (\mathbf{e}_h,\, -\nabla \cdot (\boldsymbol{\epsilon}(\mathbf{z}) - r\,\boldsymbol{\delta})) \\
&= (\nabla\,\mathbf{e}_h,\, \boldsymbol{\epsilon}(\mathbf{z}) - r\,\boldsymbol{\delta}) + \int_{\partial\Omega} \mathbf{e}_h \cdot (\mathbf{n} \cdot (\boldsymbol{\epsilon}(\mathbf{z}) - r\,\boldsymbol{\delta})) \\
&= (\boldsymbol{\epsilon}(\mathbf{e}_h),\, \boldsymbol{\epsilon}(\mathbf{z}) - r\boldsymbol{\delta}) = (\boldsymbol{\epsilon}(\mathbf{e}_h),\, \nabla\,\mathbf{z} - r\boldsymbol{\delta}) \\
&= (\mathcal{A}_\lambda E_h - \boldsymbol{\epsilon}(\mathbf{e}_h),\, r\boldsymbol{\delta} - \nabla\,\mathbf{z}) + (\mathcal{A}_\lambda E_h,\, \nabla\,\mathbf{z}) - (\mathrm{tr}(\mathcal{A}_\lambda E_h),\, r). \qquad (5.17)
\end{aligned}$$

By the definition of $\mathcal{A}_\lambda$, integration by parts, boundary conditions (5.12) and (5.14), and the second equation in (5.13), we have

$$\begin{aligned}
(\mathcal{A}_\lambda E_h,\, \nabla\,\mathbf{z}) &= \left(E_h - \frac{\lambda}{d\,\lambda + 1}(\mathrm{tr}\,E_h)\,\boldsymbol{\delta},\, \nabla\,\mathbf{z}\right) = (E_h,\, \nabla\,\mathbf{z}) - \frac{\lambda}{d\,\lambda + 1}((\mathrm{tr}\,E_h)\,\boldsymbol{\delta},\, \nabla\,\mathbf{z}) \\
&= -(\nabla\cdot\, E_h,\, \mathbf{z}) + \int_{\partial\Omega} (\mathbf{n} \cdot E_h) \cdot \mathbf{z} - \frac{\lambda}{d\,\lambda + 1}(\mathrm{tr}\,E_h,\, \nabla \cdot \mathbf{z}) \\
&= (\nabla\cdot\, E_h,\, -\mathbf{z}) + \frac{1}{d\,\lambda + 1}(\mathrm{tr}\,E_h,\, r),
\end{aligned}$$

which, together with the fact that

$$\mathrm{tr}\,(\mathcal{A}_\lambda\,E_h) = \frac{1}{d\,\lambda + 1}\mathrm{tr}\,E_h,$$

implies

$$(\mathcal{A}_\lambda E_h,\, \nabla\,\mathbf{z}) = (\nabla\cdot\, E_h,\, -\mathbf{z}) + (\mathrm{tr}(\mathcal{A}_\lambda\,E_h),\, r). \tag{5.18}$$

Now, (5.16) follows from (5.17) and (5.18).

Next, we want to find $(\boldsymbol{\gamma},\, \mathbf{w})$ such that

$$\begin{cases} \mathcal{A}_\lambda\,\boldsymbol{\gamma} - \boldsymbol{\epsilon}(\mathbf{w}) &= r\,\boldsymbol{\delta} - \nabla\,\mathbf{z} \quad \text{in} \quad \Omega, \\ \nabla \cdot \boldsymbol{\gamma} &= -\mathbf{z} \quad \text{in} \quad \Omega \end{cases} \tag{5.19}$$

and that

$$\|\mathbf{w}\|_2 + \|\boldsymbol{\gamma}\|_1 + \|\nabla \cdot \boldsymbol{\gamma}\|_1 \le C\,\|\mathbf{e}_h\|. \tag{5.20}$$

To do so, let $(\mathbf{w}, s)$ be the solution of the perturbed Stokes equation in (2.7) and (2.8) with the right-hand sides

$$\mathbf{f} = -\triangle \mathbf{z} + \mathbf{z} \quad \text{and} \quad g = -\frac{d\,\lambda + 2}{\lambda}\, r,$$

that is,

$$\left\{ \begin{array}{rcll} -\nabla \cdot (\boldsymbol{\epsilon}(\mathbf{w}) - s\,\boldsymbol{\delta}) & = & -\triangle \mathbf{z} + \mathbf{z} & \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{w} + \dfrac{1}{\lambda}\, s & = & -\dfrac{d\,\lambda + 2}{\lambda}\, r & \text{in} \quad \Omega \end{array} \right. \tag{5.21}$$

with boundary conditions

$$\mathbf{w} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot (\boldsymbol{\epsilon}(\mathbf{w}) - s\,\boldsymbol{\delta}) = \mathbf{0} \quad \text{on } \Gamma_N.$$

It then follows from Lemma 5.2 and (5.15) that

$$\begin{aligned} \|\mathbf{w}\|_2 + \|s\|_1 &\le C \left(\| - \triangle \mathbf{z} + \mathbf{z}\| + \left\| \frac{d\,\lambda + 2}{\lambda}\, r \right\|_1 \right) \\ &\le C\, (\|\mathbf{z}\|_2 + \|r\|_1) \le C\, \|\mathbf{e}_h\|. \end{aligned} \tag{5.22}$$

Now, let

$$\boldsymbol{\gamma} = -\nabla\, \mathbf{z} + \boldsymbol{\epsilon}(\mathbf{w}) - s\,\boldsymbol{\delta}. \tag{5.23}$$

Then applying the divergence operator to (5.23) and using the first equation in (5.21) yield the second equation in (5.19). Applying the trace operator to (5.23) and using the second equations in both (5.13) and (5.21), we have

$$\operatorname{tr} \boldsymbol{\gamma} = -\frac{d\,\lambda + 1}{\lambda}\, (s + r).$$

Hence,

$$s = -\frac{\lambda}{d\,\lambda + 1} \operatorname{tr} \boldsymbol{\gamma} - r,$$

which, combined with (5.23), implies the first equation in (5.19). Therefore, $(\boldsymbol{\gamma}, \mathbf{w})$ satisfies (5.19). To prove that $(\boldsymbol{\gamma}, \mathbf{w})$ also satisfies (5.20), it follows from (5.23), the triangle inequality, and (5.22) that

$$\|\boldsymbol{\gamma}\|_1 \le \|\nabla\, \mathbf{z}\|_1 + \|\boldsymbol{\epsilon}(\mathbf{w})\|_1 + d\, \|s\|_1 \le C\, (\|\mathbf{z}\|_2 + \|\mathbf{w}\|_2 + \|s\|_1) \le C\, \|\mathbf{e}_h\|$$

and from the second equation of (5.19) and (5.22) that

$$\|\nabla \cdot \boldsymbol{\gamma}\|_1 = \|\mathbf{z}\|_1 \le C\, \|\mathbf{e}_h\|,$$

which, combined with (5.22), implies (5.20).

Finally, combining (5.16) and (5.19) yields (5.1) with $(\boldsymbol{\gamma}, \mathbf{w})$ satisfying (5.2). This completes the proof of Lemma 5.1 for both the elasticity and Stokes equations. ∎

5.3. Extension to problems with low regularity. Extensions of our results in this paper to problems with low regularity can be carried out in a similar fashion. To this end, we assume the following $H^{1+\alpha}$ regularity estimates with $\alpha \in [1/2, 1)$:

$$\|u\|_{1+\alpha} \le C\,\|f\|_{1-\alpha} \quad \text{and} \quad \|z\|_{1+\alpha} \le C\,\|f\|_{1-\alpha} \tag{5.24}$$

for problems (2.3) and (2.4), respectively, and

$$\|\mathbf{u}\|_{1+\alpha} + \|p\|_\alpha \le C\,(\|\mathbf{f}\|_{1-\alpha} + \|g\|_\alpha) \quad \text{and} \quad \|\mathbf{u}\|_{1+\alpha} + \lambda\,\|\nabla\cdot\mathbf{u}\|_\alpha \le C\,\|\mathbf{f}\|_{1-\alpha} \tag{5.25}$$

for problems (2.9) and (2.10), respectively. The assumption on $\alpha \ge 1/2$ is needed in order to define the Raviart–Thomas elements. For any $\boldsymbol{\sigma} \in H^\alpha(\Omega)^{m\times m} \cap \Sigma$ with $\nabla\cdot\boldsymbol{\sigma} \in H^\alpha(\Omega)^m$, we assume the following approximation property for Σ_h^k: There exists $0 < \beta \le \alpha$ such that

$$\inf_{\boldsymbol{\tau}\in\Sigma_h^k} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} \le C\,h^\beta\,(\|\boldsymbol{\sigma}\|_\alpha + \|\nabla\cdot\boldsymbol{\sigma}\|_\alpha)\,. \tag{5.26}$$

THEOREM 5.2. *Assume that the solution* $(\boldsymbol{\sigma}, \mathbf{u})$ *of* (3.7) *is in* $H^\alpha(\Omega)^{m\times m} \times H^{\alpha+1}(\Omega)^m$ *and that the divergence of the stress* $\nabla\cdot\boldsymbol{\sigma}$ *is in* $H^\alpha(\Omega)^m$ ($m = 1$ *or* d) *with* $\alpha \in [1/2, 1)$. *Suppose that approximation property* (5.26) *holds; then with* $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^0 \times V_h^1$ *denoting the solution to* (4.3), *the following error estimate holds:*

$$|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h,\, \mathbf{u} - \mathbf{u}_h)||| \le C\,h^\beta\,(\|\boldsymbol{\sigma}\|_\alpha + \|\nabla\cdot\boldsymbol{\sigma}\|_\alpha) + C\,h^\alpha\,\|\mathbf{u}\|_{1+\alpha}. \tag{5.27}$$

Proof. The proof of (5.27) is standard and identical to that of (4.7). ∎

THEOREM 5.3. *Assume that the regularity estimates in* (5.24) *and* (5.25) *hold. Under the assumptions of Theorem* 5.2, *we have the following* L^2 *norm error estimates:*

$$\|\mathbf{u} - \mathbf{u}_h\| \le C\,h^\beta\,|||(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h,\, \mathbf{u} - \mathbf{u}_h)||| \tag{5.28}$$

$$\le C\,h^{2\,\beta}\,(\|\boldsymbol{\sigma}\|_\alpha + \|\nabla\cdot\boldsymbol{\sigma}\|_\alpha + \|\mathbf{u}\|_{1+\alpha})\,. \tag{5.29}$$

Proof. In a similar fashion, one can first establish the $H^{1+\alpha}$ regularity for the generalized Stokes equations in (2.7) and (2.8) and then prove that there exists $(\boldsymbol{\gamma}, \mathbf{w})$ such that

$$\|\mathbf{u} - \mathbf{u}_h\|^2 = b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h,\, \mathbf{u} - \mathbf{u}_h;\, \boldsymbol{\gamma},\, \mathbf{w})$$

and that

$$\|\boldsymbol{\gamma}\|_\alpha + \|\nabla\cdot\boldsymbol{\gamma}\|_\alpha + \|\mathbf{w}\|_{\alpha+1} \le C\,\|\mathbf{u} - \mathbf{u}_h\|.$$

Now, error bound (5.28) easily follows in a way similar to that of (5.3), and (5.29) is then a direct consequence of (5.28) and (5.27). ∎

REFERENCES

[1] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Multigrid in H(*div*) and H(curl)*, Numer. Math., 85 (2000), pp. 197–218.

[2] A. AZIZ, R. KELLOGG, AND A. STEPHENS, *Least-squares methods for elliptic systems*, Math. Comp., 44 (1985), pp. 53–70.

[3] P. Bochev and M. Gunzburger, *On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal., 43 (2005), pp. 340–362.

[4] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order system*, Math. Comp., 66 (1997), pp. 935–955.

[5] F. Brezzi, J. Douglas, Jr., and L. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[6] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[7] Z. Cai, B. Lee, and P. Wang, *Least-squares methods for incompressible Newtonian fluid flow: Linear stationary problems*, SIAM J. Numer. Anal., 42 (2004), pp. 843–859.

[8] Z. Cai and G. Starke, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.

[9] G. F. Carey and Y. Shen, *Convergence studies of least-squares finite elements for first order systems*, Comm. Appl. Numer. Methods, 5 (1989), pp. 427–434.

[10] C. L. Chang, *A least-squares finite element method for the Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 83 (1990), pp. 1–7.

[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[12] C. L. Cox and G. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[13] G. Fix, M. Gunzburger, and R. Nicolaides, *On the finite element methods of least-squares type*, Comput. Math. Appl., 5 (1979), pp. 87–98.

[14] D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31 (1977), pp. 873–880.

[15] B. N. Jiang and L. A. Povinelli, *Optimal least-squares finite element method for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 102 (1993), pp. 199–212.

[16] P. P. Lynn and S. K. Arya, *Use of the least squares criterion in the finite element formulation*, Internat. J. Numer. Methods Engrg., 6 (1973), pp. 75–88.

[17] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[18] A. I. Pehlivanov, G. F. Carey, R. D. Lazarov, and Y. Shen, *Convergence of least squares finite elements for first order ODE systems*, Computing, 51 (1993), pp. 111–123.

[19] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer, Berlin, 1977, pp. 292–315.

[20] D. Yang, *Analysis of least-squares mixed finite element methods for nonlinear nonstationary convection-diffusion problems*, Math. Comp., 69 (2000), pp. 929–963.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1735–1758

STRONG STABILITY FOR ADDITIVE RUNGE–KUTTA METHODS*

INMACULADA HIGUERAS†

Abstract. Space discretization of some time-dependent partial differential equations gives rise to ordinary differential equations containing additive terms with different stiffness properties. In these situations, additive Runge–Kutta (ARK) methods are used. The aim of this paper is to study monotonicity properties (also known as strong stability) for ARK methods. A new definition of absolute monotonicity for ARK methods is given and some of its properties are investigated. With this concept, monotonicity for ARK schemes under certain stepsize restrictions can be ensured. Some ARK methods from the literature are analyzed. As expected, monotonicity for each Runge–Kutte (RK) method does not ensure monotonicity for the ARK scheme. Some numerical examples show the applicability of these results.

Key words. Runge–Kutta, strong stability preserving, absolutely monotonic, radius of absolute monotonicity, CFL coefficient

AMS subject classifications. 65L06, 65L05, 65M20

DOI. 10.1137/040612968

1. Introduction.

We consider initial value problems for ordinary differential equations (ODEs) of the form

$$\begin{aligned} \frac{d}{dt}u(t) &= f\left(u(t)\right) + \tilde{f}\left(u(t)\right), \qquad t \geq t_0, \\ u(t_0) &= u_0\,. \end{aligned} \tag{1.1}$$

We assume that $t_0 \in \mathbb{R}$, $u_0 \in \mathbb{R}^m$, and f and $\tilde{f}$ are continuous functions from $\mathbb{R}^m$ to $\mathbb{R}^m$ with different stiffness properties such that for each $t_0 \in \mathbb{R}$ and $u_0 \in \mathbb{R}^m$ problem (1.1) has a unique solution $u : [t_0, \infty) \to \mathbb{R}^m$. Such systems often arise from space discretization of time-dependent partial differential equations (PDEs) by the method of lines [2, 3, 4, 7, 17, 19, 20, 22, 27]. We assume too that $\|\cdot\| : \mathbb{R}^m \to \mathbb{R}$ is a convex function (e.g., a norm, an entropy function, etc.), such that for any $t_0 \in \mathbb{R}$ and any solution $u(t)$ to (1.1) we have

$$\|u(t)\| \leq \|u(t_0)\| \qquad \text{for all } \ t \geq t_0\,. \tag{1.2}$$

In order to obtain property (1.2) for the solution of (1.1), some conditions must be imposed on the functions f and $\tilde{f}$. In the rest of the paper we assume that $(f, \tilde{f}\|\cdot\|)$ satisfy

$$\left\| y + \frac{1}{\rho}\ f(y) \right\| \leq \| y \| \qquad \text{for all } y \in \mathbb{R}^m\,, \tag{1.3}$$

$$\left\| y + \frac{1}{\tilde{\rho}}\ \tilde{f}(y) \right\| \leq \| y \| \qquad \text{for all } y \in \mathbb{R}^m \tag{1.4}$$

*Received by the editors August 6, 2004; accepted for publication (in revised form) January 9, 2006; published electronically September 26, 2006. This research was supported by the Ministerio de Ciencia y Tecnología, Project BFM2001-2188.
http://www.siam.org/journals/sinum/44-4/61296.html
†Departamento de Matemática e Informática, Universidad Pública de Navarra, Pamplona, Navarra 31006, Spain (higueras@unavarra.es).

for some fixed $\rho, \tilde{\rho} > 0$, and we denote this class of problems by $\mathcal{F}(\rho, \tilde{\rho})$. Recall that (1.3)–(1.4) imply that these inequalities also hold for any τ with $0 \le \tau \le 1/\rho$ [18, Theorem 5.1]. Under these assumptions it is straightforward to prove that (1.3)–(1.4) imply that for $u(t)$, the solution of (1.1), we have

$$D_+ \|u(t)\| \le 0\,,$$

where D_+ denotes the right-hand derivative, and hence (1.2) holds. For details see [13].

Given the initial value problem (1.1) a common class of one step methods for solving it numerically is that of the additive Runge–Kutta (ARK) methods. An s-stage ARK method is defined by two $s \times s$ real matrices $\mathcal{A}$ and $\tilde{\mathcal{A}}$, and two real vectors $b, \tilde{b} \in \mathbb{R}^s$. From u_n, the numerical approximation of the solution $u(t)$ at $t = t_n$, we obtain u_{n+1}, the numerical approximation of the solution at $t_{n+1} = t_n + h$, by

$$u_{n+1} = u_n + h \sum_{i=1}^{s} b_i f(U_i) + h \sum_{i=1}^{s} \tilde{b}_i \tilde{f}(U_i)\,, \tag{1.5}$$

where the internal stages U_i are given as

$$U_i = u_n + h \sum_{j=1}^{s} a_{ij} f(U_j) + h \sum_{j=1}^{s} \tilde{a}_{ij} \tilde{f}(U_j)\,. \tag{1.6}$$

The Runge–Kutta (RK) methods $(\mathcal{A}, b)$ and $(\tilde{\mathcal{A}}, \tilde{b})$ are chosen with the aim of integrating system (1.1) with low computational cost. For example, if f represents the nonstiff part of the system and $\tilde{f}$ the stiff part, an explicit method can be used for f and an implicit one for $\tilde{f}$. ARK methods combining implicit and explicit schemes are known in the literature as IMplicit–EXplicit (IMEX) RK methods [2, 19, 17]. We remark that the problems we have in mind have a stiff behavior and therefore, the use of an explicit method for the whole problem is not possible.

Denoting the coefficients of the ARK method by

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} \tilde{\mathcal{A}} & 0 \\ \tilde{b}^t & 0 \end{pmatrix},$$

we can write (1.5)–(1.6) in compact form as

$$U = e \otimes u_n + h(\mathbb{A} \otimes I) F(U) + h(\tilde{\mathbb{A}} \otimes I) \tilde{F}(U)\,, \tag{1.7}$$

where $e = (1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, $U = (U_1^t, \ldots, U_s^t, u_{n+1}^t)^t \in \mathbb{R}^{(s+1)m}$, $F(U) = (f(U_1)^t, \ldots, f(U_s)^t, 0)^t \in \mathbb{R}^{(s+1)m}$, and similarly $\tilde{F}(U)$. The symbol $\otimes$ denotes the Kronecker product (see, e.g., [16, section 12.1])

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mm}B \end{pmatrix}.$$

Some properties of the Kronecker product, namely $(A \otimes B)\,(C \otimes D) = (A\,B \otimes C\,D)$, $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, where the matrices involved have the proper dimensions and properties, will be used later on.

The internal stage U_i approximates $u(t_n + c_i h)$, where $c_i = \sum_{j=1}^{s} a_{ij}$. Furthermore, for many methods it holds that $c_i \geq 0$, $i = 1, \ldots, s$, and thus $t_n + c_i h \geq t_n$. Therefore, if we solve numerically an ODE (1.1) with $(f, \tilde{f}, \|\cdot\|) \in \mathcal{F}(\rho, \tilde{\rho})$ with an ARK method, a natural requirement for the internal stages and the numerical solution is

$$\|U_i\| \leq \|u_n\|, \qquad i = 1, \ldots, s, \tag{1.8}$$

$$\|u_{n+1}\| \leq \|u_n\| \tag{1.9}$$

for all $n \geq 0$, probably under a stepsize restriction $h \leq \Delta t_{MAX}$. Following the nomenclature for RK methods, we will say that an ARK scheme that satisfies (1.8)–(1.9) is strongly stable.

Monotonicity properties (1.8)–(1.9) for RK methods have been studied by several authors [5, 23, 24, 25, 10, 9, 21, 26]; see [12] for a review. The aim of this paper is to study these properties for ARK methods. As we will see in this paper, it is not true that two strongly stable RK methods give rise to a strongly stable ARK method. This fact is not surprising because in general the fact that a property holds for methods $(\mathcal{A}, b)$ and $(\tilde{\mathcal{A}}, \tilde{b})$ separately does not imply that it holds for the ARK method. This situation is well known for the order of consistency of an ARK scheme, where, together with the order conditions for each method, some extra coupling conditions are required (see, e.g., [17, 20]). Similarly, where stability issues for ARK methods have been studied by some authors [4, 17], it has been found that extra conditions must be considered.

As far as we know, strong stability has not been studied for ARK schemes. The closest research is found in [20], where, in the context of hyperbolic systems with relaxation, IMEX methods are used to solve problems of the form

$$\frac{d}{dt}\, u(t) = f(u(t)) + \frac{1}{\varepsilon}\, \tilde{f}(u(t)),$$

where ε is the stiffness parameter. For this problem, monotonicity in the stiff limit, i.e., when $\varepsilon \to 0$, is studied. For this purpose, in [20] the concept of the asymptotic preserving (AP) method is used. In this case, under certain conditions for the implicit scheme, monotonicity can be recovered under the only stability restriction introduced by the explicit part of the IMEX scheme. However, the results obtained in [20] do not cover the case of moderate values of ε.

For RK methods the radius of absolute monotonicity gives stepsize restrictions for monotonicity [18, 5]. In this paper we extend this concept for ARK methods and give sufficient conditions for monotonicity.

The rest of the paper is organized as follows. Section 2 is devoted to the extension of the radius of absolute monotonicity for ARK methods and the study of some of its properties. Stepsize restrictions for monotonicity are given in terms of this radius in section 3. This section ends with subsections 3.1 and 3.2, where the previous results are related to the Shu–Osher forms and perturbed RK schemes. With the new concept defined, the monotonicity properties of some methods from the literature are analyzed in section 4. Some numerical experiments are shown in section 5. The paper ends with some conclusions and open questions for future work in section 6.

2. Absolute monotonicity for additive RK methods. In the context of contractive and monotone RK methods, the concept of radius of absolute monotonicity

plays an important role [18, 5]. We review the definitions of absolute monotonicity and radius of absolute monotonicity as follows.

DEFINITION 2.1 (see [18, Definition 2.4]). *An s-stage RK method with coefficients* $\mathbb{A}$ *is said to be absolutely monotonic (a.m.) at a given point* $\xi \leq 0$ *if the matrix* $I - \xi\mathbb{A}$ *is nonsingular and*

$$(I - \xi\mathbb{A})^{-1}\mathbb{A} \geq 0\,, \tag{2.1}$$

$$(I - \xi\mathbb{A})^{-1}e \geq 0\,, \tag{2.2}$$

where $e = (1, 1, \ldots, 1)^t \in \mathbb{R}^{s+1}$*, and the vector inequalities are understood componentwise. Further, the method is said to be a.m. on a given set* $\Omega \subset \mathbb{R}$ *if it is a.m. at each* $\xi \in \Omega$*. The radius of absolute monotonicity* $R(\mathbb{A})$ *is defined by*

$$R(\mathbb{A}) = \sup\{\, r \,|\, r \geq 0 \text{ and } \mathbb{A} \text{ is a.m. on } [-r, 0]\,\}\,. \tag{2.3}$$

If there is no $r > 0$ *such that* $\mathbb{A}$ *is a.m. on* $[-r, 0]$*, we set* $R(\mathbb{A}) = 0$*.*

Our first aim is to extend this concept for ARK methods $(\mathbb{A}, \tilde{\mathbb{A}})$ and analyze some of its properties. For a better understanding of that extension, we briefly show how conditions (2.1)–(2.2) in Definition 2.1 arise in [18].

In [18] the scalar linear problems $u' = \lambda u$ and $u' = \lambda(t)u$, and the vectorial linear problem $u' = L(t)u(t)$, with $L(t)$ an $m \times m$ matrix, are considered. In compact form, with $U = (U_1^t, \ldots, U_s^t, u_{n+1}^t)^t \in \mathbb{R}^{(s+1)m}$ (cf. (1.7)), an RK method $\mathbb{A}$ for these problems gives, respectively, $U = \phi(h\lambda)u_n$, with

$$\phi(z) = (I_{s+1} - z\mathbb{A})^{-1}\, e;$$

$U = K\left(\text{diag}\,(h\lambda_1, \ldots, h\lambda_s, 0)\right) u_n$, with

$$K(Z) = (I_{s+1} - \mathbb{A}Z)^{-1}\, e$$

and Z a diagonal matrix; and $U = \mathbb{K}\left(\text{diag}\,(hL_1, \ldots, hL_s, 0)\right) \otimes u_n$, with $L_i = L(t_n + c_i h)$,

$$\mathbb{K}(\mathbb{Z}) = \left(I_{(s+1)\cdot m} - (\mathbb{A} \otimes I_m)\mathbb{Z}\right)^{-1} (e \otimes I_m)\,,$$

and $\mathbb{Z}$ a block diagonal matrix. The concepts of absolute monotonicity at a given point $\xi \in \mathbb{R}$ for ϕ, K, and $\mathbb{K}$ are given in [18, p. 487]. Roughly speaking they mean the nonnegativity of all coefficients of the Taylor expansion about $z = \xi$, $Z = \xi I_{s+1}$, or $\mathbb{Z} = \xi I_{(s+1)\cdot m}$, respectively. In particular, for $\mathbb{Z} = \xi I_{(s+1)\cdot m} + \mathbb{W}$, with $\mathbb{W}$ sufficiently close to zero, we obtain

$$\mathbb{K}(\mathbb{Z}) = \left[I_{(s+1)\cdot m} - (\mathsf{A}(\xi) \otimes I_m)\mathbb{W}\right]^{-1} (\mathsf{e}(\xi) \otimes I_m)\,,$$

where $\mathsf{A}(\xi) = (I_{s+1} - \xi\mathbb{A})^{-1}\mathbb{A}$ and $\mathsf{e}(\xi) = (I_{s+1} - \xi\mathbb{A})^{-1}e$. Thus with $\mathsf{A}(\xi) \geq 0$ and $\mathsf{e}(\xi) \geq 0$ (see (2.1)–(2.2)), we obtain the absolute monotonicity of $\mathbb{K}$ at $\xi I_{(s+1)\cdot m}$.

Similarly, if we consider now the additive problems $u' = -(\lambda + \tilde{\lambda})u$ and $u' = -(\lambda(t) + \tilde{\lambda}(t))u$, and the vectorial linear problem $u' = -(L(t) + \tilde{L}(t))u(t)$, with $\tilde{L}(t)$ an $m \times m$ matrix, an ARK method $(\mathbb{A}\,, \tilde{\mathbb{A}})$ gives, respectively,

$$U = \phi(h\lambda, h\tilde{\lambda})u_n,$$

with

$$\phi(z, \tilde{z}) = \left(I_{s+1} - z\mathbb{A} - \tilde{z}\tilde{\mathbb{A}}\right)^{-1} e;$$

$U = K\left(\text{diag }(h\lambda_1, \ldots, h\lambda_s, 0), \text{diag }(h\tilde{\lambda}_1, \ldots, h\tilde{\lambda}_s, 0)\right) u_n$, with

$$K(Z, \tilde{Z}) = \left(I_{s+1} - \mathbb{A}Z - \tilde{\mathbb{A}}\tilde{Z}\right)^{-1} e$$

and Z, $\tilde{Z}$ diagonal matrices; and

$$U = \mathbb{K}\left(\text{diag }(hL_1, \ldots, hL_s, 0), \text{diag }(h\tilde{L}_1, \ldots, h\tilde{L}_s, 0)\right) \otimes u_n,$$

with

$$\mathbb{K}(\mathbb{Z}, \tilde{\mathbb{Z}}) = \left(I_{(s+1)\cdot m} - (\mathbb{A} \otimes I_m)\mathbb{Z} - (\tilde{\mathbb{A}} \otimes I_m)\tilde{\mathbb{Z}}\right)^{-1} (e \otimes I_m)$$

and $\mathbb{Z}$, $\tilde{\mathbb{Z}}$ block diagonal matrices. For $\mathbb{Z} = \xi I_{(s+1)\cdot m} + \mathbb{W}$, $\tilde{\mathbb{Z}} = \tilde{\xi} I_{(s+1)\cdot m} + \tilde{\mathbb{W}}$, with $\mathbb{W}$, $\tilde{\mathbb{W}}$ sufficiently close to zero, we obtain

$$\mathbb{K}(\mathbb{Z}, \tilde{\mathbb{Z}}) = \left[I_{(s+1)\cdot m} - (\mathsf{A}(\xi, \tilde{\xi}) \otimes I_m)\mathbb{W} - (\tilde{\mathsf{A}}(\xi, \tilde{\xi}) \otimes I_m)\tilde{\mathbb{W}}\right]^{-1} \left(\mathsf{e}(\xi, \tilde{\xi}) \otimes I_m\right),$$

where now $\mathsf{A}(\xi, \tilde{\xi})$, $\tilde{\mathsf{A}}(\xi, \tilde{\xi})$, and $\mathsf{e}(\xi, \tilde{\xi})$ are defined, respectively, by (2.4), (2.5), and (2.6) below. Thus with $\mathsf{A}(\xi, \tilde{\xi}) \geq 0$, $\tilde{\mathsf{A}}(\xi, \tilde{\xi}) \geq 0$ and $\mathsf{e}(\xi, \tilde{\xi}) \geq 0$, we obtain that all the coefficients in the Taylor expansion of $\mathbb{K}(\mathbb{Z}, \tilde{\mathbb{Z}})$ at $(\xi I_{(s+1)\cdot m}, \tilde{\xi} I_{(s+1)\cdot m})$ are nonnegative. This analysis lead us to the extension of Definition 2.1 as follows.

DEFINITION 2.2. *An s-stage ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is said to be a.m. at a given point* $(\xi, \tilde{\xi})$ *with* $\xi, \tilde{\xi} \leq 0$ *if the matrix* $I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}}$ *is invertible and*

$$\mathsf{A}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\mathbb{A} \geq 0\,, \tag{2.4}$$

$$\tilde{\mathsf{A}}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}} \geq 0\,, \tag{2.5}$$

$$\mathsf{e}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}e \geq 0\,. \tag{2.6}$$

Further, the additive method is said to be a.m. on a given set $\Omega \in \mathbb{R}^2$ *if it is absolutely monotonic at each* $(\xi, \tilde{\xi}) \in \Omega$.

Observe that for RK we worked in $\mathbb{R}$, but for ARK methods we have to work in $\mathbb{R}^2$. For this reason we define the region of absolute monotonicity as follows.

DEFINITION 2.3. *The region of absolute monotonicity, denoted by* $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, *is defined by*

$$\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = \{\, (r, \tilde{r}) \,|\, r \geq 0\,, \tilde{r} \geq 0, \text{ and } (\mathbb{A}, \tilde{\mathbb{A}}) \text{ is a.m. on } [-r, 0] \times [-\tilde{r}, 0] \,\}\,.$$

Finally, for RK methods the radius of absolute monotonicity is given by the supremum (2.3), which is a part of the frontier of the set

$$\{\, r \,|\, r \geq 0 \text{ and } \mathbb{A} \text{ is a.m. on } [-r, 0] \,\}\,.$$

For ARK schemes we consider a part of the frontier of the region of a.m. $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$.

DEFINITION 2.4. *The curve of absolute monotonicity, denoted by* $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, *is the frontier of the set* $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ *excluding the coordinate axis (see Figure* 2.1*).*

If there is no $r > 0$, $\tilde{r} > 0$ *such that* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is a.m. on* $[-r, 0] \times [-\tilde{r}, 0]$, *we set* $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = (0, 0)$.

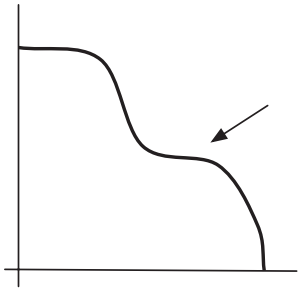


FIG. 2.1. *Curve of absolute monotonicity.*

Later on we will prove the absolute monotonicity of the ARK method at a given point $(r, \tilde{r})$ considering the absolute monotonicity of the ARK method on the semi-open line connecting the origin and the point $(r, \tilde{r})$. We give the following definition.

DEFINITION 2.5. *For $r, \tilde{r} \geq 0$ we will say that $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) \geq (r, \tilde{r})$ if $(\mathbb{A}, \tilde{\mathbb{A}})$ is a.m. at $\mathcal{L}(r, \tilde{r})$, where*

$$\mathcal{L}(r, \tilde{r}) = \left\{ (\xi, \tilde{\xi}) \,\middle|\, \xi \in (-r, 0] \,,\, \tilde{\xi} = \frac{\tilde{r}}{r} \right\} . \tag{2.7}$$

Once the basic definitions are given, we go deeper into the concept of radius of absolute monotonicity. In [18, Lemma 4.4] it is proved that under certain conditions, for the absolute monotonicity of an RK method $\mathbb{A}$ on a given interval $[-r, 0]$, it is sufficient to consider the absolute monotonicity at the left endpoint $-r$. Our next goal is to prove that a similar result is also true for ARK methods $(\mathbb{A}, \tilde{\mathbb{A}})$. We begin with some technical lemmas.

LEMMA 2.6. *Consider an order m matrix C such that $C \geq 0$. Assume too that the matrix $I + C$ is nonsingular and $C\,(I + C)^{-1} \geq 0$. Then $I - \xi C$ is nonsingular for all $\xi \in [-2, 0]$.*

Proof. As $C\,(I + C)^{-1} \geq 0$ and $C \geq 0$, then $(I + C)^{-1} = I - C\,(I + C)^{-1}$ is an M-matrix, and hence [16, section 15.2] for the spectral radius of $C\,(I + C)^{-1}$ we have spectral radius $\left(C\,(I + C)^{-1}\right) < 1$. Therefore $\left|\lambda I - C\,(I + C)^{-1}\right| \neq 0$ for all λ with $|\lambda| \geq 1$. For $\xi \neq -1$ we have that

$$\left|I - (\xi + 1)\, C\,(I + C)^{-1}\right| = (\xi + r)^m \left| \frac{1}{\xi + 1} I - C\,(I + C)^{-1} \right| ,$$

obtaining that $\left|I - (\xi + 1)\, C\,(I + C)^{-1}\right| \neq 0$ for all ξ with $|\xi + 1| \leq 1$. Finally, from

$$|I - \xi\, C| = |I - (\xi + 1)\, C\,(I + C)^{-1}| \cdot |I + C|$$

we obtain that $I - \xi\, C$ is nonsingular for all $\xi \in [-2, 0]$. ∎

Observe that we have proved that under certain assumptions, the regularity of $I - \xi\, C$ for $\xi = -1$ implies the regularity of this matrix for ξ in the interval $[-1, 0]$ (see (2.7)).

For the next result we need the concept of an a.m. function for functions $f : \mathbb{R}^2 \to \mathbb{R}$, and matrices whose elements are functions.

DEFINITION 2.7. *A function $\psi(z, \tilde{z}) = P(z, \tilde{z})/Q(z, \tilde{z})$, where P and Q are polynomials, is said to be a.m. at a given point $(\xi, \tilde{\xi}) \in \mathbb{R}^2$ if $Q(\xi, \tilde{\xi}) \neq 0$ and*

$(d^{j+k}\psi/d\tilde{z}^j dz^k)(\xi,\tilde{\xi}) \geq 0$, $k = 0, 1, \ldots$, $j = 0, 1, \ldots$. *For matrices whose elements are functions, we will say that they are a.m. at a given point* $(\xi,\tilde{\xi})$ *if each element is a.m. at* $(\xi,\tilde{\xi})$.

Observe that Definition 2.7 is an extension of the concept of an a.m. function given in [18, Definition 2.1].

As the name indicates, the concept of absolute monotonicity given in Definition 2.2 for an ARK method is closely related to the concept of absolute monotonicity for a function given in Definition 2.7.

LEMMA 2.8. *Consider the functions* $\mathsf{A}(\xi,\tilde{\xi})$, $\tilde{\mathsf{A}}(\xi,\tilde{\xi})$, *and* $\mathsf{e}(\xi,\tilde{\xi})$, *defined by* (2.4), (2.5), *and* (2.6), *respectively. Then the ARK method* $(\mathbb{A},\tilde{\mathbb{A}})$ *is a.m. at* $(\xi_0,\tilde{\xi}_0)$ *if and only if the functions* $\mathsf{A}(\xi,\tilde{\xi})$, $\tilde{\mathsf{A}}(\xi,\tilde{\xi})$, *and* $\mathsf{e}(\xi,\tilde{\xi})$ *are a.m. at* $(\xi_0,\tilde{\xi}_0)$.

Proof. Recall that from Definition 2.2, the ARK method $(\mathbb{A},\tilde{\mathbb{A}})$ is absolutely monotonic at $(\xi_0,\tilde{\xi}_0)$ if and only if $\mathsf{A}(\xi_0,\tilde{\xi}_0) \geq 0$, $\tilde{\mathsf{A}}(\xi_0,\tilde{\xi}_0) \geq 0$, and $\mathsf{e}(\xi_0,\tilde{\xi}_0) \geq 0$. The if part is trivial. For the only if part we simply have to observe that

$$\begin{aligned}
\frac{d\mathsf{A}(\xi,\tilde{\xi})}{d\xi} &= \mathsf{A}(\xi,\tilde{\xi})^2\,,\\
\frac{d\mathsf{A}(\xi,\tilde{\xi})}{d\tilde{\xi}} &= \tilde{\mathsf{A}}(\xi,\tilde{\xi})\,\mathsf{A}(\xi,\tilde{\xi})\,,\\
\frac{d\tilde{\mathsf{A}}(\xi,\tilde{\xi})}{d\xi} &= \mathsf{A}(\xi,\tilde{\xi})\,\tilde{\mathsf{A}}(\xi,\tilde{\xi})\,,
\end{aligned}$$

and hence from $\mathsf{A}(\xi_0,\tilde{\xi}_0) \geq 0$ and $\tilde{\mathsf{A}}(\xi_0,\tilde{\xi}_0) \geq 0$, after a recursion process, we obtain the absolute monotonicity of $\mathsf{A}(\xi,\tilde{\xi})$, $\tilde{\mathsf{A}}(\xi,\tilde{\xi})$ at $(\xi_0,\tilde{\xi}_0)$. We proceed in a similar way for $\mathsf{e}(\xi,\tilde{\xi})$. ∎

We go deeper into the concept of absolute monotonicity for a function in a given point. In the following lemma, $\overline{A}$ denotes the closure of the set A. See (2.7) for the definition of $\mathcal{L}(r,\tilde{r})$.

LEMMA 2.9. *Let* $\psi(z,\tilde{z}) = P(z,\tilde{z})/Q(z,\tilde{z})$ *be a rational function, where* P *and* Q *are polynomials. Suppose that* ψ *is a.m. at a given point* $(-r,-\tilde{r})$. *Assume too* $Q(z,\tilde{z}) \neq 0$ *on an open neighborhood* $\mathcal{N}$ *containing* $\overline{\mathcal{L}(r,\tilde{r})}$. *Then* ψ *is a.m. on* $\overline{\mathcal{L}(r,\tilde{r})}$.

Proof. The proof follows along the lines of Lemma 3.1 in [18]. The Taylor expansion of ψ at $(-r,-\tilde{r})$ gives

$$\psi(z,\tilde{z}) = \psi(-r,-\tilde{r}) + \frac{\partial\psi}{\partial z}(-r,-\tilde{r})(z+r) + \frac{\partial\psi}{\partial\tilde{z}}(-r,-\tilde{r})(\tilde{z}+\tilde{r}) + \cdots\,,$$

with $\psi(-r,-\tilde{r}) \geq 0$, $\frac{\partial\psi}{\partial z}(-r,-\tilde{r}) \geq 0$, $\frac{\partial\psi}{\partial\tilde{z}}(-r,-\tilde{r}) \geq 0$, ... which is valid in $\mathcal{N}$. To see this, we simply have to use the ideas in the proof of [18, Lemma 3.1]. In particular the Taylor expansion is valid for $(z,\tilde{z}) \in \overline{\mathcal{L}(r,\tilde{r})}$. Term by term differentiation shows that ψ is absolutely monotonic at $\overline{\mathcal{L}(r,\tilde{r})}$. ∎

LEMMA 2.10. *Let* $\psi(z,\tilde{z}) = P(z,\tilde{z})/Q(z,\tilde{z})$ *be a rational function, where* P *and* Q *are polynomials. Suppose that* ψ *is a.m. on* $\mathcal{L}(r,\tilde{r})$. *Then* ψ *is a.m. on* $\overline{\mathcal{L}(r,\tilde{r})}$.

Proof. The proof follows along the lines of Lemma 3.6 in [18]. We have to prove that ψ is a.m. at $(r,\tilde{r})$. If $(r,\tilde{r})$ is not a pole of ψ, then the absolute monotonicity follows from a limit argument. If $(r,\tilde{r})$ is a pole of ψ, then on a neighborhood of $(r,\tilde{r})$ on $\mathcal{L}(r,\tilde{r})$ we have $\psi(z,\tilde{z}) < 0$ or $D_{(z,\tilde{r}/r\,z)}\psi(z,\tilde{z}) < 0$, where $D_v f$ denotes the derivative of f in the direction v. In both cases we get a contradiction with the absolute monotonicity of ψ. ∎

For ARK methods, we are in a position to prove a result similar to Lemma 4.4 in [18].

PROPOSITION 2.11. *Consider an ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *and real positive numbers* r, $\tilde{r}$. *Then* $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) \geq (r, \tilde{r})$ *if and only if* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is a.m. at* $(\xi, \tilde{\xi}) = (-r, -\tilde{r})$ *and* $\mathbb{A} \geq 0$, $\tilde{\mathbb{A}} \geq 0$.

Proof. 1. We begin by assuming that $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) \geq (r, \tilde{r})$. Then $(\mathbb{A}, \tilde{\mathbb{A}})$ is a.m. at $(\xi, \tilde{\xi}) \in \mathcal{L}(r, \tilde{r})$. By Lemma 2.8 the functions $\mathsf{A}(\xi, \tilde{\xi})$, $\tilde{\mathsf{A}}(\xi, \tilde{\xi})$, and $\mathsf{e}(\xi, \tilde{\xi})$ are a.m. at $(\xi, \tilde{\xi}) \in \mathcal{L}(r, \tilde{r})$. We can now apply componentwise Lemma 2.10 to get that they are also a.m. at $(\xi, \tilde{\xi})$ with $\xi \in [-r, 0]$ and $\tilde{\xi} = \frac{r}{\tilde{r}}\,\xi$ and in particular at $(-r, -\tilde{r})$. Application of Lemma 2.8 once more gives the absolute monotonicity of $(\mathbb{A}, \tilde{\mathbb{A}})$ at $(r, \tilde{r})$. In particular the method is a.m. at $(\xi, \tilde{\xi}) = (0, 0)$, and hence $\mathbb{A} \geq 0$ and $\tilde{\mathbb{A}} \geq 0$.

2. We assume now that $(\mathbb{A}, \tilde{\mathbb{A}})$ is a.m. at $(\xi, \tilde{\xi}) = (-r, -\tilde{r})$ and $\mathbb{A} \geq 0$, $\tilde{\mathbb{A}} \geq 0$. Using Lemma 2.6 for $C = r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}}$, we get that $I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}}$ is nonsingular for all $(\xi, \tilde{\xi})$ with $\xi \in [-r, 0]$ and $\tilde{\xi} = \frac{\tilde{r}}{r}\,\xi$, and therefore the functions $\mathsf{A}(\xi, \tilde{\xi})$, $\tilde{\mathsf{A}}(\xi, \tilde{\xi})$, and $\mathsf{e}(\xi, \tilde{\xi})$ are well defined for $(\xi, \tilde{\xi}) = (\xi, \frac{\tilde{r}}{r}\,\xi)$ with $\xi \in [-r, 0]$. Hence, by Lemma 2.9 we get that the method is a.m. at $(\xi, \tilde{\xi})$ with $\xi \in [-r, 0]$ and $\tilde{\xi} = \frac{\tilde{r}}{r}\,\xi$, i.e., $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) \geq (r, \tilde{r})$. ∎

The relevance of Proposition 2.11 is that to prove the absolute monotonicity of an ARK method in the segment that connects the origin and the point $(-r, -\tilde{r})$, it is enough to check the absolute monotonicity of the method $(\mathbb{A}, \tilde{\mathbb{A}})$ at $(-r, -\tilde{r})$, and the nonnegativity of $\mathbb{A}$ and $\tilde{\mathbb{A}}$. Hence, to prove the absolute monotonicity of an ARK scheme in a region, it is enough to check the absolute monotonicity of the method at the curve of absolute monotonicity and the nonnegativity of the coefficient matrices.

We finish the section with a lemma that will justify the validity of the results for any convex function $\|\cdot\|$.

LEMMA 2.12. *Assume that the ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is a.m. at* $(-r, -\tilde{r})$ *with* $r, \tilde{r} \geq 0$. *Then*

$$0 \leq \mathsf{e}(-r, -\tilde{r}) \leq e\,, \tag{2.8}$$

$$0 \leq r\,\mathsf{A}(-r, -\tilde{r}) \leq E\,, \tag{2.9}$$

$$0 \leq \tilde{r}\,\tilde{\mathsf{A}}(-r, -\tilde{r}) \leq E\,, \tag{2.10}$$

$$\mathsf{e}(-r, -\tilde{r}) + r\,\mathsf{A}(-r, -\tilde{r})\,e + \tilde{r}\,\tilde{\mathsf{A}}(-r, -\tilde{r})\,e = e\,, \tag{2.11}$$

where E *denotes the matrix whose elements are equal to* 1 *and* $\mathsf{e}(\xi, \tilde{\xi})$, $\mathsf{A}(\xi, \tilde{\xi})$ *and* $\tilde{\mathsf{A}}(\xi, \tilde{\xi})$ *are given by* (2.4)–(2.6). *Hence, for* x, y, $z \in \mathbb{R}^{s+1}$, *componentwise, the expression*

$$\mathsf{e}(-r, -\tilde{r})\,x + r\,\mathsf{A}(-r, -\tilde{r})\,y + \tilde{r}\,\tilde{\mathsf{A}}(-r, -\tilde{r})\,z$$

is a convex combination of x_i, y_i, *and* z_i.

Proof. The lower bounds in (2.8)–(2.10) come from the absolute monotonicity at $(-r, -\tilde{r})$. To prove the upper bounds, we proceed as follows. As the method is a.m. at $(-r, -\tilde{r})$, we have

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}(r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})\,e \geq 0 \tag{2.12}$$

and as

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}(r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}}) = I - (I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}\,,$$

we can write (2.12) as

$$\left(I - (I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}\right) e \geq 0\,,$$

which gives the upper bound in (2.8). Furthermore, from

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} e \geq 0\,,$$

as

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} = I - (I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}(r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})\,,$$

we obtain

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}(r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})\, e \leq e\,. \tag{2.13}$$

The upper bounds in (2.9)–(2.10) come from inequality (2.13) and the fact that

$$(I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} r\mathbb{A} \geq 0\,, \qquad (I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} \tilde{r}\tilde{\mathbb{A}} \geq 0\,.$$

Finally, (2.11) is straightforward. ∎

3. Monotonicity for ARK methods. Once we have defined and studied the concept of absolute monotonicity for an ARK method $(\mathbb{A}, \tilde{\mathbb{A}})$, we are in a position to prove the main result that ensures monotonicity for the ARK method under certain stepsize restrictions.

THEOREM 3.1. *Assume that the ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is a.m. at* $(-r, -\tilde{r})$. *Then for*

$$h \leq r\,\frac{1}{\rho}, \qquad h \leq \tilde{r}\,\frac{1}{\tilde{\rho}} \tag{3.1}$$

it holds that

$$\|U_i\| \leq \|u_n\|\,, \qquad i = 1, \ldots, s\,, \qquad\qquad \|u_{n+1}\| \leq \|u_n\|\,.$$

Proof. The proof is similar to the one in [13]. The ARK method can be written as

$$U = e \otimes u_n + h(\mathbb{A} \otimes I)F(U) + h(\tilde{\mathbb{A}} \otimes I)\tilde{F}(U)\,. \tag{3.2}$$

Observe that the conditions on the problems imply, for h satisfying (3.1), that

$$\left\| U_i + \frac{h}{r} F(U_i) \right\| \leq \| U_i \|\,, \qquad \left\| U_i + \frac{h}{\tilde{r}} \tilde{F}(U_i) \right\| \leq \| U_i \|\,. \tag{3.3}$$

In formula (3.2) we add to both sides $((r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}}) \otimes I)U$, obtaining

$$(I + ((r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}}) \otimes I))\, U = e \otimes u_n + (\mathbb{A} \otimes I)(rU + hF(U)) + (\tilde{\mathbb{A}} \otimes I)(\tilde{r}U + h\tilde{F}(U))\,,$$

or equivalently

$$U = \mathsf{e}(-r, -\tilde{r}) \otimes u_n + (r\,\mathsf{A}(-r, -\tilde{r}) \otimes I) \left(U + \frac{h}{r} F(U)\right)$$

$$+ \left(\tilde{r}\,\tilde{\mathsf{A}}(-r, -\tilde{r}) \otimes I\right) \left(U + \frac{h}{\tilde{r}} \tilde{F}(U)\right)\,, \tag{3.4}$$

where $\mathsf{e}(\xi,\tilde{\xi})$, $\mathsf{A}(\xi,\tilde{\xi})$, and $\tilde{\mathsf{A}}(\xi,\tilde{\xi})$ are given by (2.4)–(2.6). If we take norms, the conditions on $\mathsf{e}(-r,-\tilde{r})$, $\mathsf{A}(-r,-\tilde{r})$, and $\tilde{\mathsf{A}}(-r,-\tilde{r})$ imply

$$[\|U\|] \le \mathsf{e}(-r,-\tilde{r}) \otimes \|u_n\| + (r\,\mathsf{A}(-r,-\tilde{r}) \otimes I) \left[\left\|U + \frac{h}{r} F(U)\right\|\right]$$

$$+ \left(\tilde{r}\,\tilde{\mathsf{A}}(-r,-\tilde{r}) \otimes I\right) \left[\left\|U + \frac{h}{\tilde{r}} \tilde{F}(U)\right\|\right], \tag{3.5}$$

where $[\|U\|] = (\|U_1\|, \ldots, \|U_s\|, \|u_{n+1}\|)^t \in \mathbb{R}^{s+1}$. Conditions (3.3) imply now

$$[\|U\|] \le \mathsf{e}(-r,-\tilde{r}) \otimes \|u_n\| + \left((r\mathsf{A}(-r,-\tilde{r}) + \tilde{r}\tilde{\mathsf{A}}(-r,-\tilde{r})) \otimes I\right) [\|U\|],$$

and hence

$$((I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} \otimes I)\,[\|U\|] \le ((I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1} e) \otimes \|u_n\|, \tag{3.6}$$

where we have used that $(I - r\mathsf{A}(-r,-\tilde{r}) - \tilde{r}\tilde{\mathsf{A}}(-r,-\tilde{r})) = (I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}})^{-1}$. We simply have to multiply (3.6) by $I + r\mathbb{A} + \tilde{r}\tilde{\mathbb{A}} \ge 0$ to get

$$[\|U\|] \le e \otimes \|u_n\|. \qquad \square$$

Remark 1.

(a) By Lemma 2.12, the derivation of inequality (3.5) from (3.4) is valid not only for norms $\|\cdot\|$ but also for any convex function $\psi : \mathbb{R}^m \to \mathbb{R}$.

(b) We have obtained a conditional monotonicity result. We are not interested in unconditional monotonicity because for RK methods it is known that explicit RK methods cannot be unconditionally monotone, and for implicit RK methods, unconditional monotonicity implies that the order of the method is at most one (see [18] or [12]). Hence IMEX methods cannot be unconditionally monotone. For unconditional monotone implicit ARK methods, the same order restriction holds, i.e., $p \le 1$, because to obtain an order p RK method, each RK method must have order p.

(c) If the method $\mathbb{A}$ is a.m. at $-r$, then the RK scheme is monotone under stepsize restriction $h \le r/\rho$. Similarly, if the method $\tilde{\mathbb{A}}$ is a.m. at $-r$, then the RK scheme is monotone under stepsize restriction $h \le \tilde{r}/\tilde{\rho}$. However, in order to obtain monotonicity for the ARK method $(\mathbb{A}, \tilde{\mathbb{A}})$ we require absolute monotonicity at $(-r,-\tilde{r})$. Depending on the shape of $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, sharper stepsize restrictions may occur to maintain monotonicity of the ARK method. In order to have a good monotone ARK scheme, not only must the RK methods $\mathbb{A}$ and $\tilde{\mathbb{A}}$ have a large radius of absolute monotonicity, but the methods also must be properly coupled.

(d) If the exact solution, the internal stages, and the numerical solution are in a subset $\mathcal{B} \subset \mathbb{R}^m$, then in Theorem 3.1 conditions (1.3)–(1.4) can be relaxed to

$$\left\| y + \frac{1}{\rho}\, f(y) \right\| \le \| y \|, \qquad \left\| y + \frac{1}{\tilde{\rho}}\, \tilde{f}(y) \right\| \le \| y \| \qquad \text{for all } y \in \mathcal{B}.$$

Theorem 3.1 gives us monotonicity under the stepsize restriction

$$h \le r\,\frac{1}{\rho}, \qquad h \le \tilde{r}\,\frac{1}{\tilde{\rho}},$$

with $(r, \tilde{r})$ such that the ARK method is a.m. at it. Hence, we should have ARK schemes $(\mathbb{A}, \tilde{\mathbb{A}})$ such that $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) \neq (0, 0)$ because only in this case do we have a positive stepsize restriction for monotonicity. For RK methods, in [18] it is proved that $R(\mathbb{A}) > 0$ if and only if $\mathbb{A} \geq 0$ and

$$\text{Inc}\,(\mathbb{A}^2) \leq \text{Inc}\,\mathbb{A}\,,$$

where Inc F is the incidence matrix of $F = (f_{ij})$, defined as Inc $F = (g_{ij})$, with $g_{ij} = 1$ if $f_{ij} \neq 0$ and $g_{ij} = 0$ if $f_{ij} = 0$.

Our next goal is to give algebraic criteria for $\partial \mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) \neq (0, 0)$. We begin by proving the following lemmas.

LEMMA 3.2. *Consider matrices $A = (a_{ij})$, $B = (b_{ij})$, $C = (c_{ij})$, and $D = (d_{ij})$ such that $A \geq 0$, $B \geq 0$, $C \geq 0$, $D \geq 0$. Assume that Inc $B \leq$ Inc C and Inc $(A\,C) \leq$ Inc D. Then*

$$\textit{Inc}\,(A\,B) \leq \textit{Inc}\,D\,.$$

Proof. If $(AB)_{ij} \neq 0$, then $a_{il} b_{lj} \neq 0$ for some l, and hence $a_{il} \neq 0$ and $b_{lj} \neq 0$. From $b_{lj} \neq 0$ and Inc $B \leq$ Inc C, we get that $c_{lj} \neq 0$. Now from $a_{il} \neq 0$ and $c_{lj} \neq 0$ we obtain that $(AC)_{ij} \neq 0$. Finally, from Inc $(A\,C) \leq$ Inc D we obtain that $d_{ij} \neq 0$. ∎

LEMMA 3.3. *If $\mathbb{A} \geq 0$, $\tilde{\mathbb{A}} \geq 0$, and*

$$\textit{Inc}\,(\mathbb{A}^2) \leq \textit{Inc}\,\mathbb{A}\,, \tag{3.7}$$

$$\textit{Inc}\,(\tilde{\mathbb{A}}\,\mathbb{A}) \leq \textit{Inc}\,\mathbb{A}\,, \tag{3.8}$$

then for any k_i, p_i, $i = 1, \ldots, n$, and for any n it holds that

$$\textit{Inc}\,\left((\mathbb{A})^{k_1}(\tilde{\mathbb{A}})^{p_1} \cdots (\mathbb{A})^{k_n}(\tilde{\mathbb{A}})^{p_n}\,\mathbb{A}\right) \leq \textit{Inc}\,\mathbb{A}\,. \tag{3.9}$$

Proof. Lemma 3.2 with $A = \tilde{\mathbb{A}}$ and $D = C = \mathbb{A}$ gives that Inc $B \leq$ Inc $\mathbb{A}$, together with (3.8), implies

$$\text{Inc}\,(\tilde{\mathbb{A}}\,B) \leq \text{Inc}\,\mathbb{A}\,. \tag{3.10}$$

Beginning with (3.8), we can successively use (3.10) for $B = \tilde{\mathbb{A}}\,\mathbb{A}$ to get Inc $(\tilde{\mathbb{A}}^2\mathbb{A}) \leq$ Inc $\mathbb{A}$, for $B = \tilde{\mathbb{A}}^2\,\mathbb{A}$ to get Inc $(\tilde{\mathbb{A}}^3\mathbb{A}) \leq$ Inc $\mathbb{A}$, and in this way obtain

$$\text{Inc}\,(\tilde{\mathbb{A}}^k\,\mathbb{A}) \leq \text{Inc}\,\mathbb{A}\,. \tag{3.11}$$

Similarly, Lemma 3.2 with $A = \mathbb{A}$ and $D = C = \mathbb{A}$ gives that Inc $B \leq$ Inc $\mathbb{A}$ and (3.7) imply

$$\text{Inc}\,(\mathbb{A}\,B) \leq \text{Inc}\,\mathbb{A}\,. \tag{3.12}$$

As (3.11) holds true, we can use (3.12) for $B = \tilde{\mathbb{A}}^k\,\mathbb{A}$ to get Inc $(\mathbb{A}\,\tilde{\mathbb{A}}^k\,\mathbb{A}) \leq$ Inc $\mathbb{A}$, for $B = \mathbb{A}\,\tilde{\mathbb{A}}^k\,\mathbb{A}$ to get Inc $(\mathbb{A}^2\tilde{\mathbb{A}}^k\,\mathbb{A}) \leq$ Inc $\mathbb{A}$, and in this way obtain

$$\text{Inc}\,(\mathbb{A}^p\tilde{\mathbb{A}}^k\,\mathbb{A}) \leq \text{Inc}\,\mathbb{A}\,.$$

Repeating this process, we arrive at (3.9). ∎

The following result gives algebraic criteria for obtaining $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}}) \neq (0,0)$.

PROPOSITION 3.4. *We have* $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}}) \neq (0,0)$ *if and only if* $\mathbb{A} \geq 0$, $\tilde{\mathbb{A}} \geq 0$, *and*

$$Inc\,(\mathbb{A}^2) \leq Inc\,\mathbb{A}\,, \tag{3.13}$$

$$Inc\,(\tilde{\mathbb{A}}\,\mathbb{A}) \leq Inc\,\mathbb{A}\,, \tag{3.14}$$

$$Inc\,(\tilde{\mathbb{A}}^2) \leq Inc\,\tilde{\mathbb{A}}\,, \tag{3.15}$$

$$Inc\,(\mathbb{A}\,\tilde{\mathbb{A}}) \leq Inc\,\tilde{\mathbb{A}}\,. \tag{3.16}$$

Proof. The proof is similar to the one in [13]. For real ξ, $\tilde{\xi}$ close to zero, the matrix $(I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})$ is nonsingular, and hence

$$(I-\xi\mathbb{A}-\tilde{\xi}\tilde{\mathbb{A}})^{-1}\mathbb{A} = \mathbb{A} + (\xi\mathbb{A}+\tilde{\xi}\tilde{\mathbb{A}})\mathbb{A} + \left(\xi^2\mathbb{A}^2 + \xi\tilde{\xi}\left(\mathbb{A}\tilde{\mathbb{A}}+\tilde{\mathbb{A}}\mathbb{A}\right) + \tilde{\xi}^2\tilde{\mathbb{A}}^2\right)\mathbb{A} + \cdots\,, \tag{3.17}$$

$$(I-\xi\mathbb{A}-\tilde{\xi}\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}} = \tilde{\mathbb{A}} + (\xi\mathbb{A}+\tilde{\xi}\tilde{\mathbb{A}})\tilde{\mathbb{A}} + \left(\xi^2\mathbb{A}^2 + \xi\tilde{\xi}\left(\mathbb{A}\tilde{\mathbb{A}}+\tilde{\mathbb{A}}\mathbb{A}\right) + \tilde{\xi}^2\tilde{\mathbb{A}}^2\right)\tilde{\mathbb{A}} + \cdots\,. \tag{3.18}$$

From (3.17) and (3.18) we see that $\mathbb{A} \geq 0$, $\tilde{\mathbb{A}} \geq 0$, and (3.13)–(3.16) are necessary conditions for (2.4) and (2.5) to hold on a left neighborhood of $\xi = 0$, $\tilde{\xi} = 0$. To see that they are also sufficient we use them and Lemma 3.3 to state that for any k_i, p_i, $i = 1,\ldots,n$, and for any n, it holds that

$$\begin{aligned}
\text{Inc}\,\left((\mathbb{A})^{k_1}(\tilde{\mathbb{A}})^{p_1}\cdots(\mathbb{A})^{k_n}(\tilde{\mathbb{A}})^{p_n}\,\mathbb{A}\right) &\leq \text{Inc}\,\mathbb{A},\\
\text{Inc}\,\left((\mathbb{A})^{k_1}(\tilde{\mathbb{A}})^{p_1}\cdots(\mathbb{A})^{k_n}(\tilde{\mathbb{A}})^{p_n}\,\tilde{\mathbb{A}}\right) &\leq \text{Inc}\,\tilde{\mathbb{A}}.
\end{aligned}$$

Hence, as $\mathbb{A} \geq 0$ and $\tilde{\mathbb{A}} \geq 0$, in (3.17) and (3.18) we get (2.4) and (2.5), respectively. Observe that inequality (2.6) always holds for r, $\tilde{r}$ close to zero. ∎

This criteria are extremely useful for checking if we have $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}}) \neq (0,0)$ for a given ARK method. Observe that conditions $\mathbb{A} \geq 0$ and (3.13) imply that $R(\mathbb{A}) > 0$. Similarly, conditions $\tilde{\mathbb{A}} \geq 0$ and (3.16) imply that $R(\tilde{\mathbb{A}}) > 0$. However, some extra coupling conditions, namely (3.14) and (3.15), are needed to get $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}}) \neq (0,0)$.

Example 1.

(a) The IMEX SSP2(3,3,2) stiffly accurate method [20] is

(3.19)

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$	0
$\mathbb{A}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$\frac{1}{4}$	$\frac{1}{4}$	0	0
$\frac{1}{4}$	0	$\frac{1}{4}$	0
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\tilde{\mathbb{A}}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(SSP stands for strong stability preserving). It can be computed that $R(\mathbb{A}) = 2$ and $R(\tilde{\mathbb{A}}) = \frac{12}{5}$. However, $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}}) = (0,0)$. In this case condition Inc $(\mathbb{A}\,\tilde{\mathbb{A}}) \leq$ Inc $(\tilde{\mathbb{A}})$ fails for the element $(1,2)$. We can change the implicit

method in (3.19) to

$$\begin{array}{c|ccc} \frac{1}{5} & \frac{1}{5} & 0 & 0 \\ \frac{3}{10} & \frac{1}{10} & \frac{1}{5} & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \tilde{\mathbb{A}} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}$$

It can be computed that $R(\tilde{\mathbb{A}}) = \frac{5}{9}(\sqrt{70} - 4)$. Now conditions (3.13)–(3.16) in Proposition 3.4 hold, and hence $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) \neq (0, 0)$.

(b) The L-stable IMEX SSP3(3,3,2) method [20] is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline \mathbb{A} & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \qquad \begin{array}{c|ccc} \gamma & \gamma & 0 & 0 \\ 1-\gamma & 1-2\gamma & \gamma & 0 \\ \frac{1}{2} & \frac{1}{2}-\gamma & 0 & \gamma \\ \hline \tilde{\mathbb{A}} & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}$$

with $\gamma = 1 - \frac{1}{\sqrt{2}}$. It can be computed that $R(\mathbb{A}) = 1$ and $R(\tilde{\mathbb{A}}) = \frac{30-24\sqrt{2}}{215\sqrt{2}-304}$. However, $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) = (0, 0)$. Now condition Inc $(\mathbb{A}\,\tilde{\mathbb{A}}) \leq$ Inc $(\tilde{\mathbb{A}})$ fails for the element $(3, 2)$.

3.1. Monotonicity for ARK methods in the Shu–Osher representation. Monotonicity issues for RK schemes written in the Shu–Osher form have been widely studied during the past years (see, e.g., [9, 26, 6, 13] and the references therein). If we have an RK method written in the Shu–Osher representation, the derivation of monotonicity is conceptually much easier. However, as the Shu–Osher form of an RK scheme is not unique, a previous problem, namely, the optimal Shu–Osher representation of an RK method, arises. This and some other related problems were studied in [13, 6]. In this section we briefly extend these topics for ARK methods.

We consider the following extension of the Shu–Osher form:

$$U = \alpha \otimes u_n + ((\Lambda + \tilde{\Lambda}) \otimes I)U + h\,(\Gamma \otimes I)F(U) + h\,(\tilde{\Gamma} \otimes I)\tilde{F}(U)\,, \tag{3.20}$$

where $\alpha \in \mathbb{R}^{s+1}$, and Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ are $(s+1) \times (s+1)$ matrices such that $(\Lambda + \tilde{\Lambda})\, e + \alpha = e$, the matrix $I - \Lambda - \tilde{\Lambda}$ is invertible, and the last column in Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ is zero. Scheme (3.20) is an ARK method with $\mathbb{A} = (I - \Lambda - \tilde{\Lambda})^{-1}\Gamma$ and $\tilde{\mathbb{A}} = (I - \Lambda - \tilde{\Lambda})^{-1}\tilde{\Gamma}$, and therefore we will refer to it as a representation of the ARK method.

A monotonicity result for ARK schemes in the form of (3.20) can be given. We state the following proposition, whose proof is omitted because it is similar to that in [13, Proposition 3.9].

PROPOSITION 3.5. *Consider a method in the form of* (3.20) *such that*

$$(I - (\Lambda + \tilde{\Lambda}))\, e \geq 0\,, \quad \Gamma \geq 0\,, \quad \tilde{\Gamma} \geq 0\,, \quad \Lambda - c\,\Gamma \geq 0\,, \quad \tilde{\Lambda} - \tilde{c}\,\tilde{\Gamma} \geq 0\,,$$

for some $c, \tilde{c} \geq 0$. *Then for* $h \leq \min\{c/\rho, \tilde{c}/\tilde{\rho}\}$ *it holds that*

$$\|U_i\| \leq \|u_n\|\,, \qquad i = 1, \ldots, s\,, \qquad\qquad \|u_{n+1}\| \leq \|u_n\|\,.$$

As in Theorem 3.1, Proposition 3.5 ensures monotonicity for ARK schemes. Apparently there is no relationship between the conditions imposed in both of them. However, there is a close link between them.

PROPOSITION 3.6. *Consider an ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$.

1. *Assume that the ARK method can be written as* $\mathbb{A} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\Gamma$, $\tilde{\mathbb{A}} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\tilde{\Gamma}$, *where*

$$(I - (\Lambda + \tilde{\Lambda}))\, e \geq 0\,, \quad \Gamma \geq 0\,, \quad \tilde{\Gamma} \geq 0\,, \quad \Lambda - c\,\Gamma \geq 0\,, \quad \tilde{\Lambda} - \tilde{c}\,\tilde{\Gamma} \geq 0\,,$$

   and $I - (\Lambda + \tilde{\Lambda} - c\,\Gamma - \tilde{c}\,\tilde{\Gamma})$ *is invertible for some coefficients* $c, \tilde{c} \geq 0$. *Then the method is absolutely monotonic at* $(-c, -\tilde{c})$.
2. *If the method is absolutely monotonic at* $(-c, -\tilde{c})$, *then for*

$$\begin{aligned}
\Lambda &= c\,(I + c\,\mathbb{A} + \tilde{c}\,\tilde{\mathbb{A}})^{-1}\mathbb{A}\,,\\
\tilde{\Lambda} &= c\,(I + c\,\mathbb{A} + \tilde{c}\,\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}}\,,\\
\Gamma &= \mathbb{A} - (\Lambda + \tilde{\Lambda})\mathbb{A}\,,\\
\tilde{\Gamma} &= \tilde{\mathbb{A}} - (\Lambda + \tilde{\Lambda})\tilde{\mathbb{A}}\,,
\end{aligned}$$

   we can write $\mathbb{A} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\Gamma$, $\tilde{\mathbb{A}} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\tilde{\Gamma}$, *with*

$$(I - (\Lambda + \tilde{\Lambda}))\, e \geq 0\,, \quad \Gamma \geq 0\,, \quad \tilde{\Gamma} \geq 0\,, \quad \Lambda - c\,\Gamma = 0\,, \quad \tilde{\Lambda} - \tilde{c}\,\tilde{\Gamma} = 0\,.$$

Proof. The proofs of parts 1 and 2 are similar to the ones in [13, Propositions 3.11 and 3.12], respectively. An auxiliary lemma, similar to Lemma 3.10 in [13], is required. ∎

3.2. Monotonicity for perturbed RK methods. Over the past few years [23, 24, 10, 21, 26] (see [9, 25] for a review) a great effort has been made to develop high order methods satisfying (1.8)–(1.9) when the forward Euler discretization of (1.1) satisfies (1.9), i.e.,

$$\|u_n + h f(u_n)\| \leq \|u_n\| \qquad \text{for } h \leq \Delta t_{FE}\,. \tag{3.21}$$

These are SSP methods (or total variation diminishing (TVD) methods). The class of ODEs considered in this context arises from a method of lines approximation of hyperbolic conservation laws. A simple numerical example given in [9] shows that the use of non-SSP methods for the time discretization of these ODEs produces an undesirable overshot. As the forward Euler method has the drawback of a low order of accuracy, higher order SSP methods are of great interest. The idea in [24, 23] is to derive conditions (1.8)–(1.9) from condition (1.9) for the forward Euler method by means of convex combinations.

In [23], the Shu and Osher representations are used to write explicit RK methods $\mathbb{A}$. As pointed out in [13], the Shu and Osher representations with positive coefficients correspond to $R(\mathbb{A}) > 0$, whereas the case with negative coefficients corresponds to RK methods with $R(\mathbb{A}) = 0$. In this case, together with the original problem $y' = f(y)$ with $f(y)$ such that

$$\left\| y - \frac{1}{\rho}\, f(y) \right\| \leq \| y \|\,,$$

an auxiliary problem

$$y' = -\tilde{f}(y)$$

such that

$$\left\| y - \frac{1}{\rho}\, \tilde{f}(y) \right\| \leq \| y \|$$

is also considered. The Shu and Osher representations with negative coefficients are interpreted in [13] as perturbations of an RK method. Given an RK method $\mathbb{A}$, the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is

$$U^{(p)} = e \otimes u_n^{(p)} + h(\mathbb{A} \otimes I) F(U^{(p)}) + h(\tilde{\mathbb{A}} \otimes I) \left(F(U^{(p)}) - \tilde{F}(U^{(p)}) \right)$$

or

$$U^{(p)} = e \otimes u_n^{(p)} + h \left((\mathbb{A} + \tilde{\mathbb{A}}) \otimes I \right) F(U^{(p)}) + h(\tilde{\mathbb{A}} \otimes I) \left(-\tilde{F}(U^{(p)}) \right) .$$

Recall that these equations correspond to the "additive" RK method $(\mathbb{A} + \tilde{\mathbb{A}}, \mathbb{A})$. Observe too that both functions f and $-\tilde{f}$ have the same stiffness properties, i.e., $\rho = \tilde{\rho}$.

Monotonicity issues for perturbed RK methods were studied in [13]. The tool used was the radius of absolute monotonicity for a perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ [13, Definition 3.1]. It turns out that Definition 3.1 in [13] corresponds to the concept of absolute monotonicity of the additive ARK method $(\mathbb{A} + \tilde{\mathbb{A}}, \mathbb{A})$ at (ξ, ξ) defined in this paper. We consider only values of the form (ξ, ξ) because we are dealing with f and $-\tilde{f}$ such that $\rho = \tilde{\rho}$.

A monotonicity result for a perturbed RK method was proved in [13, Theorem 3.5]. Interpreting the perturbed RK method as an "additive" RK, monotonicity under the stepsize restriction given in Theorem 3.5 in [13] can also be obtained by applying Theorem 3.1 of this paper.

Finally, algebraic criteria for nonnull stepsize restrictions, i.e., for the existence of $r > 0$ such that $(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$ is a.m. at (r, r), were also given in [13, Proposition 3.6]. More precisely, it was proved that this situation holds if and only if

$$\tilde{\mathbb{A}} \geq 0\,, \ \mathbb{A} + \tilde{\mathbb{A}} \geq 0\,, \tag{3.22}$$

$$\text{Inc}\,((\mathbb{A} + 2\tilde{\mathbb{A}})(\mathbb{A} + \tilde{\mathbb{A}})) \leq \text{Inc}\,(\mathbb{A} + \tilde{\mathbb{A}})\,, \tag{3.23}$$

$$\text{Inc}\,((\mathbb{A} + 2\tilde{\mathbb{A}})\tilde{\mathbb{A}}) \leq \text{Inc}\,(\tilde{\mathbb{A}})\,. \tag{3.24}$$

If we apply the algebraic criteria given in Proposition 3.4 for the "additive" RK method $(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$, we obtain $\partial\mathcal{R}(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}}) \neq (0, 0)$ if and only if (3.22) and

$$\text{Inc}\,((\mathbb{A} + \tilde{\mathbb{A}})^2) \leq \text{Inc}\,(\mathbb{A} + \tilde{\mathbb{A}})\,, \tag{3.25}$$

$$\text{Inc}\,(\tilde{\mathbb{A}}\,(\mathbb{A} + \tilde{\mathbb{A}})) \leq \text{Inc}\,(\mathbb{A} + \tilde{\mathbb{A}})\,, \tag{3.26}$$

$$\text{Inc}\,(\tilde{\mathbb{A}}^2) \leq \text{Inc}\,(\tilde{\mathbb{A}})\,, \tag{3.27}$$

$$\text{Inc}\,((\mathbb{A} + \tilde{\mathbb{A}})\,\tilde{\mathbb{A}}) \leq \text{Inc}\,(\tilde{\mathbb{A}})\,. \tag{3.28}$$

It is straightforward to prove that (3.25)–(3.26) imply (3.23), and (3.27)–(3.28) imply (3.24).

Consequently, the study done in [13] fits completely within the one done in this paper.

4. Examples. The next step is to analyze the regions of a.m. $\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ for some ARK methods $(\mathbb{A},\tilde{\mathbb{A}})$ from the literature with $R(\mathbb{A})>0$, $R(\tilde{\mathbb{A}})>0$. The study is summarized in the following table. The coefficients for the different methods can be seen in the appendix.

Method	$R(\mathbb{A})$	$R(\tilde{\mathbb{A}})$	$\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$
ARS(1,1,1)	1	∞	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\}$
SP(1,1,1)	1	∞	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\}$
Störmer–Verlet	2	2	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 2\,,\ 0\le \tilde{r}\le 2\}$
Peaceman–Rachford	2	2	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 2\,,\ 0\le \tilde{r}\le 2\}$
IMEX trapezoidal (7.1)	1	2	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\le 2(1-r)\}$
IMEX trapezoidal (7.2)	1	1	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\le 1-r\}$
IMEX (7.3)	2	$\frac{2}{7}(6+\sqrt{57})$	$\left\{(r,\tilde{r}) \,\vert\, 0\le r\le 2\,,\ 0\le \tilde{r}\le \frac{6+\sqrt{57}}{7}(2-r)\right\}$
CRJ	$\frac{2}{3}$	$\frac{4}{5}$	$\{(r,\tilde{r}) \,\vert\, 0\le r\le \frac{2}{3}\,,\ 0\le \tilde{r}\le \frac{2}{5}(2-3r)\}$
IMEX θ-method ($\theta=1/2$)	1	2	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\le 2(1-r)\}$
ASIRK-2A	1	$\frac{4}{3}$	$\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\le \frac{4}{3}(1-r)\}$
SSP2(2,2,2)	1	$1+\sqrt{2}$	$\left\{(r,\tilde{r}) \,\vert\, 0\le r\le 1\,,\ 0\le \tilde{r}\le \sqrt{2}(1-r)\right\}$

Observe that for some ARK methods $(\mathbb{A},\tilde{\mathbb{A}})$, namely Ascher–Ruuth–Spiteri scheme ARS(1,1,1), the splitting scheme SP(1,1,1), Störmer–Verlet, and Peaceman–Rachford, the two RK methods are perfectly coupled in the sense that the region of absolute monotonicity is the greatest one, i.e., it is the cartesian product of the regions of absolute monotonicity of both methods.

For some other methods, namely IMEX trapezoidal (7.1) and (7.2), IMEX (7.3), Caflisch–Russo–Jin (CRJ), IMEX θ-method with $\theta=1/2$, and additive semi-implicit Runge–Kutta (ASIRK-2A), the region of absolute monotonicity for the ARK methods is the triangle with vertices the points $(0,0)$, $(0,R(\mathbb{A}))$, and $(R(\tilde{\mathbb{A}}),0)$.

However, for some other methods, the region of absolute monotonicity is smaller than this triangle. For example for the method SSP2(2,2,2) we have found that $R(\mathbb{A})=\sqrt{2}$ and $R(\tilde{\mathbb{A}})=1+\sqrt{2}$, but the region of absolute monotonicity is not the triangle with vertices $(0,0)$, $(0,1)$, and $(1+\sqrt{2},0)$ but the one with vertices $(0,0)$, $(0,1)$, and $(\sqrt{2},0)$.

Besides the cartesian product of both regions, and triangles, some other shapes are possible for the region of absolute monotonicity.

Example 2. We consider the modified IMEX SSP2(3,3,2) method from Example 1 with $R(\mathbb{A})=2$ and $R(\tilde{\mathbb{A}})=\frac{5}{9}(\sqrt{70}-4)$. It can be computed that

$$\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})=\left\{(r,\tilde{r}) \,\vert\, 0\le r\le 2\,,\ 0\le \tilde{r}\le \phi(r)\right\},$$

where

$$\phi(r)=\begin{cases} \frac{1}{4}(-28+9r)+\frac{1}{4}\sqrt{1264-984r+201r^2} & \text{if} \quad 0\le r\le \frac{1}{64}(119-\sqrt{721})\,, \\ \\ \frac{5}{36}(-16+r)+\frac{5\sqrt{7}}{36}\sqrt{160-128r+31r^2} & \text{if} \quad \frac{1}{64}(119-\sqrt{721})\le r\le 2\,. \end{cases}$$

5. Numerical experiments.

5.1. Experiment 1. We have considered the following PDE of hyperbolic type [18]:

$$\begin{cases} \dfrac{\partial}{\partial t}V(\xi,t) = \dfrac{\partial}{\partial \xi}\left[a(\xi,t)V(\xi,t)\right], \\ V(0,t) = 0\,, \qquad V(\xi,0) = v_0(\xi) \qquad (0 \le \xi \le 1,\, t \ge 0), \end{cases} \tag{5.1}$$

where $a(\xi,t) = -\cos(20\xi + 80t)$, $v_0(\xi) = \xi^2\left(e^{(\xi/200)} - 3\right)$. As $v_0(\xi)$ is continuously differentiable on $[0,1]$ and $v_0(0) = v_0'(0) = 0$, it can be proved that this problem has a unique classical solution V with

$$\int_0^1 |V(\xi,t)|\, d\xi \le \int_0^1 |v_0(\xi)|\, d\xi \qquad \text{for all } t \ge 0\,.$$

We discretize the spatial derivative in (5.1) with backward differences, obtaining the ODE

$$u(t)' = L(t)\, u(t), \tag{5.2}$$

where $L(t)$ satisfies

$$\left\| u + \frac{1}{\rho}\, L(t)\, u \right\| \le \| u \|$$

for $\rho = (\Delta\xi)^{-1}$, with $\| \cdot \|$ defined as

$$\| x \| = \Delta\xi \sum_{i=1}^{n} | x_i |\,.$$

Hence the solution of (5.2) is monotone.

For the numerical solution of (5.2) we have considered the RK methods $\mathbb{A}$ and $\tilde{\mathbb{A}}$ in the IMEX trapezoidal scheme (7.1), for which $R(\mathbb{A}) = 1$ and $R(\tilde{\mathbb{A}}) = 2$. For an RK method $\mathbb{A}$, the stepsize restriction for monotonicity is given by

$$h \le \frac{1}{\rho}\, R(\mathbb{A})\,.$$

Hence, for $\Delta\xi = 1/50$ the explicit method $\mathbb{A}$ is monotone under stepsize restriction $h \le 1 \times 1/50$, whereas the implicit method $\tilde{\mathbb{A}}$ is monotone for $h \le 2 \times 1/50$. However, in our numerical computations we have found that for this initial value problem these bounds are not sharp. For example, for $h = \frac{1}{20}$ both methods are monotone. In Figure 5.1 we show the values of $\| u_n \|$ with the explicit and implicit RK methods for $t \in [0,2]$ with that stepsize.

In order to test monotonicity for the IMEX method, we have considered the artificial problem

$$U(t)' = L(t)U(t) + L(t)U(t) \tag{5.3}$$

with $L(t)$ the same matrix as in (5.2). When the IMEX method is applied to (5.3), the stepsize restrictions for monotonicity for each method do not ensure monotonicity for the IMEX scheme. Figure 5.2 shows the numerical results for $\| u_n \|$ for $t \in [0,2]$ for the IMEX method with $h = \frac{1}{20}$. Observe that, although with this stepsize we obtain monotonicity for each method (see Figure 5.1), monotonicity fails for the IMEX scheme. Monotonicity is obtained under a more severe stepsize restriction.

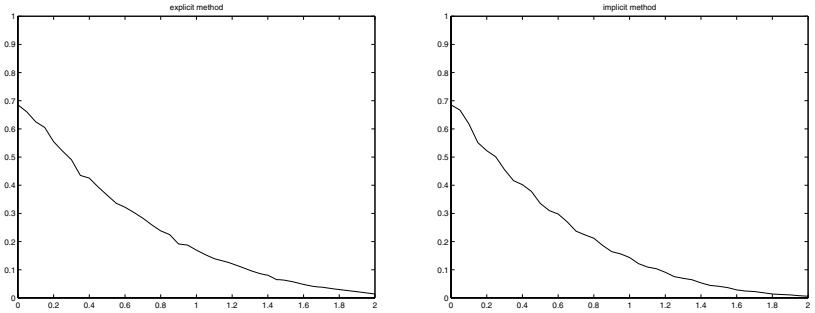


FIG. 5.1. *Left: explicit RK method. Right: implicit RK method.*

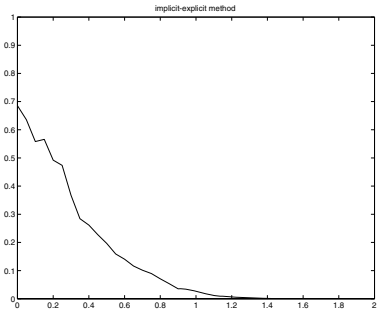


FIG. 5.2. *IMEX method.*

5.2. Experiment 2. We consider the Broadwell model [3], a hyperbolic system with relaxation. We will show how the above results can be used to obtain monotonicity for the entropy function. The problem is

$$\begin{aligned} \partial_t f + \partial_x f &= \frac{1}{\varepsilon}\left(h^2 - fg\right), \\ \partial_t h &= -\frac{1}{\varepsilon}\left(h^2 - fg\right), \\ \partial_t g - \partial_x g &= \frac{1}{\varepsilon}\left(h^2 - fg\right), \end{aligned} \tag{5.4}$$

where ε is a small parameter. We assume periodic boundary conditions. For the continuous problem there exists a function that is monotonically decreasing, namely

$$\mathcal{H}(t) = \int \hat{\mathcal{H}}(x,t)\,dx\,, \tag{5.5}$$

where

$$\hat{\mathcal{H}}(x,t) = f(x,t)\log f(x,t) + 2h(x,t)\log h(x,t) + g(x,t)\log g(x,t)\,.$$

We introduce spatial grid points $x_{j+\frac{1}{2}}$, $j = \ldots, -1, 0, 1, \ldots$ with uniform mesh spacing $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ for all j. Spatial discretization of (5.4) with the upwind scheme

gives

$$\begin{aligned} \frac{d}{dt} f_j + \frac{f_j - f_{j-1}}{\Delta x} &= \frac{1}{\varepsilon}\left(h_j^2 - f_j g_j\right), \\ \frac{d}{dt} h_j &= -\frac{1}{\varepsilon}\left(h_j^2 - f_j g_j\right), \\ \frac{d}{dt} g_j - \frac{g_{j+1} - g_j}{\Delta x} &= \frac{1}{\varepsilon}\left(h_j^2 - f_j g_j\right). \end{aligned} \tag{5.6}$$

This problem is an additive ODE in the form of (1.1) with the additive terms

$$F_1 = \begin{pmatrix} -\dfrac{f_j - f_{j-1}}{\Delta x} \\ 0 \\ \dfrac{g_{j+1} - g_j}{\Delta x} \end{pmatrix}, \qquad F_2 = \frac{1}{\varepsilon}\left(h_j^2 - f_j g_j\right) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}. \tag{5.7}$$

The discrete function for (5.5) is defined by

$$\mathcal{H}_{\Delta x}[f_j(t), h_j(t), g_j(t)] = \Delta x \sum_j \left(f_j(t)\log f_j(t) + 2h_j(t)\log h_j(t) + g_j(t)\log g_j(t)\right),$$

which is also a decreasing function. Hence, if f_j^n, h_j^n, and g_j^n denote the numerical approximations of $f_j(t), h_j(t)$, and $g_j(t)$ at time t_n, respectively, we should have $\mathcal{H}^{n+1} \le \mathcal{H}^n$, where now

$$\mathcal{H}^n = \mathcal{H}_{\Delta x}[f_j^n, h_j^n, g_j^n] = \Delta x \sum_j \left(f_j^n \log f_j^n + 2h_j^n \log h_j^n + g_j^n \log g_j^n\right).$$

In order to apply the results from section 3, we have to compute the parameters ρ and $\tilde{\rho}$ for the functions F_1 and F_2 in (5.7). Whereas the computation of the method parameter is not a difficult task, the sharp computation of problem parameters may be a hard task. In this case, for the function F_1 we have to find the values $\rho > 0$ such that

$$\mathcal{H}_{\Delta x}\left[f_j - \frac{1}{\rho}\frac{f_j - f_{j-1}}{\Delta x}, h_j, g_j + \frac{1}{\rho}\frac{g_{j+1} - g_j}{\Delta x}\right] \le \mathcal{H}_{\Delta x}[f_j, h_j, g_j].$$

It is straightforward to obtain that these inequalities hold for $\rho = 1/\Delta x$. For the function F_2 in (5.7), a value $\tilde{\rho} > 0$ such that

$$\mathcal{H}_{\Delta x}\left[f_j + \frac{1}{\tilde{\rho}\,\varepsilon}(h_j^2 - f_j g_j), h_j - \frac{1}{\tilde{\rho}\,\varepsilon}(h_j^2 - f_j g_j), g_j + \frac{1}{\tilde{\rho}\,\varepsilon}(h_j^2 - f_j g_j)\right] \le \mathcal{H}_{\Delta x}[f_j, h_j, g_j] \tag{5.8}$$

for all f, h, g such that $0 \le \Delta x\,(f + 2h + g) \le K_0$ (see Remark 1(d)) was found in [14], where K_0 is given by

$$K_0 = \Delta x \sum_j \left(f_j^0 + 2h_j^0 + g_j^0\right).$$

This value was $\tilde{\rho} = K_0/(\varepsilon\,\Delta x)$. Observe that in order to compute the discrete entropy function, the numerical solution should be positive. Hence, some extra stepsize restrictions may appear due to numerical positivity preservation.

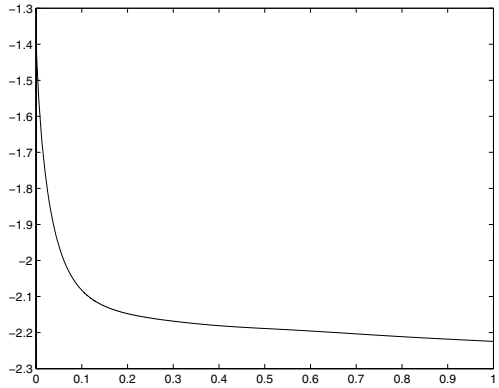


FIG. 5.3. *Entropy function for* (5.6) *($\varepsilon = 0.1$, $\Delta x = 0.01$, and $h = 0.0002$).*

According to Theorem 3.1, for $(r, \tilde{r})$ in the region of absolute monotonicity, we obtain monotonicity for the entropy function under the stepsize restrictions $h \le r\,\Delta x$, $h \le \tilde{r}\,\varepsilon\,\Delta x/K_0$. If we consider the CRJ scheme, we have to compute the intersection of the line connecting $(0, \frac{4}{5}\frac{\varepsilon\,\Delta x}{K_0})$ and $(\frac{2}{3}\Delta x, 0)$ with the bisectrix. In this way we obtain the stepsize restriction

$$h \le \frac{4\Delta x\,\varepsilon}{5K_0 + 6\,\varepsilon}\,. \tag{5.9}$$

We have integrated this problem with the CRJ scheme for $\varepsilon = 0.1$, $\Delta x = 0.01$, and $h = 0.0002$. With these parameters, stepsize restriction (5.9) is satisfied. We have checked that under this stepsize restriction, the numerical solution is positive. In Figure 5.3 we show the discrete entropy function. As expected, it is a decreasing function.

However, from our numerical experiments we have observed that the stepsize restriction (5.9) is not optimal. As pointed out above, the parameter $\tilde{\rho}$ has been obtained, imposing (5.8) for all f, h, and g such that $0 \le \Delta x\,(f + 2h + g) \le K_0$. However, for this problem, $h^2 - f\,g = \vartheta(\varepsilon)$, and this property has not been taken into account. A detailed study, out of the scope of this paper, should be done in order to find better problem parameters. In this context, the concept of the AP method introduced in [19] is an important reference.

6. Conclusions and future work. In this paper we have studied monotonicity issues for ARK methods $(\mathbb{A}, \tilde{\mathbb{A}})$. A new definition of absolute monotonicity for ARK methods has been given and some of its properties have been investigated. It has been proved that for $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) \neq 0$, it is possible to obtain monotonicity under nontrivial stepsize restrictions. As expected, monotonicity for each RK method does not ensure monotonicity for the ARK method. However, some classical methods in the PDE context, such as the Peaceman–Rachford method, are perfectly coupled and there are no extra stepsize restrictions on the use of the two RK methods as an ARK method.

With regard to the new concept $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ at least two questions remain open for future work. The first one is how to construct $(\mathbb{A}, \tilde{\mathbb{A}})$ such that $R(\mathbb{A}, \tilde{\mathbb{A}})$ is as great as possible. The second one is how to extend this concept for additive RK methods constructed with k RK methods.

7. Appendix. In this section we give the coefficients of some methods from the literature.

1. The IMEX ARS(1,1,1) method [2] (or the ASIRK-1A method in [27]) is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline \mathbb{A} & 1 & 0 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline \tilde{\mathbb{A}} & 0 & 1 \end{array}$$

2. The Störmer–Verlet method (or Lobatto IIIA–IIIB pair) [11] is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline \mathbb{A} & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \tilde{\mathbb{A}} & \frac{1}{2} & \frac{1}{2} \end{array}$$

3. The SP(1,1,1) L-stable method [19] is

$$\begin{array}{c|c} 0 & 0 \\ \hline \mathbb{A} & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline \tilde{\mathbb{A}} & 1 \end{array}$$

4. The Peaceman–Rachford ARK method [15] is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ \hline \mathbb{A} & \frac{1}{2} & 0 & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 1 & 0 \\ \hline \tilde{\mathbb{A}} & 0 & 1 & 0 \end{array}$$

5. There are different ways of combining the implicit and explicit trapezoidal rules. One way is by the IMEX method [15, p. 391]

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline \mathbb{A} & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline \tilde{\mathbb{A}} & \frac{1}{2} & \frac{1}{2} \end{array} \tag{7.1}$$

Another way is by the IMEX method [15, p. 392]

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \mathbb{A} & \frac{1}{2} & \frac{1}{2} & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ \hline \tilde{\mathbb{A}} & \frac{1}{2} & 0 & \frac{1}{2} \end{array} \tag{7.2}$$

6. In [20] different SSP$k(s,\sigma,p)$ IMEX methods are constructed. In the notation, SSP$k(s,\sigma,p)$, s, and σ are the numbers of stages of the explicit and implicit method, respectively, and k is the order of the SSP method (the explicit one).

The L-stable IMEX SSP2(2,2,2) method is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline \mathbb{A} & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline \tilde{\mathbb{A}} & \frac{1}{2} & \frac{1}{2} \end{array}$$

with $\gamma = 1 - \frac{1}{\sqrt{2}}$.

7. In [1] the following ARK scheme is considered:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 1 & 0 \\ \hline \mathbb{A} & 0 & 1 & 0 \end{array} \qquad \begin{array}{c|ccc} \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & 0 \\ \frac{1}{2} & \alpha & \frac{1-2\alpha}{2} & 0 \\ 1+\frac{\alpha}{2} & \alpha & 1-2\alpha & \frac{\alpha}{2} \\ \hline \tilde{\mathbb{A}} & \alpha & 1-2\alpha & \alpha \end{array} \tag{7.3}$$

We have studied it for $\alpha = \frac{1}{6}\left(9 - \sqrt{57}\right)$.

8. The second order IMEX method in [3] is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \hat{\alpha} & \hat{\alpha} & 0 & 0 \\ \eta\,(\hat{\alpha}+\hat{\beta}) & \eta\,\hat{\alpha} & \eta\,\hat{\beta} & 0 \\ \hline \mathbb{A} & \eta\,\hat{\alpha} & \eta\,\hat{\beta} & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma+\beta & \gamma & \beta & 0 \\ \eta\,\mu\,(\gamma+\beta) & \eta\,\gamma & \eta\,\beta & \mu \\ \hline \tilde{\mathbb{A}} & \eta\,\gamma & \eta\,\beta & \mu \end{array}$$

with

$$\beta = \frac{2\mu - 1}{2(\mu - 1)}, \qquad \gamma = -\frac{2\mu^2 - 2\mu + 1}{2\mu(\mu - 1)}, \qquad \eta = -2\mu(\mu - 1),$$
$$\hat{\alpha} = \frac{1}{2\mu}, \qquad \hat{\beta} = \frac{-1}{2(\mu - 1)}.$$

With $\mu = 1/3$, we obtain $\alpha \geq 0$, $\beta > 0$, and $\mu > 0$. We will refer to it as the CRJ IMEX method.

9. The IMEX θ-method [15, p. 383] is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline \mathbb{A} & 1 & 0 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline \tilde{\mathbb{A}} & 1-\theta & \theta \end{array}$$

Observe that for $\theta = 1$, we obtain the ARS(1,1,1) method above.

10. IMEX methods of the form

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ a_1 & a_1 & 0 & 0 & 0 \\ b_{21} & b_{21} & 0 & 0 & 0 \\ c_{21}+a_2 & c_{21} & 0 & a_2 & 0 \\ \hline \mathbb{A} & w_1 & 0 & w_2 & 0 \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ a_1 & 0 & a_1 & 0 & 0 \\ b_{21} & 0 & b_{21} & 0 & 0 \\ c_{21}+a_2 & 0 & c_{21} & 0 & a_2 \\ \hline \tilde{\mathbb{A}} & 0 & w_1 & 0 & w_2 \end{array}$$

are considered in [27]. Order 2 is obtained, e.g., for

$$w_1 = \tfrac{1}{2}\,, \quad w_2 = \tfrac{1}{2}\,, \quad b_{21} = 1\,,$$
$$a_1 = \tfrac{1}{4}\,, \quad a_2 = \tfrac{1}{3}\,, \quad c_{21} = \tfrac{5}{12}\,.$$

We will refer to this method as ASIRK-2A.

REFERENCES

[1] A. L Araújo, A. Murua, and J. M. Sanz-Serna, *Symplectic methods based on decompositions*, SIAM J. Numer. Anal., 34 (1997), pp. 1926–1947.

[2] U. M. Ascher, S. J. Ruuth, and R. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[3] R. E. Caflisch, S. Jin, and G. Russo, *Uniformly accurate schemes for hyperbolic systems with relaxation*, SIAM J. Numer. Anal., 34 (1997), pp. 246–281.

[4] M. P. Calvo, J. De Frutos, and J. Novo, *Linearly implicit Runge-Kutta methods for advection-reaction-diffusion equations*, Appl. Numer. Math., 37 (2001), pp. 535–549.

[5] L. Ferracina and M. N. Spijker, *Stepsize restrictions for the total-variation-diminishing property in general Runge–Kutta methods*, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

[6] L. Ferracina and M. N. Spijker, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp., 74 (2005), pp. 201–219.

[7] A. Gerisch, D. F. Griffiths, R. Weiner, and M. A. J. Chaplain, *A positive splitting method for mixed hyperbolic-parabolic systems*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 152–168.

[8] A. Gerisch and R. Weiner, *On the positivity of low order explicit Runge-Kutta schemes in splitting methods*, Comput. Math. Appl., 45 (2003), pp. 53–67.

[9] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[10] S. Gottlieb and C. W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[11] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Springer Ser. Comput. Math. 31, Springer-Verlag, Berlin, 2003.

[12] I. Higueras, *On strong stability preserving time discretization methods*, J. Sci. Comput., 21 (2004), pp. 193–223.

[13] I. Higueras, *Representations of Runge–Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal., 43 (2005), pp. 924–948.

[14] I. Higueras and T. Roldán, *On strong stability for additive Runge-Kutta methods: Entropy monotonicity for the Broadwell model*, submitted.

[15] W. Hundsdorfer and J. G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin, 2003.

[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, San Diego, CA, 1985.

[17] C. A. Kennedy and M. H. Carpenter, *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math., 44 (2003), pp. 139–181.

[18] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.

[19] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*, in Recent Trends in Numerical Analysis, Adv. Theory Comput. Math. 3, L. Brugnano and D. Trigiante, eds., Nova Sci. Publ., Huntington, NY, 2001, pp. 269–288.

[20] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., to appear.

[21] S. J. Ruuth and R. J. Spiteri, *Two barriers on strong stability preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211–220.

[22] J. W. Shen and X. Zhong, *Semi-Implicit Runge-Kutta Schemes for Non-Autonomous Differential Equations in Reactive Flow Computations*, Paper 1996-1969, 27th AIAA Fluid Dynamics Conference, New Orleans, LA, June 17–20, 1996.

[23] C. W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[24] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Comput., 9 (1988), pp. 1073–1084.
[25] C.-W. Shu, *A survey of strong stability preserving high order time discretizations*, in Collected Lectures on the Preservation of Stability under Discretization, D. Estep and S. Tavener, eds., Proc. Appl. Math. 109, SIAM, Philadelphia, 2002, pp. 51–65.
[26] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.
[27] X. Zhong, *Additive semi-implicit Runge-Kutta methods for computing high speed nonequilibrium reactive flows*, J. Comput. Phys., 128 (1996), pp. 19–31.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1759–1779

STABILITY ANALYSIS OF LARGE TIME-STEPPING METHODS FOR EPITAXIAL GROWTH MODELS*

CHUANJU XU† AND TAO TANG‡

Abstract. Numerical methods for solving the continuum model of the dynamics of the molecular beam epitaxy (MBE) require very large time simulation, and therefore large time steps become necessary. The main purpose of this work is to construct and analyze highly stable time discretizations which allow much larger time steps than those of a standard implicit-explicit approach. To this end, an extra term, which is consistent with the order of the time discretization, is added to stabilize the numerical schemes. Then the stability properties of the resulting schemes are established rigorously. Numerical experiments are carried out to support the theoretical claims. The proposed methods are also applied to simulate the MBE models with large solution times. The power laws for the coarsening process are obtained and are compared with previously published results.

Key words. molecular beam epitaxy, epitaxial growth, spectral method, stability, implicit-explicit method, large time-stepping

AMS subject classifications. 35Q99, 35R35, 65M12, 65M70, 74A50

DOI. 10.1137/050628143

1. Introduction.

Recently there has been significant research interest in the dynamics of molecular beam epitaxy (MBE) growth. The MBE technique is among the most refined methods for the growth of thin solid films, and it is of great importance for applied studies; see, e.g., [1, 16, 22]. The evolution of the surface morphology during epitaxial growth results in a delicate relation between the molecular flux and the relaxation of the surface profile through surface diffusion of adatoms. It occurs on time and length scales that may span several orders of magnitude. Different kinds of models have been used to describe such phenomena; these typically include *atomistic models, continuum models, and hybrid models.* The atomistic models are usually implemented in the form of molecular dynamics or kinetic Monte Carlo simulations [4, 9, 17]. The continuum models are based on partial differential equations and are appropriate mainly for investigating the temporal evolution of the MBE instability at large time and length scales [11, 24]. The hybrid models can be considered as a compromise between atomistic models and continuum models; see, e.g., [3, 8].

We are interested in the continuum models for the evolution of the MBE growth. Let $h(\mathbf{x}, t)$ be the epitaxy surface height with $\mathbf{x} \in \mathbb{R}^2$ and $t \geq 0$. Under typical conditions for MBE growth, the height evolution equation can be written under mass conservation form (see, e.g., [14]):

$$h_t = -\nabla \cdot J(\nabla h), \tag{1.1}$$

*Received by the editors March 30, 2005; accepted for publication (in revised form) February 14, 2006; published electronically September 26, 2006.
http://www.siam.org/journals/sinum/44-4/62814.html

†Department of Mathematics, Xiamen University, 361005 Xiamen, China (cjxu@xmu.edu.cn). The research of this author was partially supported by National NSF of China under grant 10531080, the Excellent Young Teachers Program by the Ministry of Education of China, and the Program of 985 Innovation Engineering on Information in Xiamen University.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, China (ttang@math.hkbu.edu.hk). The research of this author was supported in part by the Hong Kong Research Grants Council and the International Research Team on Complex System of Chinese Academy of Sciences.

where J is the surface current which can be decomposed into a sum of two currents,

$$J = J_{\mathrm{SD}} + J_{\mathrm{NE}}, \tag{1.2}$$

where J_{SD} is the equilibrium surface current describing the surface diffusion and J_{NE} is the nonequilibrium diffusion current taking into account the Ehrlich–Schwoebel effect [6, 18]. The surface diffusion current has the form

$$J_{\mathrm{SD}} = \delta\nabla(\Delta h), \tag{1.3}$$

where δ is the surface diffusion constant. By using effective free energy formulation, the nonequilibrium diffusion current under consideration can be written in the form

$$J_{\mathrm{NE}}(M) = -\frac{\partial U(M)}{\partial M}, \tag{1.4}$$

where $M = (M_1, M_2) := \nabla h$ is the slope vector and $U(M)$ is the potential depending only on the slope vector. Evidently, the term J_{NE} helps the system (1.1) to evolve toward the states in which the slope M attains the minimum of $U(M)$ because J_{NE} vanishes at the minima of $U(M)$. The minima of this potential is the preferred value of the slope. Consequently, the corresponding system is the so-called epitaxial growth model with slope selection.

The continuum model (1.1) has been extensively applied to modeling interfacial coarsening dynamics in epitaxial growth with slope selection; see, e.g., [12, 14, 24]. In (1.1), the fourth-order term models surface diffusion, and the nonlinear second-order term models the well-known Ehrlich–Schowoebel effect [6, 18], which gives rise to instabilities in the evolution of the surface morphology. The instability then leads to the formation of mounds and pyramids on the growing surface. These pyramid-like structures have been reported in many experiments and numerical simulations; see, e.g., [14, 20, 23]. It is found that the lateral width λ and the height w of these pyramids grow in time as power laws with the same component. Thus, the ratio w/λ, corresponding to the pyramid slope, approaches a constant at large times. Therefore, there is a slope selection in a typical MBE growth. The corresponding coarsening exponents were found from experiments to depend on the symmetry of the surface. Two typical values of the coarsening exponent have been found, namely 1/4 (see, e.g., [14, 20]) and 1/3 (see, e.g., [14, 23]). Some mathematical justification of such predictions was given in [10]. We also point out that the continuum model (1.1)–(1.4) has been derived by Ortiz, Repetto, and Si [15] by using a series expansion of the deposition flux in powers of the surface gradient. They also provided an explicit construction for the pyramid-like coarsening, which allows one to predict characteristic power laws for the pyramid size growth. However, it is difficult to provide growing details, especially for complex thin-film systems.

Numerical simulations with the continuum models are appropriate for investigating the surface growth instability at large times. The direct numerical simulation for (1.1)–(1.4) with different nonequilibrium diffusion currents was performed by Siegert [19], who obtained a power law close to 1/4. Moldovan and Golubovic [14] carried out very comprehensive numerical simulations by using a kinetic scaling theory and obtained a 1/3 power law. It should be pointed out that the simulations reported in [14] were not completely based on a continuum model. Instead, they solved a so-called type-A dynamics equation directly on a hexagonal grid. More recently, the well-posedness of the initial-boundary-value problem of (1.1) is studied by Li and Liu [12]

using the perturbation analysis and Galerkin spectral approximations. In [13], they used variational techniques to obtain some asymptotic results for a no-slope-selection model. Moreover, several scaling laws have been derived in [13].

The main purpose of this study is to provide efficient numerical schemes for solving (1.1), with particular emphasis on the use of large time steps. To obtain meaningful results for power laws, the integration times in simulations have to be very large (say, in the order of 10^4). As a result, it is reasonable to employ larger time steps and a small number of grid points in computations, provided that stability and accuracy can be preserved. It is observed that most of the existing continuum model simulations have used an explicit integration method in time and finite difference type approximation in space. To maintain the stability and to achieve high approximation accuracy, the number of spatial grid points must be large and the time step has to be small. Even with rapidly increasing computational resources, explicit schemes are still limited to simulating early surface evolution and therefore small length scale [12].

The main objectives of this work are threefold: First, we introduce an accurate and efficient semi-implicit Fourier pseudospectral method for solving the time-dependent nonlinear diffusion equations (1.1). To approximate the time derivatives, a backward differentiation is employed. More precisely, the fourth-order term is treated implicitly to reduce the associated stability constraints, while the nonlinear second-order terms are treated explicitly in order to avoid solving the nonlinear equations at each time step. Secondly, a stabilization second-order term is added to the discretized system, which increases the time step dramatically. In real applications, the surface diffusion constant δ may be very small after dimensional scaling. Consequently, direct use of the standard semi-implicit method still suffers from severe stability restriction on the time step. In order to overcome this difficulty, we introduce a stabilization term with constant coefficient A, which allows us to increase the time step significantly. Note that a similar technique has been used by Zhu et al. in the simulation of the Cahn–Hilliard equation [26]. Our main contribution is to show rigorously that the resulting numerical scheme is stable if an appropriate constant A is chosen. Justification of this stabilization technique is provided by considering several numerical tests. Finally, we perform some numerical simulations for the interfacial coarsening dynamics using our proposed schemes. Our numerical results yield a 1/3 power law for the isotropic symmetry surface and 1/4 for the square symmetry surface.

It is worthwhile to mention some recent papers by Feng and Prohl on the numerical analysis for Cahn–Hilliard and Allen–Cahn equations; see, e.g., [7]. They also studied stability issues—the continuous dependence of solutions on the initial data. This is different from our stability concept, which seems more related to the decay of energy. They mainly proved that the stability constant increases to infinity algebraically as a small parameter (similar to our surface diffusion constant δ in (1.3)) goes to 0. This is a big step forward, since usually the blow-up of the constant is exponential if one uses the Gronwall inequality—a standard method.

The organization of the paper is as follows. In section 2, we construct highly stable semi-implicit Fourier spectral methods for solving (1.1), which is of first-order accuracy in time. To improve the numerical stability, an $\mathcal{O}(\Delta t)$ term is added. Detailed stability analysis based on the energy method is provided to show that the proposed methods allow a large time step, and therefore are useful for large time simulations. The second-order semi-implicit methods are investigated in section 3. It will be demonstrated that the stability analysis for higher-order time-stepping methods is much more difficult. Numerical experiments for model problems are presented in

section 4. Section 5 reports some computational results for the coarsening dynamics using the numerical schemes allowing large time steps. Some concluding remarks are given in the final section.

2. Semi-implicit time discretization: First-order methods. To demonstrate the main ideas in scheme designing and stability analysis, we will use two model equations in this work. The first one is of the form

$$h_t = -\delta\Delta^2 h - \nabla\cdot[(1-|\nabla h|^2)\nabla h], \quad (\mathbf{x},t)\in\Omega\times(0,T]. \tag{2.1}$$

The second model equation is of the form

$$h_t = -\delta\Delta^2 h - ((1-|h_x|^2)h_x)_x - ((1-|h_y|^2)h_y)_y, \ \ (\mathbf{x},t)\in\Omega\times(0,T]. \tag{2.2}$$

Hereafter, we use h_t to denote $\frac{\partial h}{\partial t}$, $\nabla h = (h_x, h_y)$. Both model problems are subject to the periodic boundary conditions and suitable initial data, where $\Omega = (0,L)^2$ with $L > 0$. The model (2.1) corresponds to the isotropic surface current, while (2.2) represents the simplest square surface current.

For the MBE simulations, large computational domain is necessary in order to minimize the effect of periodicity assumption and to collect enough statistical information such as mean surface height and width of the pyramid-like structures. Moreover, sufficiently long integration time is necessary in order to detect the epitaxy growth behaviors and to reach the physical scaling regime. On the other hand, to carry out numerical simulations with large time and large computational domain, highly stable and accurate numerical methods are required. To this end, it is natural to use the Fourier spectral approach in space which has been found extremely efficient for periodic problems. As for stability issue, the implicit treatment for the fourth-order terms is employed, and more importantly, a special trick to handle the nonlinear second-order terms is used. The goal is to significantly increase the allowed time steps.

We first consider the MBE model with the isotropic symmetry current, namely, (2.1). A classical first-order semi-implicit scheme is of the form

$$\frac{h^{n+1}-h^n}{\Delta t} + \delta\Delta^2 h^{n+1} = -\nabla\cdot\left[(1-|\nabla h^n|^2)\nabla h^n\right], \ \ n\ge 0. \tag{2.3}$$

It is expected that the implicit treatment for the fourth-order term in (2.3) allows one to relax the time step restriction. However, numerical experiments demonstrate that a larger time step cannot be used for the scheme (2.3) when δ is small; see, e.g., [12]. To improve this, an $\mathcal{O}(\Delta t)$ term is added into the scheme (2.3):

$$\frac{h^{n+1}-h^n}{\Delta t} + \delta\Delta^2 h^{n+1} - A\Delta h^{n+1} = -\nabla\cdot\left[(1-|\nabla h^n|^2 + A)\nabla h^n\right], \ \ n\ge 0, \tag{2.4}$$

where A is a positive constant to be determined later and $h^n \equiv h^n(\mathbf{x})$ is an approximation of $h(\mathbf{x},t)$ at $t = t^n$. The initial data h^0 is given by the initial condition. The purpose for adding the extra terms is to improve the stability condition so that larger time steps can be used. This will be justified theoretically in this section, and will be demonstrated by our numerical results in section 4.

In order to study its stability property, we will use a discrete energy estimate. To this end, we first state the following known result.

LEMMA 2.1 (energy identities [12]). *If $h(\mathbf{x},t)$ is a solution of (2.1), then the following energy identities hold:*

$$\frac{d}{dt}\|h\|^2 + 4E(h) + \|\nabla h\|_{L^4}^4 = |\Omega|, \tag{2.5}$$

$$\frac{d}{dt}E(h) + \|h_t\|^2 = 0, \tag{2.6}$$

where $\|\cdot\|$ is the standard L^2-norm in Ω, L^p is the standard L^p-norm, and

$$E(h) = \int_\Omega \left[\frac{1}{4}(|\nabla h|^2 - 1)^2 + \frac{\delta}{2}|\Delta h|^2\right] d\mathbf{x}. \tag{2.7}$$

We briefly sketch the proof of (2.5) and (2.6), which is useful in deriving its discrete counterparts. It follows from (2.1) that

$$\langle h_t, \varphi\rangle = -\left\langle \nabla\cdot[(1-|\nabla h|^2)\nabla h + \delta\nabla\Delta h], \varphi\right\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in the L^2-space. It can be verified directly that setting $\varphi = h$ gives (2.5) and setting $\varphi = h_t$ yields (2.6).

THEOREM 2.1. *If the constant A in (2.4) is sufficiently large, then the following energy inequality holds:*

$$E(h^{n+1}) \le E(h^n), \tag{2.8}$$

where E is defined by (2.7) and h^n is computed by (2.4). Moreover, if the numerical solution is convergent in $L^\infty([0,T];W^{1,\infty}(\Omega))$ as $\Delta t\to 0$, then the constant A can be chosen to satisfy

$$A \ge \frac{3}{2}|\nabla h|^2 - \frac{1}{2} \qquad \text{a.e. in } \Omega\times(0,T], \tag{2.9}$$

where $h(\mathbf{x},t)$ is a solution of (2.1).

Proof. For any L-periodic $H^2(\Omega)$ function φ, it follows from (2.4) that

$$\frac{1}{\Delta t}\left\langle h^{n+1}-h^n, \varphi\right\rangle + \delta\left\langle \Delta h^{n+1}, \Delta\varphi\right\rangle + A\left\langle \nabla(h^{n+1}-h^n), \nabla\varphi\right\rangle + I(\varphi) = 0, \tag{2.10}$$

where

$$I(\varphi) := \left\langle (|\nabla h^n|^2 - 1)\nabla h^n, \nabla\varphi\right\rangle.$$

Letting $\varphi = \delta_t h^n := h^{n+1} - h^n$ gives

$$\frac{1}{\Delta t}\|\delta_t h^n\|^2 + \delta\left\langle \Delta h^{n+1}, \Delta\delta_t h^n\right\rangle + A\left\langle \nabla\delta_t h^n, \nabla\delta_t h^n\right\rangle + I(\delta_t h^n) = 0. \tag{2.11}$$

Observe that

$$\begin{aligned}
I(\delta_t h^n) &= \left\langle |\nabla h^n|^2 - 1, \nabla h^n\cdot\nabla h^{n+1} - |\nabla h^n|^2\right\rangle \\
&= \left\langle |\nabla h^n|^2 - 1, -\frac{1}{2}|\nabla\delta_t h^n|^2 + \frac{1}{2}|\nabla h^{n+1}|^2 - \frac{1}{2}|\nabla h^n|^2\right\rangle \\
&= -\frac{1}{2}\left\langle |\nabla h^n|^2 - 1, |\nabla\delta_t h^n|^2\right\rangle + \frac{1}{2}\left\langle (|\nabla h^n|^2-1)(|\nabla h^{n+1}|^2 - |\nabla h^n|^2), 1\right\rangle \\
&= -\frac{1}{2}\left\langle |\nabla h^n|^2 - 1, |\nabla\delta_t h^n|^2\right\rangle + \frac{1}{2}\left\langle |\nabla h^n|^2\cdot|\nabla h^{n+1}|^2, 1\right\rangle \\
&\quad + \frac{1}{2}\left\langle -|\nabla h^n|^4 - |\nabla h^{n+1}|^2 + |\nabla h^n|^2, 1\right\rangle.
\end{aligned}$$

Using the identity $2a^2b^2 = -(a^2-b^2)^2 + a^4 + b^4$ to the second last term above with $a = |\nabla h^{n+1}|$ and $b = |\nabla h^n|$, we obtain

$$\begin{aligned} I(\delta_t h^n) &= -\frac{1}{2}\left\langle |\nabla h^n|^2 - 1, |\nabla \delta_t h^n|^2 \right\rangle - \frac{1}{4}\left\langle (|\nabla h^{n+1}|^2 - |\nabla h^n|^2)^2, 1 \right\rangle \\ &\quad + \frac{1}{4}\left\langle |\nabla h^{n+1}|^4 + |\nabla h^n|^4, 1 \right\rangle + \frac{1}{2}\left\langle -|\nabla h^n|^4 - |\nabla h^{n+1}|^2 + |\nabla h^n|^2, 1 \right\rangle \\ &= \left\langle -\frac{1}{2}(|\nabla h^n|^2 - 1) - \frac{1}{4}|\nabla h^{n+1} + \nabla h^n|^2, |\nabla \delta_t h^n|^2 \right\rangle \\ &\quad + \frac{1}{4}\left\langle |\nabla h^{n+1}|^4 - |\nabla h^n|^4 - 2|\nabla h^{n+1}|^2 + 2|\nabla h^n|^2, 1 \right\rangle \\ &= \left\langle -\frac{1}{2}(|\nabla h^n|^2 - 1) - \frac{1}{4}|\nabla h^{n+1} + \nabla h^n|^2, |\nabla \delta_t h^n|^2 \right\rangle \\ &\quad + \frac{1}{4}\Big(\| |\nabla h^{n+1}|^2 - 1\|^2 - \| |\nabla h^n|^2 - 1\|^2 \Big). \end{aligned} \tag{2.12}$$

Combining (2.11) and (2.12) yields

$$\begin{aligned} &\frac{1}{\Delta t}\|\delta_t h^n\|^2 + \delta\left\langle \Delta h^{n+1}, \Delta \delta_t h^n \right\rangle + \frac{1}{4}\left(\| |\nabla h^{n+1}|^2 - 1\|^2 - \| |\nabla h^n|^2 - 1\|^2 \right) \\ &\quad + \left\langle A - \frac{1}{2}(|\nabla h^n|^2 - 1) - \frac{1}{4}|\nabla h^{n+1} + \nabla h^n|^2, |\delta_t h^n|^2 \right\rangle = 0. \end{aligned} \tag{2.13}$$

Note that the last term in (2.13) can be made nonnegative provided that

$$A \geq \max_{\mathbf{x}\in\Omega}\left\{ \frac{1}{2}(|\nabla h^n|^2 - 1) + \frac{1}{4}|\nabla h^{n+1} + \nabla h^n|^2 \right\}. \tag{2.14}$$

Observe that

$$\begin{aligned} &\delta\left\langle \Delta h^{n+1}, \Delta \delta_t h^n \right\rangle = \delta\left\langle \Delta h^{n+1}, \Delta h^{n+1} - \Delta h^n \right\rangle \\ &\geq \frac{\delta}{2}\|\Delta h^{n+1}\|^2 - \frac{\delta}{2}\|\Delta h^n\|^2. \end{aligned} \tag{2.15}$$

Consequently, Theorem 2.1 follows from (2.13)–(2.15). ∎

We now consider the MBE model with the square symmetric surface (2.2). An energy equality similar to that for the model (2.1) can be established.

LEMMA 2.2. *If $h(\mathbf{x}, t)$ is a solution of* (2.2), *then the following energy equalities hold:*

$$\frac{d}{dt}E_2(h) + \|h_t\|^2 = 0, \tag{2.16}$$

$$\frac{d}{dt}\|h\|^2 + 4E_2(h) + \|h_x\|_{L^4}^4 + \|h_y\|_{L^4}^4 = 2|\Omega|, \tag{2.17}$$

where

$$E_2(h) = \int_\Omega \left\{ \frac{\delta}{2}|\Delta h|^2 + \frac{1}{4}\left[(h_x^2 - 1)^2 + (h_y^2 - 1)^2 \right] \right\} d\mathbf{x}\,. \tag{2.18}$$

Proof. Equation (2.2) is equivalent to

$$h_t + \delta\Delta^2 h = -\nabla \cdot J, \tag{2.19}$$

where $J = (J_1, J_2)$ is given by

$$J_1 = \left(1 - h_x^2\right) h_x\,, \qquad J_2 = (1 - h_y^2) h_y\,.$$

Multiplying both sides of (2.19) with h_t gives

$$\|h_t\|^2 + \delta \langle \Delta h, (\Delta h)_t \rangle = \langle J, (\nabla h)_t \rangle\,. \tag{2.20}$$

Observe

$$\begin{aligned}\langle J, (\nabla h)_t \rangle &= \left\langle (1 - h_x^2) h_x, h_{xt} \right\rangle + \left\langle (1 - h_y^2) h_y, h_{yt} \right\rangle \\ &= -\frac{1}{4} \frac{d}{dt} \int_\Omega \left[(h_x^2 - 1)^2 + (h_y^2 - 1)^2 \right] d\mathbf{x}.\end{aligned}$$

The above result and (2.20) yield (2.16). Similarly, the energy equality (2.17) can be derived by multiplying (2.19) with h. ∎

Similar to the scheme (2.4), a first-order scheme is constructed for the MBE model (2.2):

$$\begin{aligned}&\frac{h^{n+1} - h^n}{\Delta t} + \delta \Delta^2 h^{n+1} - A \Delta h^{n+1} \\ &= -A \Delta h^n - \left[(1 - (h_x^n)^2) h_x^n \right]_x - \left[(1 - (h_y^n)^2) h_y^n \right]_y\,.\end{aligned} \tag{2.21}$$

THEOREM 2.2. *If A in* (2.21) *is chosen sufficiently large, then the following energy inequality holds:*

$$E_2(h^{n+1}) \le E_2(h^n), \tag{2.22}$$

where E_2 is defined by (2.18) *and h^n is computed by* (2.21). *Moreover, if the numerical solution of* (2.21) *is convergent, then A can be chosen to satisfy*

$$A \ge \max \left\{ \frac{3}{2} h_x^2 - \frac{1}{2},\ \frac{3}{2} h_y^2 - \frac{1}{2} \right\} \qquad \text{a.e. in } \Omega \times (0, T], \tag{2.23}$$

where $h(\mathbf{x}, t)$ is the solution of (2.2).

Proof. The proof follows in the same manner as that of Theorem 2.1. By direct computations, we can obtain

$$\begin{aligned}&\frac{1}{\Delta t} \|\delta_t h^n\|^2 + E_2(h^{n+1}) - E_2(h^n) \\ &\quad + \int \left[A - \frac{1}{2} \left((h_x^n)^2 - 1 \right) - \frac{1}{4} (h_x^{n+1} + h_x^n)^2 \right] (h_x^{n+1} - h_x^n)^2 \, d\mathbf{x} \\ &\quad + \int \left[A - \frac{1}{2} \left((h_y^n)^2 - 1 \right) - \frac{1}{4} (h_y^{n+1} + h_y^n)^2 \right] (h_y^{n+1} - h_y^n)^2 \, d\mathbf{x} = 0.\end{aligned}$$

It follows from the above result that (2.22) holds provided that

$$A \ge \max_{\mathbf{x} \in \Omega} \left\{ \frac{1}{2} \left((h_x^n)^2 - 1 \right) - \frac{1}{4} \left(h_x^{n+1} + h_x^n \right)^2 \right\}$$

and

$$A \ge \max_{\mathbf{x} \in \Omega} \left\{ \frac{1}{2} \left((h_y^n)^2 - 1 \right) - \frac{1}{4} \left(h_y^{n+1} + h_y^n \right)^2 \right\}.$$

If the numerical solution is convergent, then the above conditions become inequality (2.23). ∎

3. Semi-implicit time discretization: Higher-order methods.

3.1. Second-order scheme: BD2/EP2. By combining a second-order backward differentiation (BD2) for the time derivative term and a second-order extrapolation (EP2) for the explicit treatment of the nonlinear term, we arrive at a second-order scheme (BD2/EP2) for (2.1):

(3.1)
$$\begin{aligned}&\frac{3h^{n+1}-4h^n+h^{n-1}}{2\Delta t}+\delta\Delta^2 h^{n+1}-A\Delta h^{n+1}\\ &=-2A\Delta h^n+A\Delta h^{n-1}-\nabla\cdot[(1-|\nabla(2h^n-h^{n-1})|^2)\nabla(2h^n-h^{n-1})]\quad\forall n\geq 1.\end{aligned}$$

As usual, to start the iteration $h^0(\mathbf{x})$ is given by the initial condition, and $h^1(\mathbf{x})$ is computed by the first-order scheme (2.4).

THEOREM 3.1. *If the constant A in* (3.1) *is sufficiently large, then the following energy inequality holds:*

$$\widetilde{E}^{n+1}\leq\widetilde{E}^n+\mathcal{O}(\Delta t^2), \tag{3.2}$$

where $\widetilde{E}^n$ is defined by

$$\widetilde{E}^n=\frac{1}{\Delta t}\|h^n-h^{n-1}\|^2+\frac{1}{4}\left\||\nabla h^n|^2-1\right\|^2+\frac{\delta}{2}\|\Delta h^n\|^2+\frac{A}{2}\|\nabla(h^n-h^{n-1})\|^2. \tag{3.3}$$

In particular, we can obtain

$$E(h^n)\leq E(h^1)+\mathcal{O}(\Delta t), \tag{3.4}$$

where E is defined by (2.7). *Moreover, if the numerical solution of* (3.1) *is convergent in $L^\infty([0,T];W^{1,\infty}(\Omega))$ as $\Delta t\to 0$, then the constant A can be chosen to satisfy*

$$A\geq 3|\nabla h|^2-1\quad a.e.\ in\ \Omega\times(0,T], \tag{3.5}$$

where $h(x,t)$ is a solution of (2.1).

Proof. For ease of notation, let $\delta_t h^n=h^{n+1}-h^n$ and $\delta_{tt}h^n=h^{n+1}-2h^n+h^{n-1}$. Multiplying both sides of (3.1) by $\delta_t h^n$ and integrating the resulting equation in Ω give

$$I_1^n+I_2^n+I_3^n=I_4^n, \tag{3.6}$$

where

$$\begin{aligned}I_1^n&:=\frac{1}{2\Delta t}\langle 3\delta_t h^n-\delta_t h^{n-1}\,,\,\delta_t h^n\rangle,\\ I_2^n&:=\delta\langle\Delta^2 h^{n+1}\,,\,\delta_t h^n\rangle,\\ I_3^n&:=-A\left\langle\Delta\delta_{tt}h^n\,,\,\delta_t h^n\right\rangle,\\ I_4^n&:=-\left\langle\nabla\cdot[(1-|\nabla(2h^n-h^{n-1})|^2)\nabla(2h^n-h^{n-1})]\,,\,\delta_t h^n\right\rangle.\end{aligned}$$

We now estimate them term by term. The estimate for the first three terms is straightforward:

$$I_1^n\geq\frac{5}{4\Delta t}\|\delta_t h^n\|^2-\frac{1}{4\Delta t}\|\delta_t h^{n-1}\|^2\geq\frac{1}{\Delta t}\|\delta_t h^n\|^2-\frac{1}{\Delta t}\|\delta_t h^{n-1}\|^2, \tag{3.7}$$

$$I_2^n=\delta\left\langle\Delta h^{n+1}\,,\,\Delta h^{n+1}-\Delta h^n\right\rangle\geq\frac{\delta}{2}\|\Delta h^{n+1}\|^2-\frac{\delta}{2}\|\Delta h^n\|^2, \tag{3.8}$$

$$\begin{aligned}I_3^n&=A\left\langle\delta_{tt}\nabla h^n\,,\,\delta_t\nabla h^n\right\rangle=A\left\langle\delta_t\nabla h^n-\delta_t\nabla h^{n-1}\,,\,\delta_t\nabla h^n\right\rangle\\ &=\frac{A}{2}\|\delta_t\nabla h^n\|^2-\frac{A}{2}\|\delta_t\nabla h^{n-1}\|^2+\frac{A}{2}\|\delta_{tt}\nabla h^n\|^2.\end{aligned} \tag{3.9}$$

To estimate I_4^n, we need the following two identities. On one hand, we have

(3.10)
$$\begin{aligned}
&\nabla(2h^n - h^{n-1}) \cdot \nabla(h^{n+1} - h^n) \\
&= \nabla(2h^n - h^{n-1}) \cdot \nabla h^{n+1} - \nabla(2h^n - h^{n-1}) \cdot \nabla h^n \\
&= -\frac{1}{2}|\delta_{tt}\nabla h^n|^2 + \frac{1}{2}|\nabla(2h^n - h^{n-1})|^2 + \frac{1}{2}|\nabla h^{n+1}|^2 - \nabla(2h^n - h^{n-1}) \cdot \nabla h^n,
\end{aligned}$$

and on the other hand,

(3.11)
$$\begin{aligned}
&\nabla(2h^n - h^{n-1}) \cdot \nabla(h^{n+1} - h^n) = \nabla h^n \cdot \nabla(h^{n+1} - h^n) + \delta_t \nabla h^n \cdot \delta_t \nabla h^{n-1} \\
&= \frac{1}{2}|\nabla h^{n+1}|^2 - \frac{1}{2}|\nabla h^n|^2 - \frac{1}{2}|\delta_t \nabla h^n|^2 - \frac{1}{2}|\delta_{tt}\nabla h^n|^2 + \frac{1}{2}|\delta_t \nabla h^n|^2 + \frac{1}{2}|\delta_t \nabla h^{n-1}|^2 \\
&= \frac{1}{2}|\nabla h^{n+1}|^2 - \frac{1}{2}|\nabla h^n|^2 - \frac{1}{2}|\delta_{tt}\nabla h^n|^2 + \frac{1}{2}|\delta_t \nabla h^{n-1}|^2.
\end{aligned}$$

Using (3.10) gives

(3.12)
$$\begin{aligned}
&\left\langle -|\nabla(2h^n - h^{n-1})|^2 \nabla(2h^n - h^{n-1}),\, \nabla(h^{n+1} - h^n) \right\rangle \\
&= \frac{1}{2}\left\langle |\nabla(2h^n - h^{n-1})|^2,\, |\delta_{tt}\nabla h^n|^2 \right\rangle + J_4^n,
\end{aligned}$$

where

$$\begin{aligned}
J_4^n &:= \left\langle -|\nabla(2h^n - h^{n-1})|^2,\, \frac{1}{2}|\nabla(2h^n - h^{n-1})|^2 + \frac{1}{2}|\nabla h^{n+1}|^2 - \nabla(2h^n - h^{n-1}) \cdot \nabla h^n \right\rangle \\
&= \frac{1}{2}\left\langle 1,\, -|\nabla(2h^n - h^{n-1})|^4 \right\rangle + \frac{1}{2}\left\langle -|\nabla(2h^n - h^{n-1})|^2,\, |\nabla h^{n+1}|^2 \right\rangle \\
&\quad + \left\langle |\nabla(2h^n - h^{n-1})|^2,\, \nabla(2h^n - h^{n-1}) \cdot \nabla h^n \right\rangle \\
&= -\frac{3}{4}\left\langle 1,\, |\nabla(2h^n - h^{n-1})|^4 \right\rangle + \frac{1}{4}\| |\nabla(2h^n - h^{n-1})|^2 - |\nabla h^{n+1}|^2 \|^2 \\
&\quad - \frac{1}{4}\left\langle 1,\, |\nabla h^{n+1}|^4 \right\rangle + \left\langle |\nabla(2h^n - h^{n-1})|^2,\, \nabla(2h^n - h^{n-1}) \cdot \nabla h^n \right\rangle.
\end{aligned}$$

Using the Schwartz inequality to the last term above gives

$$\begin{aligned}
&\left\langle |\nabla(2h^n - h^{n-1})|^2,\, \nabla(2h^n - h^{n-1}) \cdot \nabla h^n \right\rangle \\
&\le \frac{1}{2}\left\langle |\nabla(2h^n - h^{n-1})|^2,\, |\nabla(2h^n - h^{n-1})|^2 \right\rangle + \frac{1}{2}\left\langle |\nabla(2h^n - h^{n-1})|^2,\, |\nabla h^n|^2 \right\rangle \\
&\le \frac{1}{2}\left\langle 1,\, |\nabla(2h^n - h^{n-1})|^4 \right\rangle + \frac{1}{4}\left\langle 1,\, |\nabla(2h^n - h^{n-1})|^4 \right\rangle + \frac{1}{4}\left\langle 1,\, |\nabla h^n|^4 \right\rangle \\
&= \frac{3}{4}\left\langle 1,\, |\nabla(2h^n - h^{n-1})|^4 \right\rangle + \frac{1}{4}\left\langle 1,\, |\nabla h^n|^4 \right\rangle.
\end{aligned}$$

Combining the above two results gives

(3.13)
$$\begin{aligned}
J_4^n &\le \frac{1}{4}\| |\nabla(2h^n - h^{n-1})|^2 - |\nabla h^{n+1}|^2 \|^2 - \frac{1}{4}\left\langle 1,\, |\nabla h^{n+1}|^4 \right\rangle + \frac{1}{4}\left\langle 1,\, |\nabla h^n|^4 \right\rangle \\
&= \frac{1}{4}\left\langle |\nabla(h^{n+1} + 2h^n - h^{n-1})|^2,\, |\delta_{tt}\nabla h^n|^2 \right\rangle - \frac{1}{4}\left\langle 1,\, |\nabla h^{n+1}|^4 \right\rangle + \frac{1}{4}\left\langle 1,\, |\nabla h^n|^4 \right\rangle.
\end{aligned}$$

Using the definition of I_4^n, together with (3.11), (3.12), and (3.13), we have

$$\begin{aligned}
I_4^n &= \left\langle (1-|\nabla(2h^n-h^{n-1})|^2)\nabla(2h^n-h^{n-1})\,,\,\nabla(h^{n+1}-h^n)\right\rangle \\
&= \frac{1}{2}\|\nabla h^{n+1}\|^2-\frac{1}{2}\|\nabla h^n\|^2-\frac{1}{2}\|\delta_{tt}\nabla h^n\|^2+\frac{1}{2}\|\delta_t\nabla h^{n-1}\|^2 \\
&\quad +\frac{1}{2}\left\langle |\nabla(2h^n-h^{n-1})|^2\,,\,|\delta_{tt}\nabla h^n|^2\right\rangle + J_4^n \\
&\le -\frac{1}{2}\left\langle 1-|\nabla(2h^n-h^{n-1})|^2\,,\,|\delta_{tt}\nabla h^n|^2\right\rangle+\frac{1}{2}\|\nabla h^{n+1}\|^2-\frac{1}{2}\|\nabla h^n\|^2+\frac{1}{2}\|\delta_t\nabla h^{n-1}\|^2 \\
&\quad +\frac{1}{4}\left\langle |\nabla(h^{n+1}+2h^n-h^{n-1})|^2\,,\,|\delta_{tt}\nabla h^n|^2\right\rangle-\frac{1}{4}\left\langle 1,|\nabla h^{n+1}|^4\right\rangle+\frac{1}{4}\left\langle 1,|\nabla h^n|^4\right\rangle \\
&= \left\langle \frac{1}{2}|\nabla(2h^n-h^{n-1})|^2-\frac{1}{2}+\frac{1}{4}|\nabla(h^{n+1}+2h^n-h^{n-1})|^2\,,\,|\delta_{tt}\nabla h^n|^2\right\rangle \\
&\quad -\frac{1}{4}\||\nabla h^{n+1}|^2-1\|^2+\frac{1}{4}\||\nabla h^n|^2-1\|^2+\frac{1}{2}\|\delta_t\nabla h^{n-1}\|^2.
\end{aligned}$$

The above result, together with (3.6) and (3.7)–(3.9), yields

$$\begin{aligned}
\widetilde{E}^{n+1} &\le \widetilde{E}^n+\frac{1}{2}\|\delta_t\nabla h^{n-1}\|^2 \\
&\quad +\left\langle -\frac{A}{2}+\frac{1}{2}|\nabla(2h^n-h^{n-1})|^2-\frac{1}{2}+\frac{1}{4}|\nabla(h^{n+1}+2h^n-h^{n-1})|^2\,,\,|\delta_{tt}\nabla h^n|^2\right\rangle.
\end{aligned}$$

The last term above can be made nonpositive provided that

$$A\ge |\nabla(2h^n-h^{n-1})|^2-1+\frac{1}{2}|\nabla(h^{n+1}+2h^n-h^{n-1})|^2 \qquad \text{a.e. in } \Omega.$$

Using the fact that

$$\|\delta_t\nabla h^{n-1}\|^2=\Delta t^2\|\nabla(h^n-h^{n-1})/\Delta t\|^2=\mathcal{O}(\Delta t^2),$$

we obtain (3.2). Summing (3.2) over n gives $\widetilde{E}^n\le\widetilde{E}^1+\mathcal{O}(\Delta t)$. In particular, by using the definition of $\widetilde{E}$ and the energy E defined by (2.7), we have

$$E(h^n)\le E(h^1)+\mathcal{O}(1)\Delta t,$$

where the $\mathcal{O}(1)$ term is given by

$$\begin{aligned}
\mathcal{O}(1) &= \|(h^1-h^0)/\Delta t\|^2+\frac{A}{2}\Delta t\|\nabla(h^1-h^0)/\Delta t\|^2 \\
&\quad +\sum_{i=1}^{n-1}\Delta t\|\nabla(h^i-h^{i-1})/\Delta t\|^2.
\end{aligned}$$

This completes the proof of this theorem. ∎

Remark 3.1. By comparing (2.9) and (3.5), we notice that the constant A used for the second-order scheme is two times larger than that for the first-order scheme.

Similarly, a second-order scheme of the BD2/EP2-type can be constructed for the square symmetry current model (2.2):

$$\begin{aligned}
&\frac{3h^{n+1}-4h^n+h^{n-1}}{2\Delta t}+\delta\Delta^2h^{n+1}-A\Delta h^{n+1} \\
&\quad =-2A\Delta h^n+A\Delta h^{n-1}-\left[\left(1-(2h_x^n-h_x^{n-1})^2\right)(2h_x^n-h_x^{n-1})\right]_x \\
&\qquad -\left[\left(1-(2h_y^n-h_y^{n-1})^2\right)(2h_y^n-h_y^{n-1})\right]_y.
\end{aligned} \tag{3.14}$$

Then an analysis similar to that of Theorem 3.1 can be carried out to obtain a stability result. The details will be omitted here.

3.2. Third-order scheme: BD3/EP3. A third-order scheme for solving the MBE model of general form (1.1) can be constructed in a similar manner as used in the last subsection. Specifically, we can obtain the BD3/EP3 scheme in the following form:

$$\begin{aligned}(3.15)\quad &\frac{11h^{n+1}-18h^n+9h^{n-1}-2h^{n-2}}{6\Delta t}+\delta\Delta^2 h^{n+1}-A\Delta h^{n+1}\\ &=-A\Delta(3h^n-3h^{n-1}+h^{n-2})-\nabla\cdot J(\nabla(3h^n-3h^{n-1}+h^{n-2}))\quad \forall n\geq 2,\end{aligned}$$

where, in order to start the iteration, h^1, h^2 are calculated via a first- and second-order scheme, respectively.

The stability analysis of the scheme (3.15) requires some very detailed energy estimates and will not be presented here. The numerical results obtained in the next two sections indicate that the third-order time discretization of type (3.15) is also stable as long as the constant A is sufficiently large.

4. Numerical experiments: Stability and accuracy tests. A complete numerical algorithm also requires a discretization strategy in space. Since the Fourier spectral method is one of the most suitable spatial approximation methods for periodic problems [2, 5, 21, 25], it will be employed to handle the spatial discretization. To demonstrate the principal ideas, we consider the full discretization for the MBE model with the isotropic current using the first-order time-stepping method, namely, we will consider only the full discretization for (2.4). It is to find an approximate solution $h_K^n(\mathbf{x})$ in form of a truncated Fourier expansion:

$$h_K^n(\mathbf{x})=\sum_{k_1,k_2=-K}^{K}\hat{h}_{\mathbf{k}}^n\exp(-\mathrm{i}\mathbf{k}\mathbf{x}),$$

where $\mathbf{k}=(k_1,k_2)$ and K is a positive integer. The above expansion is required to satisfy the following weak formulation:

$$\begin{aligned}(4.1)\quad &\frac{1}{\Delta t}\left\langle h_K^{n+1}-h_K^n,\varphi\right\rangle+\delta\left\langle\Delta h_K^{n+1},\Delta\varphi\right\rangle+A\left\langle\nabla h_K^{n+1},\nabla\varphi\right\rangle\\ &=\left\langle(1-|\nabla h_K^n|^2+A)\nabla h_K^n,\nabla\varphi\right\rangle\ \ \forall\varphi\in S_K,\end{aligned}$$

where

$$S_K=\mathrm{span}\{\exp(-\mathrm{i}\mathbf{k}\mathbf{x}),\quad -K\leq k_1,k_2\leq K\}.$$

For the full discretization problem (4.1), an energy inequality similar to that of Theorem 2.1 can be derived (its proof will be omitted here).

THEOREM 4.1. *Consider the numerical scheme* (4.1). *If*

$$(4.2)\qquad A\geq\max_{\mathbf{x}\in\Omega}\left\{\frac{1}{2}(|\nabla h_K^n|^2-1)+\frac{1}{4}|\nabla h_K^{n+1}+\nabla h_K^n|^2\right\},$$

then the solution of (4.1) *satisfies*

$$(4.3)\qquad E(h_K^{n+1})\leq E(h_K^n)\quad\forall n\geq 0,$$

where the energy E is defined by (2.7). *Moreover, if the numerical solution of* (4.1) *is convergent in* $L^\infty([0,T];W^{1,\infty}(\Omega))$ *as* $K\to\infty$ *and* $\Delta t\to 0$, *then the constant* A *can be chosen to satisfy*

$$A \ge \frac{3}{2}|\nabla h|^2 - \frac{1}{2} \qquad a.e.\ in\ \Omega\times(0,T], \tag{4.4}$$

where $h(\mathbf{x},t)$ *is a solution of* (2.1).

By applying the Fourier transformation to (2.4), we obtain a set of ordinary differential equations for each mode $\mathbf{k}$ in the Fourier space,

$$\frac{\hat h_{\mathbf{k}}^{n+1}-\hat h_{\mathbf{k}}^{n}}{\Delta t} + \delta|\mathbf{k}|^4\hat h_{\mathbf{k}}^{n+1} + A|\mathbf{k}|^2\hat h_{\mathbf{k}}^{n+1} = -\mathrm{i}\mathbf{k}\left\{(1-|\nabla h_K^n|^2+A)\nabla h_K^n\right\}_{\mathbf{k}}, \tag{4.5}$$

where $|\mathbf{k}| = \sqrt{k_1^2+k_2^2}$ is the magnitude of $\mathbf{k}$ and $\{f\}_{\mathbf{k}}$ represents the $\mathbf{k}$th-mode Fourier coefficient of the function f. The Fourier coefficients of the nonlinear term $(1-|\nabla h_K^n|^2+A)\nabla h_K^n$ are calculated by performing the discrete fast Fourier transform (FFT). It is readily seen that for a given level n the evaluation of all $\{(1-|\nabla h_K^n|^2+A)\nabla h_K^n\}_{\mathbf{k}}$ requires $8N$ one-dimensional FFT with vector length $N=2K$. This is also the total cost to compute $\hat h_{\mathbf{k}}^{n+1}$ from (4.5). In practical calculation, we work on the spectral space. At the final time level, an additional FFT is needed to recover the physical nodal values $h_K^{n+1}(\mathbf{x})$ from $\hat h_{\mathbf{k}}^{n+1}, -K\le k_1,k_2\le K$.

The purpose of this section is to verify the stability of the proposed numerical schemes in terms of the choice of the constant A. More serious applications will be reported in the next section.

Example 4.1. Consider an isotropic symmetry current model (2.1):

$$\begin{cases} h_t = -\delta\Delta^2 h - \nabla\cdot[(1-|\nabla h|^2)\nabla h], & [0,2\pi]^2\times(0,T],\\ h(\cdot,t)\ \text{is}\ 2\pi-\text{periodic} & \forall t\in(0,T],\\ h(\mathbf{x},0)=h_0(\mathbf{x}) & \forall \mathbf{x}\in[0,2\pi]^2 \end{cases} \tag{4.6}$$

with $\delta = 0.1, 0.01, 0.001$ and

$$h_0(\mathbf{x}) = 0.1(\sin 3x\sin 2y + \sin 5x\sin 5y). \tag{4.7}$$

This problem was used by Li and Liu [12] to study the most unstable modes. It was proved that with the initial condition (4.7) the most unstable modes are those with wave-vectors $\mathbf{k}$ such that $|\mathbf{k}|=\sqrt{5}$. Numerically, they showed that after short interaction of the unstable modes, the solution converges to a *steady state* which consists mainly of one mode only.

Define Δt_c as the largest possible time which allows stable numerical computation. In other words, if the time step is greater than Δt_c, then the numerical solution will blow up. In Table 1, we list the values of Δt_c for the schemes (2.4), (3.1), and (3.15) with different choices of A. All these semidiscrete schemes are approximated by the Fourier spectral methods in space. The Fourier mode number used in the calculations is $K=128$. Several observations are made from Table 1:

- If $A=0$, i.e., if a conventional implicit-explicit approach is used, then the numerical methods suffer from extremely small time steps, in particular when higher-order schemes are used or $\delta\ll 1$.
- The improvement on stability with the use of the constant A is significant. When A is sufficiently large (in this case $A\ge 2$), quite large time steps (in this case $\Delta t\ge 1$) can be used for first- and second-order time discretizations.

TABLE 1
Example 4.1: *stability comparison with different* A *and* δ. *Here* BDr *stands for* r*th-order backward differentiation and* EPr *for* r*th-order extrapolation.*

δ	A	BD1/EP1	BD2/EP2	BD3/EP3
0.1	$A=0$	$\Delta t_c \approx 1$	$\Delta t_c < 0.3$	$\Delta t_c < 0.1$
	$A=1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 0.5$
	$A=2$	$\Delta t_c \approx 1$	$\Delta t_c \approx 1$	$0.2 \le \Delta t_c < 0.5$
0.01	$A=0$	$\Delta t_c < 0.1$	$\Delta t_c < 0.01$	$\Delta t_c < 0.002$
	$A=1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 0.1$	$\Delta t_c \approx 0.002$
	$A=2$	$\Delta t_c \approx 1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 0.05$
0.001	$A=0$	$\Delta t_c < 0.01$	$\Delta t_c < 0.001$	$\Delta t_c < 10^{-4}$
	$A=1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 0.005$	$0.0005 \le \Delta t_c < 10^{-3}$
	$A=2$	$\Delta t_c \approx 1$	$\Delta t_c \approx 1$	$\Delta t_c \approx 0.005$

- The choice of A depends on the order of time discretization. For the third-order methods, a quite small time step has to be used, which is impractical for large time simulations.

We now turn to time accuracy comparison. Since the exact solution for problem (4.6) is unknown, we use numerical results of BD3/EP3 with $\Delta t = 0.0001$ and $K = 128$ as the "exact" solution. The coefficient δ is set to be 0.01 and the numerical errors are computed at $t = 1$. In this case, the "exact" solution obtained by using BD3/EP3 is plotted in Figure 1. Table 2 shows the L^2-errors using several values of A and

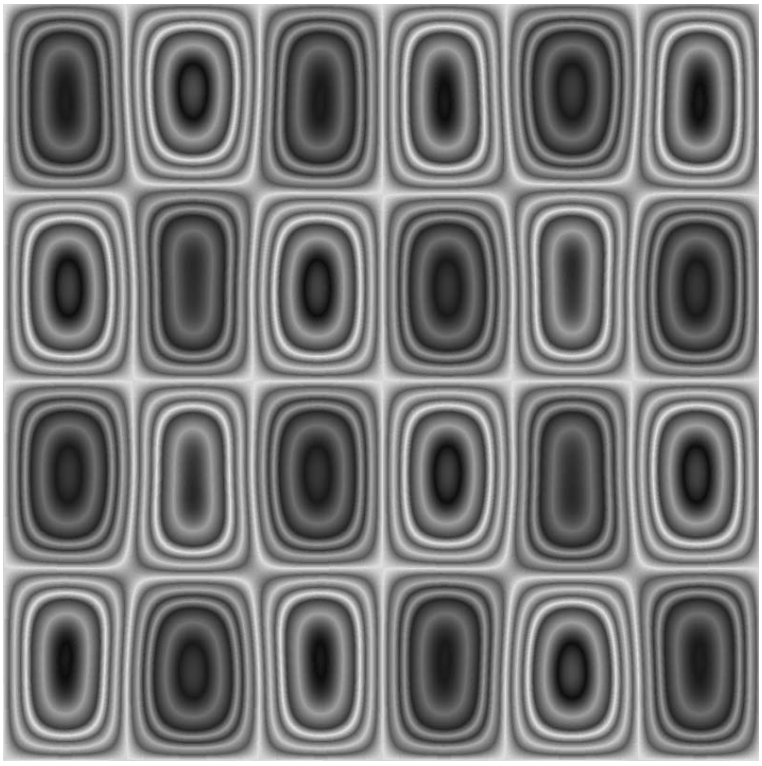


FIG. 1. *Isolines of the solution at* $t = 1$ *for* $\delta = 0.01$.

TABLE 2
Example 4.1: *accuracy with different choices of* A. $\delta = 0.01$.

A	Δt	BD1/EP1	BD2/EP2	BD3/EP3
$A = 0$	$\Delta t = 0.01$	0.72E-03	unstable	unstable
	$\Delta t = 0.005$	0.36E-03	0.24E-04	unstable
	$\Delta t = 0.0025$	0.18E-03	0.61E-05	unstable
	$\Delta t = 0.00125$	0.90E-04	0.16E-05	unstable
$A = 1$	$\Delta t = 0.01$	0.22E-02	0.21E-03	unstable
	$\Delta t = 0.005$	0.11E-02	0.56E-04	unstable
	$\Delta t = 0.0025$	0.51E-03	0.14E-04	unstable
	$\Delta t = 0.00125$	0.25E-03	0.37E-05	0.43E-06
$A = 2$	$\Delta t = 0.01$	0.43E-02	0.32E-03	0.22E-03
	$\Delta t = 0.005$	0.19E-02	0.87E-04	0.21E-04
	$\Delta t = 0.0025$	0.88E-03	0.23E-04	0.30E-05
	$\Delta t = 0.00125$	0.43E-03	0.58E-05	0.53E-06

four time steps. It is seen that once the methods are stable, the expected order of convergence (in time) is obtained.

5. Numerical experiments: Coarsening dynamics. In this section, we present the numerical results by simulating the MBE model (1.1) in cases of both isotropic surface (2.1) and square surface (2.2). The simulations are carried out in the domain $\Omega = (0, 1000)^2$, where double periodic boundary conditions are used in the spatial directions. The initial condition is a random state by assigning a random number varying from -0.001 to 0.001 to each grid point. The second-order schemes, i.e., (3.1) for the isotropic surface model and (3.14) for the square surface model, are used in our simulations. The spatial discretization is based on a Fourier pseudospectral approximation with K denoting the Fourier mode number. In order to investigate the effect of the time and space resolution, different values of Δt and K have been tested.

5.1. Growth on the isotropic symmetry surfaces. First we carry out the simulation of the growth process for the case of isotropic surfaces. In Figures 2 and 3, the isolines of the free energy $F_{free}(\mathbf{x}, t)$ at $t = 40{,}000$ and 80,000 are plotted, respectively, with $(K, \Delta t) = (512, 1)$ and $A = 1$, where $F_{free}(\mathbf{x}, t)$ is defined by

$$F_{free} = \frac{1}{4}(|\nabla h|^2 - 1)^2 + \frac{\delta}{2}|\Delta h|^2.$$

The contourlines of F_{free} are usually used to identify the edges of the pyramidal structures since the free energy is concentrated on the edges. In these two figures the temporal evolution of the morphology of the growing surface is well visualized. It is seen that the edges of the pyramids (white areas) form a random network over the surface and separate the facets of the pyramids. The pyramids grow in time via a coarsening process, as is evident from Figures 2 and 3. Also shown is the randomness of the orientation of the pyramid edges, resulting from the isotropic nature of the surface symmetry. This result is in good agreement with the published results; see, e.g., [14].

Figure 4 presents the power laws of the growth of the interface height $\tilde{h}(t)$ and width $\lambda(t)$ of the pyramidal structures. Here $\tilde{h}(t)$ is defined by

$$\tilde{h}(t) = \left(\frac{1}{|\Omega|}\int_\Omega h^2(\mathbf{x}, t)d\mathbf{x}\right)^{\frac{1}{2}}.$$

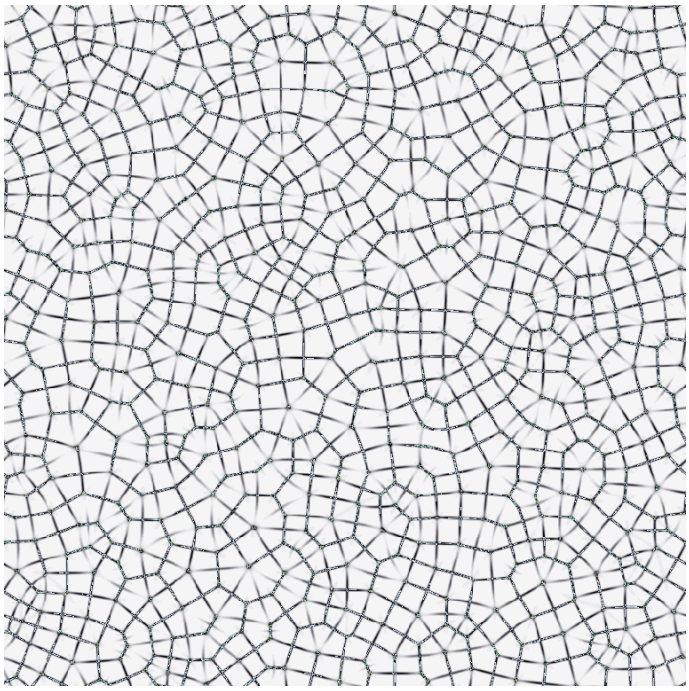


FIG. 2. *The isotropic symmetry surfaces problem: the contour plot at* $t = 40{,}000$, *obtained by using* $K = 512$ *and* $\Delta t = 1$.

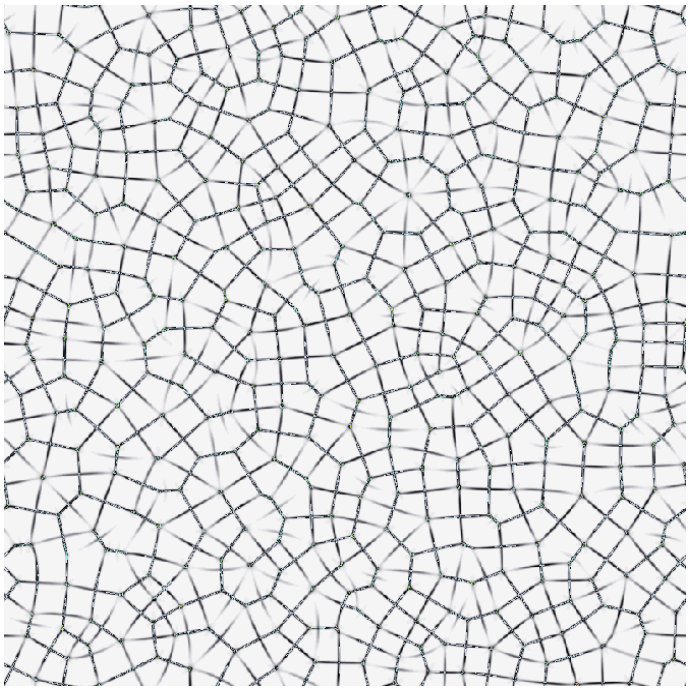


FIG. 3. *Same as Figure* 2, *except at* $t = 80{,}000$.

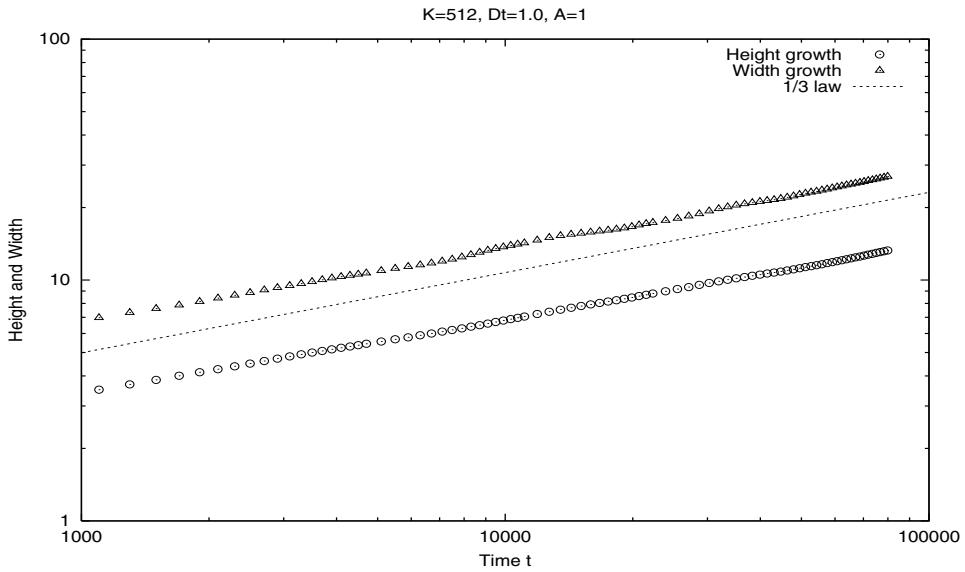


FIG. 4. *The isotropic symmetry surfaces problem: growth power law obtained by using* $K = 512$ *and* $\Delta t = 1$ *(log-to-log scale).*

The width of the pyramid edges $\lambda(t)$ measures the mean size of the network cell, which can be calculated as in [14] from the height-height correlation function

$$K_{hh}(\mathbf{r}, t) = \int_{\Omega} h(\mathbf{x} + \mathbf{r}, t) h(\mathbf{x}, t) d\mathbf{x},$$

where $\mathbf{r}$ is a positive vector. In our calculations, we used a simpler form $\mathbf{r} = (r, r)^T$. With $\mathbf{r} = (r, r)^T$, $K_{hh}(\mathbf{r}, t)$ can be considered as a function of r for fixed t, and shows an oscillatory character reflecting the presence of mound structures. For a given t, the mean pyramid width $\lambda(t)$ is defined as $r_0(t)$, which is the first zero crossing of $K_{hh}(\mathbf{r}, t)$,

$$r_0(t) = \inf\{r > 0, K_{hh}(\mathbf{r}, t) = 0\}.$$

We see from Figure 4 that both vertical height and lateral width of the pyramids grow in time as power law ct^n with exponents n close to $\frac{1}{3}$ (slope of the lines), which is again in good agreement with the existing experimental and numerical results [14, 23].

In order to check the temporal and spatial resolution, we display in Figure 5 the result obtained by using $(K, \Delta t) = (256, 0.5)$, i.e., halving the values of K and Δt. It is observed from Figures 4 and 5 that there is no significant difference between the results obtained by using the two sets of parameters.

To demonstrate the robustness of the proposed method, we plot in Figure 6 the evolution of the mean height

$$\bar{h}(t) = \frac{1}{|\Omega|} \int_{\Omega} h(\mathbf{x}, t) d\mathbf{x}.$$

It is observed that $\bar{h}(t)$ remains practically zero in the entire time intervals. This demonstrates the mass conservation which can be derived from (2.1). The energy defined in (2.7), normalized by the domain size, is plotted in Figure 7. The decay of the energy as observed in Figure 7 agrees with the theoretical result (2.6).

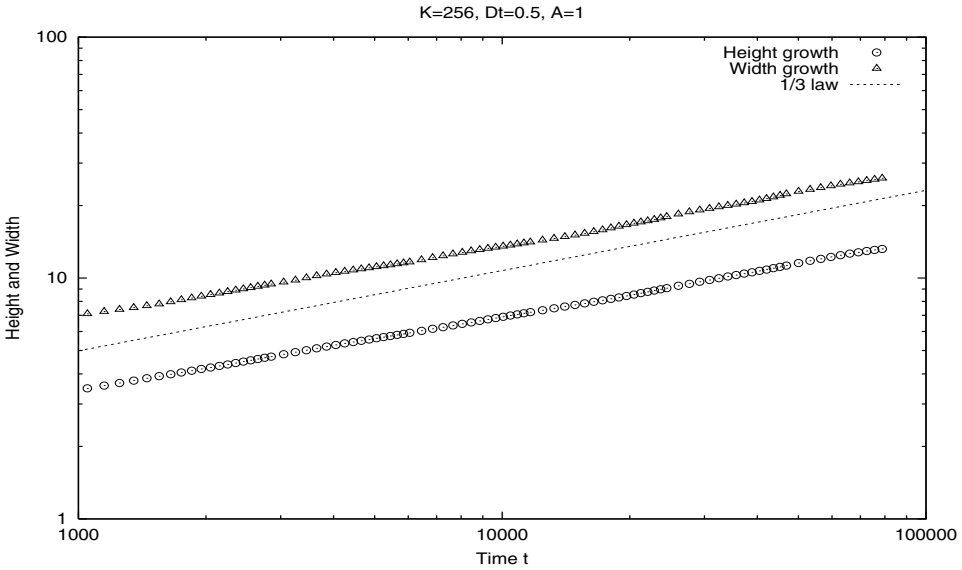


FIG. 5. *Same as Figure* 4, *except with* $K = 256$ *and* $\Delta t = 0.5$.

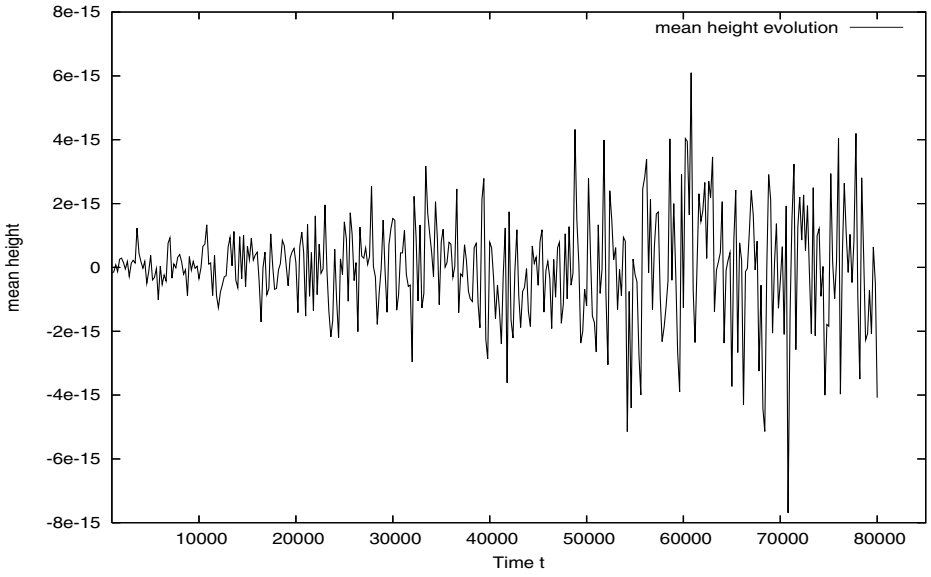


FIG. 6. *Evolution of the mean height as a function of the time.*

5.2. Growth on the square symmetry surfaces. Here we present simulation results obtained by solving the MBE model (2.2). The time discretization used in the simulation is the second-order scheme (3.14), and the space discretization is the same as in the isotropic case but here with Fourier mode number $K = 384$ and time step $\Delta t = 0.2$.

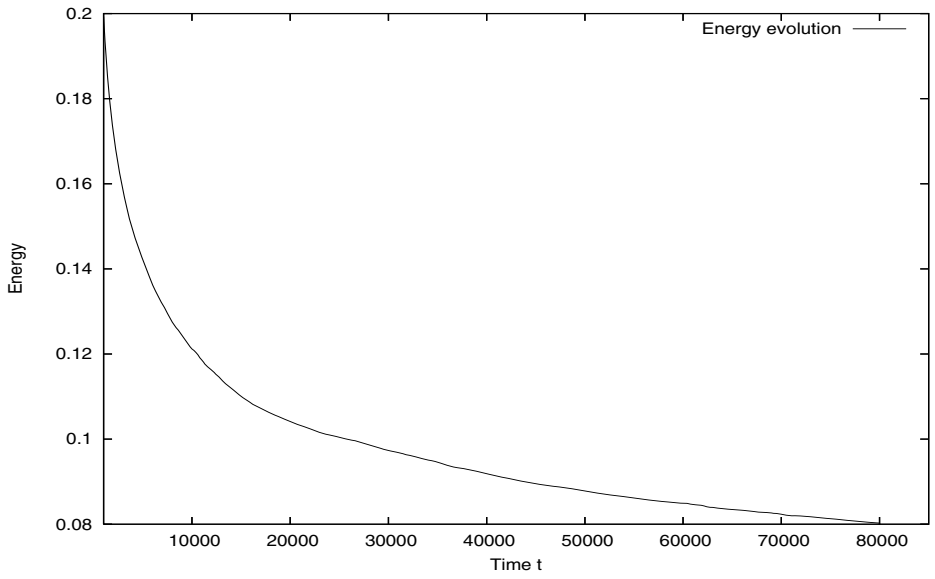


FIG. 7. *Evolution of the energy as a function of the time.*

In Figures 8 and 9, we plot the contourlines of the free energy function F'_{free} corresponding to the square symmetry model,

$$F'_{free} = \frac{\delta}{2}|\Delta h|^2 + \frac{1}{4}\left[(h_x^2-1)^2 + (h_y^2-1)^2\right].$$

As in the case of the isotropic surfaces, pyramid-like structures are growing in the surface with sharp edges carrying most of the energy, identified by the network formed by the white areas. However, in contrast to the isotropic case, the pyramid edges are well oriented toward the four preferred directions reflecting the square symmetry. A careful look at the two figures finds that the well-known dislocation feature is also presented, as reported by many experiments and simulations. Moreover, it is observed from Figure 10 that the power law obtained for the pyramid growth with the square symmetry is close to $\frac{1}{4}$. This is in good agreement with the numerical predictions of Siegert [19] and Moldovan and Golubovic [14].

6. Conclusions. In this work, we have developed and analyzed stable numerical methods for a class of nonlinear diffusion equations modeling epitaxial growth of thin films. Here, stability means that the decay of energy is preserved. In particular, we analyzed the stability properties of a class of semidiscretized (in time) schemes which are designed for large-system and long-time simulations. It is demonstrated that the classical semi-implicit method can be improved by simply adding some linear terms consistent with the truncation errors in time. The linear term consists of mixed derivatives, and the resulting numerical schemes are still semi-implicit with explicit treatment for the nonlinear terms. We also performed numerical simulations using the proposed schemes in time coupled with a Fourier spectral method in space for the molecular beam epitaxy model and determined power laws for the coarsening process. The numerical results are in good agreement with the existing ones, e.g., Moldovan and Golubovic [14] who directly solved a so-called type-A dynamics equation on a hexagonal grid.

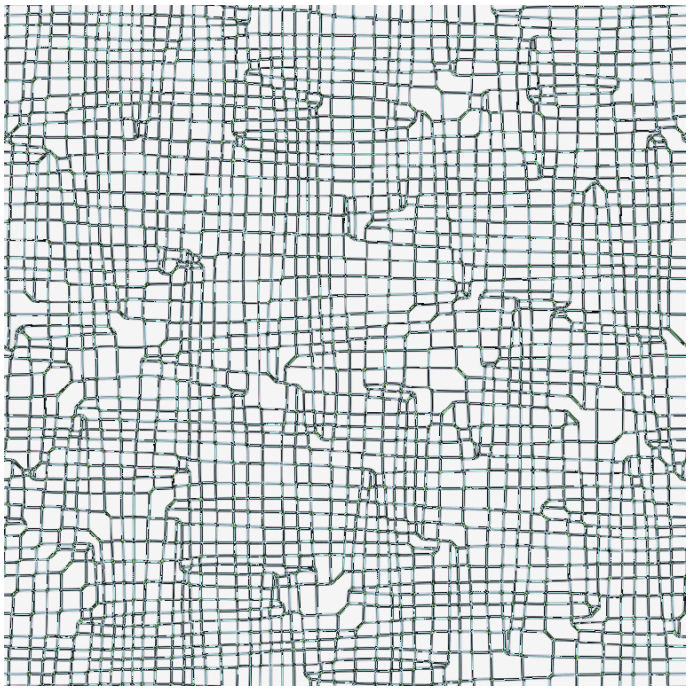


FIG. 8. *The square symmetry surface problem: the contour plot at* $t = 40{,}000$*, obtained by using* $K = 384$ *and* $\Delta t = 0.2$*.*

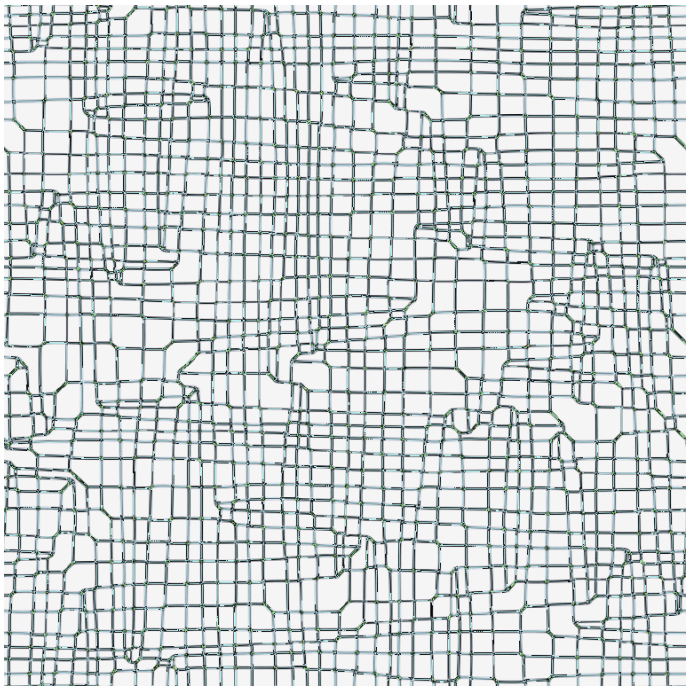


FIG. 9. *Same as Figure* 8*, except at* $t = 80{,}000$*.*

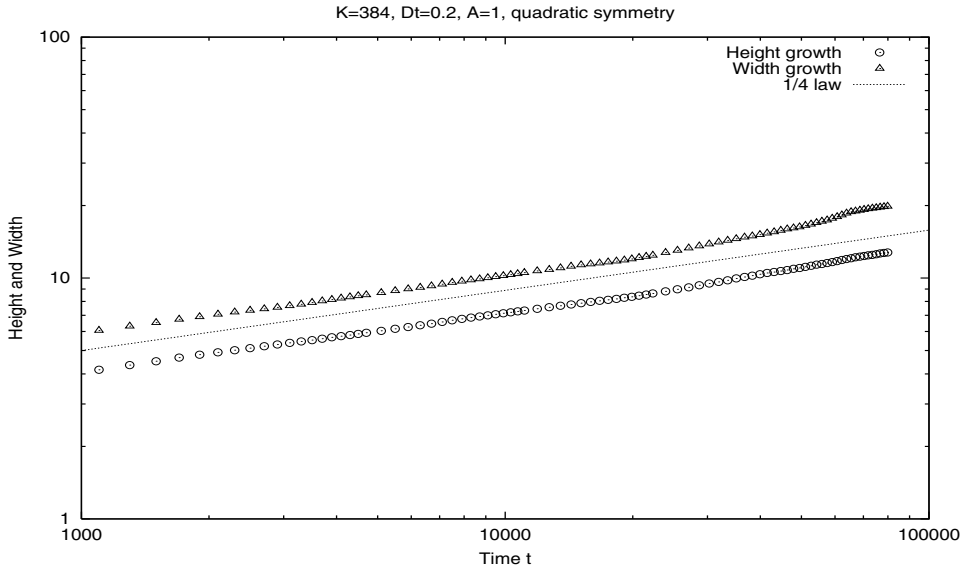


FIG. 10. *The square symmetry surface problem: growth power law obtained by using* $K = 384$ *and* $\Delta t = 0.2$.

One of the future works in this direction is to carry out more rigorous analysis for the large time-stepping techniques, including stability analysis for higher-order schemes (say, third-order time-stepping) and error analysis for the proposed schemes. Obtaining a satisfactory error bound for a numerical scheme for the MBE models seems difficult: A direct error analysis shows that the error bounds are dependent on the surface diffusion constant δ and the solution time interval, which leads to unacceptable estimates for small δ and large T. A desired bound should have weak dependence on δ and T, which seems very difficult. Other future works in this direction include adaptive time integration, i.e., treating the fast dynamics changes and slow changes separately. This is also important in improving the efficiency for the large-time simulations.

Acknowledgments. This work was motivated by several discussions and communications with Bo Li and Jian-Guo Liu during the second author's visit to the Center for Scientific Computation and Mathematical Modeling (CSCAMM) of the University of Maryland. We also thank the referees for helpful suggestions.

REFERENCES

[1] A.-L. Barabási and H. E. Stanley, *Fractal Concepts in Surface Growth*, Cambridge University Press, Cambridge, UK, 1995.

[2] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, Mineola, NY, 2001.

[3] R. E. Caflisch, M. F. Gyure, B. Merriman, S. Osher, C. Ratsch, D. D. Vvedensky, and J. J. Zinck, *Island dynamics and the level set method for epitaxial growth*, Appl. Math. Lett., 12 (1999), pp. 13–22.

[4] S. Clarke and D. D. Vvedensky, *Origin of reflection high-energy electron-diffraction intensity oscillations during molecular-beam epitaxy: A computational modeling approach*, Phys. Rev. Lett., 58 (1987), pp. 2235–2238.

[5] B. Costa, W.-S. Don, D. Gottlieb, and R. Sendersky, *Two-dimensional multi-domain hybrid spectral-WENO methods for conservation laws*, Commun. Comput. Phys., 1 (2006), pp. 550–577.

[6] G. Ehrlich and F. G. Hudda, *Atomic view of surface diffusion: Tungsten on tungsten*, J. Chem. Phys., 44 (1966), pp. 1039–1049.
[7] X. B. Feng and A. Prohl, *Error analysis of a mixed finite element method for the Cahn-Hilliard equation*, Numer. Math., 99 (2004), pp. 47–84.
[8] M. F. Gyure, C. Ratsch, B. Merriman, R. E. Caflisch, S. Osher, J. J. Zinck, and D. D. Vvedensky, *Level-set methods for the simulation of epitaxial phenomena*, Phys. Rev. E (3), 58 (1998), pp. 6927–6930.
[9] H. C. Kang and W. H. Weinberg, *Dynamic Monte Carlo with a proper energy barrier: Surface diffusion and two-dimensional domain ordering*, J. Chem. Phys., 90 (1989), pp. 2824–2830.
[10] R. V. Kohn and X. Yan, *Upper bounds on the coarsening rate for an epitaxial growth model*, Comm. Pure Appl. Math., 56 (2003), pp. 1549–1564.
[11] J. Krug, *Origins of scale invariance in growth processes*, Adv. in Phys., 46 (1997), pp. 139–282.
[12] B. Li and J. G. Liu, *Thin film epitaxy with or without slope selection*, European J. Appl. Math., 14 (2003), pp. 713–743.
[13] B. Li and J. G. Liu, *Epitaxial growth without slope selection: Energetics, coarsening, and dynamic scaling*, J. Nonlinear Sci., 14 (2004), pp. 429–451.
[14] D. Moldovan and L. Golubovic, *Interfacial coarsening dynamics in epitaxial growth with slope selection*, Phys. Rev. E (3), 61 (2000), pp. 6190–6214.
[15] M. Ortiz, E. Repetto, and H. Si, *A continuum model of kinetic roughening and coarsening in thin films*, J. Mech. Phys. Solids, 47 (1999), pp. 697–730.
[16] A. Pimpinelli and J. Villain, *Physics of Crystal Growth*, Cambridge University Press, Cambridge, UK, 1998.
[17] M. Schneider, I. K. Schuller, and A. Rahman, *Epitaxial growth of silicon: A molecular-dynamics simulation*, Phys. Rev. B, 36 (1987), pp. 1340–1343.
[18] R. L. Schwoebel and E. J. Shipsey, *Step motion on crystal surfaces*, J. Appl. Phys., 37 (1966), pp. 3682–3686.
[19] M. Siegert, *Ordering dynamics of surfaces in molecular beam epitaxy*, Phys. A, 239 (1997), pp. 420–427.
[20] M. Siegert and M. Plischke, *Slope selection and coarsening in molecular beam epitaxy*, Phys. Rev. Lett., 73 (1994), pp. 1517–1520.
[21] E. Tadmor, *Super-viscosity and spectral approximations of nonlinear conservation laws*, in Numerical Methods for Fluid Dynamics IV, M. J. Baines and K. W. Morton, eds., Oxford University Press, New York, 1993, pp. 69–81.
[22] J. Y. Tsao, *Materials Fundamentals of Molecular Beam Epitaxy*, World-Scientific, Singapore, 1993.
[23] F. Tsui, J. Wellman, C. Uher, and R. Clark, *Morphology transition and layer-by-layer growth of Rh*(111), Phys. Rev. Lett., 76 (1996), pp. 3164–3167.
[24] J. Villain, *Continuum models of critical growth from atomic beams with and without desorption*, J. Phys. I, 1 (1991), pp. 19–42.
[25] D. Xiu and J. Shen, *An efficient spectral method for acoustic scattering from rough surfaces*, Commun. Comput. Phys., to appear.
[26] J. Zhu, L.-Q. Chen, J. Shen, and V. Tikare, *Coarsening kinetics from a variable-mobility Cahn-Hilliard equation: Application of a semi-implicit Fourier spectral method*, Phys. Rev. E (3), 60 (1999), pp. 3564–3572.

SIAM J. NUMER. ANAL.
Vol. 44, No. 4, pp. 1780–1800

FINDING NUMERICAL DERIVATIVES FOR UNSTRUCTURED AND NOISY DATA BY MULTISCALE KERNELS*

LEEVAN LING†

Abstract. The recently developed multiscale kernel of R. Opfer [*Adv. Comput. Math.*, 25 (2006), pp. 357–380] is applied to approximate numerical derivatives. The proposed method is truly mesh-free and can handle unstructured data with noise in any dimension. The method of Tikhonov and the method of L-curve are employed for regularization; no information about the noise level is required. An error analysis is provided in a general setting for all dimensions. Numerical comparisons are given in two dimensions which show competitive results with recently published thin plate spline methods.

Key words. numerical differentiation, multiscale kernel, multivariate interpolation, unstructured data, inverse problems, Tikhonov regularization, L-curve.

AMS subject classifications. 65D05, 65D25, 65J20, 65J22

DOI. 10.1137/050630246

1. Introduction. Evaluating derivatives of a function using only information from discrete function values is a typical ill-posed problem. Small measurement errors, including rounding errors, will be greatly amplified during the numerical differentiation process. The problem of *numerical differentiation* arises in many branches of science and engineering. Some practical examples are the identification of discontinuities in image reconstruction [10, 13], resolution enhancement of spectra [17], solving Abel integral equations [7, 12], determination of peaks in chemical spectroscopy [24], determination of discontinuous points of the exact solutions [33], solving integral equations [8], determination of source parameter and diffusion coefficient in parabolic differential equations [6, 14], simulation of constrained mechanical systems of particles [19], singular convolution [25], and many other inverse problems in mathematical physics. The previous literature on numerical differentiation featured plenty of nicely calculated practical solutions, but most research papers on this topic are limited to one dimension or highly structured grids [4, 14, 20, 26, 27, 30, 33]. Numerical methods for higher dimensions are very limited. In particular, many existing methods are based on finite difference schemes [2], wavelet methods [5], and thin plate splines approximation [34]. The goal of this paper is to supply a new, efficient, and practical alternative for scientists and engineers who need to compute numerical differentiation from real-life, large-scale, and noisy multivariate data.

Given some set of real-life data in any dimension, multivariate functions are reconstructed from unstructured data by some specially designed multiscale kernels

$$\Phi(x,\cdot)=\sum_{j=0}^{u}\sum_{k\in\mathbb{Z}^d}\lambda_\sigma^j\varphi(2^jx-k)\varphi(2^j\cdot-k).$$

Since multiscale kernels are proven to be positive definite, for every set of data points we can solve an interpolation problem and write the interpolant in the form

*Received by the editors April 28, 2005; accepted for publication (in revised form) March 27, 2006; published electronically September 26, 2006. This research was partially supported by a Postdoctoral Fellowship from the Japan Society for the Promotion of Science.
http://www.siam.org/journals/sinum/44-4/63024.html
†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. (lling@hkbu.edu.hk).

of the *kernel representation*:

$$s = \sum_{i=1}^{n} \beta_i \Phi(x_i, \cdot). \tag{1.1}$$

The multiscale property, found in wavelet analysis, is considered a major breakthrough in the development of kernel-based mesh-free methods. We can go one step further and express (1.1) in its *frame representation*:

$$s = \sum_{j=1}^{u} \sum_{k \in \mathbb{Z}^d} \lambda_\sigma^j c_k^j \varphi(2^j \cdot - k), \tag{1.2}$$

where $c_k^j = c_k^j(\{x_i\}, \beta_i)$ are called the frame coefficients. The interpolant obtained will have a frame representation on structured grids instead of the unstructured data. The solution process involves solving a sparse matrix system if the multiscale kernel is compactly supported. Once we determine the multivariate function that interpolates the noisy data, this newly developed method has potential applications in many branches of science and engineering. The well-developed wavelet techniques (e.g., denoising, compression, shape detection, etc.) can be applied thereafter. In this paper, we focus on a classical ill-posed numerical differentiation problem. The derivative of (1.2) can be obtained by replacing φ by $D^\gamma \varphi$. An overview of multiscale kernels will be given in section 2.

In section 3, the instability of numerical differentiation is regularized by the Tikhonov regularization method that seeks a stable approximate interpolant. Error estimates in section 3.1 show that the errors of numerical derivatives blow up when the noise level is high or when the minimum separation distance of the data points is small. This agrees with the ill-posed nature of numerical differentiation. On the other hand, both errors in interpolation and in the derivatives can be minimized with an optimal regularization parameter. In section 4, the L-curve method is employed to numerically locate the optimal regularization parameter. Finally, two bivariate examples are given in section 5 to conclude the paper.

2. Finding numerical derivatives. Consider a symmetric function of the form $\Phi : \Omega \times \Omega \to \mathbb{R}$ for some $\Omega \subset \mathbb{R}^d$ and let $\mathcal{N}_\Phi$ be the *reproducing kernel* of a *native* Hilbert space [29] of Φ. It is proven in the same article that the native space $\mathcal{N}_\Phi$ for a given symmetric positive definite kernel Φ is unique if it exists, and it coincides with the closure of the space of finite linear combination of functions $\Phi(x, \cdot)$, $x \in \Omega$ under the inner product defined via

$$(\Phi(x, \cdot), \Phi(y, \cdot))_{\mathcal{N}_\Phi} = \Phi(x, y) \quad \text{for all} \quad x, y \in \Omega.$$

That is, for every fixed point $x \in \Omega$ and function $\Phi(x, \cdot)$ belongs to $\mathcal{N}_\Phi$, every $f \in \mathcal{N}_\Phi$ can be recovered by an inner product of the form $f(x) = \langle f, \Phi(x, \cdot) \rangle$, $x \in \Omega$. For a detailed treatise of reproducing kernel Hilbert spaces, see Aronszajn [3] or Meschkowski [21].

To begin, we reconstruct multivariate functions from unstructured data by a multiscale technique. The basic concepts of this technique were first investigated by Opfer [23]. The implementation of a *multiscale kernel* (MSK) is out of the scope of this paper and the developments of MSK are only sketched here. We refer the reader to the original dissertation of Opfer for the details.

A function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is called *refinable* if there is a sequence $\{\mathfrak{h}_k\}_{k\in\mathbb{Z}^d}$ of real numbers such that

$$\varphi = \sum_{k\in\mathbb{Z}^d} \mathfrak{h}_k \varphi(2\cdot - k). \tag{2.1}$$

For every level-$j \in \mathbb{Z}$ we define the *shift invariant space*

$$\mathcal{V}_j := \left\{ \sum_{k\in\mathbb{Z}^d} c_k \varphi(2^j \cdot - k) : c_k \in \mathbb{R}, \sum_{k\in\mathbb{Z}^d} (c_k)^2 < \infty \right\}. \tag{2.2}$$

By standard wavelet arguments it follows from (2.1) that the spaces $\{\mathcal{V}_j\}_{j\in\mathbb{Z}}$ form a nested sequence, i.e., $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_u$. The main idea here involves several levels of $\mathcal{V}_j$ in *one* reconstruction scheme.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a function in $L^2(\mathbb{R}^d)$ with decay $\varphi(x) = \mathcal{O}\big((1+\|x\|)^{-(d+1)/2}\big)$. Let $u \geq 0$ be a fixed integer and $\sigma > d/2$ be a positive real number. Then the kernel $\Phi_\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by

$$\Phi_\sigma(x,y) := \sum_{j=0}^{u} \lambda_\sigma^j \underbrace{\left(\sum_{k\in\mathbb{Z}^d} \varphi(2^j x - k)\varphi(2^j y - k) \right)}_{\Phi_{\sigma,j}}, \tag{2.3}$$

where $\lambda_\sigma := 2^{d-2\sigma}$, is a MSK.

THEOREM 2.1 (see [23, Theorem 5.4]). *Every MSK in the form of* (2.3) *is positive semidefinite. Let $B_\rho(c)$ be a ball of radius ρ with center $c \in \mathbb{R}^d$ such that* $\operatorname{supp}(\varphi) \subset B_\rho(c)$. *If the point set $X \subset \mathbb{R}^d$ satisfies*

$$h_{X,\min} := \min_{i\neq j} \|x_i - x_j\|_2 > \rho\, 2^{-u+1}, \tag{2.4}$$

then the matrix $A_X := \big(\Phi_\sigma(x_i,x_k)\big)_{1\leq i,k\leq n}$ is positive definite.

In this paper, we are mainly interested in compactly supported refinable functions φ that clearly satisfy the decay condition required in Theorem 2.1. The resulting MSK are therefore positive definite.

We can find to any given data Y an interpolant of the form (1.1) by solving a sparse symmetric linear collocation system for $\beta \in \mathbb{R}^n$,

$$y_j = \sum_{i=1}^{n} \beta_i \Phi_\sigma(x_i,\, x_j), \quad 1 \leq j \leq n. \tag{2.5}$$

Theorem 2.1 implies that (2.5) has a unique solution if the integer $u = u(h_{X,\min})$ is large enough with respect to the density of the data points X. The MSK scheme is based on the following idea: The kernel representation can be decomposed into a frame representation due to the specially designed structure of Φ_σ. First, $s \in \mathcal{N}_\Phi$ is decomposed into a sequence of functions $s_j \in \mathcal{V}_j$,

$$s = \sum_{i=1}^{n} \beta_i \Phi_\sigma(x_i,\cdot) = \sum_{i=1}^{n} \beta_i \sum_{j=0}^{u} \lambda_\sigma^j \Phi_{\sigma,j}(x_i,\cdot) = \sum_{j=0}^{u} \lambda_\sigma^j \underbrace{\sum_{i=1}^{n} \beta_i \Phi_{\sigma,j}(x_i,\cdot)}_{s_j} = \sum_{j=0}^{u} \lambda_\sigma^j s_j, \tag{2.6}$$

such that each $s_j \in \mathcal{V}_j$ can be further decomposed into

(2.7)
$$s_j = \sum_{i=1}^{n} \beta_i \Phi_{\sigma,j}(x_i, \cdot) = \sum_{k\in\mathbb{Z}^d} \underbrace{\left(\sum_{i=1}^{n} \beta_i \varphi(2^j x_i - k) \right)}_{c_k^j} \varphi(2^j \cdot - k) = \sum_{k\in\mathbb{Z}^d} c_k^j \varphi(2^j \cdot - k).$$

Combining (2.6) and (2.7) gives us the frame representation in the form of (1.2). Functions in lower levels capture the smooth structure of f while the higher levels contain the fine structure of f, including noise. Furthermore, the refinability of the function φ allows the *frame coefficients* c_k^j for $0 \le j \le u-1$ to be computed via

$$c_k^j = \lambda_\sigma^{-j} \sum_{\mu\in\mathbb{Z}^d} \mathfrak{h}_{\mu-2k} c_\mu^{j+1}, \quad k \in \mathbb{Z}^d.$$

Computation of frame coefficients c_k^j requires a nearest neighbor search, e.g., kd-tree [35, Chapter 14], to locate all $x \in X$ inside the support of $\varphi(2^u \cdot - k)$. Note that the number of nonzero c_k^j is finite due to the fact that $|X|$ is finite and φ is compactly supported. The native space $\mathcal{N}_\Phi$ and each $\mathcal{V}_j$ in (2.2) can be equipped with a norm, respectively,

$$\|s\|_{\mathcal{N}_\Phi}^2 = \sum_{j=0}^{u} \lambda_\sigma^{-j} \|s_j\|_{\mathcal{V}_j}^2 \quad \text{and} \quad \|s_j\|_{\mathcal{V}_j}^2 = \sum_{k\in\mathbb{Z}^d} (c_k^j)^2.$$

Let $h_{X,\Omega}$ denote the fill distance of the data points $X \subset \Omega$ given by

$$h_{X,\Omega} := \sup_{y\in\overline{\Omega}} \inf_{x_i\in X_h} \|y - x_i\|_2.$$

If φ satisfies certain smoothness and decay properties, then $\mathcal{N}_\Phi \simeq W^{\sigma,2}$ are norm equivalent and the interpolant obtained by MSK satisfies the standard native space error bound:

THEOREM 2.2 (see [23, Theorem 5.21]). *Let the MSK Φ_σ be constructed with a scaling function φ of an r-regular multiscale analysis of $L^2(\Omega^d)$ with $r > d/2$. Fix an σ with $d/2 < \sigma < r$. Further we assume that $X := \{x_1, \dots, x_n\} \subset \Omega$ is a set of points with fill distance $h_{X,\Omega}$, where $\Omega \subset \mathbb{R}^d$ is a compact set with Lipschitz boundary which satisfies an interior cone condition. Let $f \in H^\sigma(\mathbb{R}^d)$ and s be the interpolant. Let $1 \le q \le \infty$ and $\gamma = (\gamma_1, \dots, \gamma_d)$ be a multi-index such that $|\gamma| < \lfloor\sigma\rfloor - d/2$. Then, there is a constant $C > 0$ independent of f and $h_{X,\Omega}$ such that*

$$\|s - f\|_{W^{|\gamma|,q}(\Omega)} \le C_1 h_{X,\Omega}^{\sigma-|\gamma|-d\left(1/2-1/q\right)_+} \|f\|_{\mathcal{N}_\Phi},$$

where $(x)_+ = x$ if $x \ge 0$ and $(x)_+ = 0$ if $x < 0$.

2.1. Noise data. Let us assume we have points $X := \{x_1, \dots, x_n\} \subset \Omega \subset \mathbb{R}^d$ and noisy data

$$Y_\eta := \{\widetilde{y}_1, \dots, \widetilde{y}_n\} \subset \mathbb{R},$$

where

$$\widetilde{y}_i = y_i + \delta_i = f(x_i) + \eta(x_i),$$

and δ_i are random noise. The noise function η here is not necessarily classically differentiable or even continuous. Assume that we obtain an interpolant in the frame representation

$$s_{\delta,X} = \sum_{j=0}^{u} \lambda_\sigma^j s_{\delta,X,j} = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j c_k^j \varphi(2^j \cdot -k), \tag{2.8}$$

for some noisy data (X, Y_η) by MSK with the following conditions satisfied.

Assumption 2.3. The MSK in (2.3) is constructed by

1. $\sigma \geq 2$ and $\sigma > \frac{d}{2}$,
2. a r-regular φ smooth enough such that $r > \left(2 + \frac{d}{2}\right)$, i.e., $\varphi \in C^r(\Omega)$ with compact support up to order r, and
3. for any given data points X, $u = \lceil 1 + \log_2 \frac{\rho}{h_{X,\min}} \rceil$, where $h_{X,\min}$ is given in (2.4) and ρ as in Theorem 2.1.

The reasons for the above assumptions will soon become clear when we look at the error estimates in section 3.1. Throughout this paper, let γ with $|\gamma| = \gamma_1 + \cdots + \gamma_d = 1$ be a multi-index. Our interest is to approximate or reconstruct the derivatives of f from the noisy data Y_η via

$$(X,\, Y_\eta) \longrightarrow D^\gamma f.$$

From the frame representation (2.8), the numerical derivatives are given by

$$D^\gamma s_{\delta,X} = \sum_{j=0}^{u} \lambda_\sigma^j\, D^\gamma\, s_{\delta,X,j} = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j c_k^j\, D^\gamma \varphi(2^j \cdot -k). \tag{2.9}$$

This numerical procedure is highly unstable. Since the input data Y_η contains noise, the resulting approximated derivatives $D^\gamma s_{\delta,X}$ will contain large errors and therefore are not trustworthy. We select a subset of frame coefficients $\{r_k^j\} \subset \{c_k^j\}$ to *regularize* the numerical derivatives.

Any regularized interpolant g to $s_{\delta,X}$ is in the form of

$$g = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j r_k^j\, \varphi(2^j \cdot -k), \tag{2.10}$$

where $r_k^j \in \{0,\, c_k^j\}$. For some threshold $t_\sigma(j) > 0$ for $0 \leq j \leq u$ and a fixed regularization parameter α, the regularized interpolant is defined to be

$$s_\alpha = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j\, r_k^j\, \varphi(2^j \cdot -k) \;\text{ such that }\; r_k^j = \begin{cases} c_k^j & \text{if } |c_k^j| > t_\sigma(j)\,\alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

For practical problems, the optimal regularization parameter α^* is not attainable unless η is known *a priori*. In the next section, we specify our choice of threshold $t_\sigma(j)$ using the Tikhonov regularization method. After giving a concrete formula of the threshold $t_\sigma(j)$, we make sure the errors in interpolation and in the gradient of the regularized interpolant in (2.11) is both bounded and well behaved for some suitable α.

3. Regularization. The classical Tikhonov regularization method [31] is a common tool for finding a solution from an unstable system. Using some *a priori* choice strategy for regularization parameters, Hofmann and Yamamoto [18] prove convergence rates for the Tikhonov regularization method. Despite the differences with the classical problem, we seek a *regularized interpolant* s_α to $s_{\delta,X}$ (considered to be fixed here) by the Tikhonov regularization method. For any

$$\tilde{g} = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j \tilde{r}_k^j \, \varphi(2^j \cdot -k) \in \mathcal{V}_u,$$

we define the *error measure* by

$$E(\tilde{g}) = E(\tilde{g}; s_{\delta,X}) := \|s_{\delta,X}\|_{\mathcal{N}_\Phi}^2 - \|g\|_{\mathcal{N}_\Phi}^2 = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^{-j} \big((c_k^j)^2 - (\tilde{r}_k^j)^2\big), \tag{3.1}$$

and the *roughness measure* by

$$R(\tilde{g}) := \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j |\tilde{r}_k^j| \, |\varphi(2^j \cdot -k)|_{W^{2,2}(\Omega)}, \tag{3.2}$$

such that $|\tilde{g}|_{W^{2,2}(\Omega)}^2 \le R(\tilde{g})$ for any $\tilde{g} \in \mathcal{V}_u$. The error measure depends on the interpolant $s_{\delta,X}$ but both are independent of α.

Given any regularization parameter $\alpha \ge 0$ (consider to be fixed here), the *regularized interpolant* s_α is defined to be the minimizer of $E(\cdot) + \alpha\, R(\cdot)$ over all functions in the form of (2.10), i.e.,

$$E(s_\alpha) + \alpha\, R(s_\alpha) = \inf \big\{ E(g) + \alpha\, R(g) \ \text{ for all } \ g \text{ as in (2.10)} \big\}. \tag{3.3}$$

Although the number of nonzero functions in the form of (2.10) is finite, we have the following theorem to simplify our selection process.

THEOREM 3.1. *For any given $\alpha \ge 0$ the optimizer to* (3.3) *is given by* (2.11) *with*

$$t_\sigma(j) := \big(2^{d-2\sigma+4} \, |\varphi|_{W_{2,2}}^2\big)^j < \infty \ \text{ for all } \ 0 \le j \le u < \infty.$$

Proof. First by changing variables, we obtain

$$|\varphi(2^j \cdot -k)|_{W_{2,2}(\Omega)}^2 = \left\| \sum_{|\gamma|=2} D^\gamma \varphi(2^j \cdot -k) \right\|_{L^2(\Omega)}^2 = 2^{j(4-d)} \, |\varphi|_{W_{2,2}(\Omega)}^2. \tag{3.4}$$

For any g in the form of (2.10), we have

$$\begin{aligned} E(g) + \alpha\, R(g) &= \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \big(\big(\lambda_\sigma^{-j} (c_k^j)^2 - (r_k^j)^2\big) + \alpha\, \lambda_\sigma^j \, |r_k^j| \, 2^{j(4-d)} \, |\varphi|_{W_{2,2}(\Omega)}^2 \big) \\ &= \underbrace{\left(\sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^{-j} (c_k^j)^2 \right)}_{= \|s_{\delta,X}\|_{\Phi_\sigma}^2} - \left(\sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^{-j} (r_k^j)^2 - \alpha\, \lambda_\sigma^j \, |r_k^j| \, 2^{j(4-d)} \, |\varphi|_{W_{2,2}(\Omega)}^2 \right). \end{aligned}$$

Since $\|s_{\delta,X}\|^2_{\Phi_\sigma}$ is a fixed quantity, the minimizer of (3.3) corresponds to the following condition on r_k^j:

$$\lambda_\sigma^{-j}(r_k^j)^2 - \alpha\,\lambda_\sigma^j\,|r_k^j|\,2^{j(4-d)}\,|\varphi|^2_{W_{2,2}(\Omega)} > 0.$$

After simplification, we obtain $(r_k^j)^2 > t_\sigma(j)|r_k^j|\alpha$. □

Once α is determined, Theorem 3.1 allows us to select $\{r_k^j\}$ from $\{c_k^j\}$ and construct the regularized interpolant and its derivatives.

3.1. Error estimate. In general, interpolation does not make sense on $L^2(\Omega)$ and there are many possibilities of projecting $L^2(\Omega)$ to $\mathcal{N}_\Phi$. Moreover, there are many new results on interpolation in cases where f is not in the native space [22, 28]. For our problem, we will define the necessary projection by interpolation.

Let $\Omega \subset \mathbb{R}^d$ be a domain satisfying the conditions in Theorem 2.2. Suppose that the MSK Φ_σ also satisfies Assumption 2.3 and $f \in \mathcal{N}_\Phi = H^\sigma(\Omega)$. For any fixed center X and noise function $\eta \in L^2(\Omega) \cap C(\Omega)$, the *noise level* is defined as

$$\delta := \sup_{x\in\Omega} |\eta(x)|.$$

It is easy to verify that $\|\eta\|_{L^2(\Omega)} \le V^{1/2}(\Omega)\,\delta$, where $V(\Omega)$ is the volume of $\Omega \subset \mathbb{R}^d$. The noisy input data for interpolation at the points $X \subset \Omega$ is given by $Y_\eta := (f+\eta)\big|_X$ under the assumption that f and η are both well defined at all points $x \in \Omega$.

We define a finite dimensional subspace $V_X \subset \mathcal{N}_\Phi$ to be the span of $\Phi_\sigma(z,\,\cdot)$ and $V_X^{(\gamma)}$ to be the span of $D^\gamma\Phi_\sigma(z,\,\cdot)$, where differentiation acts upon the second variable of Φ_σ for all $z \in X$. Furthermore, we define a *projection map*

$$P_X : L^2(\Omega)\cap C(\Omega) \to \mathbb{R}^{|X|} \;\text{ such that }\; P_X f = \{f(x) :\; x \in X\}$$

that extracts discrete values from a function in $L^2(\Omega)\cap C(\Omega)$ at X so that interpolation is possible and makes sense, and an *interpolation map*

$$I_X : \mathbb{R}^{|X|} \to V_X \;\text{ such that }\; I_X P_X f = I_X f \;\text{ for all }\; f \in \mathcal{N}_\Phi,$$

which maps discrete function values at X to a function in V_X by interpolation using MSK. Last, we define a *truncation map*,

$$T_\alpha : \{1\}^{\mathbb{N}\times\mathbb{Z}^d} \to \{0,\,1\}^{\mathbb{N}\times\mathbb{Z}^d} \;\text{ for all }\; \alpha \ge 0$$

that smoothes out functions by truncating some of their frame coefficients. Furthermore, when no confusion arises, we treat T_α as a map from V_X and $V_X^{(\gamma)}$ onto themselves in the sense that,

$$T_\alpha\left(\sum_{j=0}^{u}\sum_{k\in\mathbb{Z}^d}\lambda_\sigma^j\,c_k^j\,\phi(2^j\cdot - k)\right) := \sum_{j=0}^{u}\sum_{k\in\mathbb{Z}^d}\lambda_\sigma^j\,T_\alpha(c_k^j)\,\phi(2^j\cdot - k),\;\; \phi = \{\varphi,\,D^\gamma\varphi\}.$$

The truncation map T_α, as in (2.11), is a nonlinear map whose actual form depends on the parameter α and the data $(X,\,Y_\eta)$. It can also be interpreted as a countable set $\{\tau_k^j\} \subset \{0,\,1\}^{\mathbb{N}\times\mathbb{Z}^d}$ such that $T_\alpha(c_k^j) = \tau_k^j(\alpha)c_k^j = r_k^j(\alpha)$, where

$$\tau_k^j = \tau_k^j(\alpha) = \begin{cases} 1 & \text{if } r_k^j = c_k^j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Since the number of nonzero $c_k^j \in \{0, 1\}^{\mathbb{N}\times\mathbb{Z}^d}$ is finite, there are infinitely many $c_k^j = 0$ and the corresponding $\tau_k^j = 1$ because $r_k^j = 0 = c_k^j$ for all $\alpha \geq 0$ by (3.5). Thus, there are infinitely many $\tau_k^j = 1$ (frame coefficients being kept) and only a finite number of $\tau_k^j = 0$ (frame coefficients being truncated) for the selected frame coefficients.

With the newly introduced notation, the unknown full interpolant can be expressed by $s := I_X P_X f$. Furthermore, we can write the regularized interpolant in Theorem 3.1 as

$$s_{\delta,X} := I_X P_X (f - \eta) \quad \text{and} \quad s_\alpha = T_\alpha s_{\delta,X}.$$

Moreover, (2.11) can be restated as

$$s_\alpha = T_\alpha s_{\delta,X} = \sum_{j=0}^{u} \sum_{k\in\mathbb{Z}^d} \lambda_\sigma^j \, \tau_k^j c_k^j \, \varphi(2^j \cdot - k).$$

Without any extra assumptions on the noise function η, the threshold $t_\sigma(j)$ and the data points X, the truncation map has the following properties.

PROPOSITION 3.2. *Let $|\gamma| = 1$ and $\mathrm{nz}_j(\cdot)$ be a function with respect to j that returns the number of zero elements in the level-j of a set in $\{0, 1\}^{\mathbb{N}\times\mathbb{Z}^d}$. Denote the $L^2(\Omega)$-induced norm for maps on V_X by $\|\cdot\|_{L^2(\Omega)}$ and define*

$$u_\alpha := \sup\left\{ j \,\middle|\, \tau_k^j \neq 0 \text{ for some } k \in \mathbb{Z}^d, 0 \leq j \leq u \right\}, \tag{3.6}$$

to be the maximum nonzero frame level after truncation. Then the truncation map T_α satisfies:

1. $\|T_\alpha\|_{L^2(\Omega)} = \|T_0 - T_\alpha\|_{L^2(\Omega)} = 1$ *for* $\alpha > 0$.
2. $\|D^\gamma T_\alpha\|_{L^2(\Omega)} = \|T_\alpha D^\gamma\|_{L^2(\Omega)} = 2^{u_\alpha} \, \|D^\gamma \varphi\|_{L^2(\Omega)} \, \|\varphi\|_{L^2(\Omega)}^{-1}$.
3. *For any given data (X, Y_η), the number of frame coefficients being truncated by T_α, denoted by $\mathrm{nz}_j(1 - \tau_k^j(\alpha)) < \infty$, is a bounded nondecreasing simple function in α and $\mathrm{nz}_j(1 - \tau_k^j(0)) = 0$.*

Proof. The perfect candidate to evaluate the above norms is the scaled function in the frame. For each nested space $\mathcal{V}_j$ $(0 \leq j \leq u)$, such function is given by

$$g_{j,k} = \left(2^{jd/2} \, \|\varphi\|_{L^2(\Omega)}^{-1}\right) \varphi(2^j \cdot - k) \in \mathcal{V}_j, \ 0 \leq j \leq u,$$

such that $\|g_{j,k}\|_{L^2(\Omega)} = 1$ and $\|D^\gamma g_{j,k}\|_{L^2(\Omega)} = 2^j \, \|D^\gamma \varphi\|_{L^2(\Omega)} \, \|\varphi\|_{L^2(\Omega)}^{-1}$.

For Proposition 3.2.1 follows directly from the fact that $T_\alpha \neq 0$ for all $\alpha \geq 0$; there exists some (j_1, k_1) and (j_2, k_2) such that $\tau_{k_1}^{j_1} = 1$ and $\tau_{k_2}^{j_2} = 0$ for $0 \leq j_i \leq u$ and $k_i \in \mathbb{Z}^d$ corresponding to a frame coefficient that is kept and truncated by T_α, respectively. Hence, we have

$$\|T_\alpha I_X P_X g_{j_1,k_1}\|_{L^2(\Omega)} = 1, \text{ and } \|(T_0 - T_\alpha) I_X P_X g_{j_2,k_2}\|_{L^2(\Omega)} = 1.$$

To prove Proposition 3.2.2, we first note that the differential operator acts on each φ independently as in (2.9); thus, c_k^j and τ_k^j are independent of the truncation process. *Differentiation after truncation* is the same as *truncation after differentiation*, namely we have $D^\gamma T_\alpha s_j = T_\alpha D^\gamma s_j$ for all $s_j \in \mathcal{V}_j$. For numerical efficiency, the operation $D^\gamma T_\alpha$ is preferred for efficiency.

Since $\|D^\gamma\varphi\|_{L^2(\Omega)}\,\|\varphi\|^{-1}_{L^2(\Omega)}$ is a fixed quantity once φ is fixed, without regularization the noise in the level-j will be greatly amplified as expected,

$$\|D^\gamma I_X P_X g_{j,k}\|_{L^2(\Omega)} = \|D^\gamma g_{j,k}\|_{L^2(\Omega)} \le 2^j\,\|D^\gamma\varphi\|_{L^2(\Omega)}\,\|\varphi\|^{-1}_{L^2(\Omega)}. \tag{3.7}$$

Let u_α be the highest nonzero frame level appearing in the regularized interpolant as in (3.6). Applying the regularization map T_α will "cut off" all levels higher than u_α exclusively and we arrive at the conclusion.

Last, Proposition 3.2.3 follows from the fact that the number of nonzero c_k^j is finite and no regularization is applied when $\alpha = 0$. □

We now turn our focus to the error estimate for $\|f - s_\alpha\|$. First of all,

$$\begin{aligned}\|f - s_\alpha\|_{L^2(\Omega)} &\le \|f - I_X P_X f\|_{L^2(\Omega)} + \|I_X P_X f - s_{\delta,X}\|_{L^2(\Omega)} + \|s_{\delta,X} - T_\alpha s_{\delta,X}\|_{L^2(\Omega)}\\ &= \underbrace{\|f - I_X P_X f\|_{L^2(\Omega)}}_{\text{interp. error}} + \underbrace{\|I_X P_X \eta\|_{L^2(\Omega)}}_{\text{noise}} + \underbrace{\|(T_0 - T_\alpha)s_{\delta,X}\|_{L^2(\Omega)}}_{\text{reg. error}}.\end{aligned}$$

The last inequality uses the fact that

$$\|I_X P_X f - s_{\delta,X}\| = \|I_X P_X f - I_X P_X (f - \eta)\| = \|I_X P_X \eta\|.$$

By Theorem 2.2 with $q = 2$ and $|\gamma| = 0$, the first term (interpolation error) can be bounded by

$$\|I_X P_X f - f\|_{L^2(\Omega)} \le C_1 h^\sigma_{X,\Omega}\,\|f\|_{\mathcal{N}_\Phi},$$

and the second term (noise) is bounded by our assumption on η,

$$\|I_X \eta\|_{L^2(\Omega)} \le V^{1/2}(\Omega)\,\delta.$$

It is straightforward to verify that

$$\|s_j\|^2_{L^2(\Omega)} \le 2^{-jd}\,\|\varphi\|^2_{L^2(\Omega)}\,\|s_j\|^2_{\mathcal{V}_j} \quad \text{for all} \quad s_j \in \mathcal{V}_j. \tag{3.8}$$

For the third term (regularization error), by Theorem 3.1 and (3.8) we have

$$\begin{aligned}\|(T_0 - T_\alpha)s_{\delta,X}\|^2_{L^2(\Omega)} &\le \sum_{j=0}^{u} \|(T_0 - T_\alpha)s_{\delta,X,j}\|^2_{L^2(\Omega)}\\ &\le \|\varphi\|^2_{L^2(\Omega)} \sum_{j=0}^{u}\sum_{k\in\mathbb{Z}^d} 2^{-jd}\big((1-\tau_k^j)c_k^j\big)^2\\ &\le \|\varphi\|^2_{L^2(\Omega)} \sum_{j=0}^{u} 2^{-jd}\mathrm{nz}_j(1-\tau_k^j)\,t_\sigma(j)^2\alpha^2\\ &\le \sum_{j=0}^{u} 2^{-2(\sigma-2)j}\mathrm{nz}_j(1-\tau_k^j)|\varphi|^{2j}_{W_{2,2}}\|\varphi\|^{2(j+1)}_{L^2(\Omega)}\alpha^2\\ &:= \big(C_2(\alpha)\alpha\big)^2.\end{aligned} \tag{3.9}$$

An immediate fact from Proposition 3.2.3 is that $C_2(\alpha)$ is a bounded positive nondecreasing simple function with $C_2(0) = 0$.

For the error in the gradient, we have

$$\begin{aligned}\|\nabla f - \nabla s_\alpha\|_{L^2(\Omega)} &\le \|\nabla f - \nabla I_X P_X f\|_{L^2(\Omega)} + \|\nabla I_X P_X f - \nabla T_\alpha I_X f\|_{L^2(\Omega)} \\ &\quad + \|\nabla T_\alpha I_X P_X f - \nabla s_\alpha\|_{L^2(\Omega)} \\ &\le \underbrace{\|\nabla f - \nabla I_X P_X f\|_{L^2(\Omega)}}_{\text{interp. error}} + \underbrace{\|\nabla (T_0 - T_\alpha) I_X P_X f\|_{L^2(\Omega)}}_{\text{reg. error}} \\ &\quad + \underbrace{\|\nabla T_\alpha I_X P_X P_X \eta\|_{L^2(\Omega)}}_{\text{noise}}.\end{aligned}$$

Using Theorem 2.2 with $q = 2$ and $|\gamma| = 1$, the interpolation error in gradient is again bounded by

$$\|\nabla I_X P_X f - \nabla f\|_{L^2(\Omega)} \le C_1 h_{X,\Omega}^{\sigma-1} \|f\|_{\mathcal{N}_\Phi}.$$

Next, we need a stronger assumption than $\sigma \ge 2$ such that $\mathcal{N}_\Phi \subseteq W^{2,2}(\Omega)$ to make use of an inequality in [1, Theorem 4.14]: For any $0 < \epsilon_0$ there exists a constant $C_3 = C_3(\epsilon_0, \Omega, d) > 0$ such that for $g \in W^{2,2}(\Omega)$ and for all $0 < \epsilon < \epsilon_0$,

$$\|\nabla g\|_{L^2(\Omega)} \le C_3\big(\epsilon\, |g|_{W^{2,2}(\Omega)} + \epsilon^{-1} \|g\|_{L^2(\Omega)}\big). \tag{3.10}$$

By assumption, the unknown function f is "smoother" than the random noise η. Hence, for all $\alpha \ge 0$ the following statement holds

$$\|\nabla (T_0 - T_\alpha) I_X P_X f\|_{L^2(\Omega)} \le \|\nabla (T_0 - T_\alpha) s_{\delta,X}\|_{L^2(\Omega)}.$$

Similar to (3.9), by (3.4) we have

$$\begin{aligned}|(T_0 - T_\alpha) s_{\delta,X}|_{W^{2,2}} &\le \sum_{j=0}^{u} |(T_0 - T_\alpha) s_{\delta,X,j}|_{W^{2,2}} \\ &\le |\varphi|_{W^{2,2}}^2 \sum_{j=0}^{u} \sum_{k \in \mathbb{Z}^d} 2^{j(2-d/2)} \big((1 - \tau_k^j) c_k^j\big)^2 \\ &\le |\varphi|_{W^{2,2}}^2 \sum_{j=0}^{u} 2^{j(2-d/2)} \mathrm{nz}_j (1 - \tau_k^j)\, t_\sigma(j) \alpha \\ &\le \sum_{j=0}^{u} 2^{(6-2\sigma+d/2)j} \mathrm{nz}_j (1 - \tau_k^j) |\varphi|_{W^{2,2}}^{2(j+1)} \alpha \\ &:= C_4(\alpha)\alpha.\end{aligned} \tag{3.11}$$

We choose $\epsilon = 1 < \epsilon_0$ for some fixed ϵ_0. Putting (3.9) and (3.11) into (3.10) yields

$$\|\nabla (T_0 - T_\alpha) I_X f\|_{L^2(\Omega)} \le C_5(\alpha)\alpha.$$

Namely, $C_5(\alpha) = C_3(C_2(\alpha) + C_4(\alpha))$ that is a bounded positive nondecreasing simple function with $C_5(0) = 0$.

All the terms considered so far are stable. Last, and most important, we consider the error in gradient due to the presence of noise. By Proposition 3.2.2, if there exist some (j,k) such that $c_k^j \ne 0$ and $\tau_k^j = 1$, we have

$$\|\nabla T_\alpha I_X P_X \eta\|_{L^2(\Omega)} \le 2^{d/2}\, 2^{u_\alpha}\, \|\nabla \varphi\|_{L^2(\Omega)}\, \|\varphi\|_{L^2(\Omega)}^{-1}\, V^{1/2}(\Omega)\, \delta := C_6(\alpha)\, \delta. \tag{3.12}$$

Otherwise $s_{\delta,X} = 0$, we clearly have $\|\nabla T_\alpha I_X P_X \eta\|_{L^2(\Omega)} = 0$ and $C_6(\alpha) = 0$.

The function $C_6(\alpha)$ in (3.12) is a bounded positive nonincreasing simple function. Since $2^u \geq \frac{2\rho}{h_{X,\min}}$ is the requirement of a positive definite kernel, the gradient error in (3.7) will blow up when one takes finer and finer data points if the noise level $\delta > 0$ is fixed and no regularization is applied.

We summarize all results by the following theorem.

THEOREM 3.3. *For any given data (X, Y_η), let s_α be the regularized interpolant obtained by a MSK satisfying Assumption* 2.3 *and regularized by Theorem* 3.1. *There exist a constant C_1, two bounded positive nondecreasing simple functions $C_2^{\nearrow}(\alpha) \geq C_5^{\nearrow}(\alpha)$ such that $C_2^{\nearrow}(0) = 0 = C_5^{\nearrow}(0)$, and a bounded nonnegative nonincreasing simple function $C_6^{\searrow}(\alpha)$ with $C_6^{\searrow}(0) > 0$ such that the errors in regularized interpolant are bounded by*

$$\|f - s_\alpha\|_{L^2(\Omega)} \leq C_1 h_{X,\Omega}^{\sigma} \|f\|_{\mathcal{N}_\Phi} + V^{1/2}(\Omega)\, \delta + C_2^{\nearrow}(\alpha)\alpha, \tag{3.13}$$

and

$$\|\nabla f - \nabla s_\alpha\|_{L^2(\Omega)} \leq C_1 h_{X,\Omega}^{\sigma-1} \|f\|_{\mathcal{N}_\Phi} + C_5^{\nearrow}(\alpha)\alpha + C_6^{\searrow}(\alpha)\, \delta, \tag{3.14}$$

for all $\alpha \geq 0$. Furthermore, if the noise level $\delta \geq K(f, \sigma)$, there exists a nonzero optimizer α^* *that minimizes the sum of the upper bounds in* (3.13) *and* (3.14).

Proof. For any given data (X, Y_η), the minimizer α^* in the theorem is also a minimizer to the function

$$\big(C_2^{\nearrow}(\alpha) + C_5^{\nearrow}(\alpha)\big)\alpha + C_6^{\searrow}(\alpha)\, \delta. \tag{3.15}$$

By the properties of $C_2^{\nearrow}(\alpha)$ and $C_5^{\nearrow}(\alpha)$, we know that the term $\big(C_2^{\nearrow}(\alpha)+C_5^{\nearrow}(\alpha)\big)\alpha$ is a monotone increasing piecewise linear function. Its jump discontinuities are governed by the term $\mathrm{nz}_j(1 - \tau_k^j(\alpha))$.

The terms $C_6^{\searrow}(\alpha)\delta$ is a nonnegative nonincreasing simple function having jump discontinuities at $0 =: \alpha_{u+1} < \alpha_u \leq \cdots \leq \alpha_0 < \infty$, where α_j is the infimum over α such that jth level is completely truncated, i.e., for all $0 \leq j \leq u$

$$\alpha_j := \inf\{\, \alpha \,\big|\, r_k^j(\alpha) = \tau_k^j c_k^j = 0 \ \text{ for all } \ k \in \mathbb{Z}^d\}.$$

Define $\Delta_k G(\alpha) = G(\alpha_{u-k}) - G(\alpha_{u-k+1})$ for all $0 \leq k < u$. If, for sufficiently large δ, the accumulated drop due to term $C_6^{\searrow}(\alpha)\delta$ is larger than the accumulated growth due to the term $\big(C_2^{\nearrow}(\alpha) + C_5^{\nearrow}(\alpha)\big)\alpha$, i.e.,

$$\delta > K(f, \sigma) := \min_{0 \leq j < u} \left\{ \sum_{k=1}^{j} \Delta_k \frac{(C_2^{\nearrow}(\alpha) + C_5^{\nearrow}(\alpha))\alpha}{C_6^{\searrow}(\alpha)} \right\}, \tag{3.16}$$

then an optimizer $\alpha^* > 0$ exists. ∎

To end this section, note that the constant term $K(f, \sigma)$ in (3.16) decreases as σ increases. If the unknown function f is sufficiently smooth with respect to the noise level δ, our MSK scheme is able to regularize the interpolant. Consider $\delta < K(f, \sigma)$. These cases correspond to small noise levels that are negligible to our regularization technique. As shown in section 5 when $\delta = 0$, while $\alpha^* = 0$ is the theoretical optimizer to (3.15), we would numerically obtain an approximation α_{LC} to α^* such that $0 < \alpha_{LC} < \epsilon_{mach}$ (machine epsilon). In these cases, we set the approximation $\alpha_{LC} = \varepsilon_{mach}$ to filter out extremely small frame coefficients for efficiency.

4. L-curve method. The theoretical existence of α^* does not help us pinpoint its whereabouts. Choosing an optimal α^*, or an approximation α_{LC}, is a separate topic that will be considered in this section.

The L-curve (LC) method was investigated by Hansen and O'Leary [16] to regularize ill-posed systems under different values of the regularization parameter α. The knowledge of the noise level δ is *not necessary*. Vogel [32] shows that the L-curve regularization parameter selection method may fail to converge for a certain class of problems. In our numerical experiments, however, we find that the LC method provides a stable algorithm to find the regularization parameter α.

Our version of the LC method is derived from simplifying both measures in (3.1) and (3.2) for the ease of computation. First, we order the frame coefficients c_k^j by defining an ordered set,

$$\{(\xi_\ell,\, \eta_\ell)\}_{\ell=1}^{\mathrm{nz}(c_k^j)} = \left\{ \left(\left\| c_k^j \varphi(2^j - k) \right\|_{L^2(\Omega)}^2,\, R\big(c_k^j \varphi(2^j - k)\big) \right) \,:\, c_k^j \neq 0 \right\}_{0 \le j \le u,\, k \in \mathbb{Z}^d}$$

such that η_ℓ/ξ_ℓ forms a monotone nondecreasing sequence where $\mathrm{nz}(\cdot)$ returns the number of nonzero elements in the set and $R(\cdot)$ is the roughness measure in (3.2). Then we compute a finite set of points in $\mathbb{R}^2$ by

$$L = \left\{ \left(\|s_{\delta,X}\|_{\Phi_\sigma}^2 - \sum_{\ell=0}^{p} \xi_\ell,\, \sum_{\ell=0}^{p} \eta_\ell \right) \subset \mathbb{R}^2,\ p = 0, 1, \ldots, \mathrm{nz}(c_k^j) \right\},$$

which is known as the *L-curve*.

A suitable regularization parameter α_{LC} is the one near the *corner* on a log-log scale of the L-curve [15]. In numerical computation, finite difference schemes are applied to (the log-values of) these discrete points in order to approximate the curvature of the L-curve. The point with maximum curvature will be labeled as the corner of the L-curve. For numerical efficiency, we impose an extra condition that

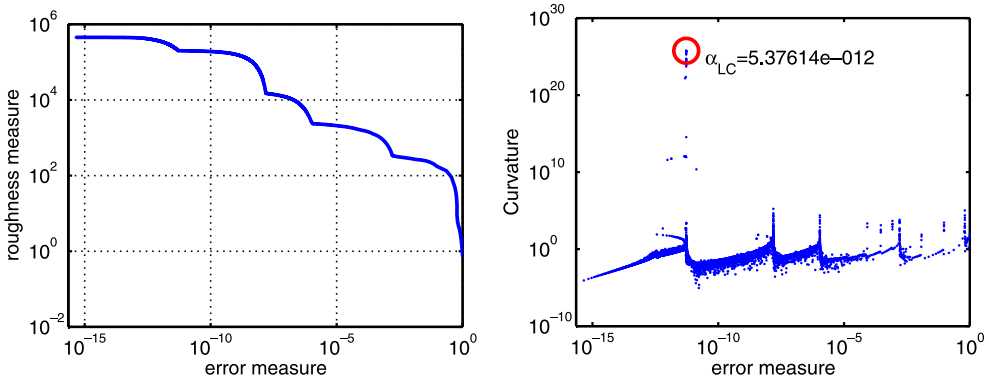
$$\alpha_{LC} \ge \epsilon_{mach}.$$

We show some results with the L-curve method in Figure 4.1. The L-curve is shown in Figure 4.1(a) with a corner at $\alpha_{LC} = 5.3761\text{e-}12$. This value is chosen from the curvature of the L-curve, see Figure 4.1(b).

The number of nonzero frame coefficients in the regularized interpolant s_α is 1735 and 520 for $\alpha = \epsilon_{mach}$ and $\alpha = \alpha_{LC}$, respectively. Figure 4.2(a) for ϵ_{mach} and Figure 4.2(b) for α_{LC} show all $|c_k^j|$ and label the selected r_k^j in boldface dots. All c_k^j are ordered by levels, from level-0 on the left to level-u on the right. In both cases, only the c_k^j in the lower few levels with large absolute values are chosen.

TABLE 4.1
MSK(3, 3) *frame coefficients among all levels on a* 41×41 *uniform grids for section* 5.1.

Level-j	0	1	2	3	4	5	6	7
$\lvert c_k^j\rvert > 0$	64	144	400	1296	4624	17424	26896	26896
$\lvert c_k^j\rvert > \epsilon_{mach}$	56	121	361	1225	4489	17161	24025	24025
$\lvert r_k^j\rvert > 0$ by α_{LC}	49	121	350	0	0	0	0	0
$\lvert r_k^j\rvert > 0$ by ϵ_{mach}	49	121	361	1204	0	0	0	0



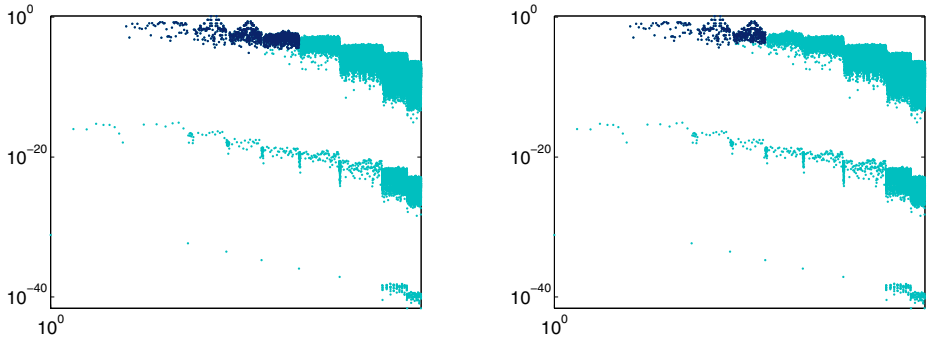
(a) L-curve in log-log scale.

(b) Corner of L-curve.

FIG. 4.1. *L-curve method applied to MSK*$(3,3)$ *in section* 5.1 *with* $\delta = 1.018\times10^{-3}$ *on a* 41×41 *uniform grids.*

At first glance, the computation of *all* nonzero c_k^j may look tremendous. In fact, we are showing all 77744 nonzero frame coefficients in Figure 4.2 but some are extremely small, e.g., 2.4e-42. If we are only interested in frame coefficients whose sizes are larger than machine epsilon, we are looking at 71463 coefficients. The distribution of the frame coefficients among all levels are in Table 4.1. After regularization, the maximum evels appears in $\{r_k^j\}$ are l$u_\alpha = 2$ for $\alpha = \alpha_{LC}$ and $u_\alpha = 3$ for $\alpha = \epsilon_{mach}$; readers may already see how this can be computed efficiently.

Our L-curve only makes use of the local property of each function $c_k^j\varphi(2^j\cdot - k)$. Pretruncation does not affect the final outcome. One could pick an intermediate value $0<\upsilon<u$ and compute frame coefficients up to level-υ only. A safeguard of this approach is that the maximum level appearing in the regularized interpolant should be strictly less than υ. If this is not the case, one can compute the frame coefficients for level-$(\upsilon+1)$ and reapply the LC method.



(a) 1735 frame coefficients for $\alpha = \epsilon_{mach}$.

(b) 520 frame coefficients for $\alpha = \alpha_{LC}$.

FIG. 4.2. *Selected frame coefficients* $\{r_k^j\}\subset\{c_k^j\}$ *corresponds to Figure* 4.1.

5. Numerical comparison and demonstration. We demonstrate some bivariate examples in this section. All codes are written in MATLAB. Random noise is generated by the built-in routine RAND with STATE reset to 0. Generated random numbers are scaled to $[-1, 1]$ and multiplied by the noise level δ. For the problem in $\mathbb{R}^2$, tested values for σ are 2 or 3, see Assumption 2.3. The MSK Φ_σ in (2.3) is constructed with the univariate B-spline of order m defined on the knot sequence $[0, 1, \ldots, m]$, denoted by b_m, see [9],

$$\varphi(x, y) = b_m(x)\, b_m(y) \;\text{ such that }\; x, y \in \mathbb{R},\; m = \{3, 4\},$$

that fulfills all assumptions in the previous discussion. Values of σ and m are specified by the notation $MSK(m, \sigma)$ throughout the section.

5.1. Comparison with TPS-based method. The recent work of Wei, Hon, and Wang [34] uses the thin plate spline (TPS) to compute numerical derivatives. The presented TPS-based method requires triangular partitions of data points; the authors claim that the method can become truly mesh-free with additional assumptions. Two regularization parameters are studied in the same paper: $\alpha_1 = \delta^2$ obtained by *a priori* rule and $\alpha_2(\delta)$ obtained by Morozov's discrepancy principle. We denote them by *TPS-AP* and *TPS-DP*, respectively, hereafter. TPS-DP is reported to be the more effective and stable method between the two.

The clear advantages of MSK with L-curve are that it is already in a truly mesh-free setting for any dimension and it does not require any *a priori* knowledge about the noise level δ. Moreover, resultant linear systems of MSK in (2.5) are sparse. To make the comparison as fair as possible, we compare the accuracies of all methods on uniformly distributed grids among many given examples in their papers. Please be reminded that there are still some differences between the problem settings here and in [34].

Let $\Omega = [-2, 2]^2$. The noise levels are chosen to be the reported $\delta = 1.018\text{e-}3$ and $\delta = 1.020\text{e-}2$. The unknown function to be approximated is given by

$$f(x, y) = \sin(\pi x)\sin(\pi y)\exp(-x^2 - y^2), \quad (x, y) \in \mathbb{R}^2,$$

with $\|f\|_{L^2(\Omega)} \approx 0.387$ and $\|\nabla f\|_{L^2(\Omega)} \approx 4.235$. Since the number of evaluation points is not reported in [34], we use the same *root mean square* (RMS) errors on a 100×100 uniformly distributed grids $x_i' \in \Omega$ to measure accuracy for interpolation,

$$\varepsilon(s_\alpha) = \frac{1}{100}\left(\sum_{i=1}^{100^2}(s_\alpha(x_i') - f(x_i'))^2\right)^{1/2},$$

and for gradient approximation,

$$\varepsilon(\nabla s_\alpha) = \frac{1}{100}\left(\sum_{i=1}^{100^2}\|\nabla s_\alpha(x_i') - \nabla f(x_i')\|_{\ell^2}^2\right)^{1/2}.$$

Table 5.1 shows the RMS errors for both tested noise levels on a 21×21 uniform grids. The differences in error should not be overinterpreted as they are influenced by the regularization parameter α_{LC} and the noise function η. It is more important to note that all choices of m and σ result in the same order of accuracy. Under this

TABLE 5.1
Comparison to TPS-based methods on a 21×21 *uniform grid with different noise levels.*

	$\delta = 1.018 \times 10^{-3}$			$\delta = 1.020 \times 10^{-2}$		
Method	$\varepsilon(s_\alpha)$	$\varepsilon(\nabla s_\alpha)$	α_{LC}	$\varepsilon(s_\alpha)$	$\varepsilon(\nabla s_\alpha)$	α_{LC}
TPS-AP	0.0028	0.0195	–	0.0699	0.3736	–
TPS-DP	0.0019	0.0157	–	0.0100	0.0659	–
MSK(3,2)	0.0011	0.0072	1.5543e-11	0.0042	0.0310	4.4899e-11
MSK(3,3)	0.0010	0.0075	9.1833e-12	0.0040	0.0260	5.1559e-13
MSK(4,2)	0.0014	0.0071	1.0479e-10	0.0042	0.0300	8.4749e-11
MSK(4,3)	0.0009	0.0048	7.4298e-11	0.0039	0.0242	7.8693e-13

TABLE 5.2
MSK$(3,2)$ *RMS errors and* α_{LC} *on* 1609 *unstructured data point with different noise levels.*

δ	α_{LC}	$\mathrm{nz}(r_k^j)$	$\varepsilon(s_\alpha)$	$\varepsilon(\nabla s_\alpha)$
0	$\alpha = 0$	100921	8.5518e-5	1.5045e-3
0	2.2204e-16	6081	1.0032e-4	1.2479e-3
1e-5	2.2204e-16	6076	1.0066e-4	1.2511e-3
1e-4	2.2204e-16	6158	1.1065e-4	1.4226e-3
1e-3	2.1649e-13	1800	4.8194e-4	4.8393e-3
1e-2	2.2794e-11	1678	3.4443e-3	3.8510e-2
1e-1	3.0885e-10	1633	3.4145e-2	3.8377e-1

point density, MSK shows competitive results and seems to outperform TPS.

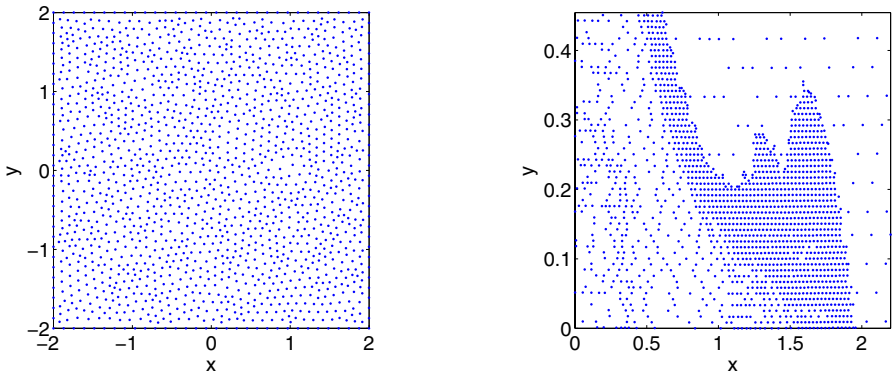
For 1609 unstructured data points, see Figure 5.1(a), with minimum separation distance $h_{X,\min} = 5.092\text{e-}2$ and fill distance $h_{X,\Omega} = 1.317\text{e-}1$. We apply MSK(3,2) to various noise levels. Results are listed in Table 5.2 and graphically demonstrated in Figure 5.2. All regularization parameters are chosen by the LC method except the first row of Table 5.2: $\alpha = 0$ indicates the result of the full interpolant without regularization. Our algorithm runs in the same way as if the data points were structured. The number of selected frame coefficients is listed under the column of $\mathrm{nz}(r_k^j)$ in the table.

Comparing the two noise-free results in Table 5.2, the interpolation error when $\alpha = 0$ is the smallest since the regularization error no longer exists. On the other hand, due to the presence of rounding errors, the regularized interpolant gives better approximation to the gradient than the unregularized full interpolant. In fact, this is true up to $\delta = 1\text{e-}4$. When $\delta \geq 1\text{e-}3$, we have $\alpha_{LC} > \epsilon_{mach}$ and our regularization technique is functioning in these examples; see Theorem 3.3. Overall, the error profile is extremely similar to the TPS-DP, see [34, Figure 5]. The monotonic trend shown in α_{LC} suggests that the proposed LC method is capable of balancing the increasing noise with an increasing regularization parameter.

Our MSK scheme performs equally well when the noise function η is smooth.[1] For completion, MSK(3,2) results in $\varepsilon(s_\alpha) = 0.0025$ and $\varepsilon(\nabla s_\alpha) = 0.0046$ on a 41×41 uniformly distributed grid. Whereas, TPS-DP results in $\varepsilon(s_\alpha) = 0.0035$ and $\varepsilon(\nabla s_\alpha) = 0.0159$.

5.2. Derivative of a landscape data. We demonstrate another example with a set of landscape data [11]; see Figure 5.1(b). The data set, containing 1669 data points, is processed by MSK(3,2) and MSK(3,3) in order to estimate its derivatives.

[1] $\eta(x,y) = 0.005 \sin\left(\frac{1}{2}\pi x\right) \sin\left(\frac{1}{2}\pi y\right)$, see [34, Table 1].



(a) 1609 unstructured points.

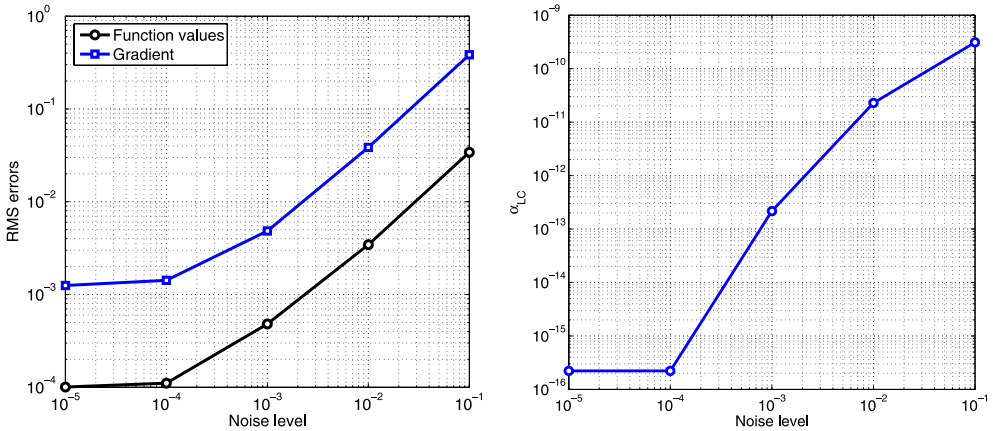
(b) Landscape data points.

FIG. 5.1. *Data points distribution for examples in sections* 5.1 *and* 5.2.

Unlike the previous example, data points are unevenly distributed and there is no exact solution for this example. Hence, the full interpolant $s_{\delta,X}$ will be used for comparison. We only demonstrate the x-derivatives; results for the y-derivatives are similar and are omitted here.

The full interpolant $s_{\delta,X}$ and its x-derivative are shown in Figure 5.3. As we see in section 3.1, the presence of noise does not introduce instability to the interpolation problem. On the other hand, we observe serious oscillations in the derivatives of the full interpolant; see Figure 5.3(b).

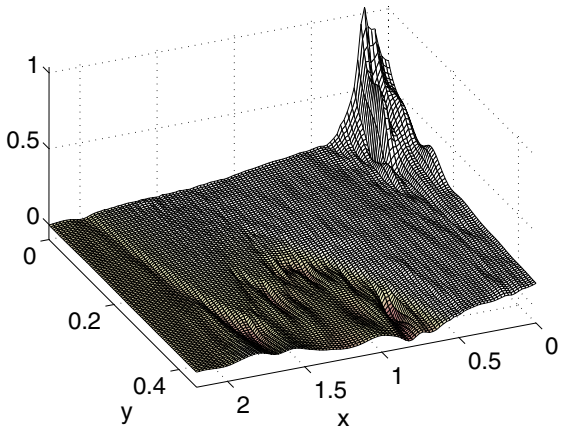
The MSK(3,2) regularized interpolants with $\alpha_{LC} = 5.0626\text{e-}14$ (566 nonzero frame coefficients) are shown in Figure 5.4. The regularized interpolant in Figure 5.4 is very similar to Figure 5.3 but with less local structures. The derivative of the regularized interpolant in Figure 5.4(b) clearly reveal the local features of the landscape.



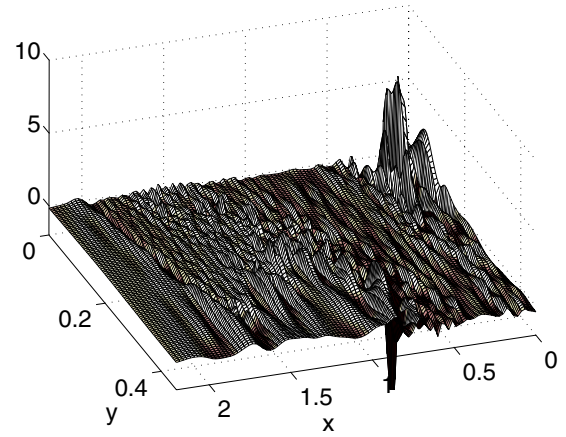
(a) RMS errors.

(b) α_{LC}.

FIG. 5.2. *RMS and* α_{LC} *errors as functions of the noise level* δ.



(a) Full interpolant $s_{\delta,X}$.

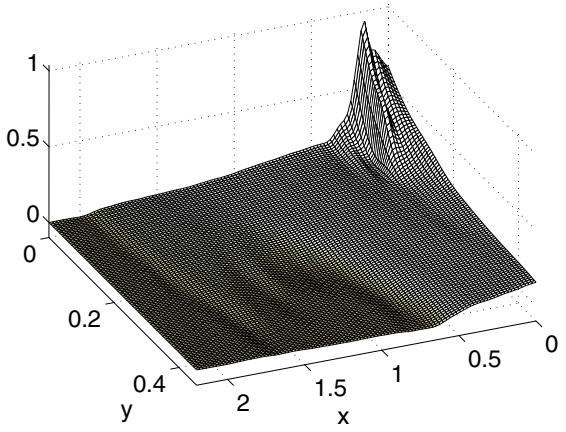


(b) x-derivatives.

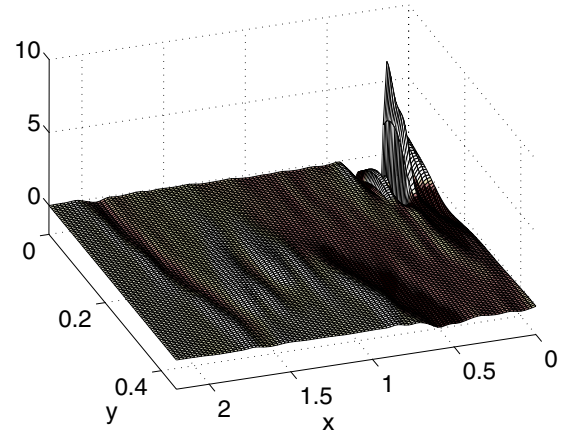
FIG. 5.3. *Full interpolant for the landscape data and its x-derivatives.*

The MSK(m,σ) method assumes the unknown function f lies in $\mathcal{N}_\Phi$ and LC regularizes the interpolant accordingly. If σ is too large, the MSK Φ_σ is very smooth and the MSK scheme will over-regularize the interpolant. Fortunately, nothing will become unbounded. To see this, if we can write the unknown function $f \notin \mathcal{N}_\Phi$ as $f = f_1 + f_2$ where $f_1 \in W^{\sigma,2}$ and $f_2 \in L^2(\Omega) \cap C(\Omega)$, then our results in section 3.1 apply consequently. As an example, Figure 5.5 shows the regularized interpolant of MSK(3,3). The regularization parameter is $\alpha_{LC} = 3.8654\text{e-}12$ resulting in 122 frame coefficients. The resulting regularized interpolant in Figure 5.5 is much smoother than that of MSK(3,2) in Figure 5.4. In fact, it seems too smooth for the landscape data.

For rough data from a function $f \notin \mathcal{N}_\Phi$, we shall treat α_{LC} as an upper estimated



(a) Regularized interpolant s_α with 566 frame coefficients.

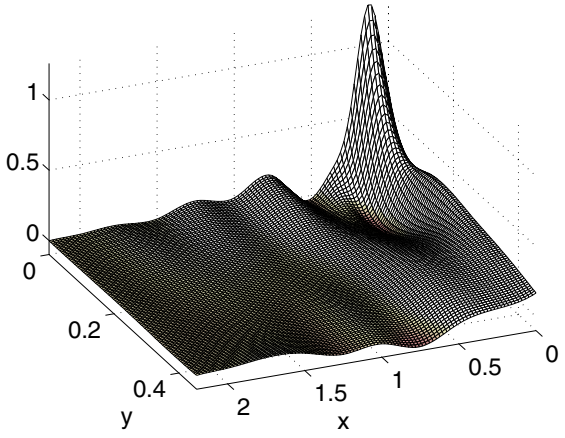


(b) x-derivatives.

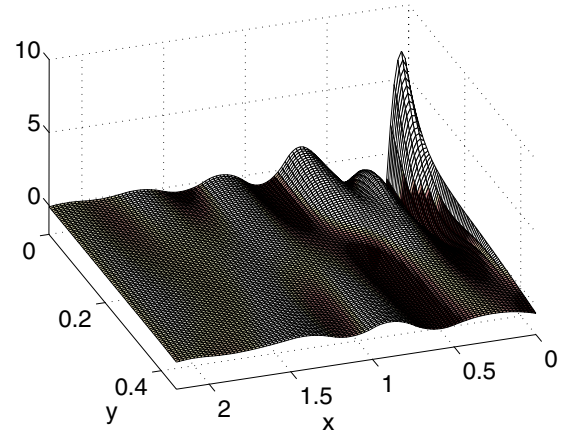
FIG. 5.4. *MSK*(3,2) *regularized interpolant for the landscape data and its x-derivatives.*

parameter. To capture more local features, we could use a regularization parameter $0 < \alpha < \alpha_{LC}$ and obtain results similar to the one from MSK(3,2). The resulting interpolant will contain more local features with any $0 < \alpha < \alpha_{LC}$, while the oscillation in its derivatives are still relatively well behaved. However, we have no robust routine for choosing an optimal regularization parameter in this case.

For unevenly distributed data points, the tolerance to roughness should be proportional to the local density of data points, e.g., a threshold of the form $t_\sigma(j,k)$. Regions with high data point density are expected to have more local features and



(a) Regularized interpolant s_α with 122 frame coefficients.



(b) x-derivatives.

FIG. 5.5. *MSK*(3,3) *regularized interpolant for the landscape data and its x-derivatives.*

higher roughness should therefore be allowed. This allows smooth kernels to capture more local features of the given data set in certain regions. An example of such a density measure is the number of data points in the support of each function $\varphi(2^j \cdot -k)$; the information is already available after computing the frame coefficients. We leave this as an open question for future study.

6. Conclusion. We solve a classical ill-posed numerical differentiation problem by a state-of-the-art matrix-free multiscale kernel based multivariate interpolation method. The theoretical stability for this ill-posed problem is investigated. The

Tikhonov regularization and the LC method are employed to obtain a regularized interpolant. The advantages of the proposed method are (1) the ability to handle problems in higher dimensions; (2) the flexibility to handle real-life, noisy, and multiple-valued data; and (3) the efficiency due to the resultant sparse matrix systems. Numerical examples are given for a bivariate test problem that shows results competitive with the TPS based method and a landscape data set that shows the stability of our scheme even when the unknown function may not be smooth enough for our assumptions.

Acknowledgements. The author would like to thank M. Yamamoto, R. Schaback, R. Opfer, M. R. Trummer, S. Ruuth, and T. Takeuchi for their helpful comments. Moreover, we thank the reviewers for improving the academic quality and readability of this manuscript.

REFERENCES

[1] R. A. Adams, *Sobolev spaces*, Academic Press, Pure and Applied Mathematics, Vol. 65, [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975.
[2] R. S. Anderssen and M. Hegland, *For numerical differentiation, dimensionality can be a blessing*, Math. Comp., 68 (1999), pp. 1121–1141.
[3] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.
[4] R. Baltensperger and M. R. Trummer, *Spectral differencing with a twist*, SIAM J. Sci. Comput., 24 (2003), pp. 1465–1487.
[5] M. Bozzini and M. Rossini, *Numerical differentiation of 2D functions from noisy data*, Comput. Math. Appl., 45 (2003), pp. 309–327.
[6] J. R. Cannon, Y. P. Lin, and S. Xu, *Numerical procedures for the determination of an unknown coefficient in semi-linear parabolic differential equations*, Inverse Problems, 10 (1994), pp. 227–243.
[7] J. Cheng, Y. C. Hon, and Y. B. Wang, *A numerical method for the discontinuous solutions of Abel integral equations*, in Inverse problems and spectral theory, Contemp. Math. 348, AMS, Providence, RI, (2004), pp. 233–243.
[8] J. Cullum, *Numerical differentiation and regularization*, SIAM J. Numer. Anal., 8 (1971), pp. 254–265.
[9] C. de Boor, *A practical guide to splines*, revised ed., Appl. Math. Sci. 27, Springer-Verlag, New York, 2001.
[10] S. R. Deans, *The Radon transform and some of its applications*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1983.
[11] R. Franke, mbay.mat. Available at http://www.math.nps.navy.mil/∼rfranke/.
[12] R. Gorenflo and M. Yamamoto, *Operator-theoretic treatment of linear Abel integral equations of first kind*, Japan J. Indust. Appl. Math., 16 (1999), pp. 137–161.
[13] C. W. Groetsch and O. Scherzer, *Iterative stabilization and edge detection*, in Inverse problems, image analysis, and medical imaging (New Orleans, LA, 2001), Contemp. Math. 313, AMS, Providence, RI, (2002), pp. 129–141.
[14] M. Hanke and O. Scherzer, *Inverse problems light: numerical differentiation*, Amer. Math. Monthly, 108 (2001), pp. 512–521.
[15] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
[16] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.
[17] M. Hegland and R. S. Anderssen, *Resolution enhancement of spectra using differentiation*, Inverse Problems, 21 (2005), pp. 915–934.
[18] B. Hofmann and M. Yamamoto, *Convergence rates for Tikhonov regularization based on range inclusions*, Inverse Problems, 21 (2005), pp. 805–820.
[19] C. Itiki and J. J. Neto, *Complete automation of the generalized inverse method for constrained mechanical systems of particles*, Appl. Math. Comput., 152 (2004), pp. 561–580.
[20] L. Ling, *Multivariate quasi-interpolation schemes for dimension-splitting multiquadric*, Appl. Math. Comput., 161 (2005), pp. 195–209.
[21] H. Meschkowski, *Hilbertsche Räume mit Kernfunktion*, Die Grundlehren der mathematischen

Wissenschaften, Bd. 113, Springer-Verlag, Berlin, 1962.
[22] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comp., 74 (2005), pp. 743–763.
[23] R. Opfer, *Multiscale kernels*, Adv. Comput. Math., 25 (2006), pp. 357–380.
[24] M. Piana, R. Barrett, J. C. Brown, and S. W. McIntosh, *A non-uniqueness problem in solar hard x-ray spectroscopy*, Inverse Problems, 15 (1999), pp. 1469–1486.
[25] D. A. Popov and D. V. Sushko, *Computation of singular convolutions*, in Applied problems of Radon transform, Amer. Math. Soc. Transl. Ser. 2, 162, AMS, Providence, RI, (1994), pp. 43–127.
[26] A. G. Ramm and A. B. Smirnova, *On stable numerical differentiation*, Math. Comp., 70 (2001), pp. 1131–1153.
[27] T. J. Rivlin, *Optimally stable Lagrangian numerical differentiation*, SIAM J. Numer. Anal., 12 (1975), pp. 712–725.
[28] R. Schaback, *Approximation by radial basis functions with finitely many centers*, Constr. Approx., 12 (1996), pp. 331–340.
[29] R. Schaback, *Native Hilbert spaces for radial basis functions I*, in New Developments in Approximation Theory, B. M.D., M. D. H., F. M., Müller, and M.W., eds., vol. 132 of Internat. Ser. Numer. Math., Birkhäuser Verlag, Basel, Switzerland, (1999), pp. 255–282.
[30] L. Tang and J. D. Baeder, *Uniformly accurate finite difference schemes for p-refinement*, SIAM J. Sci. Comput., 20 (1999), pp. 1115–1131.
[31] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, Preface by translation editor Fritz John, Scripta Series in Mathematics, V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from Russian.
[32] C. R. Vogel, *Non-convergence of the L-curve regularization parameter selection method*, Inverse Problems, 12 (1996), pp. 535–547.
[33] Y. B. Wang, X. Z. Jia, and J. Cheng, *A numerical differentiation method and its application to reconstruction of discontinuity*, Inverse Problems, 18 (2002), pp. 1461–1476.
[34] T. Wei, Y. C. Hon, and Y. B. Wang, *Reconstruction of numerical derivatives from scattered noisy data*, Inverse Problems, 21 (2005), pp. 657–672.
[35] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2005.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1801–1828

A HAMILTONIAN-PRESERVING SCHEME FOR THE LIOUVILLE EQUATION OF GEOMETRICAL OPTICS WITH PARTIAL TRANSMISSIONS AND REFLECTIONS*

SHI JIN† AND XIN WEN‡

Abstract. We construct a class of Hamiltonian-preserving numerical schemes for the Liouville equation of geometrical optics, with partial transmissions and reflections. This equation arises in the high frequency limit of the linear wave equation, with a discontinuous index of refraction. In our previous work [*Hamiltonian-preserving schemes for the Liouville equation of geometrical optics with discontinuous local wave speeds*, J. Comput. Phys. 214 (2006), pp. 672–697], we introduced the Hamiltonian-preserving schemes for the same equation when only complete transmissions or reflections occur at the interfaces. These schemes are extended in this paper to the general case of partial transmissions and reflections. The key idea is to build into the numerical flux the behavior of waves at the interface, namely, partial transmissions and reflections that satisfy Snell's law of refraction with the correct transmission and reflection coefficients. This scheme allows a hyperbolic stability condition, under which positivity, and stabilities in both l^1 and l^∞ norms, are established. Numerical experiments are carried out to study the numerical accuracy.

Key words. geometrical optics, Liouville equation, transmission and reflection, Hamiltonian-preserving schemes

AMS subject classifications. 35L45, 65M06, 70H99

DOI. 10.1137/050631343

1. Introduction.

In this paper, we construct and study a numerical scheme for the Liouville equation in d-dimension:

$$f_t + H_{\mathbf{v}} \cdot \nabla_{\mathbf{x}} f - H_{\mathbf{x}} \cdot \nabla_{\mathbf{v}} f = 0\,, \quad t > 0, \quad \mathbf{x}, \mathbf{v} \in R^d, \tag{1.1}$$

where the Hamiltonian H possesses the form

$$H(\mathbf{x}, \mathbf{v}) = c(\mathbf{x})|\mathbf{v}| = c(\mathbf{x})\sqrt{v_1^2 + v_2^2 + \cdots + v_d^2} \tag{1.2}$$

with $c(\mathbf{x})$ being the local wave speed of the medium ($1/c(\mathbf{x})$ is the index of refraction); $f(t, \mathbf{x}, \mathbf{v})$ is the density distribution of particles depending on position $\mathbf{x}$, time t, and the slowness vector $\mathbf{v}$. We are concerned with the case when $c(\mathbf{x}) \in W^{1,\infty}$ with isolated *discontinuities* due to different media. The discontinuity in c corresponds to an *interface*, and as a consequence waves crossing this interface will undergo transmissions and reflections.

*Received by the editors May 11, 2005; accepted for publication (in revised form) March 31, 2006; published electronically September 29, 2006. This research was supported in part by NSF grant DMS-0305080, NSFC under project 10228101, the Basic Research Projects of Tsinghua University under Project JC2002010, and the Knowledge Innovation Project of the Chinese Academy of Sciences K3502012D1 and K5502212F1.

http://www.siam.org/journals/sinum/44-5/63134.html

†Department of Mathematics, University of Wisconsin, Madison, WI 53706 and Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P.R. China (jin@math.wisc.edu). This author's research was also supported in part by the Institute for Mathematics and its Applications (IMA) under a New Direction Visiting Professorship.

‡Institute of Computational Mathematics, Chinese Academy of Science, P. O. Box 2719, Beijing 100080, China (wenxin@amss.ac.cn).

The bicharacteristics of this Liouville equation (1.1) satisfy the Hamiltonian system

$$\frac{d\mathbf{x}}{dt} = c(\mathbf{x})\frac{\mathbf{v}}{|\mathbf{v}|}, \quad \frac{d\mathbf{v}}{dt} = -c_{\mathbf{x}}|\mathbf{v}|. \tag{1.3}$$

In classical mechanics the Hamiltonian (1.2) of a particle remains a constant along the particle trajectory, even when it is being transmitted or reflected by the interface.

This Liouville equation arises in the phase space description of geometrical optics. It is the high frequency limit of the wave equation

$$u_{tt} - c(\mathbf{x})^2\Delta u = 0, \quad t > 0, \quad \mathbf{x} \in R^d. \tag{1.4}$$

In the past, numerous numerical methods have been proposed for the wave equation (1.4) with discontinuous coefficients c; see [32] and references therein. However, our interest is in the high frequency waves, for which many current numerical methods such as the phase space based level set methods, are based on the Liouville equation (1.1) with smooth c; see [18, 25, 34]. Semiclassical limits of wave equations with transmissions and reflections at the interfaces were studied in [1, 33, 39]. A Liouville equation based level set method for the wave front, but with only reflection, was introduced in [9].

In our previous work [28] two classes of numerical schemes that are suitable for the Liouville equation (1.1) with a discontinuous local wave speed $c(\mathbf{x})$ were constructed. The designing principle there was to build the behavior of waves at the interface—either cross over with a changed velocity according to a constant Hamiltonian, or be reflected with a negative velocity (or momentum)—into the numerical flux; see also earlier works [36, 27]. These schemes were called *Hamiltonian-preserving schemes*. By providing an interface condition, it connects the two domains of Liouville equation with smooth coefficients. This gives a physically relevant selection criterion for a unique solution to the governing equation, which is linearly hyperbolic with singular (discontinuous or measure-valued) coefficients. For a plane wave hitting a flat interface, it selects the solution at the interface governed by *Snell's law of refraction* when the wave length is much shorter than the width of the interface while both lengths go to zero. Nevertheless, this is not the only physically relevant possibility to choose a solution across the interface. When the wave length is much longer than the width of the interface, while both lengths go to zero, the waves can be *partially* transmitted and reflected, and the transmission and reflection coefficients can be analytically computed [33].

The goal of this paper is to construct the numerical scheme which is suitable to deal with partial transmissions and reflections, with computable transmission and reflection coefficients. As in [28], we still use the *Hamiltonian-preserving* principle to determine the transmitted velocity across the interface. The new contribution of this paper is to incorporate the transmission and reflection coefficients into the numerical flux, in order to treat partial transmissions and reflections. This new, explicit scheme, like those in [27, 28], allows a typical hyperbolic stability condition $\Delta t = O(\Delta x, \Delta v)$, under which we also establish the positivity, and l^1 and l^∞ stability theory for the scheme.

In geometrical optics applications, one has to solve the Liouville equation like (1.1) with *measure-valued* initial data

$$f(0, \mathbf{x}, \mathbf{v}) = \rho_0(\mathbf{x})\delta(\mathbf{v} - \mathbf{u}_0(\mathbf{x}));\tag{1.5}$$

see, for example, [38, 14, 25]. The solution at later time remains measure-valued (with finite or even infinite number of concentrations-corresponding to *multivalued* solutions in the physical space). Computation of multivalued solutions in geometrical optics and more generally in nonlinear PDEs has been a very active area of recent research; see [3, 4, 6, 5, 10, 17, 12, 13, 15, 19, 20, 21, 18, 26, 34, 37, 41].

Direct numerical methods (DNM) for the Liouville equation with measure-valued initial data (1.5), which approximate the initial delta function first, then evolve the Liouville equation, could suffer from a poor numerical resolution due to the numerical approximation of the initial data of delta function as well as numerical dissipation [24]. The level set method proposed in [24, 25] decomposes the density distribution f into the bounded level set functions obeying the same Liouville equation, which greatly enhances the numerical resolution. One only involves numerically the delta function at the output time when the moments—which has delta functions in their integrands—need to be evaluated numerically.

However, the extension of this density distribution decomposing approach to the case of partial transmission and reflection is not straightforward. In particular, as the number of transmissions and reflections increase in time, so does the number of needed level set functions satisfying (1.1). This difficulty was already pointed out in [9]. In this paper, when dealing with the measure-valued initial data (1.5) we will just use the DNM. This does not offer the same resolution as those in [28]. It remains an open question on how to extend the decomposition idea of [24, 25] to the case of partial transmissions and reflections.

This paper is organized as follows. In section 2, we present the behavior of waves at an interface, which guides the designing of our scheme. We also give an interface condition (2.5) which allows us to define the analytic solution to the Liouville equation (1.1) with singular coefficients. We present the scheme in 1d in section 3 and study its positivity and stability in both l^∞ and l^1 norms. We extend the scheme to the two space dimension in section 4 in the simple case of an interface aligning with the grids. Numerical examples are given in section 5 to verify the accuracy of the scheme. We make some concluding remarks in section 6.

2. The behavior of waves at an interface.

2.1. Transmissions and reflections at the interface. In geometrical optics, when a wave moves with its density distribution governed by the Liouville equation (1.1), *its Hamiltonian* $H = c|\mathbf{v}|$ *should be preserved across the interface*

$$c^+|\mathbf{v}^+| = c^-|\mathbf{v}^-|, \tag{2.1}$$

where the superscripts $\pm$ indicate the right and left limits of the quantity at the interface. The wave can be partly reflected and partly transmitted. The condition (2.1) can be used to determine the particle velocity on one side of the interface from its value on the other side. When a plane wave hits a flat interface, this condition is equivalent to Snell's law of refraction [28]:

$$\frac{\sin\theta_i}{c^-} = \frac{\sin\theta_t}{c^+} \tag{2.2}$$

and the reflection law

$$\theta_r = \theta_i, \tag{2.3}$$

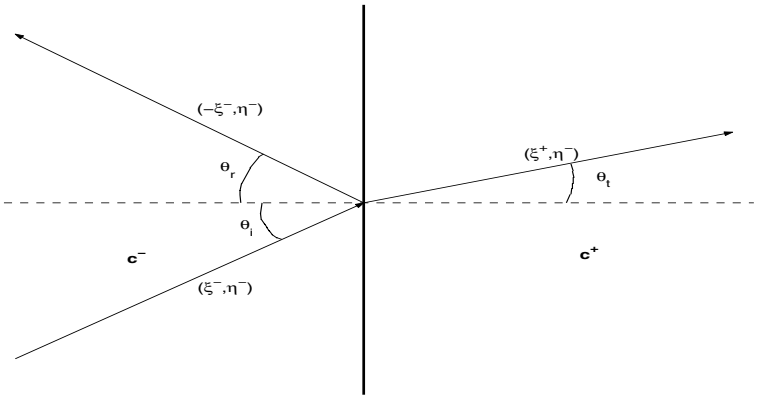


FIG. 2.1. *Wave transmission and reflection at an interface.*

where θ_i, θ_t, and θ_r stand for angles of incident and transmitted and reflected waves; see Figure 2.1. The reflection coefficient is given by

$$\alpha^R = \left(\frac{c^+\cos\theta_i - c^-\cos\theta_t}{c^+\cos\theta_i + c^-\cos\theta_t}\right)^2 \tag{2.4}$$

while the transmission coefficient is $\alpha^T = 1 - \alpha^R$; see, for example, [1, 33, 39].

We will discuss this behavior in more detail in 1D and 2D, respectively.

- The 1D case is simpler. Consider the case when, at an interface, the characteristic on the left of the interface is given by $\xi^- > 0$. Then with probability $\alpha^R = \left(\frac{c^+ - c^-}{c^+ + c^-}\right)^2$, the wave is reflected by the interface with a new velocity $-\xi^-$, and with probability $\alpha^T = 1 - \alpha^R$ it will cross the interface with the new velocity $\xi^+ = \frac{c^-}{c^+}\xi^-$ determined by (2.1).
- The 2D case, when an incident wave hits a vertical interface (see Figure 2.1). Let $\mathbf{x} = (x, y), \mathbf{v} = (\xi, \eta)$. Assume that the incident wave has a velocity (ξ^-, η^-) to the left side of the interface, with $\xi^- > 0$. Since the interface is vertical, (1.3) implies that η is not changed when the wave crosses the interface. There are two possibilities:
  1) $\left(\frac{c^-}{c^+}\right)^2 (\xi^-)^2 + \left[\left(\frac{c^-}{c^+}\right)^2 - 1\right](\eta^-)^2 > 0$. In this case the wave can partially transmit and partially be reflected. With probability $\alpha^R = \left(\frac{c^+\gamma^- - c^-\gamma^+}{c^+\gamma^- + c^-\gamma^+}\right)^2$ the wave is reflected with a new velocity $(-\xi^-, \eta^-)$, where
  $$\gamma^+ = \cos(\theta_t) = \frac{\xi^+}{\sqrt{(\xi^+)^2 + (\eta^-)^2}}, \quad \gamma^- = \cos(\theta_i) = \frac{\xi^-}{\sqrt{(\xi^-)^2 + (\eta^-)^2}}.$$
  With probability $\alpha^T = 1 - \alpha^R$ it will be transmitted with the new velocity (ξ^+, η^-), where
  $$\xi^+ = \sqrt{\left(\frac{c^-}{c^+}\right)^2 (\xi^-)^2 + \left[\left(\frac{c^-}{c^+}\right)^2 - 1\right](\eta^-)^2},$$

is obtained using (2.1).

2) $c^- < c^+$ and $\left(\frac{c^-}{c^+}\right)^2 (\xi^-)^2 + \left[\left(\frac{c^-}{c^+}\right)^2 - 1\right](\eta^-)^2 < 0$. In this case, it is impossible for the wave to transmit, so the wave will be completely reflected with velocity $(-\xi^-, \eta^-)$.

If $\xi^- < 0$, similar behavior can also be analyzed using the constant Hamiltonian condition (2.1).

2.2. The interface condition for density distribution. The solution to the Liouville equation (1.1), which is linearly hyperbolic, can be solved by the method of characteristics. Namely, the density distribution f remains a constant along a bicharacteristic. However, with partial transmissions and reflections, this is no longer true, since f needs to be determined from two bicharacteristics, one accounting for the transmission and the other for reflection. Therefore, we use the following condition at the interface:

$$f(t, \mathbf{x}^+, \mathbf{v}^+) = \alpha^T f(t, \mathbf{x}^-, \mathbf{v}^-) + \alpha^R f(t, \mathbf{x}^+, -\mathbf{v}^+), \tag{2.5}$$

where $\mathbf{v}^-$ is defined from $\mathbf{v}^+$ through the constant Hamiltonian condition (2.1), α^T and α^R are the transmission and reflection coefficients which add up to 1 and vary with $\mathbf{v}^+$ except in the 1D case. *This is the main idea of this paper*, and will be used in constructing the numerical flux across the interface in the next section. As will be seen in the next section, our scheme incorporates the interface condition into the numerical flux.

For hyperbolic systems with discontinuous coefficients, renormalized solution was introduced by DiPerna and Lions [11], and further extended in [7, 8, 22, 23] for uniqueness and stability. The renormalized solution idea cannot be applied here since the coefficients can be measure-valued. Our approach here is to use the interface condition (2.5) to connect two domains in which the Liouville equation has smooth Hamiltonians. Concretely, we define the solution for (1.1) when the local wave speed has discontinuities as follows.

DEFINITION 2.1. *The analytic solution for the Liouville equation* (1.1) *when the local wave speed c has discontinuities is constructed by method of characteristics away from the interface plus the interface condition* (2.5).

Below we justify the well-posedness of the initial value problem, for the simple case of a step function c with a vertical interface. The more general situation remains to be worked out and will be deferred to a future work.

Consider the simple case that the local wave speed $c(\mathbf{x}), \mathbf{x} \in R^d$ is piecewise constant as follows:

$$c(\mathbf{x}) = \begin{cases} c^- & x_1 < 0 \\ c^+ & x_1 > 0, \end{cases} \tag{2.6}$$

where we assume $c^- < c^+$. We will also exclude some singular points, working in the domain defined by

$$\Omega = \left\{(\mathbf{x}, \mathbf{v}) | \mathbf{x} \in R^d, \mathbf{v} \in R^d \backslash \{\mathbf{0}\}\right\} \backslash \{(\mathbf{x}, \mathbf{v}) | x_1 = v_1 = 0\}. \tag{2.7}$$

We have the following theorem.

THEOREM 2.1. *Assume the initial data $f(0, \mathbf{x}, \mathbf{v})$ has a compact support in $\mathbf{v}$. With the solution defined in Definition* 2.1*, the initial value problem to*

$$f_t + H_{\mathbf{v}} \cdot \nabla_{\mathbf{x}} f - H_{\mathbf{x}} \cdot \nabla_{\mathbf{v}} f = 0\,, \quad t > 0, \quad (\mathbf{x}, \mathbf{v}) \in \Omega\,, \tag{2.8}$$

with H given by (1.2)*, c given by* (2.6)*, and Ω given by* (2.7)*, is well-posed in l^∞ and l^1 norms.*

Proof. The proof is based on explicit construction of the analytical solution $f(T,\mathbf{x},\mathbf{v})$. The l^∞ stability follows from the maximum principle, while the key for the l^1 stability is to prove that *the Liouville theorem* (*volume preserving for a Hamiltonian flow*) *holds at the interface for partial tranmissions and reflections.*

To make the following description easier, we define a function extended from the local wave speed (2.6)

$$\widetilde{c}(\mathbf{x},\mathbf{v}) = \begin{cases} c^- & x_1 < 0 \\ c^+ & x_1 > 0 \\ c^- & x_1 = 0,\ v_1 < 0 \\ c^+ & x_1 = 0,\ v_1 > 0, \end{cases} \tag{2.9}$$

which is defined on the whole definition domain Ω. The values of $\widetilde{c}(\mathbf{x},\mathbf{v})$ on $x_1 = 0$, however, are not crucial as long as they are positive.

Split the domain Ω into two parts $\Omega = \Omega_1 \cup \Omega_2$ with

$$\Omega_1 = \left\{ (\mathbf{x},\mathbf{v}) \in \Omega \middle| \, x_1 \left(x_1 - \widetilde{c}(\mathbf{x},\mathbf{v}) \frac{v_1}{|\mathbf{v}|} T \right) > 0 \text{ or } \left(x_1 - \widetilde{c}(\mathbf{x},\mathbf{v}) \frac{v_1}{|\mathbf{v}|} T \right) = 0 \right\},$$

$$\Omega_2 = \left\{ (\mathbf{x},\mathbf{v}) \in \Omega \middle| \, x_1 \left(x_1 - \widetilde{c}(\mathbf{x},\mathbf{v}) \frac{v_1}{|\mathbf{v}|} T \right) < 0 \text{ or } x_1 = 0 \right\},$$

where Ω_1 consists of those points whose positions are not on the interface, and when tracing back along the bicharacteristics, will not hit the interface within time T, except possibly the end point. We further split domain Ω_1, Ω_2 as $\Omega_1 = \Omega_1^- \cup \Omega_1^+$, $\Omega_2 = \Omega_2^- \cup \Omega_2^+$ with

$$\begin{aligned} \Omega_1^- &= \{ (\mathbf{x},\mathbf{v}) \in \Omega_1 | \, x_1 < 0 \}, \\ \Omega_1^+ &= \{ (\mathbf{x},\mathbf{v}) \in \Omega_1 | \, x_1 > 0 \}, \\ \Omega_2^- &= \left\{ (\mathbf{x},\mathbf{v}) \in \Omega_2 \middle| \, \left(x_1 - \widetilde{c}(\mathbf{x},\mathbf{v}) \frac{v_1}{|\mathbf{v}|} T \right) > 0 \right\}, \\ \Omega_2^+ &= \left\{ (\mathbf{x},\mathbf{v}) \in \Omega_2 \middle| \, \left(x_1 - \widetilde{c}(\mathbf{x},\mathbf{v}) \frac{v_1}{|\mathbf{v}|} T \right) < 0 \right\}. \end{aligned}$$

For $(\mathbf{x},\mathbf{v}) \in \Omega_1$, one has

$$f(T,\mathbf{x},\mathbf{v}) = f\left(0, \mathbf{x} - c^- \frac{\mathbf{v}}{|\mathbf{v}|} T, \mathbf{v} \right), \quad (\mathbf{x},\mathbf{v}) \in \Omega_1^-, \tag{2.10}$$

$$f(T,\mathbf{x},\mathbf{v}) = f\left(0, \mathbf{x} - c^+ \frac{\mathbf{v}}{|\mathbf{v}|} T, \mathbf{v} \right), \quad (\mathbf{x},\mathbf{v}) \in \Omega_1^+. \tag{2.11}$$

Define a subset of Ω_2^-

$$\Omega_{2,s} = \left\{ (\mathbf{x},\mathbf{v}) \in \Omega_2^- \middle| \, \left(\frac{c^-}{c^+} \right)^2 |\mathbf{v}|^2 \le v_2^2 + \cdots + v_d^2 \right\}.$$

For $(\mathbf{x},\mathbf{v}) \in \Omega_2$, one has

$$\begin{aligned} &f(T,\mathbf{x},\mathbf{v}) = f(0,\mathbf{x_R},\mathbf{v_R}), && (\mathbf{x},\mathbf{v}) \in \Omega_{2,s}, \\ &f(T,\mathbf{x},\mathbf{v}) = \alpha_T(\mathbf{v_T}) f(0,\mathbf{x_T},\mathbf{v_T}) + \alpha_R(\mathbf{v_R}) f(0,\mathbf{x_R},\mathbf{v_R}), && (\mathbf{x},\mathbf{v}) \in \Omega_2 \backslash \Omega_{2,s}, \end{aligned} \tag{2.12}$$

(2.13)

where $\alpha_T(\mathbf{v}), \alpha_R(\mathbf{v})$ denote the transmission and reflection coefficients determined by the incident wave slowness vector $\mathbf{v}$, with condition $\alpha_T(\mathbf{v})+\alpha_R(\mathbf{v}) = 1$. In geometrical optics, the transmission coefficient also satisfies $\alpha_T(\mathbf{v_T}) = \alpha_T(\mathbf{v_R})$ for the slowness vectors $\mathbf{v_T}, \mathbf{v_R}$ appearing in (2.13), thus it holds that $\alpha_T(\mathbf{v_T}) + \alpha_R(\mathbf{v_R}) = 1$. This contributes to the maximum principle of the solution for (1.1). The positions and slowness vectors $\mathbf{x_T}, \mathbf{v_T}, \mathbf{x_R}, \mathbf{v_R}$ can be explicitly expressed by $\mathbf{x}, \mathbf{v}$ as follows

$$v_{T,1}^2 = \left[\frac{\widehat{c}\,|\mathbf{v}|}{\widehat{c}_T}\right]^2 - v_2^2 - \cdots - v_d^2, \quad v_{T,1}v_1 > 0, \tag{2.14}$$

$$v_{T,i} = v_i, \quad i = 2, \ldots, d, \tag{2.15}$$

$$x_{T,1} = \frac{(\widehat{c}_T)^2 v_{T,1}}{|\mathbf{v}|\widehat{c}}\left(\frac{x_1|\mathbf{v}|}{\widehat{c}\,v_1} - T\right), \tag{2.16}$$

$$x_{T,i} = x_i - v_i\frac{x_1}{v_1} + \frac{(\widehat{c}_T)^2 v_i}{|\mathbf{v}|\widehat{c}}\left(\frac{x_1|\mathbf{v}|}{\widehat{c}\,v_1} - T\right), \quad i = 2, \ldots, d, \tag{2.17}$$

$$v_{R,1} = -v_1, \qquad v_{R,i} = v_i, \quad i = 2, \ldots, d, \tag{2.18}$$

$$x_{R,1} = \frac{\widehat{c}\,v_1}{|\mathbf{v}|}\left(T - \frac{x_1|\mathbf{v}|}{\widehat{c}\,v_1}\right), \tag{2.19}$$

$$x_{R,i} = x_i - v_i\frac{x_1}{v_1} - \frac{\widehat{c}\,v_i}{|\mathbf{v}|}\left(T - \frac{x_1|\mathbf{v}|}{\widehat{c}\,v_1}\right), \quad i = 2, \ldots, d, \tag{2.20}$$

where $\widehat{c}, \widehat{c}_T$ are given by

$$\begin{aligned} &\widehat{c} = c^-, \quad \widehat{c}_T = c^+, \qquad \text{for } (\mathbf{x}, \mathbf{v}) \in \Omega_2^-, \\ &\widehat{c} = c^+, \quad \widehat{c}_T = c^-, \qquad \text{for } (\mathbf{x}, \mathbf{v}) \in \Omega_2^+. \end{aligned}$$

Since the solution $f(T, \mathbf{x}, \mathbf{v})$ can be explicitly expressed as (2.10), (2.11), (2.12), and (2.13), we have proved the existence and uniqueness of the solution for the initial value problem in Theorem 2.1. The l^∞ stability follows easily from the maximum principle and linearity of the Liouville equation.

In the following we prove the l^1-stability of the solution for this initial value problem. Define the l^1-norm of the solution as

$$|f|_1 = \int_\Omega |f(t, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v}.$$

Due to the linearity of the Liouville equation, one only needs to prove that when the initial value is bounded in l^1-norm, then the solution remains bounded in l^1-norm at later time. Assume $|f(0, \mathbf{x}, \mathbf{v})|_1$ exists, we now investigate the relation between $|f(T, \mathbf{x}, \mathbf{v})|_1$ and $|f(0, \mathbf{x}, \mathbf{v})|_1$.

Define the sets

$$\Omega_3^- = \left\{(\mathbf{x}, \mathbf{v}) \in \Omega \,\middle|\, \exists (\mathbf{y}, \mathbf{v}) \in \Omega_1^- \text{ s.t. } \mathbf{x} = \mathbf{y} - c(\mathbf{y})\frac{\mathbf{v}}{|\mathbf{v}|}T\right\},$$

$$\Omega_3^+ = \left\{(\mathbf{x}, \mathbf{v}) \in \Omega \,\middle|\, \exists (\mathbf{y}, \mathbf{v}) \in \Omega_1^+ \text{ s.t. } \mathbf{x} = \mathbf{y} - c(\mathbf{y})\frac{\mathbf{v}}{|\mathbf{v}|}T\right\},$$

$$\Omega_{4,s} = \left\{(\mathbf{x}, \mathbf{v}) \in \Omega \,\middle|\, x_1 < 0,\, x_1 + c^-\frac{v_1}{|\mathbf{v}|}T \geq 0,\, \left(\frac{c^-}{c^+}\right)^2 |\mathbf{v}|^2 \leq v_2^2 + \cdots + v_d^2\right\},$$

$$\Omega_4 = \Omega \backslash \{\Omega_3^- \cup \Omega_3^+ \cup \Omega_{4,s}\}.$$

One has

$$
\begin{aligned}
|f(T,\mathbf{x},\mathbf{v})|_1 &= \int_{\Omega_1^-} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v} + \int_{\Omega_1^+} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v} \\
&+ \int_{\Omega_{2,s}} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v} + \int_{\Omega_2\setminus\Omega_{2,s}} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v}.
\end{aligned} \tag{2.21}
$$

For the first part in (2.21), since the map $(\mathbf{x},\mathbf{v}) \to (\mathbf{x} + c^- \frac{\mathbf{v}}{|\mathbf{v}|}T, \mathbf{v})$ is volume-preserving, (2.10) gives

$$
\begin{aligned}
\int_{\Omega_1^-} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v} &= \int_{\Omega_1^-} \left| f\left(0, \mathbf{x} - c^- \frac{\mathbf{v}}{|\mathbf{v}|}T, \mathbf{v}\right)\right| d\mathbf{x}d\mathbf{v} \\
&= \int_{\Omega_3^-} |f(0,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v}.
\end{aligned} \tag{2.22}
$$

In the same way, the second part in (2.21) holds

$$
\int_{\Omega_1^+} |f(T,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v} = \int_{\Omega_3^+} |f(0,\mathbf{x},\mathbf{v})| d\mathbf{x}d\mathbf{v}. \tag{2.23}
$$

To calculate the last two parts in (2.21), we need to investigate the Jacobians of the maps $(\mathbf{x_T},\mathbf{v_T}) \to (\mathbf{x},\mathbf{v})$ and $(\mathbf{x_R},\mathbf{v_R}) \to (\mathbf{x},\mathbf{v})$. From (2.14)–(2.20), these two maps can be explicitly written out. The nonzero elements in the two Jacobian matrices include

$$
\begin{aligned}
&\frac{\partial x_1}{\partial x_{T,1}}, \frac{\partial x_1}{\partial v_{T,1}}, \frac{\partial x_1}{\partial v_{T,2}}, \dots, \frac{\partial x_1}{\partial v_{T,d}}, \\
&\frac{\partial x_i}{\partial x_{T,1}}, \frac{\partial x_i}{\partial x_{T,i}}, \frac{\partial x_i}{\partial v_{T,1}}, \frac{\partial x_i}{\partial v_{T,2}}, \dots, \frac{\partial x_i}{\partial v_{T,d}}, \quad i = 2,\dots,d, \\
&\frac{\partial v_1}{\partial v_{T,i}}, \quad i = 1,2,\dots,d, \qquad \frac{\partial v_i}{\partial v_{T,i}}, \quad i = 2,\dots,d, \\
&\frac{\partial x_1}{\partial x_{R,1}}, \frac{\partial x_1}{\partial v_{R,1}}, \frac{\partial x_1}{\partial v_{R,2}}, \dots, \frac{\partial x_1}{\partial v_{R,d}}, \\
&\frac{\partial x_i}{\partial x_{R,1}}, \frac{\partial x_i}{\partial x_{R,i}}, \frac{\partial x_i}{\partial v_{R,1}}, \frac{\partial x_i}{\partial v_{R,2}}, \dots, \frac{\partial x_i}{\partial v_{R,d}}, \quad i = 2,\dots,d, \\
&\frac{\partial v_i}{\partial v_{R,i}}, \quad i = 1,2,\dots,d,
\end{aligned}
$$

from which only the diagonal elements influence the Jacobians. They are

$$
\begin{aligned}
&\frac{\partial x_1}{\partial x_{T,1}} = \left(\frac{\widehat{c}}{\widehat{c}_T}\right)^2 \frac{v_1}{v_{T,1}}, \\
&\frac{\partial x_i}{\partial x_{T,i}} = 1, \quad i = 2,\dots,d, \\
&\frac{\partial v_1}{\partial v_{T,1}} = \left(\frac{\widehat{c}_T}{\widehat{c}}\right)^2 \frac{v_{T,1}}{v_1}, \\
&\frac{\partial v_i}{\partial v_{T,i}} = 1, \quad i = 2,\dots,d,
\end{aligned}
$$

$$\frac{\partial x_1}{\partial x_{R,1}} = -1, \qquad \frac{\partial x_i}{\partial x_{R,i}} = 1, \quad i = 2, \dots, d,$$
$$\frac{\partial v_1}{\partial v_{R,1}} = -1, \qquad \frac{\partial v_i}{\partial v_{R,i}} = 1, \quad i = 2, \dots, d.$$

Thus it is verified that the two maps $(\mathbf{x_T}, \mathbf{v_T}) \to (\mathbf{x}, \mathbf{v})$ and $(\mathbf{x_R}, \mathbf{v_R}) \to (\mathbf{x}, \mathbf{v})$ are *volume-preserving*.

For the third part in (2.21), from (2.12) one has

$$\int_{\Omega_{2,s}} |f(T, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} = \int_{\Omega_{2,s}} |f(0, \mathbf{x_R}, \mathbf{v_R})| d\mathbf{x} d\mathbf{v} = \int_{\Omega_{4,s}} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v}. \tag{2.24}$$

For the fourth part in (2.21), from (2.13) one has

$$\begin{aligned}
\int_{\Omega_2 \setminus \Omega_{2,s}} |f(T, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} &= \int_{\Omega_2 \setminus \Omega_{2,s}} \alpha_T(\mathbf{v_T}) \, |f(0, \mathbf{x_T}, \mathbf{v_T})| d\mathbf{x} d\mathbf{v} \\
&\quad + \int_{\Omega_2 \setminus \Omega_{2,s}} \alpha_R(\mathbf{v_R}) \, |f(0, \mathbf{x_R}, \mathbf{v_R})| d\mathbf{x} d\mathbf{v} \\
&= \int_{\Omega_4} \alpha_T(\mathbf{v}) \, |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} + \int_{\Omega_4} \alpha_R(\mathbf{v}) \, |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} \\
&= \int_{\Omega_4} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v}. \qquad (2.25)
\end{aligned}$$

Together with (2.21), (2.22), (2.23), (2.24), and (2.25), one gets

$$\begin{aligned}
|f(T, \mathbf{x}, \mathbf{v})|_1 &= \int_{\Omega_3^-} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} + \int_{\Omega_3^+} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} \\
&\quad + \int_{\Omega_{4,s}} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} + \int_{\Omega_4} |f(0, \mathbf{x}, \mathbf{v})| d\mathbf{x} d\mathbf{v} \\
&= |f(0, \mathbf{x}, \mathbf{v})|_1.
\end{aligned}$$

This is the l^1-stability—in fact l^1 preservation—of the solution for the initial value problem in Theorem 2.1. ∎

Remark 2.1. In [2], a classical-classical coupling model that connects two domains of classical mechanics with constant potentials with a classical domain $[a, b]$ in between where the potential is variable was introduced, where the interface conditions at a and b were given. When $a = b$, their interface conditions reduce to (2.5).

3. The scheme in 1D.

3.1. The numerical flux. We now describe our finite difference scheme for the 1D Liouville equation

$$f_t + c(x)\text{sign}(\xi) f_x - c_x |\xi| f_\xi = 0\,. \tag{3.1}$$

We employ a uniform mesh with grid points at $x_{i+\frac{1}{2}}, i = 0, \dots, N$, in the x-direction and $\xi_{j+\frac{1}{2}}, j = 0, \dots, M$ in the ξ-direction. The cells are centered at (x_i, ξ_j), $i = 1, \dots, N, j = 1, \dots, M$ with $x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}})$ and $\xi_j = \frac{1}{2}(\xi_{j+\frac{1}{2}} + \xi_{j-\frac{1}{2}})$. The uniform mesh size is denoted by $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta\xi = \xi_{j+\frac{1}{2}} - \xi_{j-\frac{1}{2}}$. We also assume a uniform time step Δt and the discrete time is given by $0 = t_0 < t_1 < \cdots < t_L = T$.

We introduce the mesh ratios $\lambda_x^t = \frac{\Delta t}{\Delta x}, \lambda_\xi^t = \frac{\Delta t}{\Delta \xi}$, assumed to be fixed. The cell average of f is defined by

$$f_{ij} = \frac{1}{\Delta x \Delta \xi} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x, \xi, t) d\xi dx.$$

We assume the local wave speed is Lipschitz continuous except at its isolated discontinuous points. Assume that the discontinuous points of the wave speed c are located at the grid points. Let the left and right limits of $c(x)$ at point $x_{i+1/2}$ be $c^+_{i+\frac{1}{2}}$ and $c^-_{i+\frac{1}{2}}$, respectively. Note that if c is continuous at $x_{j+1/2}$, then $c^+_{i+\frac{1}{2}} = c^-_{i+\frac{1}{2}}$. We approximate c by a piecewise linear function

$$c(x) \approx c^+_{j-1/2} + \frac{c^-_{j+1/2} - c^+_{j-1/2}}{\Delta x}(x - x_{j-1/2}).$$

We also define the average wave speed as $c_i = \frac{1}{2}(c^+_{i-\frac{1}{2}} + c^-_{i+\frac{1}{2}})$. We will adopt the flux splitting technique used in [36, 27, 28]. The semidiscrete scheme (with time continuous) reads

$$(f_{ij})_t + \frac{c_i \text{sign}(\xi_j)}{\Delta x}(f^-_{i+\frac{1}{2},j} - f^+_{i-\frac{1}{2},j}) - \frac{c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}}{\Delta x \Delta \xi}|\xi_j|(f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}) = 0, \tag{3.2}$$

where the numerical fluxes $f_{i,j+\frac{1}{2}}$ are defined using the upwind discretization. Since the characteristics of the Liouville equation may be different on the two sides of the interface, the corresponding numerical fluxes should also be different. The essential part of our algorithm is to define the split numerical fluxes $f^-_{i+\frac{1}{2},j}, f^+_{i-\frac{1}{2},j}$ at each cell interface. We will use (2.5) to define these fluxes.

Assume c is discontinuous at $x_{i+\frac{1}{2}}$. Consider the case $\xi_j > 0$. Using upwind scheme, $f^-_{i+\frac{1}{2},j} = f_{ij}$. However, by (2.5),

$$f^+_{i+\frac{1}{2},j} = \alpha^T f(t, x^-_{i+\frac{1}{2}}, \xi^-) + \alpha^R f(t, x^+_{i+\frac{1}{2}}, -\xi^+)$$

while ξ^- is obtained from $\xi^+ = \xi_j$ from (2.1). Since ξ^- may not be a grid point, we have to define it approximately. One can first locate the two cell centers that bound this velocity, and then use a linear interpolation to evaluate the needed numerical flux at ξ^-. The case of $\xi_j < 0$ is treated similarly. The detailed algorithm to generate the numerical flux is given below.

Algorithm I

- if $\xi_j > 0$

  $f^-_{i+\frac{1}{2},j} = f_{ij}$,

  $\xi' = \dfrac{c^+_{i+\frac{1}{2}}}{c^-_{i+\frac{1}{2}}}\xi_j$

  - if $\xi_k \le \xi' < \xi_{k+1}$ for some k

    $\alpha^R = \left(\dfrac{c^+_{i+\frac{1}{2}} - c^-_{i+\frac{1}{2}}}{c^+_{i+\frac{1}{2}} + c^-_{i+\frac{1}{2}}}\right)^2, \quad \alpha^T = 1 - \alpha^R$

    $f^+_{i+\frac{1}{2},j} = \alpha^T \left(\dfrac{\xi_{k+1} - \xi'}{\Delta \xi} f_{i,k} + \dfrac{\xi' - \xi_k}{\Delta \xi} f_{i,k+1}\right) + \alpha^R f_{i+1,k'}$

where $\xi_{k'} = -\xi_k$

- end
- if $\xi_j < 0$

  $f^+_{i+\frac{1}{2},j} = f_{i+1,j}$,

  $\xi' = \dfrac{c^-_{i+\frac{1}{2}}}{c^+_{i+\frac{1}{2}}}\xi_j$

  - if $\xi_k \le \xi' < \xi_{k+1}$ for some k

    $$\alpha^R = \left(\frac{c^+_{i+\frac{1}{2}} - c^-_{i+\frac{1}{2}}}{c^+_{i+\frac{1}{2}} + c^-_{i+\frac{1}{2}}}\right)^2, \quad \alpha^T = 1 - \alpha^R$$

    $$f^-_{i+\frac{1}{2},j} = \alpha^T\left(\frac{\xi_{k+1} - \xi'}{\Delta\xi} f_{i+1,k} + \frac{\xi' - \xi_k}{\Delta\xi} f_{i+1,k+1}\right) + \alpha^R f_{i,k'}$$

    where $\xi_{k'} = -\xi_k$
  - end

The above algorithm for evaluating numerical fluxes is of first order. One can obtain a second order flux by incorporating the slope limiter, such as the van Leer or minmod slope limiter [31], into the above algorithm. This can be achieved by replacing f_{ik} with $f_{ik} + \frac{\Delta x}{2} s_{ik}$, and replacing $f_{i+1,k}$ with $f_{i+1,k} - \frac{\Delta x}{2} s_{i+1,k}$ in the above algorithm for all possible index k, where s_{ik} is the slope limiter in the x-direction.

After the spatial discretization is specified, one can use any time discretization for the time derivative.

3.2. Positivity and l^∞ contraction. Since the exact solution of the Liouville equation is positive when the initial profile is, it is important that the numerical solution inherits this property.

We only consider the scheme using the first order numerical flux, and the forward Euler method in time. Without loss of generality, we consider the case $\xi_j > 0$ and $c^-_{i+\frac{1}{2}} < c^+_{i-\frac{1}{2}}$ for all i (the other cases can be treated similarly with the same conclusion). The scheme reads

$$\frac{f^{n+1}_{ij} - f^n_{ij}}{\Delta t} + c_i \frac{f_{ij} - (d_1 f_{i-1,k} + d_2 f_{i-1,k+1} + \alpha^R f_{i,k'})}{\Delta x} - \frac{c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}}{\Delta x}\xi_j \frac{f_{ij} - f_{i,j-1}}{\Delta\xi} = 0,$$

where d_1, d_2, α^R are nonnegative and $d_1 + d_2 = \alpha^T = 1 - \alpha^R$. We omit the superscript n of f. The above scheme can be rewritten as

$$f^{n+1}_{ij} = \left(1 - c_i\lambda^t_x - \frac{\left|c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}\right|}{\Delta x}|\xi_j|\lambda^t_\xi\right) f_{ij} + c_i\lambda^t_x\left(d_1 f_{i-1,k} + d_2 f_{i-1,k+1} + \alpha^R f_{i,k'}\right)$$

$$+ \frac{\left|c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}\right|}{\Delta x}|\xi_j|\lambda^t_\xi f_{i,j-1}\,. \tag{3.3}$$

Now we investigate the positivity of scheme (3.3). This is to prove that if $f^n_{ij} \ge 0$ for all (i, j), then this is also true for f^{n+1}. Clearly one just needs to show that all of the coefficients before f^n are nonnegative. A sufficient condition for this is clearly

$$1 - c_i\lambda^t_x - \frac{\left|c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}\right|}{\Delta x}|\xi_j|\lambda^t_\xi \ge 0,$$

or

$$\Delta t \max_{i,j} \left[\frac{c_i}{\Delta x} + \frac{\frac{\left|c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}\right|}{\Delta x}|\xi_j|}{\Delta \xi} \right] \le 1. \tag{3.4}$$

The quantity $\frac{\left|c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}}\right|}{\Delta x}$ now represents the wave speed gradient at its *smooth* point, which has a *finite* upper bound since $c \in W^{1,\infty}$. In addition, typically f has a compact support, so in practical computation ξ is confined in a bounded set. Thus our scheme allows a time step $\Delta t = O(\Delta x, \Delta \xi)$.

According to the study in [35], our second order scheme, which incorporates a slope limiter into the first order scheme, is positive under the half CFL condition, namely, the constant on the right-hand side of (3.4) is 1/2.

The above conclusion is drawn on the forward Euler time discretization. One can draw the same conclusion for the second order TVD Runge–Kutta time discretization [40].

The l^∞-contracting property of this scheme:

$$\|f^n\|_\infty \le \|f^0\|_\infty$$

follows easily, because the coefficients in (3.3) are positive and the sum of them is 1.

3.3. The l^1-stability of the scheme. In this section we prove the l^1-stability of the scheme (with the first order numerical flux and the forward Euler method in time). For simplicity, we consider the case when the wave speed has only one discontinuity at grid point $x_{m+\frac{1}{2}}$ with $c^-_{m+\frac{1}{2}} > c^+_{m+\frac{1}{2}}$, and $c'(x) > 0$ at smooth points. The other cases, namely, when $c'(x) \le 0$, or the wave speed having several discontinuity points with increased or decreased jumps, can be discussed similarly. Denote $\lambda_c \equiv c^+_{m+\frac{1}{2}}/c^-_{m+\frac{1}{2}} < 1$.

We consider the general case that $\xi_1 < 0, \xi_M > 0$. For this case, as adopted in [25, 28], the computational domain should exclude a set $O_\xi = \{(x,\xi) \in \mathbb{R}^2 \,|\xi = 0\}$, which causes singularity in the velocity field. For example, we can exclude the following index set:

$$D_o = \left\{ (i,j) \middle| |\xi_j| < \frac{\Delta \xi}{2} \right\},$$

from the computational domain.

Since $c(x)$ has a discontinuity, we also define an index set

$$D_l^4 = \{(i,j) | x_i \le x_m, \xi_j < \lambda_c \xi_1\}.$$

As mentioned in [28], D_l^4 represents the area where waves come from outside of the domain $[x_1, x_N] \times [\xi_1, \xi_M]$. In order to implement our scheme conveniently, this index set is also excluded from the computational domain. Thus the computational domain is chosen as

$$E_d = \{(i,j) | i = 1, \dots, N, j = 1, \dots, M\} \setminus \left\{D_o \cup D_l^4\right\}. \tag{3.5}$$

As a result of excluding the index set D_o from the computational domain, the computational domain is split into two nonoverlapping parts:

$$E_d = \{(i,j) \in E_d | \xi_j > 0\} \cup \{(i,j) \in E_d | \xi_j < 0\} \equiv E_d^+ \cup E_d^-.$$

In [28] we analyzed the l^1-stability of the scheme on E_d^+ and E_d^- separately. Here we will conduct the analysis on the full phase space E_d since transmission and reflection waves coexist at the interface.

We define the l^1-norm of a numerical solution u_{ij} in the set E_d to be

$$|f|_1 = \frac{1}{N_d} \sum_{(i,j)\in E_d} |f_{ij}| \tag{3.6}$$

with N_d being the number of elements in E_d.

Given the initial data $f_{ij}^0, (i,j) \in E_d$. Denote the numerical solution at time T to be $f_{ij}^L, (i,j) \in E_d$. To prove the l^1-stability, we need to show that $|f^L|_1 \le C|f^0|_1$.

Due to the linearity of the scheme, the equation for the error between the analytical and the numerical solutions is the same as (3.3), so in this section, f_{ij} will denote the error. We assume there is no error at the boundary, thus $f_{ij}^n = 0$ at the boundary. If the l^1-norm of the error introduced at each time step in the incoming boundary cells is ensured to be $o(1)$ part of $|f^n|_1$, our following analysis still applies.

Now denote

$$A_i = \frac{1}{\Delta x} \left| c^-_{i+\frac{1}{2}} - c^+_{i-\frac{1}{2}} \right| . \tag{3.7}$$

Since $c(x)$ is Lipschitz continuous at its smooth part, there exists an $A_u > 0$, such that $A_i < A_u, \forall i$. Assume also that there is an $C_m > 0$ such that $c_i > C_m, \forall i$. The finite difference scheme is given as follows:

- When $\xi_j > 0$
  1) if $i \neq m+1$,

$$f_{ij}^{n+1} = \left(1 - A_i|\xi_j|\lambda_\xi^t - c_i\lambda_x^t\right) f_{ij} + A_i|\xi_j|\lambda_\xi^t f_{i,j+1} + c_i\lambda_x^t f_{i-1,j}, \tag{3.8}$$

  2)

$$\begin{aligned} f_{m+1,j}^{n+1} &= \left(1 - A_{m+1}|\xi_j|\lambda_\xi^t - c_{m+1}\lambda_x^t\right) f_{m+1,j} + A_{m+1}|\xi_j|\lambda_\xi^t f_{m+1,j+1} \\ &\quad + c_{m+1}\lambda_x^t(d_{j1}f_{m,k} + d_{j2}f_{m,k+1} + \alpha^R f_{m+1,k'}). \end{aligned} \tag{3.9}$$

- When $\xi_j < 0$
  3) if $i \neq m$,

$$f_{ij}^{n+1} = \left(1 - A_i|\xi_j|\lambda_\xi^t - c_i\lambda_x^t\right) f_{ij} + A_i|\xi_j|\lambda_\xi^t f_{i,j+1} + c_i\lambda_x^t f_{i+1,j}, \tag{3.10}$$

  4)

$$\begin{aligned} f_{mj}^{n+1} &= \left(1 - A_m|\xi_j|\lambda_\xi^t - c_m\lambda_x^t\right) f_{mj} + A_m|\xi_j|\lambda_\xi^t f_{m,j+1} \\ &\quad + c_m\lambda_x^t(d_{j1}f_{m+1,k} + d_{j2}f_{m+1,k+1} + \alpha^R f_{m,k'}), \end{aligned} \tag{3.11}$$

where $0 \le d_{j1}, d_{j2} \le 1$ and $d_{j1} + d_{j2} = \alpha^T = 1 - \alpha^R = 1$. In (3.9) k is determined by $\xi_k \le \lambda_c\xi_j < \xi_{k+1}$ and $\xi_{k'} = -\xi_k$. In (3.11) k is determined by $\xi_k \le \frac{\xi_j}{\lambda_c} < \xi_{k+1}$ and $\xi_{k'} = -\xi_k$.

When summing up all absolute values of f_{ij}^{n+1} in (3.8)–(3.11), one typically gets the following inequality:

$$|f^{n+1}|_1 \le \frac{1}{N_d} \sum_{(i,j)\in E_d} \alpha_{ij}|f_{ij}^n|, \tag{3.12}$$

where the coefficients α_{ij} are positive. One can check that, under the CFL condition (3.4), $\alpha_{ij} \le 1 + 2A_u\Delta t$ except for possibly $(i,j) \in D^-_{m+1} \bigcup D^+_m$, where

$$D^-_{m+1} = \{(i,j) \in E^-_d | i = m+1\}, \quad D^+_m = \{(i,j) \in E^+_d | i = m\}.$$

We next derive the bounds for M^-, M^+ defined as

$$M^- = \max_{(m+1,j)\in D^-_{m+1}} \alpha_{m+1,j}, \quad M^+ = \max_{(m,j)\in D^+_m} \alpha_{m,j}.$$

Define the set

$$S^{m+1}_j = \left\{ j' \middle| \xi_{j'} < 0, \left| \frac{\xi_{j'}}{\lambda_c} - \xi_j \right| < \Delta\xi \right\} \quad \text{for} \quad (m+1,j) \in D^-_{m+1}.$$

Let the number of elements in S^{m+1}_j be N^{m+1}_j. One can check that $N^{m+1}_j \le 2\lambda_c + 1$ because every two elements $j'_1, j'_2 \in S^{m+1}_j$ satisfy $\left| \frac{\xi_{j'_1}}{\lambda_c} - \frac{\xi_{j'_2}}{\lambda_c} \right| \ge \frac{\Delta\xi}{\lambda_c}$.

On the other hand, one can easily check from (3.9) and (3.11), for $(m+1,j) \in D^-_{m+1}$, that

$$\alpha_{m+1,j} < 1 - c_{m+1}\lambda^t_x + c_m\lambda^t_x\,(2\lambda_c+1)\,\alpha^T + \alpha^R c_{m+1}\lambda^t_x = 1 + \alpha^T(c_m + c_{m+1})\lambda^t_x + O(\Delta x),$$

so for sufficiently small Δx, M^- can be bounded by

$$M^- < 1 + 2\alpha^T(c_m + c_{m+1})\lambda^t_x.$$

Similarly, one can prove for sufficiently small Δx, M^+ is also bounded by

$$M^+ < 1 + 2\alpha^T(c_m + c_{m+1})\lambda^t_x.$$

Denote $M' = 2\alpha^T(c_m + c_{m+1})\lambda^t_x$. From (3.12),

$$|f^{n+1}|_1 < (1 + 2A_u\Delta t)\,|f^n|_1 + \frac{M'}{N_d} \sum_{(m+1,j)\in D^-_{m+1}} |f^n_{m+1,j}| + \frac{M'}{N_d} \sum_{(m,j)\in D^+_m} |f^n_{m,j}|. \tag{3.13}$$

Consecutively using (3.13) gives

$$|f^L|_1 < (1 + 2A_u\Delta t)^L \left\{ |f^0|_1 + \frac{M'}{N_d} \sum_{n=0}^{L-1} \left[\sum_{(m+1,j)\in D^-_{m+1}} |f^n_{m+1,j}| \right] + \frac{M'}{N_d} \sum_{n=0}^{L-1} \left[\sum_{(m,j)\in D^+_m} |f^n_{m,j}| \right] \right\}. \tag{3.14}$$

Define

$$S_1 = \sum_{n=0}^{L-1} \left[\sum_{(m+1,j)\in D^-_{m+1}} |f^n_{m+1,j}| \right], \quad S_2 = \sum_{n=0}^{L-1} \left[\sum_{(m,j)\in D^+_m} |f^n_{m,j}| \right]. \tag{3.15}$$

These two terms can be proved in the same way as in [29] to get

$$S_1, S_2 < C_T N_d |f^0|_1, \tag{3.16}$$

where

$$C_T \equiv \exp\left(\frac{2A_u}{C_m}(x_N - x_1)\right)\frac{1}{C_m \lambda_x^t}. \tag{3.17}$$

Combing (3.14) and (3.16),

$$\begin{aligned} |f^L|_1 &< (1 + 2A_u\Delta t)^L \left\{|f^0|_1 + 2C_T M' |f^0|_1\right\} \\ &= \exp(2A_u T)\left[1 + 2C_T M'\right]|f^0|_1 \\ &\equiv C|f^0|_1, \end{aligned}$$

where $C \equiv \exp(2A_u T)[1 + 2C_T M']$.

Thus we prove the following theorem.

THEOREM 3.1. *Let $c(x) \in W^{1,\infty}$ have a discontinuity at one point, and be bounded below from zero, $c(x) > C_m > 0$. Assume f^0 has a finite l^1-norm defined* (3.6) *with a compact support in ξ. Then under the hyperbolic CFL condition* (3.4)*, the solution yielded by the scheme* (3.8)–(3.11) *is stable in l^1-norm:*

$$|f^L|_1 < C|f^0|_1.$$

4. The scheme in two space dimension. Consider the 2D Liouville equation

$$f_t + \frac{c(x,y)\xi}{\sqrt{\xi^2+\eta^2}} f_x + \frac{c(x,y)\eta}{\sqrt{\xi^2+\eta^2}} f_y - c_x\sqrt{\xi^2+\eta^2} f_\xi - c_y\sqrt{\xi^2+\eta^2} f_\eta = 0. \tag{4.1}$$

We employ a uniform mesh with grid points at $x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}, \xi_{k+\frac{1}{2}}, \eta_{l+\frac{1}{2}}$ in each direction. The cells are centered at $(x_i, y_j, \xi_k, \eta_l)$ with $x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}), y_j = \frac{1}{2}(y_{j+\frac{1}{2}} + y_{j-\frac{1}{2}}), \xi_k = \frac{1}{2}(\xi_{k+\frac{1}{2}} + \xi_{k-\frac{1}{2}}), \eta_l = \frac{1}{2}(\eta_{l+\frac{1}{2}} + \eta_{l-\frac{1}{2}})$. The mesh size is denoted by $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}, \Delta\xi = \xi_{k+\frac{1}{2}} - \xi_{k-\frac{1}{2}}, \Delta\eta = \eta_{l+\frac{1}{2}} - \eta_{l-\frac{1}{2}}$. We define the cell average of f as

$$f_{ijkl} = \frac{1}{\Delta x\Delta y\Delta\xi\Delta\eta}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{\xi_{k-\frac{1}{2}}}^{\xi_{k+\frac{1}{2}}}\int_{\eta_{l-\frac{1}{2}}}^{\eta_{l+\frac{1}{2}}} f(x,y,\xi,\eta,t)\,d\eta\,d\xi\,dy\,dx.$$

Similar to the 1D case, we approximate $c(x,y)$ by a piecewise bilinear function, and, for convenience, we always provide two interface values of c at each cell interface. When c is smooth at a cell interface, the two potential interface values are identical. We also define the average wave speed in a cell by averaging the four wave speed values at the cell interface:

$$c_{ij} = \frac{1}{4}(c^+_{i-\frac{1}{2},j} + c^-_{i+\frac{1}{2},j} + c^+_{i,j-\frac{1}{2}} + c^-_{i,j+\frac{1}{2}}).$$

The 2D Liouville equation (4.1) can be semidiscretized as

$$(f_{ijkl})_t + \frac{c_{ij}\xi_k}{\Delta x\sqrt{\xi_k^2+\eta_l^2}}\left(f^-_{i+\frac{1}{2},jkl} - f^+_{i-\frac{1}{2},jkl}\right)$$

$$
\begin{aligned}
&+\frac{c_{ij}\eta_l}{\Delta y\sqrt{\xi_k^2+\eta_l^2}}\left(f^-_{i,j+\frac12,kl}-f^+_{i,j-\frac12,kl}\right)\\
&-\frac{c^-_{i+\frac12,j}-c^+_{i-\frac12,j}}{\Delta x\Delta\xi}\sqrt{\xi_k^2+\eta_l^2}\left(f_{ij,k+\frac12,l}-f_{ij,k-\frac12,l}\right)\\
&-\frac{c^-_{i,j+\frac12}-c^+_{i,j-\frac12}}{\Delta y\Delta\eta}\sqrt{\xi_k^2+\eta_l^2}\left(f_{ijk,l+\frac12}-f_{ijk,l-\frac12}\right)\\
=\quad & 0,
\end{aligned}
$$

where the interface values $f_{ij,k+\frac12,l}, f_{ijk,l+\frac12}$ are provided by the upwind approximation, and the split interface values $f^-_{i+\frac12,jkl}, f^+_{i-\frac12,jkl}, f^-_{i,j+\frac12,kl}, f^+_{i,j-\frac12,kl}$ should be obtained using a similar but slightly different algorithm for the 1D case. For example, to evaluate $f^\pm_{i+\frac12,jkl}$ we can extend Algorithm I as

Algorithm I in 2D

- if $\xi_k>0$

  $f^-_{i+\frac12,jkl}=f_{ijkl},\quad \xi_{k_1}=-\xi_k$

  - if $\left(\frac{C^+_{i+\frac12,j}}{C^-_{i+\frac12,j}}\right)^2(\xi_k)^2+\left[\left(\frac{C^+_{i+\frac12,j}}{C^-_{i+\frac12,j}}\right)^2-1\right](\eta_l)^2>0$

    $\xi^-=\sqrt{\left(\frac{C^+_{i+\frac12,j}}{C^-_{i+\frac12,j}}\right)^2(\xi_k)^2+\left[\left(\frac{C^+_{i+\frac12,j}}{C^-_{i+\frac12,j}}\right)^2-1\right](\eta_l)^2}$

    - if $\xi_{k'}\le\xi^-<\xi_{k'+1}$ for some k'

      $\gamma^+=\frac{\xi_k}{\sqrt{(\xi_k)^2+(\eta_l)^2}},\quad \gamma^-=\frac{\xi^-}{\sqrt{(\xi')^2+(\eta_l)^2}}$

      $\alpha^R=\left(\frac{c^+_{i+\frac12}\gamma^- - c^-_{i+\frac12}\gamma^+}{c^+_{i+\frac12}\gamma^- + c^-_{i+\frac12}\gamma^+}\right)^2,\quad \alpha^T=1-\alpha^R$

      $f^+_{i+\frac12,jkl}=\alpha^T\left(\frac{\xi_{k'+1}-\xi^-}{\Delta\xi}f_{ij,k',l}+\frac{\xi^--\xi_{k'}}{\Delta\xi}f_{ij,k'+1,l}\right)+\alpha^R f_{i+1,j,k_1,l}$

    - end
  - else

    $f^+_{i+\frac12,jkl}=f_{i+1,j,k_1,l}$
  - end

- if $\xi_k<0$

  $f^+_{i+\frac12,jkl}=f_{i+1,jkl},\quad \xi_{k_1}=-\xi_k$

  - if $\left(\frac{C^-_{i+\frac12,j}}{C^+_{i+\frac12,j}}\right)^2(\xi_k)^2+\left[\left(\frac{C^-_{i+\frac12,j}}{C^+_{i+\frac12,j}}\right)^2-1\right](\eta_l)^2>0$

    $\xi^+=-\sqrt{\left(\frac{C^-_{i+\frac12,j}}{C^+_{i+\frac12,j}}\right)^2(\xi_k)^2+\left[\left(\frac{C^-_{i+\frac12,j}}{C^+_{i+\frac12,j}}\right)^2-1\right](\eta_l)^2}$

    - if $\xi_{k'}\le\xi^+<\xi_{k'+1}$ for some k'

      $\gamma^+=\frac{|\xi^+|}{\sqrt{(\xi^+)^2+(\eta_l)^2}},\quad \gamma^-=\frac{|\xi_k|}{\sqrt{(\xi_k)^2+(\eta_l)^2}}$

$$\alpha^R = \left(\frac{c^+_{i+\frac{1}{2}}\gamma^- - c^-_{i+\frac{1}{2}}\gamma^+}{c^+_{i+\frac{1}{2}}\gamma^- + c^-_{i+\frac{1}{2}}\gamma^+}\right)^2, \quad \alpha^T = 1 - \alpha^R$$

$$f^-_{i+\frac{1}{2},jkl} = \alpha^T\left(\frac{\xi_{k'+1} - \xi^+}{\Delta\xi} f_{i+1,j,k',l} + \frac{\xi^+ - \xi_{k'}}{\Delta\xi} f_{i+1,j,k'+1,l}\right) + \alpha^R f_{ij,k_1,l}$$

- end
- else

$$f^-_{i+\frac{1}{2},jkl} = f_{i,j,k_1,l} \text{ where } \xi_{k_1} = -\xi_k$$

- end

The flux $f^{\pm}_{i,j+\frac{1}{2},kl}$ can be constructed similarly.

As introduced in section 2, the essential difference between the 1D and 2D flux definitions is that in the 2D case, the phenomenon that a wave is completely reflected at the interface does occur, while in 1D, the transmission and reflection waves always coexist at the interface.

Since the wave speed $c \in W^{1,\infty}$, this scheme, similar to the 1D scheme, is also subject to a hyperbolic CFL condition under which the scheme is positive.

5. Numerical examples. In this section we present numerical examples to demonstrate the validity of the proposed scheme and to study the numerical accuracy. In the numerical computations the second order TVD Runge–Kutta time discretization [40] is used. We use the second order scheme with the van Leer slope limiter in constructing the numerical fluxes except for Example 5.2.

Example 5.1. A 1D problem with exact L^∞-solution. Consider the 1D Liouville equation

$$f_t + c(x)\text{sign}(\xi) f_x - c_x|\xi| f_\xi = 0 \tag{5.1}$$

with a discontinuous wave speed given by

$$c(x) = \begin{cases} 0.6 & x < 0 \\ 0.2 & x > 0. \end{cases}$$

The initial data is given by

$$f(x,\xi,0) = \begin{cases} 1 & x < 0, \xi > 0, \sqrt{x^2 + 4\xi^2} < 1, \\ 1 & x > 0, \xi < 0, \sqrt{x^2 + \xi^2} < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

In this example the reflection and transmission coefficients α^R, α^T at the interface are $\alpha^R = \dfrac{1}{4}, \alpha^T = \dfrac{3}{4}$. The exact solution for f at $t = 1$ is given by

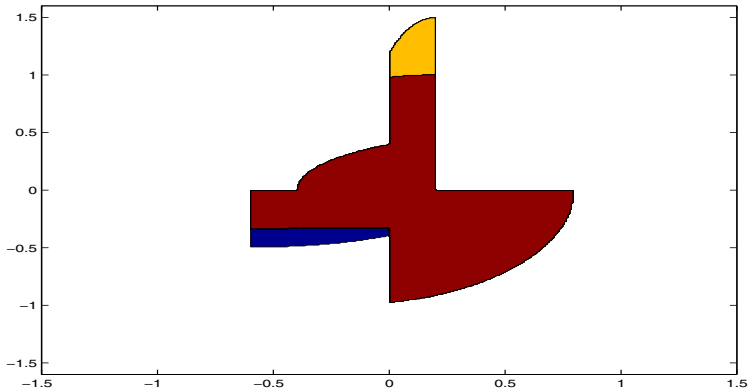


FIG. 5.1. *Example* 5.1*, the nonzero part of the exact solution* $f(x,\xi,1)$ *depicted on the* 400×400 *mesh. The horizontal axis is the position, the vertical axis is the slowness.*

$$f(x,\xi,1)=\begin{cases}\alpha^T & 0<x<0.2,\quad \sqrt{1-(0.2-x)^2}<\xi<1.5\sqrt{1-(3x-0.6)^2};\\ 1 & 0<x<0.2,\quad 0<\xi<\sqrt{1-(0.2-x)^2};\\ 1 & 0<x<0.8,\quad -\sqrt{1-(x+0.2)^2}<\xi<0;\\ 1 & -0.4<x<0,\quad 0<\xi<\dfrac{1}{2}\sqrt{1-(x-0.6)^2};\\ 1 & -0.6<x<0,\quad -\dfrac{1}{3}\sqrt{1-\left(\dfrac{x}{3}+0.2\right)^2}<\xi<0;\\ \alpha^R & -0.6<x<0,\quad -\dfrac{1}{2}\sqrt{1-(x+0.6)^2}<\xi<-\dfrac{1}{3}\sqrt{1-\left(\dfrac{x}{3}+0.2\right)^2};\\ 0 & \text{otherwise,}\end{cases}\tag{5.3}$$

as shown in Figure 5.1.

We are also interested in computing the moments of f, which include the density

$$\rho(x,t)=\int f(x,\xi,t)d\xi$$

and the averaged slowness

$$u(x,t)=\int f(x,\xi,t)\xi d\xi\Big/\rho(x,t).$$

At $t = 1$, the exact density is

$$\rho(x,1) = \begin{cases} \sqrt{1-(x+0.2)^2} & 0.2 < x < 0.8; \\ 1.5\alpha^T\sqrt{1-(3x-0.6)^2} + \alpha^R\sqrt{1-(0.2-x)^2} & \\ +\sqrt{1-(x+0.2)^2} & 0 < x < 0.2; \\ \dfrac{\alpha^T}{3}\sqrt{1-\left(\dfrac{x}{3}+0.2\right)^2} + \dfrac{\alpha^R}{2}\sqrt{1-(x+0.6)^2} & -0.6 < x < -0.4; \\ \dfrac{\alpha^T}{3}\sqrt{1-\left(\dfrac{x}{3}+0.2\right)^2} + \dfrac{\alpha^R}{2}\sqrt{1-(x+0.6)^2} & \\ +\dfrac{1}{2}\sqrt{1-(x-0.6)^2} & -0.4 < x < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

The averaged slowness only has definition in $[-0.6, 0.8]$ since the density is zero outside this interval. The exact averaged slowness in $[-0.6, 0.8]$ is

$$u(x,1) = \frac{1}{2\rho(x,1)} \begin{cases} -\left[1-(x+0.2)^2\right] & 0.2 < x < 0.8; \\ 2.25\alpha^T\left[1-(3x-0.6)^2\right] + \alpha^R\left[1-(0.2-x)^2\right] & \\ -\left[1-(x+0.2)^2\right] & 0 < x < 0.2; \\ \dfrac{-\alpha^T}{9}\left[1-\left(\dfrac{x}{3}+0.2\right)^2\right] - \dfrac{\alpha^R}{4}\left[1-(x+0.6)^2\right] & -0.6 < x < -0.4; \\ \dfrac{-\alpha^T}{9}\left[1-\left(\dfrac{x}{3}+0.2\right)^2\right] - \dfrac{\alpha^R}{4}\left[1-(x+0.6)^2\right] & \\ +\dfrac{1}{4}\left[1-(x-0.6)^2\right] & -0.4 < x < 0. \end{cases} \tag{5.5}$$

We choose the time step as $\Delta t = \frac{1}{2}\Delta\xi$. The computational domain is chosen as $[x, \xi] \in [-1.5, 1.5] \times [-1.6, 1.6]$. Table 5.1 compares the l^1-error of the numerical solutions for f, ρ on $[-1.5, 1.5]$ and u on $[-0.6, 0.8]$ computed with different meshes, respectively.

The convergence rate of f in the l^1-norm is shown to be about 0.74. This agrees with the well-established theory [30, 42], that the l^1-error by finite difference scheme for a discontinuous solution of a linear hyperbolic equation is at most half order. The convergence rate of ρ and u are shown to be about 0.74 and 0.98, respectively, since the solutions also contain discontinuities away from the interface.

Figure 5.2 shows the numerical density ρ and averaged slowness u computed with a 400× 400 cell along with the exact solutions in the physical space.

Example 5.2. Computing the physical observables of a 1D problem with measure-valued solution. Consider the 1D Liouville equation (5.1), where the wave speed is a

TABLE 5.1
l^1 error of the numerical solutions with different meshes.

meshes	50 × 50	100 × 100	200× 200	400× 400
f	0.179090	0.104788	0.064989	0.038535
ρ	0.124361	0.079007	0.043248	0.025187
u	0.143083	0.063068	0.043079	0.019870

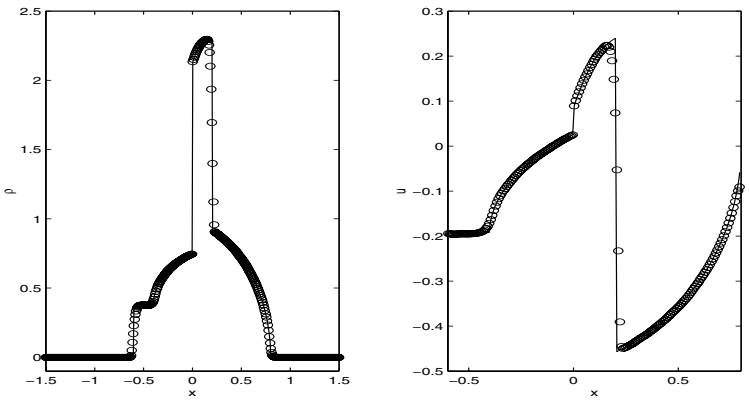


FIG. 5.2. *Example* 5.1, *the density* ρ *and averaged slowness* u *at* $t = 1$. *Solid line: the exact solution; "o": the numerical solutions using the* 400×400 *mesh. Left: the density* ρ*; Right: the averaged slowness* u.

well-shaped function

$$c(x) = \begin{cases} 0.6 & -0.4 < x < 0.4 \\ 1 & \text{else} \end{cases}$$

and the initial data is a delta-function

$$f(x, \xi, 0) = \delta(\xi - w(x)) \tag{5.6}$$

with

$$w(x) = \begin{cases} 0.5, & x \le -1.6; \\ 0.5 - \dfrac{0.4}{(1.6)^2}(x + 1.6)^2, & -1.6 < x \le 0; \\ -0.5 + \dfrac{0.4}{(1.6)^2}(x - 1.6)^2, & 0 < x < 1.6; \\ -0.5, & x \ge 1.6\,. \end{cases} \tag{5.7}$$

Figure 5.2 plots $w(x)$ in dashed lines.

In this example we are interested in the approximation of the density

$$\rho(x, t) = \int f(x, \xi, t) d\xi\,,$$

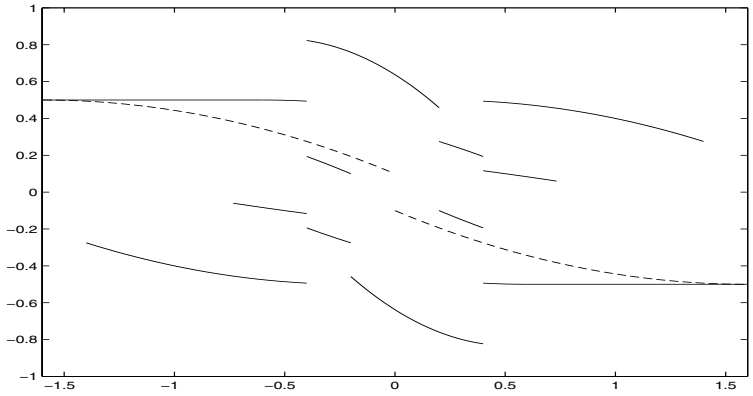


FIG. 5.3. *Example* 5.2, *slowness. Dashed line: the initial slowness* $w(x)$*; Solid line: the slowness at* $t = 1$*. The horizontal axis is the position, the vertical axis is the slowness.*

and the averaged slowness

$$u(x,t) = \frac{\int f(x,\xi,t)\xi d\xi}{\int f(x,\xi,t)d\xi}\,.$$

In the computation, we first approximate the delta function initial data (5.6) by a discrete delta function [16]:

$$\delta_\beta(x) = \begin{cases} \dfrac{1}{\beta}\left(1 - \left|\dfrac{x}{\beta}\right|\right), & |\frac{x}{\beta}| \le 1\,, \\ 0, & |\frac{x}{\beta}| > 1\,. \end{cases} \tag{5.8}$$

If $|\xi_j - w(x_i)| < \beta$, set $f_{ij}^0 = \frac{1}{\beta}\big(1 - |\frac{\xi_j - w(x_i)}{\beta}|\big)$, and $f_{ij}^0 = 0$ otherwise. The choice of the discrete delta function support size β will be made more precise later. We then use the Hamiltonian-preserving scheme to solve the Liouville equation (5.1). Then the moments are recovered by

$$\rho_i^n = \sum_j f_{ij}^n \Delta\xi, \quad u_i^n = \left(\sum_j f_{ij}^n \xi_j \Delta\xi\right) \Big/ \rho_i^n.$$

With partial transmissions and reflections, the exact multivalued slowness at $t = 1$ is depicted as the solid line in Figure 5.3.

In this example the reflection and transmission coefficients α^R, α^T at the wave speed interface are $\alpha^R = \dfrac{1}{16}, \alpha^T = \dfrac{15}{16}$. At $t = 1$, the exact density and averaged

slowness are given by

$$\rho(x,1)=\begin{cases}1, & -1.6<x<-1.4;\\ 1+\alpha^R, & -1.4<x<-0.4-1/3;\\ 1+\alpha^R+0.6\alpha^T, & -0.4-1/3<x<-0.4;\\ 1+\alpha^R+\alpha^T/0.6, & -0.4<x<-0.2;\\ \alpha^T/0.3, & -0.2<x<0.2;\\ 1+\alpha^R+\alpha^T/0.6, & 0.2<x<0.4;\\ 1+\alpha^R+0.6\alpha^T, & 0.4<x<0.4+1/3;\\ 1+\alpha^R, & 0.4+1/3<x<1.4;\\ 1, & 1.4<x<1.6;\end{cases} \tag{5.9}$$

and

$$u(x,1)=\frac{1}{\rho(x,1)}\begin{cases}0.5, & -1.6<x<-1.4;\\ 0.5-\alpha^R\Upsilon(x+0.2), & -1.4<x<-0.4-\frac{1}{3};\\ 0.5-\alpha^R\Upsilon(x+0.2)-0.36\alpha^T\Upsilon(0.6x-1.16), & -0.4-\frac{1}{3}<x<-0.6;\\ \Upsilon(x+0.6)-\alpha^R\Upsilon(x+0.2)-0.36\alpha^T\Upsilon(0.6x-1.16), & -0.6<x<-0.4;\\ \frac{\alpha^T}{0.36}\Upsilon(\frac{x}{0.6}+\frac{13}{15})-\Upsilon(x-1)+\alpha^R\Upsilon(x+1.8), & -0.4<x<-0.2;\\ \frac{\alpha^T}{0.36}\Upsilon(\frac{x}{0.6}+\frac{13}{15})-\frac{\alpha^T}{0.36}\Upsilon(\frac{x}{0.6}-\frac{13}{15}), & -0.2<x<0.2;\\ -\frac{\alpha^T}{0.36}\Upsilon(\frac{x}{0.6}-\frac{13}{15})+\Upsilon(x+1)-\alpha^R\Upsilon(x-1.8), & 0.2<x<0.4;\\ -\Upsilon(x-0.6)+\alpha^R\Upsilon(x-0.2)+0.36\alpha^T\Upsilon(0.6x+1.16), & 0.4<x<0.6;\\ -0.5+\alpha^R\Upsilon(x-0.2)+0.36\alpha^T\Upsilon(0.6x+1.16), & 0.6<x<0.4+\frac{1}{3};\\ -0.5+\alpha^R\Upsilon(x-0.2), & 0.4+\frac{1}{3}<x<1.4;\\ -0.5, & 1.4<x<1.6;\end{cases} \tag{5.10}$$

with $\Upsilon(x)=0.5-\frac{0.4}{(1.6)^2}x^2$.

The time step is chosen as $\Delta t=\frac{1}{2}\Delta\xi$. We will give, respectively, the numerical results computed by the first order Hamiltonian-preserving method and the second order method using the van Leer slope limiter. The choice of β in the first and second order methods are different. In the first order method, we use a linear relation between β and the mesh size $\Delta\xi$: $\beta=\Delta\xi$. In the second order method, this choice does not guarantee the numerical convergence, rather, β must decay to zero slower than $\Delta\xi$. Our numerical experiments indicate that $\beta\sim(\Delta\xi)^{\frac{1}{2}}$ will be appropriate.

Table 5.2 presents the l^1-error of ρ and u computed with several different meshes on the domain $[-1.6,1.6]\times[-1.2,1.2]$ by using the first order method. It can be observed that the l^1-convergence order of the numerical solutions is about 1/2 order. Tables 5.3 and 5.4 present the same errors computed by the second order method with two sets of β's. Clearly, the second order methods give more accurate solutions than the first order method. In comparison between the results by the second order methods with different choices of β, one sees that a smaller β gives more accurate

numerical solutions, but might cause mild oscillations, than a larger one. We do not have a rigorous analysis on the relation between β and $\Delta\xi$ to provide the optimal results by a second order method.

TABLE 5.2
l^1 error of the numerical moments with different meshes $\beta = \Delta\xi$, first order method.

meshes	97 × 80	197 × 160	397× 320	797× 640
ρ	3.3051E-1	2.2438E-1	1.6185E-1	1.1425E-1
u	1.1481E-1	8.4303E-2	6.0016E-2	4.2667E-2

TABLE 5.3
l^1 error of the numerical moments with different meshes $\beta = 5\Delta\xi,\ 7\Delta\xi,\ 10\Delta\xi,\ 14\Delta\xi$ for the four meshes, second order method.

meshes	97 × 80	197 × 160	397× 320	797× 640
ρ	1.8969E-1	9.2800E-2	5.5672E-2	3.3926E-2
u	6.1719E-2	3.1710E-2	1.9006E-2	1.1536E-2

Figure 5.4 shows the numerical solutions of ρ and u using the 797×640 mesh by the first order method along with the exact solutions. The numerical solution captures the correct dynamics and discontinuities, but the resolution is poor even on such a fine mesh. In contrast, Figure 5.5 shows the computed densities ρ using the 797×640 mesh by the second order method with different β's. The results have much higher resolution across the discontinuities than the first order method. However, the numerical density by using $\beta = 14\Delta\xi$ exhibits some small oscillations near the discontinuities between, while the use of a larger $\beta = 42\Delta\xi$ creates no oscillations at the expense of a slight accuracy or resolution loss.

These results show that although the second order method can give more accurate solutions than the first order method, there is a support size parameter β that needs to be properly chosen in order to compromise between convergence and accuracy of the numerical solution. It is not clear how to choose β *a priori*. In the future we will study the feasibility of introducing the decomposition technique proposed in [25] into such a problem with measure-valued data, which could avoid such an inconvenience as well as improve the numerical accuracy and resolution.

Example 5.3. Computing the physical observables of a 2D problem with a L^∞ solution. Consider the 2D Liouville equation (4.1) with a discontinuous wave speed

$$c(x,y) = \begin{cases} 2 & y > 0 \\ 1 & y < 0 \end{cases}$$

and a smooth initial data

$$f(x,y,\xi,\eta,0) = \frac{1}{\pi c_3 c_4} \exp\left(-\left(\frac{x}{c_1}\right)^2 - \left(\frac{y+0.1}{c_2}\right)^2 - \left(\frac{\xi}{c_3}\right)^2 - \left(\frac{\eta-0.1}{c_4}\right)^2 \right),$$

where $c_1 = 0.03, c_3 = 0.05, c_2 = c_4 = 0.025$.

In this example we aim at computing the density which is the zeroth moment of the density distribution

$$\rho(x,y,t) = \int\int f(x,y,\xi,\eta,t) d\xi d\eta. \tag{5.11}$$

TABLE 5.4
l^1 error of the numerical moments with different meshes $\beta = 15\Delta\xi,\ 21\Delta\xi,\ 30\Delta\xi,\ 42\Delta\xi$ for the four meshes, second order method.

meshes	97×80	197×160	397×320	797×640
ρ	4.3791E-1	2.0464E-1	9.0273E-2	3.7545E-2
u	1.3585E-1	6.0953E-2	2.9188E-2	1.2857E-2

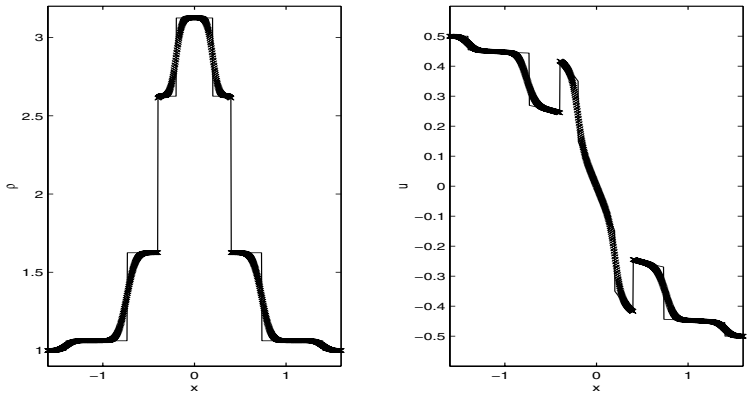


FIG. 5.4. *Example* 5.2*, density ρ and averaged slowness u at $t = 1$. Solid line: the exact solution; "x": numerical solutions by first order method using the* 797×640 *mesh. Left: ρ; Right: u.*

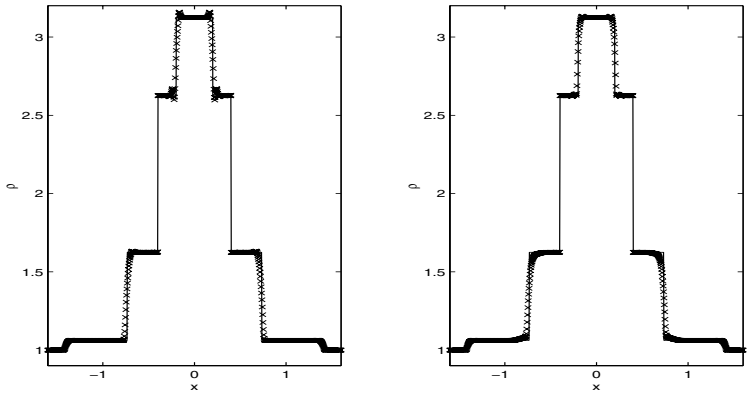


FIG. 5.5. *Example* 5.2*, density ρ at $t = 1$. Solid line: the exact solution; "x": numerical solutions by second order method using the* 797×640 *mesh. Left: $\beta = 14\Delta\xi$; Right: $\beta = 42\Delta\xi$.*

The computational domain is chosen to be $[x, y, \xi, \eta] \in [-0.12, 0.12] \times [-0.2, 0.2] \times [-0.2, 0.2] \times [-0.2, 0.2]$.

The reflection and transmission coefficients α^R, α^T at the interface are given by (2.4). The "exact" solution of ρ is obtained by first solving for $f(x, y, \xi, \eta, t)$ analytically, and then evaluating the integral (5.11) on a very fine mesh in the (ξ, η) space.

The time step is chosen as $\Delta t = \frac{1}{3}\Delta x$. Figures 5.6 and 5.8 show, respectively, the numerical density ρ at $t = 0.12, 0.15$ using different meshes along with the exact solution. Figures 5.7 and 5.9 show, respectively, the numerical density ρ on $x = 0$ at

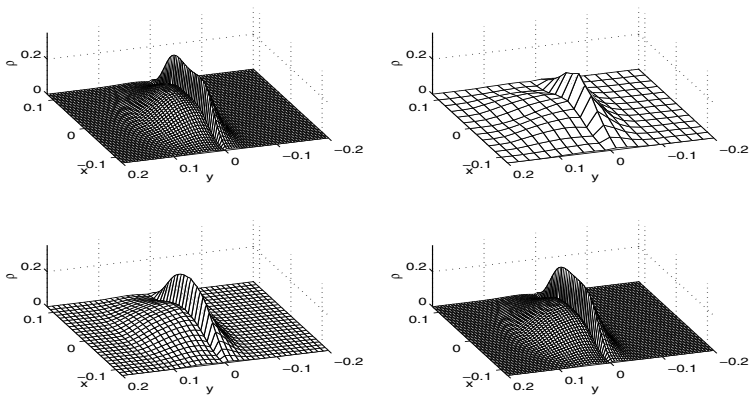


FIG. 5.6. *Example* 5.3, *density* ρ *at* $t = 0.12$. *Upper left: the exact solution; Upper right:* $13 \times 20 \times 14^2$ *mesh; Lower left:* $25 \times 40 \times 26^2$ *mesh; Lower right:* $49 \times 80 \times 50^2$ *mesh.*

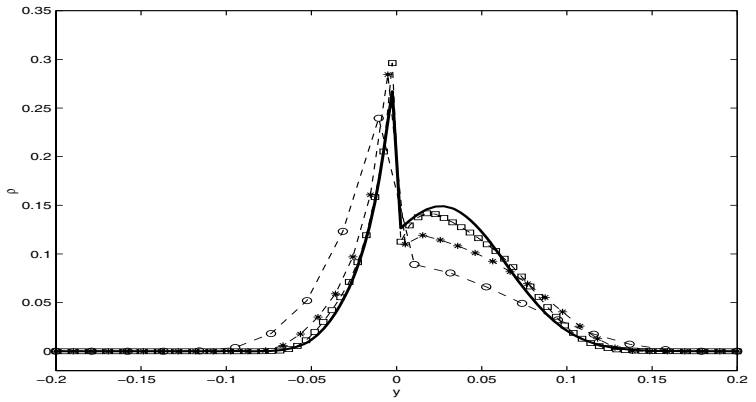


FIG. 5.7. *Example* 5.3, *density* ρ *along* $x = 0$ *at* $t = 0.12$. *Solid line: exact solution; "o":* $13 \times 20 \times 14^2$ *mesh; "*":* $25 \times 40 \times 26^2$ *mesh; "* $\square$ *":* $49 \times 80 \times 50^2$ *mesh.*

$t = 0.12, 0.15$ using different meshes along with the exact solution.

Table 5.5 presents the l^1 errors of ρ computed with different meshes in phase space at $t = 0.12, 0.15$. The convergence rate is slightly higher than first order, which does not suffer from the accuracy degeneration of an usual finite difference method for solving the discontinuous solution of a linear hyperbolic equation—which is at most 1/2 order stated by the well-established theory [30, 42]. This is because the only discontinuity in the solutions is at the interface, which has been taken care of by the Hamiltonian-preserving mechanism, and no linear discontinuity travels to the downstream direction like in the 1D case.

TABLE 5.5
l^1 *error of* ρ *using different meshes.*

meshes	$13 \times 20 \times 14^2$	$25 \times 40 \times 26^2$	$49 \times 80 \times 50^2$
$t = 0.12$	1.241556E-3	5.252852E-4	1.722251E-4
$t = 0.15$	1.244387E-3	6.621391E-4	2.617174E-4

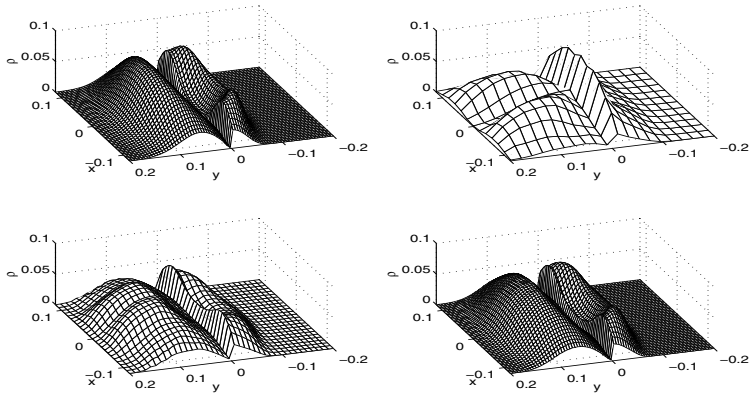


FIG. 5.8. *Example* 5.3, *density* ρ *at* $t = 0.15$. *Upper left: exact solution; Upper right:* $13 \times 20 \times 14^2$ *mesh; Lower left:* $25 \times 40 \times 26^2$ *mesh; Lower right:* $49 \times 80 \times 50^2$ *mesh.*

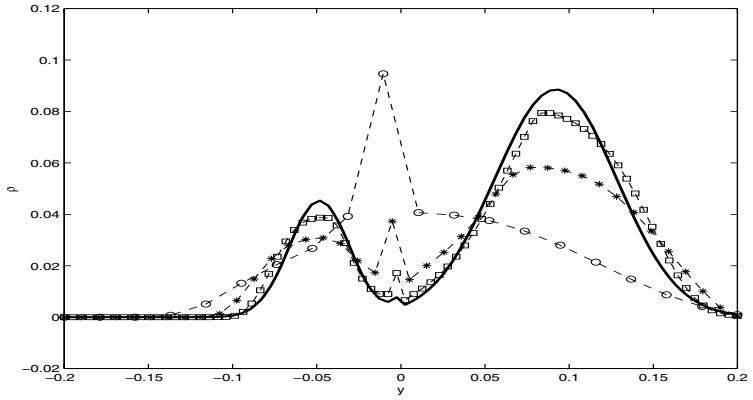


FIG. 5.9. *Example* 5.3, *density* ρ *along* $x = 0$ *at* $t = 0.15$. *Solid line: exact solution; "o":* $13 \times 20 \times 14^2$ *mesh; "*":* $25 \times 40 \times 26^2$ *mesh; "* $\square$ *":* $49 \times 80 \times 50^2$ *mesh.*

6. Conclusion. In this paper, we extended our previous work [28] to the Liouville equation of geometrical optics with partial transmissions and reflections. Such problems arise in geometrical optics through inhomogeneous media. While still utilizing the constant Hamiltonian structure in constructing the numerical flux, we also account for the transmission and reflection coefficients in the numerical flux. By doing so, the numerical flux automatically absorbs the interface condition. This gives an explicit scheme for the time dependent Liouville equation with discontinuous indices of refraction that can capture correctly the partial transmissions and reflections across the interface. This scheme is subject to a hyperbolic CFL condition, under which the scheme is positive, and stable in both l^1 and l^∞ norms. Numerical experiments are carried out to study the numerical accuracy.

We only extended a finite difference version of the Hamiltonian-preserving scheme developed in [28]. The finite volume version of the method in [28] can also be extended in a similar fashion, but will not be given here.

In the future we will consider analytical issues such as the well-posedness of the problem in a more general context than that discussed in this paper, and the convergence of the numerical scheme. We will also investigate its applications to more

complex interfaces, and develop more effective methods for the measure-valued initial value problem for the same equation.

Acknowledgement. We thank an anonymous referee for his/her valuable comments and suggestions.

REFERENCES

[1] G. Bal, J. B. Keller, G. Papanicolaou, and L. Ryzhik, *Transport theory for acoustic waves with reflection and transmission at interfaces*, Wave Motion, 30 (1999), pp. 303–327.

[2] N. Ben Abdallah, P. Degond, and I. M. Gamba, *Coupling one-dimensional time-dependent classical and quantum transport models*, J. Math. Phys., 43 (2002), pp. 1–24.

[3] J.-D. Benamou, *Big ray tracing: Multivalued travel time field computation using viscosity solutions of the Eikonal equation*, J. Comput. Phys., 128 (1996), pp. 463–474.

[4] J.-D. Benamou, *Direct computation of multivalued phase space solutions for Hamilton-Jacobi equations*, Comm. Pure Appl. Math., 52 (1999), pp. 1443–1475.

[5] J.-D. Benamou, *An introduction to Eulerian geometrical optics* (1992–2002), J. Sci. Comput., 19 (2003), pp. 63–93.

[6] J.-D. Benamou and I. Solliec, *An Eulerian method for capturing caustics*, J. Comput. Phys., 162 (2000), pp. 132–163.

[7] F. Bouchut, *Renormalized solutions to the Vlasov equation with coefficients of bounded variations*, Arch. Ration. Mech. Anal., 157 (2001), pp. 75–90.

[8] F. Bouchut and L. Desvillettes, *On two-dimensional Hamiltonian transport equations with continuous coefficients*, Differential Integral Equations, 14 (2001), pp. 1015–1024.

[9] L.-T. Cheng, M. Kang, S. Osher, H. Shim, and Y.-H. Tsai, *Reflection in a level set framework for geometric optics*, CMES Comput. Model. Eng. Sci., 5 (2004), pp. 347–360.

[10] L.-T. Cheng, H. Liu, and S. Osher, *Computational high-frequency wave propagation using the level set method, with applications to the semi-classical limit of Schrödinger equations*, Commun. Math. Sci., 1 (2003), pp. 593–621.

[11] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[12] B. Engquist and O. Runborg, *Multi-phase computations in geometrical optics*, J. Comput. Appl. Math., 74 (1996), pp. 175–192.

[13] B. Engquist and O. Runborg, *Multiphase computations in geometrical optics*, in Hyperbolic Problems: Theory, Numerics, Applications, Internat. Ser. Numer. Math. 129, M. Fey and R. Jeltsch, eds., ETH Zentrum, Zürich, Switzerland, 1998, Birkhauser-Verlag, Basel, 1999.

[14] B. Engquist and O. Runborg, *Computational high frequency wave propagation*, Acta Numer., 12 (2003), pp. 181–266.

[15] B. Engquist, O. Runborg, and A.-K. Tornberg, *High-frequency wave propagation by the segment projection method*, J. Comput. Phys., 178 (2002), pp. 373–390.

[16] B. Engquist, A.-K. Tornberg, and R. Tsai, *Discretization of Dirac delta functions in level set methods*, J. Comput. Phys., 207 (2005), pp. 28–51.

[17] E. Fatemi, B. Engquist, and S. Osher, *Numerical solution of the high frequency asymptotic expansion for the scalar wave equation*, J. Comput. Phys., 120 (1995), pp. 145–155.

[18] S. Fomel and J. A. Sethian, *Fast-phase space computation of multiple arrivals*, Proc. Natl. Acad. Sci., 99 (2002), pp. 7329–7334.

[19] L. Gosse, *Using K-branch entropy solutions for multivalued geometric optics computations*, J. Comput. Phys., 180 (2002), pp. 155–182.

[20] L. Gosse, *Multiphase semiclassical approximation of an electron in a one-dimensional crystalline lattice.* II. *Impurities, confinement, and Bloch oscillations*, J. Comput. Phys., 201 (2004), pp. 344–375.

[21] L. Gosse, S. Jin, and X. T. Li, *On two moment systems for computing multiphase semiclassical limits of the Schrödinger equation*, Math. Model Methods Appl. Sci., 13 (2003), pp. 1689–1723.

[22] M. Hauray, *On Liouville transport equation with force field in BV_{loc}*, Comm. Partial Differential Equations, 29 (2004), pp. 207–217.

[23] M. Hauray, *On two-dimensional Hamiltonian transport equations with L^p_{loc} coefficients*, Ann. Inst. H. Poincare Anal. Non Linéaire, 20 (2003), pp. 625–644.

[24] S. Jin, H. Liu, S. Osher, and Y.-S.R. Tsai, *Computing multivalued physical observables for the semiclassical limit of the Schrödinger equation*, J. Comput. Phys., 205 (2005), pp. 222–241.

[25] S. Jin, H. Liu, S. Osher, and Y.-S.R. Tsai, *Computing multi-valued physical observables for high frequency limit of symmetric hyperbolic systems*, J. Comput. Phys., 210 (2005), pp. 497–518.

[26] S. Jin and S. Osher, *A level set method for the computation of multivalued solutions to quasi-linear hyperbolic PDEs and Hamilton-Jacobi equations*, Commun. Math. Sci., 1 (2003), pp. 575–591.

[27] S. Jin and X. Wen, *Hamiltonian-preserving schemes for the Liouville equation with discontinuous potentials*, Commun. Math. Sci., 3 (2005), pp. 285–315.

[28] S. Jin and X. Wen, *Hamiltonian-preserving schemes for the Liouville equation of geometrical optics with discontinuous local wave speeds*, J. Comput. Phys. 214 (2006), pp. 672–697.

[29] S. Jin and X. Wen, *The l^1-stability of a Hamiltonian-preserving scheme for the Liouville equation with discontinuous potentials*, Math. Comp., submitted.

[30] N. N. Kuznetsov, *On stable methods for solving nonlinear first order partial differential equations in the class of discontinuous functions*, Topics in Numerical Analysis, III, Proc. Roy. Irish Acad. Conf., J. J. H. Miller, ed., Academic Press, London, 1977, pp. 183–197.

[31] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser-Verlag, Basel, 1990.

[32] R. J. LeVeque and C. Zhang, *The immersed interface method for acoustic wave equations with discontinuous coefficients*, Wave Motion, 25 (1997), pp. 237–263.

[33] L. Miller, *Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary*, J. Math. Pures Appl. (9), 79 (2000), pp. 227–269.

[34] S. Osher, L.-T. Cheng, M. Kang, H. Shim, and Y.-H. Tsai, *Geometric optics in a phase-space-based level set and Eulerian framework*, J. Comput. Phys., 179 (2002), pp. 622–648.

[35] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130.

[36] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.

[37] O. Runborg, *Some new results in multiphase geometrical optics*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1203–1231.

[38] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.

[39] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for waves in a half space*, Comm. Partial Differential Equations, 22 (1997), pp. 1869–1910.

[40] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock capturing scheme*, J. Comput. Phys., 77 (1988), pp. 439–471.

[41] W. W. Symes and J. Qian, *A slowness matching Eulerian method for multivalued solutions of Eikonal equations*, J. Sci. Comput., 19 (2003), pp. 501–526. Special issue in honor of the sixtieth birthday of Stanley Osher.

[42] T. Tang and Z. H. Teng, *The sharpness of Kuznetsov's $O(\sqrt{\Delta x})$ L^1-error estimate for monotone difference schemes*, Math. Comp., 64 (1995), pp. 581–589.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1829–1853

NUMERICAL ANALYSIS OF CONVECTION-DIFFUSION-REACTION PROBLEMS WITH HIGHER ORDER CHARACTERISTICS/FINITE ELEMENTS. PART I: TIME DISCRETIZATION*

ALFREDO BERMÚDEZ†, MARIA R. NOGUEIRAS†, AND CARLOS VÁZQUEZ‡

Abstract. In this paper a higher order characteristics time discretization scheme is analyzed for a variable coefficient convection-(possibly degenerate)diffusion-reaction equation with mixed Dirichlet–Robin boundary conditions. First, the proposed second order time discretization scheme is rigorously introduced for exact and approximate characteristics. Next, under not very restrictive hypotheses on the data, the $l^\infty(L^2)$ stability is proved and $l^\infty(L^2)$ error estimates of order $O(\Delta t^2)$ are obtained. Lagrange–Galerkin schemes using different finite elements spaces will be analyzed in the second part of this work [to appear in *SIAM J. Numer. Anal.*], where quadrature formulas are proposed for practical implementation.

Key words. convection-diffusion-reaction equation, characteristics method, stability, error estimates, second order schemes

AMS subject classifications. 65M12, 65M25

DOI. 10.1137/040612014

1. Introduction.

Linear convection-diffusion-reaction equations arise in the mathematical modeling of many important problems from different fields of engineering and applied sciences, such as thermodynamics, fluid mechanics, and finance (see [18], for example). In many cases the diffusive term is smaller than the convective one, giving rise to the so-called convection dominated problems (see [15]). Furthermore, in some convection dominated cases, the diffusive term becomes degenerate, as in some financial models (see, for instance, [26]).

In the framework of numerical solutions of convection dominated problems, a possible strategy is provided by the method of characteristics for time discretization (see the review paper [15]). This approach is based on the discretization of the total (or material) derivative, i.e., the time derivative along the characteristics lines of the convective part of the equation. Many authors have mathematically analyzed and applied the characteristics method to different problems. In [13] and [21], error estimates for the combination of a first order characteristics scheme with both finite differences and classical Lagrange finite elements have been given for time dependent convection-diffusion problems. Its adaptation to steady state convection-diffusion equations has been developed in [8] and its application to Navier–Stokes equations proposed in [21] (see also [7]). The use of standard quadrature formulas to compute the terms appearing in the formulation leads to conditional stability, as opposed to the unconditionally stable exact integrated schemes. This aspect has been studied for

*Received by the editors July 20, 2004; accepted for publication (in revised form) April 17, 2006; published electronically September 29, 2006. This work was partially supported by project VEM2003-20069-C03-03 of MCYT.

http://www.siam.org/journals/sinum/44-5/61201.html

†Dep. de Matemática Aplicada, Universidade de Santiago, Campus Sur s/n, 15706 Santiago, Spain (mabermud@usc.es, marianog@usc.es). The second author was supported by Ministerio de Educacion, Cultura y Deporte.

‡Dep. de Matemáticas, Universidade da Coruña, Campus Elviña s/n, 15071-A Coruña, Spain (carlosv@udc.es).

first order schemes in [19] and [25]. More recently, the combination of the classical first order scheme with discontinuous Galerkin methods has been analyzed in [4, 3, 5].

The present paper falls into the frame of higher order Lagrange–Galerkin methods. Increasing the order of time and space approximations can be obtained by using higher order schemes for the discretization of the material derivative and higher order finite element spaces. In [14] multistep Galerkin methods for constant coefficient convection-diffusion problems are studied and the need for analyzing the variable coefficient case is pointed out. Also, in [11, 12] multistep methods for approximating the material time derivative, combined with either mixed finite element or spectral methods, are proposed to solve incompressible Navier–Stokes equations. Stability is proved and optimal error estimates for the fully discretized problem obtained. More recently, in [24], a second order Runge–Kutta method is proposed to approximate the material time derivative when solving a constant coefficient convection-diffusion equation with Dirichlet boundary conditions. Second order in time is maintained by the Crank–Nicolson scheme and an adequate upwinding of the diffusive term.

The present paper extends [24] in four aspects: first, it deals with a (possibly degenerate) variable coefficient diffusive term instead of the simpler Laplacian one. Second, nonzero reaction functions are allowed. Third, a general mixed Dirichlet–Robin boundary condition is considered. Fourth, non-divergence-free velocity fields are handled. While the first two extensions are quite straightforward because we still assume the solutions are smooth, dealing with boundary conditions other than Dirichlet requires nonstandard Green's formulas. Similarly, the fact that the velocity field is not divergence-free makes it necessary to introduce some new terms in the weak formulation. Moreover, although the analysis is given only for velocity fields null on the boundary, the stated Green's formulas allow us to write the weak formulation of problems for which this assumption is not satisfied and, thereby, to write second order schemes for their numerical solution. Actually, these schemes have been successfully applied to some interesting applications in mathematical finance (see [10]).

In this wider setting, the mathematical formalism of continuum mechanics (see, for instance, [16]) is used to express the results and notation related not only to the approximate characteristics proposed in [24] but also to the exact ones. The latter rather than the former can be used in some particularly interesting applications (see [10]), so our analysis also includes, as a novelty, the case where the characteristic lines are exactly computed. Moreover, by using Taylor expansions we rigorously justify some approximations of the characteristics, velocity gradients and their determinants and inverses. A technical proof of an $l^\infty(L^2)$ stability inequality is developed so that it can be appropriately used to obtain $l^\infty(L^2)$ error estimates of order $O(\Delta t^2)$ between the solutions of the time discretized problem and the continuous one. The fully discretized Lagrange–Galerkin scheme with different finite element spaces is analyzed in part two of this work (see [9]). Moreover, in [9] adequate quadrature formulas are proposed for different triangular and quadrangular finite element spaces.

The present paper is organized as follows. In section 2 the strong formulation of the convection-diffusion-reaction problem is posed and notation concerning functional spaces is introduced. In section 3 the characteristics curves associated to the velocity field are rigorously defined and useful results and notation related to them are stated. It is worth mentioning that notation and results are inspired in those handled by classical continuum mechanics textbooks (see, for instance, [16]). In section 4 the variational formulation of the problem is posed. In section 5, first the second order time discretization scheme is introduced for both exact and second order approximate characteristics. Next, under suitable but not very restrictive hypotheses on data

functions, the $l^\infty(L^2)$ stability result is proved for small enough time step. Finally, assuming greater regularity on the data functions, $l^\infty(L^2)$ error estimates of order $O(\Delta t^2)$ for the solution of the time discretized problem are derived.

2. Statement of the problem and functional spaces. Let Ω be a bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$ with Lipschitz boundary, Γ, divided into two parts: $\Gamma = \Gamma_D \cup \Gamma_R$, with $\Gamma_D \cap \Gamma_R = \emptyset$. Let T be a positive constant. We consider the following initial boundary value problem.

(SP) STRONG PROBLEM. *Find a function* $\phi : \Omega \times (0,T) \longrightarrow \mathbb{R}$ *such that*

$$\phi'(\mathbf{x},t) - \operatorname{div}(\mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x},t)) + \mathbf{v}(\mathbf{x},t)\cdot\nabla\phi(\mathbf{x},t) + r(\mathbf{x})\phi(\mathbf{x},t) = f(\mathbf{x},t) \tag{2.1}$$

for $(\mathbf{x},t) \in \Omega \times (0,T)$*, subject to boundary conditions*

$$\phi(\mathbf{x},t) = 0 \quad \text{on } \Gamma_D \times (0,T), \tag{2.2}$$

$$\alpha\ \phi(\mathbf{x},t) + \mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = g(\mathbf{x},t) \quad \text{on } \Gamma_R \times (0,T) \tag{2.3}$$

and initial condition

$$\phi(\mathbf{x},0) = \phi^0(\mathbf{x}) \quad \text{in } \Omega. \tag{2.4}$$

In the above equations, ϕ' denotes the partial derivative with respect to t, $\mathbf{A} : \overline{\Omega} \longrightarrow \mathcal{S}_d$ denotes the diffusion matrix function, where $\mathcal{S}_d$ is the space of symmetric $d \times d$ matrices, $\mathbf{v} : \overline{\Omega} \times [0,T] \longrightarrow \mathbb{R}^d$ is the convection vector field, $r : \overline{\Omega} \longrightarrow \mathbb{R}$ is the reaction function, $f : \overline{\Omega} \times [0,T] \longrightarrow \mathbb{R}$ and $g : \Gamma_R \times [0,T] \longrightarrow \mathbb{R}$ are given scalar functions, and $\mathbf{n}$ is the outward unit normal vector to Γ.

Let us introduce the Lebesgue spaces $L^p(\Omega)$ and the Sobolev spaces $W^{m,p}(\Omega)$ for $p = 1, 2, \ldots, \infty$ and m an integer. For the particular case $p = 2$, we consider the Hilbert space $L^2(\Omega)$ with the usual inner product $\langle\ ,\ \rangle$, which induces the norm $\|\ \|_0$, and spaces $H^m(\Omega) = W^{m,2}(\Omega)$ equipped with the usual norms $\|\ \|_m$ (see [1] for details). Moreover, we denote by $H^1_{\Gamma_D}(\Omega)$ the closed subspace of $H^1(\Omega)$ defined by

$$H^1_{\Gamma_D}(\Omega) := \{\varphi \in H^1(\Omega),\ \varphi|_{\Gamma_D} = 0\}.$$

For a Banach space X and an integer m, spaces $C^m([0,T],X)$ and $H^m((0,T),X)$ will be abbreviated as $C^m(X)$ and $H^m(X)$, respectively, and endowed with norms

$$\|\varphi\|_{C^m(X)} := \max_{t\in[0,T]}\left\{\max_{j=0,\ldots,m}\|\varphi^{(j)}(t)\|_X\right\},\ \|\varphi\|_{H^m(X)} := \left(\int_0^T \sum_{j=0}^m \left\|\varphi^{(j)}(t)\right\|_X^2 dt\right)^{\frac{1}{2}}.$$

In the above definitions, $\varphi^{(j)}$ denotes the jth derivative of φ with respect to time.

Next, we introduce the Banach space $Z^m = \{\varphi \in C^j(H^{m-j}(\Omega));\ j = 0, \ldots, m\}$ for $m \in Z^+$, equipped with the norm $\|\varphi\|_{Z^m} := \max\{\|\varphi\|_{C^j(H^{m-j})};\ 0 \le j \le m\}$. Similar spaces are considered for the boundary sets Γ_R and Γ_D. For example, for Γ_R we use the notation $\|\cdot\|_{m,\Gamma_R}$, $\|\cdot\|_{Z^m,\Gamma_R}$, etc.

3. Characteristic curves. In this section we define the characteristic lines associated with vector field $\mathbf{v}$ and recall some classical properties satisfied by them.

Thus, for given $(\mathbf{x},t) \in \overline{\Omega} \times [0,T]$, the characteristic line through $(\mathbf{x},t)$ is the vector function $X_e(\mathbf{x},t;\cdot)$ solving the initial value problem

$$\frac{\partial X_e}{\partial\tau}(\mathbf{x},t;\tau) = \mathbf{v}(X_e(\mathbf{x},t;\tau),\tau), \quad X_e(\mathbf{x},t;t) = \mathbf{x}. \tag{3.1}$$

It represents the trajectory described by a material point that is placed at position $\mathbf{x}$ at time t and is driven by the velocity field $\mathbf{v}$. If $\mathbf{v} \in C^0(\overline{\Omega} \times [0,T])$, it is Lipschitz continuous with respect to the first variable and vanishes on Γ; then the characteristic line solving (3.1) is well defined in the whole domain, $[0,T]$, and it is unique for each initial condition $(\mathbf{x},t)$. In this case, as a function on $(\mathbf{x},t;\tau)$, it is Lipschitz continuous in $\Omega \times [0,T] \times [0,T]$. Moreover (see [22]), if $\tau_1,\tau_2,\tau_3 \in [0,T]$ and $\mathbf{x} \in \Omega$, then

$$X_e(\mathbf{x},\tau_1;\tau_3) = X_e(X_e(\mathbf{x},\tau_1;\tau_2),\tau_2;\tau_3). \tag{3.2}$$

Indeed, by replacing $\tau_1 = \tau_3$ we deduce that the mapping $X_e(\cdot,\tau_1;\tau_2) : \overline{\Omega} \longrightarrow \overline{\Omega}$ is one-to-one, with inverse $X_e(\cdot,\tau_2;\tau_1)$.

Next, assuming they exist, we denote by $\mathbf{F}_e$ (respectively, by $\mathbf{L}$) the gradient of X_e (respectively, of $\mathbf{v}$) with respect to the space variable $\mathbf{x}$, i.e.,

$$(F_e)_{rs}(\mathbf{x},t;\tau) := \frac{\partial (X_e)_r}{\partial x_s}(\mathbf{x},t;\tau), \qquad L_{rs}(\mathbf{x},t) := \frac{\partial v_r}{\partial x_s}(\mathbf{x},t).$$

PROPOSITION 3.1. *If $\mathbf{v} \in C^0(C^n(\overline{\Omega}))$ for $n \geq 1$ an integer, then $X_e \in C^0(\overline{\Omega} \times [0,T] \times [0,T])$ and it is C^n with respect to the $\mathbf{x}$ variable.*

Proof. See, for instance, [2]. ∎

It will be useful to compute second order approximations of matrices $\mathbf{F}_e$ and $\mathbf{F}_e^{-1}$. For this, we need the following equations (see [16]):

$$\frac{\partial \mathbf{F}_e}{\partial \tau}(\mathbf{x},t;\tau) = \mathbf{L}(X_e(\mathbf{x},t;\tau),\tau)\mathbf{F}_e(\mathbf{x},t;\tau), \tag{3.3}$$

$$\frac{\partial^2 \mathbf{F}_e}{\partial \tau^2}(\mathbf{x},t;\tau) = \nabla \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{L}\mathbf{v} \right)(X_e(\mathbf{x},t;\tau),\tau)\mathbf{F}_e(\mathbf{x},t;\tau). \tag{3.4}$$

By using Gronwall's lemma and (3.3) we can prove the following result.

PROPOSITION 3.2. *If $\mathbf{v} \in C^0(C^1(\overline{\Omega}))$, then*

$$\|\mathbf{F}_e(\mathbf{x},t;\tau)\| \leq e^{\|\mathbf{v}\|_{C^0(C^1(\overline{\Omega}))}|\tau - t|} \quad \forall \mathbf{x} \in \Omega,\ t,\tau \in [0,T]. \tag{3.5}$$

PROPOSITION 3.3. *If $\mathbf{v} \in C^0(C^2(\overline{\Omega})) \cap C^1(C^1(\overline{\Omega}))$, then $\mathbf{F}_e$ satisfies the Taylor expansions*

$$\begin{aligned} \mathbf{F}_e(\mathbf{x},t;s) &= \mathbf{I} + (s-t)\ \mathbf{L}(\mathbf{x},t) \\ &\quad + \int_s^t (\tau - s)\ \nabla \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{L}\mathbf{v} \right)(X_e(\mathbf{x},t;\tau),\tau)\mathbf{F}_e(\mathbf{x},t;\tau)\ d\tau, \end{aligned} \tag{3.6}$$

and its inverse, $\mathbf{F}_e^{-1}$, satisfies the Liouville theorem

$$\begin{aligned} \mathbf{F}_e^{-1}(\mathbf{x},t;s) &= \mathbf{I} + (t-s)\ \mathbf{L}(X_e(\mathbf{x},t;s),s) \\ &\quad - \int_s^t (\tau - t)\ \nabla \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{L}\mathbf{v} \right)(X_e(\mathbf{x},t;\tau),\tau)\mathbf{F}_e(X_e(\mathbf{x},t;s),s;\tau)\ d\tau. \end{aligned} \tag{3.7}$$

Proof. Expression (3.6) derives from the Taylor expansion, (3.3) and (3.4). To prove (3.7) we differentiate (3.2) for $\tau_1 = \tau_3 = s$ and $\tau_2 = t$ to obtain $\mathbf{F}_e^{-1}(\mathbf{x},t;s) = \mathbf{F}_e(X_e(\mathbf{x},t;s),s;t)$. Then, we use (3.6) and $X_e(X_e(\mathbf{x},t;s),s;\tau) = X_e(\mathbf{x},t;\tau)$. ∎

Now, we develop analogous computations for $\mathsf{det}\ \mathbf{F}_e^{-1}$. To do that, we first use Liouville's theorem (see [2]) and the chain rule, obtaining

$$\frac{\partial}{\partial \tau}\mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau) = -\mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\ \mathrm{div}\ \mathbf{v}(X_e(\mathbf{x},t;\tau),\tau). \tag{3.8}$$

Differentiating (3.8) again we get

$$\frac{\partial^2}{\partial\tau^2}\ \mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau) = \mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\left((\mathrm{div}\ \mathbf{v})^2(X_e(\mathbf{x},t;\tau),\tau)\right.$$
$$\left.\mathrm{div}\left(\frac{\partial\mathbf{v}}{\partial t}+\mathbf{L}\mathbf{v}\right)(X_e(\mathbf{x},t;\tau),\tau)-\left(\mathbf{L}\cdot\mathbf{L}^T\right)(X_e(\mathbf{x},t;\tau),\tau)\right). \tag{3.9}$$

The following propositions can be easily proved.

PROPOSITION 3.4. *If* $\mathbf{v}\in C^0(C^1(\overline{\Omega}))$, *then*

$$\mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\le e^{\|\mathbf{v}\|_{C^0(C^1(\overline{\Omega}))}|\tau-t|}\quad\forall\mathbf{x}\in\Omega,\ t,\tau\in[0,T]. \tag{3.10}$$

PROPOSITION 3.5. *If* $\mathbf{v}\in C^0(C^2(\Omega))\cap C^1(C^1(\overline{\Omega}))$, *then* $\mathsf{det}\ \mathbf{F}_e^{-1}$ *satisfies*

$$\mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;s)=1-(s-t)\ \mathrm{div}\ \mathbf{v}(\mathbf{x},t)+\int_s^t(\tau-s)\ \frac{\partial^2}{\partial\tau^2}\left(\mathsf{det}\ \mathbf{F}_e^{-1}\right)(\mathbf{x},t;\tau)\ d\tau. \tag{3.11}$$

4. Variational formulation. We are going to develop some formal computations in order to give a variational formulation of problem (SP). From the definition of the characteristic curves and by using the chain rule, it follows that

$$\frac{\mathsf{d}\phi}{\mathsf{d}\tau}(X_e(\mathbf{x},t;\tau),\tau)=\phi'(X_e(\mathbf{x},t;\tau),\tau)+\mathbf{v}(X_e(\mathbf{x},t;\tau),\tau)\cdot\nabla\phi(X_e(\mathbf{x},t;\tau),\tau). \tag{4.1}$$

By writing equation (2.1) at point $X_e(\mathbf{x},t;\tau)$ and time τ, and using (4.1), we have

$$\frac{\mathsf{d}\phi}{\mathsf{d}t}(X_e(\mathbf{x},t;\tau),\tau)-\mathrm{div}(\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau),\tau))$$
$$+r(X_e(\mathbf{x},t;\tau))\phi(X_e(\mathbf{x},t;\tau),\tau)=f(X_e(\mathbf{x},t;\tau),\tau). \tag{4.2}$$

Before giving a weak formulation of (4.2), we state two lemmas. The first one can be considered as a Green's formula.

LEMMA 4.1. *Let* $X:\overline{\Omega}\longrightarrow\overline{X(\Omega)}$, $X\in C^2(\overline{\Omega})$, *be an invertible vector valued function. Let* $\mathbf{F}=\nabla X$ *and assume that* $\mathbf{F}^{-1}\in C^1(\overline{\Omega})$. *Then*

$$\int_\Omega \mathrm{div}\ \mathbf{w}(X(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}=\int_\Gamma \mathbf{F}^{-T}(\mathbf{x})\mathbf{n}(\mathbf{x})\cdot\mathbf{w}(X(\mathbf{x}))\ \psi(\mathbf{x})\ dA_{\mathbf{x}}$$
$$-\int_\Omega \mathbf{F}^{-1}(\mathbf{x})\mathbf{w}(X(\mathbf{x}))\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x}-\int_\Omega \mathrm{div}\ \mathbf{F}^{-T}(\mathbf{x})\cdot\mathbf{w}(X(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}, \tag{4.3}$$

with $\mathbf{w}\in H^1(X(\Omega))$ *a vector valued function and* $\psi\in H^1(\Omega)$ *a scalar function.*

Proof. First, by the Gauss theorem, we have

$$\int_\Gamma \mathbf{F}^{-T}(\mathbf{x})\mathbf{n}(\mathbf{x})\cdot\mathbf{w}(X(\mathbf{x}))\ \psi(\mathbf{x})\ dA_{\mathbf{x}}=\int_\Omega \mathrm{div}\left(\mathbf{F}^{-1}(\mathbf{w}\circ X)\psi\right)(\mathbf{x})\ d\mathbf{x}. \tag{4.4}$$

Finally, identity (4.3) is obtained by developing the divergence term in (4.4) with

$$\mathrm{div}\left(\mathbf{F}^{-1}(\mathbf{w}\circ X)\psi\right)(\mathbf{x})=\psi(\mathbf{x})\ \mathrm{div}\left(\mathbf{F}^{-1}(\mathbf{w}\circ X)\right)(\mathbf{x})+\nabla\psi(\mathbf{x})\cdot\mathbf{F}^{-1}(\mathbf{x})\mathbf{w}(X(\mathbf{x})),$$
$$\mathrm{div}\left(\mathbf{F}^{-1}(\mathbf{w}\circ X)\right)(\mathbf{x})=\mathbf{F}^{-T}(\mathbf{x})\cdot\nabla\left(\mathbf{w}\circ X\right)(\mathbf{x})+\mathbf{w}(X(\mathbf{x}))\cdot\mathrm{div}\ \mathbf{F}^{-T}(\mathbf{x}),$$
$$\mathbf{F}^{-T}(\mathbf{x})\cdot\nabla\left(\mathbf{w}\circ X\right)(\mathbf{x})=\mathsf{tr}\left(\nabla\left(\mathbf{w}\circ X\right)(\mathbf{x})\mathbf{F}^{-1}(\mathbf{x})\right)$$
$$=\mathsf{tr}\left(\nabla\mathbf{w}(X(\mathbf{x}))\right)=\mathrm{div}\ \mathbf{w}(X(\mathbf{x})).\qquad\square$$

Now, we can multiply (4.2) by a test function $\psi \in H^1_{\Gamma_D}(\Omega)$, integrate in Ω, and apply the usual Green's formula and (4.3) with $X(\mathbf{x}) = X_e(\mathbf{x}, t; \tau)$, obtaining

$$
\begin{aligned}
(4.5)\qquad & \int_\Omega \frac{\mathsf{d}\phi}{\mathsf{d}t}(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau))\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega \operatorname{div}\ \mathbf{F}_e^{-T}(\mathbf{x},t;\tau)\cdot\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega r(X_e(\mathbf{x},t;\tau))\phi(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_{\Gamma_R} \mathbf{F}_e^{-T}(\mathbf{x},t;\tau)\mathbf{n}(\mathbf{x})\cdot\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau))\ \psi(\mathbf{x})dA_{\mathbf{x}} \\
& = \int_\Omega f(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x}.
\end{aligned}
$$

Remark 4.1. Notice that as long as the involved functions $\mathbf{v}$, $\mathbf{A}$, r, and f are defined in a wider (time dependent) domain, equations (4.2) and (4.5) are valid without assuming the velocity field vanishes on the boundary. This is the case, for instance, with some interesting problems arising in mathematical finance. Actually, the above formulation has been necessary for the numerical solution of Asian options pricing problems (see [10]). However, the analysis in the present work covers only velocity fields which are null at the boundary, and time independent spatial domains.

LEMMA 4.2. *Let $X : \overline{\Omega} \longrightarrow \overline{\Omega}$, $X \in C^2(\overline{\Omega})$, be an invertible function satisfying $X(\mathbf{x}) = \mathbf{x}\ \forall \mathbf{x} \in \Gamma$. Let $\mathbf{F} = \nabla X$ such that $\mathbf{F}^{-1} \in C^1(\overline{\Omega})$. Then, we have*

$$
(4.6)\qquad \int_\Gamma \mathbf{F}^{-T}(\mathbf{x})\mathbf{n}(\mathbf{x})\cdot\mathbf{w}(\mathbf{x})\ \psi(\mathbf{x})\ dA_{\mathbf{x}} = \int_\Gamma \mathbf{n}(\mathbf{x})\cdot\mathbf{w}(\mathbf{x})\ \psi(\mathbf{x})\ \mathsf{det}\ \mathbf{F}^{-1}(\mathbf{x})\ dA_{\mathbf{x}}
$$

for $\mathbf{w} \in H^1(\Omega)$ and $\psi \in H^1(\Omega)$, where $\mathbf{n}$ is the outward unit normal vector to Γ.

Proof. First we apply the change of variable $\mathbf{x} = X^{-1}(\mathbf{y})$ to get (see [16, p. 53])

$$
\int_\Gamma \mathbf{F}^{-T}(\mathbf{x})\mathbf{n}(\mathbf{x})\cdot\mathbf{w}(\mathbf{x})\ \psi(\mathbf{x})\, dA_{\mathbf{x}} = \int_{\partial X(\Omega)} \mathbf{m}(\mathbf{y})\cdot\mathbf{w}(X^{-1}(\mathbf{y}))\ \psi(X^{-1}(\mathbf{y}))\mathsf{det}\ \mathbf{F}^{-1}(X^{-1}(\mathbf{y}))\, dA_{\mathbf{y}},
$$

where $\partial X(\Omega)$ denotes the boundary of $X(\Omega)$ and $\mathbf{m}$ is the unit normal vector to $\partial X(\Omega)$. Thus, $X(\mathbf{x}) = \mathbf{x}\ \forall \mathbf{x} \in \Gamma$ implies (4.6). ∎

Now, replacing in (4.5) formula (4.6) with $X(\mathbf{x}) = X_e(\mathbf{x}, t; \tau)$, and replacing the Robin condition (2.3), we have

$$
\begin{aligned}
(4.7)\qquad & \int_\Omega \frac{\mathsf{d}\phi}{\mathsf{d}t}(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau))\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega \operatorname{div}\ \mathbf{F}_e^{-T}(\mathbf{x},t;\tau)\cdot\mathbf{A}(X_e(\mathbf{x},t;\tau))\nabla\phi(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} \\
& + \int_\Omega r(X_e(\mathbf{x},t;\tau))\phi(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} + \int_{\Gamma_R} \alpha\ \phi(\mathbf{x},\tau)\ \psi(\mathbf{x})\ \mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\ dA_{\mathbf{x}} \\
& = \int_\Omega f(X_e(\mathbf{x},t;\tau),\tau)\ \psi(\mathbf{x})\ d\mathbf{x} + \int_{\Gamma_R} g(\mathbf{x},\tau)\ \psi(\mathbf{x})\ \mathsf{det}\ \mathbf{F}_e^{-1}(\mathbf{x},t;\tau)\ dA_{\mathbf{x}}.
\end{aligned}
$$

For these previous computations, we have assumed appropriate regularity on the involved data (i.e., $\mathbf{v} \in C^0(C^2(\Omega))$ and $\mathbf{A}\nabla(\cdot)\phi(\cdot,t) \in H^1(\Omega)$ for each $t \in (0,T)$) and that the velocity field, $\mathbf{v}$, vanishes on the boundary, so that the differentiable mapping, $\mathbf{x} \longrightarrow X_e(\mathbf{x},t;\tau)$, satisfies the result stated in Lemmas 4.1 and 4.2.

5. Time discretization. In this section we present a second order characteristics scheme for time semidiscretization of (4.7). Keeping in mind more general applications, it extends the scheme proposed in [24] to the case where the diffusion matrix depends on the space variable and can be degenerate, there are reaction and source terms, and the velocity field is not divergence-free. Moreover, mixed Dirichlet–Robin boundary conditions are allowed instead of merely Dirichlet ones.

In the first part, we develop some computations to motivate the scheme assuming that the characteristic lines are exactly computed. This can be useful in some cases (see, for instance, [10]). Then, after having studied results similar to those in section 3 concerning Euler and Runge–Kutta approximations of the characteristic lines, we propose the scheme. Finally, stability and error estimates are rigorously stated.

5.1. Second order semidiscretized scheme with exact characteristic lines. We propose a time semidiscretization of (4.7) for which we introduce the number of time steps, N, the time step $\Delta t = T/N$, and the mesh points $t_n = n\Delta t$ for $n = 0, 1/2, 1, 3/2, \ldots, N$. Throughout this work, we use the notation $\psi^n(\mathbf{x}) := \psi(\mathbf{x}, t_n)$ for a function $\psi(\mathbf{x},t)$. Moreover, for $n = 0, 1, 2, \ldots$, we define

$$
\begin{aligned}
X_e^n(\mathbf{x}) &:= X_e(\mathbf{x}, t_{n+1}; t_n), & \mathbf{F}_e^n(\mathbf{x}) &:= \mathbf{F}_e(\mathbf{x}, t_{n+1}; t_n),\\
X_e^{n+\frac{1}{2}}(\mathbf{x}) &:= X_e(\mathbf{x}, t_{n+1}; t_{n+\frac{1}{2}}), & \mathbf{F}_e^{n+\frac{1}{2}}(\mathbf{x}) &:= \mathbf{F}_e(\mathbf{x}, t_{n+1}; t_{n+\frac{1}{2}}).
\end{aligned} \tag{5.1}
$$

We recall that $X_e(\mathbf{x}, t_{n+1}; \tau)$ is the unique solution of the Cauchy problem

$$
\frac{\mathsf{d}X_e}{\mathsf{d}\tau}(\mathbf{x}, t_{n+1}; \tau) = \mathbf{v}\left(X_e(\mathbf{x}, t_{n+1}; \tau), \tau\right), \quad X_e(\mathbf{x}, t_{n+1}; t_{n+1}) = \mathbf{x}. \tag{5.2}
$$

The scheme we study arises from fixing $t = t_{n+1}$, $n = 0, 1, \ldots, N-1$, in (4.7) and using a Crank–Nicolson method with respect to τ. Thus, from (5.1), we have

$$
\begin{aligned}
&\int_\Omega \frac{\phi^{n+1}(\mathbf{x}) - \phi^n(X_e^n(\mathbf{x}))}{\Delta t}\ \psi(\mathbf{x})\ d\mathbf{x} + \frac{1}{2}\int_\Omega \mathbf{A}(\mathbf{x})\nabla\phi^{n+1}(\mathbf{x})\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x}\\
&+ \frac{1}{2}\int_\Omega (\mathbf{F}_e^n)^{-1}(\mathbf{x})\mathbf{A}(X_e^n(\mathbf{x}))\nabla\phi^n(X_e^n(\mathbf{x}))\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x}\\
&+ \frac{1}{2}\int_\Omega \operatorname{div}\ (\mathbf{F}_e^n)^{-T}(\mathbf{x})\cdot\mathbf{A}(X_e^n(\mathbf{x}))\nabla\phi^n(X_e^n(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}\\
&+ \frac{1}{2}\int_\Omega r(\mathbf{x})\phi^{n+1}(\mathbf{x}) + \frac{1}{2}\int_\Omega r(X_e^n(\mathbf{x}))\phi^n(X_e^n(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}\\
&+ \frac{1}{2}\int_{\Gamma_R} \alpha\left(\phi^{n+1}(\mathbf{x}) + \phi^n(\mathbf{x})\ \mathsf{det}\ (\mathbf{F}_e^n)^{-1}(\mathbf{x})\right)\psi(\mathbf{x})\ dA_{\mathbf{x}}\\
&= \frac{1}{2}\int_\Omega \left(f^{n+1}(\mathbf{x}) + f^n(X_e^n(\mathbf{x}))\right)\psi(\mathbf{x}) + \frac{1}{2}\int_{\Gamma_R}\left(g^{n+1}(\mathbf{x}) + g^n(\mathbf{x})\,\mathsf{det}\ (\mathbf{F}_e^n)^{-1}(\mathbf{x})\right)\psi(\mathbf{x})\ dA_{\mathbf{x}}.
\end{aligned} \tag{5.3}
$$

Remark 5.1. In section 5.4 we will prove that approximations involved in scheme (5.3) are of order $O(\Delta t^2)$ at point $(X_e^{n+\frac{1}{2}}(\mathbf{x}), t_{n+\frac{1}{2}})$.

The error of the scheme (5.3) does not change if we replace both $\mathbf{F}_e^{-1}$ and $\mathsf{det}\ \mathbf{F}_e^{-1}$ by their $O(\Delta t^2)$ approximations below, avoiding the matrix inversion computations. That is, by considering $(\mathbf{x}, t; s) = (\mathbf{x}, t_{n+1}; t_n)$ we deduce from (3.7) and (3.11) that

$$(\mathbf{F}_e^n)^{-1}(\mathbf{x}) = \mathbf{I}(\mathbf{x}) + \Delta t\ \mathbf{L}^n(X_e^n(\mathbf{x})) + O(\Delta t^2),$$
$$\mathsf{det}\ (\mathbf{F}_e^n)^{-1}(\mathbf{x}) = 1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1}(\mathbf{x}) + O(\Delta t^2).$$

Moreover, we can also use that $\mathrm{div}(\mathbf{F}_e^n)^{-T}(\mathbf{x}) = \Delta t\ \nabla \mathrm{div}\ \mathbf{v}^n(X_e^n(\mathbf{x})) + O(\Delta t^2)$. Thus, scheme (5.3) is replaced by

$$\begin{aligned}(5.4)\quad &\int_\Omega \frac{\phi^{n+1} - \phi^n(X_e^n)}{\Delta t}\ \psi\ d\mathbf{x} + \int_\Omega \frac{\mathbf{A}\nabla\phi^{n+1} + \mathbf{A}(X_e^n)\nabla\phi^n(X_e^n)}{2} \cdot \nabla\psi\ d\mathbf{x}\\ &+ \frac{\Delta t}{2}\int_\Omega \mathbf{L}^n(X_e^n)\mathbf{A}(X_e^n)\nabla\phi^n(X_e^n)\cdot\nabla\psi\ d\mathbf{x} + \frac{\Delta t}{2}\int_\Omega \nabla \mathrm{div}\ \mathbf{v}^n(X_e^n)\cdot\mathbf{A}(X_e^n)\nabla\phi^n(X_e^n)\ \psi\ d\mathbf{x}\\ &+ \int_\Omega \frac{r\phi^{n+1} + r(X_e^n)\phi^n(X_e^n)}{2}\ \psi\ d\mathbf{x} + \int_{\Gamma_R} \alpha\ \frac{\phi^{n+1} + \phi^n(1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}\ \psi\ dA_{\mathbf{x}}\\ &= \int_\Omega \frac{f^{n+1} + f^n(X_e^n)}{2}\ \psi\ d\mathbf{x} + \int_{\Gamma_R} \frac{g^{n+1} + g^n(1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}\ \psi\ dA_{\mathbf{x}}.\end{aligned}$$

5.2. Second order semidiscretized scheme with approximate characteristic lines. In most cases the Cauchy problem (5.2) cannot be exactly solved. Instead, following [24], we propose two explicit numerical schemes to approximate $X_e^n(\mathbf{x})$:

$$X_E^n(\mathbf{x}) := \mathbf{x} - \Delta t\ \mathbf{v}^{n+1}(\mathbf{x}) \qquad \text{(first order Euler scheme)},$$
$$X_{RK}^n(\mathbf{x}) := \mathbf{x} - \Delta t\ \mathbf{v}^{n+\frac{1}{2}}\left(\mathbf{x} - \frac{\Delta t}{2}\mathbf{v}^{n+1}(\mathbf{x})\right) \quad \text{(second order Runge–Kutta scheme)}.$$

A similar notation to the one in section 3 is used for the Jacobian of X_E^n, namely,

$$\mathbf{F}_E^n(\mathbf{x}) := \nabla X_E^n(\mathbf{x}) = \mathbf{I}(\mathbf{x}) - \Delta t\ \mathbf{L}^{n+1}(\mathbf{x}). \tag{5.5}$$

Now, we state three lemmas concerning properties of the characteristic line approximations, similar to those satisfied by the exact characteristics. For this, we require the time step to be bounded and the velocity to satisfy the following assumption.

Hypothesis 1. The velocity field $\mathbf{v} \in C^0\left(W^{1,\infty}(\Omega)\right)$ and satisfies $\mathbf{v} = 0$ on Γ.

LEMMA 5.1. *Under Hypothesis* 1, *if* $\|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))}\Delta t < 1/2$, *we have* $X_E^n(\overline{\Omega}) = X_{RK}^n(\overline{\Omega}) = \overline{\Omega}$.

Proof. See Proposition 1 in [24]. ∎

LEMMA 5.2. *Under Hypothesis* 1, *if* $\|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))}\Delta t < 1/2$, *then*

$$(\mathbf{F}_E^n)^{-1}(\mathbf{x}) = \mathbf{I} + \Delta t\ \mathbf{L}^{n+1}(\mathbf{x}) + \Delta t^2\ \left(\mathbf{L}^{n+1}(\mathbf{x})\right)^2 + O(\Delta t^3). \tag{5.6}$$

Proof. By applying norms to (5.5) we have that $\|\mathbf{I} - \mathbf{F}_E^n(\mathbf{x})\| < 1\ \forall \mathbf{x} \in \Omega$. Thus, $\mathbf{F}_E^n$ is invertible with $(\mathbf{F}_E^n)^{-1}(\mathbf{x}) = \sum_{j=0}^\infty (\mathbf{I}(\mathbf{x}) - \mathbf{F}_E^n(\mathbf{x}))^j$, and (5.6) follows. ∎

COROLLARY 5.3. *Under the assumptions of Lemma* 5.2, $\forall \mathbf{x} \in \Omega$, *we have*

$$\mathsf{det}(\mathbf{F}_E^n)^{-1}(\mathbf{x}) = 1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1}(\mathbf{x}) + O(\Delta t^2), \tag{5.7}$$
$$|\mathsf{det}(\mathbf{F}_E^n)^{-1}(\mathbf{x})| \le 1 + \Delta t\ \|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))} + O(\Delta t^2). \tag{5.8}$$

Proof. First, we have $\mathsf{det}(\mathbf{I}+\mathbf{D}) = 1 + \mathsf{tr}\ \mathbf{D} - \frac{1}{2}\left((\mathsf{tr}\ \mathbf{D})^2 - \mathsf{tr}\ \mathbf{D}^2\right) + \mathsf{det}\ \mathbf{D}$ for every tensor $\mathbf{D}$ (see, for instance, [16]). Thus, the result directly follows by replacing, in the previous formula, $\mathbf{D} = (\mathbf{F}_e^n)^{-1}(\mathbf{x}) - \mathbf{I}$ and using (5.6). ∎

LEMMA 5.4. *Under Hypothesis* 1, *if* $\psi \in L^2(\Omega)$ *and* $\|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))}\Delta t < 1/2$, *then there exists a positive constant* c *such that*

$$\|\psi \circ X_i^n\|_0^2 \leq (1 + \Delta t\ c)\,\|\psi\|_0^2 \quad \textit{for } n = 0, \ldots, N \textit{ and } i = E, RK. \tag{5.9}$$

Proof. See Lemma 1 in [24]. ∎

Thus, in the case where the characteristic lines and their gradients are not explicitly known, we propose the following approximation of (5.3):

$$\begin{aligned}
&\int_\Omega \frac{\phi^{n+1}(\mathbf{x}) - \phi^n(X_{RK}^n(\mathbf{x}))}{\Delta t}\ \psi(\mathbf{x})\ d\mathbf{x} + \frac{1}{2}\int_\Omega \mathbf{A}(\mathbf{x})\nabla\phi^{n+1}(\mathbf{x})\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x}\\
&\quad + \frac{1}{2}\int_\Omega (\mathbf{F}_E^n)^{-1}(\mathbf{x})\mathbf{A}(X_E^n(\mathbf{x}))\nabla\phi^n(X_E^n(\mathbf{x}))\cdot\nabla\psi(\mathbf{x})\ d\mathbf{x}\\
&\quad + \frac{1}{2}\int_\Omega \mathrm{div}\ (\mathbf{F}_E^n)^{-T}(\mathbf{x})\cdot\mathbf{A}(X_E^n(\mathbf{x}))\nabla\phi^n(X_E^n(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}\\
&\quad + \frac{1}{2}\int_\Omega r(\mathbf{x})\phi^{n+1}(\mathbf{x}) + \frac{1}{2}\int_\Omega r(X_E^n(\mathbf{x}))\phi^n(X_E^n(\mathbf{x}))\ \psi(\mathbf{x})\ d\mathbf{x}\\
&\quad + \frac{1}{2}\int_{\Gamma_R} \alpha\left(\phi^{n+1}(\mathbf{x}) + \phi^n(\mathbf{x})\ \mathsf{det}\ (\mathbf{F}_E^n)^{-1}(\mathbf{x})\right)\psi(\mathbf{x})\ dA_{\mathbf{x}}\\
&= \frac{1}{2}\int_\Omega \left(f^{n+1}(\mathbf{x}) + f^n(X_E^n(\mathbf{x}))\right)\psi(\mathbf{x})\ d\mathbf{x}\\
&\quad + \frac{1}{2}\int_{\Gamma_R} \left(g^{n+1}(\mathbf{x}) + g^n(\mathbf{x})\ \mathsf{det}\ (\mathbf{F}_E^n)^{-1}(\mathbf{x})\right)\psi(\mathbf{x})\ dA_{\mathbf{x}}.
\end{aligned} \tag{5.10}$$

Notice that we have used the lowest order characteristics approximation formula preserving second order time accuracy. The error of the semidiscretized scheme (5.10) does not change if we replace both $\mathbf{F}_E^{-1}$ and $\mathsf{det}\ \mathbf{F}_E^{-1}$ by their $O(\Delta t^2)$ approximations given in (5.6) and (5.7), which avoid the inversion of matrix $\mathbf{F}_E^n$.

Finally, we describe the semidiscretized scheme to be analyzed hereafter. Let us introduce $\mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\phi \in (H^1(\Omega))'$ and $\mathcal{F}_{\Delta t}^{n+\frac{1}{2}} \in (H^1(\Omega))'$, defined by

$$\begin{aligned}
\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\phi, \psi\right\rangle &:= \left\langle \frac{\phi^{n+1} - \phi^n \circ X_{RK}^n}{\Delta t}, \psi\right\rangle + \left\langle \frac{\mathbf{A}\nabla\phi^{n+1} + (\mathbf{A}\nabla\phi^n)\circ X_E^n}{2}, \nabla\psi\right\rangle\\
&\quad + \frac{\Delta t}{2}\left\langle (\mathbf{L}^n\mathbf{A}\nabla\phi^n)\circ X_E^n, \nabla\psi\right\rangle + \frac{\Delta t}{2}\left\langle (\nabla\mathrm{div}\ \mathbf{v}^n\cdot\mathbf{A}\nabla\phi^n)\circ X_E^n, \psi\right\rangle\\
&\quad + \left\langle \frac{r\phi^{n+1} + (r\phi^n)\circ X_E^n}{2}, \psi\right\rangle + \alpha\left\langle \frac{\phi^{n+1} + \phi^n(1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}, \psi\right\rangle_{\Gamma_R},
\end{aligned} \tag{5.11}$$

$$\left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}}, \psi\right\rangle := \left\langle \frac{f^{n+1} + f^n \circ X_E^n}{2}, \psi\right\rangle + \left\langle \frac{g^{n+1} + g^n(1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}, \psi\right\rangle_{\Gamma_R} \tag{5.12}$$

for $\phi \in C^0(H^1(\Omega))$ and $\psi \in H^1(\Omega)$.

Remark 5.2. Regarding the definitions of $\mathcal{L}_{\Delta t}^{n+\frac{1}{2}}$ and $\mathcal{F}_{\Delta t}^{n+\frac{1}{2}}$, only the values of function ϕ at discrete time steps $\{t_n\}_{n=0}^N$ are required. Thus, the above definitions can also be stated for a sequence of functions $\widehat{\phi} = \{\phi^n\}_{n=1}^N \in \left[H^1(\Omega)\right]^N$.

Then, the semidiscretized time scheme can be written as follows:

$$\text{(5.13)}\qquad \begin{cases} \text{Given } \phi^0_{\Delta t}, \text{ find } \widehat{\phi_{\Delta t}} = \{\phi^n_{\Delta t}\}_{n=1}^N \in \left[H^1_{\Gamma_D}(\Omega)\right]^N \text{ such that} \\ \left\langle \mathcal{L}^{n+\frac{1}{2}}_{\Delta t}\widehat{\phi_{\Delta t}}, \psi \right\rangle = \left\langle \mathcal{F}^{n+\frac{1}{2}}_{\Delta t}, \psi \right\rangle \quad \forall \psi \in H^1_{\Gamma_D}(\Omega) \text{ for } n = 0, \ldots, N-1. \end{cases}$$

Remark 5.3. The stability and convergence properties we are going to study in the sections that follow remain valid if we replace the approximations of characteristics appearing in scheme (5.13) by higher order ones or by the exact value. In particular, X^n_E can be replaced by X^n_{RK} or X^n_e, and X^n_{RK} can be replaced by X^n_e in (5.11) and (5.12).

5.3. Stability of the semidiscretized scheme. In order to develop the stability analysis, some assumptions on the different terms of (2.1) are required.

Hypothesis 2. The velocity field $\mathbf{v} \in C^0(W^{2,\infty}(\Omega))$ and satisfies $\mathbf{v} = 0$ on Γ.

Remark 5.4. Throughout this paper c_1 denotes the maximum between the positive constant appearing in Lemma 5.4 and the norm of the velocity in $C^0(W^{2,\infty}(\Omega))$.

Hypothesis 3. The diffusion matrix coefficients, A_{ij}, belong to $W^{1,\infty}(\Omega)$. Moreover, $\mathbf{A}$ is an $m \times m$ symmetric matrix satisfying

$$\text{(5.14)}\qquad \mathbf{A} = \begin{pmatrix} \mathbf{A}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix},$$

with $\mathbf{A}_{m_1}$ being a positive definite symmetric $m_1 \times m_1$ matrix ($m_1 \geq 1$), and where Θ denotes an appropriate zero matrix. Moreover, there exists a strictly positive constant δ which is a uniform lower bound for the eigenvalues of $\mathbf{A}_{m_1}$.

As a consequence of Hypothesis 3, there exists a unique positive definite symmetric $m_1 \times m_1$ matrix function, $\mathbf{C}_{m_1}$, such that $\mathbf{A}_{m_1} = (\mathbf{C}_{m_1})^2$. Let us denote by $\mathbf{C}$ the symmetric and positive semidefinite $m \times m$ matrix

$$\text{(5.15)}\qquad \mathbf{C} = \begin{pmatrix} \mathbf{C}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix}.$$

Notice that $\mathbf{A} = \mathbf{C}^2$ and $C_{ij} \in W^{1,\infty}(\Omega)$. At this point, let us introduce the constant $c_2 := \max_{i,j} \left\{ \|C_{ij}\|^2_{W^{1,\infty}(\Omega)} \right\}$. Next, let us denote by $\mathbf{B}$ the $m \times m$ matrix

$$\text{(5.16)}\qquad \mathbf{B} = \begin{pmatrix} \mathbf{I}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix},$$

where $\mathbf{I}_{m_1}$ is the $m_1 \times m_1$ identity matrix. Clearly, under Hypothesis 3 we have

$$\text{(5.17)}\qquad \delta \|\mathbf{B}\mathbf{w}\|^2_0 \leq \langle \mathbf{A}\mathbf{w}, \mathbf{w} \rangle = \|\mathbf{C}\mathbf{w}\|^2_0 \leq c_2 \|\mathbf{B}\mathbf{w}\|^2_0 \quad \forall \mathbf{w} \in \mathbb{R}^m.$$

Hypothesis 4. The velocity field satisfies $(\mathbf{I} - \mathbf{B})\mathbf{L}(\mathbf{x}, t)\mathbf{B} = \mathbf{0}\ \forall (\mathbf{x}, t) \in \Omega \times [0, T]$.

Remark 5.5. Hypothesis 4 is equivalent to having a velocity field $\mathbf{v}$ whose $m - m_1$ last components depend only on the last $m - m_1$ variables.

Remark 5.6. Under Hypotheses 3 and 4, for every $m \times m$ matrix $\mathbf{E}$ and vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^m$ it is easy to verify that $\langle \mathbf{E}\mathbf{A}\mathbf{w}_1, \mathbf{w}_2 \rangle = \langle \mathbf{E}\mathbf{A}\mathbf{w}_1, \mathbf{B}\mathbf{w}_2 \rangle$.

Hypothesis 5. The reaction function, $r \in W^{1,\infty}(\Omega)$, satisfies $0 < \gamma \leq r(\mathbf{x})$ in Ω, where γ is a constant.

Under the previous hypothesis, let $c_3 := \|\sqrt{r}\|^2_{W^{1,\infty}(\Omega)}$.

Hypothesis 6. The source function $f \in C^0(L^2(\Omega))$.

Hypothesis 7. In Robin boundary condition (2.3), $g \in C^0(L^2(\Gamma_R))$ and $\alpha > 0$.

It is convenient now to note that Hypothesis 3 also covers the nondegenerate case. This hypothesis is a common assumption in ultraparabolic equations (see, for instance, [23]), which represent a wide class of degenerate diffusion equations arising in many applications (see, for instance, [6]). Furthermore, as stated in [17], ultraparabolic problems either present C^∞ solutions or can be reduced to nondegenerate problems posed in a lower spatial dimension. This is an important point, as the stability and error estimates will be obtained under regularity assumptions on the solution.

Corresponding to the semidiscretized scheme, we have to deal with sequences of functions $\widehat{\psi} = \{\psi^n\}_{n=0}^N$. Thus, we consider spaces $l^\infty((0,T), L^2(\Omega))$, $l^2((0,T), L^2(\Omega))$ (abbreviated to $l^\infty(L^2(\Omega))$ and $l^2(L^2(\Omega))$, respectively) equipped with the norms

$$\left\|\widehat{\psi}\right\|_{l^\infty(L^2(\Omega))} := \max_{n=0}^{N} \|\psi^n\|_0, \qquad \left\|\widehat{\psi}\right\|_{l^2(L^2(\Omega))} := \sqrt{\Delta t \sum_{n=0}^{N} \|\psi^n\|_0^2}.$$

Similar definitions are considered for functional spaces $l^\infty(L^2(\Gamma_R))$ and $l^2(L^2(\Gamma_R))$ associated to the Robin boundary condition. Moreover, let us introduce the notation

$$D_{\Delta t}^n \widehat{\psi} := \frac{\psi^{n+1} - \psi^n}{\Delta t}. \tag{5.18}$$

For the sequence $\|\widehat{\psi}\|_0 := \{\|\psi^n\|_0\}$, let us define

$$D_{\Delta t}^n \left(\|\widehat{\psi}\|_0\right) := \frac{\|\psi^{n+1}\|_0 - \|\psi^n\|_0}{\Delta t}, \tag{5.19}$$

and for $\widehat{D_{\Delta t}\psi} := \left\{D_{\Delta t}^n \widehat{\psi}\right\}_{n=0}^{N-1}$ we define

$$\left\|\widehat{D_{\Delta t}\psi}\right\|_{l^2(L^2(\Gamma_R))} = \sqrt{\Delta t \sum_{n=0}^{N-1} \left\|\frac{\psi^{n+1} - \psi^n}{\Delta t}\right\|_{0,\Gamma_R}^2}. \tag{5.20}$$

Before establishing some technical lemmas, let us recall the Young inequality

$$ab \le \frac{1}{2}\left(ca^2 + \frac{1}{c}b^2\right) \tag{5.21}$$

for $a, b \in R$ and $c > 0$, which will be extensively used in what follows.

LEMMA 5.5. *Let us assume Hypotheses* 2, 3, 4, *and* 5. *Let us suppose* $c_1 \Delta t < 1/2$. *If* $\widehat{\phi_{\Delta t}} = \{\phi_{\Delta t}^n\}_{n=1}^N$ *denotes the solution of* (5.13) *and* $\alpha > 0$, $\delta > 0$ *are the constants appearing, respectively, in Hypothesis* 3 *and* (2.3), *then*

$$\begin{aligned}
&\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\phi_{\Delta t}}, \phi_{\Delta t}^{n+1} \right\rangle \\
&\quad \ge D_{\Delta t}^n \left(\frac{1}{2} \left\|\widehat{\phi_{\Delta t}}\right\|_0^2 + \frac{\Delta t}{4} \left\|\mathbf{C}\nabla \widehat{\phi_{\Delta t}}\right\|_0^2 + \frac{\Delta t}{4} \left\|\sqrt{r}\, \widehat{\phi_{\Delta t}}\right\|_0^2 + \frac{\alpha\, \Delta t}{4} \|\widehat{\phi_{\Delta t}}\|_{0,\Gamma_R}^2 \right) \\
&\qquad + \frac{1}{2\Delta t} \left\|\phi_{\Delta t}^{n+1} - \phi_{\Delta t}^n \circ X_{RK}^n\right\|_0^2 + \frac{1}{4} \left\|\mathbf{C}\nabla \phi_{\Delta t}^{n+1} + (\mathbf{C}\nabla \phi_{\Delta t}^n) \circ X_E^n\right\|_0^2 \\
&\qquad + \frac{1}{4} \left\|\sqrt{r}\phi_{\Delta t}^{n+1} + (\sqrt{r}\phi_{\Delta t}^n) \circ X_E^n\right\|_0^2 + \frac{\alpha}{4} \left\|\phi^{n+1} + \phi^n(1 + \Delta t \text{ div } \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2 \\
&\qquad - \frac{c}{2} \left(\|\phi_{\Delta t}^n\|_0^2 + \left\|\phi_{\Delta t}^{n+1}\right\|_0^2 \right) - c\Delta t\delta \left(\|\mathbf{B}\nabla \phi_{\Delta t}^n\|_0^2 + \left\|\mathbf{B}\nabla \phi_{\Delta t}^{n+1}\right\|_0^2 \right) \\
&\qquad - c\Delta t \left(\left\|\sqrt{r}\phi_{\Delta t}^n\right\|_0^2 + \left\|\sqrt{r}\phi_{\Delta t}^{n+1}\right\|_0^2 + \alpha \|\phi_{\Delta t}^n\|_{0,\Gamma_R}^2 \right),
\end{aligned} \tag{5.22}$$

where $\mathbf{C}\nabla\widehat{\phi_{\Delta t}} := \{\mathbf{C}\nabla\phi^n_{\Delta t}\}$, $\mathbf{B}\nabla\widehat{\phi_{\Delta t}} := \{\mathbf{B}\nabla\phi^n_{\Delta t}\}$, *and with the constant* c *given by* $c = \max\left\{1, c_1, c_2, (2c_1c_2 + c_1c_2^2)/\delta, c_1c_3/\gamma\right\}$.

Proof. First, we decompose $\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\widehat{\phi_{\Delta t}}, \phi^{n+1}_{\Delta t}\right\rangle = I_1 + I_2 + I_3 + I_4 + I_5 + I_6$, with

$$
\begin{aligned}
I_1 &= \left\langle \frac{\phi^{n+1}_{\Delta t} - \phi^n_{\Delta t}\circ X^n_{RK}}{\Delta t}, \phi^{n+1}_{\Delta t}\right\rangle,\\
I_2 &= \frac{1}{2}\left\langle \mathbf{A}\nabla\phi^{n+1}_{\Delta t} + (\mathbf{A}\nabla\phi^n_{\Delta t})\circ X^n_E, \nabla\phi^{n+1}_{\Delta t}\right\rangle,\\
I_3 &= \frac{\Delta t}{2}\left\langle (\mathbf{L}^n\mathbf{A}\nabla\phi^n_{\Delta t})\circ X^n_E, \nabla\phi^{n+1}_{\Delta t}\right\rangle,\\
I_4 &= \frac{\Delta t}{2}\left\langle (\nabla\ \mathrm{div}\ \mathbf{v}^n\cdot\mathbf{A}\nabla\phi^n_{\Delta t})\circ X^n_E, \phi^{n+1}_{\Delta t}\right\rangle,\\
I_5 &= \frac{1}{2}\left\langle r\phi^{n+1}_{\Delta t} + (r\phi^n_{\Delta t})\circ X^n_E, \phi^{n+1}_{\Delta t}\right\rangle,\\
I_6 &= \alpha\left\langle \frac{\phi^{n+1} + \phi^n(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}, \phi^{n+1}_{\Delta t}\right\rangle_{\Gamma_R}.
\end{aligned}
$$

For I_1 we can use Lemma 2 in [24] to obtain

$$
I_1 \geq D^n_{\Delta t}\left(\frac{1}{2}\widehat{\|\phi_{\Delta t}\|_0^2}\right) + \frac{1}{2\Delta t}\left\|\phi^{n+1}_{\Delta t} - \phi^n_{\Delta t}\circ X^n_{RK}\right\|_0^2 - \frac{c_1}{2}\left\|\phi^n_{\Delta t}\right\|_0^2. \tag{5.23}
$$

For I_2 the following lower bound can be stated:

$$
\begin{aligned}
I_2 \geq D^n_{\Delta t}&\left(\frac{\Delta t}{4}\left\|\mathbf{C}\nabla\widehat{\phi_{\Delta t}}\right\|_0^2\right) + \frac{1}{4}\left\|\mathbf{C}\nabla\phi^{n+1}_{\Delta t} + (\mathbf{C}\nabla\phi^n_{\Delta t})\circ X^n_E\right\|_0^2\\
&- \frac{3c_1c_2\Delta t}{4}\left(\|\mathbf{B}\nabla\phi^n_{\Delta t}\|_0^2 + \left\|\mathbf{B}\nabla\phi^{n+1}_{\Delta t}\right\|_0^2\right).
\end{aligned} \tag{5.24}
$$

To prove (5.24) we first use the definition of $D^n_{\Delta t}$, Lemma 5.4, and (5.21) to get

$$
\begin{aligned}
&D^n_{\Delta t}\left(\frac{\Delta t}{4}\left\|\mathbf{C}\nabla\widehat{\phi_{\Delta t}}\right\|_0^2\right) + \frac{1}{4}\left\|\mathbf{C}\nabla\phi^{n+1}_{\Delta t} + (\mathbf{C}\nabla\phi^n_{\Delta t})\circ X^n_E\right\|_0^2 - \frac{c_1c_2\Delta t}{4}\|\mathbf{B}\nabla\phi^n_{\Delta t}\|_0^2\\
&\quad\leq \frac{1}{4}\left(\left\|\mathbf{C}\nabla\phi^{n+1}_{\Delta t}\right\|_0^2 - \|(\mathbf{C}\nabla\phi^n_{\Delta t})\circ X^n_E\|_0^2 + \left\|\mathbf{C}\nabla\phi^{n+1}_{\Delta t} + (\mathbf{C}\ \nabla\ \phi^n_{\Delta t})\circ X^n_E\right\|_0^2\right)\\
&\quad= \frac{1}{2}\left\langle \mathbf{C}\nabla\phi^{n+1}_{\Delta t} + (\mathbf{C}\nabla\phi^n)\circ X^n_E, \mathbf{C}\nabla\phi^{n+1}_{\Delta t}\right\rangle.
\end{aligned}
$$

Next, we introduce the function $Y^n_E(\mathbf{x},\cdot) : [t_n, t_{n+1}] \longrightarrow \overline{\Omega}$, defined by $Y^n_E(\mathbf{x}, s) := \mathbf{x} - (t_{n+1} - s)\mathbf{v}^{n+1}(\mathbf{x})$, which satisfies $Y^n_E(\mathbf{x}, t_n) = X^n_E(\mathbf{x})$ and $Y^n_E(\mathbf{x}, t_{n+1}) = \mathbf{x}$. Moreover, since $\mathbf{A}_{m_1}$ is symmetric and positive definite, $\mathbf{C}_{m_1} = \sqrt{\mathbf{A}_{m_1}}$ is a differentiable function ($\nabla\mathbf{C}$ is the appropriate completion by zeros of $\nabla\mathbf{C}_{m_1}$). Then, by Barrow's rule and the chain rule, the following identity holds:

$$
\mathbf{C}(\mathbf{x}) = \mathbf{C}(X^n_E(\mathbf{x})) + \mathbf{D}^n(\mathbf{x}) \qquad \text{for a.e. } \mathbf{x}\in\Omega, \tag{5.25}
$$

where we have denoted by $\mathbf{D}^n$ the $m\times m$ symmetric matrix defined by

$$
D^n_{ij}(\mathbf{x}) := \int_{t_n}^{t_{n+1}} \nabla C_{ij}(Y^n_E(\mathbf{x}, s))\cdot\mathbf{v}^{n+1}(\mathbf{x})\ ds \qquad \text{for a.e. } \mathbf{x}\in\Omega, \tag{5.26}
$$

which verifies $\left|D_{ij}^n(\mathbf{x})\right| \le c_1\sqrt{c_2}\Delta t$. So, from the previous notation, we have

$$\begin{aligned}\frac{1}{2}\left\langle \mathbf{C}\nabla\phi_{\Delta t}^{n+1} + (\mathbf{C}\nabla\phi_{\Delta t}^{n})\circ X_E^n, \mathbf{C}\nabla\phi_{\Delta t}^{n+1}\right\rangle &= I_2 + \frac{1}{2}\left\langle \mathbf{D}^n(\mathbf{C}\nabla\phi_{\Delta t}^{n})\circ X_E^n, \nabla\phi_{\Delta t}^{n+1}\right\rangle \\ &\le I_2 + \frac{1}{2}\sqrt{1+c_1\Delta t}\sqrt{c_2}\,\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0\ \Delta t c_1\sqrt{c_2}\left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0 \\ &\le I_2 + \frac{c_1c_2\Delta t}{2}\left(\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2 + \left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0^2\right),\end{aligned}$$

where we have used the Cauchy–Schwarz inequality, Hypotheses 2 and 3, Lemma 5.4, inequality (5.21), and the fact that $c_1\Delta t < 1$. Finally, result (5.24) follows.

Next, by first using Remark 5.6 and then the Cauchy–Schwarz inequality, Hypotheses 2 and 3, Lemma 5.4, inequality (5.21), and $c_1\Delta t < 1$, we obtain

$$|I_3| = \frac{\Delta t}{2}\left|\langle(\mathbf{L}^n\mathbf{A}\nabla\phi_{\Delta t}^{n})\circ X_E^n, \mathbf{B}\nabla\phi_{\Delta t}^{n+1}\rangle\right| \le \frac{\Delta t\ c_1c_2}{2}\left(\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2 + \left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0^2\right).$$

Then when both $I_3 \ge 0$ and $I_3 < 0$, we have

$$I_3 \ge -\frac{c_1c_2\ \Delta t}{2}\left(\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2 + \left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0^2\right). \tag{5.27}$$

Similarly, for I_4 we obtain the estimate

$$I_4 \ge -\frac{c_1c_2^2\ \Delta t}{2}\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2 - \frac{1}{4}\left\|\phi_{\Delta t}^{n+1}\right\|_0^2. \tag{5.28}$$

For the reaction term, we can obtain

$$\begin{aligned}I_5 \ge D_{\Delta t}^n\left(\frac{\Delta t}{4}\left\|\sqrt{r}\,\widehat{\phi_{\Delta t}}\right\|_0^2\right) + \frac{1}{4}\left\|\sqrt{r}\phi_{\Delta t}^{n+1} + (\sqrt{r}\phi_{\Delta t}^{n})\circ X_E^n\right\|_0^2 \\ - \max\{c_1, c_1c_3/\gamma\}\ \frac{\Delta t}{2}\left(\left\|\sqrt{r}\phi_{\Delta t}^{n}\right\|_0^2 + \left\|\sqrt{r}\phi_{\Delta t}^{n+1}\right\|_0^2\right).\end{aligned} \tag{5.29}$$

The proof of (5.29) is analogous to the one of (5.24), but using Hypothesis 5 instead of Hypotheses 2 and 3.

For boundary integral term I_6, we first use some properties of the inner product in the space $L^2(\Gamma_R)$ and the inequality $(1+c_1\Delta t)^2 \le 1 + 3c_1\Delta t$ to get the estimate

$$\|\psi(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\|_{0,\Gamma_R}^2 \le (1+c_1\Delta t)^2\|\psi\|_{0,\Gamma_R}^2 \le (1+3c_1\Delta t)\,\|\psi\|_{0,\Gamma_R}^2$$

for $\psi \in L^2(\Gamma_R)$. Thus, we obtain

$$\begin{aligned}I_6 \ge D_{\Delta t}^n\left(\frac{\alpha\Delta t}{4}\left\|\widehat{\phi_{\Delta t}}\right\|_{0,\Gamma_R}^2\right) + \frac{\alpha}{4}\left\|\phi_{\Delta t}^{n+1} + \phi_{\Delta t}^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2 \\ - \frac{3}{4}c_1\alpha\Delta t\,\|\phi_{\Delta t}^{n}\|_{0,\Gamma_R}^2.\end{aligned} \tag{5.30}$$

Then, by summing up (5.23), (5.24), (5.27), (5.28), (5.29), and (5.30), inequality (5.22) follows. ∎

Now, we study the stability of a scheme with a more general right-hand side, i.e.,

$$\begin{cases}\text{Given } \phi_{\Delta t}^0, \text{ find } \widehat{\phi_{\Delta t}} = \{\phi_{\Delta t}^n\}_{n=1}^N \in \left[H_{\Gamma_D}^1(\Omega)\right]^N \text{ such that} \\ \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\widehat{\phi_{\Delta t}}, \psi\right\rangle = \left\langle \mathcal{H}_{\Delta t}^{n+\frac{1}{2}}, \psi\right\rangle \quad \forall\psi \in H_D^1(\Omega) \text{ for } n = 0,\dots,N-1,\end{cases} \tag{5.31}$$

with $\left\langle \mathcal{H}_{\Delta t}^{n+\frac{1}{2}}, \psi \right\rangle = \left\langle F^{n+1}, \psi \right\rangle + \left\langle G^{n+1}, \psi \right\rangle_{\Gamma_R}$.

Let us denote $\widehat{F} = \{F^n\}_{n=0}^N \in \left[L^2(\Omega)\right]^{N+1}$ and $\widehat{G} = \{G^n\}_{n=0}^N \in \left[L^2(\Gamma_R)\right]^{N+1}$. Notice that constant $\alpha > 0$ related to Robin boundary condition (2.3) appears explicitly in the following lemmas; however, they remain valid for any positive constant.

LEMMA 5.6. *Let us assume Hypothesis* 2, $F^{n+1} \in L^2(\Omega)$, *and* $G^{n+1} \in L^2(\Gamma_R)$. *If* $c_1 \Delta t < 1$, *then*

$$\begin{aligned} \left\langle F^{n+1}, \psi \right\rangle + \left\langle G^{n+1}, \psi \right\rangle_{\Gamma_R} &\leq \frac{1}{2}\left(\|F^{n+1}\|_0^2 + \|\psi\|_0^2\right) + \left\langle H^{n+1}, \psi - \varphi \right\rangle_{\Gamma_R} \\ &+ \frac{21}{\alpha}\|H^{n+1}\|_{0,\Gamma_R}^2 + \frac{\alpha c_1 \Delta t}{2}\|\varphi\|_{0,\Gamma_R}^2 + \frac{\alpha}{16}\|\psi + \varphi(1 + \Delta t \text{ div } \mathbf{v}^{n+1})\|_{0,\Gamma_R}^2, \end{aligned} \tag{5.32}$$

with $H^{n+1} := G^{n+1}/(2 + \Delta t \text{ div } \mathbf{v}^{n+1})$ $\forall \varphi, \psi \in H^1(\Omega)$ *and* $\alpha > 0$.

Proof. Let us introduce the notation $I_1 = \left\langle F^{n+1}, \psi \right\rangle$ and $I_2 = \left\langle G^{n+1}, \psi \right\rangle_{\Gamma_R}$. For I_1 we only need to apply the Cauchy–Schwarz inequality. For I_2 let us note first that function H^{n+1} is well defined under hypothesis $c_1 \Delta t < 1$. Then I_2 is decomposed into three terms, namely, $I_2 = I_2^1 + I_2^2 + I_2^3$, where

$$\begin{aligned} I_2^1 &= \left\langle H^{n+1}, \psi \right\rangle_{\Gamma_R} - \left\langle H^{n+1}, \varphi \right\rangle_{\Gamma_R}, \\ I_2^2 &= \left\langle H^{n+1}, \varphi \right\rangle_{\Gamma_R} - \left\langle H^{n+1}(1 + \Delta t \text{ div } \mathbf{v}^{n+1}), \varphi(1 + \Delta t \text{ div } \mathbf{v}^{n+1}) \right\rangle_{\Gamma_R}, \\ I_2^3 &= \left\langle H^{n+1}(1 + \Delta t \text{ div } \mathbf{v}^{n+1}), \varphi(1 + \Delta t \text{ div } \mathbf{v}^{n+1}) \right\rangle_{\Gamma_R} + \left\langle H^{n+1}(1 + \Delta t \text{ div } \mathbf{v}^{n+1}), \psi \right\rangle_{\Gamma_R}. \end{aligned}$$

In order to estimate I_2^2 it is easy to show that

$$\begin{aligned} I_2^2 &\leq 3c_1 \Delta t \, \|\varphi\|_{0,\Gamma_R} \|H^{n+1}\|_{0,\Gamma_R} \leq \frac{c_1 \Delta t}{2}\left(\alpha \|\varphi\|_{0,\Gamma_R}^2 + \frac{9}{\alpha}\|H^{n+1}\|_{0,\Gamma_R}^2\right) \\ &\leq \frac{c_1 \alpha \Delta t}{2}\|\varphi\|_{0,\Gamma_R}^2 + \frac{9}{2\alpha}\|H^{n+1}\|_{0,\Gamma_R}^2, \end{aligned}$$

where we have used (5.21) with $a = \|\varphi\|_{0,\Gamma_R}$, $b = 3\|H^{n+1}\|_{0,\Gamma_R}$, and $c = \alpha$.

For I_2^3, (5.21) and estimate $(1 + c_1 \Delta t)^2 \leq 4$ lead to

$$\begin{aligned} I_2^3 &= \left\langle H^{n+1}(1 + \Delta t \text{ div } \mathbf{v}^{n+1}), \psi + \varphi(1 + \Delta t \text{ div } \mathbf{v}^{n+1}) \right\rangle_{\Gamma_R} \\ &\leq \frac{16}{\alpha}\|H^{n+1}\|_{0,\Gamma_R}^2 + \frac{\alpha}{16}\|\psi + \varphi(1 + \Delta t \text{ div } \mathbf{v}^{n+1})\|_{0,\Gamma_R}^2. \end{aligned}$$

Finally, by jointly considering the above inequalities we get (5.32). ∎

The following lemma involves functions defined on Γ_R and velocity field $\mathbf{v}$.

LEMMA 5.7. *Let us assume Hypothesis* 2. *Let* $\{G^n\}_{n=0}^N \in \left[L^2(\Gamma_R)\right]^{N+1}$ *and* $\{H^n\}_{n=0}^N$ *be as in Lemma* 5.6. *If* $c_1 \Delta t < 1$, *then* $\{H^n\}_{n=0}^N \in \left[L^2(\Gamma_R)\right]^{N+1}$ *and*

$$\|H^n\|_{0,\Gamma_R} \leq \|G^n\|_{0,\Gamma_R}. \tag{5.33}$$

Moreover, for any sequence $\{\psi^n\}_{n=0}^N \in \left[L^2(\Gamma_R)\right]^{N+1}$ *and any* $m \in \{0, \ldots, N-1\}$, *the following inequality holds:*

$$\begin{aligned} \left|\sum_{n=0}^{m-1} \left\langle H^{n+1}, \psi^{n+1} - \psi^n \right\rangle_{\Gamma_R}\right| &\leq \frac{\alpha}{16}\|\psi^m\|_{0,\Gamma_R}^2 + \frac{4}{\alpha}\|G^m\|_{0,\Gamma_R}^2 + \frac{\alpha}{16}\left\|\psi^0\right\|_{0,\Gamma_R}^2 \\ + \frac{4}{\alpha}\left\|G^1\right\|_{0,\Gamma_R}^2 + 3\alpha \Delta t \sum_{n=1}^{m-1}\|\psi^n\|_{0,\Gamma_R}^2 &+ \frac{\Delta t}{\alpha}\sum_{n=1}^{m-1}\left\|\frac{G^{n+1} - G^n}{\Delta t}\right\|_{0,\Gamma_R}^2 + \frac{c_1 \Delta t}{\alpha}\sum_{n=1}^{m}\|G^n\|_{0,\Gamma_R}^2 \end{aligned} \tag{5.34}$$

for any $\alpha > 0$ related to condition (2.3).

Proof. First, since $c_1\Delta t < 1$, then $1/3 \le (2 + \Delta t \text{ div } \mathbf{v}^n)^{-1} \le 1$. So, $|H^n(\mathbf{x})| \le |G^n(\mathbf{x})|$ for a.e. $\mathbf{x} \in \Gamma_R$, $H^n \in L^2(\Gamma_R)$, and (5.33) holds. For (5.34), we use the identity

(5.35)
$$\sum_{n=0}^{m-1} \left\langle H^{n+1}, \psi^{n+1} - \psi^n \right\rangle_{\Gamma_R} = \langle H^m, \psi^m \rangle_{\Gamma_R} - \left\langle H^1, \psi^0 \right\rangle_{\Gamma_R} - \Delta t \sum_{n=1}^{m-1} \left\langle \frac{H^{n+1} - H^n}{\Delta t}, \psi^n \right\rangle_{\Gamma_R}.$$

The third term in (5.35) can be bounded as follows:

$$\Delta t \left| \sum_{n=1}^{m-1} \left\langle \frac{H^{n+1} - H^n}{\Delta t}, \psi^n \right\rangle_{\Gamma_R} \right| \le \frac{\Delta t}{\alpha} \sum_{n=1}^{m-1} \left\| \frac{G^{n+1} - G^n}{\Delta t} \right\|_{0,\Gamma_R}^2 + \frac{c_1 \Delta t}{\alpha} \sum_{n=0}^{m-1} \|G^n\|_{0,\Gamma_R}^2 + 3\alpha \Delta t \sum_{n=1}^{m-1} \|\psi^n\|_{0,\Gamma_R}^2, \tag{5.36}$$

where we have used the equality

$$\frac{H^{n+1} - H^n}{\Delta t} = \frac{2(G^{n+1} - G^n)}{\Delta t\ (2 + \Delta t \text{ div } \mathbf{v}^{n+1})(2 + \Delta t \text{ div } \mathbf{v}^n)} + \frac{\text{div } \mathbf{v}^n\ G^{n+1} - \text{div } \mathbf{v}^{n+1}\ G^n}{(2 + \Delta t \text{ div } \mathbf{v}^{n+1})(2 + \Delta t \text{ div } \mathbf{v}^n)},$$

together with (5.21), the stated bound for the term $(2 + \Delta t \text{ div } \mathbf{v}^n)^{-1}$, and Hypothesis 2. The first two terms in (5.35) can be bounded by using (5.21) and (5.33), obtaining

$$\left| \left\langle H^i, \psi^j \right\rangle_{\Gamma_R} \right| \le \frac{4}{\alpha} \|G^i\|_{0,\Gamma_R}^2 + \frac{\alpha}{16} \|\psi^j\|_{0,\Gamma_R}^2 \quad \text{for } (i,j) = (1,0) \text{ and } (m,m). \tag{5.37}$$

Finally, by jointly considering (5.35), (5.36), and (5.37) we get (5.34). ∎

THEOREM 5.8. *Let us assume Hypotheses* 2, 3, 4, *and* 5. *Let* $\widehat{F} \in \left[L^2(\Omega)\right]^{N+1}$, $\widehat{G} \in \left[L^2(\Gamma_R)\right]^{N+1}$, *and* $\widehat{\phi_{\Delta t}} = \{\phi_{\Delta t}^n\}_{n=1}^N$ *be the solution of* (5.31) *subject to initial value* $\phi_{\Delta t}^0 \in H^1(\Omega)$. *Let* $\alpha > 0$ *be the constant appearing in* (2.3). *Then there exist two positive constants* c *and* $d = d(c_1, c_2, \delta, c_3, \gamma)$, *such that if* $\Delta t < d$, *then*

$$\begin{aligned} &\frac{1}{\sqrt{2}} \left\| \widehat{\phi_{\Delta t}} \right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\delta\ \Delta t}{4}} \left\| \mathbf{B}\nabla \widehat{\phi_{\Delta t}} \right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\Delta t}{4}} \left\| \sqrt{r}\, \widehat{\phi_{\Delta t}} \right\|_{l^\infty(L^2(\Omega))} \\ &+ \sqrt{\frac{\alpha \Delta t}{16}} \left\| \widehat{\phi_{\Delta t}} \right\|_{l^\infty(L^2(\Gamma_R))} \le c \left(\frac{1}{2} \left\| \phi_{\Delta t}^0 \right\|_0 + \sqrt{\frac{\delta\ \Delta t}{4}} \left\| \mathbf{B}\nabla \phi_{\Delta t}^0 \right\|_0 + \sqrt{\frac{\Delta t}{4}} \left\| \sqrt{r} \phi_{\Delta t}^0 \right\|_0 \right. \\ &\left. + \sqrt{\frac{\alpha \Delta t}{16}} \left\| \phi_{\Delta t}^0 \right\|_{0,\Gamma_R} + \left\| \widehat{F} \right\|_{l^2(L^2(\Omega))} + \left\| \widehat{G} \right\|_{l^2(L^2(\Gamma_R))} + \Delta t \left\| \widehat{D_{\Delta t} G} \right\|_{l^2(L^2(\Gamma_R))} \right), \end{aligned} \tag{5.38}$$

where $\mathbf{B}\nabla \widehat{\phi_{\Delta t}} := \{\mathbf{B}\nabla \phi_{\Delta t}^n\}$, $\widehat{F} = \{F^n\}_{n=0}^N$, *and* $\widehat{G} = \{G^n\}_{n=0}^N$.

Proof. Sequence $\widehat{\phi_{\Delta t}} = \{\phi_{\Delta t}^n\}_{n=1}^N$ satisfies $\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\phi_{\Delta t}}, \phi_{\Delta t}^{n+1} \right\rangle = \left\langle \mathcal{H}_{\Delta t}^{n+\frac{1}{2}}, \phi_{\Delta t}^{n+1} \right\rangle$.

So, we first use Lemma 5.5 to obtain a lower bound left-hand side of this expression:

$$
\begin{aligned}
&\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\widehat{\phi_{\Delta t}},\phi_{\Delta t}^{n+1}\right\rangle \\
&\geq D_{\Delta t}^{n}\left(\frac{1}{2}\left\|\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\Delta t}{4}\left\|\mathbf{C}\nabla\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\Delta t}{4}\left\|\sqrt{r}\,\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\alpha\Delta t}{4}\left\|\widehat{\phi_{\Delta t}}\right\|_{0,\Gamma_R}^2\right)\\
&+\frac{\alpha}{4}\|\phi_{\Delta t}^{n+1}+\phi_{\Delta t}^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\|_{0,\Gamma_R}^2-c\left(\|\phi_{\Delta t}^{n}\|_0^2+\left\|\phi_{\Delta t}^{n+1}\right\|_0^2\right)\\
&-c\Delta t\left(\delta\left(\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2+\left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0^2\right)+\left\|\sqrt{r}\phi_{\Delta t}^{n}\right\|_0^2+\left\|\sqrt{r}\phi_{\Delta t}^{n+1}\right\|_0^2+\alpha\left\|\phi_{\Delta t}^{n}\right\|_{0,\Gamma_R}^2\right).
\end{aligned}
$$

Second, we use Lemma 5.6 for $\psi=\phi_{\Delta t}^{n+1}$ and $\varphi=\phi_{\Delta t}^{n}$ to obtain the upper bound

$$
\begin{aligned}
\left\langle \mathcal{H}_{\Delta t}^{n+\frac{1}{2}},\phi_{\Delta t}^{n+1}\right\rangle &\leq \frac{1}{2}\left(\left\|F^{n+1}\right\|_0^2+\left\|\phi^{n+1}\right\|_0^2\right)+\left\langle H^{n+1},\phi^{n+1}-\phi^{n}\right\rangle_{\Gamma_R}\\
&+\frac{21}{\alpha}\left\|H^{n+1}\right\|_{0,\Gamma_R}^2+\frac{\alpha c_1\Delta t}{2}\left\|\phi^{n}\right\|_{0,\Gamma_R}^2+\frac{\alpha}{16}\left\|\phi^{n+1}+\phi^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2.
\end{aligned}
$$

Next, by jointly considering both estimates, regrouping, and simplifying terms we get

$$
\begin{aligned}
&D_{\Delta t}^{n}\left(\frac{1}{2}\left\|\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\Delta t}{4}\left\|\mathbf{C}\nabla\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\Delta t}{4}\left\|\sqrt{r}\,\widehat{\phi_{\Delta t}}\right\|_0^2+\frac{\alpha\Delta t}{4}\left\|\widehat{\phi_{\Delta t}}\right\|_{0,\Gamma_R}^2\right)\\
&\leq\frac{1}{2}\left\|F^{n+1}\right\|_0^2+\frac{21}{\alpha}\left\|H^{n+1}\right\|_{0,\Gamma_R}^2+\left\langle H^{n+1},\phi_{\Delta t}^{n+1}-\phi_{\Delta t}^{n}\right\rangle_{\Gamma_R}+c\left(\|\phi_{\Delta t}^{n}\|_0^2+\left\|\phi_{\Delta t}^{n+1}\right\|_0^2\right)\\
&+c\Delta t\left(\delta\left(\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2+\left\|\mathbf{B}\nabla\phi_{\Delta t}^{n+1}\right\|_0^2\right)+\left\|\sqrt{r}\phi_{\Delta t}^{n}\right\|_0^2+\left\|\sqrt{r}\phi_{\Delta t}^{n+1}\right\|_0^2+2\alpha\left\|\phi_{\Delta t}^{n}\right\|_{0,\Gamma_R}^2\right),
\end{aligned}
\tag{5.39}
$$

with $c=\max\left\{1,c_1,c_2,(2c_1c_2+c_1c_2^2)/\delta,c_1c_3/\gamma\right\}$. Now, for fixed integer $m\geq 1$, let us sum (5.39) multiplied by Δt from $n=0$ to $n=m-1$. We obtain

$$
\begin{aligned}
&\frac{1}{2}\|\phi_{\Delta t}^{m}\|_0^2+\frac{\Delta t}{4}\|\mathbf{C}\nabla\phi_{\Delta t}^{m}\|_0^2+\frac{\Delta t}{4}\left\|\sqrt{r}\phi_{\Delta t}^{m}\right\|_0^2+\frac{\alpha\Delta t}{4}\|\phi_{\Delta t}^{m}\|_{0,\Gamma_R}^2\\
&-\frac{1}{2}\|\phi_{\Delta t}^{0}\|_0^2-\frac{\Delta t}{4}\|\mathbf{C}\nabla\phi_{\Delta t}^{0}\|_0^2-\frac{\Delta t}{4}\left\|\sqrt{r}\phi_{\Delta t}^{0}\right\|_0^2-\frac{\alpha\Delta t}{4}\|\phi_{\Delta t}^{0}\|_{0,\Gamma_R}^2\\
&\leq\frac{\Delta t}{2}\sum_{n=1}^{m}\|F^{n}\|_0^2+\frac{21\Delta t}{\alpha}\sum_{n=1}^{m}\|H^{n}\|_{0,\Gamma_R}^2+\Delta t\sum_{n=0}^{m-1}\left\langle H^{n+1},\phi_{\Delta t}^{n+1}-\phi_{\Delta t}^{n}\right\rangle_{\Gamma_R}\\
&+2c\Delta t\sum_{n=0}^{m}\|\phi_{\Delta t}^{n}\|_0^2+2c\Delta t^2\left(\sum_{n=0}^{m}\delta\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2+\sum_{n=0}^{m}\left\|\sqrt{r}\phi_{\Delta t}^{n}\right\|_0^2+\sum_{n=0}^{m-1}\alpha\|\phi_{\Delta t}^{n}\|_{0,\Gamma_R}^2\right).
\end{aligned}
$$

Now, by using (5.17) and Lemma 5.7, for $\psi^n=\phi^n$ we get

$$
\begin{aligned}
&\frac{1}{2}\|\phi_{\Delta t}^{m}\|_0^2+\frac{\delta\Delta t}{4}\|\mathbf{B}\nabla\phi_{\Delta t}^{m}\|_0^2+\frac{\Delta t}{4}\left\|\sqrt{r}\phi_{\Delta t}^{m}\right\|_0^2+\frac{3\alpha\Delta t}{16}\|\phi_{\Delta t}^{m}\|_{0,\Gamma_R}^2\\
&\leq\frac{1}{2}\left\|\phi_{\Delta t}^{0}\right\|_0^2+\frac{c_2\Delta t}{4}\left\|\mathbf{B}\nabla\phi_{\Delta t}^{0}\right\|_0^2+\frac{\Delta t}{4}\left\|\sqrt{r}\phi_{\Delta t}^{0}\right\|_0^2+\frac{5\alpha\Delta t}{16}\left\|\phi_{\Delta t}^{0}\right\|_{0,\Gamma_R}^2\\
&+\frac{\Delta t}{2}\sum_{n=1}^{m}\|F^{n}\|_0^2+\frac{25\Delta t}{\alpha}\sum_{n=1}^{m}\|G^{n}\|_{0,\Gamma_R}^2+\frac{c_1\Delta t^2}{\alpha}\sum_{n=1}^{m}\|G^{n}\|_{0,\Gamma_R}^2+\frac{\Delta t^2}{\alpha}\sum_{n=1}^{m-1}\left\|\frac{G^{n+1}-G^{n}}{\Delta t}\right\|_{0,\Gamma_R}^2\\
&+2c\Delta t\sum_{n=0}^{m}\|\phi_{\Delta t}^{n}\|_0^2+2c\Delta t^2\left(\sum_{n=0}^{m}\delta\|\mathbf{B}\nabla\phi_{\Delta t}^{n}\|_0^2+\sum_{n=0}^{m}\left\|\sqrt{r}\phi_{\Delta t}^{n}\right\|_0^2+\frac{5}{2}\sum_{n=1}^{m-1}\alpha\|\phi_{\Delta t}^{n}\|_{0,\Gamma_R}^2\right).
\end{aligned}
$$

Let us introduce, for $n = 0, \dots, N$, the notation

$$\theta_n := \frac{1}{2} \|\phi^n_{\Delta t}\|^2_0 + \frac{\delta \Delta t}{4} \|\mathbf{B}\nabla\phi^n_{\Delta t}\|^2_0 + \frac{\Delta t}{4} \left\|\sqrt{r}\ \phi^n_{\Delta t}\right\|^2_0, \quad \overline{\theta}_n := \frac{\alpha \Delta t}{16} \|\phi^n_{\Delta t}\|^2_{0,\Gamma_R}.$$

With the above notation we have

$$\begin{aligned}(1 - 8c\Delta t)\theta_m + \overline{\theta}_m &\le 8c\Delta t \sum_{n=0}^{m-1} \theta_n + 80c\Delta t \sum_{n=0}^{m-1} \overline{\theta}_n \\ &+ \widetilde{c}\left(\theta_0 + \overline{\theta_0} + \left\|\widehat{F}\right\|^2_{l^2(L^2(\Omega))} + \left\|\widehat{G}\right\|^2_{l^2(L^2(\Gamma_R))} + \Delta t \left\|\widehat{D_{\Delta t} G}\right\|^2_{l^2(L^2(\Gamma_R))}\right),\end{aligned}$$

with $c = \max\left\{1, c_1, c_2, (2c_1c_2 + c_1^2c_2^2)/\delta, c_1c_3/\gamma\right\}$ and $\widetilde{c}$ a positive constant. For Δt small enough, we can apply the discrete Gronwall inequality (see, for instance, [22]) and take the maximum in $q \in \{1, \dots, N\}$. Thus, estimate (5.38) follows. ∎

COROLLARY 5.9 (stability). *Let us assume Hypotheses* 2 *to* 7. *Let* $\widehat{\phi_{\Delta t}} = \{\phi^n_{\Delta t}\}_{n=1}^N$ *be the solution of* (5.13) *subject to initial value* $\phi^0_{\Delta t}$. *Then there exist two positive constants,* c *and* $d = d(c_1, c_2, \delta, c_3, \gamma)$, *such that, for* $\Delta t < d$, *we have*

$$\begin{aligned}&\frac{1}{\sqrt{2}} \left\|\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\delta\ \Delta t}{4}} \left\|\mathbf{B}\nabla\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\Delta t}{4}} \left\|\sqrt{r}\,\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} \\ &+ \sqrt{\frac{\alpha\ \Delta t}{16}} \left\|\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Gamma_R))} \le c\left(\frac{1}{2}\left\|\phi^0_{\Delta t}\right\|_0 + \sqrt{\frac{\delta\ \Delta t}{4}}\left\|\mathbf{B}\nabla\phi^0_{\Delta t}\right\|_0 + \sqrt{\frac{\Delta t}{4}}\left\|\sqrt{r}\phi^0_{\Delta t}\right\|_0\right. \\ &\left. + \sqrt{\frac{\alpha\ \Delta t}{16}}\left\|\phi^0_{\Delta t}\right\|_{0,\Gamma_R} + \left\|\widehat{f}\right\|_{l^2(L^2(\Omega))} + \|\widehat{g}\|_{l^2(L^2(\Gamma_R))} + \Delta t \left\|\widehat{D_{\Delta t} g}\right\|_{l^2(L^2(\Gamma_R))}\right),\end{aligned} \tag{5.40}$$

where $\mathbf{B}\nabla\widehat{\phi_{\Delta t}} := \{\mathbf{B}\nabla\phi^n_{\Delta t}\}$.

Proof. The result follows directly by replacing F^{n+1} with $f^{n+1} + f^n \circ X^n_E$ and G^{n+1} with $g^{n+1} + g^n(1 + \Delta t \text{ div } \mathbf{v}^{n+1})$ in (5.38). ∎

5.4. Error estimate for the semidiscretized scheme. The aim of the present section is to estimate the difference between the *discrete solution* of (5.13), $\widehat{\phi_{\Delta t}}$, and $\widehat{\phi} = \{\phi^n\}$, the exact solution of the continuous problem. According to (4.7) for $t = t_{n+1}$ and $\tau = t_{n+\frac{1}{2}}$, the latter solves the problem

$$\left\langle \mathcal{L}^{n+\frac{1}{2}}\widehat{\phi}, \psi\right\rangle = \left\langle \mathcal{F}^{n+\frac{1}{2}}, \psi\right\rangle \quad \forall \psi \in H^1_{\Gamma_D}(\Omega), \tag{5.41}$$

where $\mathcal{L}^{n+\frac{1}{2}}\widehat{\phi} \in (H^1(\Omega))'$ and $\mathcal{F}^{n+\frac{1}{2}} \in (H^1(\Omega))'$ are defined by

$$\begin{aligned}\left\langle \mathcal{L}^{n+\frac{1}{2}}\widehat{\phi}, \psi\right\rangle &:= \left\langle \left(\frac{\mathsf{d}\phi}{\mathsf{d}t}\right)^{n+\frac{1}{2}} \circ X_e^{n+\frac{1}{2}}, \psi\right\rangle + \left\langle \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1}\left(\mathbf{A}\nabla\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}},\ \nabla\ \psi\right\rangle \\ &+ \left\langle \text{div } \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-\mathrm{T}} \cdot \left(\mathbf{A}\nabla\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}}, \psi\right\rangle \\ &+ \left\langle \left(r\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}}, \psi\right\rangle + \alpha \left\langle \mathsf{det}\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1} \phi^{n+\frac{1}{2}}, \psi\right\rangle_{\Gamma_R}, \\ \left\langle \mathcal{F}^{n+\frac{1}{2}}, \psi\right\rangle &:= \left\langle f^{n+\frac{1}{2}} \circ X_e^{n+\frac{1}{2}}, \psi\right\rangle + \left\langle \mathsf{det}\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1} g^{n+\frac{1}{2}}, \psi\right\rangle_{\Gamma_R} \quad \forall \psi \in H^1(\Omega).\end{aligned}$$

The error estimate to be stated in Theorem 5.21 is proved by means of Theorem 5.8 and the forthcoming Lemmas 5.19 and 5.20. Before doing this, we give some results with sketched proofs (see [20] for further details). Moreover, in what follows, $\widetilde{c_1}$ denotes a generic positive constant related to the norm of the velocity field $\mathbf{v}$ and not necessarily the same at each occurrence.

LEMMA 5.10. *Let us assume that* $\mathbf{v} \in C^2(L^\infty(\Omega)) \cap C^1(W^{1,\infty}(\Omega)) \cap C^0(W^{2,\infty}(\Omega))$ *and vanishes on the boundary,* $\Delta t \|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))} < 1/2$, *and* $\varphi \in Z^3$. *Let us define the function* $\xi^{n+\frac{1}{2}}$ *by*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) := \frac{\mathsf{d}\varphi}{\mathsf{d}t}^{n+\frac{1}{2}} \left(X_e^{n+\frac{1}{2}}(\mathbf{x})\right) - \frac{\varphi^{n+1}(\mathbf{x}) - \varphi^n(X_{RK}^n(\mathbf{x}))}{\Delta t} \quad \forall \mathbf{x} \in \Omega.$$

Then $\xi^{n+\frac{1}{2}} \in L^2(\Omega)$ *and* $\|\xi^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1} \Delta t^2 \|\varphi\|_{Z^3}$, $n = 0, \ldots, N-1$.

Proof. Let us first write $\xi^{n+\frac{1}{2}}(\mathbf{x}) = \xi_1^{n+\frac{1}{2}}(\mathbf{x}) + \xi_2^{n+\frac{1}{2}}(\mathbf{x})$ with

$$\xi_1^{n+\frac{1}{2}}(\mathbf{x}) := \frac{\mathsf{d}\varphi}{\mathsf{d}t}^{n+\frac{1}{2}} \left(X_e^{n+\frac{1}{2}}(\mathbf{x})\right) - \frac{\varphi^{n+1}(\mathbf{x}) - \varphi^n(X_e^n(\mathbf{x}))}{\Delta t} \quad \forall \mathbf{x} \in \Omega,$$
$$\xi_2^{n+\frac{1}{2}}(\mathbf{x}) := \frac{\varphi^n(X_e^n(\mathbf{x})) - \varphi^n(X_{RK}^n(\mathbf{x}))}{\Delta t} \quad \forall \mathbf{x} \in \Omega.$$

The result follows by applying Taylor expansions to the above functions, noting that $|X_e^n(\mathbf{x}) - X_{RK}^n(\mathbf{x})| \leq \widetilde{c_1} \Delta t^3$. ∎

LEMMA 5.11. *Let us assume that* $\mathbf{v} \in C^1(L^\infty(\Omega)) \cap C^0(W^{1,\infty}(\Omega))$ *and vanishes on the boundary. Let* $\varphi \in Z^2$ *be a given function and* $\xi^{n+\frac{1}{2}} : \Omega \longrightarrow \mathbb{R}^m$ *be defined by*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) := \varphi(X_e^{n+\frac{1}{2}}(\mathbf{x}), t_{n+\frac{1}{2}}) - \frac{\varphi(\mathbf{x}, t_{n+1}) + \varphi(X_e^n(\mathbf{x}), t_n)}{2}.$$

Then $\xi^{n+\frac{1}{2}} \in L^2(\Omega)$ *and we have*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) = -\frac{1}{2} \int_{t_{n+\frac{1}{2}}}^{t_{n+1}} (t_{n+1} - s) \, \frac{\mathsf{d}^2\varphi}{\mathsf{d}t^2}(X_e(\mathbf{x}, t_{n+1}; s), s) \, ds$$
$$- \frac{1}{2} \int_{t_{n+\frac{1}{2}}}^{t_n} (t_n - s) \, \frac{\mathsf{d}^2\varphi}{\mathsf{d}t^2}(X_e(\mathbf{x}, t_{n+1}; s), s) \, ds, \quad a.e.\ \mathbf{x} \in \Omega, \quad n = 0, \ldots, N-1.$$

Proof. For the function $G(\tau) := \varphi(X_e(\mathbf{x}, t_{n+1}; \tau), \tau)$, $\tau \in (0, T)$, we have

$$G(\tau) = G(t_{n+\frac{1}{2}}) + (\tau - t_{n+\frac{1}{2}}) G'(t_{n+\frac{1}{2}}) + \int_{t_{n+\frac{1}{2}}}^{\tau} (\tau - s) G''(s) ds.$$

Thus, the result follows by taking successively $\tau = t_n$ and $\tau = t_{n+1}$ and adding both expressions. ∎

LEMMA 5.12. *Let us assume that* $\mathbf{v} \in C^1(W^{1,\infty}(\Omega)) \cap C^0(W^{2,\infty}(\Omega))$ *and vanishes on the boundary. Let* $\mathbf{w} : \Omega \times [0, T] \longrightarrow \mathbb{R}^m$, $\mathbf{w} \in Z^2$, *be a given function, and let* $\vartheta^{n+\frac{1}{2}} : \Omega \longrightarrow \mathbb{R}^m$ *be defined by*

$$\vartheta^{n+\frac{1}{2}}(\mathbf{x}) := \mathbf{F}_e^{-1}(\mathbf{x}, t_{n+1}; t_{n+\frac{1}{2}}) \mathbf{w}(X_e^{n+\frac{1}{2}}(\mathbf{x}), t_{n+\frac{1}{2}})$$
$$- \frac{\mathbf{w}(\mathbf{x}, t_{n+1}) + (\mathbf{F}_e^n)^{-1}(\mathbf{x}) \mathbf{w}(X_e^n(\mathbf{x}), t_n)}{2}.$$

Then $\vartheta^{n+\frac{1}{2}} \in L^2(\Omega)$ *and* $\|\vartheta^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1}\Delta t^2 \|\mathbf{w}\|_{Z^2}$, $n = 0, \ldots, N-1$. *Moreover, if* $\mathbf{v} \in C^1(W^{2,\infty}(\Omega)) \cap C^0(W^{3,\infty}(\Omega))$ *and* $\mathbf{w} \in C^i(H^{3-i}(\Omega))$, $i = 0, 1, 2$, *then* $\vartheta^{n+\frac{1}{2}} \in H^1(\Omega)$ *and* $\|\text{div } \vartheta^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1}\ \Delta t^2\ \|\text{div } \mathbf{w}\|_{Z^2}$, $n = 0, \ldots, N-1$.

Proof. The proof follows from applying the Taylor expansion to the auxiliary vector function $G(\tau) := \mathbf{F}_e^{-1}(\mathbf{x}, t_{n+1}; \tau)\mathbf{w}(X_e(\mathbf{x}, t_{n+1}; \tau), \tau)$. ∎

LEMMA 5.13. *Let us assume that* $\mathbf{v} \in C^1(W^{2,\infty}(\Omega)) \cap C^0(W^{3,\infty}(\Omega))$ *and vanishes on the boundary. Let* $\mathbf{w} : \Omega \times [0, T] \longrightarrow \mathbb{R}^m$, $\mathbf{w} \in Z^2$, *be a given function, and let* $\xi^{n+\frac{1}{2}} : \Omega \longrightarrow \mathbb{R}^m$ *be defined by*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) := \text{div } \mathbf{F}_e^{-\mathrm{T}}(\mathbf{x}, t_{n+1}; t_{n+\frac{1}{2}}) \cdot \mathbf{w}(X_e^{n+\frac{1}{2}}(\mathbf{x}), t_{n+\frac{1}{2}}) - \frac{\text{div } (\mathbf{F}_e^n)^{-\mathrm{T}}(\mathbf{x})\mathbf{w}(X_e^n(\mathbf{x}), t_n)}{2}.$$

Then $\xi^{n+\frac{1}{2}} \in L^2(\Omega)$ *and* $\|\xi^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1}\Delta t^2 \|\mathbf{w}\|_{Z^2}$, $n = 0, \ldots, N-1$.

Proof. The proof follows from applying the Taylor formula to the auxiliary scalar function $G(\tau) := \text{div } \mathbf{F}_e^{-\mathrm{T}}(\mathbf{x}, t_{n+1}; \tau) \cdot \mathbf{w}(X_e(\mathbf{x}, t_{n+1}; \tau), \tau)$. ∎

LEMMA 5.14. *Assume that* $\mathbf{v} \in C^1(L^\infty(\Omega)) \cap C^0(W^{1,\infty}(\Omega))$ *vanishes on the boundary and* $\Delta t\|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))} < 1$. *Let* $\varphi \in H^1(\Omega)$ *and let* ξ^n *be defined by*

$$\xi^n(\mathbf{x}) := \varphi(X_e^n(\mathbf{x})) - \varphi(X_E^n(\mathbf{x})) \quad \textit{for a.e. } \mathbf{x} \in \Omega, \quad n = 0, \ldots, N-1.$$

Then $\xi^n \in L^2(\Omega)$ *and*

$$\xi^n(\mathbf{x}) = \nabla\ \varphi(X_E^n(\mathbf{x}) + \theta\ \mathbf{y}) \cdot \mathbf{y} \quad \textit{for a.e. } \mathbf{x} \in \Omega, \tag{5.42}$$

with $\theta \in (0, 1)$, *and*

$$\mathbf{y} = \int_{t_n}^{t_{n+1}} (s - t_n)\ \frac{\mathsf{d}\mathbf{v}}{\mathsf{d}t}(X_e(\mathbf{x}, t_{n+1}; s), s)\ ds. \tag{5.43}$$

Moreover, if $\mathbf{v} \in C^1(W^{1,\infty}(\Omega)) \cap C^0(W^{2,\infty}(\Omega))$ *and* $\varphi \in H^2(\Omega)$, *then* $\xi^{n+\frac{1}{2}} \in H^1(\Omega)$.

Proof. We use the following first order Taylor expansion of function $\varphi \in H^1(\Omega)$:

$$\varphi(X_e^n(\mathbf{x}), t_n) = \varphi(X_E^n(\mathbf{x}), t_n) + \nabla\ \varphi\,(X_E^n(\mathbf{x}) + \theta(X_e^n(\mathbf{x}) - X_E^n(\mathbf{x}))) \cdot (X_e^n(\mathbf{x}) - X_E^n(\mathbf{x}))$$

for some number $\theta \in (0, 1)$. Next, by applying again a Taylor expansion to function $X_e(\mathbf{x}, t_{n+1}; \tau)$ it follows that

$$X_e^n(\mathbf{x}) - X_E^n(\mathbf{x}) = \int_{t_{n+1}}^{t_n} (t_n - s)\ \frac{\mathsf{d}\mathbf{v}}{\mathsf{d}t}(X_e(\mathbf{x}, t_{n+1}; s), s)\ ds.$$

By jointly considering both Taylor expansions, (5.42) and (5.43) follow, and the regularity of ξ^n is a consequence of the regularity of φ and $\mathbf{v}$. ∎

COROLLARY 5.15. *Let us assume that* $\mathbf{v} \in C^1(L^\infty(\Omega)) \cap C^0(W^{1,\infty}(\Omega))$ *and vanishes on the boundary, and* $\Delta t\|\mathbf{v}\|_{C^0(W^{1,\infty}(\Omega))} < 1$. *Let* $\varphi : \Omega \times [0, T] \longrightarrow \mathbb{R}$, $\varphi \in Z^2$, *be a given function. Let* $\xi^{n+\frac{1}{2}}$ *be the function defined by*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) := \varphi(X_e^{n+\frac{1}{2}}(\mathbf{x}), t_{n+\frac{1}{2}}) - \frac{\varphi(\mathbf{x}, t_{n+1}) + \varphi(X_E^n(\mathbf{x}), t_n)}{2} \quad \forall \mathbf{x} \in \Omega.$$

Then $\xi^{n+\frac{1}{2}} \in L^2(\Omega)$ *and* $\|\xi^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1}\ \Delta t^2\ \|\varphi\|_{Z^2}$, $n = 0, \ldots, N-1$, *where* $\widetilde{c_1}$ *is independent of* Δt.

Proof. The proof directly follows from applying Lemmas 5.11 and 5.14. ∎

LEMMA 5.16. *Let us assume that* $\mathbf{v} \in C^1(W^{1,\infty}(\Omega)) \cap C^0(W^{2,\infty}(\Omega))$ *and vanishes on the boundary. Let* Ψ^n *be defined by*

$$\Psi^n(\mathbf{x}) := (\mathbf{F}_e^n)^{-1}(X_e^n(\mathbf{x})) - (\mathbf{I} + \Delta t\ \mathbf{L}^n)(X_e^n(\mathbf{x})) \quad \forall \mathbf{x} \in \Omega.$$

Then $\Psi^n \in L^2(\Omega)$ *for* $n = 0, \ldots, N-1$ *and*

(5.44)
$$\Psi^n(\mathbf{x}) = -\int_{t_n}^{t_{n+1}} (\tau - t_{n+1})\ \nabla\ \frac{\mathsf{d}\mathbf{v}}{\mathsf{d}t}(X_e(\mathbf{x}, t_{n+1}; \tau), \tau)\mathbf{F}_e(X_e^n(\mathbf{x}), t_n; \tau)\ d\tau \text{ for a.e. } \mathbf{x} \in \Omega.$$

Moreover, if $\mathbf{v} \in C^1(W^{2,\infty}(\Omega)) \cap C^0(W^{3,\infty}(\Omega))$, *then* $\Psi^n \in H^1(\Omega)$.

Proof. The result directly follows from replacing $t = t_{n+1}$ and $s = t_n$ in (3.7). ∎

LEMMA 5.17. *Let us assume that* $\mathbf{v} \in C^1(W^{2,\infty}(\Omega)) \cap C^0(W^{3,\infty}(\Omega))$ *and vanishes on the boundary. Let* ϑ^n *be defined by*

$$\vartheta^n(\mathbf{x}) := \text{div}\ (\mathbf{F}_e^n)^{-\mathrm{T}}(X_e^n(\mathbf{x})) - \Delta t \nabla \text{div}\ \mathbf{v}^n(X_e^n(\mathbf{x})) \quad \forall \mathbf{x} \in \Omega.$$

Then $\vartheta^n \in L^2(\Omega)$ *for* $n = 0, \ldots, N-1$ *and*

$$\vartheta^n(\mathbf{x}) = -\int_{t_n}^{t_{n+1}} (\tau - t_{n+1})\ \text{div}\ \left((\mathbf{F}_e)^{\mathrm{T}} \left(\nabla\ \frac{\mathsf{d}\mathbf{v}}{\mathsf{d}t} \right)^{\mathrm{T}} \right) (X_e(\mathbf{x}, t_{n+1}; \tau)\ d\tau \text{ for a.e. } \mathbf{x} \in \Omega.$$

Proof. The assumed regularity for $\mathbf{v}$ allows us to apply Lemma 5.16 and to define $\Psi^n \in H^1(\Omega)$ by (5.44). The result directly follows from applying the div operator to Ψ^n and by taking into account that $\text{div}\left(\mathbf{I} + \Delta t\ (\mathbf{L}^n)^{\mathrm{T}}\right)(\mathbf{y}) = \Delta t \nabla \text{div}\ \mathbf{v}^n(\mathbf{y})$. ∎

Whereas the previous lemmas concern the approximation of functions defined on the whole domain Ω, the one below gives a second order approximation of a function defined on the boundary Γ_R.

LEMMA 5.18. *Let* $\mathbf{v} \in C^0(W^{2,\infty}(\Omega)) \cap C^1(W^{1,\infty}(\Omega))$ *and vanish on the boundary. Let* $\varphi : \Gamma_R \times [0,T] \longrightarrow \mathbb{R}$, $\varphi \in C^2(L^2(\Gamma_R))$, *and let* $\xi^{n+\frac{1}{2}}$ *be a function defined on the boundary* Γ_R *by*

$$\xi^{n+\frac{1}{2}}(\mathbf{x}) := \mathsf{det}\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1}(\mathbf{x})\ \varphi(\mathbf{x}, t_{n+\frac{1}{2}}) - \frac{\varphi(\mathbf{x}, t_{n+1}) + \left(1 + \Delta t\ \text{div}\ \mathbf{v}^{n+1}(\mathbf{x})\right)\varphi(\mathbf{x}, t_n)}{2}$$

for a.e. $\mathbf{x} \in \Gamma_R$. *Then* $\xi^{n+\frac{1}{2}} \in L^2(\Gamma_R)$ *and* $\|\xi^{n+\frac{1}{2}}\|_{0,\Gamma_R} \leq \widetilde{c_1} \Delta t^2 \|\varphi\|_{C^2(L^2(\Gamma_R))}$, $n = 0, \ldots, N-1$, *where* $\widetilde{c_1}$ *is independent of* Δt.

Proof. Let us first write $\xi^{n+\frac{1}{2}}(\mathbf{x}) = \xi_1^{n+\frac{1}{2}}(\mathbf{x}) - \xi_2^n(\mathbf{x})$ with

$$\xi_1^{n+\frac{1}{2}}(\mathbf{x}) := \mathsf{det}\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1}(\mathbf{x})\varphi(\mathbf{x}, t_{n+\frac{1}{2}}) - \frac{\varphi(\mathbf{x}, t_{n+1}) + \mathsf{det}\ (\mathbf{F}_e^n)^{-1}(\mathbf{x})\varphi(\mathbf{x}, t_n)}{2},$$
$$\xi_2^n(\mathbf{x}) := \frac{\mathsf{det}\ (\mathbf{F}_e^n)^{-1}(\mathbf{x})\varphi(\mathbf{x}, t_n) - \left(1 + \Delta t\ \text{div}\ \mathbf{v}^{n+1}(\mathbf{x})\right)\varphi(\mathbf{x}, t_n)}{2}.$$

The proof is achieved by means of Taylor expansions, expressions (3.8) and (3.9) for $t = t_{n+1}$, and Proposition 3.5 for $t = t_{n+1}$ and $s = t_n$. ∎

LEMMA 5.19. *Assume Hypotheses* 3, 4, *and* 5 *hold, and that the coefficients of the problem satisfy* $\mathbf{v} \in C^0(W^{3,\infty}(\Omega)) \cap C^1(W^{2,\infty}(\Omega)) \cap C^2(L^\infty(\Omega))$, $\mathbf{v}|_\Gamma = 0$, $\mathbf{A} \in W^{3,\infty}(\Omega)$, $r \in W^{2,\infty}(\Omega)$, *and that* $\Delta t \|\mathbf{v}\|_{C^0(W^{1,\infty})} < 1/2$. *Let the solution of* (5.41)

satisfy $\phi \in Z^3$, $\nabla\phi \in Z^3$, $\phi|_{\Gamma_R} \in C^2(L^2(\Gamma_R))$. *Then, for each* $n = 0, 1, \ldots, N-1$, *there exist two functions* $\xi_{\mathcal{L}_1}^{n+\frac{1}{2}} : \Omega \longrightarrow \mathbb{R}$ *and* $\xi_{\mathcal{L}_2}^{n+\frac{1}{2}} : \Gamma_R \longrightarrow \mathbb{R}$, *such that*

$$\left\langle \left(\mathcal{L}^{n+\frac{1}{2}} - \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \right) \widehat{\phi}, \psi \right\rangle = \left\langle \xi_{\mathcal{L}_1}^{n+\frac{1}{2}}, \psi \right\rangle + \left\langle \xi_{\mathcal{L}_2}^{n+\frac{1}{2}}, \psi \right\rangle_{\Gamma_R} \tag{5.45}$$

$\forall \psi \in H^1_{\Gamma_D}(\Omega)$. *Moreover,* $\xi_{\mathcal{L}_1}^{n+\frac{1}{2}} \in L^2(\Omega)$ *and* $\xi_{\mathcal{L}_2}^{n+\frac{1}{2}} \in L^2(\Gamma_R)$ *and the following estimates hold:*

$$\left\| \xi_{\mathcal{L}_1}^{n+\frac{1}{2}} \right\|_0 \leq \widetilde{c_1} \Delta t^2 \left(\|\phi\|_{Z^3} + \|\mathbf{A}\nabla\phi\|_{Z^3} + \|r\phi\|_{Z^2} \right),$$
$$\left\| \xi_{\mathcal{L}_2}^{n+\frac{1}{2}} \right\|_{0,\Gamma_R} \leq \widetilde{c_1} \Delta t^2 \left(\|\mathbf{A}\nabla\phi \cdot \mathbf{n}\|_{Z^2,\Gamma_R} + \alpha \|\phi\|_{C^2(L^2(\Gamma_R))} \right),$$

where $\widetilde{c_1}$ *denotes a constant independent of* Δt *and* $\alpha > 0$ *appears in* (2.3).

Proof. The left-hand side of (5.45) is equal to $I_1 + I_2 + I_3 + I_4 + I_5$, with

$$I_1 = \left\langle \left(\frac{\mathsf{d}\phi}{\mathsf{d}t} \right)^{n+\frac{1}{2}} \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle - \left\langle \frac{\phi^{n+1} - \phi^n \circ X_{RK}^n}{\Delta t}, \psi \right\rangle,$$
$$I_2 = \left\langle \left(\mathbf{F}_e^{n+\frac{1}{2}} \right)^{-1} \left(\mathbf{A} \nabla \phi^{n+\frac{1}{2}} \right) \circ X_e^{n+\frac{1}{2}}, \nabla\psi \right\rangle$$
$$- \left\langle \frac{\mathbf{A}\nabla\phi^{n+1} + ((\mathbf{I} + \Delta t\, \mathbf{L}^n)\mathbf{A}\nabla\phi^n) \circ X_E^n}{2}, \nabla\psi \right\rangle,$$
$$I_3 = \left\langle \mathrm{div} \left(\mathbf{F}_e^{n+\frac{1}{2}} \right)^{-\mathrm{T}} \cdot \left(\mathbf{A}\nabla\phi^{n+\frac{1}{2}} \right) \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle - \left\langle \frac{\Delta t \left(\nabla \mathrm{div}\ \mathbf{v}^n \cdot \mathbf{A}\nabla\phi^n \right) \circ X_E^n}{2}, \psi \right\rangle,$$
$$I_4 = \left\langle (r\phi^{n+\frac{1}{2}}) \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle - \left\langle \frac{r\phi^{n+1} + (r\phi^n) \circ X_E^n}{2}, \psi \right\rangle,$$
$$I_5 = \alpha \left\langle \mathsf{det} \left(\mathbf{F}_e^{n+\frac{1}{2}} \right)^{-1} \phi^{n+\frac{1}{2}}, \psi \right\rangle_{\Gamma_R} - \alpha \left\langle \frac{\phi^{n+1} + \phi^n (1 + \Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})}{2}, \psi \right\rangle_{\Gamma_R}.$$

The bound for I_1 directly follows from Lemma 5.10 for $\varphi = \phi$, so we can define a function $\xi_{I_1}^{n+\frac{1}{2}} \in L^2(\Omega)$ such that

$$I_1 = \left\langle \xi_{I_1}^{n+\frac{1}{2}}, \psi \right\rangle \quad \text{with} \quad \|\xi_{I_1}^{n+\frac{1}{2}}\|_0 \leq \widetilde{c_1} \Delta t^2 \|\phi\|_{Z^3}. \tag{5.46}$$

Term I_2 is written as $I_2 = I_2^1 + I_2^2 + I_2^3$, where

$$I_2^1 = \left\langle \left(\mathbf{F}_e^{n+\frac{1}{2}} \right)^{-1} \left(\mathbf{A} \nabla \phi^{n+\frac{1}{2}} \right) \circ X_e^{n+\frac{1}{2}} - \frac{\mathbf{A}\nabla\phi^{n+1} + (\mathbf{F}_e^n)^{-1} (\mathbf{A}\nabla\phi^n) \circ X_e^n}{2}, \nabla\ \psi \right\rangle,$$
$$I_2^2 = \left\langle \frac{(\mathbf{F}_e^n)^{-1} (\mathbf{A}\nabla\phi^n) \circ X_e^n - ((\mathbf{I} + \Delta t\, \mathbf{L}^n)\mathbf{A}\nabla\phi^n) \circ X_e^n}{2}, \nabla\psi \right\rangle,$$
$$I_2^3 = \left\langle \frac{((\mathbf{I} + \Delta t\, \mathbf{L}^n)\mathbf{A}\nabla\phi^n) \circ X_e^n - ((\mathbf{I} + \Delta t\, \mathbf{L}^n)\mathbf{A}\nabla\phi^n) \circ X_E^n}{2}, \nabla\psi \right\rangle.$$

In order to estimate I_2^1 we apply Lemma 5.12 to $\mathbf{w} = \mathbf{A} \nabla \phi \in C^i(H^{3-i}(\Omega))$ for $i = 0, 1, 2$, so a vector valued function $\vartheta_{I_2^1}^{n+\frac{1}{2}} \in H^1(\Omega)$ can be defined and Green's formula can be applied. Thus, we have

$$I_2^1 = \left\langle \vartheta_{I_2^1}^{n+\frac{1}{2}}, \nabla\ \psi \right\rangle = \left\langle \vartheta_{I_2^1}^{n+\frac{1}{2}} \cdot \mathbf{n}, \psi \right\rangle_{\Gamma_R} - \left\langle \mathrm{div}\ \vartheta_{I_2^1}^{n+\frac{1}{2}}, \psi \right\rangle,$$

where the involved functions are bounded as follows:

(5.47)
$$\left\|\vartheta_{I_2^1}^{n+\frac{1}{2}}\cdot\mathbf{n}\right\|_{0,\Gamma_R}\leq\widetilde{c_1}\Delta t^2\|\mathbf{A}\nabla\phi\cdot\mathbf{n}\|_{Z^2,\Gamma_R},\qquad\left\|\text{div }\vartheta_{I_2^1}^{n+\frac{1}{2}}\right\|_0\leq\widetilde{c_1}\Delta t^2\left\|\text{div }\mathbf{A}\ \nabla\ \phi\right\|_{Z^2}.$$

For I_2^2 we apply Lemma 5.16 for $\mathbf{v}\in C^1(W^{2,\infty}(\Omega))\cap C^0(W^{3,\infty}(\Omega))$, finding a matrix valued function $\Psi_{I_2^2}^n\in H^1(\Omega)$ satisfying

$$\begin{aligned}I_2^2&=\left\langle\Psi_{I_2^2}^n\left(\mathbf{A}\nabla\phi\right)\circ X_e^n,\nabla\ \psi\right\rangle\\&=\left\langle\Psi_{I_2^2}^n\left(\mathbf{A}\nabla\phi\right)\circ X_e^n\cdot\mathbf{n},\ \psi\right\rangle_{\Gamma_R}-\left\langle\text{div}\left(\Psi_{I_2^2}^n\left(\mathbf{A}\nabla\phi\right)\circ X_e^n\right),\ \psi\right\rangle,\end{aligned}$$

where the Green formula has been used for the last equality and with functions on the left-hand side of the inner products bounded as in (5.47). For I_2^3 we apply Lemma 5.14 componentwise to $\varphi^n(\mathbf{x})=[(\mathbf{I}+\Delta t\ \mathbf{L}^n(\mathbf{x}))\mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x})]_i\in H^2(\Omega)$, and we find a function $\vartheta_{I_2^3}^n\in H^1(\Omega)$, to which we can apply Green's formula obtaining bounds analogous to (5.47). Summing up, we can write I_2 as

$$I_2=\left\langle\xi_{I_2^A}^{n+\frac{1}{2}},\psi\right\rangle+\left\langle\xi_{I_2^B}^{n+\frac{1}{2}},\psi\right\rangle_{\Gamma_R},\tag{5.48}$$

where $\|\xi_{I_2^A}^{n+\frac{1}{2}}\|_0\leq\widetilde{c_1}\Delta t^2\ \|\mathbf{A}\ \nabla\ \phi\|_{Z^3}$ and $\|\xi_{I_2^B}^{n+\frac{1}{2}}\|_{0,\Gamma_R}\leq\widetilde{c_1}\Delta t^2\ \|\mathbf{A}\nabla\phi\cdot\mathbf{n}\|_{Z^2,\Gamma_R}$.

Similar computations with I_3, by using Lemmas 5.13, 5.17, and 5.14, lead to

$$I_3=\left\langle\xi_{I_3}^{n+\frac{1}{2}},\psi\right\rangle,\quad\text{with}\quad\left\|\xi_{I_3}^{n+\frac{1}{2}}\right\|_0\leq\widetilde{c_1}\Delta t^2\ \|\mathbf{A}\nabla\phi\|_{Z^2}.\tag{5.49}$$

For $\xi=r\phi\in Z^2$ we can apply Corollary 5.15 to I_4, obtaining

$$I_4=\left\langle\xi_{I_4}^{n+\frac{1}{2}},\psi\right\rangle,\quad\text{with}\quad\left\|\xi_{I_4}^{n+\frac{1}{2}}\right\|_0\leq\widetilde{c_1}\Delta t^2\ \|r\phi\|_{Z^2}.\tag{5.50}$$

The estimate for I_5 follows from Lemma 5.18 for $\xi=\alpha\phi|_{\Gamma_R}\in C^2(L^2(\Gamma_R))$:

$$I_5=\left\langle\xi_{I_5}^{n+\frac{1}{2}},\psi\right\rangle_{\Gamma_R},\quad\text{with}\quad\left\|\xi_{I_5}^{n+\frac{1}{2}}\right\|_{0,\Gamma_R}\leq\widetilde{c_1}\Delta t^2\|\phi\|_{C^2(L^2(\Gamma_R))}.\tag{5.51}$$

Finally, partial results (5.46), (5.48), (5.49), (5.50), and (5.51) imply (5.45). ∎

LEMMA 5.20. *Assume that* $\mathbf{v}\in C^0(W^{2,\infty}(\Omega))\cap C^1(W^{1,\infty}(\Omega))$ *vanishes on the boundary and* $\Delta t\|\mathbf{v}\|_{C^0(W^{1,\infty})}<1/2$. *Let* $f\in Z^2$ *and* $g\in C^2(L^2(\Gamma_R))$. *Then, for each* $n=0,1,\ldots,N-1$, *there exist* $\xi_f^{n+\frac{1}{2}}:\Omega\longrightarrow\mathbb{R}$ *and* $\xi_g^{n+\frac{1}{2}}:\Gamma_R\longrightarrow\mathbb{R}$, *satisfying*

$$\left\langle\mathcal{F}^{n+\frac{1}{2}}-\mathcal{F}_{\Delta t}^{n+\frac{1}{2}},\psi\right\rangle=\left\langle\xi_f^{n+\frac{1}{2}},\psi\right\rangle+\left\langle\xi_g^{n+\frac{1}{2}},\psi\right\rangle_{\Gamma_R}\quad\forall\psi\in H^1(\Omega).\tag{5.52}$$

Moreover, $\xi_f\in L^2(\Omega)$, $\xi_g\in L^2(\Gamma_R)$, *and the following estimates hold:*

$$\left\|\xi_f^{n+\frac{1}{2}}\right\|_0\leq\widetilde{c_1}\Delta t^2\left\|f\right\|_{Z^2},\qquad\left\|\xi_g^{n+\frac{1}{2}}(\mathbf{x})\right\|_{0,\Gamma_R}\leq\widetilde{c_1}\Delta t^2\left\|g\right\|_{C^2(L^2(\Gamma_R))},$$

with constant $\widetilde{c_1}$ *independent of* Δt.

Proof. The proof follows from Corollary 5.15 and Lemma 5.18. ∎

Lemmas in this section hold under Hypotheses 3, 4, 5 and the following one.

Hypothesis 8. Functions appearing in problem (2.1)–(2.4) satisfy

- $\mathbf{A} \in W^{3,\infty}(\Omega)$, $r \in W^{2,\infty}(\Omega)$;
- $\mathbf{v} \in C^0(W^{3,\infty}(\Omega)) \cap C^1(W^{2,\infty}(\Omega)) \cap C^2(L^\infty(\Omega))$ and $\mathbf{v}|_\Gamma = 0$;
- $f \in Z^2$, $g \in Z^3(\Gamma_R)$, and $\alpha > 0$.

Remark 5.7. Although in Lemma 5.15 only $g \in C^2(L^2(\Gamma_R))$ was required, more smoothness will be necessary in the following theorem.

THEOREM 5.21 (error estimate). *Assume Hypotheses* 3, 4, 5, *and* 8. *Let* $\phi \in Z^3$ *be the solution of* (5.41), $\nabla\phi \in Z^3$, $\phi|_{\Gamma_R} \in Z^3(\Gamma_R)$. *Let* $\phi_{\Delta t} = \{\phi^n_{\Delta t}\}$ *be the solution of* (5.13) *subject to initial value* $\phi^0_{\Delta t} = \phi^0$ *and* $\mathbf{B}\nabla\widehat{\phi_{\Delta t}} = \{\mathbf{B}\nabla\phi^n_{\Delta t}\}$. *Then there exist two positive constants* c *and* $d = d(c_1, c_2, \delta, c_3, \gamma)$ *such that if* $\Delta t < d$, *we have*

$$
\begin{aligned}
(5.53)\qquad & \sqrt{\frac{1}{2}}\left\|\widehat{\phi} - \widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\Delta t\ \delta}{4}}\left\|\mathbf{B}\nabla\widehat{\phi} - \mathbf{B}\nabla\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} \\
& + \sqrt{\frac{\Delta t}{4}}\left\|\sqrt{r}\,\widehat{\phi} - \sqrt{r}\,\widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\Delta t\ \alpha}{16}}\left\|\widehat{\phi} - \widehat{\phi_{\Delta t}}\right\|_{l^\infty(L^2(\Gamma_R))} \\
& \le c\ \Delta t^2\ \left(\|\phi\|_{Z^3} + \|\mathbf{A}\nabla\phi\|_{Z^3} + \|r\phi\|_{Z^2} + \|\phi\|_{Z^2,\Gamma_R} + \|f\|_{Z^2} + \|g\|_{Z^2,\Gamma_R}\right) \\
& \quad + c\ \Delta t^3\left(\|\phi\|_{Z^3,\Gamma_R} + \|g\|_{Z^3,\Gamma_R}\right).
\end{aligned}
$$

Proof. Let us denote by $e_{\Delta t}$ the difference between the continuous and the discrete solutions, i.e., $e_{\Delta t} = \{e^n_{\Delta t}\}$, with $e^n_{\Delta t} = \phi^n - \phi^n_{\Delta t}$. From (5.41) and (5.13) we have

$$
\left\langle \mathcal{L}^{n+\frac{1}{2}}_{\Delta t}\widehat{e_{\Delta t}}, \psi\right\rangle = \left\langle \left(\mathcal{L}^{n+\frac{1}{2}}_{\Delta t} - \mathcal{L}^{n+\frac{1}{2}}\right)\widehat{\phi}, \psi\right\rangle + \left\langle \mathcal{F}^{n+\frac{1}{2}} - \mathcal{F}^{n+\frac{1}{2}}_{\Delta t}, \psi\right\rangle,
$$

and, as a consequence of Lemmas 5.19 and 5.20, we are led to the following scheme:

$$
(5.54)\quad \left\langle \mathcal{L}^{n+\frac{1}{2}}_{\Delta t}\widehat{e_{\Delta t}}, \psi\right\rangle = \left\langle \xi^{n+\frac{1}{2}}_{\mathcal{L}_1} - \xi^{n+\frac{1}{2}}_{f}, \psi\right\rangle + \left\langle \xi^{n+\frac{1}{2}}_{\mathcal{L}_2} - \xi^{n+\frac{1}{2}}_{g}, \psi\right\rangle_{\Gamma_R} \quad \forall \psi \in H^1_{\Gamma_D}(\Omega).
$$

Next, we apply Theorem 5.8 to (5.54), noting that $e^0_{\Delta t} = 0$. Thus, we obtain

$$
\begin{aligned}
(5.55)\quad & \frac{1}{\sqrt{2}}\|\widehat{e_{\Delta t}}\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\delta\ \Delta t}{4}}\|\mathbf{B}\nabla\widehat{e_{\Delta t}}\|_{l^\infty(L^2(\Omega))} + \sqrt{\frac{\Delta t}{4}}\left\|\sqrt{r}\widehat{e_{\Delta t}}\right\|_{l^\infty(L^2(\Omega))} \\
& + \sqrt{\frac{\alpha\ \Delta t}{16}}\|\widehat{e_{\Delta t}}\|_{l^\infty(L^2(\Gamma_R))} \le c\ \Delta t\left(\left\|\widehat{D_{\Delta t}\xi_{\mathcal{L}_2}}\right\|_{l^2(L^2(\Gamma_R))} + \left\|\widehat{D_{\Delta t}\xi_g}\right\|_{l^2(L^2(\Gamma_R))}\right) \\
& \quad + c\left(\left\|\widehat{\xi_{\mathcal{L}_1}}\right\|_{l^2(L^2(\Omega))} + \left\|\widehat{\xi_f}\right\|_{l^2(L^2(\Omega))} + \left\|\widehat{\xi_{\mathcal{L}_2}}\right\|_{l^2(L^2(\Gamma_R))} + \left\|\widehat{\xi_g}\right\|_{l^2(L^2(\Gamma_R))}\right).
\end{aligned}
$$

Thus, error estimate (5.53) follows from the upper bounds for $\xi_{\mathcal{L}_1}$, ξ_f, $\xi_{\mathcal{L}_2}$, and ξ_g given in Lemmas 5.19 and 5.20 and replacing the Robin boundary condition (5.41). ∎

6. Conclusions. We have performed the numerical analysis of second order characteristic semidiscretized schemes for solving linear convection-diffusion-reaction equations, extending the work in [24]. More precisely, we allow for degenerate diffusion coefficients, reaction terms, non-divergence-free velocity fields, and general Dirichlet–Robin boundary conditions. The method has been introduced by using the formalism of continuum mechanics, and weak formulations by means of an appropriate Green's formula are obtained. Although our analysis considers only velocity fields which are null at the boundary and use approximate characteristic lines, we could also deal with more general situations. Second order error estimates have been obtained when smooth enough data and solutions are available.

This paper is completed in [9], where the fully discretized Lagrange–Galerkin scheme is theoretically studied. Moreover, the effect of different proposed quadrature formulas and some numerical examples illustrating the predicted behavior are shown.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. Amann, *Ordinary Differential Equations. An Introduction to Nonlinear Analysis*, de Gruyter Stud. Math. 13, Walter de Gruyter, Berlin, 1990.

[3] J. Baranger, D. Esslaoui, and A. Machmoum, *Error estimate for convection problem with characteristics method*, Numer. Algorithms, 21 (1999), pp. 49–56.

[4] J. Baranger and A. Machmoum, *Une norme "naturelle" pour la méthode des caractéristiques en éléments finis discontinus: Cas* 1-*D*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 549–574.

[5] J. Baranger and A. Machmoum, *A "natural" norm for the method of characteristics using discontinuous finite elements:* 2*D and* 3*D case*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1223–1240.

[6] E. Barucci, S. Polidoro, and V. Vespri, *Some results on partial differential equations and Asian options*, Math. Models Methods Appl. Sci., 11 (2001), pp. 475–497.

[7] M. Bercovier, O. Pironneau, and V. Sastri, *Finite elements and characteristics for some parabolic-hyperbolic problems*, Appl. Math. Modelling, 7 (1983), pp. 89–96.

[8] A. Bermúdez and J. Durany, *La méthode des charactéristiques pour les problèmes de convection-diffusion stationnaires*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 7–26.

[9] A. Bermúdez, M. R. Nogueiras, and C. Vázquez, *Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. Part* II: *Fully discretized scheme and quadrature formulas*, SIAM J. Numer. Anal., 44 (2006), pp. 1854–1876.

[10] A. Bermúdez, M. R. Nogueiras, and C. Vázquez, *Numerical solution of variational inequalities for pricing Asian options by higher order Lagrange-Galerkin methods*, Appl. Numer. Math., 56 (2006), pp. 1256–1270.

[11] K. Boukir, Y. Maday, and B. Métivet, *A high order characteristics method for the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 116 (1994), pp. 211–218.

[12] K. Boukir, Y. Maday, B. Métivet, and E. Razafindrakoto, *A high-order characteristics/finite element method for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 25 (1997), pp. 1421–1454.

[13] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[14] R. E. Ewing and T. F. Russel, *Multistep Galerkin methods along characteristics for convection-diffusion problems*, in Advances in Computer Methods for Partial Differential Equations IV, R. Vichtneveski and R. S. Stepleman, eds., IMACS, New Brunswick, NJ, 1981, pp. 28–36.

[15] R. E. Ewing and H. Wang, *A summary of numerical methods for time-dependent advection-dominated partial differential equations*, J. Comput. Appl. Math., 128 (2001), pp. 423–445.

[16] M. Gurtin, *An Introduction to Continuum Mechanics*, Math. Sci. Engrg. 158, Academic Press, San Diego, 1981.

[17] L. Hörmander, *Hypoelliptic second order differential equations*, Acta Math., 119 (1967), pp. 147–171.

[18] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Appl. Math. Math. Comput. 12, Chapman & Hall, London, 1996.

[19] K. W. Morton, A. Priestley, and E. Süli, *Stability of the Lagrange-Galerkin method with nonexact integration*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 625–653.

[20] M. R. Nogueiras, *Numerical Analysis of Second Order Lagrange-Galerkin Schemes. Application to Option Pricing Problems*, Ph.D. thesis, University of Santiago de Compostela, 2005.

[21] O. Pironneau, *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations*, Numer. Math., 38 (1981/1982), pp. 309–332.

[22] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[23] M. A. Ragusa, *On weak solutions of ultraparabolic equations*, Nonlinear Anal., 47 (2001), pp.

503–511.

[24] H. Rui and M. Tabata, *A second order characteristic finite element scheme for convection-diffusion problems*, Numer. Math., 92 (2002), pp. 161–177.

[25] E. Süli, *Stability and convergence of the Lagrange-Galerkin method with nonexact integration*, in The Mathematics of Finite Elements and Applications, VI (Uxbridge, 1987), Academic Press, London, 1988, pp. 435–442.

[26] P. Wilmott, J. Dewynne, and S. Howison, *Option Pricing. Mathematical Models and Computation*, Oxford Financial Press, Oxford, 1993.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1854–1876

NUMERICAL ANALYSIS OF CONVECTION-DIFFUSION-REACTION PROBLEMS WITH HIGHER ORDER CHARACTERISTICS/FINITE ELEMENTS. PART II: FULLY DISCRETIZED SCHEME AND QUADRATURE FORMULAS*

ALFREDO BERMÚDEZ†, MARIA R. NOGUEIRAS†, AND CARLOS VÁZQUEZ‡

Abstract. In this paper a higher order Lagrange–Galerkin discretization method is analyzed when applied to a variable coefficient convection-(possibly degenerated) diffusion-reaction equation with mixed Dirichlet–Robin boundary conditions. In a previous paper [A. Bermúdez, M. R. Nogueiras, and C. Vázquez, *SIAM J. Numer. Anal.*, to appear], the proposed second order time discretization scheme has been rigorously introduced for exact and approximated characteristics. Moreover, the $l^\infty(L^2)$ stability property and $l^\infty(L^2)$ error estimates of order $O(\Delta t^2)$ have been obtained. As a continuation of that work, consistency error estimates of order $O(\Delta t^2 + h^k)$ are obtained for the fully discretized Lagrange–Galerkin scheme. Moreover, adequate quadrature formulas are proposed for the practical implementation of the method with particular finite element spaces. Finally, some numerical tests illustrate the theoretical results and the performance of the combination of second order Lagrange–Galerkin schemes with quadrature formulas.

Key words. convection-diffusion equation, Lagrange–Galerkin methods, stability, error estimates, second order schemes, quadrature formulas

AMS subject classifications. 65M12, 65M25, 65M60

DOI. 10.1137/040615109

1. Introduction.

In the framework of numerical solution of convection dominated problems (including the degenerate diffusion case arising, for example,in finance [20]) a possible upwinding strategy is provided by the method of characteristics for time discretization (see [11]). This approach is based on the discretization of the total (or material) time derivative. Many authors have mathematically analyzed and applied the characteristics method to different problems [9, 15, 5, 4, 13, 19, 2, 1, 3].

The increase in the order of time and space approximations can be obtained by using higher order schemes for the discretization of the material derivative and higher order finite element spaces. In [10] multistep Galerkin methods for constant coefficients convection-diffusion problems are studied and the need for analyzing the variable coefficient case is pointed out. In [7, 8] multistep methods to approximate the material time derivative, combined with either mixed finite elements or spectral methods for spatial discretization, are analyzed to solve incompressible Navier–Stokes equations. More recently, in [18], a second order Runge–Kutta method is proposed to approximate the material time derivative when solving a constant coefficient

*Received by the editors September 15, 2004; accepted for publication (in revised form) February 27, 2006; published electronically September 29, 2006. This work was partially supported by Project VEM2003-20069-C03-03 of MCYT.

http://www.siam.org/journals/sinum/44-5/61510.html

†Dep. de Matemática Aplicada, Universidade de Santiago, Campus Sur s/n, 15706-Santiago, Spain (mabermud@usc.es, marianog@usc.es). The second author was supported by Ministerio de Educacion, Cultura y Deporte.

‡Dep. de Matemáticas, Universidade da Coruña, Campus Elviña s/n, 15071-A Coruña, Spain (carlosv@udc.es).

convection-diffusion equation with Dirichlet boundary conditions. Second order in time is achieved by the Crank–Nicholson scheme and an adequate upwinding of the diffusive term.

Our contribution, in both [6] and the present paper, is to extend [18] in several aspects: first, we deal with a (possibly degenerate) variable coefficient diffusive term instead of the more classical Laplacian one. Second, nonzero reaction functions are allowed. Third, a general mixed Dirichlet–Robin boundary condition is considered. Fourth, nondivergence-free velocity fields are handled. Fifth, a complete analysis of the influence of quadrature formulas for different finite element spaces is developed.

In [6] the mathematical formalism of continuum mechanics (see [12]) and Taylor expansions are used to express the results and notations related not only to the approximate characteristics proposed in [18] but also to the exact ones, and an appropriate variational formulation of the problem is obtained. Then, second order characteristics time discretization schemes are introduced. Moreover, the $l^\infty(L^2)$ stability property is stated and $l^\infty(L^2)$ error estimates of order $O(\Delta t^2)$ are obtained.

As a logical continuation of [6], the fully discretized Lagrange–Galerkin scheme with a wide class of finite element spaces is analyzed in the present paper, where the results in [18] are again extended to a more general partial differential equation (PDE) problem. Moreover, adequate quadrature formulas are proposed for the practical implementation with Lagrange finite elements on triangular and quadrangular meshes. Notice that in [18], for piecewise linear Lagrange finite elements, just low order formulas on each element obtained by dividing each mesh triangle were performed. In the present paper, the stability of some of the proposed quadrature formulas is rigorously studied by using Fourier analysis. In this aspect, previous studies about the influence of quadratures in the case of the classical first order Lagrange–Galerkin method applied to transport [13] and convection-diffusion [19] equations are here extended to the second order Lagrange–Galerkin one. Furthermore, some numerical tests illustrate both the theoretical results and the performance of the combination of higher order Lagrange–Galerkin schemes with quadrature formulas.

The present paper is organized as follows. In section 2 the strong formulation of the convection-diffusion-reaction problem is established. In section 3 we introduce the hypotheses on the finite element spaces to be considered for spatial discretization, pose the corresponding fully discretized schemes, and state their stability properties. In section 4, the main result concerning error estimates of the fully discretized schemes is proved. In section 5, stability is analyzed for adequate quadrature formulas combined with particular finite element spaces. Finally, in section 6, several numerical test examples are introduced to illustrate the above theoretical results about the combination of quadrature formulas with second order Lagrange–Galerkin schemes; comparison with first order ones are also included.

2. Statement of the problem, weak formulation, and some hypothesis. Let Ω be a bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$ with Lipschitz boundary, Γ, divided into two parts: $\Gamma = \Gamma_D \cup \Gamma_R$, with $\Gamma_D \cap \Gamma_R = \emptyset$. Let T be a positive constant. We consider the following initial boundary value problem:

(SP) STRONG PROBLEM: *Find a function* $\phi : \Omega \times (0, T) \longrightarrow \mathbb{R}$ *such that*

$$\phi'(\mathbf{x}, t) - \operatorname{div}\left(\mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x}, t)\right) + \mathbf{v}(\mathbf{x}, t) \cdot \nabla\phi(\mathbf{x}, t) + r(\mathbf{x})\phi(\mathbf{x}, t) = f(\mathbf{x}, t) \tag{2.1}$$

for $(\mathbf{x}, t) \in \Omega \times (0, T)$*, subject to boundary conditions*

$$\phi(\mathbf{x},t) = 0 \text{ on } \Gamma_D \times (0,T), \tag{2.2}$$

$$\alpha\ \phi(\mathbf{x},t) + \mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = g(\mathbf{x},t) \text{ on } \Gamma_R \times (0,T), \tag{2.3}$$

and initial condition

$$\phi(\mathbf{x},0) = \phi^0(\mathbf{x}) \text{ in } \Omega. \tag{2.4}$$

In the above equations, ϕ' denotes the partial derivative with respect to t, $\mathbf{A} : \overline{\Omega} \longrightarrow \mathcal{S}_d$ denotes the diffusion matrix function where $\mathcal{S}_d$ is the space of symmetric $d \times d$ matrices, $\mathbf{v} : \overline{\Omega} \times [0,T] \longrightarrow \mathbb{R}^d$ is the velocity vector field, $r : \overline{\Omega} \longrightarrow \mathbb{R}$ is the reaction function, $f : \overline{\Omega} \times [0,T] \longrightarrow \mathbb{R}$ and $g : \Gamma_R \times [0,T] \longrightarrow \mathbb{R}$ are given scalar functions, and $\mathbf{n}$ is the outward unit normal vector to Γ.

Throughout this paper, we use the notation $\psi^i = \psi(t_i)$ for a time-dependent function and $i = 0, \frac{1}{2}, 1, \ldots$. In [6] the following weak formulation of the above problem has been obtained:

(WP) WEAK PROBLEM: *Find a function* $\phi : \Omega \times (0,T) \longrightarrow \mathbb{R}$ *such that*

$$\left\langle \mathcal{L}^{n+\frac{1}{2}}\widehat{\phi}, \psi \right\rangle = \left\langle \mathcal{F}^{n+\frac{1}{2}}, \psi \right\rangle \forall \psi \in H^1_{\Gamma_D}(\Omega), \tag{2.5}$$

where $\mathcal{L}^{n+\frac{1}{2}}\widehat{\phi} \in (H^1(\Omega))'$ *and* $\mathcal{F}^{n+\frac{1}{2}} \in (H^1(\Omega))'$ *are defined by*

$$\begin{aligned}
\left\langle \mathcal{L}^{n+\frac{1}{2}}\widehat{\phi}, \psi \right\rangle :=& \left\langle \left(\frac{d\phi}{dt}\right)^{n+\frac{1}{2}} \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle + \left\langle \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1} \left(\mathbf{A}\nabla\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}},\ \nabla\ \psi \right\rangle \\
&+ \left\langle \text{div}\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-\mathrm{T}} \cdot \left(\mathbf{A}\nabla\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle \\
&+ \left\langle \left(r\phi^{n+\frac{1}{2}}\right) \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle + \alpha \left\langle \det\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1} \phi^{n+\frac{1}{2}}, \psi \right\rangle_{\Gamma_R}, \\
\left\langle \mathcal{F}^{n+\frac{1}{2}}, \psi \right\rangle :=& \left\langle f^{n+\frac{1}{2}} \circ X_e^{n+\frac{1}{2}}, \psi \right\rangle + \left\langle \det\ \left(\mathbf{F}_e^{n+\frac{1}{2}}\right)^{-1} g^{n+\frac{1}{2}}, \psi \right\rangle_{\Gamma_R} \qquad \forall \psi \in H^1(\Omega),
\end{aligned}$$

where $X_e(\mathbf{x},t;\cdot)$ *denotes the characteristic line (associated to* $\mathbf{v}$*) through* $(\mathbf{x},t)$ *and* $\mathbf{F}_e$ *denotes the gradient of* X_e *with respect to* $\mathbf{x}$.

We will adopt the usual notation for the functional spaces involved, which has been recalled in section 1 of [6]. Let us only recall that, for a nonnegative integer m, $Z^m = \left\{\varphi \in C^j(H^{m-j}(\Omega));\ j = 0, \ldots, m\right\}$ is a Banach space when equipped with the norm $\|\varphi\|_{Z^m} := \max\left\{\|\varphi\|_{C^j(H^{m-j})};\ 0 \le j \le m\right\}$.

In [6] a second order characteristics semidiscretized scheme has been proposed and analyzed, obtaining stability and consistency error results under the following hypothesis on the data of the problem:

Hypothesis 1. The velocity field $\mathbf{v} \in C^0\left(W^{2,\infty}(\Omega)\right)$ satisfies $\mathbf{v} = 0$ on Γ.

Remark 2.1. Throughout this paper c_1 denotes the maximum between the positive constant appearing in Lemma 5.4 in [6] and the norm of the velocity in $C^0(W^{2,\infty}(\Omega))$.

Hypothesis 2. The diffusion matrix coefficients, A_{ij}, belong to $W^{1,\infty}(\Omega)$. Moreover, $\mathbf{A}$ is a $m \times m$ symmetric matrix satisfying

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix}, \tag{2.6}$$

with $\mathbf{A}_{m_1}$ being a positive definite symmetric $m_1 \times m_1$ matrix ($m_1 \geq 1$), and where Θ denotes an appropriate zero matrix. Moreover, there exists a strictly positive constant δ, which is a uniform lower bound for the eigenvalues of $\mathbf{A}_{m_1}$.

As a consequence of Hypothesis 2, there exists a unique positive definite symmetric $m_1 \times m_1$ matrix function, $\mathbf{C}_{m_1}$, such that $\mathbf{A}_{m_1} = (\mathbf{C}_{m_1})^2$. Let us denote by $\mathbf{C}$ the symmetric and positive semidefinite $m \times m$ matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix}. \tag{2.7}$$

Notice that $\mathbf{A} = \mathbf{C}^2$ and $C_{ij} \in W^{1,\infty}(\Omega)$. Then, let us introduce the constant $c_2 := \max_{i,j} \left\{ \|C_{ij}\|^2_{W^{1,\infty}(\Omega)} \right\}$. Next, let us denote by $\mathbf{B}$ the $m \times m$ matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{I}_{m_1} & \Theta \\ \Theta & \Theta \end{pmatrix}, \tag{2.8}$$

where $\mathbf{I}_{m_1}$ is the $m_1 \times m_1$ identity matrix. Clearly, under Hypothesis 2 we have

$$\delta \|\mathbf{B}\mathbf{w}\|_0^2 \leq \langle \mathbf{A}\mathbf{w}, \mathbf{w} \rangle = \|\mathbf{C}\mathbf{w}\|_0^2 \leq c_2 \|\mathbf{B}\mathbf{w}\|_0^2 \quad \forall \mathbf{w} \in \mathbb{R}^m. \tag{2.9}$$

Hypothesis 3. The velocity field satisfies $(\mathbf{I} - \mathbf{B})\mathbf{L}(\mathbf{x}, t)\mathbf{B} = \mathbf{0}$ for all $(\mathbf{x}, t) \in \Omega \times [0, T]$, where $\mathbf{L}$ denotes the gradient of $\mathbf{v}$ with respect to $\mathbf{x}$.

Remark 2.2. Under Hypotheses 2 and 3, for every $m \times m$ matrix $\mathbf{E}$ and vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^m$ it is easy to verify that $\langle \mathbf{E}\mathbf{A}\mathbf{w}_1, \mathbf{w}_2 \rangle = \langle \mathbf{E}\mathbf{A}\mathbf{w}_1, \mathbf{B}\mathbf{w}_2 \rangle$.

Hypothesis 4. The reaction function, $r \in W^{1,\infty}(\Omega)$, satisfies $0 < \gamma \leq r(\mathbf{x})$ in Ω, where γ is a constant.

Under the previous hypothesis, let $c_3 := \|\sqrt{r}\|^2_{W^{1,\infty}(\Omega)}$.

Hypothesis 5. Functions appearing in problem (2.1)–(2.4) satisfy

- $\mathbf{A} \in \mathbf{W}^{3,\infty}(\Omega)$, $r \in W^{2,\infty}(\Omega)$,
- $\mathbf{v} \in C^0(\mathsf{W}^{3,\infty}(\Omega)) \cap C^1(\mathsf{W}^{2,\infty}(\Omega)) \cap C^2(\mathsf{L}^\infty(\Omega))$ and $\mathbf{v}|_\Gamma = 0$,
- $f \in Z^2$, $g \in Z^3(\Gamma_R)$, and $\alpha > 0$.

In the next section we introduce spatial finite elements discretizations of the time semidiscretized scheme proposed in [6]. In other words, we propose and analyze different Lagrange–Galerkin schemes.

3. Space discretization: Finite element method. We propose a spatial discretization by using finite element spaces V_h^k, where h denotes the mesh parameter and the positive integer k is the "approximation degree" in the following sense:

Hypothesis 6. There exists an interpolation operator $\pi_h : C^0(\overline{\Omega}) \longrightarrow V_h^k$ satisfying

$$\|\pi_h \psi - \psi\|_s \leq K\, h^{k+1-s} \|\psi\|_{k+1} \qquad \forall \psi \in C^0(\overline{\Omega}) \cap H^{k+1}(\Omega),\ s = 1, 2, \tag{3.1}$$

for a positive constant K independent of h.

The fully discrete scheme reads

$$\begin{cases} \text{Given } \phi_h^0 \in V_h^k, \text{ find } \widehat{\phi_h} := \{\phi_h^n\}_{n=1}^N \in \left[V_h^k\right]^N \text{ such that} \\ \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\phi_h}, \psi_h \right\rangle = \left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}}, \psi_h \right\rangle \quad \forall \psi_h \in V_h^k, \text{ for } n = 0, \ldots, N-1. \end{cases} \tag{3.2}$$

The involved operators are defined as follows:

$$\begin{aligned}\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\phi,\psi\right\rangle :=& \left\langle \frac{\phi^{n+1}-\phi^n\circ X_{RK}^n}{\Delta t},\psi\right\rangle + \left\langle \frac{\mathbf{A}\nabla\phi^{n+1}+(\mathbf{A}\nabla\phi^n)\circ X_E^n}{2},\nabla\psi\right\rangle \\ &+ \frac{\Delta t}{2}\left\langle (\mathbf{L}^n\mathbf{A}\nabla\phi^n)\circ X_E^n,\nabla\psi\right\rangle + \frac{\Delta t}{2}\left\langle (\nabla \text{div } \mathbf{v}^n\cdot\mathbf{A}\nabla\phi^n)\circ X_E^n,\psi\right\rangle \\ &+ \left\langle \frac{r\phi^{n+1}+(r\phi^n)\circ X_E^n}{2},\psi\right\rangle \\ &+ \alpha\left\langle \frac{\phi^{n+1}+\phi^n(1+\Delta t \text{ div } \mathbf{v}^{n+1})}{2},\psi\right\rangle_{\Gamma_R}, \end{aligned} \tag{3.3}$$

$$\left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}},\psi\right\rangle := \left\langle \frac{f^{n+1}+f^n\circ X_E^n}{2},\psi\right\rangle + \left\langle \frac{g^{n+1}+g^n(1+\Delta t \text{ div } \mathbf{v}^{n+1})}{2},\psi\right\rangle_{\Gamma_R}, \tag{3.4}$$

for $\phi\in C^0(H^1(\Omega))$ and $\psi\in H^1(\Omega)$. Functions X_E^n and X_{RK}^n are, respectively, the Euler and Runge–Kutta approximations of the characteristics lines, X_e^n, (see [6]).

The following stability result for the fully discretized scheme can be analogously obtained to the one stated in [6] for the semidiscretized scheme.

THEOREM 3.1 (stability). *Let us assume Hypotheses* 1, 2, 3, *and* 4 *and let* $f\in C^0(L^2(\Omega))$, $g\in C^0(L^2(\Gamma_R))$ *and* $\alpha>0$. *Moreover, let* $\widehat{\phi_h}=\{\phi_h^n\}_{n=1}^N$ *be the solution of* (3.2) *subject to initial value* ϕ_h^0 *and* $\mathbf{B}\nabla\widehat{\phi_h}:=\{\mathbf{B}\nabla\phi_h^n\}_{n=1}^N$. *Then, there exist two positive constants,* c *and* d, $d=d(c_1,c_2,c_3,\delta,\gamma)$, *such that for* $\Delta t<d$ *we have*

$$\begin{aligned}&\frac{1}{\sqrt{2}}\left\|\widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))}+\sqrt{\frac{\delta\ \Delta t}{4}}\left\|\mathbf{B}\nabla\widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))}+\sqrt{\frac{\Delta t}{4}}\left\|\sqrt{r}\widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))} \\ &+\sqrt{\frac{\alpha\ \Delta t}{8}}\left\|\widehat{\phi_h}\right\|_{l^\infty(L^2(\Gamma_R))} \quad\le\quad c\left(\frac{1}{2}\left\|\phi_h^0\right\|_0+\sqrt{\frac{\delta\ \Delta t}{4}}\left\|\mathbf{B}\nabla\phi_h^0\right\|_0+\sqrt{\frac{\Delta t}{4}}\left\|\sqrt{r}\phi_h^0\right\|_0\right. \\ &\left.+\sqrt{\frac{\alpha\ \Delta t}{8}}\left\|\phi_h^0\right\|_{0,\Gamma_R}+\|f\|_{l^2(L^2(\Omega))}+\|g\|_{l^2(L^2(\Gamma_R))}+\Delta t\left\|\widehat{D_{\Delta t}g}\right\|_{l^2(L^2(\Gamma_R))}\right).\end{aligned}$$

4. Error estimates for the fully discretized scheme. In order to study consistency errors of the fully discretized scheme (3.2) let us introduce the notations $\widehat{e_h}:=\widehat{\phi_h}-\widehat{\pi_h\phi}$ and $\widehat{\eta_h}:=\widehat{\phi}-\widehat{\pi_h\phi}$, and state the following lemma.

LEMMA 4.1. *Under Hypotheses* 1, 2, 3, 4, *and* 6, *if* $\phi\in C^0(C^0(\overline{\Omega}))\cap C^0(H^{k+1}(\Omega))\cap H^1(H^k(\Omega))$ *and* $c_1\Delta t<1/2$, *the following inequality holds:*

$$\begin{aligned}&\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}}\widehat{\eta_h},e_h^{n+1}\right\rangle \\ &\le\frac{1}{8}\left\|\mathbf{C}\ \nabla\ e_h^{n+1}+(\mathbf{C}\ \nabla\ e_h^n)\circ X_E^n\right\|_0^2+D_{\Delta t}^n\left(\frac{\Delta t}{2}\left\langle \mathbf{C}\ \nabla\ \widehat{\eta_h},\mathbf{C}\ \nabla\ \widehat{e_h}\right\rangle\right) \\ &+\frac{1}{8}\left\|\sqrt{r}e_h^{n+1}+(\sqrt{r}e_h^n)\circ X_E^n\right\|_0^2+D_{\Delta t}^n\left(\frac{\Delta t}{2}\left\langle \sqrt{r}\widehat{\eta_h},\sqrt{r}e_h^n\right\rangle\right) \\ &+\frac{\alpha}{8}\left\|e_h^{n+1}+e_h^n(1+\Delta t \text{ div } \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2+D_{\Delta t}^n\left(\frac{\alpha\ \Delta t}{2}\left\langle \widehat{\eta_h},\widehat{e_h}\right\rangle_{\Gamma_R}\right)+c\left\|e_h^{n+1}\right\|_0^2 \\ &+c\Delta t\left(\delta\left(\|\mathbf{B}\ \nabla\ e_h^n\|_0^2+\left\|\mathbf{B}\ \nabla\ e_h^{n+1}\right\|_0^2\right)+\left\|\sqrt{r}e_h^n\right\|_0^2+\left\|\sqrt{r}e_h^{n+1}\right\|_0^2+\alpha\left\|e_h^n\right\|_{0,\Gamma_R}^2\right) \\ &+\widetilde{c}K^2h^{2k}\left(\frac{1}{\Delta t}\|\phi'\|_{L^2((t_n,t_{n+1}),H^k(\Omega))}^2+\frac{1}{\Delta t}\|\phi\|_{L^2((t_n,t_{n+1}),H^{k+1}(\Omega))}^2+\Delta t\left\|\phi^n\right\|_{k+1}^2\right)\end{aligned} \tag{4.1}$$

with $c = \max\{1, (c_1c_2+1)/4\delta, c_1c_3/4, c_1\}$, $\widetilde{c}$ *a positive constant, and* $\delta > 0$ *and* $\alpha > 0$ *being the constants appearing, respectively, in Hypothesis* 2 *and* (2.3).

Proof. First, the left-hand side in (4.1) is decomposed as a sum of the terms

$$
\begin{aligned}
I_1 &= \left\langle \frac{\eta_h^{n+1} - \eta_h^n \circ X_{RK}^n}{\Delta t}, e_h^{n+1} \right\rangle, \\
I_2 &= \left\langle \frac{\mathbf{A} \nabla \eta_h^{n+1} + (\mathbf{A} \nabla \eta_h^n) \circ X_E^n}{2}, \nabla e_h^{n+1} \right\rangle, \\
I_3 &= \frac{\Delta t}{2} \left\langle (\mathbf{L}^n \mathbf{A} \nabla \eta_h^n) \circ X_E^n, \nabla e_h^{n+1} \right\rangle, \\
I_4 &= \frac{\Delta t}{2} \left\langle (\nabla \operatorname{div} \mathbf{v}^n \cdot \mathbf{A} \nabla \eta_h^n) \circ X_E^n, e_h^{n+1} \right\rangle, \\
I_5 &= \left\langle \frac{r\eta_h^{n+1} + (r\eta_h^n) \circ X_E^n}{2}, e_h^{n+1} \right\rangle, \\
I_6 &= \alpha \left\langle \frac{\eta_h^{n+1} + \eta_h^n (1 + \Delta t \operatorname{div} \mathbf{v}^{n+1})}{2}, e_h^{n+1} \right\rangle_{\Gamma_R}.
\end{aligned}
\tag{4.2}
$$

In order to work with I_1 we introduce the function $Y_{RK}^n(\mathbf{y}, \cdot) : [t_n, t_{n+1}] \longrightarrow \overline{\Omega}$, defined by $Y_{RK}^n(\mathbf{y}, s) := \mathbf{y} - \mathbf{v}\left(\mathbf{y} - \mathbf{v}^{n+1}(\mathbf{y}) \frac{t_{n+1}-s}{2}, \frac{t_{n+1}+s}{2}\right)(t_{n+1} - s)$. First, applying the chain rule in order to compute its partial derivative with respect to s and its gradient, for $c_1 \Delta t < 1$ it is easy to obtain the following bounds:

$$
\left| \frac{\partial Y_{RK}^n}{\partial s}(\mathbf{y}, s) \right| \leq c_1 + \frac{\Delta t}{2} c_1 + \frac{\Delta t}{2} c_1^2 \leq 1 + 2c_1 \quad \forall (\mathbf{y}, s) \in \Omega \times [t_n, t_{n+1}], \tag{4.3}
$$

$$
\left| \det \left(\nabla Y_{RK}^n \right)^{-1} (\mathbf{y}, s) \right| \leq 1 + c_1 \Delta t \quad \forall (\mathbf{y}, s) \in \Omega \times [t_n, t_{n+1}], \tag{4.4}
$$

where Lemma 5.2 in [6] and Corollary 5.3 in [6] have been used. Moreover, noting that $Y_{RK}^n(\mathbf{y}, t_{n+1}) = \mathbf{y}$ and $Y_{RK}^n(\mathbf{y}, t_n) = X_{RK}^n(\mathbf{y})$, and by using Barrow's rule we have

$$
\begin{aligned}
\frac{\eta_h^{n+1}(\mathbf{y}) - \eta_h^n \left(X_{RK}^n(\mathbf{y}) \right)}{\Delta t} &= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \frac{d\eta_h}{ds} \left((Y_{RK}^n(\mathbf{y}, s), s) \right) \\
= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \eta_h'(Y_{RK}^n(\mathbf{y}, s), s) \, ds &+ \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \nabla \eta_h(Y_{RK}^n(\mathbf{y}, s), s) \cdot \frac{\partial Y_{RK}^n}{\partial s}(\mathbf{y}, s) \, ds,
\end{aligned}
\tag{4.5}
$$

where the chain rule has been used for the last equality. Next, by applying Holder's inequality we get

$$
\frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \eta_h'(Y_{RK}^n(\mathbf{y}, s), s) \, ds \leq \frac{1}{\sqrt{\Delta t}} \left(\int_{t_n}^{t_{n+1}} \left(\eta_h'(Y_{RK}^n(\mathbf{y}, s), s) \right)^2 \, ds \right)^{\frac{1}{2}},
$$

and then

$$
\begin{aligned}
&\int_\Omega \left(\frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \eta_h'(Y_{RK}^n(\mathbf{y}, s), s) \, ds \right)^2 d\mathbf{y} \leq \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \int_\Omega \eta_h'(Y_{RK}^n(\mathbf{y}, s), s)^2 \, d\mathbf{y} \, ds \\
&= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \int_\Omega \eta_h'(\mathbf{z}, s)^2 \det \left(\nabla Y_{RK}^n \right)^{-1} (\mathbf{z}, s) \, d\mathbf{z} \, ds \leq \frac{(1 + c_1 \Delta t)}{\Delta t} \left\| \eta_h' \right\|^2_{L^2((t_n, t_{n+1}), L^2(\Omega))},
\end{aligned}
$$

where the change of variable $\mathbf{z} = Y_{RK}^n(\mathbf{y}, s)$ and estimate (4.4) have been used. Analogously, for the last term of (4.5) we get

$$\int_\Omega \left(\frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \nabla\, \eta_h(Y_{RK}^n(\mathbf{y}, s), s) \cdot \frac{\partial Y_{RK}^n}{\partial s}(\mathbf{y}, s)\, ds \right)^2 d\mathbf{y}$$
$$\leq \frac{(1 + c_1\Delta t)\,(1 + 2c_1)^2}{\Delta t} \|\nabla\, \eta_h\|^2_{L^2((t_n,t_{n+1}),L^2(\Omega))},$$

where we have also considered estimate (4.3). Finally, by applying Young's inequality to term I_1 and using the above results and Hypothesis 6 for $s = 0$ and $r = k$ we obtain

(4.6)
$$I_1 \leq \frac{(1 + c_1\Delta t)\,(1 + 2c_1)^2\, K^2\, h^{2k}}{2\Delta t} \left(\|\phi'\|^2_{L^2((t_n,t_{n+1}),H^k(\Omega))} + \|\phi\|^2_{L^2((t_n,t_{n+1}),H^{k+1}(\Omega))} \right)$$
$$+ \frac{1}{2}\|e_h^{n+1}\|_0^2.$$

Next, we decompose term I_2 into $I_2 = I_2^1 + I_2^2 + I_2^3$, with

$$I_2^1 = \frac{1}{2}\left\langle \mathbf{C}\,\nabla\,\eta_h^{n+1},\ \mathbf{C}\,\nabla\, e_h^{n+1} \right\rangle - \frac{1}{2}\left\langle \mathbf{C}\,\nabla\,\eta_h^{n},\ \mathbf{C}\,\nabla\, e_h^{n} \right\rangle,$$
$$I_2^2 = \frac{1}{2}\left\langle \mathbf{C}\,\nabla\,\eta_h^{n},\ \mathbf{C}\,\nabla\, e_h^{n} \right\rangle - \frac{1}{2}\left\langle (\mathbf{C}\,\nabla\,\eta_h^{n})\circ X_E^n,\ (\mathbf{C}\,\nabla\, e_h^{n})\circ X_E^n \right\rangle,$$
$$I_2^3 = \frac{1}{2}\left\langle (\mathbf{C}\,\nabla\,\eta_h^{n})\circ X_E^n,\ (\mathbf{C}\circ X_E^n)(\nabla\, e_h^{n+1} + \ \nabla\, e_h^n \circ X_E^n) \right\rangle.$$

Notice that I_2^1 explicitly appears in (4.1). For I_2^2 we apply first the change of variable $\mathbf{y} = X_E^n(\mathbf{x})$ in the integral, Lemma 5.1 in [6], and Hypotheses 2 and 6 obtaining

$$I_2^2 = \frac{1}{2}\int_\Omega (\mathbf{C}\,\nabla\,\eta_h^n)(\mathbf{x}) \cdot (\mathbf{C}\,\nabla\, e_h^n)(\mathbf{x})\, d\mathbf{x} - \frac{1}{2}\int_\Omega (\mathbf{C}\,\nabla\,\eta_h^n \cdot \mathbf{C}\,\nabla\, e_h^n)\circ X_E^n(\mathbf{x})\, d\mathbf{x}$$
$$= \frac{1}{2}\int_\Omega \left(1 - \det((\mathbf{F}_E^n)^{-1}(\mathbf{x}))\right)(\mathbf{C}\,\nabla\,\eta_h^n)(\mathbf{x}) \cdot (\mathbf{C}\,\nabla\, e_h^n)(\mathbf{x})\, d\mathbf{x}$$
$$\leq \frac{c_1 c_2 \Delta t}{4}\left(K^2 h^{2k}\, \|\phi^n\|^2_{k+1} + \|\mathbf{B}\,\nabla\, e_h^n\|_0^2 \right).$$

Now, we replace in I_2^3 equality $\mathbf{C}(X_E^n(\mathbf{x})) = \mathbf{C}(\mathbf{x}) - \mathbf{D}^n(\mathbf{x})$, where

$$D_{ij}^n(\mathbf{x}) := \int_{t_n}^{t_{n+1}} \nabla C_{ij}(Y_E^n(\mathbf{x}, s)) \cdot \mathbf{v}^{n+1}(\mathbf{x})\, ds \ \ a.e.\ \mathbf{x} \in \Omega,$$

with $\left|D_{ij}^n(\mathbf{x})\right| \leq c_1\sqrt{c_2}\Delta t$, and the function $Y_E^n(\mathbf{x}, \cdot) : [t_n, t_{n+1}] \longrightarrow \overline{\Omega}$ is defined by $Y_E^n(\mathbf{x}, s) := \mathbf{x} - (t_{n+1} - s)\mathbf{v}^{n+1}(\mathbf{x})$. Thus, we get

$$I_2^3 = \frac{1}{2}\left\langle (\mathbf{C}\,\nabla\,\eta_h^n)\circ X_E^n\ ,\ \mathbf{C}\,\nabla\, e_h^{n+1} - \mathbf{D}\,\nabla\, e_h^{n+1} + (\mathbf{C}\,\nabla\, e_h^n)\circ X_E^n \right\rangle,$$

and then

$$I_2^3 \leq \|(\mathbf{C}\,\nabla\,\eta_h^n)\circ X_E^n\|_0^2 + \frac{1}{8}\left\|\mathbf{C}\,\nabla\, e_h^{n+1} + (\mathbf{C}\,\nabla\, e_h^n)\circ X_E^n\right\|_0^2 + \frac{1}{8}\left\|\mathbf{D}\,\nabla\, e_h^{n+1}\right\|_0^2.$$

Moreover, we have

$$\|(\mathbf{C}\ \nabla\ \eta_h^n)\circ X_E^n\|_0^2 \le (1+c_1\Delta t)c_2K^2h^{2k}\,\|\phi^n\|_{k+1}^2\,, \tag{4.7}$$

$$\left\|\mathbf{D}\ \nabla\ e_h^{n+1}\right\|_0^2 \le c_1^2c_2\Delta t^2\left\|\mathbf{B}\ \nabla\ e_h^{n+1}\right\|_0^2, \tag{4.8}$$

where, Lemma 5.4 in [6] and Hypotheses 2 and 6 with $r=k+1$ and $s=1$ have been required. Now, by jointly considering estimates for I_2^2 and I_2^3, and the lower bound δ of Hypothesis 2 we can state

$$\begin{aligned} I_2 \le I_2^1 &+ \left(\frac{c_1c_2\Delta t}{4}+(1+c_1\Delta t)c_2\right)K^2h^{2k}\,\|\phi^n\|_{k+1}^2 \\ &+ \frac{(2c_1+c_1^2\Delta t)c_2\Delta t}{8\delta}\,\delta\left(\|\mathbf{B}\ \nabla\ e_h^n\|_0^2+\left\|\mathbf{B}\ \nabla\ e_h^{n+1}\right\|_0^2\right) \\ &+ \frac{1}{8}\left\|\mathbf{C}\ \nabla\ e_h^{n+1}+(\mathbf{C}\ \nabla\ e_h^n)\circ X_E^n\right\|_0^2. \end{aligned} \tag{4.9}$$

Similar reasoning, i.e., Lemma 5.4 in [6] and Hypotheses 1, 2, and 6 lead to

$$\begin{aligned} \|(\mathbf{L}^n\mathbf{A}\ \nabla\eta_h^n)\circ X_E^n\|_0^2 &\le (1+c_1\Delta t)c_1^2c_2^2K^2h^{2k}\ \ \|\phi^n\|_{k+1}^2\,, \\ \|(\nabla\ \mathrm{div}\ \mathbf{v}^n\cdot\mathbf{A}\ \nabla\eta_h^n)\circ X_E^n\|_0^2 &\le (1+c_1\Delta t)c_1^2c_2^2K^2h^{2k}\ \ \|\phi^n\|_{k+1}^2\,. \end{aligned} \tag{4.10}$$

Using these inequalities, I_3 and I_4 can be bounded as follows:

$$I_3 \le \frac{(1+c_1\Delta t)c_1^2c_2^2\Delta t\ K^2\ h^{2k}}{4}\,\|\phi^n\|_{k+1}^2+\frac{\Delta t}{4\delta}\delta\left\|\mathbf{B}\ \nabla\ e_h^{n+1}\right\|_0^2, \tag{4.11}$$

$$I_4 \le \frac{(1+c_1\Delta t)c_1^2c_2^2\Delta t^2\ K^2\ h^{2k}}{4}\,\|\phi^n\|_{k+1}^2+\frac{1}{4}\left\|e_h^{n+1}\right\|_0^2, \tag{4.12}$$

where, in order to estimate I_3, we have used Remark 2.2 and Hypothesis 2.

Next, term I_5 can be decomposed like term I_2, namely, $I_5=I_5^1+I_5^2+I_5^3$, where

$$\begin{aligned} I_5^1 &= \frac{1}{2}\left\langle\sqrt{r}\eta_h^{n+1}\ ,\ \sqrt{r}e_h^{n+1}\right\rangle-\frac{1}{2}\left\langle\sqrt{r}\eta_h^n\ ,\ \sqrt{r}e_h^n\right\rangle \\ I_5^2 &= \frac{1}{2}\left\langle\sqrt{r}\eta_h^n\ ,\ \sqrt{r}e_h^n\right\rangle-\frac{1}{2}\left\langle(\sqrt{r}\eta_h^n)\circ X_E^n\ ,\ (\sqrt{r}e_h^n)\circ X_E^n\right\rangle, \\ I_5^3 &= \frac{1}{2}\left\langle(\sqrt{r}\eta_h^n)\circ X_E^n\ ,\ \sqrt{r}\circ X_E^n\left(e_h^{n+1}+e_h^n\circ X_E^n\right)\right\rangle. \end{aligned} \tag{4.13}$$

Moreover, by using again the definition of Y_E^n, we can rewrite I_5^3 as

$$\begin{aligned} \frac{1}{2}\Bigg\langle(\sqrt{r}\eta_h^n)\circ X_E^n, \sqrt{r}e_h^{n+1}-\left(\int_{t_n}^{t_{n+1}}\nabla\ \sqrt{r}(Y_E^n(\mathbf{x},s))\cdot\mathbf{v}^{n+1}(\mathbf{x})\ ds\right)e_h^{n+1} \\ +\ \left(\sqrt{r}e_h^n\right)\circ X_E^n\Bigg\rangle. \end{aligned}$$

Thus, the same kind of computations used for I_2 lead to the following estimate:

$$\begin{aligned} I_5 \le I_5^1 &+ \left(\frac{c_1c_3\Delta t}{4}+(1+c_1\Delta t)c_3\right)K^2h^{2k}\,\|\phi^n\|_k^2 \\ &+ \frac{(2c_1+c_1^2\Delta t)c_3\Delta t}{8}\left(\left\|\sqrt{r}e_h^n\right\|_0^2+\left\|\sqrt{r}e_h^{n+1}\right\|_0^2\right)+\frac{1}{8}\left\|\sqrt{r}e_h^{n+1}+(\sqrt{r}e_h^n)\circ X_E^n\right\|_0^2. \end{aligned} \tag{4.14}$$

Finally, the boundary integral term I_6 is decomposed as a sum of the three terms

$$I_6^1 = \frac{\alpha}{2}\left\langle \eta_h^{n+1},\ e_h^{n+1}\right\rangle_{\Gamma_R} - \frac{\alpha}{2}\left\langle \eta_h^{n},\ e_h^{n}\right\rangle_{\Gamma_R},$$

$$I_6^2 = \frac{\alpha}{2}\left\langle \eta_h^{n},\ e_h^{n}\right\rangle_{\Gamma_R} - \frac{\alpha}{2}\left\langle \eta_h^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1}),\ e_h^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\rangle_{\Gamma_R},$$

$$I_6^3 = \frac{\alpha}{2}\left\langle \eta_h^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1}),\ e_h^{n+1}+e_h^{n}(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\rangle_{\Gamma_R}.$$

Term I_6^1 appears explicitly in (4.1). For I_6^2 and I_6^3 we use Hypothesis 1 and that $c_1\Delta t<1$ to establish the bounds $\left|1-(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})^2(\mathbf{x})\right| \le c_1\Delta t(2+c_1\Delta t)$ and $(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})^2(\mathbf{x}) \le 1+2c_1\Delta t+c_1^2\Delta t^2$, for all $\mathbf{x}\in\Omega$. Thus, we have

$$I_6^2 \le \frac{\alpha c_1\Delta t(2+c_1\Delta t)}{4}\left(\|\eta_h^n\|_{0,\Gamma_R}^2+\|e_h^n\|_{0,\Gamma_R}^2\right),$$

$$I_6^3 \le \frac{\alpha(1+2c_1\Delta t+c_1^2\Delta t^2)}{2}\|\eta_h^n\|_{0,\Gamma_R}^2+\frac{\alpha}{8}\left\|e_h^{n+1}+e_h^n(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2.$$

Next, by using the continuity of the trace mapping, there is $c_\Omega>0$ such that $\|\eta_h^n\|_{0,\Gamma_R}^2 \le c_\Omega\ \|\eta_h^n\|_1^2$. We deduce, by also using Hypothesis 6, that

(4.15)
$$I_6 \le I_6^1+\alpha\left(\frac{c_1\Delta t(2+c_1\Delta t)}{4}+\frac{(1+2c_1\Delta t+c_1^2\Delta t^2)}{2}\right)c_\Omega K^2\ h^{2k}\,\|\phi^n\|_{k+1}^2$$
$$+\frac{\alpha}{8}\left\|e_h^{n+1}+e_h^n(1+\Delta t\ \mathrm{div}\ \mathbf{v}^{n+1})\right\|_{0,\Gamma_R}^2+\frac{\alpha c_1\Delta t(2+c_1\Delta t)}{4}\|e_h^n\|_{0,\Gamma_R}^2.$$

Finally, by jointly considering (4.6), (4.9), (4.11), (4.12), (4.14), and (4.15), and taking into account that $c_1\Delta t<1$, result (4.1) follows. ∎

THEOREM 4.2 (error estimate). *Let us assume Hypotheses* 2, 3, 4, 5, *and* 6. *Let* $\phi\in Z^3\cap C^0(H^{k+1}(\Omega))\cap H^1(H^k(\Omega))$ *be the solution of* (2.5), *with* $\nabla\ \phi\in \mathsf{Z}^3$ *and* $\phi|_{\Gamma_R}\in Z^3(\Gamma_R)$, *and* $\widehat{\phi_h}$ *be the solution of* (3.2) *subject to the initial value* $\phi_h^0=\pi_h\phi^0$. *Let* $\mathbf{B}\ \nabla\ (\widehat{\phi}-\widehat{\phi_h}) := \{\mathbf{B}\ \nabla\ (\phi^n-\phi_h^n)\}_{n=1}^N$. *Then, there exist two positive constants,* c *and* d, *independent of* h *and* Δt, *such that, if* $\Delta t<d$ *we have*

(4.16)
$$\sqrt{\frac{1}{2}}\left\|\widehat{\phi}-\widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))}+\sqrt{\frac{\Delta t\ \delta}{8}}\left\|\mathbf{B}\ \nabla\ \widehat{\phi}-\mathbf{B}\ \nabla\ \widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))}$$
$$+\sqrt{\frac{\Delta t}{8}}\left\|\sqrt{r}\widehat{\phi}-\sqrt{r}\widehat{\phi_h}\right\|_{l^\infty(L^2(\Omega))}+\sqrt{\frac{\Delta t\ \alpha}{16}}\left\|\widehat{\phi}-\widehat{\phi_h}\right\|_{l^\infty(L^2(\Gamma_R))}$$
$$\le ch^k\left(\|\phi\|_{H^1(H^k(\Omega))}+\|\phi\|_{C^0(H^{k+1}(\Omega))}\right)$$
$$+c\Delta t^2\left(\|\phi\|_{Z^3}+\|\mathbf{A}\ \nabla\ \phi\|_{\mathsf{Z}^3}+\ \|r\phi\|_{Z^2}+\|\phi\|_{Z^2,\Gamma_R}+\|f\|_{Z^2}+\|g\|_{Z^2,\Gamma_R}\right)$$
$$+c\Delta t^{\frac{5}{2}}\left(\|\phi\|_{Z^3,\Gamma_R}+\|g\|_{Z^3,\Gamma_R}\right).$$

Proof. First, recall that $\widehat{e_h} = \widehat{\eta_h} - \widehat{\phi} + \widehat{\phi_h}$. By also using the definitions of schemes (3.2) and (2.5) the following identity holds:

(4.17)

$$\begin{aligned}
&\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{e_h}, e_h^{n+1} \right\rangle \\
&\quad = \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \left(\widehat{\eta_h} - \widehat{\phi} + \widehat{\phi_h} \right), e_h^{n+1} \right\rangle \\
&\quad = \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\eta_h}, e_h^{n+1} \right\rangle - \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\phi}, e_h^{n+1} \right\rangle + \left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}}, e_h^{n+1} \right\rangle \\
&\quad = \left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{\eta_h}, e_h^{n+1} \right\rangle + \left\langle \left(\mathcal{L}^{n+\frac{1}{2}} - \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \right) \widehat{\phi}, e_h^{n+1} \right\rangle + \left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}} - \mathcal{F}^{n+\frac{1}{2}}, e_h^{n+1} \right\rangle .
\end{aligned}$$

A lower bound for (4.17) is given by Lemma 5.5 in [6], namely,

(4.18)

$$\begin{aligned}
&\left\langle \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \widehat{e_h}, e_h^{n+1} \right\rangle \\
&\quad \geq D_{\Delta t}^n \left(\frac{1}{2} \left\| \widehat{e_h} \right\|_0^2 + \frac{\Delta t}{4} \left\| \mathbf{C} \nabla \widehat{e_h} \right\|_0^2 + \frac{\Delta t}{4} \left\| \sqrt{r} \widehat{e_h} \right\|_0^2 + \frac{\alpha \Delta t}{4} \left\| \widehat{e_h} \right\|_{0,\Gamma_R}^2 \right) \\
&\qquad + \frac{1}{2\Delta t} \left\| e_h^{n+1} - e_h^n \circ X_{RK}^n \right\|_0^2 + \frac{1}{4} \left\| \mathbf{C} \nabla e_h^{n+1} + (\mathbf{C} \nabla e_h^n) \circ X_E^n \right\|_0^2 \\
&\qquad + \frac{1}{4} \left\| \sqrt{r} e_h^{n+1} + (\sqrt{r} e_h^n) \circ X_E^n \right\|_0^2 + \frac{\alpha}{4} \left\| \phi^{n+1} + \phi^n (1 + \Delta t \ \mathrm{div}\ \mathbf{v}^{n+1}) \right\|_{0,\Gamma_R}^2 \\
&\qquad - \frac{c}{2} \left(\left\| e_h^n \right\|_0^2 + \left\| e_h^{n+1} \right\|_0^2 \right) - c\alpha \Delta t \left\| e_h^n \right\|_{0,\Gamma_R}^2 \\
&\qquad - c\Delta t \left(\left\| \sqrt{r} e_h^n \right\|_0^2 + \left\| \sqrt{r} e_h^{n+1} \right\|_0^2 \right) - c\delta \Delta t \left(\left\| \mathbf{B} \nabla e_h^n \right\|_0^2 + \left\| \mathbf{B} \nabla e_h^{n+1} \right\|_0^2 \right),
\end{aligned}$$

with $c = \max\left\{1, c_1, c_2, (2c_1c_2 + c_1c_2^2)/\delta, c_1c_3/\gamma\right\}$. Now, by using Lemmas 5.19 and 5.20 in [6] we have

(4.19)

$$\begin{aligned}
&\left\langle \left(\mathcal{L}^{n+\frac{1}{2}} - \mathcal{L}_{\Delta t}^{n+\frac{1}{2}} \right) \widehat{\phi}, e_h^{n+1} \right\rangle + \left\langle \mathcal{F}_{\Delta t}^{n+\frac{1}{2}} - \mathcal{F}^{n+\frac{1}{2}}, e_h^{n+1} \right\rangle \\
&\quad = \left\langle \xi_{\mathcal{L}_1}^{n+\frac{1}{2}} - \xi_f^{n+\frac{1}{2}}, e_h^{n+1} \right\rangle + \left\langle \xi_{\mathcal{L}_2}^{n+\frac{1}{2}} - \xi_g^{n+\frac{1}{2}}, e_h^{n+1} \right\rangle_{\Gamma_R} \\
&\quad \leq \left(\left\| \xi_{\mathcal{L}_1}^{n+\frac{1}{2}} \right\|_0^2 + \left\| \xi_f^{n+\frac{1}{2}} \right\|_0^2 \right) + \left\langle H^{n+\frac{1}{2}}, e_h^{n+1} - e_h^n \right\rangle \\
&\qquad + \frac{21}{\alpha} \left(\left\| \xi_{\mathcal{L}_2}^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 + \left\| \xi_g^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 \right) \\
&\qquad + \left\| e_h^{n+1} \right\|_0^2 + \alpha c_1 \ \Delta t \left\| e_h^n \right\|_{0,\Gamma_R}^2 + \frac{\alpha}{8} \left\| e_h^{n+1} + e_h^n (1 + \Delta t \ \mathrm{div}\ \mathbf{v}^{n+1}) \right\|_{0,\Gamma_R}^2 ,
\end{aligned}$$

with $H^{n+\frac{1}{2}}(\mathbf{x}) := (\xi_{\mathcal{L}_2}^{n+\frac{1}{2}}(\mathbf{x}) - \xi_g^{n+\frac{1}{2}}(\mathbf{x}))/\left(2 + \Delta t \ \mathrm{div}\ \mathbf{v}^{n+1}(\mathbf{x})\right)$ a.e. $\mathbf{x} \in \Omega$. Lemma 5.6 in [6] has been applied to the last inequality for the choices $\psi = e_h^{n+1}$ and $\varphi = e_h^n$, first for $F^{n+1} = \xi_{\mathcal{L}_1}^{n+\frac{1}{2}}$, $G^{n+1} = \xi_{\mathcal{L}_2}^{n+\frac{1}{2}}$ and then for $F^{n+1} = -\xi_f^{n+\frac{1}{2}}$, $G^{n+1} = -\xi_g^{n+\frac{1}{2}}$.

By jointly considering the lower bound of (4.17) given in (4.18), the upper bound given in (4.19) and Lemma 4.1 we deduce

(4.20)

$$
\begin{aligned}
&D_{\Delta t}^n\left(\frac{1}{2}\ \|\widehat{e_h}\|_0^2+\frac{\Delta t}{4}\,\|\mathbf{C}\ \nabla\ \widehat{e_h}\|_0^2+\frac{\Delta t}{4}\,\left\|\sqrt{r}\widehat{e_h}\right\|_0^2+\frac{\alpha\Delta t}{4}\,\|\widehat{e_h}\|_{0,\Gamma_R}^2\right)\\
&\quad\le \widetilde{c}K^2h^{2k}\\
&\qquad\times\left(\frac{1}{\Delta t}\,\|\phi'\|^2_{L^2((t_n,t_{n+1}),H^k(\Omega))}+\frac{1}{\Delta t}\,\|\phi\|^2_{L^2((t_n,t_{n+1}),H^{k+1}(\Omega))}+\Delta t\ \|\phi^n\|^2_{k+1}\right)\\
&\qquad+\widetilde{c}\left(\left\|\xi_{\mathcal{L}_1}^{n+\frac12}\right\|_0^2+\left\|\xi_f^{n+\frac12}\right\|_0^2+\left\|\xi_{\mathcal{L}_2}^{n+\frac12}\right\|_{0,\Gamma_R}^2+\left\|\xi_g^{n+\frac12}\right\|_{0,\Gamma_R}^2\right)\\
&\qquad+\left\langle H^{n+\frac12},e_h^{n+1}-e_h^n\right\rangle_{\Gamma_R}\\
&\qquad+\frac{\Delta t}{2}D_{\Delta t}^n\left(\langle\mathbf{C}\ \nabla\ \widehat{\eta_h},\mathbf{C}\ \nabla\ \widehat{e_h}\rangle+\left\langle\sqrt{r}\widehat{\eta_h},\sqrt{r}\widehat{e_h}\right\rangle+\alpha\,\langle\widehat{\eta_h},\widehat{e_h}\rangle_{\Gamma_R}\right)\\
&\qquad+c\left(\|e_h^n\|_0^2+\left\|e_h^{n+1}\right\|_0^2\right)+c\Delta t\left(\left\|\sqrt{r}e_h^n\right\|_0^2+\left\|\sqrt{r}e_h^{n+1}\right\|_0^2\right)\\
&\qquad+c\Delta t\left(\delta\left(\|\mathbf{B}\ \nabla\ e_h^n\|_0^2+\left\|\mathbf{B}\ \nabla\ e_h^{n+1}\right\|_0^2\right)+\alpha\,\|e_h^n\|_{0,\Gamma_R}^2\right),
\end{aligned}
$$

where we omit some positive terms on the left-hand side. In (4.20), $\widetilde{c}$ is a positive constant and $c=\max\left\{2,(3c_1c_2+c_1c_2^2+1)/\delta,c_1c_3(1/\gamma+1/4),c_1,c_2\right\}$. Now, for fixed integer $q\ge 1$, we multiply (4.20) by Δt and sum it from $n=0$ to $n=q-1$. We get

(4.21)

$$
\begin{aligned}
&\frac{1}{2}\,\|e_h^q\|_0^2+\frac{\Delta t}{4}\,\|\mathbf{C}\ \nabla\ e_h^q\|_0^2+\frac{\Delta t}{4}\,\left\|\sqrt{r}e_h^q\right\|_0^2+\frac{\alpha\Delta t}{4}\,\|e_h^q\|_{0,\Gamma_R}^2\\
&\le\frac{1}{2}\,\left\|e_h^0\right\|_0^2+\frac{\Delta t}{4}\,\left\|\mathbf{C}\ \nabla\ e_h^0\right\|_0^2+\frac{\Delta t}{4}\,\left\|\sqrt{r}e_h^0\right\|_0^2+\frac{\alpha\Delta t}{4}\,\left\|e_h^0\right\|_{0,\Gamma_R}^2\\
&\quad+\widetilde{c}K^2h^{2k}\\
&\quad\times\left(\sum_{n=0}^{q-1}\|\phi'\|^2_{L^2((t_n,t_{n+1}),H^k(\Omega))}+\sum_{n=0}^{q-1}\|\phi\|^2_{L^2((t_n,t_{n+1}),H^{k+1}(\Omega))}+\sum_{n=0}^{q-1}\Delta t^2\ \|\phi^n\|^2_{k+1}\right)\\
&\quad+\widetilde{c}\sum_{n=0}^{q-1}\Delta t\left(\left\|\xi_{\mathcal{L}_1}^{n+\frac12}\right\|_0^2+\left\|\xi_f^{n+\frac12}\right\|_0^2+\left\|\xi_{\mathcal{L}_2}^{n+\frac12}\right\|_{0,\Gamma_R}^2+\left\|\xi_g^{n+\frac12}\right\|_{0,\Gamma_R}^2\right)\\
&\quad+\sum_{n=0}^{q-1}\Delta t\left\langle H^{n+\frac12},e_h^{n+1}-e_h^n\right\rangle_{\Gamma_R}+\frac{\Delta t}{2}\left(\langle\mathbf{C}\ \nabla\ \eta_h^q,\mathbf{C}\ \nabla\ e_h^q\rangle-\langle\mathbf{C}\ \nabla\ \eta_h^0,\mathbf{C}\ \nabla\ e_h^0\rangle\right)\\
&\quad+\frac{\Delta t}{2}\left(\left\langle\sqrt{r}\eta_h^q,\sqrt{r}e_h^q\right\rangle-\left\langle\sqrt{r}\eta_h^0,\sqrt{r}e_h^0\right\rangle\right)+\frac{\alpha\Delta t}{2}\left(\langle\eta_h^q,e_h^q\rangle_{\Gamma_R}-\langle\eta_h^0,e_h^0\rangle_{\Gamma_R}\right)\\
&\quad+2c\Delta t\sum_{n=0}^{q}\|e_h^n\|_0^2+2c\Delta t^2\left(\sum_{n=0}^{q}\|\mathbf{B}\ \nabla\ e_h^n\|_0^2+\sum_{n=0}^{q}\left\|\sqrt{r}e_h^n\right\|_0^2+\sum_{n=0}^{q-1}\|e_h^n\|_{0,\Gamma_R}^2\right).
\end{aligned}
$$

Moreover, some terms on the right-hand side of (4.21) can also be bounded. We have

$$
\begin{aligned}
(4.22)\quad & \sum_{n=0}^{q-1} \|\phi'\|^2_{L^2((t_n,t_{n+1}),H^k(\Omega))} + \sum_{n=0}^{q-1} \|\phi\|^2_{L^2((t_n,t_{n+1}),H^{k+1}(\Omega))} + \sum_{n=0}^{q-1} \Delta t^2 \ \|\phi^n\|^2_{k+1} \\
& \leq \|\phi'\|^2_{L^2(H^k(\Omega))} + \|\phi\|^2_{L^2(H^{k+1}(\Omega))} + \Delta t \ T \ \|\phi\|^2_{C^0(H^{k+1}(\Omega))} \\
& \leq \|\phi'\|^2_{L^2(H^k(\Omega))} + (1+\Delta t) T \ \|\phi\|^2_{C^0(H^{k+1}(\Omega))},
\end{aligned}
$$

where T is the measure of the time interval.

Second, by using Lemmas 5.19 and 5.20 in [6] and the fact that $\mathbf{A} \ \nabla \ \phi \cdot \mathbf{n} = g - \alpha\phi$ on the boundary Γ_R we get

$$
\begin{aligned}
& \sum_{n=0}^{q-1} \Delta t \left(\left\| \xi_{\mathcal{L}_1}^{n+\frac{1}{2}} \right\|_0^2 + \left\| \xi_f^{n+\frac{1}{2}} \right\|_0^2 + \left\| \xi_{\mathcal{L}_2}^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 + \left\| \xi_g^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 \right) \\
& \leq \Delta t^4 \ \widetilde{c_1} \ T \ \left(\|\phi\|_{Z^3} + \|\mathbf{A} \ \nabla \ \phi\|_{\mathsf{Z}^3} + \ \|r\phi\|_{Z^2} + \|\alpha\phi\|_{Z^2,\Gamma_R} + \|f\|_{Z^2} + \|g\|_{Z^2,\Gamma_R} \right).
\end{aligned}
$$

Third, by applying Lemma 5.7 in [6], we obtain the estimate

$$
\begin{aligned}
& \left| \sum_{n=0}^{q-1} \Delta t \left\langle H^{n+\frac{1}{2}}, e_h^{n+1} - e_h^n \right\rangle_{\Gamma_R} \right| \leq \frac{\alpha\Delta t}{8} \|e_h^q\|^2_{0,\Gamma_R} + \frac{\alpha\Delta t}{8} \|e_h^0\|^2_{0,\Gamma_R} + 6\alpha\Delta t^2 \sum_{n=0}^{q-1} \|e_h^n\|^2_{0,\Gamma_R} \\
& + \frac{4 + c_1\Delta t}{\alpha} \Delta t \left(\sum_{n=0}^{q} \left\| \xi_{\mathcal{L}_2}^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 + \sum_{n=0}^{q} \left\| \xi_g^{n+\frac{1}{2}} \right\|_{0,\Gamma_R}^2 \right) \\
& + \frac{\Delta t^2}{\alpha} \left(\sum_{n=0}^{q-1} \left\| \frac{\xi_{\mathcal{L}_2}^{n+\frac{1}{2}} - \xi_{\mathcal{L}_2}^{n-\frac{1}{2}}}{\Delta t} \right\|_{0,\Gamma_R}^2 + \sum_{n=0}^{q-1} \left\| \frac{h_g^{n+\frac{1}{2}} - h_g^{n-\frac{1}{2}}}{\Delta t} \right\|_{0,\Gamma_R}^2 \right).
\end{aligned}
$$

Thus, by using again the bounds given in Lemmas 5.19 and 5.20 in [6] we get

$$
\begin{aligned}
& \left| \sum_{n=0}^{q-1} \Delta t \left\langle H^{n+\frac{1}{2}}, e_h^{n+1} - e_h^n \right\rangle_{\Gamma_R} \right| \leq \frac{\alpha\Delta t}{8} \|e_h^q\|^2_{0,\Gamma_R} + \frac{\alpha\Delta t}{8} \|e_h^0\|^2_{0,\Gamma_R} + 6\alpha\Delta t^2 \sum_{n=0}^{q-1} \|e_h^n\|^2_{0,\Gamma_R} \\
& + \frac{5}{\alpha} \Delta t^4 \ T \ \widetilde{c_1} \left(\|\phi\|_{Z^3} + \|\mathbf{A} \ \nabla \ \phi\|_{\mathsf{Z}^3} + \ \|r\phi\|_{Z^2} + \|\alpha \ \phi\|_{Z^2,\Gamma_R} + \|f\|_{Z^2} + \|g\|_{Z^2,\Gamma_R} \right) \\
& + \frac{\Delta t^5 \ T}{\alpha} \left(\|\alpha \ \phi\|_{Z^3,\Gamma_R} + \|g\|_{Z^3,\Gamma_R} \right).
\end{aligned}
$$

Fourth, analogous computations to those developed in Lemma 4.1 give

$$
\begin{aligned}
& \frac{\Delta t}{2} \left\langle \mathbf{C} \ \nabla \ \eta_h^j, \mathbf{C} \ \nabla \ e_h^j \right\rangle \leq \frac{K^2 \ c_2 \ h^{2k} \ \Delta t}{2} \left\| \phi^j \right\|^2_{k+1} + \frac{\Delta t}{8} \left\| \mathbf{C} \ \nabla \ e_h^j \right\|_0^2, \\
& \frac{\Delta t}{2} \left\langle \sqrt{r}\eta_h^j, \sqrt{r} e_h^j \right\rangle \leq \frac{K^2 \ c_3 \ h^{2k} \ \Delta t}{2} \left\| \phi^j \right\|^2_{k+1} + \frac{\Delta t}{8} \left\| \sqrt{r} e_h^j \right\|_0^2, \\
& \frac{\alpha\Delta t}{2} \left\langle \eta_h^j, e_h^j \right\rangle_{\Gamma_R} \leq \alpha K^2 \ c_\Omega \ h^{2k} \ \Delta t \left\| \phi^j \right\|^2_{k+1} + \frac{\alpha \ \Delta t}{16} \left\| e_h^j \right\|^2_{0,\Gamma_R} \quad \text{for } j = 0, q.
\end{aligned}
$$

Next, for $n = 0, \ldots, N$, let us introduce the notation

$$\theta_n := \frac{1}{2} \|e_h^n\|_0^2 + \frac{\delta \Delta t}{8} \|\mathbf{B} \nabla e_h^n\|_0^2 + \frac{\Delta t}{8} \left\|\sqrt{r}\, e_h^n\right\|_0^2, \quad \overline{\theta}_n := \frac{\alpha \Delta t}{16} \|e_h^n\|_{0,\Gamma_R}^2 .$$

With the above notation, the previous estimates lead to

$$(1 - 16c\Delta t)\theta_q + \overline{\theta}_q \leq 16c\Delta t \sum_{n=0}^{q-1} \theta_n + 128c\Delta t \sum_{n=0}^{q-1} \overline{\theta}_n + \widetilde{c}\left(\theta_0 + \overline{\theta_0} + C\right),$$

where C contains the constant terms multiplied by h^{2k}, by Δt^4 and by Δt^5.

Finally, taking into account that $e_h^0 = 0$, the result is concluded by discrete Gronwall's inequality (see, for instance, [17]). ∎

5. Finite element spaces and quadrature formulas. Results concerning stability and consistency of scheme (3.2), when the inner products are exactly integrated, have been proved above for a wide class of finite element spaces (we only require the interpolation property in Hypothesis 6). Nevertheless, numerical integration has to be used in practice to approximate the involved integrals. It is well known that, for the classical first order in time Lagrange–Galerkin method, numerical quadrature can lead to conditional stability [13, 19, 16]. Moreover, new terms may appear in the error estimates (see [15, 13] and the two last paragraphs in section 6).

In the present section we analyze the stability of (3.2) when combined with some finite element spaces and quadrature formulas, extending the studies in the literature regarding the classical scheme. In the next section we present some numerical tests showing the influence of quadrature formulas in both stability and consistency errors. Most of the papers in the literature study the classical Lagrange–Galerkin method for piecewise linear finite elements. Indeed, conditional instability is shown in [13] for Gauss–Legendre, Gauss–Lobatto (with more than three points), Radau and Newton–Cotes formula, when applied to the linear convection equation. This work was extended to the linear convection-diffusion equation in [19] and to a wider class of quadrature formulas in [16]. For both convection and convection-diffusion equations, Gauss–Lobatto quadrature formulas lead to the most stable schemes. However, only the Trapezium rule (or two points Gauss–Lobatto) preserves unconditional stability. We have not found in the bibliography any positive statement concerning the classical Lagrange–Galerkin method when using quadrature formulas for quadratic elements. The above results are established by using Fourier analysis for constant coefficients equations and only in the one dimensional case. In [13] the analysis has been generalized to d dimensions for the linear convection equation under some particular conditions.

In the present paper, for ν and $\mathbf{v}$ constant and $\nu \geq 0$, we consider the linear convection (and convection-diffusion) equation with constant coefficients

$$\frac{\partial \phi}{\partial t} - \nu \Delta \phi + \mathbf{v} \cdot \nabla \phi = 0, \qquad (\mathbf{x}, t) \in \Omega \times [0, T]. \tag{5.1}$$

Only the one dimensional equation with strictly positive velocity is treated. Similar results can be obtained for negative velocity. Moreover, our analysis can be generalized to d dimensions as we will establish in Lemma 5.1. A similar result has been proved in [13] for the classical Lagrange–Galerkin method and constant linear convection equation (i.e., $\nu = 0$ in (5.1)).

Remark 5.1. Although we have stated the previous results for $d = 2, 3$, it is easy to prove similar results for the analogous scheme with $d = 1$. The only difference appears when writing boundary conditions, which are not required in Fourier analysis.

Now, for a family of quadrangular meshes of parameter h, $\mathcal{T}_h$, let us introduce the finite element spaces

$$\mathcal{Q}_h^k := \{f \in C^0(\overline{\Omega}), f|_K \in \mathcal{Q}^k \ \forall K \in \mathcal{T}_h\}, \tag{5.2}$$

with $\mathcal{Q}^k$ being the space of polynomials of degree less than or equal to k in each variable separately. Analogously, for triangular meshes, let us introduce the finite element spaces

$$\mathcal{P}_h^k := \{f \in C^0(\overline{\Omega}), f|_K \in \mathcal{P}^k \ \forall K \in \mathcal{T}_h\}, \tag{5.3}$$

with $\mathcal{P}^k$ the space of polynomials of degree less than or equal to k.

5.1. Study of the one dimensional problem. Let us first introduce the notation $T_x[\psi] := \psi(x - v\Delta t)$, for $x \in \Omega \subset \mathbb{R}$ and $\psi : \Omega \to \mathbb{R}$. Similarly, $T_{(x_1,x_2)}[\psi] := \psi(x_1 - v_1\Delta t, x_2 - v_2\Delta t)$, $T_{x_1}[\psi] := \psi(x_1 - v_1\Delta t, x_2)$ and $T_{x_2}[\psi] := \psi(x_1, x_2 - v_2\Delta t)$, for $\psi : (x_1, x_2) \in \Omega_1 \times \Omega_2 \subset \mathbb{R}^2 \to \mathbb{R}$, and $\mathbf{v} = (v_1, v_2)$. In the previous definitions, Ω, Ω_1, and Ω_2 are intervals of $\mathbb{R}$. Thus, scheme (3.2) applied to the one dimensional version of (5.1) has the form

$$\begin{aligned} &\frac{1}{\Delta t}\int_\Omega \phi_h^{n+1}\psi_j dx + \frac{\nu}{2}\int_\Omega \frac{d\phi_h^{n+1}}{dx}\frac{d\psi_j}{dx}dx \\ &\qquad = \frac{1}{\Delta t}\int_\Omega T_x\left[\phi_h^n\right]\psi_j dx - \frac{\nu}{2}\int_\Omega T_x\left[\frac{d\phi_h^n}{dx}\right]\frac{d\psi_j}{dx}dx, \end{aligned} \tag{5.4}$$

where ψ_j is the jth basis function of the chosen one dimensional finite element space.

Notice that, for $\mathbf{v}$ constant, Euler and Runge–Kutta approximations lead to the same scheme.

LEMMA 5.1. *For a linear convection-diffusion equation in d dimensions with constant coefficients, the second order Lagrange–Galerkin method is just a tensor product of one dimensional second order Lagrange–Galerkin methods assuming that the basis functions themselves are tensor products of the corresponding one dimensional basis functions on a grid which is uniform in each coordinate direction.*

Proof. It is a straightforward adaptation of the one given in [13] for the first order Lagrange–Galerkin method in the case $d = 2$, by adding the new terms (see [14]). The general case can be solved by induction on d. ∎

In order to develop Fourier analysis, we recall the definition of the Courant number, $\mu := v\Delta t/h$, and the Peclet number, $\rho := \nu\Delta t/h^2$. Moreover, in Table 5.1 we write some difference operators with their corresponding Fourier transforms.

Once we have stated Lemma 5.1, we consider the one dimensional Lagrange finite elements of degree k (notice that $\mathcal{Q}_h^k = \mathcal{P}_h^k$ in one dimension). In particular, we study cases $k = 1$ and $k = 2$ combined with the following Gauss–Lobatto quadrature formulas:

$$\int_{x_1}^{x_2} \psi(x)\, dx \approx \frac{x_2 - x_1}{2}\left(\psi(x_1) + \psi(x_2)\right) \qquad \text{(Trapezium)}, \tag{5.5}$$

$$\int_{x_1}^{x_2} \psi(x)\, dx \approx \frac{x_2 - x_1}{6}\left(\psi(x_1) + 4\psi\left(\frac{x_1 + x_2}{2}\right) + \psi(x_2)\right) \qquad \text{(Simpson)}. \tag{5.6}$$

TABLE 5.1
Operators and corresponding Fourier transforms, with $s = \sin(\theta/2)$ *and* $c = \cos(\theta/2)$.

Operator	Fourier transform
$\delta^2[U_j] := U_{j+1} - 2U_j + U_{j-1}$	$-4s^2$
$\Delta_0[U_j] = (U_{j+1} - U_{j-1})/2$	$2isc$
$\Delta_-[U_j] = U_j - U_{j-1}$	$2(isc + s^2)$
$E_r[U_i] = U_{i-r}$	$e^{-ir\theta}$

PROPOSITION 5.2. *If scheme* (3.2) *with* $\mathcal{P}_h^1$ *finite elements on a uniform mesh is applied to the one dimensional equation* (5.1) *combined with the two point Gauss–Lobatto quadrature* (5.5) *in all of the terms, then the method is unconditionally stable.*

Proof. First, let us compute the terms appearing in the jth equation. We use the notation $(\phi_h)_j^n := \phi_h(x_j, t_n)$ for meshpoint (x_j, t_n).

- The mass term is approximated by (5.5) in the form

$$\frac{1}{\Delta t}\int_\Omega \phi_h^{n+1}(x)\ \psi_j(x)\ dx \approx \frac{h}{\Delta t}\ E_0\left[(\phi_h)_j^{n+1}\right].$$

- The stiffness term is exactly integrated by (5.5), giving rise to

$$\frac{\nu}{2}\int_\Omega \frac{d\phi_h^{n+1}}{dx}(x)\ \frac{d\psi_j}{dx}(x)\ dx = -\frac{\nu}{2}\ \frac{1}{h}\delta^2\left[(\phi_h)_j^{n+1}\right].$$

- The integral of the second member term associated to the first order characteristics method, approximated by (5.5), depends on μ in the form

$$\frac{1}{\Delta t}\int_\Omega T_x[\phi_h^n(x)]\ \psi_j(x)\ dx \approx \frac{h}{\Delta t}\ \left(E_0 + (m-\mu)\Delta_- E_{m-1}\right)\left[(\phi_h)_j^n\right],$$

  for a positive integer m such that $m - 1 < |\mu| < m$.
- The integral of the second member term associated to the second order characteristics method is

$$-\frac{\nu}{2}\int_\Omega \frac{dT_x[\phi_h^n]}{dx}(x)\ \frac{d\psi_j}{dx}(x)\ dx \approx \frac{\nu}{2}\ \frac{1}{h}\ \Delta_0\Delta_- E_{m-1}\left[(\phi_h)_j^n\right],$$

  for a positive integer m such that $m - 1 < |\mu| < m$.

In order to apply von Neumann analysis the amplification factor is needed. Considering the approximate integrals computed above and replacing $(\phi_h)_j^n$ with $g^n e^{i\theta j}$, the following expression for the amplification factor is obtained when $|\mu| < 1$:

$$g_{\mu,\rho}(\theta) = \frac{1 - 2\mu(s^2 + isc) + 2\rho(is^3c - s^2c^2)}{1 + 2\rho s^2}, \tag{5.7}$$

with $s = \sin(\theta/2)$ and $c = \cos(\theta/2)$. Now, since

$$|g_{\mu,\rho}(\theta)|^2 = \frac{1 + 4s^2\mu(\mu-1) + 4c^2(\rho^2 s^4 - \rho s^2)}{1 + 4\rho^2 s^4 + 4\rho s^4}, \tag{5.8}$$

easy computations lead to $|g_{\mu,\rho}(\theta)| \le 1$ for all $\theta \in [-\pi/2, \pi/2]$. We can proceed analogously for arbitrary μ. Thus, unconditional stability is stated. ∎

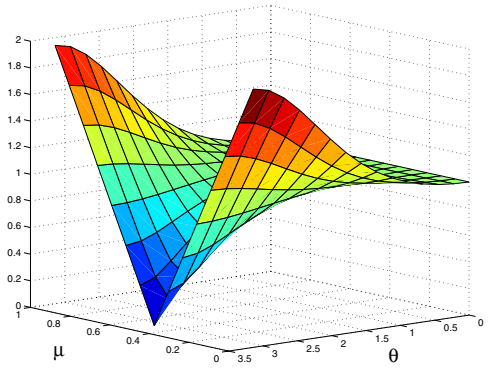


FIG. 5.1. *Modulus of the amplification factor for second order Lagrange–Galerkin method with* $\mathcal{P}_h^1$ *when the exact mass matrix is used.*

Remark 5.2. Note that Proposition 5.2 generalizes results in [13] and [19] for the first order Lagrange–Galerkin method, for which we would have obtained

$$g_{\mu,\rho}(\theta) = \frac{1 - 2\mu(s^2 + isc)}{1 + 4\rho s^2}. \tag{5.9}$$

Remark 5.3. An exact integration of the mass matrix yields

$$I_{m1}^j = \frac{h}{\Delta t}\left(\frac{1}{6}\delta^2\left[(\phi_h)_j^{n+1}\right] - \frac{2}{3}E_0\left[(\phi_h)_j^{n+1}\right]\right),$$

which, combined with the other terms computed in Proposition 5.2, leads to regions of instability. This fact is illustrated in Figure 5.1 by plotting the modulus of the amplification factor as a function of μ and θ for fixed $\rho = 0.19$.

Next, we study quadratic elements for which we have only found the following negative result in the literature (see [13]).

LEMMA 5.3. *The classical Lagrange–Galerkin method applied to the one dimensional version of* (5.1) *with* $\nu = 0$ *and using piecewise quadratic elements has regions of instability if the mass matrix is exactly computed and the right-hand side is evaluated by using a quadrature formula with only interior nodes.*

In the following proposition we prove that the combination of the classical Lagrange–Galerkin method with quadratic elements and Simpson quadrature preserves the unconditional stability of the scheme when applied to the linear convection equation. It is straightforward to prove that a similar result holds for the linear convection-diffusion equation, since the diffusion term is evaluated implicitly.

PROPOSITION 5.4. *If the classical Lagrange–Galerkin scheme with* $\mathcal{P}_h^2$ *finite elements on a uniform mesh is applied to the one dimensional convection equation (*(5.1) *with* $\nu = 0$*) combined with* (5.6) *in all of the terms, then the method is unconditionally stable.*

Proof. First, let us introduce the notation for the basis functions on the reference element: $p_1(x) = 1 - 3x + 2x^2$, $p_2(x) = 4(x - x^2)$, and $p_3(x) = -x + 2x^2$, with the corresponding derivatives denoted by dp_1, dp_2, and dp_3, respectively.

In this case we must take into account that the integrals depend on whether the basis function, ψ_j, corresponds to an interior or to a vertex node. Moreover, for the sake of simplicity, we only consider the case $|\mu| < 1$.

- The mass term provides the approximation $\dfrac{1}{\Delta t}\ \dfrac{2h}{3}E_0\left[(\phi_h)_j^{n+1}\right]$, when j is an interior node, and $\dfrac{1}{\Delta t}\ \dfrac{h}{3}E_0\left[(\phi_h)_j^{n+1}\right]$, otherwise.
- For the second member related to the first order characteristics method we distinguish two cases. First, when j is an interior node and $|\mu| < \frac{1}{2}$ we have

$$\frac{1}{\Delta t}\ \frac{2h}{3}\ \left[(\phi_h)_{j-1}^n\ p_1\left(\frac{1}{2}-\mu\right) + (\phi_h)_j^n\ p_2\left(\frac{1}{2}-\mu\right) + (\phi_h)_{j+1}^n\ p_3\left(\frac{1}{2}-\mu\right)\right];$$

and when $\frac{1}{2} < |\mu| < \frac{3}{2}$, we have

$$\frac{1}{\Delta t}\ \frac{2h}{3}\ \left((\phi_h)_{j-3}^n\, p_1\left(\frac{3}{2}-\mu\right) + (\phi_h)_{j-2}^n\ p_2\left(\frac{3}{2}-\mu\right) + (\phi_h)_{j-1}^n\ p_3\left(\frac{3}{2}-\mu\right)\right).$$

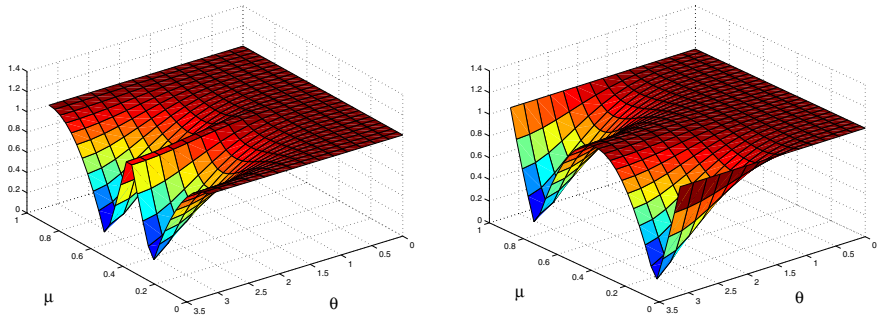
Second, when j is a vertex node and $|\mu| < 1$, we have

$$\frac{1}{\Delta t}\ \frac{h}{3}\ \left((\phi_h)_{j-2}^n\ p_1(1-\mu) + \phi_{j-1}^n\ p_2(1-\mu) + (\phi_h)_j^n\ p_3(1-\mu)\right).$$

The corresponding amplification factors satisfy condition $|g| \leq 1$ (see Figure 5.2), so unconditional stability is reached. □

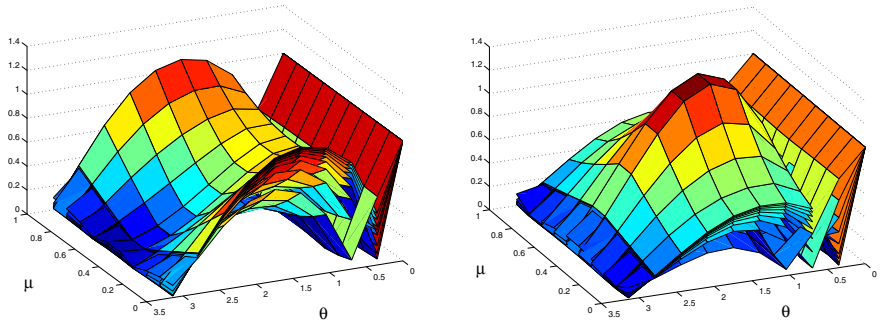
Remark 5.4. Notice that, when applied to the linear convection equation, the classical scheme and the second order Lagrange–Galerkin scheme (3.2) are exactly the same (assuming the same approximation of the characteristic lines).

In the case of scheme (3.2) combined with quadratic elements and Simpson quadrature formula, only conditional stability has been obtained when $\nu > 0$ (equivalently, when $\rho > 0$). Moreover, the region of instability grows up with the Peclet number and it is very small for low Peclet numbers (see, in Figure 5.3, the norm of the amplification factor as a function of θ and μ for different ρ). Thus, it could be used when convection-dominated features are present. In fact, as Priestley [16] pointed out, "results concerning stability are largely academic in that, for the schemes using the higher order quadratures it can be very hard to generate signs of instability. We know of no examples where, in a physical situation, the quadrature instability has caused any problems." We were also unable to make second order Lagrange–Galerkin method with quadratic elements and Simpson quadrature go unstable. We will present some numerical results in the next section.



Interior node. Vertex node.

FIG. 5.2. *Modulus of the amplification factor for the Lagrange–Galerkin method with P_h^2 applied to linear convection equation when Simpson quadrature is used.*



Interior node ($\rho \in (0, 10]$). Vertex node ($\rho \in (0, 5]$).

FIG. 5.3. *Norm of the amplification factor for different Peclet numbers when second order Lagrange–Galerkin method with Q_h^2 and Simpson quadrature is applied to the convection-diffusion equation.*

$\mathcal{Q}_h^1$. $\mathcal{Q}_h^2$.

FIG. 5.4. *Nodes of quadrature for $d = 2$ and $\mathcal{Q}_h^k$ finite elements.*

TABLE 5.2
Quadrature formula (vertex formula) used for space $\mathcal{P}_h^1$.

Node number	Barycentric coordinates	Weights
1	(1,0,0)	1/3
2	(0,1,0)	1/3
3	(0,0,1)	1/3

TABLE 5.3
Quadrature formula used for space $\mathcal{P}_h^2$.

Node number	Barycentric coordinates	Weights
1	(1,0,0)	3/60
2	(0,1,0)	3/60
3	(0,0,1)	3/60
4	(0.5,0.5,0)	8/60
5	(0,0.5,0.5)	8/60
6	(0.5,0,0.5)	8/60
7	(1/3,1/3,1/3)	27/60

5.2. Analysis in d dimensions. Since spaces $\mathcal{Q}_h^k$ satisfy hypotheses of Lemma 5.1, results given in Lemma 5.2 for $k = 1$ and Lemma 5.4 for $k = 2$ can be extended to d dimensions for the corresponding tensor product finite element space and quadrature formulas. In Figure 5.4 nodes of quadrature corresponding to $k = 1$ (left) and $k = 2$ (right) for the two dimensional case are shown.

However, spaces $\mathcal{P}_h^k$ do not satisfy the product property, thus, the theoretical Fourier analysis developed in the present section cannot be applied. Nonetheless, we propose, for the two dimensional case, vertex quadrature for $\mathcal{P}_h^1$ (see Table 5.2) and a seven points quadrature formula for $\mathcal{P}_h^2$ (see Table 5.3). With these formulas we have obtained satisfactory results, as we will illustrate in the next section. Let us notice that, with quadratic elements, also the mid-edges formula has provided good results; however, the mass matrix becomes singular (see [14] for more details).

6. Numerical results. We show numerical results for two numerical examples in two space dimensions. We have tested the above properties of the proposed schemes. We have not found any sign of instability when using scheme (3.2) combined with either $\mathcal{Q}_h^2$ and Simpson rule (only conditionally stable) or $\mathcal{P}_h^2$ and the proposed quadrature formulas (for which, the developed Fourier analysis does not apply).

We notice that, instead of the theoretical $l^\infty(L^2(\Omega))$ norm, we use an approximation denoted by $l^\infty(l^2(\Omega))$, obtained by using quadrature formula in the integrals.

Example 1 (*the rotating Gaussian hill problem*). We choose $\mathbf{A} = \sigma_1 I$, $\mathbf{v} = (-y, x)$, $r = 0$, and $f = 0$ in the domain $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$, and $T = 2$. Dirichlet boundary conditions and initial condition are chosen so that the solution is

$$\phi(x, y, t) = \frac{\sigma_2}{\sigma_2 + 4\sigma_1 t} \exp\left\{ -\frac{(\bar{x}(t) - x_c)^2 + (\bar{y}(t) - y_c)^2}{\sigma_2 + 4\sigma_1 t} \right\}, \tag{6.1}$$

where $\bar{x} = x \cos t + y \sin t$, $\bar{y} = -x \sin t + y \cos t$, $(x_c, y_c) = (0.25, 0)$, $\sigma_1 = 0.001$, and $\sigma_2 = 0.01$. Moreover, we have artificially imposed $\mathbf{v} = 0$ on Γ (as in [18]) and chosen spatio-temporal meshes in such a way that $X_{RK}^n(\overline{\Omega}) \subset \overline{\Omega}$ $\forall n$. In [18], results for $\mathcal{P}_h^1$ finite elements and a quadrature formula with 10 nodes per triangle are shown.

If Figure 6.1 we represent the computed $l^\infty\left((0,T); l^2(\Omega)\right)$ error obtained versus the number of time steps for two uniform spatial meshes with $N_{dof} = N_{x_1} = N_{x_2}$ degrees of freedom in each direction. We denote by $(\mathcal{LG})_1$ the classical Lagrange–Galerkin scheme and by $(\mathcal{LG})_2$ the second order one given by (3.2). In view of the

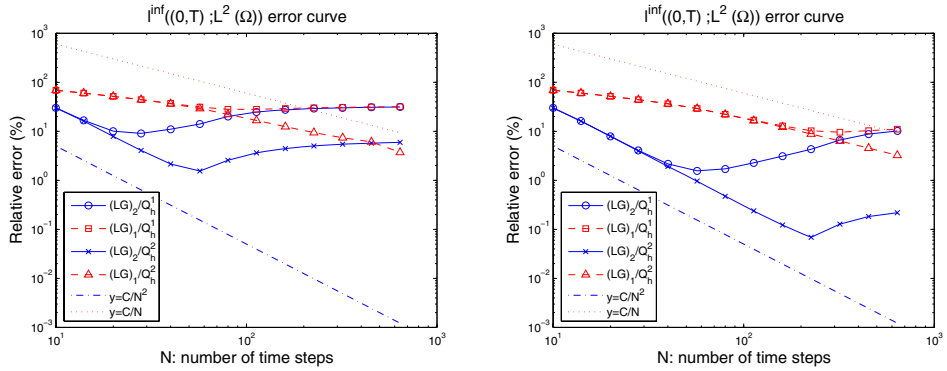
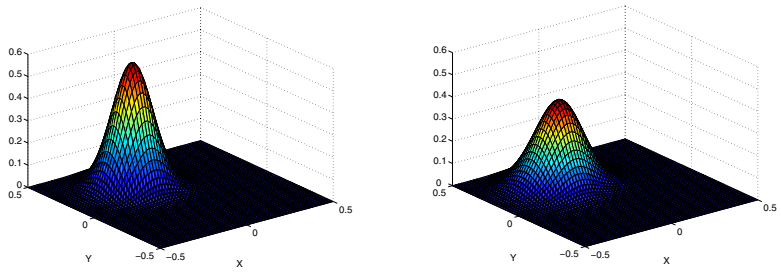


FIG. 6.1. *Computed* $l^\infty((0,T); l^2(\Omega))$ *errors, in log-log scale, for Example* 1 *versus the number of time steps for two fixed spatial meshes: on the left with* $N_{dof} = 67$ *and on the right with* $N_{dof} = 265$.



Exact solution. $\mathcal{Q}_h^1$ computed solution.

FIG. 6.2. *Exact and computed solution of Example* 1 *at time* $T = 2$ *with* second order Lagrange–Galerkin method *with* $\mathcal{Q}_h^1$ *and mesh parameters* $h = 0.015625$ *and* $\Delta t = 0.01$.

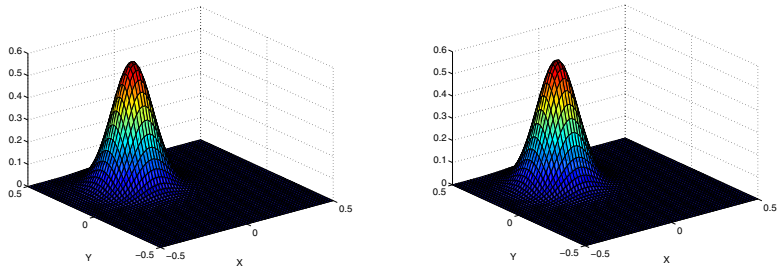
results, we have the following comments:

- The second order Lagrange–Galerkin method proposed ($(\mathcal{LG})_2$) reduces the time error and allows for a lesser number of time steps.
- In both $(\mathcal{LG})_2\mathcal{Q}_h^1$ and $(\mathcal{LG})_2\mathcal{Q}_h^2$, a $O(1/\Delta t)$ term is observed for fixed h.
- Quadratic finite elements lead to a smaller $O(h)$ term than linear ones. Notice that we have used meshes with the same number of degrees of freedom (d.o.f.), or, equivalently, the linear mesh has four times the number of elements of the quadratic mesh. For this reason, and for the same mesh, quadratic elements lead to better algorithms than linear ones.

We can see the exact solution compared to the computed solutions in Figures 6.2 and 6.3, with $\mathcal{Q}_h^1$ and $\mathcal{Q}_h^2$ finite elements, respectively, and mesh parameters $h = 0.015625$ and $\Delta t = 0.01$. Particularly, to be noticed is the reduction of numerical diffusion when using quadratic finite elements.

Example 2 (*a convection-(degenerated) diffusion-reaction problem with variable coefficients*). The spatial domain is $\Omega = (0,1) \times (0,1)$ and $T = 1$. The only nonnull coefficient of the diffusion matrix is $\mathbf{A}_{22}(x_1, x_2) = x_1^2 + 0.5$. Moreover, $\mathbf{v}(x_1, x_2) = (0, x_2)$, $r(x_1, x_2) = x_2$. Neumann boundary conditions are imposed on $\Gamma_{2,-} := \Gamma \cap \{x_2 = 0\}$ (i.e., Robin condition with $\alpha = 0$) and Dirichlet boundary conditions on $\Gamma \setminus \Gamma_{2,-}$. Functions f and g are chosen so that the solution is $\phi(x_1, x_2, t) = e^{x_1+x_2+t}$.

For this problem we want not only to show the performance of scheme (3.2) with different finite element spaces, but also to compare it with the more classical



Exact solution. $\mathcal{Q}_h^2$ computed solution.

FIG. 6.3. *Exact and computed solution of Example* 1 *at time* $T = 2$ *with* second order Lagrange–Galerkin method *with* $\mathcal{Q}_h^2$ *and mesh parameters* $h = 0.015625$ *and* $\Delta t = 0.01$.

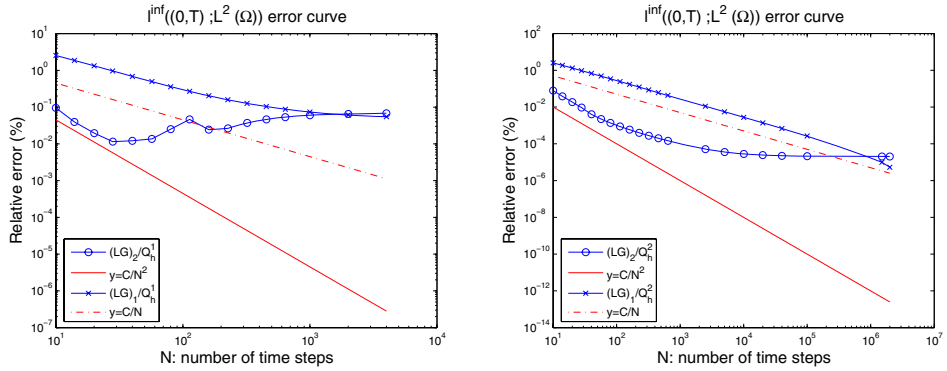


FIG. 6.4. *Computed* $l^\infty((0,T);l^2(\Omega))$ *errors, in log-log scale, for Example* 2 *versus the number of time steps for a fixed spatial mesh. On the left* $\mathcal{Q}_h^1$ *finite elements with* 49 *d.o.f. in each spatial direction. On the right* $\mathcal{Q}_h^2$ *finite elements with* 25 *d.o.f. in each spatial direction.*

characteristics method, when the diffusion, reaction, and source terms are totally implicit in time. In fact, we have observed better errors for $(\mathcal{LG})_2/\mathcal{Q}_h^1$ than for the corresponding first order method. Moreover, for fixed h, a term $1/\Delta t$ is added by the quadrature formula to the error. This behavior is illustrated in Figure 6.4 (left). Notice that an analogous term has already been observed for the $(\mathcal{LG})_1/\mathcal{Q}_h^1$ method in [15, 13].

Similar comments also hold when using $\mathcal{Q}_h^2$ finite elements. However, in this case, it seems that the quadrature formula does not add any error term in the case of the $(\mathcal{LG})_1/\mathcal{Q}_h^2$ method or, at least, a very small time step would be needed to observe it (see Figure 6.4 (right)).

Notice that, at the boundary where the Neumann boundary condition is imposed, the velocity field vanishes, div $\mathbf{v} = 1$, and the term

$$\mathbf{A}(\mathbf{x})\nabla\phi(\mathbf{x},t)\cdot\mathbf{n}|_{\Gamma_{2,-}} = -(x_1^2+0.5)e^{x_1+t}$$

is not null. The necessity of including the $(1+\Delta t \text{ div } \mathbf{v})$ term at the boundary condition is illustrated in Figure 6.5, where for the referred as "Bad $(\mathcal{LG})_2/\mathcal{Q}_h^2$" method we replace this term by 1 (as if div $\mathbf{v} = 0$).

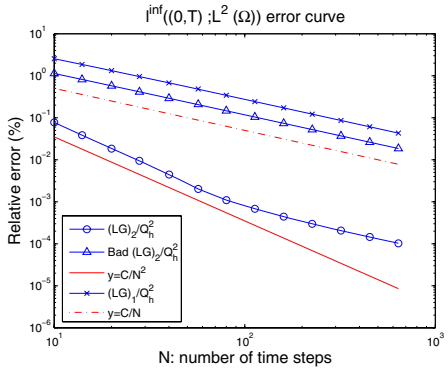


FIG. 6.5. *Computed* $l^\infty((0,T);l^2(\Omega))$ *errors, in log-log scale, for Example* 2 *by using* $\mathcal{Q}_h^2$ *finite elements and different Lagrangian methods.*

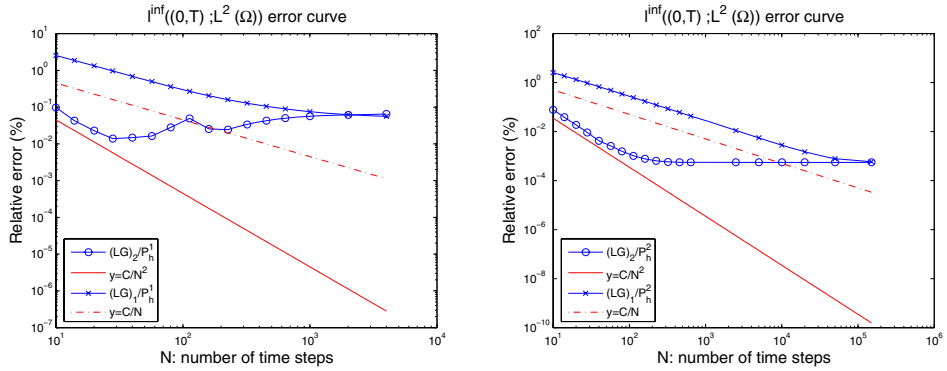


FIG. 6.6. *Computed* $l^\infty((0,T);l^2(\Omega))$ *errors, in log-log scale, for Example* 2 *versus the number of time steps for a fixed spatial mesh. On the left* $\mathcal{P}_h^1$ *FE with* 49 *d.o.f. in each spatial direction. On the right* $\mathcal{P}_h^2$ *with* 25 *d.o.f. in each spatial direction.*

With respect to the computational cost of the algorithms, we remark that first order and second order Lagrange–Galerkin methods take approximately the same time for the same meshes. Moreover, for the same number of degrees of freedom, the $(\mathcal{LG})_2/\mathcal{Q}_h^2$ is quicker than $(\mathcal{LG})_2/\mathcal{Q}_h^1$ due to the different amount of mesh elements (for the same number of nodes, a mesh of linear elements have four times the number of elements of a mesh of quadratic elements).

The conclusion after this second test is that we have obtained better results with $(\mathcal{LG})_2$ than with $(\mathcal{LG})_1$ for similar computing times. Moreover, we have obtained better (accuracy) and quicker (less computing time) results with $(\mathcal{LG})_2/\mathcal{Q}_h^2$ than with $(\mathcal{LG})_2/\mathcal{Q}_h^1$ for the same number of degrees of freedom.

Finally, we have observed an analogous behavior for $\mathcal{P}_h$ finite elements. In Figure 6.6 we show the error versus the number of time steps for fixed spatial meshes.

REFERENCES

[1] J. Baranger, D. Esslaoui, and A. Machmoum, *Error estimate for convection problem with characteristics method*, Numerical Methods for Partial Differential Equations (Marrakech, 1998) Numer. Algorithms, 21 (1999), pp. 49–56.

[2] J. Baranger and A. Machmoum, *Une norme "naturelle" pour la méthode des caractéristiques en éléments finis discontinus: Cas* 1*-D*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 549–574.

[3] J. Baranger and A. Machmoum, *A "natural" norm for the method of characteristics using discontinuous finite elements: 2D and 3D case*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1223–1240.

[4] M. Bercovier, O. Pironneau, and V. Sastri, *Finite elements and characteristics for some parabolic-hyperbolic problems*, Appl. Math. Model., 7 (1983), pp. 89–96.

[5] A. Bermúdez and J. Durany, *La méthode des charactéristiques pour les problèmes de convection-diffusion stationnaires*, RAIRO Math. Model. Numer. Anal., 21 (1987), pp. 7–26.

[6] A. Bermúdez, M. R. Nogueiras, and C. Vázquez, *Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. Part* I: *Time discretization*, SIAM J. Numer. Anal., 44 (2006), pp. 1829–1853.

[7] K. Boukir, Y. Maday, and B. Métivet, *A high order characteristics method for the incompressible Navier-Stokes equations*, ICOSAHOM' 92 (Montpellier, 1992) Comput. Methods Appl. Mech. Engrg., 116 (1994), pp. 211–218.

[8] K. Boukir, Y. Maday, B. Métivet, and E. Razafindrakoto, *A high-order characteristics/finite element method for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 25 (1997), pp. 1421–1454.

[9] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[10] R. E. Ewing and T. F. Russel, *Multistep Galerkin methods along characteristics from convection-diffusion problems*, in Advances in Computer Methods for PDEs IV, R. Vichtneveski and R. S. Stepleman, eds., IMACS Publ., New Brunswick, NJ, 1981, pp. 28–36.

[11] R. E. Ewing and H. Wang, *A summary of numerical methods for time-dependent advection-dominated partial differential equations*, J. Comput. Appl. Math., 128 (2001), pp. 423–445.

[12] M. Gurtin, *An Introduction to Continuum Mechanics*, Math. Sci. and Eng. 158, Academic Press, San Diego, 1981.

[13] K. W. Morton, A. Priestley, and E. Süli, *Stability of the Lagrange-Galerkin method with nonexact integration*, RAIRO Math. Model. Numer. Anal., 22 (1988), pp. 625–653.

[14] M. R. Nogueiras, *Numerical Analysis of Second Order Lagrange-Galerkin Schemes. Application to Option Pricing Problems*, Ph.D. thesis, University of Santiago de Compostela, Santiago de Compostela, Spain, 2005.

[15] O. Pironneau, *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations*, Numer. Math., 38 (1981/82), pp. 309–332.

[16] A. Priestley, *Exact projections and the Lagrange-Galerkin method: A realistic alternative to quadrature*, J. Comput. Phys., 112 (1994), pp. 316–333.

[17] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Series in Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[18] H. Rui and M. Tabata, *A second order characteristic finite element scheme for convection-diffusion problems*, Numer. Math., 92 (2002), pp. 161–177.

[19] E. Süli, *Stability and convergence of the Lagrange-Galerkin method with nonexact integration*, in The Mathematics of Finite Elements and Applications, VI (Uxbridge, 1987), Academic Press, London, 1988, pp. 435–442.

[20] P. Wilmott, J. Dewynne, and S. Howison, *Option Pricing. Mathematical Models and Computation*, Oxford Financial Press, Oxford, 1993.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1877–1902

SPHERICAL INTERFACE DYNAMOS: MATHEMATICAL THEORY, FINITE ELEMENT APPROXIMATION, AND APPLICATION*

KIT HUNG CHAN†, KEKE ZHANG†, AND JUN ZOU‡

Abstract. Stellar magnetic activities such as the 11-year sunspot cycle are the manifestation of magnetohydrodynamic dynamo processes taking place in the deep interiors of stars. This paper is concerned with the mathematical theory and finite element approximation of mean-field spherical dynamos and their astrophysical application. We first investigate the existence, uniqueness, and stability of the dynamo system governed by a set of nonlinear PDEs with discontinuous physical coefficients in spherical geometry, and characterize the system by a saddle-point type variational form. Then we propose a fully discrete finite element approximation to the dynamo system and study its convergence and stability. For the astrophysical application, we perform some fully three-dimensional numerical simulations of a solar interface dynamo using the proposed algorithm, which successfully generates the equatorially propagating dynamo wave with a period of about 11 years similar to that of the Sun.

Key words. spherical interface dynamo, well-posedness, finite element analysis

AMS subject classifications. 65M60, 65M12, 65M15

DOI. 10.1137/050635596

1. Introduction.

Many astrophysical bodies possess intrinsic magnetic fields. The radio signals in connection with Jupiter's magnetic field were first observed more than a half century ago [8] and Jupiter's magnetic field was later measured by the Pioneer spacecraft [1]; the Sun's magnetic field has been observed for a long time [31] and undergone nearly periodic variations with a period of about 11 years. It has been widely accepted that large-scale planetary and stellar magnetic activities represent the manifestation of magnetohydrodynamic dynamo processes taking place in the deep interiors of planets and stars [23, 34, 36, 4]. Though significant progress has been made toward the understanding of quantitative features of stellar magnetic activities, more realistic dynamo simulations in the parameter regime pertaining to stars and planets remain a tough challenge.

Nearly all current stellar and planetary numerical dynamo models employ spectral methods with spherical harmonic functions [35, 19, 22, 7]. The slow Legendre transform and its global nature are computationally inefficient and severely limit the application of spectral methods to general dynamo models, especially to the models with variable physical parameters of space and time. It is becoming increasingly clear that, in order to simulate astrophysical and planetary dynamos using more realistic physical parameters [36], developing other numerically more efficient methods is necessary. The first attempt using finite element methods for numerical dynamo

*Received by the editors July 9, 2005; accepted for publication (in revised form) March 31, 2006; published electronically September 29, 2006.

http://www.siam.org/journals/sinum/44-5/63559.html

†School of Mathematical Sciences, University of Exeter, Exeter, UK (k.h.chan@exeter.ac.uk, k.zhang@exeter.ac.uk). The work of the second author was supported by UK PPARC, Leverhulme and NERC grants.

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (zou@math.cuhk.edu.hk). The work of this author was fully supported by Hong Kong RGC grants (Project 403403 and Project 4048/02P).

simulations was made in [11] and proved to be very promising. The current work presents the first mathematical theory and numerical analysis for mean-field spherical dynamos and their application to astrophysical and planetary problems.

Many stars and planets like the Sun and Jupiter are convectively unstable, which drive small-scale turbulent flows as well as large-scale global circulations in their interiors. The small-scale turbulent convective flows are capable of generating large-scale magnetic fields by the complex dynamo processes [24, 9]. A widely accepted theory for the generation of large-scale magnetic fields through the effect of small-scale turbulence in a conducting fluid is called the mean-field dynamo theory [23], in which a key quantity is the turbulent electromotive force defined as

$$\mathcal{E} =< \hat{\mathbf{u}} \times \hat{\mathbf{B}} >\approx \alpha \mathbf{B}, \tag{1.1}$$

where $< . >$ indicates an average in the dynamo domain, $\mathbf{B}$ is the large-scale mean field, $\hat{\mathbf{u}}$ and $\hat{\mathbf{B}}$ denote the fluctuating small-scale velocity and magnetic fields, and α is typically a tensor describing how the small-scale flows generate the large-scale mean field. Furthermore, the small-scale dynamo simulations suggest that the turbulent electromotive force obeys the following relation [9]:

$$\mathcal{E} = \frac{\alpha_0 \, \mathbf{B}}{1 + (\hat{R}_m)^n |\mathbf{B}|^2 / B_{eq}^2}, \tag{1.2}$$

where α_0 is constant, $0 \le n \le 2$ and B_{eq} is the stellar equipartition field and $\hat{R}_m$ is the magnetic Reynolds number measuring the magnitude of the small-scale flow. The factor $(1 + (\hat{R}_m)^n |\mathbf{B}/B_{eq}|^2)$ represents the nonlinear process of alpha quenching (the catastrophic quenching) which saturates the growing magnetic field. It should be noted that the $\hat{R}_m$-dependent quenching expression should be regarded as a simplified steady state expression for the nonlinear dynamo [4]. On the basis of the quenching relation (1.2), one can investigate the dynamo process of large-scale stellar magnetic fields without being complicated by the dynamic effect such as Lorentz forces. In consequence, (1.2) has been frequently used in the numerical study of astrophysical dynamos [26].

In the present study, we consider a general nonlinear kinematic dynamo for stars and planets consisting of three major zones in spherical geometry; see Figure 1. An inner radiative sphere Ω_1 of radius r_1, with magnetic diffusivity $\lambda_1(\mathbf{x})$, rotates uniformly. Magnetic field $\mathbf{B}_1$ cannot be generated in the radiative region by dynamo action. On the top of the radiative core, there exists a turbulent convection zone Ω_2, $r_1 \le r \le r_2$, in which thermal instabilities drive global circulations $\mathbf{u}$ and small-scale turbulent flows $\hat{\mathbf{u}}$. Note that the effect of the small-scale turbulence in the convection zone is described by α. In the current mean-field dynamo model, we shall use a conventional quenching formula by ignoring the $\hat{R}_m$-dependence in the quenching expression. The magnetic diffusivity in the convection zone is denoted by λ_2 while the nonlinear alpha quenching is assumed to be of the form

$$\alpha = \frac{\alpha_0 f(\mathbf{x}, t)}{1 + \sigma |\mathbf{B}/B_{eq}|^2}, \tag{1.3}$$

where $f(\mathbf{x}, t)$ is a model-oriented function, α_0 and σ are constant parameters, and $\mathbf{B}$ is the generated large-scale magnetic field in the convection zone. The outer region Ω_3, $r_2 \le r \le r_3$, exterior to the convection zone is assumed to be nearly electrically

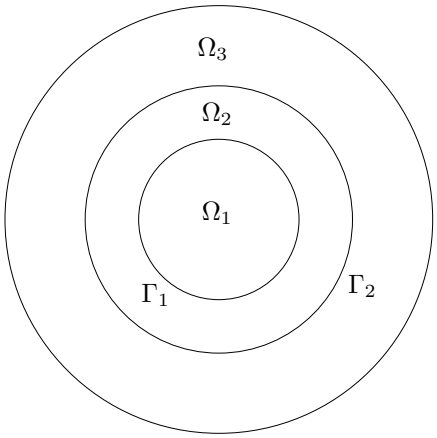


FIG. 1. *Domain Ω, with its inner core Ω_1, convection zone Ω_2, and exterior region Ω_3.*

insulating. We nondimensionalize length by the thickness of the convection zone $d = (r_2 - r_1)$, the magnetic field by the equipartition field B_{eq}, and time by the magnetic diffusion time d^2/λ_2 of the convection zone. This leads to the three sets of dimensionless equations for the three zones in a magnetic star. For the convection fluid shell zone, we have

$$\frac{\partial \mathbf{B}_2}{\partial t} + \nabla \times (\nabla \times \mathbf{B}_2) = R_\alpha \nabla \times \Big(\frac{f(\mathbf{x},t)}{1+\sigma|\mathbf{B}_2|^2} \mathbf{B}_2 \Big) + R_m \nabla \times (\mathbf{u} \times \mathbf{B}_2) \quad \text{in} \quad \Omega_2 \times (0,T) \tag{1.4}$$

$$\nabla \cdot \mathbf{B}_2 = 0 \quad \text{in} \quad \Omega_2 \times (0,T), \tag{1.5}$$

where R_α is a dynamo parameter in connection with the generation process of small-scale turbulence $\hat{\mathbf{u}}$ and R_m is the magnetic Reynolds number associated with the global circulation. For dynamo action to occur, either R_α or R_m must be sufficiently large. The diffusion of the magnetic field $\mathbf{B}_1$ in the inner radiative core with a magnetic diffusivity β_1 can be described by

$$\frac{\partial \mathbf{B}_1}{\partial t} + \nabla \times (\beta_1(\mathbf{x}) \nabla \times \mathbf{B}_1) = 0 \quad \text{in} \quad \Omega_1 \times (0,T), \tag{1.6}$$

$$\nabla \cdot \mathbf{B}_1 = 0 \quad \text{in} \quad \Omega_1 \times (0,T)\,. \tag{1.7}$$

The outer exterior region is usually nearly electrically insulating and governed by

$$\frac{\partial \mathbf{B}_3}{\partial t} + \nabla \times (\beta_3(\mathbf{x}) \nabla \times \mathbf{B}_3) = 0 \quad \text{in} \quad \Omega_3 \times (0,T), \tag{1.8}$$

$$\nabla \cdot \mathbf{B}_3 = 0 \quad \text{in} \quad \Omega_3 \times (0,T)\,, \tag{1.9}$$

where $\beta_3(\mathbf{x})$ is the magnetic diffusivity of the zone.

The above model system will be complemented with the initial conditions

$$\mathbf{B}(\mathbf{x},0) = \mathbf{B}_0(\mathbf{x}) \quad \text{in} \quad \Omega \tag{1.10}$$

and the boundary conditions

$$(\beta_3(\mathbf{x})\nabla \times \mathbf{B}_3) \times \mathbf{n} = 0, \quad \mathbf{B}_3 \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \times (0,T), \tag{1.11}$$

here and in what follows, $\mathbf{n}$ stands for the unit outward normal to the boundary $\partial\Omega$ of the entire physical domain Ω, which consists of the inner core Ω_1, the convection zone Ω_2, and the outer exterior region Ω_3. It should be mentioned that the shear near the solar surface, the effect of which is neglected in our interface solar dynamo model in section 6, may play an important role [3].

We shall use Γ_1 and Γ_2 to denote, respectively, the interface between the inner core and outer convection zone and between the convection zone and the outer exterior; see Figure 1. Since the magnetic diffusivity $\beta(\mathbf{x})$ has jumps across the interfaces Γ_1 and Γ_2 the magnetic field must fulfill some physical interface conditions. We shall take the following standard physical jump conditions adopted in the geodynamo modelling across the interfaces:

$$[(\beta(\mathbf{x})\,\nabla \times \mathbf{B}) \times \mathbf{n}] = 0\,, \quad [\mathbf{B}] = 0 \quad \text{on} \quad (\Gamma_1 \cup \Gamma_2) \times (0,T), \tag{1.12}$$

here and in what follows we use $[\mathbf{A}]$ to denote the quantity of jumps of $\mathbf{A}$ across the interfaces, and $\mathbf{n}$ is the outward normal of $\partial\Omega_2$.

Physically speaking, function $f(\mathbf{x},t)$ and the convective flow $\mathbf{u}$ in (1.4) appear only in the fluid shell region. We shall assume that the velocity $\mathbf{u}$ is nonslip on the boundaries of the fluid shell, i.e., both $f(\mathbf{x},t)$ and $\mathbf{u}$ vanish on Γ_1 and Γ_2. Then by viewing $f(\mathbf{x},t)$ and $\mathbf{u}$ to be extended by zero onto the whole physical domain Ω, we can unify (1.4)–(1.5), (1.6)–(1.7), and (1.8)–(1.9) in three regions Ω_1, Ω_2, and Ω_3 as the following mean-field dynamo system:

$$\begin{aligned} \frac{\partial \mathbf{B}}{\partial t} + \nabla \times (\beta(\mathbf{x})\nabla \times \mathbf{B}) &= R_\alpha \nabla \times \Big(\frac{f(\mathbf{x},t)}{1+\sigma|\mathbf{B}|^2}\mathbf{B}\Big) \\ &\quad + R_m \nabla \times (\mathbf{u} \times \mathbf{B}) \quad \text{in} \quad \Omega \times (0,T) \end{aligned} \tag{1.13}$$

$$\nabla \cdot \mathbf{B} = 0 \quad \text{in} \quad \Omega \times (0,T), \tag{1.14}$$

where $\beta(\mathbf{x})$ represents the magnetic diffusivity $\beta_1(\mathbf{x})$, $\beta_2(\mathbf{x})$, and $\beta_3(\mathbf{x})$ in Ω_1, Ω_2, and Ω_3,respectively, with $\beta_2(\mathbf{x})$ normalized to be 1, so $\beta(\mathbf{x})$ is piecewise smooth and may have large jumps across the interfaces.

The rest of this paper is arranged as follows. Section 2 addresses the well-posedness of the mean-field dynamo system, which is then characterized in terms of a saddle-point type formulation in section 3 for the convenient approximation by finite element methods. The existing convergence theory on saddle-point systems is first generalized in section 4, and a fully discrete finite element method is then proposed and the stability and unique existence are studied. The convergence of the fully discrete scheme is established in section 5, for which the key steps are the introduction of a discrete projection operator and a modification of the Scott–Zhang operator as well as the derivations of their approximation error estimates for piecewise smooth functions. The application of the proposed numerical method to a solar interface dynamo is carried out in section 6. Finally some concluding remarks are given in section 7 to summarize the main contributions of the paper.

2. Well-posedness of the mean-field dynamo system. In this section, we shall investigate the existence, uniqueness, and stability of the solutions to the mean-field dynamo system (1.13)–(1.14) with the initial-boundary conditions (1.10)–(1.11)

and the interface conditions (1.12). Due to space limitations, some proof details may be omitted from time to time throughout the paper but can be found in [10].

2.1. Preliminaries. The most frequently used spaces in the subsequent analysis are the following two Sobolev spaces:

$$H(\mathbf{curl};\Omega) = \left\{\mathbf{A} \in L^2(\Omega)^3; \quad \mathbf{curl}\mathbf{A} \in L^2(\Omega)^3\right\},$$
$$H(\mathrm{div};\Omega) = \left\{\mathbf{A} \in L^2(\Omega)^3; \quad \mathrm{div}\mathbf{A} \in L^2(\Omega)\right\},$$

as well as their subspaces

$$H_0(\mathbf{curl};\Omega) = \left\{\mathbf{A} \in L^2(\Omega)^3; \quad \mathbf{curl}\mathbf{A} \in L^2(\Omega)^3,\ \mathbf{A}\times\mathbf{n} = 0 \text{ on } \partial\Omega\right\},$$
$$H_0(\mathrm{div};\Omega) = \left\{\mathbf{A} \in L^2(\Omega)^3, \quad \mathrm{div}\mathbf{A} \in L^2(\Omega),\ \mathbf{A}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\right\},$$

equipped with the norms

$$\|\mathbf{A}\|_{H(\mathbf{curl};\Omega)} = \left\{\|\mathbf{A}\|^2 + \|\nabla\times\mathbf{A}\|^2\right\}^{\frac{1}{2}}; \quad \|\mathbf{A}\|_{H(\mathrm{div};\Omega)} = \left\{\|\mathbf{A}\|^2 + \|\nabla\cdot\mathbf{A}\|^2\right\}^{\frac{1}{2}}.$$

In the case that the magnetic field is continuous across the interfaces, the intersection of the spaces $H(\mathbf{curl};\Omega)$ and $H(\mathrm{div};\Omega)$ is the natural Sobolev space to be adopted:

$$H(\mathbf{curl},\mathrm{div};\Omega) = \left\{\mathbf{A} \in L^2(\Omega)^3; \quad \mathbf{curl}\mathbf{A} \in L^2(\Omega)^3,\ \mathrm{div}\mathbf{A} \in L^2(\Omega)\right\},$$

$$H_0(\mathbf{curl},\mathrm{div};\Omega) = \left\{\mathbf{A} \in H(\mathbf{curl},\mathrm{div};\Omega); \quad \mathbf{A}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\right\},$$

both equipped with the norm

$$\|\mathbf{A}\|_{H(\mathbf{curl},\mathrm{div};\Omega)} = \left\{\|\mathbf{A}\|^2 + \|\nabla\times\mathbf{A}\|^2 + \|\nabla\cdot\mathbf{A}\|^2\right\}^{\frac{1}{2}}.$$

As the spaces $H(\mathbf{curl},\mathrm{div};\Omega)$ and $H_0(\mathbf{curl},\mathrm{div};\Omega)$ will be frequently used, we shall write

$$H = H(\mathbf{curl},\mathrm{div};\Omega), \quad H_0 = H_0(\mathbf{curl},\mathrm{div};\Omega).$$

To treat the constraint equation $\nabla\cdot\mathbf{B} = 0$, we shall need the following subspace of $H_0(\mathbf{curl},\mathrm{div};\Omega)$:

$$V = \left\{\mathbf{A} \in H_0(\mathbf{curl},\mathrm{div};\Omega); \quad \nabla\cdot\mathbf{A} = 0 \text{ in } \Omega\right\}.$$

Due to the smoothness of the spherical domain Ω, it is known that the space $H_0(\mathbf{curl},\mathrm{div};\Omega)$ is equivalent to the usual Sobolev space $H^1(\Omega)^3$ (see, e.g., [18]). Therefore the Sobolev space V can also be written equivalently as

$$V = \left\{\mathbf{A} \in H^1(\Omega)^3; \quad \nabla\cdot\mathbf{A} = 0 \text{ in } \Omega,\ \mathbf{A}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\right\}, \tag{2.1}$$

and the following equivalence holds:

$$\|\mathbf{A}\|_H^2 = \|\mathbf{A}\|^2 + \|\nabla \times \mathbf{A}\|^2 + \|\nabla \cdot \mathbf{A}\| \eqsim \|\mathbf{A}\|_1^2 \quad \forall\, \mathbf{A} \in H_0\,. \tag{2.2}$$

In the previous statement and what follows, $\|\cdot\|_{s,O}$ are always used to stand for the norm in Sobolev space $H^s(O)$ or $H^s(O)^3$ for any real number $s \geq 0$ and open bounded domain O. We will simply write $\|\cdot\|_s$ when $O = \Omega$ and $\|\cdot\|$ when $s = 0$. The notation $(\cdot,\cdot)$ is used for the scalar product in $L^2(\Omega)$ or $L^2(\Omega)^3$, while $\langle\cdot,\cdot\rangle$ is used to denote the dual pairing between any two Hilbert spaces, and it is the extension of the scalar product $(\cdot,\cdot)$. For a nonnegative function $\beta(x)$, we will often use the notation $\|\cdot\|_\beta = (\beta\cdot,\cdot)^{1/2}$. We may also write $Q_T = \Omega\times(0,T)$ sometimes. In various estimates, we shall frequently use C to stand for a generic constant that is independent of the mesh size h, time stepsize τ, and relevant functions involved.

We end this subsection with a collection of some auxiliary results and formulae for later use.

(1) The space V in (2.1) is a closed subspace of $H^1(\Omega)^3$; see [10].

(2) Young's inequality:

$$a\,b \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \quad \forall\, a, b \in R^1 \ \text{ and } \varepsilon > 0\,.$$

(3) Integration by parts formula which hold for $\mathbf{B} \in H(\mathbf{curl};\Omega)$, $\mathbf{A} \in H^1(\Omega)^3$ and $q \in H^1(\Omega)$:

$$\int_\Omega (\nabla \times \mathbf{B}) \cdot \mathbf{A} d\mathbf{x} = \int_\Omega \mathbf{B} \cdot (\nabla \times \mathbf{A}) d\mathbf{x} - \int_{\partial\Omega} (\mathbf{B} \times \mathbf{n}) \cdot \mathbf{A} ds, \tag{2.3}$$

$$\int_\Omega (\nabla \cdot \mathbf{B}) q d\mathbf{x} = -\int_\Omega \mathbf{B} \cdot \nabla q d\mathbf{x} + \int_{\partial\Omega} (\mathbf{B} \cdot \mathbf{n}) q ds\,. \tag{2.4}$$

(4) Compact embedding lemma [32]. Suppose that X, B, and Y are Banach spaces satisfying $X \subset B \subset Y$ with compact embedding $X \to B$. Then for any $q > 1$, each set bounded both in $L^q(0,T;X)$ and $W^{1,q}(0,T;Y)$ is relatively compact in the space $L^q(0,T;B)$.

(5) Gronwall's inequality. Suppose $h(t)$, $g(t)$ are two nonnegative and square integrable functions on [a, b], $c(t)$ is nondecreasing, and

$$g(t) \leq c(t) + \int_a^t h(s)\, g(s) ds \quad \forall\, t \in [a,b]\,,$$

then the following holds

$$g(t) \leq c(t) \exp\Big(\int_a^t h(s) ds\Big) \quad \forall\, t \in [a,b].$$

2.2. Well-posedness of the mean-field dynamo system. This section is mainly devoted to the well-posedness of the dynamo system (1.4)–(1.12). Due to the jumps in the coefficients, it is not desirable for the system to have classical type solutions. Instead, we shall seek the weak solutions to the mean-field system.

Let us first derive the variational formulation. By multiplying both sides of (1.13) by an $\mathbf{A} \in V$, integrating over Ω and making use of formula (2.3) we obtain

$$\int_\Omega \frac{\partial \mathbf{B}}{\partial t}\cdot \mathbf{A}d\mathbf{x} + \sum_{i=1}^3 \int_{\Omega_i} (\beta\,\nabla\times\mathbf{B})\cdot(\nabla\times\mathbf{A})d\mathbf{x} - \sum_{i=1}^3 \int_{\partial\Omega_i} (\beta\,\nabla\times\mathbf{B})\times\mathbf{n}_i\cdot\mathbf{A}ds$$
$$= R_\alpha \sum_{i=1}^3 \int_{\Omega_i} \Big(\frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}\Big)\cdot(\nabla\times\mathbf{A})d\mathbf{x} - R_\alpha \sum_{i=1}^3 \int_{\partial\Omega_i} \Big(\frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}\times\mathbf{n}_i\Big)\cdot\mathbf{A}ds$$
$$+ R_m \sum_{i=1}^3 \int_{\Omega_i} (\mathbf{u}\times\mathbf{B})\cdot(\nabla\times\mathbf{A})d\mathbf{x} - R_m \sum_{i=1}^3 \int_{\partial\Omega_i} (\mathbf{u}\times\mathbf{B})\times\mathbf{n}_i\cdot\mathbf{A}ds\,.$$

Using the boundary and interface conditions (1.11) and (1.12), we deduce the variational formulation for the dynamo system (1.10)–(1.14).

Find $\mathbf{B}(t)\in V$ such that $\mathbf{B}(0)=\mathbf{B}_0$ and for almost all $t\in(0,T)$,

$$\begin{aligned}&(\mathbf{B}'(t),\mathbf{A}) + (\beta\,\nabla\times\mathbf{B}(t),\nabla\times\mathbf{A})\\ &= R_\alpha\Big(\frac{f(t)}{1+\sigma|\mathbf{B}|^2}\mathbf{B}(t),\nabla\times\mathbf{A}\Big) + R_m(\mathbf{u}(t)\times\mathbf{B}(t),\nabla\times\mathbf{A}) \quad \forall\,\mathbf{A}\in V,\end{aligned} \tag{2.5}$$

here and in what follows, functions of $\mathbf{x}$ and t may be written as functions of t only for simplicity.

The following theorem summarizes the well-posedness of the system (2.5).

THEOREM 2.1. *Assume that $\mathbf{B}_0\in V$, $f\in H^1(0,T;L^\infty(\Omega))$ and $\mathbf{u}\in H^1(0,T;L^\infty(\Omega))$. Then there exists a unique solution $\mathbf{B}$ to the system (2.5) with the regularity*

$$\mathbf{B}\in L^\infty(0,T;V)\cap H^1(0,T;L^2(\Omega)), \tag{2.6}$$

and the solution $\mathbf{B}$ is stable with the following stability estimate:

$$\begin{aligned}&\|\mathbf{B}\|_{L^\infty(0,T;V)} + \|\mathbf{B}\|_{H^1(0,T;L^2(\Omega)^3)}\\ &\le C\,(\|\nabla\times\mathbf{B}(0)\|^2 + \|\mathbf{B}(0)\|^2)\max_{0\le t\le T}(\|f(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}(t)\|^2_{L^\infty(\Omega)})\\ &\quad\cdot\exp\Big(C\int_0^T\Big\{\|f(t)\|^2_{L^\infty(\Omega)} + \|f'(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}'(t)\|^2_{L^\infty(\Omega)}\Big\}dt\Big),\end{aligned} \tag{2.7}$$

where the constant C depends only on the magnetic diffusivity coefficient $\beta(\mathbf{x})$.

Proof. We shall only outline the proof and refer to [10] for the details. As $H^1(\Omega)$ is separable, we know that V is separable. Let $\{\mathbf{w}_k\}_{k=1}^\infty$ be a base of V, and

$$V_m = \{\mathbf{w}_1,\mathbf{w}_2,\dots,\mathbf{w}_m\}\,,\quad m=1,2,3,\dots.$$

Choose $\psi_m\in V_m$ such that $\psi_m\to\mathbf{B}_0$ in V. Then we consider the following approximation of the problem (2.5): Find $\mathbf{B}_m(t)\in V_m$ such that $\mathbf{B}_m(0)=\psi_m$ and for any $\mathbf{A}\in V_m$,

$$\begin{aligned}&(\mathbf{B}_m'(t),\mathbf{A}) + (\beta\nabla\times\mathbf{B}_m(t),\nabla\times\mathbf{A})\\ &= R_\alpha\left(\frac{f}{1+\sigma|\mathbf{B}_m|^2}\mathbf{B}_m(t),\nabla\times\mathbf{A}\right) + R_m(\mathbf{u}\times\mathbf{B}_m(t),\nabla\times\mathbf{A})\,.\end{aligned} \tag{2.8}$$

We claim that the sequence $\{\mathbf{B}_m(t)\}$ is well defined. To see this, we write

$$\mathbf{B}_m(t) = \sum_{j=1}^m \alpha_{j,m}(t)\mathbf{w}_j,\quad \psi_m = \sum_{j=1}^m \gamma_{j,m}\mathbf{w}_j$$

and substitute into (2.8) to get

$$\mathcal{M}\frac{d\alpha_m}{dt} = \mathcal{G}(\alpha_m, t) \quad \text{with} \quad \alpha_m(0) = \gamma_m\,, \tag{2.9}$$

where $\mathcal{M} = (m_{ij})$ with $m_{ij} = (\mathbf{w}_j, \mathbf{w}_i)$, $\mathcal{G}(\alpha_m, t)$ is a vector-valued function of α_m and t, and

$$\gamma_m(t) = (\gamma_{1,m}, \gamma_{2,m}, \ldots, \gamma_{m,m})^t, \quad \alpha_m(t) = (\alpha_{1,m}, \alpha_{2,m}, \ldots, \alpha_{m,m})^t.$$

As $\{\mathbf{w}_m\}$ is linearly independent, the matrix $\mathcal{M}$ is symmetric and positive definite, so it is invertible. Then using the Lipschitz continuity of $\mathcal{G}(\alpha_m, t)$ with respect to α_m, and our subsequent a priori estimates on the solutions to the system (2.8) that ensures the boundedness of $\mathbf{B}_m(t)$ independent of m, one can show (cf. [10]) that the solutions $\{\mathbf{B}_m(t)\}$ of the system (2.8) is well defined in $[0, T]$.

Next, we derive some a priori estimates on the solution to (2.8). By taking $\mathbf{A} = \mathbf{B}_m(t)$ in (2.8), then integrating over $(0, t)$ and using the Cauchy–Schwartz and Gronwall inequality, one can obtain

$$\begin{aligned} &\|\mathbf{B}_m\|^2_{L^\infty(0,T;L^2(\Omega)^3)} + \|\nabla\times\mathbf{B}_m\|^2_{L^2(0,T;L^2(\Omega))} \\ &\le \|\mathbf{B}_m(0)\|_0^2\, \exp\Big(C\int_0^T \Big\{\|f(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}(t)|^2_{L^\infty(\Omega)}\Big\}dt\Big)\,. \end{aligned} \tag{2.10}$$

On the other hand, letting $\mathbf{A} = \mathbf{B}'_m(t)$ in (2.8), then integrating over $(0, t)$, applying the integration by parts and the Gronwall's inequality, we have

$$\begin{aligned} &\|\mathbf{B}'_m\|^2_{L^2(Q_T)} + \|\nabla\times\mathbf{B}_m\|^2_{L^\infty(0,T;L^2(\Omega)^3)} \\ &\le \|\mathbf{B}(0)\|^2\, \exp\Big(C\Big\{\|f\|^2_{L^\infty(Q_T)} + \|\mathbf{u}|^2_{L^\infty(Q_T)}\Big\}\Big) \\ &\quad\cdot \exp\Big(C\int_0^T \Big\{\|f(t)\|^2_{L^\infty(\Omega)} + \|f'(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}(t)\|^2_{L^\infty(\Omega)} + \|\mathbf{u}'(t)\|^2_{L^\infty(\Omega)}\Big\}dt\Big)\,. \end{aligned} \tag{2.11}$$

Using the estimates (2.10)–(2.11), we can extract a subsequence $\{\mathbf{B}_n\}$ from $\{\mathbf{B}_m\}$ such that

$$\mathbf{B}_n \to \mathbf{B} \text{ weakly star in } L^\infty(0,T;V); \quad \mathbf{B}'_n \to \tilde{\mathbf{B}} \text{ weakly in } L^2(0,T;L^2(\Omega)^3)\,. \tag{2.12}$$

By the compact embedding lemma of subsection 2.1, we know that $H^1(0,T;V') \cap L^2(0,T;V)$ is compactly embedded in $L^2(0,T;L^2(\Omega)^3)$, so we have

$$\mathbf{B}_n \to \mathbf{B} \quad \text{in } L^2(0,T;L^2(\Omega)^3). \tag{2.13}$$

We can show that this $\mathbf{B}(t)$ solves the system (2.5). Therefore (2.5) has at least one solution $\mathbf{B}$, which has the regularity (2.6). ∎

3. Characterization of the dynamo system in terms of a saddle-point type problem. The Sobolev space V in the weak formulation (2.5) involves the solenoidal functions, and it is well known that the solenoidal conditions are difficult to enforce in finite element spaces, especially in three dimensions. Hence the variational formulation (2.5) is inconvenient and ineffective for the use in a fully discrete finite element approximation. Instead, we shall transform the variational problem (2.5) into an equivalent saddle-point type system, which can be more easily adopted for its approximations by finite element methods.

3.1. Characterization of solenoidal functions. Let $\mathcal{D}(\Omega)$ be the set of all infinitely differentiable functions with compact supports in Ω, and $\mathcal{V}$ be the subspace of $\mathcal{D}$ with all solenoidal functions:

$$\mathcal{V} = \Big\{ \mathbf{w} \in \mathcal{D}(\Omega);\ \operatorname{div} \mathbf{w} = 0 \text{ in } \Omega \Big\}.$$

We start with the characterization of the gradient of a distribution. For any distribution function p in Ω, written as $p \in \mathcal{D}'(\Omega)$, it is easy to verify that

$$\langle \nabla p, \mathbf{w} \rangle = \sum_{i=1}^{n} \langle \partial_{x_i} p, w_i \rangle = -\sum_{i=1}^{n} \langle p, \partial_{x_i} w_i \rangle = \langle p, \nabla \cdot \mathbf{w} \rangle = 0 \quad \forall\, \mathbf{w} \in \mathcal{V}.$$

That is, ∇p lies in the polar set of $\mathcal{V}$. The following lemma indicates that the converse of this property is also true (cf. [18]).

LEMMA 3.1. *Let Ω be a bounded Lipschitz domain in R^n and $\mathbf{f} = (f_1, f_2, \ldots, f_n)^t$ with $f_i \in \mathcal{D}'(\Omega)$. Then $\mathbf{f} = \nabla p$ for some $p \in \mathcal{D}'(\Omega)$ if and only if*

$$\langle \mathbf{f}, \mathbf{w} \rangle = 0 \quad \forall\, \mathbf{w} \in \mathcal{V}.$$

If $\partial_{x_i} p \in H^{-1}(\Omega)$, then $p \in L^2(\Omega)$ and

$$\|p\|_{L^2(\Omega)/R} \le C(\Omega) \|\nabla p\|_{H^{-1}(\Omega)}.$$

Moreover, if $\partial_{x_i} p \in L^2(\Omega)$, then $p, \nabla p \in L^2(\Omega)$ and

$$\|p\|_{L^2(\Omega)/R} \le C(\Omega) \|\nabla p\|_{L^2(\Omega)}.$$

3.2. Saddle-point formulation of the mean-field dynamo system. In this subsection, we are going to show that the solution $\mathbf{B}(t)$ of the system (2.5) also solves the following saddle-point type problem for some $p \in L^2(0,T; L^2(\Omega))$:

$$\begin{aligned} &\frac{\partial \mathbf{B}}{\partial t} + \nabla \times (\beta(\mathbf{x})\, \nabla \times \mathbf{B}) + \nabla p \\ &= R_\alpha \nabla \times \Big(\frac{f(\mathbf{x},t)}{1+\sigma|\mathbf{B}|^2} \mathbf{B} \Big) + R_m \nabla \times (\mathbf{u} \times \mathbf{B}) \quad \text{in } \Omega \times (0,T); \end{aligned} \tag{3.1}$$

$$\nabla \cdot \mathbf{B} = 0 \quad \text{in} \quad \Omega \times (0,T), \tag{3.2}$$

where a pressure-like term, namely a Lagrange multiplier p, is introduced in (3.1) to ensure that the divergence condition (3.2) is satisfied. This formulation is done purely for the convenience of the subsequent construction of some stable and convergent finite element approximations; the approach is widely employed in numerical solutions of Maxwell equations (see, e.g., [2]).

To do so, we introduce

$$\tilde{\mathbf{B}}(t) = \int_0^t \mathbf{B}(t) dt, \quad \tilde{\mathbf{F}}(t) = \int_0^t \frac{f}{1+\sigma|\mathbf{B}|^2} \mathbf{B}(t) dt, \quad \tilde{\mathbf{U}}(t) = \int_0^t \mathbf{u}(t) \times \mathbf{B}(t) dt\,.$$

Using the regularity of $\mathbf{B}(t)$ from Theorem 2.1 we have

$$\tilde{\mathbf{B}}(t) \in H^1(0,T;V), \quad \tilde{\mathbf{F}}(t) \in H^1(0,T; L^2(\Omega)^3), \quad \tilde{\mathbf{U}}(t) \in H^1(0,T;V)\,.$$

Clearly, both $\tilde{\mathbf{B}}(t)$ and $\tilde{\mathbf{F}}(t)$ are absolutely continuous with respect to t as $\mathbf{B}(t)$ and $f\,\mathbf{B}(t)/(1+\sigma|\mathbf{B}|^2)$ are integrable in $L^1(0,T)$, and we have

$$\tilde{\mathbf{B}}'(t)=\mathbf{B}(t),\quad \tilde{\mathbf{F}}'(t)=\frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}(t),\quad \tilde{\mathbf{U}}'(t)=\mathbf{u}\times\mathbf{B}(t)\,.$$

Now, integrating both sides of (2.5), we obtain for all $t\in[0,T]$ and $\mathbf{A}\in V$ that

$$(\mathbf{B}(t)-\mathbf{B}_0,\mathbf{A})+(\beta\nabla\times\tilde{\mathbf{B}}(t),\nabla\times\mathbf{A})=R_\alpha(\tilde{\mathbf{F}}(t),\nabla\times\mathbf{A})+R_m(\tilde{\mathbf{U}},\nabla\times\mathbf{A})\,. \tag{3.3}$$

We remark that this equation is defined for every $t\in[0,T]$ as $\mathbf{B}(t)$, $\tilde{\mathbf{B}}(t)$, and $\tilde{\mathbf{F}}(t)$ are all continuous with respect to t. This is why we do not treat the system (2.5) directly but instead its integrated form.

For all $t\in[0,T]$, (3.3) can be written as

$$\langle\mathbf{B}(t)-\mathbf{B}_0+\nabla\times(\beta\nabla\times\tilde{\mathbf{B}}(t))-R_\alpha\nabla\times\tilde{\mathbf{F}}(t)-R_m\nabla\times\tilde{\mathbf{U}}(t),\mathbf{A}\rangle=0\quad\forall\,\mathbf{A}\in V,$$

this with Lemma 3.1 indicates that there exists a $P(t)\in L^2(\Omega)$, for every $t\in[0,T]$, such that

$$\mathbf{B}(t)-\mathbf{B}_0+\nabla\times(\beta\nabla\times\tilde{\mathbf{B}}(t))+\nabla P(t)=R_\alpha\nabla\times\tilde{\mathbf{F}}(t)+R_m\nabla\times\tilde{\mathbf{U}}(t), \tag{3.4}$$

or we can write

$$\nabla P(t)=\mathbf{B}_0-\mathbf{B}(t)-\nabla\times(\beta\nabla\times\tilde{\mathbf{B}}(t))+R_\alpha\nabla\times\tilde{\mathbf{F}}(t)+R_m\nabla\times\tilde{\mathbf{U}}(t). \tag{3.5}$$

Noting the right-hand side of (3.5) lies in $(H_0(\mathbf{curl},\mathrm{div};\Omega))'$, we have

$$\nabla P(t)\in H^1(0,T;H_0(\mathbf{curl},\mathrm{div};\Omega)')\subset H^1(0,T;H^{-1}(\Omega)), \tag{3.6}$$

then by Lemma 3.1 we obtain

$$\|P(t)\|_{L^2(\Omega)/R}\le C\|\nabla P(t)\|_{H^{-1}(\Omega)}\quad\forall\,t\in[0,T],$$

this proves $P(t)\in H^1(0,T;L^2(\Omega))$.

Now (3.1) follows immediately by letting $p(t)=\frac{\partial P(t)}{\partial t}$ and differentiating (3.4) with respect to t, and $p(t)\in L^2(0,T;L^2(\Omega))$.

Adding a term $\gamma(\nabla\cdot\mathbf{B},\nabla\cdot\mathbf{A})$ for some constant $\gamma>0$ in (2.5), an important stabilization term in the subsequent numerical approximation, we are then led to the following theorem.

THEOREM 3.1. *The system* (2.5) *is equivalent to the following variational problem:*

Find $\mathbf{B}(t)\in H_0\equiv H_0(\mathbf{curl},\mathrm{div};\Omega)$ *and* $p(t)\in L^2_0(\Omega)$ *such that* $\mathbf{B}(0)=\mathbf{B}_0$ *and*

$$\begin{cases}(\mathbf{B}'(t),\mathbf{A})+(\beta\nabla\times\mathbf{B}(t),\nabla\times\mathbf{A})+\gamma(\nabla\cdot\mathbf{B}(t),\nabla\cdot\mathbf{A})+(p,\nabla\cdot\mathbf{A})\\ =R_\alpha\Big(\frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}(t),\nabla\times\mathbf{A}\Big)+R_m(\mathbf{u}\times\mathbf{B}(t),\nabla\times\mathbf{A})\quad\forall\,\mathbf{A}\in H_0\\ (\nabla\cdot\mathbf{B},q)=0\quad\forall\,q\in L^2_0(\Omega)\end{cases} \tag{3.7}$$

for a.e. $t\in(0,T)$. *Moreover, we have the following stability estimates for the solution* $(\mathbf{B},p)$*:*

$$\begin{aligned}&\|\mathbf{B}\|_{L^\infty(0,T;V)}+\|\mathbf{B}\|_{H^1(0,T;L^2(\Omega)^3)}+\|p\|_{L^2(0,T;L^2_0(\Omega))}\\ &\le C\,(\|\nabla\times\mathbf{B}(0)\|^2+\|\mathbf{B}(0)\|^2)\max_{0\le t\le T}(\|f(t)\|^2_{L^\infty(\Omega)}+\|\mathbf{u}(t)\|^2_{L^\infty(\Omega)})\\ &\quad\cdot\exp\Big(C\int_0^T\Big\{\|f(t)\|^2_{L^\infty(\Omega)}+\|f'(t)\|^2_{L^\infty(\Omega)}+\|\mathbf{u}(t)\|^2_{L^\infty(\Omega)}+\|\mathbf{u}'(t)\|^2_{L^\infty(\Omega)}\Big\}dt\Big).\end{aligned} \tag{3.8}$$

Proof. From the previous derivations, we know the solution $(\mathbf{B}, p)$ to (2.5) also satisfies (3.1)–(3.2). Then using the interface conditions (1.12), we can directly derive (3.7) from (3.1)–(3.2) by integration by parts. On the other hand, one can readily check by integration by parts that a solution $(\mathbf{B}, p)$ of (3.7) is a solution of (3.1)–(3.2) or (2.5). This proves the equivalence of (2.5) and (3.7).

The uniqueness of the solutions to (3.7) can be done similarly to the proof of Theorem 2.1 (cf. [10]).

The estimates of the first two terms on the left-hand side of (3.8) follow from Theorem 2.1. We next derive the estimate of the last term on the left of (3.8) for p. For this, we introduce a $\phi \in H^1(\Omega) \cap L_0^2(\Omega)$ which satisfies

$$\Delta\phi = p \quad \text{in } \Omega; \quad \frac{\partial \phi}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega.$$

Then it is easy to see by Poincare's inequality that

$$\|\nabla\phi\|^2 \le \|\phi\| \, \|p\| \le C\|p\| \, \|\nabla\phi\|,$$

which gives

$$\|\nabla\phi\| \le C \, \|p\|.$$

Letting $b(\mathbf{A}, q) = \int_\Omega q \nabla \cdot \mathbf{A} d\mathbf{x}$ for any $\mathbf{A} \in H_0$ and $q \in L_0^2(\Omega)$, then we take a special $\mathbf{A} = \nabla\phi$. It is easy to verify that $\mathbf{A} \in H_0$, and

$$\|\mathbf{A}\|_{H_0(\mathbf{curl},\mathrm{div};\Omega)} = \left(\|\nabla\phi\|^2 + \|p\|^2\right)^{\frac{1}{2}} \le C\|p\|\,, \tag{3.9}$$

$$\frac{b(\mathbf{A}, p)}{\|\mathbf{A}\|_{H_0}} = \frac{(p, p)}{\|\mathbf{A}\|_{H_0}} \ge C\|p\|\,. \tag{3.10}$$

But we know from (3.7) that

$$\begin{aligned} b(\mathbf{A}, p) = {} & R_\alpha\Big(\frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}(t), \nabla\times\mathbf{A}\Big) + R_m(\mathbf{u}\times\mathbf{B}(t), \nabla\times\mathbf{A}) \\ & - (\mathbf{B}'(t), \mathbf{A}) - (\beta\nabla\times\mathbf{B}(t), \nabla\times\mathbf{A})\,, \end{aligned}$$

from which and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} b(\mathbf{A}, p) &\le R_\alpha\|f(t)\|_{L^\infty(\Omega)}\|\mathbf{B}(t)\| \, \|\nabla\times\mathbf{A}\| + R_m\|\mathbf{u}\times\mathbf{B}(t)\| \, \|\nabla\times\mathbf{A}\| \\ &\quad + \|\mathbf{B}'(t)\| \, \|\mathbf{A}\| + \|\nabla\times\mathbf{B}(t)\|_\beta \, \|\nabla\times\mathbf{A}\|_\beta \\ &\le C\left(\|f(t)\|_{L^\infty(\Omega)}\|\mathbf{B}(t)\| + \|\mathbf{u}(t)\|_{L^\infty(\Omega)}|\mathbf{B}(t)| + \|\mathbf{B}'(t)\| + \|\nabla\times\mathbf{B}(t)\|\right)\|\mathbf{A}\|_{H_0}. \end{aligned}$$

Now the estimate for p follows from this, (3.10), and the estimates of first two terms in (3.8). ∎

4. Finite element approximations. In this section we shall address the finite element approximation of the nonlinear dynamo system (1.4)–(1.12), based on its saddle-point type variational formulation (3.7). As we know, the so-called edge element methods are widely used in numerical solutions of the Maxwell systems [12, 13, 14]. Their main advantages lie in the convenience to incorporate the divergence constraints implicitly and the easy satisfaction of the usual interface conditions which involve tangential components of the fields. But the nonlinear geodynamo system of our current interest requires the continuity of all the components of the magnetic field across the interfaces (see (1.12)), not just the tangential components as in other nondynamo modelling systems. This fact makes the edge element methods inconvenient for the approximation of the geodynamo system (1.4)–(1.12). Instead we shall make use of the standard Lagrange nodal finite element methods.

4.1. Saddle-point system and its approximation. We first recall some existing well-posedness about the general saddle-point system and its approximation. Let X and M be two Hilbert spaces, with scalar products $(\cdot,\cdot)_X$ and $(\cdot,\cdot)_M$, respectively, and $a(v,w)$ and $b(v,q)$ be two continuous bilinear forms on $X\times X$ and $X\times M$, i.e., there exist two positive constants $\|a\|$ and $\|b\|$ such that

$$|a(v,w)|\le \|a\|\,\|v\|_X\|w\|_X \quad \forall\, v,w\in X, \tag{4.1}$$

$$|b(v,q)|\le \|b\|\,\|v\|_X\|q\|_M \quad \forall\, v\in X, q\in M. \tag{4.2}$$

We shall need the kernel space V associated with $b(\cdot,\cdot)$ and the polar set of V:

$$V=\{w\in X;\ b(w,q)=0 \quad \forall\, q\in M\},\quad V^0=\{g\in X';\ \langle g,v\rangle=0 \quad \forall\, v\in X\}\,.$$

Consider the saddle-point system: Find $(u,p)\in X\times M$ such that

$$a(u,v)+b(v,p)=f(v) \quad \forall\, v\in X, \tag{4.3}$$

$$b(u,q)=g(q) \quad \forall\, q\in M, \tag{4.4}$$

where $f\in X'$ and $g\in M'$. The following well-posedness results about this saddle-point system can be found in [6, 18].

LEMMA 4.1. *Assume* (4.1) *and* (4.2)*, and*

$$\sup_{w\in V}\frac{a(v,w)}{\|w\|_X}\ge\alpha\|v\|_X \quad \forall\, v\in V; \tag{4.5}$$

$$\sup_{v\in V}a(v,w)>0 \quad \forall\, w\in V,\ w\neq 0, \tag{4.6}$$

$$\sup_{v\in X}\frac{b(v,q)}{\|v\|_X\|q\|_M}\ge\beta \quad \forall\, q\in M,\ q\neq 0\,. \tag{4.7}$$

Then there exists a unique solution $(u,p)\in X\times M$ *to the saddle-point problem* (4.3)–(4.4).

Now we discuss the approximation of the saddle-point system (4.3)–(4.4). Let $X_h\subset X$ and $M_h\subset M$ be two finite dimensional spaces, and define

$$V_h=\{w_h\in X_h;\ b(w_h,q_h)=0 \quad \forall\, q_h\in M_h\}\,.$$

We then introduce a bilinear form $a_h(\cdot,\cdot)$ defined on $X_h\times M_h$ satisfying

$$a_h(v_h,v_h)\ge\alpha^*\|v_h\|_X^2 \quad \forall v_h\in V_h\,, \tag{4.8}$$

$$|a_h(v_h,w_h)|\le\|a_h\|\,\|v_h\|_X\|w_h\|_X \quad \forall\, v_h,w_h\in X_h \tag{4.9}$$

for two positive constants α^* and $\|a_h\|$. In our later applications, $a_h(\cdot,\cdot)$ comes from some approximation of $a(\cdot,\cdot)$ and is formed from $a(\cdot,\cdot)$ in such a way that numerical integrations on polyhedra with curved faces are replaced by much easier integrations on polyhedra with planar faces.

Then we introduce the approximation of the saddle-point system (4.3)–(4.4).

Find $(u_h,p_h)\in X_h\times M_h$ such that

$$a_h(u_h,v_h)+b(v_h,p_h)=f(v_h) \quad \forall\, v_h\in X_h \tag{4.10}$$

$$b(u_h,q_h)=g(q_h) \quad \forall\, q_h\in M_h\,. \tag{4.11}$$

We have the following convergence theorem (cf. [6]), whose detailed proof can be found in [10].

THEOREM 4.1. *In addition to the assumptions* (4.8)–(4.9)*, we assume that the* inf-sup *condition*

$$\sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X \|q_h\|_M} \geq \beta^* \quad \forall\, q_h \in M_h,\ q_h \neq 0\,. \tag{4.12}$$

is also satisfied. Then the system (4.10)–(4.11) *has a unique solution* $(u_h, p_h) \in X_h \times M_h$ *and the following error estimate holds:*

$$\begin{aligned} \|u - u_h\|_X \leq \Big(1 + \frac{\|a_h\|}{\alpha^*}\Big)\Big(1 + \frac{\|b\|}{\beta^*}\Big) \inf_{v_h \in X_h} \|u - v_h\|_X + \|b\| \inf_{\mu_h \in M_h} \|p - \mu_h\|_M \\ + \frac{1}{\alpha^*} \sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(u, v_h)}{\|v_h\|_X}. \end{aligned} \tag{4.13}$$

4.2. A fully discrete finite element method and its stability. In this section, we will propose a fully discrete finite element method for the variational system (3.7). For this purpose, we have to approximate the problem in both time and space. We shall use the backward Euler scheme for time discretization and the popular Hood–Taylor finite elements (cf. [20]) for space discretization.

We start with the partition of the time interval [0, T] and the triangulation of the physical spherical domain Ω. We divide the time interval [0, T] into M equally spaced subintervals using the following nodal points:

$$0 = t_0 < t_1 < t_2 < \cdots < t_M = \mathrm{T},$$

where $t_n = n\,\tau$ for $n = 0, 1, \ldots, M$ and $\tau = T/M$. For any given discrete time sequence $\{u^n\}_{n=0}^M$ with each u^n lying in $L^2(\Omega)$ or $L^2(\Omega)^3$, we define the first order backward finite differences and the averages as follows:

$$\partial_\tau u^n = \frac{u^n - u^{n-1}}{\tau}, \quad \bar{u}^n = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} u(\cdot, s)ds.$$

If $u(\mathbf{x}, t)$ is a function which is continuous with respect to t, we shall often write $u^n(\cdot) = u(\cdot, t_n)$ for $n = 0, 1, \ldots, M$. For the ease of exposition, we may also use the function values for $t \leq 0$, by assuming the convention that $u(\mathbf{x}, t) = u(\mathbf{x}, 0)$ for all $t \leq 0$.

We now introduce the triangulation of the domain Ω, consisting of the inner core Ω_1, the outer core Ω_2, and the exterior zone Ω_3. For the sake of technical treatments, we shall assume that *the outer boundary of the exterior zone* Ω_3 *is a closed convex polygon*; the actual curved boundary case can be treated in the same manner as we handle in this and next section the curved interfaces Γ_1 and Γ_2; see Figure 1.

We first triangulate the inner core Ω_1 using a quasi-uniform triangulation $\mathcal{T}_h^1$ with tetrahedral elements of mesh size h, which form a polyhedral domain $\Omega_h^1 \subset \Omega_1$. The triangulation is done such that the boundary vertices of Ω_h^1 all lie on the boundary of Ω_1.

Then we triangulate the exterior zone Ω_3 using a triangulation $\mathcal{T}_h^3$ with tetrahedral elements, which form a polyhedral domain Ω_h^3. The triangulation is done such that all the vertices on the outer polygonal boundary $\partial\Omega$ are also vertices of Ω_h^3, and the inner boundary vertices of Ω_h^3 all lie on the inner boundary of Ω_3.

Finally we triangulate the outer core Ω_2 using a triangulation $\mathcal{T}_h^2$ with tetrahedral elements, which form a polyhedral domain Ω_h^2. The triangulation is done such that all the vertices on the outer boundary of Ω_h^2 match those vertices on the inner boundary of Ω_h^3, while all the vertices on the inner boundary of Ω_h^2 match the boundary vertices of Ω_h^1.

Now the three individual triangulations $\mathcal{T}_h^1$, $\mathcal{T}_h^2$, and $\mathcal{T}_h^3$ form a global triangulation $\mathcal{T}_h$ of Ω. By $\mathcal{N}_h$ we shall denote the set of all the nodal points of the triangulation $\mathcal{T}_h$, and by $\mathcal{F}_h$ the set of all faces of elements in $\mathcal{T}_h$.

For convenience, any element K of $\mathcal{T}_h$ whose interior has nonempty intersection with the interface Γ_1 and Γ_2 will be called *an interface element*. The set of all interface elements is denoted by $\mathcal{T}_h^*$. Let us introduce some notation needed in the subsequent error estimates. For each interface element $K \in \mathcal{T}_h^*$, we know that K must lie either in Ω_2^h or Ω_3^h according to the construction of the triangulation $\mathcal{T}_h$. And each interface element K is divided by the interface into two parts, written as $\mathcal{K}_1$ and $\mathcal{K}_2$. Since the interfaces Γ_1 and Γ_2 are smooth spheric surfaces, one can show (cf. [17]) that one of the two parts $\mathcal{K}_1$ and $\mathcal{K}_2$, denoted always by $\mathcal{K}$, has a volume of order h_K^4, that is,

$$|\mathcal{K}| \eqsim h_K^4 \,. \tag{4.14}$$

Here and in what follows, we shall often use the symbols $\lesssim$ and $\eqsim$, and $x \lesssim y$ means that $x \le Cy$ for some generic constant C, and $x \eqsim y$ means $x \lesssim y$ and $y \lesssim x$.

Also we may absorb in the generic constant C the upper and lower bounds β_m, β_M, f_M, and u_M of the functions $\beta(\mathbf{x})$, $f(\mathbf{x},t)$, and $\mathbf{u}(\mathbf{x},t)$ over $\Omega \times (0,T)$:

$$\beta_m \le \beta(\mathbf{x}) \le \beta_M \,; \quad |f(\mathbf{x},t)| \,,\ |f_t(\mathbf{x},t)| \le f_M \,; \quad |\mathbf{u}(\mathbf{x},t)| \,,\ |\mathbf{u}_t(\mathbf{x},t)| \le u_M \,.$$

Noting the coefficient $\beta(\mathbf{x})$ in (1.13) has large jumps across the interfaces Γ_1 and Γ_2, hence it may be strongly discontinuous inside each interface element $K \in \mathcal{T}_h^*$, namely when crossing the (curved) common face of two curved polyhedra parts $\mathcal{K}_1$ and $\mathcal{K}_2$ of K. To avoid numerical integrations on polyhedra with curved faces in forming the finite element stiffness matrix, we introduce the following approximations of the coefficients $\beta(\mathbf{x})$, $f(\mathbf{x},t)$, and $\mathbf{u}(\mathbf{x},t)$:

$$\beta_h(\mathbf{x}) = \beta(\mathbf{x}),\ \mathbf{x} \in K \in \mathcal{T}_h \setminus \mathcal{T}_h^* \,; \quad \beta_h(\mathbf{x}) = \beta_i(\mathbf{x}),\ \mathbf{x} \in K \in \mathcal{T}_h^* \cap \Omega_i^h \ (i = 2 \text{ or } 3) \,,$$

$$f_h(\mathbf{x},t) = \begin{cases} 0, & \mathbf{x} \in K \in \mathcal{T}_h^* \cap \Omega_3^h \,; \\ f(\mathbf{x},t), & \text{otherwise} \end{cases} \qquad \mathbf{u}_h(\mathbf{x},t) = \begin{cases} 0, & \mathbf{x} \in K \in \mathcal{T}_h^* \cap \Omega_3^h \\ \mathbf{u}(\mathbf{x},t), & \text{otherwise} \end{cases} .$$

We shall use the Hood–Taylor finite elements (cf. [18, 15, 33]) to approximate the system (3.7), namely the piecewise quadratic polynomials for the magnetic field $\mathbf{B}$ and the piecewise linear polynomials for the Lagrange multiplier p. These spaces can be defined as follows:

$$\begin{aligned} H_h &= \Big\{ \mathbf{w} \in C(\bar{\Omega}) : \ \mathbf{w}|_K \in P_2(K)^3 \ \forall \ K \in \mathcal{T}_h \Big\}, \\ H_{0h} &= \Big\{ \mathbf{w} \in H_h; \ \mathbf{w} \cdot \mathbf{n}_F = 0 \ \forall \, F \in \mathcal{F}_h \cap \partial\Omega \Big\}, \\ Q_h &= \Big\{ q_h \in C(\bar{\Omega}); \ q_h|_K \in P_1(K) \ \forall \ K \in \mathcal{T}_h \Big\} , \end{aligned}$$

where $\mathbf{n}_F$ is the unit normal vector of a face $F \in \mathcal{F}_h$. And the following subspaces of H_{0h} and Q_h will be also needed:

$$\tilde{H}_{0h} = \Big\{ \mathbf{w}_h \in H_{0h}; \ \mathbf{w}_h = 0 \text{ on } \partial\Omega \Big\}, \quad Q_{0h} = \Big\{ q_h \in Q_h; \ \int_\Omega q_h dx = 0 \Big\}.$$

Now, we are ready to propose the fully discrete finite element approximation of the variational problem (3.7) using the approximate functions β_h, f_h, and $\mathbf{u}_h$.

Find $\mathbf{B}_h^n \in H_{0h}$, $p_h^n \in Q_{0h}$ for $n = 1, 2, \ldots, M$ such that $\mathbf{B}_h^0 = S_h\mathbf{B}_0$ and

(4.15)
$$\begin{cases} (\partial_\tau \mathbf{B}_h^n, \mathbf{A}_h) + (\beta_h \nabla\times \mathbf{B}_h^n, \nabla\times \mathbf{A}_h) + \gamma(\nabla\cdot \mathbf{B}_h^n, \nabla\cdot \mathbf{A}_h) + (p_h^n, \nabla\cdot \mathbf{A}_h) \\ \quad = R_\alpha\Big(\frac{f_h^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}_h^n, \nabla\times \mathbf{A}_h\Big) + R_m(\mathbf{u}_h^n\times \mathbf{B}_h^n, \nabla\times \mathbf{A}_h\Big) \quad \forall\, \mathbf{A}_h \in H_{0h}; \\ (\nabla\cdot \mathbf{B}_h^n, q_h) = 0 \quad \forall\, q_h \in Q_{0h}, \end{cases}$$

where S_h is the modified Scott-Zhang interpolation to be defined in section 5. One may replace S_h here by the computationally less expensive standard interpolation operator Π_h induced by the finite element space H_h, but as it will be seen in the subsequent analysis, with Π_h one requires a stronger regularity on the initial data $\mathbf{B}_0$.

We remark that the discrete system (4.15) cannot ensure $\nabla\cdot\mathbf{B}_h^n = 0$, different from the continuous case. The next lemma verifies the well-posedness of the fully discrete scheme (4.15).

LEMMA 4.2. *There exists a unique solution* $(\mathbf{B}_h^n, p_h^n)$ *to the discrete system* (4.15) *for each fixed* n $(1 \le n \le M)$ *and the sequence* $\{\mathbf{B}_h^n\}_{n=0}^M$ *has the following stability estimates:*

$$\max_{1\le n\le M} \|\mathbf{B}_h^n\|^2 + \tau\sum_{n=1}^{M}(\|\nabla\times\mathbf{B}_h^n\|^2 + \|\nabla\cdot\mathbf{B}_h^n\|^2) \lesssim \|\mathbf{B}_h^0\|^2. \tag{4.16}$$

Proof. Inequality (4.16) follows by taking $\mathbf{A}_h = \tau\mathbf{B}_h^n$ in (4.15) and the discrete Gronwall's inequality.

We now verify the existence of solutions to (4.15) for each fixed $n = T/M$ and h by applying the Brouwer fixed point theorem. To this aim, we define a mapping $F_h : (\bar{\mathbf{B}}_h, \bar{p}_h) \to (\mathbf{B}_h, p_h)$ by

$$\begin{cases} \tilde{a}_h(\mathbf{B}_h, \mathbf{A}_h) + \tilde{b}(\mathbf{A}_h, p_h) = \tilde{g}(\bar{\mathbf{B}}_h, \mathbf{A}_h) \quad \forall\, \mathbf{A}_h \in H_{0h}, \\ \tilde{b}(\mathbf{B}_h, q_h) = 0 \quad \forall\, q_h \in Q_{0h}, \end{cases} \tag{4.17}$$

where $\tilde{a}_h$, $\tilde{b}$ and $\tilde{g}$ are given by

$$\tilde{a}_h(\mathbf{B}, \mathbf{A}) = (\mathbf{B}, \mathbf{A}) + \tau(\beta_h\nabla\times\mathbf{B}, \nabla\times\mathbf{A}) + \gamma\tau(\nabla\cdot\mathbf{B}, \nabla\cdot\mathbf{A}),\ \tilde{b}(\mathbf{A}, q) = \tau(q, \nabla\cdot\mathbf{A}),$$
$$\tilde{g}(\mathbf{B}, \mathbf{A}) = (\mathbf{B}_h^{n-1}, \mathbf{A}) + \tau\, R_\alpha\Big(\frac{f_h^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}, \nabla\times\mathbf{A}\Big) + \tau R_m\,(\mathbf{u}_h^n\times\mathbf{B}, \nabla\times\mathbf{A}).$$

By applying Theorem 4.1 one can show that the mapping F_h is well defined; see [10] for details.

We next show that F_h maps a bounded subset of $H_{0h}\times Q_{0h}$ into itself. In fact, taking $\mathbf{A}_h = \mathbf{B}_h$ in the first equation of (4.17), using the second equation and Young's inequality we can obtain

$$\|\mathbf{B}_h\|^2 + \tau\|\nabla\times\mathbf{B}_h\|_{\beta_h}^2 + \gamma\tau\|\nabla\cdot\mathbf{B}_h\|^2 \le \|\mathbf{B}_h^{n-1}\|^2 + \frac{2\tau}{\beta_m}(R_\alpha^2 f_M^2 + 4R_m^2 u_M^2)\|\bar{\mathbf{B}}_h\|^2\,.$$

Thus for any $\bar{\mathbf{B}}_h$ lying in the ball $B(0, r_0) = \{\mathbf{A}_h;\ \|\mathbf{A}_h\|_{H_0} \le r_0\}$ with $r_0 = \sqrt{2}\|\mathbf{B}_h^{n-1}\|$, we have

$$\|\mathbf{B}_h\|^2 + \tau\|\nabla\times\mathbf{B}_h\|_{\beta_h}^2 + \gamma\tau\|\nabla\cdot\mathbf{B}_h\|^2 \le r_0^2$$

when τ is appropriately small such that $4\tau(R_\alpha^2 f_M^2 + 4R_m^2 u_M^2) \le \beta_m$. Next we show that p_h lies in the ball $B(0,\bar{r}_0)$ with

$$\bar{r}_0 = C_0^{-1}\Big(\frac{2}{\tau} + \frac{\sqrt{\gamma}+\sqrt{\beta_M}}{\sqrt{\tau}} + R_\alpha f_M + 2R_m u_M\Big)\, r_0\,.$$

To see this, for any $\mathbf{A}_h \in H_{0h}$, we obtain from (4.17) using the Cauchy–Schwarz inequality that

$$\begin{aligned}\tau(\nabla\cdot\mathbf{A}_h, p_h) \le \Big(&\|\mathbf{B}_h\| + \tau\beta_M^{1/2}\|\nabla\times\mathbf{B}_h\|_\beta + \gamma\tau\|\nabla\cdot\mathbf{B}_h\| \\ &+ \|\mathbf{B}_h^{n-1}\| + \tau R_\alpha f_M\|\bar{\mathbf{B}}_h\| + 2\tau R_m u_M\|\bar{\mathbf{B}}_h\|\Big)\|\mathbf{A}_h\|_{H_0}.\end{aligned}$$

This, combined with the *inf-sup* condition for $\tilde{b}(\cdot,\cdot)$ (cf. [10]), leads to the conclusion that p_h lies in the ball $B(0,\bar{r}_0)$. Thus we have proved that F_h maps the bounded subset $B(0,r_0)\times B(0,\bar{r}_0)$ of $H_{0h}\times Q_{0h}$ into itself. Therefore by the Brouwer fixed point theorem, F_h has a fixed point $(\mathbf{B}_h,p_h)\in B(0,\bar{r}_0)\times B(0,\bar{r}_0)$. This proves the existence of solutions to the system (4.15).

The uniqueness of the solutions can be shown in the same manner as in Theorem 3.1. ∎

5. Convergence analysis of the fully discrete finite element method. This section will be devoted to the convergence analysis on the fully discrete finite element approximation (4.15) to the variational problem (3.7). As we shall see, one of the crucial tools in the analysis relies on the following projection operator P_h which maps functions from the space $H_0\times Q_0 \equiv H_0(\mathbf{curl},\mathrm{div};\Omega)\times L_0^2(\Omega)$ into $H_{0h}\times Q_{0h}$: for any $(\mathbf{B},p)\in H_0\times Q_0$, $(\mathbf{B}_h,p_h) = P_h(\mathbf{B},p)\in H_{0h}\times Q_{0h}$ solves the following saddle-point system:

$$\text{(5.1)}\qquad \begin{cases} (\mathbf{B}_h,\mathbf{A}_h) + a_h(\mathbf{B}_h,\mathbf{A}_h) + (p_h,\nabla\cdot\mathbf{A}_h) \\ \qquad = (\mathbf{B},\mathbf{A}_h) + a(\mathbf{B},\mathbf{A}_h) + (p,\nabla\cdot\mathbf{A}_h) \quad \forall\,\mathbf{A}_h\in H_{0h}, \\ (\nabla\cdot\mathbf{B}_h,q_h) = 0 \quad \forall\, q_h\in Q_{0h}, \end{cases}$$

where for any $\mathbf{B},\mathbf{A}\in H_0$, $a(\mathbf{B},\mathbf{A})$ and $a_h(\mathbf{B},\mathbf{A})$ are given by

$$\begin{aligned} a(\mathbf{B},\mathbf{A}) &= (\beta\,\nabla\times\mathbf{B},\nabla\times\mathbf{A}) + \gamma(\nabla\cdot\mathbf{B},\nabla\cdot\mathbf{A})\,,\\ a_h(\mathbf{B},\mathbf{A}) &= (\beta_h\,\nabla\times\mathbf{B},\nabla\times\mathbf{A}) + \gamma(\nabla\cdot\mathbf{B},\nabla\cdot\mathbf{A})\,.\end{aligned}$$

By taking $\mathbf{A}_h = \mathbf{B}_h$ in (5.1) and using Young's inequality and the bounds of $\beta(\mathbf{x})$ and $\beta_h(\mathbf{x})$, we can directly establish the following stability estimates on the projection P_h (cf. [10]):

LEMMA 5.1. *For any $\mathbf{B}\in H_0$ and $p\in Q_0$, let $(\mathbf{B}_h,p_h)$ be the projection of $(\mathbf{B},p)$ defined by* (5.1)*, then we have*

$$\|\mathbf{B}_h\|_{H_0(\mathbf{curl},\mathrm{div};\Omega)} \lesssim \|\mathbf{B}\|_{H_0(\mathbf{curl},\mathrm{div};\Omega)} + \|p\|\,.$$

Considering the discontinuity of coefficient $\beta(\mathbf{x})$ across the interfaces Γ_1 and Γ_2, the solution $(\mathbf{B},p)$ to the system (3.7) often has higher regularity locally inside each medium subdomain Ω_i $(i=1,2,3)$ than in the entire domain Ω. To make full use of the better local regularities of $(\mathbf{B},p)$ to establish the error estimates of the projection P_h, we can introduce a specially constructed interpolation operator by modifying the

Scott–Zhang operator [29] such that it preserves the boundary condition in H_0: for any $\mathbf{B} \in H_0$, we have $S_h\mathbf{B} \in H_{0h}$ (see [10] for details); and S_h has the following local approximation property (cf. [29, 10]):

$$\|w - S_h w\|_{W^{m,p}(K)} \lesssim h_K^{l-m} \|w\|_{W^{l,p}(S_K)} \quad \forall\, w \in W^{l,p}(S_K), \tag{5.2}$$

where $0 < m \le l$ and S_K is the union of all elements in $\mathcal{T}_h$, whose closure has nonempty intersection with K. We now establish the error estimates of form (5.2) in the entire domain Ω for functions with higher regularities locally in each subdomain Ω_k $(k = 1, 2, 3)$.

LEMMA 5.2. *For any $s \ge 0$, and $u \in X = H^1(\Omega) \cap H^{1+s}(\Omega_k)$ $(k = 1, 2, 3)$,*

$$\|u - S_h u\| \lesssim h^{1+\frac{2s}{3}} \sum_{k=1}^{3} \|u\|_{1+s,\Omega_k}, \quad \|u - S_h u\|_1 \lesssim h^{\frac{2s}{3}} \sum_{k=1}^{3} \|u\|_{1+s,\Omega_k}.$$

Proof. For any $u \in X$, let u_k be the restriction of u on Ω_k $(k = 1, 2, 3)$. Noting the interfaces Γ_1 and Γ_2 are smooth, one can extend (cf. [30]) $u_k \in H^{1+s}(\Omega_k)$ onto the whole domain Ω such that the extended function $\tilde{u}_k \in H^{1+s}(\Omega)$ and

$$\|\tilde{u}_k\|_{1+s,\Omega} \lesssim \|u_k\|_{1+s,\Omega_k} \quad \text{for } k = 1, 2, 3. \tag{5.3}$$

First, we consider the estimate on any noninterface element $K \notin \mathcal{T}_h^*$. Since u has H^{1+s}-regularity in such element K, one can follow the standard error estimate in [29] and make use of our construction of the face τ_i associated with each node a_i to derive

$$\|u - S_h u\|_{\mu,K} \lesssim h^{1+s-\mu} \sum_{i=1}^{3} \|u\|_{1+s,S_K \cap \Omega_i}, \quad \mu = 0, 1. \tag{5.4}$$

The tricky case happens to the interface elements. Without loss of generality, consider an interface element $K \in \mathcal{T}_h^*$ near the interface Γ_1. We analyze the errors in $\mathcal{K}_1$ and $\mathcal{K}_2$ separately. Clearly, $\mathcal{K}_1 \subset \Omega_1$, $\mathcal{K}_2 \subset \Omega_2$, $|\mathcal{K}_1| \eqsim h_K^4$ by (4.14). Then by Hölder's inequality, Sobolev embedding, and (5.2), we derive for any $2 \le p \le 6/(3-2s)$ and $\mu = 0, 1$ that

$$\begin{aligned} \|u - S_h u\|_{\mu,\mathcal{K}_1}^2 &\lesssim h_K^{\frac{4(p-2)}{p}} \|u - S_h u\|_{W^{\mu,p}(\mathcal{K}_1)}^2 \\ &\lesssim h_K^{\frac{4(p-2)}{p}} \|u - S_h u\|_{W^{\mu,p}(K)}^2 \lesssim h_K^{6-2\mu-\frac{8}{p}} \|u\|_{W^{1,p}(S_K)}^2. \end{aligned}$$

But on $\mathcal{K}_2$, by the choice of the face τ_i associated with the node a_i in the definition of S_h we know that

$$\tilde{u}_2 = u_2 \ \text{ on } \ \mathcal{K}_2, \quad S_h\, \tilde{u}_2 = S_h\, u \ \text{ on } \ \mathcal{K}_2.$$

Using this and (5.2), we derive

$$\|u - S_h u\|_{\mu,\mathcal{K}_2}^2 \lesssim \|\tilde{u}_2 - S_h \tilde{u}_2\|_{\mu,\mathcal{K}_2}^2 \lesssim \|\tilde{u}_2 - S_h \tilde{u}_2\|_{\mu,K}^2 \lesssim h_K^{2(1+s-\mu)} \|\tilde{u}_2\|_{1+s,S_K}^2, \tag{5.5}$$

combined with the previous estimate on $\mathcal{K}_1$ and (5.3) yields

$$\sum_{K \in \mathcal{T}_h^*} \|u - S_h u\|_{\mu,K}^2 \lesssim h^{2(1+s-\mu)} \sum_{k=1}^{3} \|\tilde{u}_k\|_{1+s,\Omega_k}^2 + \sum_{K \in \mathcal{T}_h^*} h_K^{6-2\mu-\frac{8}{p}} \|u\|_{W^{1,p}(S_K)}^2. \tag{5.6}$$

Then by Hölder's inequality and the fact that the number of interface elements in $\mathcal{T}_h^* \lesssim h^{-2}$,

$$\sum_{K\in\mathcal{T}_h^*} \|u - S_h u\|_{\mu,K}^2 \lesssim h^{2(1+s-\mu)} \sum_{k=1}^{3} \|u\|_{1+s,\Omega_k}^2 + h^{4-2\mu-\frac{4}{p}} \Big(\sum_{K\in\mathcal{T}_h^*} \|u\|_{W^{1,p}(S_K)}^p \Big)^{2/p}.$$

Now the desired estimate follows by taking $p = 6/(3-2s)$ above and using (5.4). ∎

The following lemma provides a crucial observation needed in the subsequent analysis.

LEMMA 5.3. *Let $\mathcal{K}$ be the interface part of any interface element $K \in \mathcal{T}_h^*$ such that $|\mathcal{K}| \eqsim h_K^4$ (cf. (4.14)), then the following estimates hold:*

$$\|\nabla \times \mathbf{A}_h\|_{0,\mathcal{K}}^2 \lesssim h_K \|\nabla \times \mathbf{A}_h\|_{0,K}^2, \quad \|\mathbf{A}_h\|_{0,\mathcal{K}}^2 \lesssim h_K \|\mathbf{A}_h\|_{0,K}^2 \quad \forall\, \mathbf{A}_h \in H_{0h}. \tag{5.7}$$

Proof. We prove only the first inequality in (5.7), the second is similar. Let K be an interface element with 4 vertices, v_1, v_2, v_3, and v_4, and d_K be the largest distance from the curved side of $\mathcal{K}$ to its opposite face of K. Since the interfaces Γ_1 and Γ_2 are C^∞-smooth, we can easily show that $d_K \le Ch_K^2$. Then we can construct a cube $\mathcal{C}(\mathcal{K})$ such that $\mathcal{K} \subset C(\mathcal{K})$ and $\mathcal{C}(\mathcal{K})$ has a height d_K and a rectangular base of length $\alpha_1 h_K$ and width $\alpha_2 h_K$, where α_1 and α_2 are two positive constants independent of mesh size h. Then we divide $\mathcal{C}(\mathcal{K})$ into 6 small tetrahedra $\mathcal{K}^1, \ldots, \mathcal{K}^6$.

By scaling arguments, one can easily verify the equivalence

$$\|q\|_{0,A}^2 \eqsim |A| \sum_{i=1}^{4} (q(a_i))^2 \quad \forall\, q \in P_1(A) \tag{5.8}$$

for any tetrahedron A with vertices a_1, a_2, a_3, and a_4. Let p be a component of $\nabla \times \mathbf{A}_h$ for some $\mathbf{A}_h \in H_{0h}$, then $p \in P_1(K)$. Clearly we also see $p \in P_1(\mathcal{K})$, and p can be naturally extended to $\tilde{p} \in P_1(\mathcal{C}(\mathcal{K}))$, thus

$$\|p\|_{0,\mathcal{K}}^2 \le \|\tilde{p}\|_{0,\mathcal{C}(\mathcal{K})}^2 \le \sum_{i=1}^{6} \|\tilde{p}\|_{0,\mathcal{K}^i}^2.$$

But using (5.8) and the fact that the value $\tilde{p}$ at any point in $\mathcal{K}^i$ can be expressed as a convex combination of the values of p at the 4 vertices of K, we obtain

$$\|p\|_{0,\mathcal{K}}^2 \lesssim h_K^4 \sum_{j=1}^{4} (p(v_j))^2 \eqsim h_K |K| \sum_{j=1}^{4} (p(v_j))^2 \eqsim h_K \|p\|_{0,K}^2.$$

This proves the first estimate in (5.7). ∎

Using the interpolation error estimates in Lemma 5.2 and the convergence theory in Theorem 4.1, we can now derive the error estimate for the projection operator P_h defined in (5.1).

LEMMA 5.4. *Let $\mathbf{B} \in H_0(\mathbf{curl}, \mathrm{div}; \Omega)$ and $p \in L_0^2(\Omega)$ be given such that $\mathbf{B} \in H^{1+s_1}(\Omega_k)$ in each Ω_k $(k = 1, 2, 3)$ for some $0 \le s_1 < 1$ and $p \in H^{s_2}(\Omega)$ for some $0 \le s_2 < 1$. Then the following error estimates hold for the projection $(\mathbf{B}_h, p_h)$ of $(\mathbf{B}, p)$ defined in (5.1):*

$$\sum_{i=1}^{3} \|\mathbf{B} - \mathbf{B}_h\|_{H(\mathbf{curl},\mathrm{div};\Omega_i)}^2 \lesssim h^{\frac{4s_1}{3}} \sum_{i=1}^{3} \|\mathbf{B}\|_{1+s_1,\Omega_i}^2 + h^{2s_2} \|p\|_{s_2,\Omega}^2.$$

Proof. Let $X = H_0(\mathbf{curl}, \mathrm{div}; \Omega)$ and $M = L^2_0(\Omega)$, $X_h = H_{0h}$, and $M_h = Q_{0h}$, then we can apply Theorem 4.1 to the system (5.1) to obtain

$$\|\mathbf{B} - \mathbf{B}_h\|_{H_0} \lesssim \inf_{\mathbf{A}_h \in H_{0h}} \|\mathbf{B} - \mathbf{A}_h\|_{H_0} + \inf_{q_h \in Q_{0h}} \|p - q_h\| + \sup_{\mathbf{A}_h \in V_h} \frac{a(\mathbf{B}, \mathbf{A}_h) - a_h(\mathbf{B}, \mathbf{A}_h)}{\|\mathbf{A}_h\|_{H_{0h}}} . \tag{5.9}$$

Noting that for $\mathbf{B} \in H_0$, we have $S_h\mathbf{B} \in H_{0h}$. On the other hand, for $p \in L^2_0(\Omega)$, let $\pi_h p$ be its standard L^2 projection in Q_h. Clearly $\pi_h p$ may not be in Q_{0h}. But if we set $\tilde{p}_h = \pi_h p - \overline{\pi_h p}$, where $\overline{q}$ stands for the average of q over Ω for any $q \in L^2(\Omega)$, then we have $\tilde{p}_h \in Q_{0h}$, and the following estimates hold using the standard approximation property of the L^2 projection:

$$\|p - \tilde{p}_h\| = \|(p - \pi_h p) - \overline{(p - \pi_h p)}\| \le \|p - \pi_h p\| \lesssim h^{s_2} \|p\|_{s_2,\Omega} .$$

Using this and Lemma 5.2, we derive by taking $\mathbf{A}_h = S_h\mathbf{B}$ and $q_h = \tilde{p}_h$ in (5.9) that

$$\inf_{\mathbf{A}_h \in H_{0h}} \|\mathbf{B} - \mathbf{A}_h\|_{H_0} + \inf_{q_h \in Q_{0h}} \|p - q_h\| \lesssim h^{\frac{2s_1}{3}} \sum_{i=1}^{3} \|\mathbf{B}\|_{1+s_1,\Omega_i} + h^{s_2} \|p\|^2_{s_2,\Omega} . \tag{5.10}$$

It remains to estimate the last term in (5.9). Let $\mathcal{K}$ be the same as in Lemma 5.3, then by the definition of $a(\cdot,\cdot)$ and $a_h(\cdot,\cdot)$, we can write for any $\mathbf{A}_h \in H_{0h}$ (cf. [10]),

$$a(\mathbf{B}, \mathbf{A}_h) - a_h(\mathbf{B}, \mathbf{A}_h) = \sum_{K \in \mathcal{T}_h^*} \int_{\mathcal{K}} (\beta(x) - \beta_h(x)) \nabla \times \mathbf{B} \cdot \nabla \times \mathbf{A}_h dx .$$

Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} &|a(\mathbf{B}, \mathbf{A}_h) - a_h(\mathbf{B}, \mathbf{A}_h)| \\ &\quad \lesssim \sum_{K \in \mathcal{T}_h^*} \left\{ \|\nabla \times S_h\mathbf{B}\|_{0,\mathcal{K}} \|\nabla \times \mathbf{A}_h\|_{0,\mathcal{K}} + \|\nabla \times (\mathbf{B} - S_h\mathbf{B})\|_{0,\mathcal{K}} \|\nabla \times \mathbf{A}_h\|_{0,\mathcal{K}} \right\} . \end{aligned}$$

By Lemmas 5.3 and 5.2, we deduce

$$\begin{aligned} |a(\mathbf{B}, \mathbf{A}_h) - a_h(\mathbf{B}, \mathbf{A}_h)| &\lesssim \sum_{K \in \mathcal{T}_h^*} h_K \|\nabla \times S_h\mathbf{B}\|_{0,K} \|\nabla \times \mathbf{A}_h\|_{0,K} \\ &\quad + \sum_{K \in \mathcal{T}_h^*} h_K^{1/2} \|\nabla \times (\mathbf{B} - S_h\mathbf{B})\|_{0,K} \|\nabla \times \mathbf{A}_h\|_{0,K} \\ &\lesssim h \|\nabla \times S_h\mathbf{B}\|_{0,\Omega} \|\nabla \times \mathbf{A}_h\|_{0,\Omega} \\ &\quad + h^{1/2} \|\nabla \times (\mathbf{B} - S_h\mathbf{B})\|_{0,\Omega} \|\nabla \times \mathbf{A}_h\|_{0,\Omega} \\ &\lesssim (h + h^{1/2+2s_1/3}) \left(\sum_{k=1}^{3} \|\mathbf{B}\|_{1+s_1,\Omega_k} \right) \|\nabla \times \mathbf{A}_h\|_{0,\Omega} , \end{aligned}$$

and this completes the proof of Lemma 5.4. ∎

Now, the above preparations enable us to make full use of the better local regularity of $\mathbf{B}$ in each subdomain Ω_k to derive the main results of this section, the finite element convergence.

THEOREM 5.1. *Let* $(\mathbf{B}, p) \in H^2(0,T;H_0) \times L^2(0,T;L_0^2(\Omega))$ *be the solution to the variational problem* (3.7) *such that* $\mathbf{B} \in H^1(0,T;H^{1+s_1}(\Omega_k))$ *in each* Ω_k $(k=1,2,3)$ *for some* $0 \le s_1 < 1$ *and* $p \in H^1(0,T;H^{s_2}(\Omega))$ *for some* $0 \le s_2 < 1$. *And let* $(\mathbf{B}_h, p_h)$ *be the finite element solution to the fully discrete finite element approximation* (4.15), *then we have the following error estimates:*

$$\begin{aligned}
&\max_{1\le n\le M} \|\mathbf{B}_h^n - \mathbf{B}^n\|^2 + \tau \sum_{n=1}^{M} \Big\{ \|\nabla\times(\mathbf{B}_h^n - \mathbf{B}^n)\|^2 + \|\nabla\cdot(\mathbf{B}_h^n - \mathbf{B}^n)\|^2 \Big\} \\
&\lesssim h^{\frac{4s_1}{3}} \sum_{k=1}^{3} \|\mathbf{B}\|^2_{H^1(0,T;H^{1+s_1}(\Omega_k))} + h^{2s_2}\|p\|^2_{H^1(0,T;H^{s_2}(\Omega))} \\
&\quad + \tau^2\{\|\mathbf{B}\|^2_{H^2(0,T;H_0)} + \|p\|^2_{L^2(Q_T)}\}.
\end{aligned}$$

Proof. Our aim is to estimate the error $(\mathbf{B}^n - \mathbf{B}_h^n)$. Using the relation

$$\mathbf{B}^n - \mathbf{B}_h^n = (\mathbf{B}^n - \bar{\mathbf{B}}^n) + (\bar{\mathbf{B}}^n - P_h\bar{\mathbf{B}}^n) + (P_h\bar{\mathbf{B}}^n - \mathbf{B}_h^n), \tag{5.11}$$

and the projection results from Lemma 5.4, it suffices to estimate the difference $\xi_h^n = (P_h\bar{\mathbf{B}}^n - \mathbf{B}_h^n)$ in the specified norms. To do so, letting $\mathbf{A} = \tau^{-1}\mathbf{A}_h \in H_{0h}$ and $q = q_h \in Q_{0h}$ in (3.7), then integrating over $[t_{n-1}, t_n]$ we obtain

$$\begin{cases} (\partial_\tau \mathbf{B}^n, \mathbf{A}_h) + (\beta\nabla\times\bar{\mathbf{B}}^n, \nabla\times\mathbf{A}_h) + \gamma(\nabla\cdot\bar{\mathbf{B}}^n, \nabla\cdot\mathbf{A}_h) + (\bar{p}^n, \nabla\cdot\mathbf{A}_h) \\ \qquad = R_\alpha(\bar{f}_B^n, \nabla\times\mathbf{A}_h) + R_m(\overline{\mathbf{u}\times\mathbf{B}}^n, \nabla\times\mathbf{A}_h) \quad \forall\, \mathbf{A}_h \in H_{0h}; \\ (\nabla\cdot\bar{\mathbf{B}}^n, q_h) = 0 \quad \forall\, q_h \in Q_{0h}, \end{cases} \tag{5.12}$$

where

$$\bar{f}_B^n = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} \frac{f}{1+\sigma|\mathbf{B}|^2}\mathbf{B}(t)dt.$$

Subtracting (4.15) from (5.12) yields

$$\begin{aligned}
&(\partial_\tau \xi_h^n, \mathbf{A}_h) + (\beta_h\nabla\times\xi_h^n, \nabla\times\mathbf{A}_h) + \gamma(\nabla\cdot\xi_h^n, \nabla\cdot\mathbf{A}_h) \\
&= (\partial_\tau(P_h\bar{\mathbf{B}}^n - \mathbf{B}^n), \mathbf{A}_h) + R_\alpha\Big(\bar{f}_B^n - \frac{f_h^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}_h^n, \nabla\times\mathbf{A}_h\Big), \\
&\quad + R_m(\overline{\mathbf{u}\times\mathbf{B}}^n - \mathbf{u}_h^n\times\mathbf{B}_h^n, \nabla\times\mathbf{A}_h) \\
&\quad + (\beta_h\nabla\times P_h\bar{\mathbf{B}}^n - \beta\nabla\times\bar{\mathbf{B}}^n, \nabla\times\mathbf{A}_h) + (\nabla\cdot(P_h\bar{\mathbf{B}}^n - \bar{\mathbf{B}}^n), \nabla\cdot\mathbf{A}_h) \\
&\quad + (P_h\bar{p}^n - \bar{p}^n, \nabla\cdot\mathbf{A}_h) + (p_h^n - P_h\bar{p}^n, \nabla\cdot\mathbf{A}_h) \quad \forall\, \mathbf{A}_h \in H_{0h}\,.
\end{aligned}$$

Letting $\mathbf{A}_h = \tau\xi_h^n \in H_{0h}$ above, then using the second equations in both (4.15) and (5.12) and the definition of the projection P_h, we come to (cf. [10])

$$\begin{aligned}
&\tau(\partial_\tau\xi_h^n, \xi_h^n) + \tau(\beta_h\nabla\times\xi_h^n, \nabla\times\xi_h^n) + \gamma\tau(\nabla\cdot\xi_h^n, \nabla\cdot\xi_h^n) \\
&= R_\alpha\tau\Big(\bar{f}_B^n - \frac{f^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}_h^n, \nabla\times\xi_h^n\Big) + R_m\tau(\overline{\mathbf{u}\times\mathbf{B}}^n - \mathbf{u}^n\times\mathbf{B}_h^n, \nabla\times\xi_h^n) \\
&\quad + \tau(\partial_\tau(P_h\bar{\mathbf{B}}^n - \mathbf{B}^n), \xi_h^n) + \tau(\bar{\mathbf{B}}^n - P_h\bar{\mathbf{B}}^n, \xi_h^n) \\
&\quad + \tau\Big(\frac{R_\alpha(f^n - f_h^n)}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}_h^n + R_m(\mathbf{u}^n - \mathbf{u}_h^n)\times\mathbf{B}_h^n, \nabla\times\xi_h^n\Big) \\
&\equiv: (\mathrm{I})_1 + (\mathrm{I})_2 + (\mathrm{I})_3 + (\mathrm{I})_4 + (\mathrm{I})_5\,.
\end{aligned} \tag{5.13}$$

Obviously, $(\mathrm{I})_4$ can be estimated immediately by the projection property. Below, we shall analyze $(\mathrm{I})_1$, $(\mathrm{I})_2$, $(\mathrm{I})_3$, and $(\mathrm{I})_5$ one by one. For the estimation of $(\mathrm{I})_1$, we first consider

$$(\mathrm{II})_1 \equiv: \bar{f}_B^n - \frac{f^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\mathbf{B}_h^n.$$

By direct manipulations, we have (cf. [10])

$$\begin{aligned}(\mathrm{II})_1 = {} & \frac{1}{\tau}\int_{t_{n-1}}^{t_n} \frac{(f-f^n)+\sigma f\,(|\mathbf{B}_h^{n-1}|^2-|\mathbf{B}|^2)+\sigma(f-f^n)|\mathbf{B}|^2}{(1+\sigma|\mathbf{B}|^2)(1+\sigma|\mathbf{B}_h^{n-1}|^2)}\mathbf{B}(t)dt \\ & + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} \frac{f^n}{1+\sigma|\mathbf{B}_h^{n-1}|^2}\Big\{(\mathbf{B}(t)-\bar{\mathbf{B}}^n)+(\bar{\mathbf{B}}^n-P_h\bar{\mathbf{B}}^n)+(P_h\bar{\mathbf{B}}^n-\mathbf{B}_h^n)\Big\}dt,\end{aligned}$$

which can be easily bounded by

(5.14)

$$\begin{aligned}|(\mathrm{II})_1| \le {} & \frac{2}{\tau}\int_{t_{n-1}}^{t_n} |f|\,\{|\mathbf{B}_h^{n-1}-P_h\bar{\mathbf{B}}^{n-1}|+|P_h\bar{\mathbf{B}}^{n-1}-\bar{\mathbf{B}}^{n-1}|+|\bar{\mathbf{B}}^{n-1}-\mathbf{B}|\}dt \\ & + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} |f-f^n|\,|\mathbf{B}(t)|dt + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} |f^n|\,\Big(|\mathbf{B}(t)-\bar{\mathbf{B}}^n|+|\bar{\mathbf{B}}^n-P_h\bar{\mathbf{B}}^n|+|\xi_h^n|\Big)dt.\end{aligned}$$

By the standard error estimates [5, 14, 10], we obtain from (5.14) that

(5.15)

$$|(\mathrm{II})_1| \le 5f_M\Big\{\sqrt{\tau}\|\mathbf{B}\|_{L^2(t_{n-1},t_n)} + \sqrt{\tau}\|\mathbf{B}_t\|_{L^2(t_{n-2},t_n)} + \sum_{k=n-1}^{n}|\xi_h^k| + \sum_{k=n-1}^{n}|P_h\bar{\mathbf{B}}^k-\bar{\mathbf{B}}^k|\Big\}.$$

Similarly, for the estimation of $(\mathrm{I})_2$, we first analyze $(\mathrm{II}_2) := \overline{\mathbf{u}\times\mathbf{B}}^n - \mathbf{u}^n\times\mathbf{B}_h^n$. We write

$$(\mathrm{II})_2 = \frac{1}{\tau}\int_{t_{n-1}}^{t_n}\{(\mathbf{u}(t)-\mathbf{u}^n)\times\mathbf{B}(t)+\mathbf{u}^n\times(\mathbf{B}(t)-\mathbf{B}_h^n)\}dt\,,$$

this leads readily to (with $I_n = (t_{n-1}, t_n]$)

$$|(\mathrm{II})_2| \le u_M\sqrt{\tau}\|\mathbf{B}\|_{L^2(I_n)} + 2u_M(\sqrt{\tau}\|\mathbf{B}_t\|_{L^2(I_n)} + |\bar{\mathbf{B}}^n - P_h\bar{\mathbf{B}}^n| + |\xi_h^n|)\,. \tag{5.16}$$

Next, we estimate the following term needed in (5.13):

$$\begin{aligned}(\mathrm{II})_3 \equiv: {} & \partial_\tau(P_h\bar{\mathbf{B}}^n - \mathbf{B}^n) = \partial_\tau(P_h\mathbf{B}^n-\mathbf{B}^n) + P_h\partial_\tau(\bar{\mathbf{B}}^n-\mathbf{B}^n) \\ = {} & \frac{1}{\tau}\int_{t_{n-1}}^{t_n}(P_h\mathbf{B}_t(s)-\mathbf{B}_t(s))ds + P_h\partial_\tau(\bar{\mathbf{B}}^n-\mathbf{B}^n). \qquad (5.17)\end{aligned}$$

For the second term above, we can write after some manipulations [10] that

$$\partial_\tau(\bar{\mathbf{B}}^n-\mathbf{B}^n) = \frac{1}{\tau^2}\int_{t_{n-1}}^{t_n}\int_{s-\tau}^{s}\int_{s}^{\mu}\mathbf{B}_{tt}(\lambda)d\lambda d\mu ds,$$

this enables us to rewrite $(\mathrm{II})_3$ as

$$(\mathrm{II})_3 = \frac{1}{\tau}\int_{t_{n-1}}^{t_n}(P_h\mathbf{B}_t(s)-\mathbf{B}_t(s))ds + \frac{1}{\tau^2}\int_{t_{n-1}}^{t_n}\int_{s-\tau}^{s}\int_{s}^{\mu}P_h\mathbf{B}_{tt}(\lambda)d\lambda d\mu ds,$$

and so,

$$|(\mathrm{II})_3| \le \frac{1}{\sqrt{\tau}}\|P_h\mathbf{B}_t - \mathbf{B}_t\|_{L^2(I_n)} + \sqrt{\tau}\|P_h\mathbf{B}_{tt}\|_{L^2(I_n)}. \tag{5.18}$$

For the last term $(\mathrm{I})_5$ in (5.13), by Lemma 5.3 and Young's inequality there exists some constant $\tilde{C}$ independent of τ and h such that [10]

$$\begin{aligned} |(\mathrm{I})_5| &\le 2\tau\, h\, \tilde{C}(R_\alpha f_M + 2R_m u_M) \sum_{K\in\mathcal{T}_h^*} \|\mathbf{B}_h^n\|_{0,K}\|\nabla\times\xi_h^n\|_{0,K} \\ &\le \frac{\beta_m}{4}\tau\|\nabla\times\xi_h^n\|^2 + C\tau\, h^2\, \|\mathbf{B}_h^n\|^2\,. \end{aligned} \tag{5.19}$$

Now, we obtain from (5.13) and (5.19) and Young's inequality that

$$\|\xi_h^n\|^2 - \|\xi_h^{n-1}\|^2 + \tau\beta_m\|\nabla\times\xi_h^n\|^2 + \gamma\tau\|\nabla\cdot\xi_h^n\|^2 \lesssim (\mathrm{III})_1, \tag{5.20}$$

where the term $(\mathrm{III})_1$ is given by

$$(\mathrm{III})_1 = \tau\left\{\|(\mathrm{II})_1\|^2 + \|(\mathrm{II})_2\|^2 + \|(\mathrm{II})_3\|^2 + \|\bar{\mathbf{B}}^n - P_h\bar{\mathbf{B}}^n\|^2 + \|\xi_h^n\|^2 + h^2\|\mathbf{B}_h^n\|^2\right\}.$$

Using the estimates for $(\mathrm{II})_1$, $(\mathrm{II})_2$, and $(\mathrm{II})_3$ in (5.15)–(5.18), we can further estimate $(\mathrm{III})_1$ by

$$\begin{aligned} (\mathrm{III})_1 \lesssim{}& \int_{t_{n-1}}^{t_n}\|P_h\mathbf{B}_t(s) - \mathbf{B}_t(s)\|^2 ds + \tau^2\int_{t_{n-1}}^{t_n}\|P_h\mathbf{B}_{tt}(t)\|^2 dt + \tau\{\|\xi_h^{n-1}\|^2 + \|\xi_h^n\|^2\} \\ &+ \tau\left\{\sum_{k=n-1}^{n}\|P_h\bar{\mathbf{B}}^k - \bar{\mathbf{B}}^k\|^2 + \tau\int_{t_{n-2}}^{t_n}(\|\mathbf{B}(t)\|^2 + \|\mathbf{B}_t(t)\|^2)dt\right\} + \tau\, h^2\|\mathbf{B}_h^n\|^2. \end{aligned}$$

Summing both sides of (5.20) over $n = 1, 2, \ldots, k \le M$ yields

$$\begin{aligned} &\|\xi_h^k\|^2 + \beta_m\tau\sum_{n=1}^{k}\|\nabla\times\xi_h^n\|^2 + 2\tau\sum_{n=1}^{k}\|\nabla\cdot\xi_h^n\|^2 \\ &\lesssim \|\xi_h^0\|^2 + \|P_h\mathbf{B}(0) - \mathbf{B}(0)\|^2 + \tau\sum_{n=1}^{k}\|\xi_h^n\|^2 + \tau\, h^2\sum_{n=1}^{k}\|\mathbf{B}_h^n\|^2 \\ &\quad + \int_0^T\|P_h\mathbf{B}_t(t) - \mathbf{B}_t(t)\|^2 dt + \tau^2\int_0^T\|P_h\mathbf{B}_{tt}(t)\|^2 dt \\ &\quad + \tau\sum_{n=1}^{k}\|P_h\bar{\mathbf{B}}^n - \bar{\mathbf{B}}^n\|^2 + \tau\int_0^T(\|\mathbf{B}(t)\|^2 + \|\mathbf{B}_t(t)\|^2)dt\Big\}. \end{aligned} \tag{5.21}$$

Finally using Lemmas 5.1, 5.2, and 5.4, and applying the discrete Gronwall inequality, we are led to the error estimates in Theorem 5.1. ∎

6. Application to a solar interface dynamo. For the astrophysical application of the mathematical theory, we shall concentrate on the numerical modelling of solar interface dynamos. Helioseismology reveals the existence of a highly differentially rotating transition zone at the bottom of the convection zone, which is usually referred to as the solar tachocline [28]. It is thought that the tachocline offers an ideal location for the generation and storage of the Sun's strong toroidal magnetic fields.

In other words, the large-scale solar surface magnetic activities can be interpreted as a result of the rising and emerging of tachocline-seated, strong toroidal magnetic fields driven by magnetic buoyancy [34]. The existence of the tachocline leads to development of the solar interface dynamo first proposed by Parker [25], in which the generation of a weak poloidal magnetic field and a strong toroidal magnetic field takes place in separate fluid regions. Parker's interface dynamo concept depicts an attractive picture of generating a strong toroidal magnetic field within the tachocline while avoiding the dilemma relating to the strong alpha quenching in the convection zone. We shall apply the finite element dynamo theory and algorithm discussed in the previous sections to the problem of solar interface dynamo modelling.

In the solar interface dynamo model, we shall take $\mathbf{u}$ as the solar-like internal differential rotation profile, a result of the helioseismic inversion (e.g., [28]) while the function f is assumed to be given by

$$f(\mathbf{x},t) = \sin^2\theta\cos\theta\sin\left[\pi\frac{(r-r_1)}{(r_2-r_1)}\right],$$

where (r,θ,ϕ) is the spherical polar coordinates. Similar forms have been used in the previous solar dynamo simulations [27, 21, 16]. Furthermore, the weaker pole-equator differential rotation in the convection zone is neglected and the amplification of the toroidal magnetic field only occurs in the tachocline. It follows that the two magnetic induction sources, the generation of a poloidal field in the convection zone, and the amplification of the toroidal field in the tachocline, is spatially separated, as suggested by Parker [25].

We have simulated three nonlinear finite element dynamos at $R_\alpha = 30$ for different magnetic Reynolds numbers, $R_m = 100$, 200, and 500. Figure 2 displays magnetic energies of the three nonlinear dynamo solutions as a function of time. The corresponding butterfly diagram, contours of the azimuthal magnetic field evaluated at the bottom of the convection zone plotted against time, is also shown in Figure 2 for the case with $R_m = 200$. In Figure 3, we illustrate the time-dependent spatial structure of the generated magnetic field in a meridional plane for $R_m = 200$, showing an equatorially propagating dynamo wave similar to that of the solar cycle.

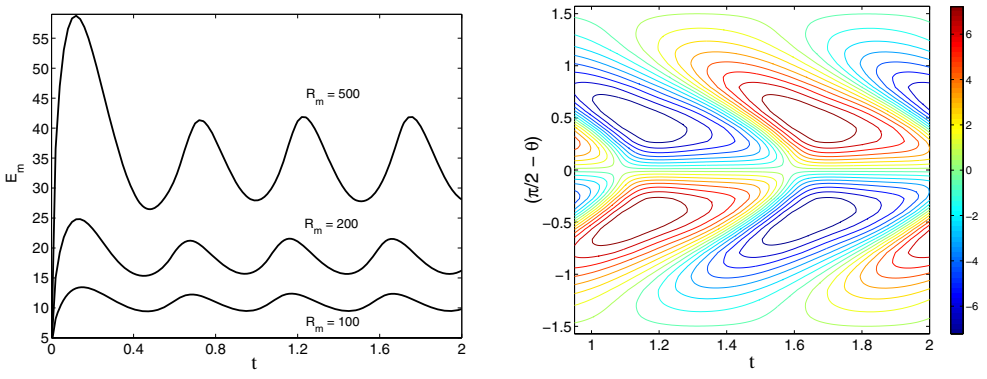


FIG. 2. *The left panel shows magnetic energy E_m of the dynamo as a function of time with a steady tachocline for different values of R_m at $R_\alpha = 30$ with $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 150$. The right panel displays "a butterfly diagram" for the solution $R_m = 200$ with the azimuthal magnetic field evaluated at the bottom of the convection zone.*

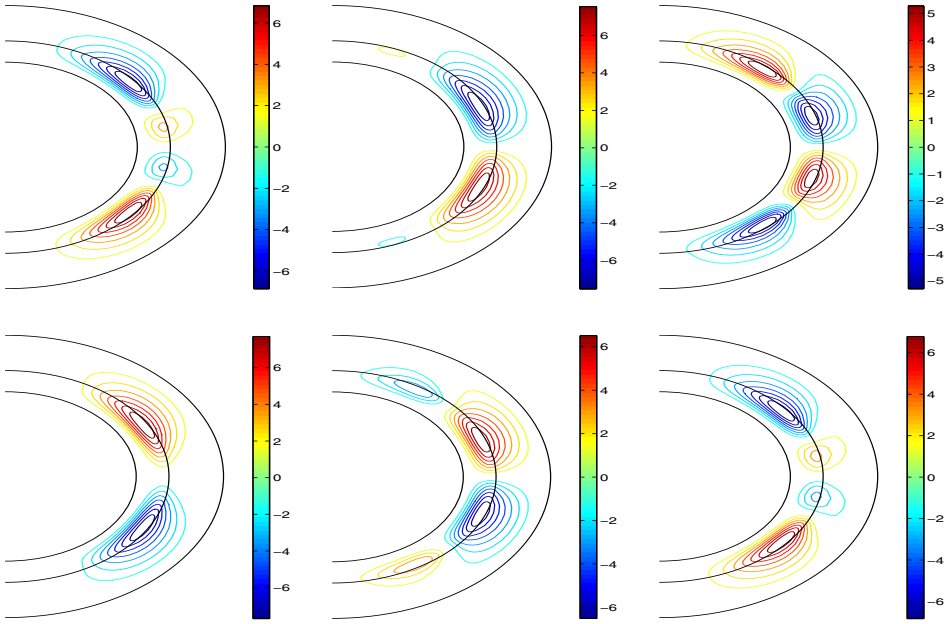


FIG. 3. *Contours of the azimuthal field B_ϕ in a meridional plane plotted at six different instants for $t = 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$ (from top left to right and then from lower left to right) for $R_m = 200$, $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 150$.*

There are a number of important features shown in our finite element dynamo solutions. First, the effect of the large-scale differential rotation $\mathbf{u}$ in the tachocline always gives rise to an oscillatory dynamo with a period of about one magnetic diffusion unit, which is about 10 years if we adopt $\lambda_2 = 10^8 \mathrm{m}^2\mathrm{s}^{-1}$, approximate to what has been observed in the solar magnetic field. Second, the interface dynamo solutions always select dipolar symmetry and propagate equator ward though the numerical simulation is fully three-dimensional, which is again consistent with the observed feature of the solar magnetic field. Finally, the generated magnetic field mainly concentrates in the vicinity of the interface between the tachocline and the convection zone. A strong toroidal magnetic field in the tachocline is likely to be susceptible to magnetic buoyancy instabilities leading to a quick eruption of the field into the surface of the Sun in the form of sunspots.

7. Concluding remarks. Modelling stellar and planetary dynamos represents an important, highly active research front in astrophysics and planetary physics. Nearly all current stellar dynamo models are based on spectral methods in terms of spherical harmonic expansions, which are computationally inefficient on modern parallel computers and limit the application to general dynamo models, especially to the models with variable physical coefficients of space and time. The finite element method discussed in this paper offers an attractive alternative for simulating dynamos in spherical geometry.

The first attempt at using finite element methods for numerical simulations of spherical dynamos was made in [11]. The current work presents the first mathematical theory and numerical analysis for mean-field spherical dynamos, and it has made contributions in the following aspects: (1) The well-posedness of the mean-field

dynamo system is rigorously demonstrated; the dynamo system is characterized in terms of a saddle-point type formulation which can be conveniently approximated by finite element methods. (2) The existing convergence theory on saddle-point systems is improved and generalized so that the symmetric part of the bilinear form allows the approximations of curved interfaces by straight polygons and numerical integrations on polyhedra with curved faces are replaced by much easier integrations on polyhedra with planar faces; and this is the first work of such type on saddle-point systems. (3) A fully discrete finite element method is proposed for the interface dynamo system with discontinuous coefficients, and error estimates are established under very weak global and local regularity assumptions on the solutions, and this work seems to be the first in achieving error estimates of numerical methods for three-dimensional interface PDEs with curved interfaces, especially for nonlinear PDEs. (4) The application of the proposed numerical method to a solar interface dynamo verifies some important physical observations.

We believe that the mathematical theory and finite element methods for spherical dynamos and their successful application to the solar interface dynamo presented in this paper open up an exciting opportunity for future numerical simulation of stellar dynamos. We also believe that the finite element theory and method developed in the paper would also benefit other research communities in geophysics, planetary physics, and astrophysics where the magnetic field and spherical geometry play an essential role.

REFERENCES

[1] M. H. Acuna and N. F. Ness, *Jupiter's main magnetic field measured by Pioneer* 11, Nature, 253 (1975), pp. 327–328.

[2] F. Assous, P. Degond, E. Heintzé, P. A. Raviart, and J. Segré, *On a finite element method for solving the three-dimensional Maxwell equations*, J. Comput. Phys., 109 (1993), pp. 222–237.

[3] A. Brandenburg, *The case for a distributed solar dynamo shaped by near-surface shear*, Astrophy. J., 625 (2005), pp. 539–547.

[4] A. Brandenburg and K. Subramanian, *Astrophyiscal magnetic fields and nonlinear dynamo theory*, Physics Reports, 417 (2005), pp. 1–209.

[5] H. T. Banks and J. Zou, *Regularity and approximation of systems arising in electromagnetic interrogation of dielectric material*, Numer. Funct. Anal. Optim., 20 (1999), pp. 609–627.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[7] A. S. Brun and J. Toomre, *Turbulent convection under the influence of rotation: sustaining a strong differential rotation*, Astrophy. J., 570 (2002), pp. 865–885.

[8] B. F. Burke and K. L. Franklin, *Observations of a variable radio source associated with the planet Jupiter*, J. Geophys. Res., 60 (1952), pp. 213–217.

[9] F. Cattaneo and D. W. Hughes, *Nonlinear saturation of the turbulent alpha effect where a large scale field is imposed*, Phys. Rev. E, 54 (1996), pp. R4532–R4535.

[10] K. H. Chan, K. Zhang, and J. Zou, *Spherical Interface Dynamos: Mathematical Theory, Finite Element Approximation and Application*, Technical Report CUHK-2005-10 (331), Department of Mathematics, The Chinese University of Hong Kong, Hong Kong, 2005, available online at http://www.math.cuhk.edu.hk/en/report/index.php.

[11] K. H. Chan, K. Zhang, J. Zou, and G. Schubert, *A nonlinear* 3*-D spherical alpha-square dynamo using a finite element method*, Phys. Earth Planet. Int., 128 (2001), pp. 35–50.

[12] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[13] P. Ciarlet, Jr., and J. Zou, *Finite element convergence for the Darwin model to Maxwell's equations*, RAIRO Math. Model. Numer. Anal., 31 (1997), pp. 213–249.

[14] P. Ciarlet, Jr., and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell's equations*, Numer. Math., 82 (1999), pp. 193–219.

[15] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland Publishing, Amsterdam-New York-Oxford, 1978.
[16] M. Dikpati and P. Charbonneau, *A Babcock-Leighton flux transport dynamo with solar-like differential rotation*, Astrophy. J., 518 (1999), pp. 508–520.
[17] M. Feistauer and A. Zenisek, *Finite element solution of nonlinear elliptic problems*, Numer. Math., 50 (1987), pp. 451–475.
[18] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[19] G. Glatzmaier and P. Roberts, *A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle*, Phys. Earth Planet. Int., 91 (1995), pp. 63–75.
[20] P. Hood and G. Taylor, *Navier-Stokes equations using mixed interpolation*, in Finite Element Methods in Flow Problems, J. Oden, ed., UAH Press, Huntsville, AL, 1974.
[21] J. A. Markiel and J. H. Thomas, *Solar interface dynamo models with a realistic rotation profile*, Astrophy. J., 523 (1999), pp. 827–837.
[22] M. S. Miesch, J. R. Elliott, J. Toomre, T. L. Clune, G. A. Glatzmaier, and P. A. Gilman, *Three-dimensional spherical simulations of solar convection.* I. *Differential rotation and pattern evolution achieved with laminar and turbulent states*, Astrophy. J., 532 (2000), pp. 593–615.
[23] H. K. Moffatt, *Magnetic Field Generation in Electrically Conducting Fluids*, Cambridge University Press, Cambridge, UK, 1978.
[24] E. N. Parker, *Cosmical Magnetic Fields*, Clarendon Press, Oxford, 1979.
[25] E. N. Parker, *A solar dynamo surface wave at the interface between convection and nonuniform rotation*, Astrophy. J., 408 (1993), pp. 707–719.
[26] P. H. Roberts and A. M. Soward, *Dynamo theory*, in Annual Review of Fluid Mechanics, Vol. 24, Annual Reviews, Palo Alto, CA, 1992, pp. 459–512.
[27] G. Ruediger and A. Brandenburg, *A solar dynamo in the overshoot layer: Cycle period and butterfly diagram*, Astronomy and Astrophysics, 296 (1995), pp. 557–566.
[28] J. Schou et al., *Helioseismic studies of differential rotation in the solar envelope by the solar oscillations investigation using Michelson Doppler Imager*, Astrophy. J, 505 (1998), pp. 390–417.
[29] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[30] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[31] M. Stix, *The Sun: An Introduction*, Springer-Verlag, Berlin, 2002.
[32] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland Publishing, Amsterdam, 1977.
[33] R. Verfürth, *Error estimates for a mixed finite element approximation of the Stokes equations*, RAIRO Model. Math. Numer. Anal., 18 (1984), pp. 175–182.
[34] N. O. Weiss, in Lectures on Solar and Planetary Dynamos, M. R. E. Proctor and A. D. Gilbert, eds., Cambridge University Press, England, 1994.
[35] K. Zhang and F. Busse, *Convection driven magnetohydrodynamic dynamos in rotating spherical shells*, Geophys. Astrophys. Fluid Dyn., 49 (1989), pp. 97–116.
[36] K. Zhang and G. Schubert, *Magnetohydrodynamics in rapidly rotating spherical systems*, in Annual Review of Fluid Mechanics, Vol. 32, Annual Reviews, Palo Alto, CA, 2000, pp. 411–445.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1903–1920

OPTIMAL CONTROL OF THE STOKES EQUATIONS: A PRIORI ERROR ANALYSIS FOR FINITE ELEMENT DISCRETIZATION WITH POSTPROCESSING*

ARND RÖSCH† AND BORIS VEXLER†

Abstract. An optimal control problem for 2d and 3d Stokes equations is investigated with pointwise control constraints. This paper is concerned with the discretization of the control by piecewise constant functions. The state and the adjoint state are discretized by finite element schemes. In the paper a postprocessing strategy is suggested, which allows for significant improvement of the accuracy.

Key words. PDE-constrained optimization, finite elements, error estimates, Stokes equations, numerical approximation, control constraints.

AMS subject classifications. 49K20, 49M25, 65N30

DOI. 10.1137/050637364

1. Introduction.

The paper is concerned with the discretization of the optimal control problem

$$\text{minimize} \quad J(v,q) = \frac{1}{2}\|v - v_d\|^2_{L^2(\Omega)^d} + \frac{\nu}{2}\|q\|^2_{L^2(\Omega)^d} \tag{1.1}$$

subject to the Stokes equations (state equation)

$$\begin{aligned} -\Delta v + \nabla p &= f + q && \text{in } \Omega, \\ \nabla \cdot v &= 0 && \text{on } \Omega, \\ v &= 0 && \text{on } \Gamma \end{aligned} \tag{1.2}$$

and subject to the control constraints

$$a \le q(x) \le b \qquad \text{for a.a. } x \in \Omega, \tag{1.3}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with $d = 2, 3$ and Γ is the boundary of Ω. The quantities $a, b \in \mathbb{R}^d$ are constant vectors, the inequality (1.3) is understood componentwise and $\nu > 0$ is a given regularization (or control cost) parameter. We denote by $u = (v, p)$ the solution of (1.2). Moreover, we assume for the desired velocity field v_d and the right-hand side f to be from $L^\infty(\Omega)^d$.

The set of admissible controls Q_{ad} is given by

$$Q_{ad} = \{v \in Q := L^2(\Omega)^d : \ a \le q \le b \ \text{ a.e. in } \Omega\}.$$

We discuss here the discretization of the control and state variables by finite elements. The asymptotic behavior of the discretized problem is studied.

First results in the context of a priori error analysis of optimal control problems go back to papers by Falk [15], Geveci [16], and Malanowski [24]. In the past few

*Received by the editors August 2, 2005; accepted for publication (in revised form) April 6, 2006; published electronically September 29, 2006.
http://www.siam.org/journals/sinum/44-5/63736.html
†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (arnd.roesch@oeaw.ac.at, boris.vexler@oeaw.ac.at).

years the theory has been extended to semilinear problems; see Arada, Casas, and Tröltzsch [1] and Casas, Mateos, and Tröltzsch [8]. Error estimates of order h in the L^2-norm and in the L^∞-norm are established in these articles.

Piecewise linear control discretizations for elliptic optimal control problems are studied by Casas and Tröltzsch [9] and Casas [7], containing error estimates of order h and $o(h)$ in the L^2-norm for general cases. For more regular cases an approximation order of $h^{3/2}$ can be proved; see Rösch [27, 28]. An error estimate of order h in the L^∞-norm for an elliptic problem is proved by Meyer and Rösch [26].

However, new discretization concepts have been developed in recent years. The *variational approach* by Hinze [20] and the *superconvergence approach* of Meyer and Rösch [25] can achieve approximation order h^2 in the L^2-norm.

In this paper, we will generalize the superconvergence approach of Meyer and Rösch [25]. The controls are discretized by piecewise constant functions. Clearly, the approximation order of the control cannot be better than h for the optimal control. However, we will show that the point values of the control variable in the barycenter of the elements are approximated with order h^2. Moreover, we will prove that the state and adjoint variable are approximated with order h^2 with respect to the L^2-norm. This allows for a postprocessing step, which leads to h^2 approximation of the control variable in the L^2-norm, too.

Apart from the fact that the Stokes equations have a more complex structure than the equation investigated in [25], this paper contains an essential generalization in the theory. The theory presented in [25] works only for piecewise linear finite elements. The fact that the second derivative of each ansatz function vanishes identically on each triangle is used in a very explicit manner. Consequently, only piecewise linear finite elements defined on triangles can be handled by that technique. We will prove superconvergence results without such restrictions, i.e., only stability and interpolation properties of the elements are requested. Therefore, our results include many different finite element discretization schemes for the 2d and 3d Stokes equations.

To the best of the authors knowledge this is the first paper discussing the discretization error for the optimal control of the Stokes equations with pointwise control constraints. In principle, the classical approach [15, 16] as well as the variational approach [20] can be generalized to the Stokes equations. Of course, several papers are published for the optimal control of the Stokes equations and the Navier–Stokes equations without control constraints; see, e.g., Gunzburger, Hou, and Svobodny [18, 19], Bochev and Gunzburger [4], and Deckelnick and Hinze [14].

Let us remark that the investigated optimal control problems governed by the Stokes equations occur as subproblems in several Newton-type methods for control constrained optimal control problems for the Navier–Stokes equations. The convergence theory of such Newton-type methods requires sufficiently accurate numerical solutions of the subproblems.

The paper is organized as follows: In section 2 a general discretization concept is introduced and the main results are stated. Section 3 contains results from the finite element theory. The proofs of the superconvergence results are placed in section 4. The assumption of the general discretization concept are verified for a specific discretization in section 5. The paper ends with numerical experiments shown in section 6.

2. Discretization and superconvergence results. Throughout this paper, Ω denotes a bounded convex and polygonal domain in $\mathbb{R}^d$, $d = 2, 3$. Moreover, for the case $d = 3$ we assume that edge openings of the domain Ω are smaller than $2\pi/3$.

This will ensure the $W^{1,\infty}$-regularity of the velocity field.

We denote by V and L the Hilbert spaces

$$V := H_0^1(\Omega)^d,$$
$$L := \left\{ p \in L^2(\Omega) : \int_\Omega p(x)\,dx = 0 \right\}.$$

In all that follows, we will omit the subscript L^2 in the norms and inner products if there is no risk of misunderstanding. We look for solutions of the Stokes equations (1.2) in the sense of a weak formulation: the following equation has to be satisfied for arbitrary $\phi = (\psi, \xi) \in V \times L$:

$$a(u,\phi) := (\nabla v, \nabla\psi) - (p, \nabla\cdot\psi) + (\nabla\cdot v, \xi) = (f+q,\psi) =: (F(q),\phi). \tag{2.1}$$

LEMMA 2.1. *Let g be a given function in $L^\infty(\Omega)^d$. Then there exists a unique solution $u = (v,p) \in V \times L$ of*

$$\begin{aligned} -\Delta v + \nabla p &= g \quad && \text{in } \Omega,\\ \nabla\cdot v &= 0 && \text{on } \Omega,\\ v &= 0 && \text{on } \Gamma. \end{aligned} \tag{2.2}$$

Moreover, there exist positive constant c and with $v \in H^2(\Omega)\cap W^{1,\infty}(\Omega)^d$, $p \in H^1(\Omega)$ and

$$\|v\|_{H^2(\Omega)^d} + \|v\|_{W^{1,\infty}(\Omega)^d} + \|p\|_{H^1(\Omega)} \le c\|g\|_{L^\infty(\Omega)^d} \tag{2.3}$$

for $d = 2$ or for $d = 3$ and all edge openings are smaller than $2\pi/3$.

For the proof of this result on polygonal domains especially for $d = 3$, we refer to [13, Theorem 6.3].

We will *assume* that (2.3) is valid for the investigated domain Ω. However, it would be enough for the theory presented here to have this regularity for the optimal adjoint velocity $\bar w$ introduced below.

In order to formulate the optimality system, we introduce the adjoint equation

$$\begin{aligned} -\Delta w - \nabla r &= v - v_d \quad && \text{in } \Omega,\\ \nabla\cdot w &= 0 && \text{on } \Omega,\\ w &= 0 && \text{on } \Gamma. \end{aligned} \tag{2.4}$$

We denote by $z = (w,r) \in V \times L$ the adjoint state. Due to Lemma 2.1 the adjoint velocity w belongs to $H^2(\Omega)^d \cap W^{1,\infty}(\Omega)^d$.

We say that $u = (v,p)$ is the state associated to q if u is the solution of (1.2). Analogously, we call the solution $z = (w,r)$ of (2.4) adjoint state associated to q.

LEMMA 2.2. *There exists a uniquely determined solution $\bar q$ of the optimal control problem* (1.1)–(1.3). *Moreover, a necessary and sufficient condition for the optimality of a control $\bar q$ with associated state $\bar u$ and associated adjoint state $\bar z$, respectively, is that the variational inequality*

$$(\bar w + \nu\bar q, q - \bar q)_Q \ge 0 \quad \text{for all } q \in Q_{ad} \tag{2.5}$$

holds.

The optimal control problem (1.1)–(1.3) is strictly convex and radially unbounded. Hence, there exists a uniquely determined optimal solution and the first order necessary conditions are also sufficient for optimality. Such basic results and an introduction in optimal control theory governed by partial differential equations can be

found for instance in Lions [23]. We remark that the variational inequality (2.5) can be equivalently formulated; see, e.g., Malanowski [24], as

$$\bar{q} = \Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{w}\Big), \tag{2.6}$$

where the projection Π is defined by

$$\Pi_{[a,b]}(f(x)) = \max(a, \min(b, f(x))).$$

Again, all functions are defined componentwise.

In order to discretize the optimal control problem, we consider a 2- or 3-d mesh $\mathcal{T}_h$ consisting of open cells T, which constitute a nonoverlapping covering of the domain Ω. The cells are either triangles, tetrahedra, quadrilaterals or hexahedra. The mesh parameter h is defined as a cellwise constant function by setting $h_{|T} = h_T$ and h_T is the diameter of K. Usually we use the symbol h also for the maximal cell size, i.e.,

$$h = \max_{T\in\mathcal{T}_h} h_T.$$

The straight parts which make up the boundary ∂T of a cell T are called *faces*. For the mesh $\mathcal{T}_h$ we require to be regular in the following sense; see, e.g., [10], i.e.:

(A1)
- $\bar{\Omega} = \cup_{T\in\mathcal{T}_h} \bar{T}$,
- $T_1 \cap T_2 = \emptyset$ or $T_1 = T_2$, for all $T_1, T_2 \in \mathcal{T}_h$,
- Any face of any cell $T_1 \in \mathcal{T}_h$ is either a subset of the boundary $\partial\Omega$, or a face of another cell $T_2 \in \mathcal{T}_h$.

The control variable q is discretized by piecewise constant elements on the mesh $\mathcal{T}_h$ using the following discrete space:

$$Q_h := \{q_h \in L^2(\Omega)^d : \, q_h|_T \in (\mathcal{P}_0)^d \quad \text{for all} \quad T \in \mathcal{T}_h\}.$$

Here $\mathcal{P}_k$ denotes the polynoms with degree less than or equal to k.

Next, we introduce a general conforming finite element setting for the discretization of the state equation. Let $V_h \subset V$ and $L_h \subset L$ be finite dimensional subspaces with the following properties:

(A2) The space V_h and the mesh $\mathcal{T}_h$ fit in the following sense: Every function $v_h \in V_h$ is piecewise polynomial on $\mathcal{T}_h$

$$v_h|_T \in \mathcal{P}^d \quad \text{for all} \quad T \in \mathcal{T}_h,$$

where $\mathcal{P}$ is a polynomial space.

For a given control $q \in Q$, the state equation (2.1) is discretized using the spaces V_h and L_h as follows: Find $u_h = (v_h, p_h)$ such that

$$a(u_h, \phi_h) + s_h(p_h, \xi_h) = (F(q), \phi_h) \quad \text{for all } \phi_h = (\psi_h, \xi_h) \in V_h \times L_h. \tag{2.7}$$

Here, the term $s_h(\cdot,\cdot)$ denotes a stabilization (continuous, symmetric) bilinear form on $L_h \times L_h$. Such stabilization terms are needed if, e.g., finite elements of equal order for the velocities and pressure are used; see [2] or [6]. For this discretization we require the following conditions: We introduce the space of cellwise H^2 functions

$$V_h^2(\Omega)^d := \{v \in V : \, v|_T \in H^2(T)^d \text{ for all } T \in \mathcal{T}_h\},$$

with a discrete H^2 norm defined by

$$\|\psi_h\|'_{H^2(\Omega)^d} := \left(\|\psi_h\|^2_{H^1(\Omega)^d} + \sum_{T\in\Omega} \|\nabla^2 \psi_h\|^2_{L^2(T)^d} \right)^{1/2}.$$

(A3) There exist interpolation operators $i_h^v : H^2(\Omega)^d \cap V \to V_h$ and $i_h^p : H^1(\Omega) \cap L \to L_h$ with the following approximation properties:

$$\begin{aligned} \|v - i_h^v v\|_{L^2(\Omega)^d} + h\|\nabla(v - i_h^v v)\|_{L^2(\Omega)^d} &\le c_v h^2 \, \|\nabla^2 v\|_{L^2(\Omega)^{d\times d}}, \\ \|v - i_h^v v\|_{L^\infty(\Omega)^d} + h^{2-d/2}\|v - i_h^v v\|'_{H^2(\Omega)^d} &\le c_v h^{2-d/2} \, \|\nabla^2 v\|_{L^2(\Omega)^{d\times d}}, \\ \|p - i_h^p p\|_{L^2(\Omega)} &\le c_p h \, \|\nabla p\|_{L^2(\Omega)}. \end{aligned}$$

For the existence of operators i_h^p we refer to Clément [11].

(A4) There exists a finite dimensional space $\tilde{L}_h \subset L_h$ and a continuous projection operator $\pi\colon L_h \to \tilde{L}_h$ such that

- For the pair $(V_h, \tilde{L}_h)$ the inf-sup condition holds, i.e., there exists a positive constant γ independent of h with

$$\sup_{\phi_h \in V_h} \frac{(p_h, \nabla\cdot\phi_h)}{\|\nabla\phi_h\|_{L^2(\Omega)^d}} \ge \gamma \|p_h\|_{L^2(\Omega)} \quad \text{for all} \quad p_h \in \tilde{L}_h.$$

- There is a positive constant c independent of h such that

$$\|\pi p_h\|_{L^2(\Omega)} \le c\|p_h\|_{L^2(\Omega)} \quad \text{for all} \quad p_h \in L_h.$$

- There is a positive constant c independent of h such that

$$\|p_h - \pi p_h\|^2_{L^2(\Omega)} \le c\, s_h(p_h, p_h) \quad \text{for all} \quad p_h \in L_h.$$

- There is a positive constant c independent of h such that

$$s_h(i_h^p p, i_h^p p) \le ch^2 \|\nabla p\|_{L^2(\Omega)^d} \quad \text{for all} \quad p \in L \cap H^1(\Omega).$$

Remark 2.3. If the inf-sup condition is fulfilled for the pair (V_h, L_h) itself, there is no need for stabilization and we can set $s(p,\xi) \equiv 0$ and $\pi = id_{L_h}$.

Remark 2.4. In the presence of the regularization term $s_h(p_h, \xi_h)$, the discretization (2.7) is not a pure Galerkin scheme for (2.1) any more. Therefore, the question arises, if the approaches "discretize-then-optimize" and "optimize-then-discretize" coincide; see the discussion in Collis and Heinkenschloss [12]. In our setting these two approaches coincide due to the fact, that $s_h(\cdot,\cdot)$ is a symmetric bilinear form.

Moreover, we require the following inverse inequalities:

(A5) There is a positive constant c independent of h such that for all $v_h \in V_h$ holds:

$$\|v_h\|'_{H^2(\Omega)^d} \le ch^{-1}\|v_h\|_{H^1(\Omega)^d} \quad \text{and} \quad \|v_h\|_{L^\infty(\Omega)^d} \le ch^{-d/2}\|v_h\|_{L^2(\Omega)^d}.$$

Let T be an arbitrary element of the mesh $\mathcal{T}_h$. We define an operator $R_h : C(\Omega) \to Q_h$ by

$$(R_h g)(x) = g(S_T),$$

where S_T denotes the barycenter of the element T. The operator R_h is defined componentwise in the case of a vector valued function.

(A6) Let $T \in \mathcal{T}_h$ be an arbitrary element of the discretization and $g \in H^2(T)$ an arbitrary function. We require the following estimates:

$$\left| \int_T g(x) - (R_h g)(x)\, dx \right| \leq ch^2 |T|^{1/2} \|\nabla^2 g\|_{L^2(T)},$$
$$\|g - R_h g\|_{L^\infty(\Omega)} \leq ch\, \|\nabla g\|_{L^\infty(\Omega)}$$

with a positive constant c independent of h.

Remark 2.5. Assumptions (A1)–(A6) are standard properties of finite element discretizations. They are fulfilled for many different conforming element pairs with and without stabilization. We will verify these conditions for one specific discretization in section 5.

For our superconvergence result an additional assumption is needed. It follows from Lemma 2.1, that the optimal adjoint velocity $\bar{w}$ belongs to $H^2(\Omega)^d \cap W^{1,\infty}(\Omega)^d$. However, the regularity of the optimal control is weaker because of the occurrence of kinks caused by the max-function in (2.6). Nevertheless, we can group all elements $T \in \mathcal{T}_h$ into two classes:

$$K_1 := \bigcup_{T \in \mathcal{T}_h,\, \bar{q} \notin H^2(T)^d} T, \qquad K_2 := \bigcup_{T \in \mathcal{T}_h,\, \bar{q} \in H^2(T)^d} T. \tag{2.8}$$

We remark that the properties of the projection operator and Lemma 2.1 imply $\bar{q} \in W^{1,\infty}(\Omega)^d$,

(A7)

$$|K_1| \leq ch.$$

Assumption (A7) is difficult to verify, but is valid in many practical cases.

The discrete optimization problem is given by the minimization of the cost functional (1.1) subject to the discretized state equation (2.7) and subject to $q_h \in Q_h^{ad} = Q_h \cap Q_{ad}$. Similar to the notation for the continuous problem, we denote by $\bar{q}_h \in Q_h^{ad}$, $\bar{u}_h \in V_h \times L_h$, and $\bar{z}_h \in V_h \times L_h$ the optimal control, the associated state, and the adjoint state of the discretized optimal control problem. In the following theorems we formulate our main results.

THEOREM 2.6. *Assume that* (A1)–(A7) *holds. Then the estimate*

$$\|\bar{q}_h - R_h \bar{q}\|_Q \leq ch^2 (\|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d} + \|\bar{q}\|_{L^\infty(\Omega)^d}) \tag{2.9}$$

is valid with a positive constant c independent of h.

THEOREM 2.7. *The estimates*

$$\|\bar{v} - \bar{v}_h\|_Q \leq ch^2 (\|\bar{q}\|_{L^\infty(\Omega)^d} + \|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d}) \tag{2.10}$$

$$\|\bar{w} - \bar{w}_h\|_Q \leq ch^2 (\|\bar{q}\|_{L^\infty(\Omega)^d} + \|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d}) \tag{2.11}$$

are valid provided that (A1)–(A7) *hold.*

THEOREM 2.8. *Assume that* (A1)–(A7) *holds. Then the estimate*

$$\|\tilde{q}_h - \bar{q}\|_Q \leq ch^2 (\|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d} + \|\bar{q}\|_{L^\infty(\Omega)^d}) \tag{2.12}$$

is valid with

$$\tilde{q}_h = \Pi_{[a,b]} \left(-\frac{1}{\nu} \bar{w}_h \right) \tag{2.13}$$

and a positive constant c independent of h.

The proofs of the Theorems 2.6 and 2.8 are contained in section 4.

Let us briefly explain why these are superconvergence results: The best possible rate for approximation of the optimal solution by a piecewise constant function is h. Therefore we can only expect

$$\|\bar{q} - \bar{q}_h\|_Q = \mathcal{O}(h).$$

However, we will show in Theorem 2.6 that the values in the barycenter are approximated with order h^2. A direct implication of this result will be that the velocity and the adjoint velocity is approximated with order h^2 in the L^2-norm. The projection in (2.13) increases the accuracy of the calculated control to order h^2. Hence, the result of Theorem 2.8 provides a possibility to significantly improve the behavior of the error by a simple postprocessing step (2.13).

Remark 2.9. Theorem 2.7 provides error bounds for the optimal velocity and the adjoint velocity in L^2-norm. As a direct consequence one obtains the corresponding estimates for the velocity with respect to H^1-norm and for the pressure with respect to L^2-norm of order $\mathcal{O}(h)$.

Remark 2.10. Assumption (A7) is essential for quadratic approximation results, i.e., for $\|q - \tilde{q}_h\|_Q = \mathcal{O}(h^2)$. In the absence of this assumption one can obtain by classical techniques:

$$\|\bar{q} - \bar{q}_h\| = \mathcal{O}(h) \quad \text{and} \quad \|\bar{q} - \tilde{q}_h\| = \mathcal{O}(h).$$

3. Results from finite element theory. In this section, we collect results from finite element theory. We define (linear) solution mappings S and S^p of the continuous state equation such that there holds for all $\phi = (\psi, \xi) \in V \times L$, $g \in Q$ and $v^g = S(g)$, $p^g = S^p(g)$:

$$u^g = (v^g, p^g) \in V \times L \,:\, a(u^g, \phi) = (g, \psi). \tag{3.1}$$

In a similar way, we define the solution mappings S_h and S_h^p of the discretized state equation such that there holds for all $\phi_h = (\psi_h, \xi_h) \in V_h \times L_h$, $g \in Q$ and $v_h^g = S_h(g)$, $p_h^g = S_h^p(g)$:

$$u_h^g = (v_h^g, p_h^g) \in V_h \times L_h \,:\, a(u_h^g, \phi_h) + s_h(p_h^g, \xi_h) = (g, \psi_h). \tag{3.2}$$

Although the solution operators S and S_h have better regularity properties, it is more convenient (in particular for section 4) to consider them in the space Q:

$$S\colon Q \to Q, \quad S_h\colon Q \to Q.$$

In the following we provide some properties of these operators based on the assumptions (A1)–(A7). The following lemma ensures the stability of the discretization of the state equation.

LEMMA 3.1. *Under assumptions* (A1)–(A6) *the following modified inf-sup condition holds: There exist positive constants $\tilde{\gamma}$ and c independent of h with*

$$\sup_{\phi_h \in V_h} \frac{(p_h, \nabla \cdot \phi_h) + c\, s(p_h, p_h)}{\|\nabla \phi_h\|_{L^2(\Omega)^d}} \geq \tilde{\gamma} \|p_h\|_{L^2(\Omega)} \quad \textit{for all } p_h \in L_h.$$

Proof. For the proof we refer to [2]. ∎

Next, we define the affine linear operators $P : Q \to Q$ and $P_h : Q \to Q$ by

$$Pq = S^*(S(q+f) - v_d), \quad P_h q = S_h^*(S_h(q+f) - v_d),$$

where S^* and S_h^* denote the adjoint operators of S and S_h, respectively.

LEMMA 3.2. *Assume that the assumption* (A1)–(A6) *hold. Let $q \in Q$ be an arbitrary control. Then, the discretization error of the state equation and the adjoint equation can be estimated by*

(i) $h\|S_h(q+f) - S(q+f)\|_{H^1(\Omega)^d} + \|S_h(q+f) - S(q+f)\|_Q \le ch^2\,(\|q\|_Q + \|f\|_Q)$,
(ii) $\|S_h(q+f) - S(q+f)\|_{L^\infty(\Omega)^d} \le ch^{2-d/2}\,(\|q\|_Q + \|f\|_Q)$,
(iii) $\|S_h(q+f) - S(q+f)\|'_{H^2(\Omega)^d} \le c\,(\|q\|_Q + \|f\|_Q)$,
(iv) $\|P_h q - Pq\|_Q \le ch^2\,(\|q\|_Q + \|f\|_Q + \|v_d\|_Q)$.

Proof. The proof of the error estimate (i) relies on Lemma 3.1 and is given in [2]. The result concerning L^2-estimate can be obtained by standard techniques; see, e.g., [17] for the application of the Aubin–Nitsche trick to the Stokes problem.

For the proof of (ii) we set $g = f + q$ and use the second inverse inequality from (A5) and an interpolation estimate from (A3):

$$\begin{aligned}
\|(S - S_h)(g)\|_{L^\infty(\Omega)^d} &\le \|S(g) - i_h^v S(g)\|_{L^\infty(\Omega)^d} + \|S_h(g) - i_h^v S(g)\|_{L^\infty(\Omega)^d} \\
&\le ch^{2-d/2}\,\|\nabla^2 S(g)\|_{L^2(\Omega)^d} + ch^{-d/2}\|S_h(g) - i_h^v S(g)\|_{L^2(\Omega)^d} \\
&\le ch^{2-d/2}\,\|\nabla^2 S(g)\|_{L^2(\Omega)^d} \\
&\quad + ch^{-d/2}\big\{\|S(g) - i_h^v S(g)\|_{L^2(\Omega)^d} + \|S_h(g) - S(g)\|_{L^2(\Omega)^d}\big\} \\
&\le ch^{2-d/2}\,\|\nabla^2 S(g)\|_{L^2(\Omega)^d} \le ch^{2-d/2}\,\|g\|_{L^2(\Omega)^d}.
\end{aligned}$$

The estimate (iii) follows in the same manner using the first inverse inequality from (A5) and an interpolation estimate from (A3).

The error estimate (iv) is obtained in a similar way as (i). ∎

LEMMA 3.3. *The discretization operators S_h and S_h^* are bounded in the following sense:*

(i) $\|S_h\|_{Q\to V} \le c$, $\|S_h^*\|_{Q\to V} \le c$,
(ii) $\|S_h\|_{Q\to L^\infty(\Omega)^d} \le c$, $\|S_h^*\|_{Q\to L^\infty(\Omega)^d} \le c$,
(iii) $\|S_h\|_{Q\to V_h^2(\Omega)^d} \le c$, $\|S_h^*\|_{Q\to V_h^2(\Omega)^d} \le c$.

Proof. We sketch only the proof for the operator S_h. The results for the adjoint operator S_h^* can be derived by the same techniques. In order to prove the first estimate we set $\phi_h = (S_h(g), S_h^p(g))$ in (3.2) and obtain

$$\|\nabla S_h(g)\|^2_{L^2(\Omega)^d} + s_h(S_h^p(g), S_h^p(g)) = (g, S_h(g)).$$

Due to (A4) we have

$$\|\nabla S_h(g)\|^2_{L^2(\Omega)^d} \le (g, S_h(g)).$$

The assertion follows then by Poincaré inequality.

The second estimate is obtained using (ii) from Lemma 3.2:

$$\|S_h(g)\|_{L^\infty(\Omega)^d} \le \|S(g) - S_h(g)\|_{L^\infty(\Omega)^d} + \|S(g)\|_{L^\infty(\Omega)^d} \le c(1+h)\|g\|_{L^2(\Omega)}.$$

The third estimate is obtained similarly using (iii) from Lemma 3.2. ∎

LEMMA 3.4. *Let the conditions of Lemma* 2.1 *be fulfilled, i.e., in particular,* $w \in H^2(\Omega)^d \cap W^{1,\infty}(\Omega)^d$. *Then the inequality*

$$(\psi_h, \bar{q} - R_h\bar{q}) \le ch^2(\|\psi_h\|_{L^\infty(\Omega)^d} + \|\psi_h\|'_{H^2(\Omega)^d})(\|\bar{q}\|_{L^\infty(\Omega)^d} + \|\bar{w}\|_{W^{2,p'}(\Omega)^d})$$

is satisfied for all $\psi_h \in V_h$ *provided that the assumptions* (A1)–(A7) *are fulfilled.*

Proof. With the sets K_1 and K_2 introduced by (2.8), we obtain

$$(\psi_h, \bar{q} - R_h\bar{q}) = \int_{K_1} \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx + \int_{K_2} \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx. \tag{3.3}$$

Using the $W^{1,\infty}$-regularity of $\bar{q}$, the K_1-part can be estimated as follows:

$$\begin{aligned}\int_{K_1} \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx &= \sum_{T\subset K_1} \int_T \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx \\ &\le \|\psi_h\|_{L^\infty(\Omega)^d} \sum_{T\subset K_1} \|\bar{q} - R_h\bar{q}\|_{L^\infty(T)^d} \int_T dx \\ &\le ch\|\psi_h\|_{L^\infty(\Omega)^d} |\bar{q}|_{W^{1,\infty}(\Omega)^d} |K_1|. \end{aligned} \tag{3.4}$$

Assumption (A7) and the properties of the projection (2.6) yield

$$\begin{aligned}\int_{K_1} \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx &\le ch^2\|\psi_h\|_{L^\infty(\Omega)^d} \|\bar{w}\|_{W^{1,\infty}(\Omega)^d} \\ &\le ch^2\|\psi_h\|_{L^\infty(\Omega)^d} \|\bar{w}\|_{W^{2,p'}(\Omega)^d}. \end{aligned} \tag{3.5}$$

On the K_2-part, we proceed as follows:

$$\begin{aligned}\left|\int_{K_2} \psi_h \cdot (\bar{q} - R_h\bar{q})\,dx\right| &\le \left|\int_{K_2} (\psi_h \cdot R_h\bar{q} - R_h(\psi_h \cdot \bar{q}))\,dx\right| \\ &\quad + \left|\int_{K_2} \psi_h \cdot \bar{q} - R_h(\psi_h \cdot \bar{q})\,dx\right|. \end{aligned} \tag{3.6}$$

Using $R_h(\psi_h \cdot \bar{q}) = R_h\psi_h \cdot R_h\bar{q}$, we find for the first integral

$$\left|\int_{K_2} (\psi_h \cdot R_h\bar{q} - R_h(\psi_h \cdot \bar{q}))\,dx\right| \le \sum_{T\subset K_2} \left|\int_T (\psi_h - R_h\psi_h) \cdot R_h\bar{q}\,dx\right|.$$

Note, that $R_h\bar{q}$ is constant on every element T. Hence, we can continue with

$$\left|\int_{K_2} (\psi_h \cdot R_h\bar{q} - R_h(\psi_h \cdot \bar{q}))\,dx\right| \le \sum_{T\subset K_2} \left|R_h\bar{q} \cdot \int_T (\psi_h - R_h\psi_h)\,dx\right|.$$

Consequently, we find by means of (A6)

$$\begin{aligned}\left|\int_{K_2} (\psi_h \cdot R_h\bar{q} - R_h(\psi_h \cdot \bar{q}))\,dx\right| &\le ch^2 \sum_{T\subset K_2} |T|^{1/2}\|\bar{q}\|_{L^\infty(T)^d} |\psi_h|_{H^2(T)^d} \\ &\le c|\Omega|^{1/2}h^2\|\bar{q}\|_{L^\infty(\Omega)^d} \|\psi_h\|'_{H^2(\Omega)^d} \\ &\le ch^2\|\bar{w}\|_{L^\infty(\Omega)^d} \|\psi_h\|'_{H^2(\Omega)^d}. \end{aligned} \tag{3.7}$$

It remains the second integral in (3.6). Again, we can use (A6):

$$\left|\int_{K_2} \psi_h\cdot\bar q - R_h(\psi_h\cdot\bar q)\,dx\right| \le \sum_{T\subset K_2}\left|\int_T \psi_h\cdot\bar q - R_h(\psi_h\cdot\bar q)\,dx\right| \le ch^2\sum_{T\subset K_2}|T|^{1/2}|\psi_h\cdot\bar q|_{H^2(T)}. \tag{3.8}$$

We will estimate this seminorm by

$$\begin{aligned}|\psi_h\cdot\bar q|_{H^2(T)} &\le \|\psi_h\|_{L^\infty(T)^d}|\bar q|_{H^2(T)^d} + |\psi_h|_{H^2(T)^d}\|\bar q\|_{L^\infty(T)^d}\\ &\quad + 2|\psi_h|_{H^1(T)^d}|\bar q|_{W^{1,\infty}(T)^d}.\end{aligned} \tag{3.9}$$

The projection formula (2.6) and the fact that $\bar q$ smooth is on every $T\subset K_2$ imply

$$\begin{aligned}|\psi_h\cdot\bar q|_{H^2(T)} &\le c(\|\psi_h\|_{L^\infty(T)^d}|\bar w|_{H^2(T)^d}+|\psi_h|_{H^2(T)^d}\|\bar q\|_{L^\infty(T)^d}+|\psi_h|_{H^1(T)^d}|\bar w|_{W^{1,\infty}(T)^d})\\ &\le c(\|\psi_h\|_{L^\infty(\Omega)^d}|\bar w|_{H^2(T)^d}+|\psi_h|_{H^2(T)^d}\|\bar q\|_{L^\infty(\Omega)^d}+|\psi_h|_{H^1(T)^d}|\bar w|_{W^{1,\infty}(\Omega)^d})\end{aligned} \tag{3.10}$$

Combining (3.8) with (3.10) we find

$$\begin{aligned}\left|\int_{K_2} \psi_h\cdot\bar q - R_h(\psi_h\cdot\bar q)\,dx\right| &\le ch^2\sum_{T\subset K_2}|T|^{1/2}(\|\psi_h\|_{L^\infty(\Omega)^d}|\bar w|_{H^2(T)^d}\\ &\quad + |\psi_h|_{H^2(T)^d}\|\bar q\|_{L^\infty(\Omega)^d} + |\psi_h|_{H^1(T)^d}|\bar w|_{W^{1,\infty}(\Omega)^d})\\ &\le ch^2(\|\psi_h\|_{L^\infty(\Omega)^d}\|\bar w\|_{H^2(\Omega)^d}\\ &\quad + \|\psi_h\|'_{H^2(\Omega)^d}\|\bar q\|_{L^\infty(\Omega)^d} + \|\psi_h\|_{H^1(\Omega)^d}\|\bar w\|_{W^{1,\infty}(\Omega)^d}).\end{aligned}$$

By imbedding arguments we end up with

$$\begin{aligned}\left|\int_{K_2} \psi_h\cdot\bar q - R_h(\psi_h\cdot\bar q)\,dx\right| &\le ch^2(\|\psi_h\|_{L^\infty(\Omega)^d}\\ &\quad + \|\psi_h\|'_{H^2(\Omega)^d})(\|\bar q\|_{L^\infty(\Omega)^d} + \|\bar w\|_{W^{2,p'}(\Omega)^d})\end{aligned} \tag{3.11}$$

Inserting (3.7) and (3.11) into (3.6), we obtain

$$\begin{aligned}\left|\int_{K_2} (\psi_h\cdot R_h\bar q - R_h(\psi_h\cdot\bar q))\,dx\right| &\le ch^2(\|\psi_h\|_{L^\infty(\Omega)^d} + \|\psi_h\|'_{H^2(\Omega)^d})\\ &\quad (\|\bar q\|_{L^\infty(\Omega)^d} + \|\bar w\|_{W^{2,p'}(\Omega)^d}).\end{aligned} \tag{3.12}$$

From (3.3), (3.5), and (3.12), the assertion follows immediately. ∎

LEMMA 3.5. *Let $p'>d$ the regularity parameter of Lemma* 2.1 *and* (A1)–(A7) *be fulfilled. Then the estimates*

$$\|S_h(\bar q+f) - S_h(R_h\bar q+f)\|_Q \le ch^2(\|\bar q\|_{L^\infty(\Omega)^d} + \|\bar w\|_{W^{2,p'}(\Omega)^d}) \tag{3.13}$$

$$\|P_h\bar q - P_hR_h\bar q\|_Q \le ch^2(\|\bar q\|_{L^\infty(\Omega)^d} + \|\bar w\|_{W^{2,p'}(\Omega)^d}) \tag{3.14}$$

are valid.

Proof. We start with

$$\begin{aligned}\|S_h(\bar q+f) - S_h(R_h\bar q+f)\|_Q^2 &= (S_h(\bar q+f) - S_h(R_h\bar q+f), S_h(\bar q+f) - S_h(R_h\bar q+f))_Q\\ &= (S_h(\bar q - R_h\bar q), (S_h(\bar q+f) - v_d) - (S_h(R_h\bar q+f) - v_d))_Q\\ &= (\bar q - R_h\bar q, P_h\bar q - P_hR_h\bar q)_Q\\ &\le ch^2(\|P_h\bar q - P_hR_h\bar q\|_{L^\infty(\Omega)^d} + \|P_h\bar q - P_hR_h\bar q\|'_{H^2(\Omega)^d})\\ &\quad \cdot(\|\bar q\|_{L^\infty(\Omega)^d} + \|\bar w\|_{W^{2,p'}(\Omega)^d}),\end{aligned} \tag{3.15}$$

where we have used Lemma 3.4 with $\psi_h = P_h\bar{q} - P_hR_h\bar{q}$. We benefit now from the fact that $P_h\bar{q}$ and $P_hR_h\bar{q}$ are solutions of the discretized adjoint equation, that means $\psi_h = P_h\bar{q} - P_hR_h\bar{q} = S_h^*(S_h(\bar{q}+f) - v_d) - S_h^*(S_h(R_h\bar{q}+f) - v_d) = S_h^*(S_h(\bar{q}+f) - S_h(R_h\bar{q}+f))$. Therefore we obtain by Lemma 3.3

$$\|P_h\bar{q} - P_hR_h\bar{q}\|_{L^\infty(\Omega)^d} \le c\|S_h(\bar{q}+f) - S_h(R_h\bar{q}+f)\|_Q \tag{3.16}$$

and

$$\|P_h\bar{q} - P_hR_h\bar{q}\|'_{H^2(\Omega)^d} \le c\|S_h(\bar{q}+f) - S_h(R_h\bar{q}+f)\|_Q. \tag{3.17}$$

Inserting (3.16) and (3.17) in (3.15) and dividing by $\|S_h(\bar{q}+f) - S_h(R_h\bar{q}+f)\|_Q$, the assertion (3.13) is obtained. Inequality (3.13) and the continuity of S_h in Q yield (3.14). ∎

COROLLARY 3.6. *Let $p' > d$ the regularity parameter of Lemma* 2.1 *and* (A1)–(A7) *be fulfilled. Then the inequality*

$$\|\bar{w} - P_hR_h\bar{q}\|_Q \le ch^2(\|\bar{q}\|_{L^\infty(\Omega)^d} + \|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d} + \|\bar{w}\|_{W^{2,p'}(\Omega)^d}) \tag{3.18}$$

is valid.

Proof. We apply Lemma 3.2 for $q = \bar{q}$. Using $\bar{w} = P\bar{q}$, we obtain

$$\|\bar{w} - P_h\bar{q}\|_Q \le ch^2(\|\bar{q}\|_{L^\infty(\Omega)^d} + \|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d})$$

The assertion follows now from (3.14) and the triangle inequality. ∎

4. Superconvergence properties. In this section, we prove the main results stated in section 2. We start with an auxiliary result.

LEMMA 4.1. *The inequality*

$$\nu\|R_h\bar{q} - \bar{q}_h\|_Q^2 \le (R_h\bar{w} - \bar{w}_h, \bar{q}_h - R_h\bar{q})_Q \tag{4.1}$$

is valid provided that the assumptions (A1)–(A7) *hold.*

Proof. First, we recall the optimality condition (2.5):

$$(\bar{w} + \nu\bar{q}, q - \bar{q})_Q \ge 0 \quad \text{for all } q \in Q_{ad}.$$

This formula holds also pointwise a.e. in Ω:

$$(\bar{w}(x) + \nu\bar{q}(x)) \cdot (q(x) - \bar{q}(x)) \ge 0 \quad \text{for all } q \in Q_{ad}.$$

Consider any element T with center of gravity S_T and apply this formula for $x = S_T$ and $q = \bar{q}_h$. This can be done because of the continuity of the functions $\bar{w}$, $\bar{q}$, and $\bar{q}_h$ in this point:

$$(\bar{w}(S_T) + \nu\bar{q}(S_T)) \cdot (\bar{q}_h(S_T) - \bar{q}(S_T)) \ge 0 \quad \text{for all } T \in \mathcal{T}_h.$$

Due to the definition of R_h, this is equivalent to

$$(R_h\bar{w}(S_T) + \nu R_h\bar{q}(S_T)) \cdot (\bar{q}_h(S_T) - R_h\bar{q}(S_T)) \ge 0 \quad \text{for all } T \in \mathcal{T}_h.$$

We integrate this formula over T, add over all T, and get

$$(R_h\bar{w} + \nu R_h\bar{q}, \bar{q}_h - R_h\bar{q})_Q \ge 0. \tag{4.2}$$

Otherwise, the optimal control $\bar{q}_h$ of the discretized problem fulfills the optimality condition

$$(\bar{w}_h + \nu\bar{q}_h, q - \bar{q}_h)_Q \geq 0 \quad \text{for all } q \in Q_h^{ad}. \tag{4.3}$$

We apply this formula for $q = R_h\bar{q}$:

$$(\bar{w}_h + \nu\bar{q}_h, R_h\bar{q} - \bar{q}_h)_Q \geq 0. \tag{4.4}$$

Adding (4.2) and (4.4), we obtain

$$(R_h\bar{w} - \bar{w}_h + \nu(R_h\bar{q} - \bar{q}_h), \bar{q}_h - R_h\bar{q})_Q \geq 0 \tag{4.5}$$

This completes the proof. □

Remark 4.2. Lemma 4.1 is the key to prove our main results. The presented technique benefits from the fact that the controls are discretized by piecewise constant functions. The derivation of the estimate (4.1) motivates our choice for the control discretization.

Now we are able to prove Theorem 2.6.

Proof of Theorem 2.6. We begin by rewriting formula (4.1):

$$\begin{aligned}\nu\|R_h\bar{q} - \bar{q}_h\|_Q^2 &\leq (R_h\bar{w} - \bar{w}_h, \bar{q}_h - R_h\bar{q})_Q \\ &= (R_h\bar{w} - \bar{w}, \bar{q}_h - R_h\bar{q})_Q + (\bar{w} - P_hR_h\bar{q}, \bar{q}_h - R_h\bar{q})_Q \\ &\quad + (P_hR_h\bar{q} - \bar{w}_h, \bar{q}_h - R_h\bar{q})_Q. \end{aligned} \tag{4.6}$$

Let us now estimate these three terms. We start with the first term using (A6) and the fact that $\bar{q}_h - R_h\bar{q}$ is piecewise constant on each element,

$$\begin{aligned}(R_h\bar{w} - \bar{w}, \bar{q}_h - R_h\bar{q})_Q &= \sum_{T\in\mathcal{T}_h}\int_T (R_h\bar{w}(x) - \bar{w}(x)) \cdot (\bar{q}_h(x) - R_h\bar{q}(x))\,dx \\ &= \sum_{T\in\mathcal{T}_h} (\bar{q}_h(S_T) - \bar{q}(S_T)) \cdot \int_T (\bar{w}(S_T) - \bar{w}(x))\,dx \\ &\leq \sum_{T\in\mathcal{T}_h} ch^2|\bar{q}_h(S_T) - \bar{q}(S_T)||T|^{1/2}|\bar{w}|_{H^2(T)^d} \\ &\leq ch^2\|\bar{q}_h - R_h\bar{q}\|_Q\|\bar{w}\|_{W^{2,p'}(\Omega)^d}. \end{aligned} \tag{4.7}$$

The second term in (4.6) is estimated by Corollary 3.6 and the Cauchy–Schwartz inequality:

$$\begin{aligned}(\bar{w} - P_hR_h\bar{q}, \bar{q}_h - R_h\bar{q})_Q \leq ch^2(&\|\bar{q}\|_{L^\infty(\Omega)^d} + \|v_d\|_{L^\infty(\Omega)^d} + \|f\|_{L^\infty(\Omega)^d} \\ &+ \|\bar{w}\|_{W^{2,p'}(\Omega)^d})\|\bar{q}_h - R_h\bar{q}\|_Q. \end{aligned} \tag{4.8}$$

The third term can be omitted because of

$$\begin{aligned}(P_hR_h\bar{q} - \bar{w}_h, \bar{q}_h - R_h\bar{q})_Q &= (P_hR_h\bar{q} - P_h\bar{q}_h, \bar{q}_h - R_h\bar{q})_Q \\ &= (S_h^*(S_h(R_h\bar{q} + f) - v_d) - S_h^*(S_h(\bar{q}_h + f) - v_d), \bar{q}_h - R_h\bar{q})_Q \\ &= (S_h^*S_h(R_h\bar{q} - \bar{q}_h), \bar{q}_h - R_h\bar{q})_Q \\ &= (S_h(R_h\bar{q} - \bar{q}_h), S_h(\bar{q}_h - R_h\bar{q}))_Q \\ &\leq 0. \end{aligned} \tag{4.9}$$

Inserting (4.7)–(4.9) into (4.6), we end up with

$$\nu\|R_h\bar{q}-\bar{q}_h\|_Q^2 \le ch^2(\|\bar{q}\|_{L^\infty(\Omega)^d}+\|v_d\|_{L^\infty(\Omega)^d}+\|f\|_{L^\infty(\Omega)^d} +\|\bar{w}\|_{W^{2,p'}(\Omega)^d})\|\bar{q}_h-R_h\bar{q}\|_Q. \tag{4.10}$$

This inequality is equivalent to the assertion (2.9). ∎

Proof of Theorem 2.7. Using the triangle inequality, we find

$$\begin{aligned}\|\bar{v}-\bar{v}_h\|_Q &= \|S(\bar{q}+f)-S_h(\bar{q}_h+f)\|_Q\\ &\le \|S(\bar{q}+f)-S_h(\bar{q}+f)\|_Q+\|S_h(\bar{q}+f)-S_h(R_h\bar{q}+f)\|_Q+\|S_h(R_h\bar{q}-\bar{q}_h)\|_Q.\end{aligned}$$

The first term is estimated using Lemma 3.2, for the second term we use the assertion from Lemma 3.5, and for the third term we apply Theorem 2.6 and the boundedness of S_h. This yields estimate (2.10).

Inequality (2.11) can be similarly obtained by Corollary 3.6, Theorem 2.6, and the boundedness of S_h and S_h^* in $\mathcal{L}(Q)$. ∎

Next, we prove Theorem 2.8.

Proof of Theorem 2.8. We use the Lipschitz continuity of the projection operator and find

$$\begin{aligned}\|\tilde{q}_h-\bar{q}\|_Q &= \left\|\Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{w}_h\Big)-\Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{w}\Big)\right\|_Q\\ &\le \frac{1}{\nu}\|\bar{w}_h-\bar{w}\|_Q.\end{aligned}$$

Inequality (2.11) now implies the assertion. ∎

5. Verification of the assumptions for concrete numerical schemes. In this section we check the assumptions (A1)–(A6) for some discretization schemes. Let $\mathcal{T}_h$ be a shape regular quasi-uniform mesh (see, e.g., Braess [5]) consisting of triangles or quadrilaterals for $d=2$ or tetrahedrons or hexahedrons for $d=3$. Then, the assumption (A1) is automatically fulfilled. If the control is defined on the same mesh, then assumption (A2) is fulfilled, too.

Let P_h^k denote the space of finite elements of order k on a triangle/tetrahedron mesh $\mathcal{T}_h$ and Q_h^k denote the space of finite elements of order k (bi/trilinear, bi/triquadratic etc.) on a quadrilateral/hexahedron mesh $\mathcal{T}_h$.

If $(P_h^k)^d \subset V_h$ and $P_h^l \subset L_h$ (or $(Q_h^k)^d \subset V_h$ and $Q_h^l \subset L_h$) for $k \ge 1$, $l \ge 0$, then the assumptions (A3) and (A5) follow by standard arguments; see, e.g., [5] or [10]. Assumption (A6) is also fulfilled on shape regular quasi-uniform meshes, which can be seen by virtue of the Bramble–Hilbert lemma and a transformation argument.

It still remains to discuss the assumption (A4). As mentioned in Remark 2.3, this assumption is obviously fulfilled, if the pair (V_h, L_h) is stable, i.e., if the inf-sup condition is directly fulfilled for (V_h, L_h). Therefore, our results are justified for all such pairs, as, e.g., "Taylor–Hood element" P_2/P_1, Q_2/Q_1/ or higher order "Taylor–Hood element" P_{k+1}/P_k, Q_{k+1}/Q_k; see [21], different bubble elements (P_1^b/P_0, Q_1^b/Q_0 etc.; see, e.g., [17]) etc.

In what follows we want to recall another discretization scheme, introduced in Becker and Braack [2], which also fulfills the assumption (A4). This scheme will be used in the next section for the numerical example.

For this discretization we assume that the (quadrilateral or hexahedron) mesh $\mathcal{T}_h$ is organized in a patchwise manner. This means, that it results from a coarser regular

mesh $\mathcal{T}_{2h}$ by one uniform refinement. By a "patch" of elements we denote a group of four cells (in 2D) or eight cells (in 3D) in $\mathcal{T}_h$ which results from a common coarser cell in $\mathcal{T}_{2h}$. The finite element spaces are chosen as

$$V_h = (Q_h^1)^d \qquad L_h = Q_h^1.$$

The space $\tilde{L}_h$ is defined as the space of bilinear/trilinear elements on the patch-mesh $\mathcal{T}_{2h}$, i.e., $\tilde{L}_h = Q_{2h}^1$. The stabilization form $s_h(\cdot,\cdot)$ is defined as

$$s_h(p_h, \xi_h) = \delta_0 h^2 \sum_{P \in \mathcal{T}_{2h}} (\nabla p_h - \widetilde{\nabla p_h}, \nabla \xi_h - \widetilde{\nabla \xi_h})_{L^2(P)},$$

where

$$\widetilde{\nabla p_h} = \frac{1}{|P|} \int_P \nabla p_h \, dx.$$

We refer to [2] for the proof that this scheme fulfills the assumption (A4). We note, also that other equal order stabilized schemes are included in our setting; see, e.g., [6].

6. Numerical examples. In this section we present two numerical examples (2D and 3D) illustrating our results. In both examples the Stokes equation are discretized by equal order elements (bilinear in 2D and trilinear in 3D) with a stabilization term as described in the previous section. The resulting finite dimensional optimal control problem is solved by primal-dual active set method; see, e.g., [3] or [22].

6.1. Example in 2D. We consider an optimal control problem as stated in section 1 with

$$\Omega = (0,1)^2, \quad \nu = 1, \quad a = (-0.1, -0.1)^t, \quad b = (0.25, 0.25)^t,$$

and a given solution

$$\begin{aligned} \bar{v}_1 &= \bar{w}_1 = \sin^2(\pi x)\, \sin(\pi y)\, \cos(\pi y), \\ \bar{v}_2 &= \bar{w}_2 = -\sin^2(\pi y)\, \sin(\pi x)\, \cos(\pi x), \\ \bar{p} &= \bar{r} = \sin(2\pi x)\, \sin(2\pi y), \end{aligned}$$

and

$$\bar{q} = \Pi_{[a,b]}\Big(-\frac{1}{\nu} \bar{w} \Big).$$

The data of the problem is then given by

$$\begin{aligned} f &= -\Delta \bar{v} + \nabla \bar{p} - \bar{q}, \\ v_d &= \bar{v} + \Delta \bar{w} + \nabla \bar{r}. \end{aligned}$$

In Figure 6.1 we show the first component of the optimal solution $\bar{q}$. The second component of $\bar{q}$ has a similar structure.

Let us remark that the assumption (A7) is fulfilled for this example. Let

$$\gamma_{ex} := \{x \in \Omega : (\bar{w}_1(x) - a_1)(\bar{w}_1(x) - b_1)(\bar{w}_2(x) - a_2)(\bar{w}_2(x) - b_2) = 0\},$$

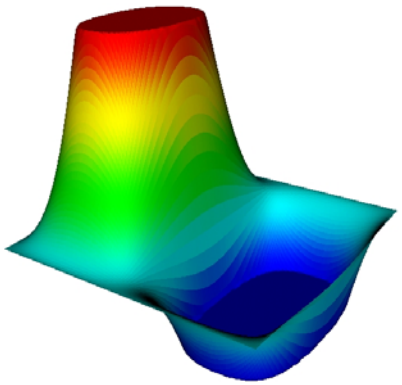


FIG. 6.1. *The first component of the optimal solution* $\bar{q}$.

TABLE 6.1
Convergence of $\|\bar{q} - \bar{q}_h\|_Q$ *and* $\|\bar{q} - \tilde{q}_h\|_Q$ *for* h *tending to zero.*

$h/\sqrt{2}$	$\|\bar{q} - \bar{q}_h\|_Q$	reduction rate	$\|\bar{q} - \tilde{q}_h\|_Q$	reduction rate
2^{-2}	7.65e-2	–	3.84e-2	–
2^{-3}	5.06e-2	1.50	8.59e-3	4.47
2^{-4}	2.73e-2	1.85	1.82e-3	4.72
2^{-5}	1.39e-2	1.97	4.23e-4	4.29
2^{-6}	6.99e-3	1.99	1.02e-4	4.15
2^{-7}	3.51e-3	1.99	2.51e-5	4.06

i.e., the curve (consisting of four connected parts) that separates active and inactive parts of the optimal control. Then we find the estimate

$$|K_1| \leq 2h|\gamma_{ex}|$$

and consequently (A7) holds.

In Table 6.1 we show the behavior of the error $\|\bar{q} - \bar{q}_h\|_Q$ and the error after the postprocessing step, i.e., $\|\bar{q} - \tilde{q}_h\|_Q$ on a sequence of uniformly refined meshes. As expected, we observe first order convergence for $\|\bar{q} - \bar{q}_h\|_Q$ and second order convergence for $\|\bar{q} - \tilde{q}_h\|_Q$.

In Table 6.2 we show the corresponding results concerning the error behavior with respect to $\|\cdot\|_{L^\infty(\Omega)}$. Although we only proved the results concerning the convergence with respect to $\|\cdot\|_{L^2(\Omega)}$, we observe similar behavior also for $\|\cdot\|_{L^\infty(\Omega)}$.

The results concerning the error behavior for the optimal velocity and the optimal pressure are given in Table 6.3. The pressure shows better order of convergence as $\mathcal{O}(h)$. Such effects are known for equal order finite elements on uniform meshes; see, e.g., [2].

6.2. Example in 3D. For the 3D case we construct a similar example as in 2D by setting

$$\Omega = (0,1)^3, \quad \nu = 1, \quad a = (-0.1, -0.1, -0.1)^t, \quad b = (0.25, 0.25, 0.01)^t.$$

TABLE 6.2
Convergence of $\|\bar{q} - \bar{q}_h\|_{L^\infty(\Omega)^d}$ and $\|\bar{q} - \tilde{q}_h\|_{L^\infty(\Omega)^d}$ for h tending to zero.

$h/\sqrt{2}$	$\|\bar{q} - \bar{q}_h\|_{L^\infty(\Omega)^d}$	reduction rate	$\|\bar{q} - \tilde{q}_h\|_{L^\infty(\Omega)^d}$	reduction rate
2^{-2}	2.60e-1	–	1.20e-1	–
2^{-3}	2.19e-1	1.18	2.43e-2	4.96
2^{-4}	1.27e-1	1.73	7.62e-3	3.19
2^{-5}	6.47e-2	1.96	1.77e-3	4.30
2^{-6}	3.25e-2	1.99	4.74e-4	3.74
2^{-7}	1.63e-2	2.00	1.22e-4	3.88

TABLE 6.3
Convergence of $\|\bar{v} - \bar{v}_h\|_{L^2(\Omega)^d}$ and $\|\bar{p} - \bar{p}_h\|_{L^2(\Omega)}$ for h tending to zero.

$h/\sqrt{2}$	$\|\bar{v} - \bar{v}_h\|_{L^2(\Omega)^d}$	reduction rate	$\|\bar{p} - \bar{p}_h\|_{L^2(\Omega)}$	reduction rate
2^{-2}	6.81e-2	–	4.94e-1	–
2^{-3}	1.68e-2	4.05	1.65e-1	2.99
2^{-4}	4.08e-3	4.12	5.49e-2	3.01
2^{-5}	9.95e-4	4.10	1.91e-2	2.87
2^{-6}	2.47e-4	4.02	6.67e-3	2.86
2^{-7}	6.12e-5	4.04	2.35e-3	2.84

The exact solution is given by

$$\begin{aligned}
\bar{v}_1 &= \bar{w}_1 = 2\,\sin^2(\pi x)\,\sin(2\pi y)\,\sin(2\pi z),\\
\bar{v}_2 &= \bar{w}_2 = -\sin^2(\pi y)\,\sin(2\pi x)\,\sin(2\pi z),\\
\bar{v}_3 &= \bar{w}_3 = -\sin^2(\pi z)\,\sin(2\pi x)\,\sin(2\pi y),\\
\bar{p} &= \bar{r} = \sin(2\pi x)\,\sin(2\pi y)\,\sin(2\pi z),
\end{aligned}$$

and

$$\bar{q} = \Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{w}\Big).$$

The data of the problem is then determined by

$$\begin{aligned}
f &= -\Delta\bar{v} + \nabla\bar{p} - \bar{q},\\
v_d &= \bar{v} + \Delta\bar{w} + \nabla\bar{r}.
\end{aligned}$$

For this problem, Assumption (A7) is valid for similar reasons as in the previous example. As for the 2D example, we present the behavior of error $\|\bar{q} - \bar{q}_h\|_Q$ and the error after the postprocessing step, i.e., $\|\bar{q} - \tilde{q}_h\|_Q$ in Table 6.4, and for the corresponding L^∞-norm in Table 6.5. In Table 6.6 the error for optimal pressure and velocity are shown.

TABLE 6.4
Convergence of $\|\bar{q} - \bar{q}_h\|_Q$ and $\|\bar{q} - \tilde{q}_h\|_Q$ for h tending to zero.

$h/\sqrt{3}$	$\|\bar{q} - \bar{q}_h\|_Q$	reduction rate	$\|\bar{q} - \tilde{q}_h\|_Q$	reduction rate
$1/3 \cdot 2^{-1}$	8.68e-2	–	2.13e-2	–
$1/3 \cdot 2^{-2}$	5.81e-2	1.49	5.27e-3	4.04
$1/3 \cdot 2^{-3}$	3.49e-2	1.66	1.08e-3	4.87
$1/3 \cdot 2^{-4}$	1.83e-2	1.91	2.43e-4	4.44

TABLE 6.5
Convergence of $\|\bar{q} - \bar{q}_h\|_{L^\infty(\Omega)^d}$ and $\|\bar{q} - \tilde{q}_h\|_{L^\infty(\Omega)^d}$ for h tending to zero.

$h/\sqrt{3}$	$\|\bar{q} - \bar{q}_h\|_{L^\infty(\Omega)^d}$	reduction rate	$\|\bar{q} - \tilde{q}_h\|_{L^\infty(\Omega)^d}$	reduction rate
$1/3 \cdot 2^{-1}$	3.29e-1	–	7.96e-2	–
$1/3 \cdot 2^{-2}$	2.99e-1	1.10	2.76e-2	2.88
$1/3 \cdot 2^{-3}$	2.59e-1	1.15	5.69e-3	4.85
$1/3 \cdot 2^{-4}$	1.39e-1	1.86	1.30e-3	4.37

TABLE 6.6
Convergence of $\|\bar{v} - \bar{v}_h\|_{L^2(\Omega)^d}$ and $\|\bar{p} - \bar{p}_h\|_{L^2(\Omega)}$ for h tending to zero.

$h/\sqrt{3}$	$\|\bar{v} - \bar{v}_h\|_{L^2(\Omega)^d}$	reduction rate	$\|\bar{p} - \bar{p}_h\|_{L^2(\Omega)}$	reduction rate
$1/3 \cdot 2^{-1}$	7.59e-2	–	8.72e-1	–
$1/3 \cdot 2^{-2}$	1.81e-2	4.19	2.88e-1	3.03
$1/3 \cdot 2^{-3}$	4.40e-3	4.11	1.02e-1	2.82
$1/3 \cdot 2^{-4}$	1.08e-3	4.07	3.65e-2	2.79

REFERENCES

[1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic optimal control problem*, Comput. Optim. Appr., 23 (2002), pp. 201–229.

[2] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 137–199.

[3] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[4] P. BOCHEV AND M. GUNZBURGER, *Least-squares finite-element methods for optimization and control problems for the Stokes equations*, Comput. Math. Appl., 48 (2004), pp. 1035–1057.

[5] D. BRAESS, *Finite Elemente*, Springer, Berlin, 1992.

[6] E. BURMAN AND P. HANSBO, *Edge stabilization for the generalized Stokes problem: a continuous interior penalty method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2393–2410.

[7] E. CASAS, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, submitted.

[8] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–219.

[9] E. CASAS AND F. TRÖLTZSCH, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, V. B. et al, ed., Boston, Kluwer Academic Publishers, Boston, 2003, pp. 89–100.

[10] P. CIARLET, *Basic error estimates for elliptic problems*, in Finite Element Methods, vol. II of Handbook of Numerical Analysis, North-Holland, Amsterdam, 1991, pp. 17–351.

[11] P. CLÉMENT, *Approximation by finite element functions using local regularization*, Revue Franc. Automat. Inform. Rech. Operat., 9 (1975), pp. 77–84.

[12] S. S. COLLIS AND M. HEINKENSCHLOSS, *Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems*, CAAM TR02-01, Rice University, Houston, TX, (2002).

[13] M. DAUGE, *Stationary Stokes and Navier–Stokes systems on two- or three-dimensional domains with corners* I, SIAM J. Math. Anal., 20 (1989), pp. 74–97.

[14] K. DECKELNICK AND M. HINZE, *Semidiscretization and error estimates for distributed control of the instationary Navier–Stokes equations*, Numer. Math., 97 (2004), pp. 297–320.

[15] R. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[16] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.

[17] V. GIRAULT AND P.-A. RAVIART, *Finite element methods for Navier–Stokes equations. Theory and algorithms*, Springer Ser. Comput. Math. 5, Berlin, 1986.

[18] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier–Stokes equations with Dirichlet controls*, Math. Model. Numer. Anal., 25 (1991), pp. 711–748.

[19] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier–Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.

[20] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

[21] P. Hood and C. Taylor, *A numerical solution of the Navier–Stokes equations using the finite element technique*, Internat. J. Comput. and Fluids, 1 (1973), pp. 73–100.

[22] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.

[23] J. L. Lions, *Contrôle Optimal de Systems Gouvernés par des Équations aux Dérivées Partielles*, Dunvol Gauthier–Villars, Paris, 1968.

[24] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal control problems*, Appl. Math. Opt., 8 (1982), pp. 69–95.

[25] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.

[26] C. Meyer and A. Rösch, *L^∞-estimates for approximated optimal control problems*, SIAM J. Control Optim., 44 (2005), pp. 1636–1649.

[27] A. Rösch, *Error estimates for parabolic optimal control problems with control constraints*, Z. Anal. Anwendungen, 23 (2004), pp. 353–376.

[28] A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., 21 (2006), pp. 121–134.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1921–1953

CONVERGENCE OF AN IMPLICIT SPACETIME GODUNOV FINITE VOLUME METHOD FOR A CLASS OF HYPERBOLIC SYSTEMS*

KATARINA JEGDIC† AND ROBERT L. JERRARD‡

Abstract. We study an implicit spacetime Godunov method for a class of hyperbolic conservation laws known as Temple class systems. We establish the well-posedness of this method, a discrete entropy inequality, a property analogous to the total variation diminishing property of certain numerical schemes for scalar conservation laws, and, as a consequence, the convergence of the numerical method.

Key words. conservation laws, spacetime discontinuous, Temple systems

AMS subject classifications. 65M12, 35L65

DOI. 10.1137/040613731

1. Introduction.

In this paper we prove the convergence of an implicit spacetime Godunov finite volume method for solving certain systems of hyperbolic conservation laws:

$$u_t + f(u)_x = 0, \qquad u : [0,\infty) \times \mathbb{R} \to \mathcal{D} \subset \mathbb{R}^n. \tag{1.1}$$

Here u is the vector of unknown densities of conserved variables, and $f : \mathcal{D} \to \mathbb{R}^n$ is the spatial flux defined on a domain of conservation states $\mathcal{D} \subset \mathbb{R}^n$. The system is supplemented by the initial condition

$$u(0,x) = u_0(x), \qquad x \in \mathbb{R}, \tag{1.2}$$

with the assumption that u_0 is a function of bounded variation.

In the numerical scheme that we consider, we fix a partition $\mathcal{T}_h$ of $[0,\infty)\times\mathbb{R}$ into a union of closed triangles T with diameter $\approx h$ and with pairwise disjoint interiors. We seek an approximate solution u_h in the space $\mathcal{P}_h$ of functions $u : [0,\infty) \times \mathbb{R} \to \mathcal{D}$ that are constant on the interior of each triangle. This approximate solution is required to satisfy an equation of the form

$$\int_{\partial T} F^{num}\, d\mathcal{H}^1 = 0 \tag{1.3}$$

for all triangles T in the triangulation $\mathcal{T}_h$. Here $\mathcal{H}^1$ denotes the one-dimensional Hausdorff measure on ∂T, and F^{num} denotes a numerical flux function, taking values in $\mathbb{R}^n$ and depending on the values of u_h on both sides of ∂T and on the unit normal

*Received by the editors August 22, 2004; accepted for publication (in revised form) April 17, 2006; published electronically October 4, 2006.
http://www.siam.org/journals/sinum/44-5/61373.html

†Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL 61801. Current address: Department of Mathematics, University of Houston, Houston, TX 77204-3008 (kjegdic@math.uh.edu). The first author was partially supported by the U.S. National Science Foundation and the U.S. Defense Advanced Research Projects Agency via an NSF grant, NSF DMS 98-73945. Continuing support is provided by NSF grant NSF DMR 01-21695.

‡Department of Mathematics, University of Toronto, Toronto, ON M5S 3G3, Canada (rjerrard@math.toronto.edu). The second author was partially supported by the National Science and Engineering Research Council of Canada under operating grant 261955.

to ∂T. The approximate solution u_h is also required to satisfy $u_h(0,\cdot) = u_{0h}$ in a suitable sense, where u_{0h} is a discretization of the initial data.

In this paper we always take the numerical flux in (1.3) to be the Godunov flux, the definition of which is recalled in section 3.2; this is arguably the most natural choice on physical grounds. In general, (1.3) is a massively coupled system of nonlinear equations, since values of u_h on adjacent triangles are coupled through the Godunov flux. Following [18], [20], we define *causal* and *patchwise causal* triangulations (see section 3.4) for which these equations largely decouple and take on a simpler form. Our main results, which are summarized in Theorem 3.3, prove that for these sorts of triangulations, and for a special class of conservation laws known as Temple class systems, approximate solutions generated by (1.3) converge to solutions that verify certain entropy inequalities.

With causal or patchwise causal triangulations and the Godunov flux, (1.3) is suitable for adaptive meshing and parallelization. The computational complexity is $O(N)$, where N is the number of triangles on which the approximate solution is computed. (For compactly supported initial data on a finite time interval, effectively $N \sim h^{-2}$.) In addition, it will be clear that (1.3) with the Godunov flux is consistent with the system (1.1).

Temple class systems were introduced in [27]. We recall their definition in section 2.2. Their chief properties are, first, that they come equipped with a coordinate system of Riemann invariants, in our notation given by a diffeomorphism $b = (b^1, \dots, b^n) : \mathcal{D} \to \mathcal{R} \subset \mathbb{R}^n$; and, second, that the set

$$S_{\alpha,c} := \{u \in \mathcal{D} : b^{\alpha_1}(u) = c_1, \dots, b^{\alpha_k}(u) = c_k\}$$

is an invariant set for solutions of (1.1), for any $\alpha = (\alpha_1, \dots, \alpha_k)$ and $c = (c_1, \dots, c_k)$ with $1 \le \alpha_1 < \cdots < \alpha_k \le n$. (A particular form of this invariance property, adapted to our purposes, is given in Lemma 5.3.) In [27] it was shown that Temple class systems of the sort that we consider are the only genuinely nonlinear, strictly hyperbolic systems with this abundance of invariant submanifolds.

A key point in the proof is an estimate analogous to the well-known total variation diminishing (TVD) property enjoyed by certain numerical schemes for scalar conservation laws. We show that when (1.1) is a Temple class system, for the approximation scheme (1.3) with the Godunov flux and a suitable triangulation $\mathcal{T}_h$, the approximate solution u_h is such that if $w_h^i := b^i(u_h)$, then

$$t \mapsto TV(w_h^i(t,\cdot)) \quad \text{is a nonincreasing function} \tag{1.4}$$

for every $i \in \{1, \dots, n\}$, where TV denotes the total variation. This conclusion is deduced from a related statement that holds on individual triangles, and it easily implies that sequences of approximate solutions are precompact in appropriate norms.

Note that a system that admits a convergent numerical method satisfying (1.4) must itself have the same property, and this implies in particular that $S_{\alpha,c}$, as defined above, is an invariant set for such a system (1.1). Thus (1.4) can hold only if (1.1) is a Temple class system, and hence our analysis relies very heavily on specific properties of Temple class systems.

Related work. The method (1.3) is the $k = 0$ case of various spacetime discontinuous Galerkin (DG) finite element methods, as proposed in [8], [18], [20], for example. Spacetime DG methods were first devised and studied for linear hyperbolic equations; see [14], [11]. They are now well understood for scalar conservation laws, starting with work on the shock-capturing DG method. Convergence of this method (with piecewise

polynomial approximants of arbitrary degree) for scalar conservation laws in d space dimensions was proven by Jaffre, Johnson, and Szepessy [8], and error estimates were established by Cockburn and Gremaud [3]. Related earlier work established convergence of the similar shock-capturing streamline diffusion finite element method (see [24], [25]), which, however, is not a DG method, and error estimates for this method were also given in [3]. The only work that we know of that proves convergence of any DG method for any system of conservation laws is a very recent paper of Arvanitis, Makridakis, and Tzavaras [1] that considers a DG method based on a relaxation approximation. This method, which does not reduce to (1.3) when $k = 0$, is shown in [1] to converge for systems that can be controlled using compensated compactness techniques.

As remarked earlier, our method can also be seen as a Godunov finite volume method, and our analysis is thus related to numerous results in the large literature on Godunov finite volume methods; see [15] for a survey. In particular, LeVeque and Temple [16] and Serre [22] establish the stability of Godunov's method for Temple class systems of two equations, via TVD estimates of the form (1.4). (Serre also proves similar results for the Lax–Friedrichs method and the Glimm scheme.) Our method differs from the basic Godunov scheme and other standard Godunov finite volume methods in that it is a spacetime method, and one which is implicit in the sense that, even in the easiest case, a nonlinear system of n equations in n unknowns must be solved to determine the approximate solution on each spacetime triangle. This causes genuine difficulties. For example, even existence and uniqueness of solutions of the scheme (1.3) is not immediate. Indeed, at the end of section 6 we show by example that the solution of scheme (1.3) need not be unique without suitable restrictions on the mesh, even in the scalar case.[1] It should be noted that, despite these theoretical difficulties, in practice approximate solutions can be computed quite efficiently using Newton's method; see [20] and [18] for related numerical results.

Temple class systems have also been studied by the front tracking method; see, for example, Baiti and Bressan [2], which uses front tracking approximations to establish the existence of a unique Lipschitz semigroup for large data. Heibig [7] proves that entropy solutions of Temple class systems with small total variation are unique.

Contents of this paper. We sketch the organization of this paper. Section 2 presents some notation and summarizes background on hyperbolic conservation laws, including the definition of Temple class systems and related notation that we will use throughout the paper. Section 3 gives a precise formulation of our approximation scheme and states our main results.

In section 4 we present results valid for hyperbolic systems not necessarily of Temple class, including the well-posedness of the numerical method (1.3) for causal triangulations when the flux is globally Lipschitz, and a discrete entropy inequality.

The rest of the paper deals exclusively with Temple class systems. The proof of well-posedness of the approximation scheme on a single causal triangle, with Riemann invariant bounds, is presented in section 5. (Well-posedness does not follow from the general result in section 4, as Temple class systems do not satisfy the global Lipschitz condition assumed there.) In section 6 we prove the corresponding results on a single causal patch. Finally, in section 7 we complete the proof of the main theorem by converting local Riemann invariant bounds (on a single element or patch)

[1] This issue did not arise in [8], [3], as these papers employ the Lax–Friedrichs flux where we use the Godunov flux, and in the scalar case approximate solutions are unique for *strictly* monotone numerical fluxes.

to global estimates that imply compactness sufficient to pass to limits and obtain weak solutions.

2. Definitions, notation, and background.

2.1. Some notation. We write $M^{k\times\ell}$ for the space of $k\times\ell$ matrices. We identify $\mathbb{R}^n$ with $M^{n\times 1}$, except where explicitly noted otherwise. In particular, the state variable u, flux f, and numerical flux F^{num} are understood to be column vectors. If $\zeta:\mathbb{R}^n\to\mathbb{R}^k$ is any function, then $D\zeta$ denotes its gradient, which is understood to be a $k\times n$ matrix. In particular, if ζ is a scalar function on $\mathcal{D}\subset\mathbb{R}^n$, then $D\zeta$ is a row vector.

If M is a matrix, then M^t denotes its transpose.

We write $\|\cdot\|$ to denote the Euclidean norm.

Given $v_1,v_2\in\mathbb{R}^n$, we write (v_1,v_2) to denote the $n\times 2$ matrix whose first and second columns are v_1 and v_2, respectively.

Given $\zeta:\mathbb{R}\to\mathbb{R}^k$, we write $TV(\zeta)$ to denote the total variation of ζ, i.e.,

$$TV(\zeta):=\sup\left\{\sum\|\zeta(x_i)-\zeta(x_{i-1})\|:\cdots<x_{i-1}<x_i<x_{i+1}<\cdots\right\}. \tag{2.1}$$

For $h>0$, $\mathcal{T}_h$ will always denote a triangulation of $[0,\infty)\times\mathbb{R}$, that is, a collection of closed triangles with pairwise disjoint interiors whose union is $[0,\infty)\times\mathbb{R}$. Here h denotes a length scale; we will be more precise about this later. We write T to denote a generic triangle in $\mathcal{T}_h$. We write $\mathcal{P}_h$ to denote the space of functions $u\in L^\infty([0,\infty)\times\mathbb{R};\mathbb{R}^n)$ such that, for every $T\in\mathcal{T}_h$, u is constant on the interior of T.

We write e to denote a generic edge of a triangle $T\in\mathcal{T}_h$. We use the notation

$$\nu_{e,T}:=\text{ outer unit normal to } T \text{ along edge } e.$$

Given $v\in\mathcal{P}_h$, we write v_T to indicate the value of v on the interior of triangle T.

We always make the following assumption about triangulations $\mathcal{T}_h$: for every edge e of a triangle $T\in\mathcal{T}_h$, if e is not contained in $\{t=0\}\times\mathbb{R}$, then there is a unique triangle, denoted T_e, that shares edge e with T, so that $T\cap T_e=e$. Thus v_{T_e} is the value of v "across edge e from triangle T." It will sometimes be convenient to consider a single triangle T, not as part of a triangulation $\mathcal{T}_h$. When we do this, we will write u_{T_e} to denote prescribed data on the exterior of edge e.

2.2. Systems of conservation laws. Consider a system of conservation laws (1.1). We always assume that the system (1.1) is *strictly hyperbolic*, which means that the matrix $Df(u)$ has real and distinct eigenvalues. Let us suppose that $\lambda^1(u)<\cdots<\lambda^n(u)$ for all $u\in\mathcal{D}$. Further, let us denote by $r_i(u)\in\mathbb{R}^n$ and $l_i(u)\in M^{1\times n}$, $i\in\{1,\dots,n\}$, the corresponding right and left eigenvectors, respectively, of the matrix $Df(u)$. We normalize these eigenvectors by requiring that $\|r_i(u)\|=\|l_i(u)\|=1$ for all $i\in\{1,\dots,n\}$ and $u\in\mathcal{D}$.

The ith characteristic field λ^i is called *genuinely nonlinear* if

$$D\lambda^i(u)\,r_i(u)\neq 0 \quad \text{for all } u\in\mathcal{D}. \tag{2.2}$$

If all characteristic fields of the gradient matrix Df are genuinely nonlinear, the system (1.1) is genuinely nonlinear.

We say that system (1.1) has a coordinate system of *Riemann invariants* in $\mathcal{D}$ if there exist a subset $\mathcal{R}\subset\mathbb{R}^n$ and a diffeomorphism $b:\mathcal{D}\to\mathcal{R}$ such that

$$Db(u)\,Df(u)\,Db(u)^{-1}=\Lambda(u) \quad \text{for } u\in\mathcal{D}, \tag{2.3}$$

where $\Lambda(u)$ is the diagonal matrix given by

$$\Lambda(u) := \operatorname{diag}(\lambda^1(u), \dots, \lambda^n(u)).$$

Let us define $a := b^{-1}$. For $w := b(u) \in \mathcal{R}$, we will sometimes use the following notation:

$$\begin{aligned} &\tilde{f}(w) := f(a(w)), \\ &\tilde{\lambda}^i(w) := \lambda^i(a(w)), \quad i \in \{1, \dots, n\}, \\ &\tilde{\Lambda}(w) := \Lambda(a(w)). \end{aligned} \tag{2.4}$$

Notice that (2.3) is equivalent to

$$D\tilde{f}(w) = Da(w)\,\tilde{\Lambda}(w), \quad \text{for } w \in \mathcal{R}. \tag{2.5}$$

Suppose that (1.1) is a strictly hyperbolic system of conservation laws which is genuinely nonlinear and equipped with the coordinate system of Riemann invariants. If in addition, the set

$$a(\{w \in \mathcal{R} : w^i = c\}) \tag{2.6}$$

is contained in a hyperplane in $\mathcal{D}$, for every $i \in \{1, \dots, n\}$ and every constant $c \in \mathbb{R}$, the system (1.1) is said to be a *genuinely nonlinear Temple class* system. As remarked in the introduction, Temple class systems are characterized by the fact that they have numerous invariant submanifolds.

Throughout this paper, when we study genuinely nonlinear Temple class systems, we will assume that the domain of conservation states $\mathcal{D}$ is such that

$$\mathcal{R} = b(\mathcal{D}) \subset \mathbb{R}^n \quad \text{is a rectangle} \tag{2.7}$$

with sides parallel to the coordinate hyperplanes. We will also assume that there exists K_0 (depending on f and $\mathcal{D}$) such that

$$\|Da\|_{L^\infty(\mathcal{R})} + \|Db\|_{L^\infty(\mathcal{D})} \le K_0. \tag{2.8}$$

2.3. Entropy solutions. We next recall the notion of entropy and of an entropy solution. Let (η, ψ) be an *entropy-entropy flux pair* for the system (1.1); this means that $\eta, \psi : \mathcal{D} \to \mathbb{R}$ are smooth functions, η is convex, and

$$D\eta(z)\,Df(z) = D\psi(z) \quad \text{for all } z \in \mathcal{D}. \tag{2.9}$$

Because $D^2\psi$ is symmetric, any entropy η for (1.1) must satisfy

$$D^2\eta(z)\,Df(z) = Df(z)^t\,D^2\eta(z) \quad \text{for all } z \in \mathcal{D}. \tag{2.10}$$

In the case $n = 2$, the PDE system (2.10) reduces to a hyperbolic equation whose solutions form an infinite-dimensional space of entropy functions. When $n > 2$, the system (2.10) is, in general, overdetermined. However, existence of a coordinate system of Riemann invariants is sufficient to nullify this overdeterminancy (see [5, section 7.4]). In particular, Temple class systems are endowed with large numbers of entropy-entropy flux pairs.

Suppose that u is a weak solution of the initial value problem (1.1), (1.2). If the inequality

$$\eta(u)_t + \psi(u)_x \le 0 \tag{2.11}$$

holds in the weak sense on the spacetime domain $[0, \infty) \times \mathbb{R}$ for all entropy-entropy flux pairs (η, ψ) of the system (1.1), then u is called an *entropy solution* of (1.1), (1.2).

3. Main results.

3.1. More notation. We first introduce notation that will enable us to give a precise formulation of our approximation scheme (i.e., (1.3) with the Godunov flux).

We define

$$\mathcal{T}_h^0 := \{T \in \mathcal{T}_h \ : \ \partial T \cap (\{t=0\} \times \mathbb{R}) \text{ is an interval}\}, \tag{3.1}$$

$$\mathcal{T}_h^+ := \mathcal{T}_h \setminus \mathcal{T}_h^0. \tag{3.2}$$

We write $g(u_T, u_{T_e}; \nu)$ to denote the Godunov flux. We define this below, first in full generality, and then in an important and simpler special case. (In fact, we will rarely need the general definition.) With this notation, we seek an approximate solution $u_h \in \mathcal{P}_h$ of (1.1) by requiring that u_h satisfy

$$\int_{\partial T} g(u_{h,T}, u_{h,T_e}; \nu_{e,T}) \, d\mathcal{H}^1 = 0 \quad \text{for every } T \in \mathcal{T}_h^+. \tag{3.3}$$

We also require that the initial condition (1.2) be approximately satisfied in the sense that

$$u_{h,T} = \frac{1}{\mathcal{H}^1(e)} \int_e u_0 \, dx \quad \text{for } \ e := \partial T \cap (\{t=0\} \times \mathbb{R}), \quad \text{for every } T \in \mathcal{T}_h^0. \tag{3.4}$$

3.2. Godunov flux. Formally, the *Godunov flux* associated with $u_1, u_2 \in \mathbb{R}^n$ and a unit (column) vector $\nu = (\nu_t, \nu_x)^t$ is the flux through a line element with spacetime unit normal ν, with values u_1 and u_2 on the two sides of the line element. More precisely, given u_1, u_2, ν as above, if $\nu_x \neq 0$, let $u : (0,\infty) \times \mathbb{R} \to \mathbb{R}^n$ denote the standard entropy solution of the Riemann problem[2] for (1.1) with initial data

$$u(0,x) = \begin{cases} u_1 & \text{if } \operatorname{sign} x = -\operatorname{sign} \nu_x, \\ u_2 & \text{if } \operatorname{sign} x = \operatorname{sign} \nu_x. \end{cases} \tag{3.5}$$

We define

$$\xi_-(u_1, u_2; \nu) := \lim_{(t,x)\cdot\nu \nearrow 0, \ t>0} u(t,x), \qquad \xi_+(u_1, u_2; \nu) := \lim_{(t,x)\cdot\nu \searrow 0, \ t>0} u(t,x).$$

These limits exist, since the solution u of the Riemann problem has bounded variation in x for every t and depends only on $\frac{x}{t}$. We will typically omit the dependence on u_1, u_2, ν and write ξ_-, ξ_+. The fact that u is a weak solution of (1.1) implies that $(\xi_-, f(\xi_-))\,\nu = (\xi_+, f(\xi_+))\,\nu$. Thus the definition of the Godunov flux as

$$g(u_1, u_2; \nu) := (\xi_-, f(\xi_-))\,\nu = (\xi_+, f(\xi_+))\,\nu$$

makes sense. (In fact, in the concrete situations we consider later, it will always be the case that $\xi_+ = \xi_-$, and so we will always write simply ξ and call it the *Godunov value.*) If $\nu = (\nu_t, \nu_x)^t \in \mathbb{R}^2$ is a unit vector with $\nu_x = 0$, then we define

$$g(u_1, u_2; \nu) := (u_1, f(u_1))\,\nu \quad \text{if } \nu_t = 1, \qquad g(u_1, u_2; \nu) := -(u_2, f(u_2))\,\nu \quad \text{if } \nu_t = -1.$$

[2]The Godunov flux $g(u_1, u_2; \nu)$ is generally defined only if the solution of the Riemann problem is defined.

3.3. Causal and semicausal edges. We will study the initial-value problem (1.1), (1.2) in situations for which we know a priori that the solution takes values in some set $\mathcal{D} \subset \mathbb{R}^n$. Let $e \subset \partial T$ be an edge of a triangle. If

$$\big(1,\, \lambda^i(u)\big)\, \nu_{e,T} < 0, \quad i \in \{1, \dots, n\}, \qquad \text{for all } u \in \mathcal{D}, \tag{3.6}$$

then we say that e is an *inflow* edge for element T (see Figure 1(a)). Similarly, if

$$\big(1,\, \lambda^i(u)\big)\, \nu_{e,T} > 0, \quad i \in \{1, \dots, n\}, \qquad \text{for all } u \in \mathcal{D}, \tag{3.7}$$

then we say that e is an *outflow* edge for element T (see Figure 1(b)). We also use the terminology

$$\text{an edge } e \text{ is causal if it is either outflow or inflow,} \tag{3.8}$$

and we further say that e is *semicausal* if it is not causal and

$$(1, \lambda^i(u))\, \nu_{e,T} \neq 0 \quad \text{for all } i \in \{1, \dots, n\} \text{ and } u \in \mathcal{D}. \tag{3.9}$$

We will sometimes write "causal (inflow, outflow, ...) for f, $\mathcal{D}$" to emphasize the dependence on the nonlinearity and the domain of conservation states.

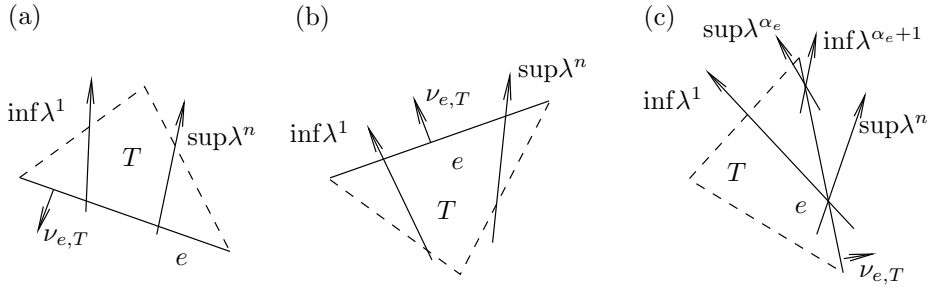


FIG. 1. *Inflow, outflow, and semicausal edges.*

If e is semicausal, then the strict hyperbolicity assumption and the continuity of $u \mapsto \lambda^i(u)$ for all i imply that there exists some $\alpha_e \in \{1, \dots, n-1\}$ such that

$$\lambda^{\alpha_e}(u) < -\frac{\nu_t}{\nu_x} < \lambda^{\alpha_e+1}(u) \tag{3.10}$$

for all $u \in \mathcal{D}$, where $\nu = \nu_{e,T}$ denotes the outer normal to T along edge e (see Figure 1(c)). Thus, there can exist semicausal edges only if $\sup_{u\in\mathcal{D}} \lambda^i(u) \leq \inf_{u\in\mathcal{D}} \lambda^{i+1}(u)$ for some i.

It follows from well-known features of the solution of the Riemann problem that if $u_1, u_2 \in \mathcal{D}$, then

$$\text{if } e \text{ is an inflow edge, then } g(u_1, u_2; \nu_{e,T}) = (u_2, f(u_2))\, \nu_{e,T}, \tag{3.11}$$

$$\text{if } e \text{ is an outflow edge, then } g(u_1, u_2; \nu_{e,T}) = (u_1, f(u_1))\, \nu_{e,T}. \tag{3.12}$$

We will take these last two equations as the *definition* of Godunov flux in situations where the Riemann problem is not necessarily solvable but (3.6) or (3.7) is satisfied.

We will see in Lemma 6.1 that for Temple class systems the Godunov flux also has a relatively simple form across semicausal edges.

3.4. Causal and patchwise causal triangulations. We say that a triangle T is *causal* if all of its edges are causal, and a triangulation $\mathcal{T}_h$ is *causal* if every triangle T in $\mathcal{T}_h$ is causal.

A triangle T is *semicausal* if every edge of T is either causal or semicausal. Our analysis extends to certain semicausal triangulations that we define as follows. Let $\mathcal{T}_h$ be a triangulation of $[0,\infty)\times\mathbb{R}$. A *patch* is defined to be a union of triangles in $\mathcal{T}_h$. A patch $P=\cup_{i=1}^k T_i$ is said to be *causal* if

$$\text{every edge } e\subset\partial P \text{ is causal,} \tag{3.13}$$

$$\text{every triangle } T_i\subset P \text{ has an outflow edge,} \tag{3.14}$$

$$\text{if } e \text{ is an edge of a triangle } T_i\subset P, \text{ and } e\not\subseteq\partial P, \text{ then } e \text{ satisfies (3.9).} \tag{3.15}$$

A triangulation $\mathcal{T}_h$ is said to be *patchwise causal* if it is a union of causal patches. These causal patches can always be assumed to be minimal, which we define to mean that

$$P \text{ does not have any proper subpatch satisfying (3.13)–(3.15).} \tag{3.16}$$

A proper subpatch is defined in the natural way: if $P=\cup_{i=1}^k T_i$ is a patch, then a proper subpatch is a set of the form $P'=\cup_{\ell=1}^j T_{i_\ell}$ for some $j<k$.

For a causal or semicausal triangle T, we write

$$\partial T^- = \text{the inflow portion of } \partial T, \tag{3.17}$$

$$\partial T^+ = \text{the outflow portion of } \partial T, \tag{3.18}$$

$$\partial T^0 = \text{the semicausal portion of } \partial T, \tag{3.19}$$

some of which may be empty.

For causal triangles or patches, the system (3.3) decouples to a certain extent. For example, in view of (3.11), (3.12), on a causal triangle T the scheme (3.3) takes the form

$$(u,f(u))\,n \;+\; \sum_{e\ \text{inflow}} (u_{T_e},f(u_{T_e}))\,n_e = 0, \tag{3.20}$$

where we are writing simply u instead of u_h, and

$$n := \int_{\partial T^+} \nu_{e,T}\; d\mathcal{H}^1, \qquad n_e := \int_e \nu_{e,T}\; d\mathcal{H}^1 = \nu_{e,T}\;\mathcal{H}^1(e). \tag{3.21}$$

We note for future reference that

$$n \;+\; \sum_{e\ \text{inflow}} n_e = \int_{\partial T} \nu\; d\mathcal{H}^1 = 0. \tag{3.22}$$

If the data u_{T_e} is known on all inflow edges, then (3.20) is a system of n equations in unknowns $u=(u^1,\dots,u^n)$, depending on the inflow data and triangle geometry as parameters. Thus, once (3.20) is known to be solvable for suitable inflow data in a single triangle, one can solve the whole system (3.3) one triangle at a time, always considering triangles for which the inflow data have already been found. This strategy is discussed at greater length in numerous references; see, for example, [18], [20]. Note that we require that on every triangle T the approximate solution u_T satisfy $u_T\in\mathcal{D}$,

since this condition is assumed in our definition of causality and indeed is needed to ensure that (3.3) reduces to (3.20).

Similarly, for a patchwise causal triangulation, we write ∂P^- and ∂P^+ to denote the inflow and outflow portions, respectively, of ∂P. We will show that if P is a causal patch, then one can solve (3.3) in P, once inflow data on ∂P^- is known; thus a patch-by-patch solution strategy is possible.

We note the following result.

LEMMA 3.1. *If T is a causal triangle, then it must have either one or two inflow edges.*

Proof. Multiply (3.22) on the left by the vector $(1, \lambda^i(u))$ for some $i \in \{1, \dots, n\}$ and some $u \in \mathcal{D}$. We deduce that $\sum_{j=1}^{3} (1, \lambda^i(u))\ \nu_{e_j,T}\ \mathcal{H}^1(e_j) = 0$, where e_1, e_2, e_3 denote the three edges. This shows that (3.6) cannot hold for all $e_j, j = 1, 2, 3$, and similarly (3.7) cannot hold for all three edges. ∎

LEMMA 3.2. *Let P be a causal patch, and let $T \subset P$ be a triangle with two causal edges. Then T must have an inflow edge.*

Proof. Let e_1, e_2, and e_3 be edges for triangle T. Let us suppose that e_1 is an outflow edge, and let e_3 be an edge which satisfies (3.10) for some $\alpha_{e_3} \in \{1, \dots, n-1\}$. We need to show that e_2 is inflow. Exactly as in the proof of Lemma 3.1, we find that $\sum_{j=1}^{3} (1, \lambda^i(u))\ \nu_{e_j,T}\ \mathcal{H}^1(e_j) = 0$ for all $i \in \{1, \dots, n\}$ and $u \in \mathcal{D}$. Since e_1 is outflow and e_3 satisfies (3.10), there must be some i such that $(1, \lambda^i(u))\ \nu_{e_1,T}$ and $(1, \lambda^i(u))\ \nu_{e_3,T}$ are both positive. Hence $(1, \lambda^i(u))\ \nu_{e_2,T} < 0$ for this i and therefore (since e_2 is causal) for all $i \in \{1, \dots, n\}$. Thus e_2 is an inflow edge, as required. ∎

3.5. Additional mesh-related considerations. We will see that the restriction of the system (3.3) to a causal patch or causal triangle always has a solution that is unique in a certain sense and that is bounded by the inflow data. In order to convert these local statements to global control over an approximate solution u_h, we need to impose some additional restrictions on the triangulations that we will consider.

To derive the TVD-like estimate (1.4) from bounds on the total variation of the Riemann invariants of an approximate solution on individual patches, we will need to assume that for every edge e of every triangle $T \in \mathcal{T}$

$$\text{if } \min_{(t,x)\in e} t > \min_{(t,x)\in T} t, \quad \text{then } e \text{ is an outflow edge for } T. \tag{3.23}$$

The necessity of this assumption is illustrated in Example 7.1. The point is that if an edge e violates (3.23), then information that enters a triangle along this edge can propagate backwards in time within the triangle.

Next, one can easily see from (3.20) and (3.22) that if T is a causal triangle with a single inflow edge e and inflow data u_{T_e}, then $u = u_{T_e}$ is a solution of (3.20). We will later prove that for Temple class systems, this is the only solution in $\mathcal{D}$.

Motivated by this, we say that an edge e is *trivial* if e forms the entire inflow boundary ∂T^- for some causal triangle T, and nontrivial otherwise. Many of our results will assume that there exists a constant $\alpha > 0$ such that

$$\text{if } e \text{ is any nontrivial edge,} \quad \text{then } |\nu_x| \geq \alpha |\nu_t| \quad \text{for } \nu = \nu_{e,T}. \tag{3.24}$$

Assumption (3.24), together with the fact that our solution u_h is constant except along nontrivial edges, is used to obtain bounds on $|u_{h,t}|$ from control over $|u_{h,x}|$. (Here and in what follows, $|u_{h,t}|, |u_{h,x}|$ are understood as nonnegative measures on $(0, \infty) \times \mathbb{R}$).) Bounds on $|u_{h,x}|$ in turn are an immediate consequence of the TVD estimate (1.4).

3.6. Main results. The main result of this paper is the following.

THEOREM 3.3. *Let* (1.1) *be a strictly hyperbolic, genuinely nonlinear Temple class system of conservation laws, and assume that the domain of conservation states* $\mathcal{D} \subset \mathbb{R}^n$ *satisfies* (2.7). *Let* $\mathcal{T}_h$ *be a causal or patchwise causal triangulation. Then there exists a unique solution* $u_h \in \mathcal{P}_h$ *of* (3.3), (3.4) *satisfying*

$$\min_{e \subset (\partial T^- \cup \partial T^0)} w^i_{h,T_e} \le w^i_{h,T} \le \max_{e \subset (\partial T^- \cup \partial T^0)} w^i_{h,T_e} \quad \text{for } w^i_h := b^i(u_h) \tag{3.25}$$

for every $i \in \{1, \dots, n\}$ *and* $T \in \mathcal{T}_h$.

If $\mathcal{T}_h$ *also satisfies* (3.23) *and* (3.24), *then in addition*

$$t \mapsto TV(w^i_h(t, \cdot)) \quad \text{is a nonincreasing function} \tag{3.26}$$

for every $i \in \{1, \dots, n\}$.

Finally, let $\mathcal{T}_h$ *be a sequence of causal or patchwise causal triangulations such that* (3.23) *holds and* (3.24) *is satisfied with a positive constant* α *independent of* h. *Assume also that there exists a constant* C *independent of* h *such that*

$$C^{-1}h \le |e| \le Ch \qquad \text{for every edge } e \text{ of every triangle } T \in \mathcal{T}_h. \tag{3.27}$$

Then the approximate solutions $\{u_h\}$ *are precompact in* $L^1_{loc}([0, \infty) \times \mathbb{R}; \mathbb{R}^n)$, *and any limit of any convergent subsequence* u_{h_k} *with* $h_k \to 0$ *is a distributional solution of* (1.1), (1.2). *Moreover, given any strictly convex entropy* η *with entropy flux* ψ, *there exists a subset* $\mathcal{D}' \subset \mathcal{D}$ *such that if* $u_0(x) \in \mathcal{D}'$ *for almost every* x, *then any limiting solution satisfies the entropy inequality* (2.11) *in the sense of distributions.*

Remark 3.4. One can find $\mathcal{D}' \subset \mathcal{D}$ such that, if $u_0(x) \in \mathcal{D}'$ for almost every x and u_0 has sufficiently small total variation, then there is a unique entropy solution u of (1.1) with initial data u_0, and the whole sequence $\{u_h\}$ of approximate solutions converges to this entropy solution.

The uniqueness assertion has been proven by Heibig [7]. The convergence $u_h \to u$ follows from two considerations. First, the theorem shows that for suitable $\mathcal{D}'$ any limit of a convergent subsequence must satisfy the entropy inequality (2.11) for some fixed strictly convex η. Second, it is well known that if a weak solution of (1.1) has only weak shocks (this can be arranged by a choice of $\mathcal{D}'$) and satisfies the entropy inequality (2.11) for a single strictly convex entropy, it is in fact an entropy solution. A proof of this last fact can be found, for example, in [23, Vol. 1, section 4.3].

Remark 3.5. Numerical simulations in Palaniappan, Haber, and Jerrard (see [20]) in the scalar case employ causal patches in which every patch consists of two triangles separated by a vertical semicausal edge.

Remark 3.6. Implicit in the numerical method (3.3), (3.4) is that the Godunov flux is well defined, and hence that the Riemann problem with initial data (3.5) is solvable for all $(u_1, u_2) \in \mathcal{D} \times \mathcal{D}$. Solutions of the Riemann problem for Temple class systems (see, for example, [22]) as far as we know all assume

$$\max_{u \in \mathcal{D}} \lambda^i(u) \le \min_{u \in \mathcal{D}} \lambda^{i+1}(u)$$

for all i. In fact, our results remain valid when this condition does not hold, provided that we take (3.11), (3.12), and (6.3) as the definitions of $g(u_1, u_2; \nu)$ on inflow, outflow, and semicausal edges, respectively.

We believe that, without (3.23), the numerical method is still convergent, but the proof would be somewhat more complicated. In particular, we give an example in section 7 to show that (3.26) can fail if (3.23) is not assumed.

In section 6 we also give an example showing that approximate solutions can fail to be unique on patches of elements on which we do not assume properties (3.14) and (3.15) from the definition of a causal patch.

4. General hyperbolic systems with causal triangulations. In this section we prove some results that do not use the special structure of Temple class systems, but that do, however, impose causality-type assumptions related to those introduced earlier.

4.1. Well-posedness of the approximation scheme. In this section we prove that, for a hyperbolic conservation law with flux f that is globally Lipschitz on $\mathbb{R}^n$ (but satisfying no other structural conditions), there is a unique solution of the approximation scheme (3.20) on suitable triangles. This is not needed for other results in this paper, but it indicates that one can at least hope to extend our stability and convergence results to more general classes of hyperbolic systems.

Our discussion will focus on triangles with two inflow edges (Figure 2), since the case of triangles with a single inflow edge is much easier. For triangles with two inflow edges, we write e_l and e_r to denote inflow edges on the left and right, respectively, and e to denote the single outflow edge. Defining $n_l = n_{e_l} = \nu_{e_l}\mathcal{H}^1(e_l)$ as in (3.21), and similarly n_r and n, we rewrite (3.20) in the form

$$\mathcal{F}(u_l, u_r, u) := F(u_l)\ n_l + F(u_r)\ n_r + F(u)\ n = 0, \tag{4.1}$$

where u_l, u_r denote prescribed inflow data on edges e_l, e_r. We also use the notation

$$F(z) := (z, f(z)) \in M^{n\times 2}. \tag{4.2}$$

We first prove the following result.

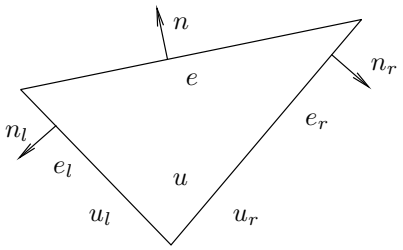


FIG. 2. *An element with two inflow faces.*

THEOREM 4.1. *Consider a system of hyperbolic conservation laws* (1.1) *with a domain of conservation states* $\mathcal{D} \subset \mathbb{R}^n$*, and assume that there exists a constant* $K_1 > 0$ *with the property*

$$\|f(u) - f(v)\| \le K_1 \|u - v\| \quad \text{for every } u, v \in \mathcal{D}. \tag{4.3}$$

Let T *be a triangle such that each edge* e *of* T *satisfies (writing* ν *for* $\nu_{e,T}$*)*

$$|\nu_t| - K_1|\nu_x| > 0. \tag{4.4}$$

Then there exists a constant K_2 *such that if* $u, \tilde{u} \in \mathcal{D}$ *solve* (3.20) *on* T *with inflow data* $\{u_{T_e} \in \mathcal{D} : e \subset \partial T^-\}$ *and* $\{\tilde{u}_{T_e} \in \mathcal{D} : e \subset \partial T^-\}$*, respectively, then*

$$||u - \tilde{u}|| \le K_2 \sum_{e\subset\partial T^-} \|u_{T_e} - \tilde{u}_{T_e}\|. \tag{4.5}$$

In particular, there is at most one solution in $\mathcal{D}$ *for inflow data in* $\mathcal{D}$*.*

Moreover, if $\mathcal{D} = \mathbb{R}^n$, then for any inflow data $\{u_{T_e} \in \mathbb{R}^n : e \subset \partial T^-\}$ there exists a unique solution of (3.20) on triangle T.

Proof. Step 0. First note that (4.3) and (4.4) together imply that T is causal for $f, \mathcal{D}$. This follows from noting that for any $u \in \mathcal{D}$ and all $i \in \{1, \dots, n\}$

$$|\lambda^i(u)| = \|Df(u)\ r_i(u)\| = \left\| \lim_{h \to 0} \frac{1}{h}\ [f(u + hr_i(u)) - f(u)] \right\| \leq K_1,$$

by (4.3). Thus (4.4) implies that if e is an edge with normal $\nu_{e,T} = \nu = (\nu_t, \nu_x)^t$ with $\nu_t > 0$, say, then

$$\nu_t + \lambda^i(u)\, \nu_x \ \geq\ |\nu_t| - K_1 |\nu_x| \ >\ 0,$$

so that e is an outflow edge for T. Similarly if $\nu_t < 0$, then e is an inflow edge for T.

Step 1. We now prove uniqueness and continuous dependence. We start by noting that, for n as defined in (3.21) and any $u, \tilde{u} \in \mathcal{D}$,

$$\begin{aligned} \|(F(u) - F(\tilde{u}))\ n\| &\geq \|u - \tilde{u}\|\ |n_t| - \|f(u) - f(\tilde{u})\|\ |n_x| \\ &\geq \|u - \tilde{u}\|\ (|n_t| - K_1 |n_x|). \end{aligned} \tag{4.6}$$

In view of assumption (4.4) and the definition (3.21) of n (as a multiple of ν), it follows that $u \mapsto F(u)\ n$ is one-to-one, which proves that there can be at most one solution u of (3.20), once the inflow data is specified.

Since the trivial solution $u = u_{T_e}$, where $e = \partial T^-$, always exists on a triangle with only one inflow edge, the uniqueness of solutions proves that (4.5) holds in this case. We next establish (4.5) for triangles with two inflow edges. We use the notation (4.1). Writing out these equations for both u and $\tilde{u}$ and subtracting, we obtain

$$(F(\tilde{u}) - F(u))\ n = (F(u_l) - F(\tilde{u}_l))\ n_l + (F(u_r) - F(\tilde{u}_r))\ n_r.$$

Since f is Lipschitz, the right-hand side is bounded by $C\,(\|u_l - \tilde{u}_l\| + \|u_r - \tilde{u}_r\|)$. Thus (4.6) implies that

$$(|n_t| - K_1 |n_x|)\, \|u - \tilde{u}\| \ \leq C\ (\|u_l - \tilde{u}_l\| + \|u_r - \tilde{u}_r\|).$$

Appealing again to (4.4), this implies (4.5).

Step 2. We now prove the existence of solutions when $\mathcal{D} = \mathbb{R}^n$. We have already done this for triangles with a single inflow edge, so we consider triangles with two inflow edges, and we use the notation (4.1).

Define the set

$$S := \{(u_l, u_r) \in \mathcal{D} \times \mathcal{D} :\ \text{there exists a solution } u(u_l, u_r) \in \mathcal{D} \text{ of (4.1)}\}.$$

Note that S is nonempty, since $\mathcal{F}(u, u, u) = 0$ for every $u \in \mathcal{D}$, due to (3.22).

We next claim that S is open. Let $(u_l, u_r) \in S$, and suppose that $\mathcal{F}(u_l, u_r, u) = 0$. To prove that S contains an open neighborhood of (u_l, u_r), it suffices (via the implicit function theorem) to verify that $D_u \mathcal{F}(u_l, u_r, u)$ is nonsingular, where D_u denotes the gradient with respect to the u variable. Clearly

$$D_u \mathcal{F}(u_l, u_r, u) = DF(u)\ n = n_t\, I + n_x\, Df(u),$$

where I denotes the identity matrix. The eigenvalues of this matrix are $n_t + n_x \lambda^i(u)$. Since e is an outflow edge, the definition (3.7) states that these eigenvalues are all positive, and we conclude that S is open as claimed.

We next show that S is closed. Let $\{(u_{l,k}, u_{r,k})\} \subset S$ be a sequence, and suppose that $(u_{l,k}, u_{r,k}) \to (u_l, u_r)$. For each k, let u_k satisfy $\mathcal{F}(u_{l,k}, u_{r,k}, u_k) = 0$, and let $u = \lim_{k\to\infty} u_k$. This limit exists on account of (4.5), and the continuity of $\mathcal{F}$ implies that $\mathcal{F}(u_l, u_r, u) = 0$. Thus S is closed.

Since S is nonempty, open, and closed, it follows that S is all of $\mathcal{D}$, proving the existence assertion. □

Note that, in Step 2 of the above proof, the verification that S is open uses the fact that $\mathcal{D}$ is open, and similarly our proof that S is closed requires that $\mathcal{D}$ be closed. Thus the argument works only for $\mathcal{D} = \mathbb{R}^n$.

4.2. Discrete entropy condition. The main result of this section is the following discrete entropy condition, which we emphasize holds for general symmetrizable strictly hyperbolic systems.

THEOREM 4.2. *Consider a system of strictly hyperbolic conservation laws* (1.1) *with a domain of conservation states $\mathcal{D} \subset \mathbb{R}^n$, and assume that T is a causal triangle for $f, \mathcal{D}$. Let η be a strictly convex entropy for* (1.1)*, with associated entropy flux ψ. Then there exists a subdomain $\mathcal{D}' \subset \mathcal{D}$, depending on η, such that if $u \in \mathcal{D}'$ solves* (3.20) *with inflow data $\{u_{T_e} \in \mathcal{D}' : e \subset \partial T^-\}$, then*

$$\int_{\partial T^+} (\eta(u_T), \psi(u_T))\, \nu \, d\mathcal{H}^1 \; + \sum_{e \;\; inflow} \int_{\partial T^-} (\eta(u_{T_e}), \psi(u_{T_e}))\; \nu_{e,T} \; d\mathcal{H}^1 \;\leq\; 0. \tag{4.7}$$

The set $\mathcal{D}'$ is characterized in (4.14). In particular, we can take $\mathcal{D}' = \mathcal{D}$ if $\{D\eta(u) \;:\; u \in \mathcal{D}\}$ is convex.

Theorem 4.2 shows that the net entropy flux is given by an integral along the inflow boundary of a quantity that is pointwise positive. To verify this positivity we employ an idea that goes back to Harten and Lax [6] and perhaps earlier, integrating along a suitable path in configuration space; see also Osher [19] for related arguments, and [26], for example, for more recent developments and numerous references.

For the convenience of the reader, we give the complete proof, although parts of it essentially reproduce arguments from [6].

Proof. Fix T and (η, ψ) as in the statement of the theorem. Along an inflow edge $e \subset \partial T^-$ we use the notation $u_- := u_{T_e}$, $f_- := f(u_-)$, and so on. Also, (u, f) denotes $(u_T, f(u_T))$. Then we can rewrite (3.20) in the form

$$\int_{\partial T^-} (u^i, f^i)_-\, \nu \, d\mathcal{H}^1 \; + \int_{\partial T^+} (u^i, f^i)\, \nu \, d\mathcal{H}^1 = 0, \quad i = 1, \ldots, n. \tag{4.8}$$

Define auxiliary functions ξ and ζ by

$$(\xi, \zeta) := \eta_{u^i}(u^i, f^i) - (\eta, \psi), \quad u \in \mathcal{D}, \tag{4.9}$$

where subscripts denote partial differentiation.[3] Using (2.9), we deduce that

$$(\xi, \zeta)_{u^j} = \eta_{u^i u^j}(u^i, f^i) \tag{4.10}$$

for $j \in \{1, \ldots, n\}$. We claim that

$$\int_{\partial T^-} \big(\eta_{u^i}(u^i, f^i)_- - (\xi, \zeta)\big)\; \nu \, d\mathcal{H}^1 + \int_{\partial T^+} \big(\eta_{u^i}(u^i, f^i) - (\xi, \zeta)\big)\; \nu \, d\mathcal{H}^1 = 0. \tag{4.11}$$

[3]Note that in this proof ξ is not the same as the Godunov value defined in section 3.2.

Here (ξ,ζ) denotes $(\xi(u_T),\zeta(u_T))$, which is constant on T, so it is clear that $\int_{\partial T}(\xi,\zeta)\,\nu\,d\mathcal{H}^1 = 0$. The other terms in (4.11) are just the left-hand side of (4.8) multiplied by the constant $\eta_{u_i} = \eta_{u_i}(u_T)$ and implicitly summed over i. Thus (4.11) follows from (4.8). Using the definition (4.9) of (ξ,ζ), we infer from (4.11) that

$$\int_{\partial T^-}(\eta,\psi)_-\ \nu\,d\mathcal{H}^1 + \int_{\partial T^+}(\eta,\psi)\ \nu\,d\mathcal{H}^1 + E = 0, \tag{4.12}$$

where

$$\begin{aligned} E &= \int_{\partial T^-}\left(\eta_{u^i}(u^i,f^i)_- - (\xi,\zeta) - (\eta,\psi)_-\right)\ \nu\,d\mathcal{H}^1 \\ &= \int_{\partial T^-}\left[(\eta_{u^i}-\eta_{u^i,-})(u^i,f^i)_- - (\xi,\zeta) + (\xi,\zeta)_-\right]\ \nu\,d\mathcal{H}^1. \end{aligned}$$

To establish the discrete entropy inequality (4.7), we need to show $E \geq 0$.

Let E_p denote the integrand in E at some fixed point $p \in \partial T^-$. We will prove that $E_p \geq 0$ at every p. To do this, let $u : [0,1] \to \mathcal{D}$ be a path (to be chosen later) such that $u(0) = u_-(p)$ and $u(1) = u(p) = u_T$. Then, using the fundamental theorem of calculus and (4.10),

$$\begin{aligned} E_p &= \int_0^1 \frac{d}{ds}\left\{\eta_{u^i}(u(s))\,(u^i,f^i)_- - (\xi,\zeta)(u(s))\right\}\ \nu\,ds \\ &= \int_0^1 \eta_{u^iu^j}(u(s))\left\{(u^i,f^i)_- - (u^i(s),f^i(u(s)))\right\}\ \nu\,\dot{u}^j(s)\,ds \\ &= \int_0^1 \eta_{u^iu^j}(u(s))\left\{-\int_0^s \frac{d}{dr}(u^i(r),f^i(u(r)))dr\right\}\ \nu\,\dot{u}^j(s)\,ds \\ &= \int_0^1\int_0^s \eta_{u^iu^j}(u(s))\ \left(-\delta_{ik},-f^i_{u_k}(u(r))\right)\ \nu\ \dot{u}^j(s)\,\dot{u}^k(r)\,dr\,ds. \end{aligned}$$

We write $D^2\eta(s) = D^2\eta(u(s))$ and $A(r) = -\nu_t\,I - \nu_x\,Df(u(r))$, where I is the identity matrix. In this notation the integrand above has the form

$$\dot{u}^t(s)\ D^2\eta(s)\ A(r)\ \dot{u}(r). \tag{4.13}$$

A difficulty in analyzing this expression arises from the fact that $D^2\eta$ and $A = -\nu_t\,I - \nu_x\,Df$ are evaluated at distinct points $u(r), u(s)$ and therefore need not be related in any special way. We eliminate this problem by a suitable choice of the path $u(\cdot)$. We define

$$D\eta(u(s)) := s\,D\eta(u_T) + (1-s)\,D\eta(u_-(p)).$$

Since η is strictly convex, $D\eta$ is a bijection onto its image, and so $u(\cdot)$ is well defined as long as the right-hand side above lies in the image of $D\eta$. We therefore select $\mathcal{D}'$ to be any subset of $\mathcal{D}$ such that

$$\mathrm{co}\left(\{D\eta(u) : u \in \mathcal{D}'\}\right) \subset \{D\eta(u) :\ u \in \mathcal{D}\}, \tag{4.14}$$

where $\mathrm{co}(\{\cdots\})$ denotes the convex hull. Then $u(s)$ is well defined, and clearly $u(0) = u_-(p), u(1) = u_T$, as required. By differentiating with respect to s, we find that

$$D^2\eta(s)\ \dot{u}(s) = D\eta(u_T) - D\eta(u_-(p))\ =:\ \gamma.$$

Thus, for this choice of $u(s)$, the expression in (4.13) becomes

$$\dot{u}^t(s)\ D^2\eta(s)\ A(r)\ \dot{u}(r) = \gamma^t\ A(r)\,(D^2\eta(r))^{-1}\ \gamma. \tag{4.15}$$

Note that only the variable r appears on the right-hand side. We conclude the proof by showing that

$$\gamma^t A(r)(D^2\eta(r))^{-1}\gamma \geq 0 \tag{4.16}$$

for every $\gamma \in \mathbb{R}^n$ and every $r \in [0,1]$. We fix r and write simply A and $D^2\eta$. Let r_i denote the ith right eigenvector of $Df(u(r))$, and note that r_i is also an eigenvector for $A(r)$, with eigenvalue $-\nu_t - \nu_x\lambda^i(u(r)) =: \mu^i > 0$, using the inflow assumption (3.6). Let $\gamma_i := D^2\eta\, r_i$, $i = 1,\dots,n$. Then for all i,j,

$$\gamma_j^t\ A\,(D^2\eta)^{-1}\ \gamma_i = r_j^t\ D^2\eta\, A\ r_i = \mu^i(r_j^t\ D^2\eta\ r_i) = \mu^j(r_i^t\ D^2\eta\ r_j)^t.$$

The last identity follows from the fact that $D^2\eta\, A$ is symmetric, which is a consequence of the necessary condition (2.10) for the entropy η. (This is where we use the fact that $(D^2\eta(\cdot))^{-1}$ and $A(\cdot)$ are evaluated at *the same* point r on the right-hand side of (4.15).) Recalling the convexity of η and noting that the eigenvalues of A are distinct (on account of the strict hyperbolicity of (1.1)), we conclude that

$$\gamma_j^t\ A\,(D^2\eta)^{-1}\gamma_i \geq 0 \quad \text{if } i = j, \qquad \gamma_j^t\ A\,(D^2\eta)^{-1}\gamma_i = 0 \quad \text{if } i \neq j.$$

Since $\{\gamma_i\}_{i=1}^n$ form a basis for $\mathbb{R}^n$, this is easily seen to establish (4.16). ∎

5. A single causal triangle. In this section we prove the part of Theorem 3.3 that deals with a single causal triangle, in particular the existence of a unique solution satisfying the Riemann invariant bounds (3.25) in the causal case. In some sense this is the main point of our analysis; we will deduce (3.25) for causal patches from the corresponding estimate for a single causal triangle, and compactness and convergence will then be relatively easy consequences.

The proof uses induction on n, the number of equations in the system. In section 5.1 we consider case $n = 1$ of scalar equations. Some properties of Temple class systems that are used in the induction argument are given in sections 5.2 and 5.3, and the induction argument is carried out in section 5.4.

5.1. Scalar conservation laws. In this section we prove the following.

LEMMA 5.1. *Consider a* scalar *conservation law* (1.1), (1.2), *and assume that the spatial flux function f is Lipschitz continuous in an open interval $\mathcal{D}$. Let T be a causal triangle for $f,\mathcal{D}$. Assume that we are given inflow data satisfying $u_{T_e} \in \mathcal{D}$ for all $e \subset \partial T^-$. Then there exists a unique $u \in \mathcal{D}$ satisfying the approximate equation* (3.20). *Moreover,*

$$\min_{e\subset\partial T^-} u_{T_e} \leq u \leq \max_{e\subset\partial T^-} u_{T_e}, \tag{5.1}$$

with both inequalities in (5.1) *strict unless* $\min_{e\subset\partial T^-} u_{T_e} = \max_{e\subset\partial T^-} u_{T_e}$.

This sort of local maximum principle is well known for Godunov finite volume methods, and more generally for monotone schemes for scalar conservation laws; such results date back to [12], [4]. Similar results are established in [16] for Temple class systems. A lot of recent work has been devoted to constructing higher-order accurate schemes that enjoy a local maximum principle; see, for example, [17], [10]. Since we

do not know of any result exactly of this sort in the implicit spacetime setting that we consider here, we present the straightforward proof for the reader's convenience.

Proof. We first assume that T is a triangle with two inflow edges, e_l on the left and e_r on the right, and we use the notation as in (4.1). In view of (3.22) we can rewrite the approximation scheme (4.1) in the form

$$\mathcal{F}(u_l, u_r, u) = (F(u_l) - F(u))\ n_l + (F(u_r) - F(u))\ n_r = 0. \tag{5.2}$$

Recalling the definition (4.2) of F and fact that e_l is an inflow edge (see (3.6)), we obtain

$$\frac{d}{d\,z}\,(F(z)\ n_l) = n_{l,t} + n_{l,x}\,f'(z) = (1, f'(z))\ n_l < 0, \tag{5.3}$$

and likewise

$$\frac{d}{d\,z}\,(F(z)\ n_r) = (1, f'(z))\ n_r < 0, \tag{5.4}$$

for all $z \in \mathcal{D}$. Hence $u \mapsto \mathcal{F}(u_l, u_r, u)$ is strictly increasing on $\mathcal{D}$. If $u_l = u_r$, it follows that $u = u_l = u_r$ is the unique solution of (5.2). If $u_l \neq u_r$, then let us write $u_* = \min(u_l, u_r), u^* = \max(u_l, u_r)$. It follows from (5.2), (5.3), (5.4) that

$$\mathcal{F}(u_l, u_r, u_*) < 0, \qquad \mathcal{F}(u_l, u_r, u^*) > 0.$$

Thus, again by monotonicity, there exists a unique $u \in (u_*, u^*)$ solving (5.2).

The proof for a triangle with only one inflow edge is similar, but easier. ∎

5.2. Some facts about Temple class systems. The main result of this section is Lemma 5.3, which makes precise the statement that, if in a genuinely nonlinear Temple class system of n equations we fix k coordinates in the coordinate system provided by Riemann invariants, then the system reduces to a genuinely nonlinear Temple class system of $n - k$ equations. This is well known, but we have not seen the exact statement we prove (which is needed in the next section) anywhere in the literature. In order to prove it we need the following.

LEMMA 5.2. *Suppose that* (1.1) *is a genuinely nonlinear Temple class system with domain of conservation states* $\mathcal{D}$. *Then for every* $i \in \{1, \dots, n\}$ *and every constant* $c \in \mathbb{R}$ *there exist nonzero row vectors* $p \in M^{1\times n}$ *and* $q \in M^{1\times 2}$ *such that*

$$p\ F(u) = q \quad \text{for all} \quad u \in b^{-1}\left(\{w \in \mathcal{R} : w^i = c\}\right), \tag{5.5}$$

where $F(u) = (u, f(u))$ *and* b *denotes the diffeomorphism, defined in* (2.3), *from conservation states* u *onto Riemann invariants* w.

Proof. Fix some $i \in \{1, \dots, n\}$ and $c \in \mathbb{R}$, and let

$$\mathcal{C} := b^{-1}\left(\{w \in \mathcal{R} : w^i = c\}\right).$$

The assertion that there exists a nonzero row vector p and a number q_1 such that $pu = q_1$ for all $u \in \mathcal{C}$ is simply the defining attribute (2.6) of genuinely nonlinear Temple class systems.

Since p is orthogonal to $\mathcal{C}$, which is a level set of b^i, it is clear that p is parallel to Db^i on $\mathcal{C}$. However, (2.3) asserts that $Db^i(u)$ is a left eigenvector of $Df(u)$, with eigenvalue $\lambda^i(u)$. Thus p is a left eigenvector of $Df(u)$ for every $u \in \mathcal{C}$. As a result,

$p\ r_j(u) = 0$ for $j \neq i$. It follows that $\{r_j(u) : j \neq i\}$ spans the tangent space to $\mathcal{C}$ at any point $u \in \mathcal{C}$. However, for $j \neq i$ and $u \in \mathcal{C}$,

$$0 = p\ Df(u)\ r_j(u) = D(p\ f(u))\ r_j(u),$$

which asserts that the directional derivative of $p\ f(u)$ in the $r_j(u)$ direction vanishes for $j \neq i$. It follows that all tangential derivatives of $p\ f(u)$ along $\mathcal{C}$ vanish, so that $p\ f(u)$ is constant on $\mathcal{C}$. ∎

We now present our main result of this section.

LEMMA 5.3. *Let* (1.1) *be a genuinely nonlinear Temple class system with domain of conservation states* $\mathcal{D}$ *satisfying* (2.7). *Define*

$$\mathcal{C} := b^{-1}(\{w \in \mathcal{R}\ :\ w^{i_1} = c_1, \ldots, w^{i_k} = c_k\})$$

for some $i_j \in \{1, \ldots, n\}$ *and* $c_j \in \mathbb{R}$. *Then there exist a domain of conservation states* $\overline{\mathcal{D}} \subset \mathbb{R}^{n-k}$, *an affine map* $L : \mathbb{R}^{n-k} \to \mathbb{R}^n$ *that maps* $\overline{\mathcal{D}}$ *onto* $\mathcal{C}$, *and a mapping* $\bar{f} : \overline{\mathcal{D}} \to \mathbb{R}^{n-k}$ *satisfying*

$$f(L\bar{u}) = L\bar{f}(\bar{u}) + K, \quad \bar{u} \in \overline{\mathcal{D}}, \tag{5.6}$$

for some constant vector $K \in \mathbb{R}^n$ *(depending on* i_j, c_j, $j = 1, \ldots, k$*) and such that for the domain of conservation states* $\overline{\mathcal{D}}$

$$\bar{u}_t + \bar{f}(\bar{u})_x = 0 \tag{5.7}$$

is a genuinely nonlinear Temple class system satisfying (2.7). *Furthermore, if* $\bar{\lambda}^j, j = 1, \ldots, n-k$, *denotes the eigenvalues of* $D\bar{f}$, *then*

$$\{\bar{\lambda}^j(\bar{u})\}_{j=1}^{n-k} = \{\lambda^j(L\bar{u})\}_{j \notin \{i_1, \ldots, i_k\}}, \quad \bar{u} \in \overline{\mathcal{D}}. \tag{5.8}$$

Using the lemma, it is easy to show that if $\bar{u} : (0, \infty) \times \mathbb{R} \to \overline{\mathcal{D}}$ is a solution of $\bar{u}_t + \bar{f}(\bar{u})_x = 0$, then $u := L\bar{u}$ solves (1.1) in the same sense. This is clear for classical solutions,

$$u_t + f(u)_x = (L\bar{u})_t + f(L\bar{u})_x = L\bar{u}_t + (L\bar{f}(\bar{u}) + K)_x = L(\bar{u}_t + \bar{f}(\bar{u})_x) = 0,$$

and can similarly be justified for weak solutions. Thus, informally, the lemma states that the system for u reduces in $\mathcal{C}$ to a smaller Temple class system for the new dependent variable $\bar{u}$. Note that this also implies that $\mathcal{C}$ is an invariant submanifold for (1.1), as mentioned in the introduction.

Proof. We give the proof in the case $k = 1$. The general case follows by a straightforward induction argument.

Let $p = (p_1, \ldots, p_n) \in M^{1 \times n}$ and $q = (q_1, q_2) \in M^{1 \times 2}$ be the vectors provided by Lemma 5.2, satisfying (5.5). For concreteness we assume that $p_n \neq 0$; this does not entail any loss of generality. Then for all $u \in \mathcal{C}$ we have

$$u^n = \frac{q_1}{p_n} - \frac{1}{p_n} \sum_{j=1}^{n-1} p_j\, u^j \quad \text{and} \quad f^n(u) = \frac{q_2}{p_n} - \frac{1}{p_n} \sum_{j=1}^{n-1} p_j\, f^j(u). \tag{5.9}$$

For $\bar{u} = (\bar{u}^1, \ldots, \bar{u}^{n-1})^t \in \mathbb{R}^{n-1}$ we define

$$L\bar{u} := \left(\bar{u}^1, \ldots, \bar{u}^{n-1}, \frac{q_1}{p_n} - \frac{1}{p_n} \sum_{j=1}^{n-1} p_j\, \bar{u}^j \right)^t. \tag{5.10}$$

In view of (5.9), the image of L contains $\mathcal{C}$. We next define

$$\overline{\mathcal{D}} := \left\{ \bar{u} \in \mathbb{R}^{n-1} : L\,\bar{u} \in \mathcal{C} \right\}, \tag{5.11}$$

and $\bar{f} := (\bar{f}^1, \dots, \bar{f}^{n-1}) : \overline{\mathcal{D}} \to \mathbb{R}^{n-1}$ is defined by

$$\bar{f}^j(\bar{u}) = f^j(L\bar{u}), \qquad j = 1, \dots, n-1. \tag{5.12}$$

The remainder of the proof simply consists of verifying that $L, \bar{f}, \overline{\mathcal{D}}$ have all the required properties.

First, (5.9) and the definitions of $L, \bar{f}$ imply that

$$f(L\bar{u}) - L\bar{f}(\bar{u}) = \left(0, \dots, 0, \frac{q_2 - q_1}{p_n}\right)^t,$$

so that (5.6) holds. Note that differentiation of (5.6) with respect to $\bar{u}$ gives

$$D_u f(L\bar{u})\; D_{\bar{u}} L = D_{\bar{u}} L\; D_{\bar{u}} \bar{f}(\bar{u}). \tag{5.13}$$

We define $\overline{\mathcal{R}} := \{\bar{w} \in \mathbb{R}^{n-1} \,:\, (\bar{w}^1, \dots, \bar{w}^{i-1}, c, \bar{w}^i, \dots, \bar{w}^{n-1}) \in \mathcal{R}\}$ and

$$\bar{b}(\bar{u}) := (b^1(L\bar{u}), \dots, b^{i-1}(L\bar{u}), b^{i+1}(L\bar{u}), \dots, b^n(L\bar{u})) \qquad \text{for } \bar{u} \in \overline{\mathcal{D}}. \tag{5.14}$$

It is clear that $\bar{b} : \overline{\mathcal{D}} \to \overline{\mathcal{R}}$ is a diffeomorphism and that $\overline{\mathcal{R}}$ is a rectangle in $\mathbb{R}^{n-1}$ such that (2.7) holds for $\overline{\mathcal{D}}$.

We next check that $\bar{b}$ provides a coordinate system of Riemann invariants. Let $j \in \{1, \dots, n-1\}$ and $\bar{u} \in \overline{\mathcal{D}}$ be fixed. By (5.14), there exists $s \in \{1, \dots, n\} \setminus \{i\}$ such that

$$\bar{b}^j(\bar{u}) = b^s(L\bar{u}), \tag{5.15}$$

which further gives

$$D_{\bar{u}} \bar{b}^j(\bar{u}) = D_u b^s(L\bar{u})\; D_{\bar{u}} L. \tag{5.16}$$

Since the original system (1.1) satisfies (2.3), we have that

$$D_u b^s(L\bar{u})\; D_u f(L\bar{u}) = \lambda^s(L\bar{u})\; D_u b^s(L\bar{u}), \tag{5.17}$$

and using (5.16) and (5.13), we obtain

$$D_{\bar{u}} \bar{b}^j(\bar{u})\, D_{\bar{u}} \bar{f}(\bar{u}) = \lambda^s(L\bar{u})\, D_{\bar{u}} \bar{b}^j(\bar{u}), \tag{5.18}$$

implying $\bar{\lambda}^j(\bar{u}) = \lambda^s(L\bar{u})$. Thus (5.7) admits a coordinate system of Riemann invariants, as claimed. Note that the above also implies that (5.8) holds, and moreover it follows from (5.8) that the system (5.7) is strictly hyperbolic.

To verify that (5.7) is a Temple class system, we must also check that (2.6) is satisfied. This is an easy consequence of the definitions; see [9] for more details.

Finally we verify that the system (5.7) is genuinely nonlinear. By arguing as in the verification of (5.18) above, one can check that for j and s as in (5.15), and for $\bar{u} \in \overline{\mathcal{D}}$ with $Lu =: u \in \mathcal{D}$, there exists a nonzero constant c such that $r_s(u) = cD_{\bar{u}} L\; \bar{r}_j(\bar{u})$ and, moreover, $D_u \lambda^s(u) r_s(u) = cD_{\bar{u}} \bar{\lambda}^j(\bar{u}) \bar{r}_j(\bar{u})$; we refer again to [9] for more details. Thus the desired conclusion follows from the fact that (1.1) is genuinely nonlinear. ∎

Remark 5.4. Suppose that (η, ψ) is an entropy-entropy flux pair for system (1.1). Define

$$\bar{\eta}(\bar{u}) := \eta(L\bar{u}) \quad \text{and} \quad \bar{\psi}(\bar{u}) := \psi(L\bar{u}) \tag{5.19}$$

for $\bar{u} \in \overline{\mathcal{D}}$. By the chain rule we have

$$D_{\bar{u}}\bar{\eta}(\bar{u}) = D_u\eta(L\bar{u})\, D_{\bar{u}}L \quad \text{and} \quad D_{\bar{u}}\bar{\psi}(\bar{u}) = D_u\psi(L\bar{u})\, D_{\bar{u}}L. \tag{5.20}$$

With (5.13), this implies $D_{\bar{u}}\bar{\eta}(\bar{u})\ D_{\bar{u}}\bar{f}(\bar{u}) = D_{\bar{u}}\bar{\psi}(\bar{u})$, which means that $(\bar{\eta}, \bar{\psi})$ is an entropy-entropy flux pair for system (5.7).

5.3. Invariant submanifolds for the approximation scheme. As mentioned several times, level sets of the Riemann invariants b^i (defined in (2.3)) form invariant submanifolds for Temple class systems. In this section we prove in effect that these same sets are also invariant submanifolds for the discretized Temple class system (3.20) on a single causal or semicausal triangle.

LEMMA 5.5. *Suppose that* (1.1) *is a genuinely nonlinear Temple class system in domain* $\mathcal{D}$. *Assume that* $n_i = (n_{i,t}, n_{i,x})^t$, $i = 1, 2, 3$, *are such that* $n_1 + n_2 + n_3 = 0$ *and that* $u_1, u_2, u_3 \in \mathcal{D}$ *satisfy the following equation:*

$$F(u_1)\ n_1 + F(u_2)\ n_2 + F(u_3)\ n_3 = 0. \tag{5.21}$$

If for some $i \in \{1, \dots, n\}$ *and some* $c \in \mathbb{R}$, $b^i(u_1) = b^i(u_2) = c$ *and the expression* $(1, \lambda^i(u))\ n^3$ *is of constant nonzero sign for all* $u \in \mathcal{D}$, *then* $b^i(u_3) = c$.

Proof. Let us define set $\mathcal{C} := b^{-1}\left(\{w \in \mathcal{R} : w^i = c\}\right)$, which, by (2.6), is contained in a hyperplane in $\mathcal{D}$. By Lemma 5.2 there exist matrices p and q such that $p\, F(u) = q$ for all $u \in \mathcal{C}$. Multiplying (5.21) on the left by p and using the assumption that $n_1 + n_2 + n_3 = 0$, we obtain

$$(p\ F(u_3) - q)\ n_3 = 0.$$

This implies that for every $v \in \mathcal{C}$ we have

$$p\,((u_3 - v),\ (f(u_3) - f(v)))\ n_3 = 0. \tag{5.22}$$

Define $v_c = b^{-1}(w_c) \in \mathcal{C}$, where $w_c \in \mathcal{R}$ satisfies $w_c^i = c$, $w_c^j = b^j(u_3)$ for all $j \neq i$. Let $v(s) = s\, u_3 + (1 - s)\, v_c$ for $s \in [0, 1]$. We claim that

$$Df(v(s)(u_3 - v_c) = \lambda_i(v(s))(u_3 - v_c) \qquad \text{for all } s \in [0, 1]. \tag{5.23}$$

This is essentially just the well-known fact that, for Temple class systems, integral curves are straight lines. We will first use (5.23) to complete the proof of the lemma, and then for the convenience of the reader we present a proof of (5.23).

Using (5.23), we obtain

$$\begin{aligned} f(u_3) - f(v_c) = \int_0^1 \frac{d}{ds} f(v(s))\, ds &= \int_0^1 Df(v(s))\ (u_3 - v_c)\, ds \\ &= \int_0^1 \lambda^i(v(s))\, ds\, (u_3 - v_c). \end{aligned}$$

Substituting this into (5.22), we find that

$$p\ (u_3 - v_c) \int_0^1 (1, \lambda^i(v(s)))\ n_3\, ds = 0. \tag{5.24}$$

Since $(1, \lambda^i(v(s)))\ n_3$ is of constant nonzero sign for all $s \in [0,1]$, we obtain that $p\ (u_3 - v_c) = 0$ and therefore $p\ u_3 = q_1$. Since $u_3 \in \mathcal{D}$, we conclude that $u_3 \in \mathcal{C}$, which completes the proof.

To prove (5.23), we first recall that the defining property (2.6) of Temple class systems states that for each j the set $b^{-1}(\{w \in \mathcal{R} \ : \ w^j = b^j(u_3)\})$ is a hyperplane. Since this set also contains u_3 and v_c, it must contain the line segment joining them. Thus all Riemann invariants $b^j(v(s))$, $j \neq i$, are constant for $s \in [0,1]$. It follows that

$$\frac{d}{ds} b(v(s)) = Db(v(s))\dot{v}(s) = (0, \dots, 0, \sigma_i(s), 0, \dots 0)^t$$

for some σ_i (in the ith component.) Hence

$$\dot{v}(s) = \sigma_i(s) \quad \big(\text{the } i\text{th column of } Db^{-1}(v(s))\ \big).$$

However, according to (2.3), the ith column of $Db^{-1}(v(s))$ is a multiple of the normalized right eigenvector $r^i(v(s))$. Since $\dot{v}(s) = u_3 - v_c$, this proves (5.23). ∎

5.4. Well-posedness and Riemann invariant bounds. The central part of the proof of Theorem 3.3 is carried out in the following.

LEMMA 5.6. *Let* (1.1) *be a strictly hyperbolic, genuinely nonlinear Temple class system of conservation laws, and assume that the domain of conservation states* $\mathcal{D} \subset \mathbb{R}^n$ *satisfies* (2.7). *Let* T *be a causal triangle for* $f, \mathcal{D}$. *Then for any inflow data* $\{u_{T_e} \in \mathcal{D} \ : \ e \subset \partial T_-\}$ *there exists a unique solution* u *of* (3.20) *satisfying*

$$\min_{e \subset \partial T^-} w^i_{T_e} \ \leq \ w^i_T \ \leq \ \max_{e \subset \partial T^-} w^i_{T_e}, \qquad \text{for } w^i := b^i(u),$$

(5.25) *with both inequalities strict unless equality holds throughout,*

for all $i = 1, \dots, n$.

Similar local maximum principles are proved in [16] and [22] for the Godunov scheme for Temple class systems of two equations.

Proof. The theorem is obvious for triangles with only one inflow edge, so we consider only triangles with two inflow edges, say e_l and e_r on the left and right, respectively, and we write (3.20) in terms of inflow data u_l and u_r as in (4.1).

We will prove the existence of a solution satisfying (5.25) by induction on n, the number of equations in the system. (Uniqueness will be established at the end of the proof). The case $n = 1$ is the case of a scalar equation with strictly convex or concave spatial flux f; this is covered by results of section 5.1.

To do the induction step, let us assume that the conclusion holds for all genuinely nonlinear Temple class systems of $n-1$ equations, and let (1.1) be a genuinely nonlinear Temple class system of n equations. Let us define

$$S := \{(u_l, u_r) \in \mathcal{D} \times \mathcal{D} : \text{ there exists } u \in \mathcal{D} \text{ satisfying (4.1) and (5.25)}\}.$$

We will show $S = \mathcal{D} \times \mathcal{D}$.

Step 1. We prove that for every $i \in \{1, \dots, n\}$ and every constant $c \in \mathbb{R}$,

$$\{(u_l, u_r) \in \mathcal{D} \times \mathcal{D} : b^i(u_l) = b^i(u_r) = c\} \subset S.$$

Fix some $c \in \mathbb{R}$ and $i \in \{1, \dots, n\}$, and let $\mathcal{C} = b^{-1}(\{w \in \mathcal{R} \ : \ w^i = c\})$. According to Lemma 5.3, there exists a set $\overline{\mathcal{D}} \subset \mathbb{R}^{n-1}$ and maps $\bar{f} : \overline{\mathcal{D}} \to \mathbb{R}^{n-1}$ and $L : \mathcal{D} \to \mathcal{C}$ such that (5.6), (5.7), and (5.8) hold. It follows from (5.8) that any triangle that is

causal for $f, \mathcal{D}$ is also causal for $\bar{f}, \overline{\mathcal{D}}$. Thus the induction hypothesis and (5.7) imply that there exists a solution $\bar{u}$ of the equation

$$(\bar{u}_l, \bar{f}(\bar{u}_l))\ n_l + (\bar{u}_r, \bar{f}(\bar{u}_r))\ n_r + (\bar{u}, \bar{f}(\bar{u}))\ n = 0, \tag{5.26}$$

where $\bar{u}_l, \bar{u}_r \in \overline{\mathcal{D}}$ and $L\bar{u}_l = u_l$ and $L\bar{u}_r = u_r$. In addition, $\bar{u}$ satisfies the analogue of (5.25). Let $u := L\bar{u} \in \mathcal{C}$. Then, applying L to (5.26) and recalling from (5.6) that $L\bar{f}(\bar{u}) = f(L\bar{u}) + k$ with k constant, we deduce that

$$\begin{aligned} 0 &= (u_l, f(u_l) + k)\ n_l + (u_r, f(u_r) + k)\ n_r + (u, f(u) + k)\ n \\ &= (u_l, f(u_l))\ n_l + (u_r, f(u_r))\ n_r + (u, f(u))\ n. \end{aligned}$$

We have used (3.22) for the second equality. Also, it is clear from the construction in Lemma 5.3 that L preserves Riemann invariants, and hence that u satisfies (5.25).

Step 2. Let us define

$$\tilde{S} := \{(w_l, w_r) \in \mathcal{R} \times \mathcal{R} : w_l = b(u_l), w_r = b(u_r) \text{ for some } (u_l, u_r) \in S\}.$$

To prove the existence of a solution satisfying (5.25), it suffices to show that $\tilde{S} = \mathcal{R} \times \mathcal{R}$.

Define

$$Q := \{(w_l, w_r) \in \mathcal{R} \times \mathcal{R} : w_r^i > w_l^i \text{ for all } i\}.$$

We will show that

$$Q \subset \tilde{S}. \tag{5.27}$$

Exactly the same arguments can be used to show that

$$\{(w_l, w_r) \in \mathcal{R} \times \mathcal{R} : \pm w_r^i > w_l^i \text{ for all } i\} \subset \tilde{S}$$

for any choice of signs, and so the proof of (5.27) will show that

$$\{(w_l, w_r) \in \mathcal{R} \times \mathcal{R} : w_r^i \neq w_l^i \text{ for all } i\} \subset \tilde{S}.$$

In view of Step 1, the last inclusion implies that $\mathcal{R} \times \mathcal{R} \subset \tilde{S}$. Hence the proof of (5.27) will complete the proof of the existence of a solution satisfying (5.25).

We break the proof of (5.27) into three parts. For all three parts, it is convenient to write (4.1) in Riemann invariant coordinates, in which it takes the form

$$\tilde{F}(w_l)\ n_l + \tilde{F}(w_r)\ n_r + \tilde{F}(w)\ n = 0, \tag{5.28}$$

where $\tilde{F}(w) := F(a(w)) = (a(w), \tilde{f}(w))$ for $w \in \mathcal{R}$. Here we are writing $a = b^{-1}$ and using notation from (2.4).

Step 2a. The set $\tilde{S} \cap Q$ is open in Q.

This follows from exactly the same argument given in Step 2 of the proof of Theorem 4.1. The point is that the outflow condition (3.7) on edge e implies that $D(\tilde{F}(w)\ n)$ is nonsingular for all $w \in \mathcal{R}$, so that if (w_l, w_r, w) solves (5.28), then the implicit function theorem asserts that for w_l', w_r' sufficiently near w_l, w_r, there is a unique solution w' near w, depending smoothly on w_l', w_r'. If $(w_l, w_r) \in Q$, then strict inequality holds in (5.25) for all i, and so strict inequality will also hold in (5.25) for all (w_l', w_r', w') in a neighborhood of (w_l, w_r, w).

Step 2b. The set $\tilde{S} \cap Q$ is nonempty.

Let $w_l \in \mathcal{R}$ be fixed. In view of the considerations in Step 2a, for w_r sufficiently close to w_l there exists a unique w in a neighborhood of w_l such that (5.28) is satisfied. We write $w(w_r)$ to denote this solution w. We must show that there exist w_r such that the solution $w(w_r)$ satisfies (5.25).

We claim that at the point $w = w_r = w_l$, the matrix $\left(\frac{\partial w}{\partial w_r}\right)$ is diagonal, with all entries $\alpha_i, i = 1 \dots, n$, positive and strictly less than 1.

The claim asserts that $w^i(w_r) = w_l^i + \alpha_i(w_r^i - w_l^i) + o(\|w_l - w_r\|)$ as $w_r \to w_l$, which since $0 < \alpha_i < 1$ for all i, implies that $w(w_r)$ satisfies (5.25) for all w_r of the form $w_l + (h, \dots, h)$ when h is positive and sufficiently small. Thus the conclusion of Step 2a follows from the claim about the form of $\partial w/\partial w_r$.

To prove this claim, we differentiate (5.28) with respect to w_r to obtain

$$\Big\{n_{r,x}\, D\tilde{f}(w_r) \,+\, n_{r,t}\, Da(w_r)\Big\} + \Big\{n_x\, D\tilde{f}(w) \,+\, n_t\, Da(w)\Big\}\left(\frac{\partial w}{\partial w_r}\right) = 0.$$

Multiplying on the left by $(Da(w))^{-1}$ and using (2.5), we find that

$$Da(w)^{-1}\; Da(w_r)\; \Big\{n_{r,x}\, \tilde{\Lambda}(w_r) \,+\, n_{r,t}\, I\Big\} + \Big\{n_x\, \tilde{\Lambda}(w) \,+\, n_t\, I\Big\}\left(\frac{\partial w}{\partial w_r}\right) = 0.$$

We now set $w_l = w_r = w$. Then $Da(w)^{-1}\; Da(w_r) = I$. Due to the inflow and outflow conditions (3.6) and (3.7), for every $i \in \{1, \dots, n\}$ we have $(1, \tilde{\lambda}^i(w_r))\; n_r < 0$ and $(1, \tilde{\lambda}^i(w))\; n > 0$. Thus at the point in question,

$$\left(\frac{\partial w}{\partial w_r}\right) = -\begin{bmatrix} \frac{(1,\tilde{\lambda}^1(w_l))\; n_r}{(1,\tilde{\lambda}^1(w_l))\; n} & \cdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \cdots & \frac{(1,\tilde{\lambda}^n(w_l))\; n_r}{(1,\tilde{\lambda}^n(w_l))\; n} \end{bmatrix},$$

and all the diagonal elements are positive. To show that $\frac{\partial w^i}{\partial w_r^i} < 1$, we need to show $-(1, \tilde{\lambda}^i(w_l))\; n_r < (1, \tilde{\lambda}^i(w_l))\; n$, which, in view of (3.22), is equivalent to showing $(1, \tilde{\lambda}^i(w_l))\; n_l < 0$. This is satisfied, again as a result of the inflow constraint (3.6).

Step 2c. The set $\tilde{S} \cap Q$ is closed in Q.

Let $(w_{l,k}, w_{r,k}) \in \tilde{S} \cap Q$, for $k \in \mathbb{N}$, and suppose $\lim_{k\to\infty}(w_{l,k}, w_{r,k}) = (w_l, w_r) \in Q$. We need to show $(w_l, w_r) \in \tilde{S}$.

For each $k \in \mathbb{N}$, there exists a unique $w_k \in \mathcal{R}$ such that

$$\tilde{F}(w_{l,k})\; n_l + \tilde{F}(w_{r,k})\; n_r + \tilde{F}(w_k)\; n = 0 \tag{5.29}$$

and

$$w_{l,k}^i \;<\; w_k^i \;<\; w_{r,k}^i \quad \text{for } i \in \{1, \dots, n\}. \tag{5.30}$$

From (5.30), we deduce that there exists a convergent subsequence $\{w_{k_m}\}_{k_m}$. Let us denote $w = \lim_{k_m\to\infty} w_{k_m}$. From (5.29) and (5.30) we get that

$$\tilde{F}(w_l)\; n_l + \tilde{F}(w_r)\; n_r + \tilde{F}(w)\; n = 0$$

and

$$w_l^i \;\le\; w^i \le\; w_r^i \quad \text{for } i \in \{1, \dots, n\}. \tag{5.31}$$

Since $(w_l, w_r) \in Q$, we have $w_l^i \neq w_r^i$ for all i. Suppose that at least one of the inequalities in (5.31) is not strict. Then either $w_l^i = w^i$ or $w_r^i = w^i$. By Lemma 5.5, we must have $w_l^i = w^i = w_r^i$, which is contradiction to the assumption. Hence, $(w_l, w_r) \in \tilde{S}$.

Step 3. Finally we prove that there can be only one solution of (3.20) that satisfies (5.25).

Suppose that

$$\tilde{F}(w_l)\ n_l + \tilde{F}(w_r)\ n_r + \tilde{F}(w_k)\ n = 0$$

for $k = 1, 2$, and that both w_1 and w_2 satisfy (5.25). Define $w_r(s) = w_r + s(w_l - w_r)$. The arguments of Step 2 imply that there exist functions $w_k(s), 0 \leq s \leq 1$, such that $s \mapsto w_k(s)$ is C^1,

$$w_k(0) = w_k, \qquad (w_l, w_r(s), w_k(s)) \quad \text{satisfies (5.28) and (5.25)}$$

for $k = 1, 2$. Clearly $w_r(1) = w_l$, and so (5.25) implies that $w_1(1) = w_2(1) = w_l$. Let $S_1 := \{s \in [0,1] : w_1(s) = w_2(s)\}$. We have just shown that S_1 is nonempty. The continuity of w_1, w_2 implies that S_1 is closed, and the same implicit function theorem arguments used above demonstrate that S_1 is relatively open in $[0,1]$. It follows that $S_1 = [0,1]$, and hence that $w_1 = w_1(0) = w_2(0) = w_2$. This completes the proof. ∎

6. A single causal patch. The main result of this section is Lemma 6.2, which establishes those parts of Theorem 3.3 that describe the unique solvability, with bounds on the Riemann invariants, of the approximation scheme (3.3) on a single causal patch. The idea of the proof is to show that on such a patch the individual triangles in fact decouple, in that we can determine a priori the values of the Godunov flux across interior edges, in terms of only the patch geometry and inflow data; then the problem on each triangle can be reduced to the causal case studied earlier.

We first obtain a formula for the Godunov flux that is valid in particular for semicausal edges, such as those that occur in the interior of causal patches.

LEMMA 6.1. *Assume that* (1.1) *is a strictly hyperbolic, genuinely nonlinear Temple class system for which the domain of conservation states* $\mathcal{D} \subset \mathbb{R}^n$ *satisfies* (2.7)*, and suppose also that*

$$\sup_{u \in \mathcal{D}} \lambda^i(u) < \inf_{u \in \mathcal{D}} \lambda^{i+1}(u) \quad \text{for all } i. \tag{6.1}$$

Let $\nu = (\nu_t, \nu_x)^t$ *be a unit vector such that* $\nu_x > 0$ *and satisfying* (3.10) *for some* $\alpha \in \{1, \ldots, n-1\}$*, i.e.,*

$$\lambda^\alpha(u) < -\frac{\nu_t}{\nu_x} < \lambda^{\alpha+1}(u) \quad \text{for all } u \in \mathcal{D}. \tag{6.2}$$

Then for any $u_l, u_r \in \mathcal{D}$*, the Godunov flux is given by* $g(u_l, u_r; \nu) = (\xi, f(\xi))\,\nu$*, where* $\xi = \xi(u_l, u_r; \nu)$ *is expressed in Riemann invariant coordinates by*

$$b(\xi) = (w_r^1, \ldots, w_r^\alpha, w_l^{\alpha+1}, \ldots, w_l^n) \tag{6.3}$$

and $w_l = b(u_l), w_r = b(u_r)$ *denote the Riemann invariants associated with* u_l, u_r*.*

Proof. To determine the Godunov flux $g(u_l, u_r; \nu)$ we need to solve the Riemann problem for (1.1) with initial data

$$u(0, x) = u_l \text{ if } x < 0 \quad \text{and} \quad u(0, x) = u_r \text{ if } x > 0.$$

The solution of the Riemann problem for Temple class systems satisfying (6.1) is described[4] in [22, section 2]. This solution has the form $u(t,x) = v(x/t)$ for some $v : \mathbb{R} \to \mathcal{D}$, which among other properties satisfies

$$v(s) = v_i \quad \text{for all } s \text{ in an interval } (a_i, b_i) \supseteq \left(\sup_{u\in\mathcal{D}} \lambda^i(u), \inf_{u\in\mathcal{D}} \lambda^{i+1}(u) \right),$$

where

$$b(v_i) = (w_r^1, \dots, w_r^i, w_l^{i+1}, \dots, w_l^n).$$

In particular, in view of (6.2) and definitions from section 3.2, this implies that $\xi_-(u_l, u_r; \nu) = \xi_+(u_l, u_r; \nu) = v_\alpha$, which is what we need to show. □

It is not clear how to determine $g(u_l, u_r; \nu)$ from the basic definition in terms of the Riemann problem if condition (6.1) is not satisfied, since as far as we know there is no standard solution of the Riemann problem in this situation. However, our results remain valid if we take (6.3) as the definition of $g(u_l, u_r; \nu)$ in cases when (6.2) holds but (6.1) does not.

We now fix some notation: Let P be a causal patch, minimal in the sense of (3.16), consisting of $k \geq 2$ triangles $T_1, \dots, T_k$. We label the triangles and edges as shown in Figure 3.[5]

Let u_l denote the inflow data on e_0, and u_r the inflow data on e_k, and let u_i denote the solution we seek on triangle T_i. On the patch P the system (3.3) can be rewritten as

$$-F(\xi_{j-1})\, n_{j-1} + F(u_j)\, n_j^o + F(\xi_j)\, n_j = 0, \quad j = 1, \dots, k, \tag{6.4}$$

where

$$\xi_j = \xi(u_j, u_{j+1}; \nu_j), \quad j \in \{1, \dots, k-1\}, \tag{6.5}$$

and where we use the notation

$$\xi_0 = u_l, \qquad \xi_k = u_r. \tag{6.6}$$

Here we use the notation

$$n_j = \nu_j\, \mathcal{H}^1(e_j), \quad n_j^o = \nu_j^o\, \mathcal{H}^1(e_j^o), \quad j = 0, \dots, k. \tag{6.7}$$

We now prove the following result.

[4]Serre discusses only systems of two equations, but the general case is exactly the same.

[5]Note that this is always possible. Indeed, by the definition of a causal patch (see (3.14)) and Lemma 3.2, each triangle T_i must have exactly one outflow edge, which we denote e_i^o. The non-outflow edges will be denoted e_j. The patch inflow boundary must be nonempty, by the argument of Lemma 3.2. We may therefore assume that some triangle, say T_1, has an inflow edge, which we denote e_0. Let us write e_1 to denote the third edge of T_1. This edge cannot be causal, since then $P' = T_1$ would be a proper causal subpatch, in violation of the minimality condition (3.16). Thus the edge must be part of the interior of P, and so there must exist another triangle, say T_2, such that $e_1 = T_1 \cap T_2$. Proceeding in this way, we see that the triangles can be labeled such that T_i and T_{i+1} intersect along an edge e_i for $i = 1, \dots, k-1$. The final triangle T_k can intersect only T_{k-1} (as all other edges of all other triangles are already accounted for), and so its final edge, which we denote e_k, must be inflow.

Let ν_i for $i = 1, \dots, k$ denote the outer unit normal to T_i along e_i or, equivalently, the inner unit normal to T_{i+1}, and let ν_0 denote the inner unit normal to T_1 along e_0. By reversing the ordering of the triangles if necessary, we can assume that

$$-\frac{\nu_{i,t}}{\nu_{i,x}} < -\frac{\nu_{i+1,t}}{\nu_{i+1,x}}, \quad i = 0, \dots, k-1.$$

After fixing notation in this way we arrive at Figure 3.

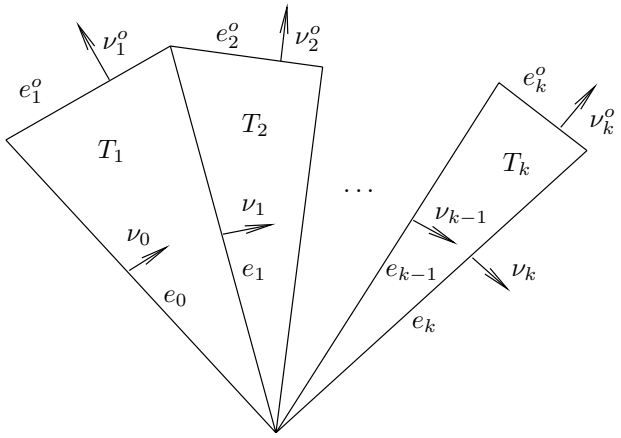


FIG. 3. *A causal patch of $k \geq 2$ triangles.*

LEMMA 6.2. *Let* (1.1) *be a strictly hyperbolic, genuinely nonlinear Temple class system, and assume that* (2.7) *holds for the domain of conservation states $\mathcal{D} \subset \mathbb{R}^n$. Let P be a causal patch of k triangles for $f, \mathcal{D}$ (see Figure* 3*). Then for any inflow data $u_l, u_r \in \mathcal{D}$, there exist unique $u_1, \dots, u_k \in \mathcal{D}$ satisfying* (6.4)–(6.6) *and*

(6.8)
$$\min\{w_{j-1}^i, w_{j+1}^i\} \leq w_j^i \leq \max\{w_{j-1}^i, w_{j+1}^i\} \quad \text{for } i \in \{1, \dots, n\}, j \in \{1, \dots, k\}.$$

Here we use the notation

$$w_0 := b(u_l), \quad w_{k+1} := b(u_r), \quad \text{and } w_j := b(u_j), \qquad j = 1, \dots, k. \tag{6.9}$$

Finally, the solution also satisfies

$$|w_{k+1}^i - w_0^i| = \sum_{j=1}^{k+1} |w_j^i - w_{j-1}^i|, \qquad i = 1, \dots, n. \tag{6.10}$$

Note that (6.8) is exactly conclusion (3.25) of the main Theorem 3.3.

Proof. We have fixed the orientation of the normals on the nonoutflow faces so that $\nu_{j,x} > 0$ for all j. Since e_j is semicausal for $j \in \{1, \dots, k-1\}$, there exist $\alpha_1, \dots, \alpha_{k-1}$ such that $\alpha_0 := 0 < \alpha_1 \leq \cdots \leq \alpha_{k-1} < \alpha_k := n$ such that

$$\lambda^{\alpha_j}(u) < -\frac{\nu_{j,t}}{\nu_{j,x}} < \lambda^{\alpha_j+1}(u) \tag{6.11}$$

for all $u \in \mathcal{D}$ and for all $j \in \{0, \dots, k\}$ (where the left-hand inequality for $j = 0$ and the right-hand inequality for $j = k$ are understood to hold trivially, since λ^0 and λ^{n+1} are not defined.)

Step 1. We first prove that if $u_1, \dots, u_k \in \mathcal{D}$ satisfy (6.4)–(6.6) for inflow data $u_l, u_r \in \mathcal{D}$, then necessarily

$$\xi_j = \xi(u_l, u_r; \nu_j), \quad j \in \{1, \dots, k-1\}, \tag{6.12}$$

where ξ_j is as defined in (6.5). Fix any j. Using the notation (6.9), the formula for the Godunov value $\xi(u_l, u_r; \nu_j)$ from Lemma 6.1, and (6.11), we can write

$$b(\xi_j) = (w^1_{j+1}, \dots, w^{\alpha_j}_{j+1}, w^{\alpha_j+1}_j, \dots, w^n_j). \tag{6.13}$$

In particular, in the coordinate system provided by Riemann invariants, the ith coordinates of ξ_j and u_j are equal for $i = \alpha_j + 1, \dots, n$. Thus, in view of (6.4) and the invariant submanifold property of Lemma 5.5, it must be the case that the same coordinates are shared by ξ_{j-1}. In view of (6.13) (with j replaced by $j-1$), this implies that $w^i_{j-1} = w^i_j$ for $i = \alpha_j + 1, \dots, n$. Since this holds for all j, and since $\alpha_{j-1} \le \alpha_j$, we deduce that

$$b^i(\xi_j) = b^i(\xi_0) = b^i(u_l) = w^i_0 \quad \text{for } i = \alpha_j + 1, \dots, n. \tag{6.14}$$

Similarly, from (6.13) we see that, in the coordinate system given by Riemann invariants, the ith coordinates of ξ_j and u_{j+1} are equal for $i = 1, \dots, \alpha_j$, and so using (6.4) and Lemma 5.5 again, we find that

$$b^i(\xi_j) = b^i(\xi_k) = b^i(u_r) = w^i_{k+1} \quad \text{for } i = 1, \dots, \alpha_j. \tag{6.15}$$

Again appealing to Lemma 6.1 to find an explicit formula for the Godunov value $\xi(u_l, u_r; \nu_j)$, we find that (6.14) and (6.15) together prove (6.12).

Step 2. In view of Step 1, the triangles in the causal patch in effect decouple, and on each triangle T_j, $j = 1, \dots, k$, we must show the existence of a unique solution (with Riemann invariant bounds) for the problem

$$-F(\xi_{j-1})\, n_{j-1} + F(u_j)\, n^o_j + F(\xi_j)\, n_j = 0, \tag{6.16}$$

where now all the values ξ_j are known:

$$\xi_j = \xi(u_l, u_r, \nu_j), \quad j \in \{1, \dots, k-1\}, \tag{6.17}$$

$$\xi_0 = u_l, \quad \xi_k = u_r. \tag{6.18}$$

Fix a triangle T_j. Let us define

$$\mathcal{C}_j := b^{-1}\left(\left\{ w \in \mathcal{R} \ : w^i = \begin{cases} w^i_r, & i = 1, \dots, \alpha_{j-1}, \\ w^i_l, & i = \alpha_j + 1, \dots, n. \end{cases} \right\} \right)$$

From (6.17) and Lemma 6.1 we see that $\xi_{j-1}, \xi_j \in \mathcal{C}_j$. Thus Lemma 5.5 implies that any solution $u_j \in \mathcal{D}$ of (6.16) must also satisfy $u_j \in \mathcal{C}_j$. If $\alpha_{j-1} = \alpha_j$, then we have $\xi_{j-1} = u_j = \xi_j$, and if $\alpha_{j-1} < \alpha_j$, we show that we can find a solution in $\mathcal{C}_j$ by reducing (1.1) to a smaller *causal* Temple class system.

Let $\ell = \alpha_j - \alpha_{j-1}$, so that $n - \ell$ is the number of coordinates in the coordinate system of Riemann invariants that are specified for elements of $\mathcal{C}_j$. By Lemma 5.3, there exist a domain of conservation states $\overline{\mathcal{D}} \subset \mathbb{R}^\ell$, an affine bijection $L : \overline{\mathcal{D}} \to \mathcal{C}_j$, and a mapping $\bar{f} : \overline{\mathcal{D}} \to \mathbb{R}^\ell$ such that

$$\bar{u}_t + \bar{f}(\bar{u})_x = 0 \tag{6.19}$$

is a genuinely nonlinear Temple class system for $\bar{u} \in \overline{\mathcal{D}}$, and

$$f(L\bar{u}) = L\bar{f}(\bar{u}) + K, \qquad \bar{u} \in \overline{\mathcal{D}}, \tag{6.20}$$

where $K \in \mathbb{R}^n$ is some constant. Furthermore, if $\bar{\lambda}^j, j = 1, \dots, \ell$, denotes the eigenvalues of $D\bar{f}$, then

$$\{\bar{\lambda}^j(\bar{u})\}_{j=1}^{\ell} = \{\lambda^j(L\bar{u})\}_{j=\alpha_{j-1}+1}^{\alpha_j}, \quad \bar{u} \in \overline{\mathcal{D}}. \tag{6.21}$$

We assert that edges e_{j-1} and e_j are inflow edges for T_j. Indeed, our choice of notation implies that $\nu_{j,x} > 0$ for all j, so that (6.11) implies that $\nu_{j,t} + \nu_{j,x}\lambda^i(u) < 0$ for all $i \le \alpha_j$ and all $u \in \mathcal{D}$. Also, ν_j is the outer normal to T_j along edge e_j, so e_j satisfies the definition (3.6) of an inflow edge for T_j. Similarly, $\nu_{j-1,t} + \nu_{j-1,x}\lambda^i(u) > 0$ for all $i \ge \alpha_{j-1} + 1$ and all $u \in \mathcal{D}$, and $-\nu_{j-1}$ is the outer normal to T_j along e_{j-1}, from which it follows that e_{j-1} is also an inflow edge for T_j.

Thus T_j is a causal triangle for $\bar{f}, \overline{\mathcal{D}}$, and so Lemma 5.6 implies that there exists a unique solution $\bar{u}_j$ of the equation

$$-(\bar{\xi}_{j-1}, \bar{f}(\bar{\xi}_{j-1}))\, n_{j-1} + (\bar{u}_j, \bar{f}(\bar{u}_j))\, n_j^o + (\bar{\xi}_j, \bar{f}(\bar{\xi}_j))\, n_j = 0, \tag{6.22}$$

where $\bar{\xi}_j \in \overline{\mathcal{D}}$ satisfies $L(\bar{\xi}_j) = \xi_j$, and similarly $\bar{\xi}_{j-1}$. This solution $\bar{u}_j$ also satisfies Riemann invariant bounds (5.25), and is the only solution with this property.

We claim that $u_j := L\bar{u}_j$ satisfies (6.16). Indeed, from (6.20), (6.22), and the fact that $n_j^o + n_{j-1} + n_j = 0$ (see (3.22)) we infer that

$$\begin{aligned}
-F(\xi_{j-1})\, n_{j-1} + F(u_j)\, n_j^o + F(\xi_j)\, n_j & \\
&= -(L\bar{\xi}_{j-1}, f(L\bar{\xi}_{j-1}))\, n_{j-1} + (L\bar{u}_j, f(L\bar{u}_j))\, n_j^o + (L\bar{\xi}_j, f(L\bar{\xi}_j))\, n_j \\
&= -(L\bar{\xi}_{j-1}, L\bar{f}(\bar{\xi}_{j-1}))\, n_{j-1} + (L\bar{u}_j, L\bar{f}(\bar{u}_j))\, n_j^o + (L\bar{\xi}_j, L\bar{f}(\bar{\xi}_j))\, n_j \\
&= -L(\bar{\xi}_{j-1}, \bar{f}(\bar{\xi}_{j-1}))\, n_{j-1} + L(\bar{u}_j, \bar{f}(\bar{u}_j))\, n_j^o + L(\bar{\xi}_j, \bar{f}(\bar{\xi}_j))\, n_j \\
&= 0.
\end{aligned}$$

Step 3. We next claim that u_j satisfies (6.8) and is the only solution of (6.16) with this property. The Riemann invariant bounds satisfied by $\bar{u}_j$ and properties of the map L imply that u_j satisfies

$$\min\{b^i(\xi_{j-1}), b^i(\xi_j)\} \le w_j^i \le \max\{b^i(\xi_{j-1}), b^i(\xi_j)\} \tag{6.23}$$

if $i \in \{\alpha_{j-1} + 1, \dots, \alpha_j\}$. From (6.13) we see that $b^i(\xi_{j-1}) = w_{j-1}^i$ for $i \ge \alpha_{j-1} + 1$, and $b^i(\xi_j) = w_{j+1}^i$ for $i \le \alpha_j$, so that (6.23) becomes (6.8) for $i \in \{\alpha_{j-1} + 1, \dots, \alpha_j\}$. If $i > \alpha_j$, then similarly from (6.14) we see that $b^i(\xi_j) = w_0^i = w_j^i$ (using the fact that $u_j \in \mathcal{C}_j$). Thus in this case (6.8) clearly holds. Similarly, in the case $i \le \alpha_{j-1}$, we deduce (6.8) by noting that $b^i(\xi_{j-1}) = w_{k+1}^i = w_j^i$.

Note in addition that if v_j is any solution of (6.16), then necessarily $v_j \in \mathcal{C}_j$, and by undoing the above arguments one can see that $v_j = L\bar{v}_j$ for some $\bar{v}_j \in \overline{\mathcal{D}}$ solving (6.22) and satisfying suitable Riemann invariant bounds. It follows that $\bar{v}_j = \bar{u}_j$, so that u_j is the only solution of (6.16) for which (6.8) holds.

Step 4. Finally, to prove (6.10), fix some $i \in \{1, \dots, n\}$. There exists some j such that $\alpha_{j-1} < i \le \alpha_j$. It follows from Step 2 (in particular the fact that $u_j \in \mathcal{C}_j$) that

$$w_m^i = w_0^i \quad \text{for } m = 1, \dots, j-1, \qquad w_m^i = w_{k+1}^i \quad \text{for } m = j+1, \dots, k.$$

Together with (6.8), this establishes (6.10). ∎

Remark 6.3. Suppose that all assumptions of the previous lemma hold. Let (η, ψ) be an entropy-entropy flux pair for (1.1), and let a set $\mathcal{D}'$ be characterized by (4.14).

By taking $\mathcal{D}'$ smaller, we can assume that it satisfies (2.7). Consider a triangle T_j, for some $j \in \{1, \dots, k\}$, within the patch, and let $\mathcal{C}_j, \overline{\mathcal{D}}$, and L be as in Step 2 of the previous proof. By Remark 5.4 we have that $(\bar{\eta}, \bar{\psi})$ defined by (5.19) forms an entropy-entropy flux pair for the Temple class system (6.19). Note that, by (5.20) and the fact that the map $L : \overline{\mathcal{D}} \to \mathcal{D}$ is linear, we have $\mathcal{D}' \cap \mathcal{C}_j \subset L(\overline{\mathcal{D}}')$.

By Theorem 4.2 for the system (6.19) and the causal triangle T_j for $\bar{f}, \overline{\mathcal{D}}$, we have that if $\bar{\xi}_{j-1}, \bar{\xi}_j, \bar{u}_j \in \overline{\mathcal{D}}'$, then

$$\int_{e_j^o} \left(\bar{\eta}(\bar{u}_j), \bar{\psi}(\bar{u}_j)\right) \nu_j^o \, d\mathcal{H}^1 + \int_{e_{j-1}} \left(\bar{\eta}(\bar{\xi}_{j-1}), \bar{\psi}(\bar{\xi}_{j-1})\right) (-\nu_{j-1}) \, d\mathcal{H}^1 + \int_{e_j} \left(\bar{\eta}(\bar{\xi}_j), \bar{\psi}(\bar{\xi}_j)\right) \nu_j \, d\mathcal{H}^1 \le 0.$$

Hence, if the inflow data on a causal patch P is such that $\{u_{T_e} \in \mathcal{D}' : e \subset \partial P^-\}$, then for every triangle T within P,

$$\int_{\partial T^+} (\eta(u_T), \psi(u_T)) \, \nu \, d\mathcal{H}^1 + \int_{e \text{ inflow}} (\eta(u_{T_e}), \psi(u_{T_e})) \, \nu_{e,T} \, d\mathcal{H}^1 + \sum_{e \text{ semicausal}} \int_e (\eta(\xi_e), \psi(\xi_e)) \, \nu_{e,T} \, d\mathcal{H}^1 \le 0,$$

where ξ_e denotes the Godunov value along a semicausal edge e of triangle T. Using (3.22) and letting ξ_e denote the inflow value u_{T_e} along an inflow edge, this can be written more concisely in the form

$$\sum_{e \text{ noncausal}} \int_e \left[(\eta(\xi_e), \psi(\xi_e)) - (\eta(u_T), \psi(u_T))\right] \nu_{e,T} \, d\mathcal{H}^1 \le 0. \tag{6.24}$$

Example 6.4. In this example we show that uniqueness of solutions of the approximation scheme can fail if both conditions (3.14) and (3.15) are violated. We do not know whether uniqueness can fail if either condition alone is not satisfied.

Let $P = T_1 \cup T_2 \cup T_3$, as in Figure 4. Consider Burgers' equation (scalar conservation law where $f(u) = \frac{u^2}{2}$) with $\mathcal{D} = [-a, a]$ for any $0 < a < \frac{4}{3}$. It is clear that every edge $e \subset \partial P$ is causal. Let u_i^- denote the inflow data along $T_i \cap \partial P^-$, where ∂P^- denotes the patch inflow boundary, and let u_i denote the approximate solution on T_i. Suppose that

$$u_1^- = 1, \quad u_2^- = 0, \quad \text{and} \quad u_3^- = -1.$$

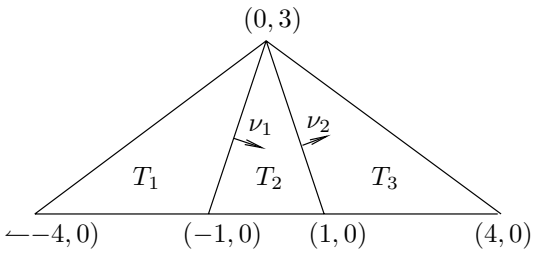


FIG. 4.

We claim that an approximate solution on P is given by

$$u_1 = 1, \quad u_3 = -1, \quad \text{and} \quad u_2 = \sigma, \quad \text{for arbitrary } \sigma \in \left(-\tfrac{1}{3}, \tfrac{1}{3}\right). \tag{6.25}$$

Fix $\sigma \in \left(-\frac{1}{3}, \frac{1}{3}\right)$. Let $e_i = T_i \cap T_{i+1}$, and let ν_i denote the outer unit normal to triangle T_i along the edge e_i for $i = 1, 2$. Note that

$$\xi_1 := \xi(1, \sigma; \nu_1) = 1. \tag{6.26}$$

Indeed, the entropy solution of the Riemann problem for the Burgers' equation with initial data $u(0, x) = 1$, $x < 0$, and $u(0, x) = \sigma$, $x > 0$, is a shock propagating with speed $s = \frac{1}{2}(1 + \sigma)$ and having values 1 and σ on the left and on the right, respectively. With the choice of σ we have $s > \frac{1}{3} = -\frac{\nu_{1,t}}{\nu_{1,x}}$, which implies (6.26). Similarly, $\xi_2 := \xi(\sigma, -1; \nu_2) = -1$. One can then check by an easy calculation that the equation is satisfied; indeed, this is almost obvious by symmetry. Hence, (6.25) is a solution for all $\sigma \in (-1/3, 1/3)$, as asserted.

7. Convergence and entropy inequalities. In this section we give the remainder of the arguments needed for the proof of Theorem 3.3. We also show at the end of the section, in Example 7.1, why the assumption (3.23) is needed.

Proof of Theorem 3.3. Recall that for this theorem we assume that $\mathcal{T}_h$ is a causal or a patchwise causal partition for $f, \mathcal{D}$, where f is the flux function for a genuinely nonlinear Temple class system and $\mathcal{D}$ satisfies (2.7). We also assume that $\mathcal{T}_h$ satisfies (3.23), (3.24), and (3.27), with constants independent of h. Since a causal triangulation is also patchwise causal (in which each causal patch consists of a single triangle), we consider only the case when $\mathcal{T}_h$ is patchwise causal.

Step 1. First, results in Lemma 6.2 about existence and uniqueness of solutions on a single causal patch (once the inflow data is known), together with remarks in section 3.4 about the patch-by-patch solution strategy, imply that there exists a unique solution $u_h \in \mathcal{P}_h$ of (3.3) satisfying the elementwise Riemann invariant bounds (3.25) on each triangle. (Recall that $\mathcal{P}_h$ denotes the space of piecewise constants on the triangulation $\mathcal{T}_h$.)

Step 2. Next, let u_{0h} be the discretization of the data u_0 implicit in the discrete initial condition (3.4), so that u_{0h} is constant on intervals $I \subset \mathbb{R}$ such that $\{t = 0\} \times I = (\partial T \cap \{t = 0\})$ for some $T \in \mathcal{T}_h^0$, and on each such interval

$$u_{0h} = \frac{1}{\mathcal{H}^1(I)} \int_I u_0 \, dx.$$

It is easy to check that for x in an interval I as above, $\|u_{0h}(x) - u_0(x)\| \le TV(u_0; I)$, where the right-hand side denotes the total variation of u_0 in the interval I. From this one can check that

$$\|u_{0h} - u_0\|_{L^1(\mathbb{R})} \to 0 \quad \text{as } h \to 0 \qquad \text{and} \tag{7.1}$$

$$TV(u_{0h}) \le 3\, TV(u_0) \quad \text{for all } h. \tag{7.2}$$

Step 3. Fix a patchwise causal partition $\mathcal{T}_h$, and let $u_h \in \mathcal{P}_h$ be the corresponding approximate solution satisfying local Riemann invariant bounds. Let $w_h = b(u_h)$ denote the approximate solution expressed in terms of the Riemann invariant coordinates, and similarly $w_{0h} = b(u_{0h})$. We next establish the claim (3.26) that $TV(w_h^i(t, \cdot))$ is nonincreasing in t for every i. The conclusion will follow only from

assumptions about the triangulation and from the local Riemann invariant bounds (6.10). We view this as *nearly* geometrically obvious, but we give a detailed proof for the sake of completeness.

Fix $i \in \{1, \dots, n\}$. To simplify the notation we drop the sub- and superscripts for this part of our argument and simply write w.

Our assumptions about the mesh (in particular, (3.24) and (3.27)) imply that there exists $\delta > 0$ such that for every T in $\mathcal{T}_h$ and every nontrivial edge $e \subset \partial T$,

$$\sup\{t : (t,x) \in e\} - \inf\{t : (t,x) \in e\} \geq \delta h. \tag{7.3}$$

We will show that

$$TV(w(\sigma,\cdot)) \geq TV(w(\tau,\cdot)) \quad \text{whenever } 0 \leq \sigma < \tau \leq \sigma + \delta h. \tag{7.4}$$

This clearly suffices to prove (3.26).

Fix $0 \leq \sigma < \tau < \sigma + \delta h$ and choose a sequence of points $\cdots < y_j < y_{j+1} < \cdots$ with the property that, for each triangle T such that $T^o \cap \{t = \tau\} \neq \emptyset$, there exists exactly one point $(\tau, y_j) \in T^o$. (Here T^o denotes the interior of T.) Then

$$TV(w(\tau,\cdot)) = \sum_j |w(\tau, y_j) - w(\tau, y_{j-1})|.$$

We will write T_j to denote the triangle containing the point (τ, y_j).

Next, for triangles $S, T \in \mathcal{T}_h$, we write $S \sim T$ if S and T are separated by a trivial edge, as defined immediately before (3.24). In particular, if $S \sim T$, then $w_S = w_T$. We define

$$I := \left\{ j \in \mathbb{Z} : \text{either } T_j^o \cap \{t = \sigma\} \neq \emptyset \text{ or } \exists S \text{ with } S \sim T_j \text{ and } S^o \cap \{t = \sigma\} \neq \emptyset \right\}.$$

We can index the elements of I by the integers, with $\cdots < j_{-1} < j_0 < j_1 < \cdots$. For $j_\ell \in I$, let

$$S_\ell := \begin{cases} T_{j_\ell} & \text{if } T_{j_\ell}^o \cap \{t = \sigma\} \neq \emptyset, \\ \text{the triangle } S \text{ such that } S \sim T_{j_\ell},\ S^o \cap \{t = \sigma\} \neq \emptyset & \text{if not,} \end{cases}$$

and for each ℓ let x_ℓ be a point such that $(\sigma, x_\ell) \in S_\ell^o$. The definitions imply that $x_\ell < x_{\ell+1}$ and that $w(\sigma, x_\ell) = w(\tau, y_{j_\ell})$ for all ℓ. Thus

$$TV(w(\sigma,\cdot)) \geq \sum_\ell |w(\sigma, x_\ell) - w(\sigma, x_{\ell-1})| = \sum_\ell |w(\tau, y_{j_\ell}) - w(\tau, y_{j_{\ell-1}})|.$$

Thus to prove (7.4) it suffices to show that

$$|w(\tau, y_{j_\ell}) - w(\tau, y_{j_{\ell-1}})| = \sum_{m=j_{\ell-1}+1}^{j_\ell} |w(\tau, y_m) - w(\tau, y_{m-1})| \tag{7.5}$$

when $j_{\ell-1}, j_\ell$ are adjacent points in I. To simplify the notation, let us assume that $j_\ell = 0$, and let us also write $k+1 := j_{\ell+1}$. We may assume that $k \geq 1$, as otherwise (7.5) is trivial.

We claim that if $j \notin I$, then T_j must have *exactly* one vertex, say Q_j, in $[\sigma, \tau) \times \mathbb{R}$. It is easy to see from the definition of I that T_j must have at least one such vertex.

And if T_j has two vertices in this slab, then (7.3) and the assumption that $\tau < \sigma + \delta h$ imply that the edge e joining these two vertices must be trivial, and so $T_j \sim T_{j,e}$. Again from (7.3) and the choice of σ, τ, it follows that $T^o_{j,e} \cap \{t = \sigma\} \neq \emptyset$, which would imply that $j \in I$.

Next note that for $j = 1, \dots, k-1$, T_j and T_{j+1} intersect along an edge that terminates at a shared vertex of T_j, T_{j+1} in the set $(0, \tau) \times \mathbb{R}$, and hence in the slab $[\sigma, \tau) \times \mathbb{R}$. It follows that $Q_1 = Q_2 = \cdots = Q_k =: Q$. Thus triangles $T_1, \dots, T_k$ appear exactly as pictured in Figure 3. For the duration of this discussion, we adopt the notation for edges and normals displayed in Figure 3. Assumption (3.23) implies that e^o_j is outflow for $j = 1, \dots, k$, as in Figure 3.

Now consider triangle T_1. We claim that e_0 is an inflow edge for T_1, or equivalently an outflow edge for T_0. To prove this, recall that either $T^o_0 \cap \{t = \sigma\} \neq \emptyset$ or $T_0 \sim S_\ell$ for some S_ℓ such that $S^o_\ell \cap \{t = \sigma\} \neq \emptyset$. In the former case, the claim then follows from assumption (3.23). In the latter case, it follows from the fact that (by definition of $\sim$) the edge $T_0 \cap S_\ell$ is a trivial edge, which (by the definition of a trivial edge) means that the other edges of T_0 are outflow.

The same considerations show that e_k is an inflow edge for T_k. Now let $k_* = \max\{j : e_j \text{ is inflow for } T_{j+1}\}$ and $k^* = \min\{j : e_{j+1} \text{ is inflow for } T_{j+1}\}$, and note that $k_* < k^*$. It is now straightforward to check the following.

First, if $k_* > 1$, then edges $e_0, \dots, e_{k_*-1}$ are all trivial, and as a result $w(\tau, y_0) = \cdots = w(\tau, y_{k_*-1})$. Similarly, if $k^* < k$, then $w(\tau, y_{k^*+1}) = \cdots = w(\tau, y_{k+1})$.

Second, $T_{k_*} \cup \cdots \cup T_{k^*}$ form a causal patch (consisting possibly of a single triangle), and so (6.10) implies that

$$|w_{k^*+1} - w_{k_*-1}| = \sum_{j=k_*}^{k^*+1} |w_j - w_{j-1}|.$$

Combining the last observations, we obtain (7.5), so this completes the proof of (3.26).

Step 4. We next show that there exists a constant C, depending only on $TV(u_0)$ and the constants K_0 from (2.8) and α from (3.24), such that for any $\tau > 0$,

$$\int_0^\tau \int_{\mathbb{R}} (|u_{h,x}| + |u_{h,t}|) \;\le\; C\,\tau. \tag{7.6}$$

Here $|u_{h,x}|$ and $|u_{h,t}|$ are understood as measures on $(0, \infty) \times \mathbb{R}$. To prove (7.6), first note that, for any $t > 0$,

$$\begin{aligned} TV(u_h(t, \cdot)) \le K_0\, TV(w_h(t, \cdot)) \le K_0 \sum_{i=1}^n TV(w^i_h(t, \cdot)) &\le K_0 \sum_{i=1}^n TV(w^i_{0h}) \\ &\le n\, K_0^2\, TV(u_{0h}), \end{aligned}$$

using (2.8) and (3.26). Thus $\int_0^\tau \int_{\mathbb{R}} |u_{h,x}| = \int_0^\tau TV(u_h(t, \cdot))\, dt \le C\,\tau$, by (7.2) and the assumption that $TV(u_0)$ is finite. Also, (3.24) together with standard facts about functions of bounded variation show that

$$\int_A |u_{h,t}| \;\le\; \frac{1}{\alpha} \int_A |u_{h,x}|$$

for every $A \subset [0, \infty) \times \mathbb{R}$. By combining these last two inequalities, we arrive at (7.6).

From (7.6), a version of Rellich's compactness theorem implies that the sequence $\{u_h\}$ is precompact in $L^1_{loc}([0, \infty) \times \mathbb{R})$.

Completion of the proof. The proof that any limit of a convergent subsequence is an entropy solution of (1.1) follows by standard arguments, along the lines of the proof of the Lax–Wendroff theorem. A detailed exposition is presented in [9]. □

Example 7.1. In this example we show why assumption (3.23) is needed in order to deduce from elementwise Riemann invariant bounds (3.25) that property (3.26) holds.

Consider Burgers' equation with $\mathcal{D} = [a, b]$, for $0 < a < b < 1$. Then a triangulation is causal if and only if no triangle has an edge that is a subset of a line with slope in the interval $[\frac{1}{b}, \frac{1}{a}]$. In particular the mesh shown in Figure 5 is causal (assuming that the vertical and horizontal scales are related in a suitable way.) Let the initial data be given by $u_0 = b$ to the left of point Q and $u_0 = a$ to the right of Q. Note that the total variation of the initial data is $b - a$. Let u_i denote the approximate solution in triangle T_i. Clearly, $u_1 = b$, $u_3 = a$, and by Lemma 5.1 we have $u_2 \in (a, b)$. Since the edge separating triangles T_2 and T_4 is trivial, we conclude $u_4 = u_2$. Finally, again by Lemma 5.1, we obtain that $a < u_5 < u_2 < b$. Thus the total variation of the approximate solution at time $t = t_1$ is at least

$$|u_1 - u_2| + |u_2 - u_3| + |u_3 - u_5| = b - a + (u_5 - a),$$

which is strictly greater than the total variation at $t = 0$.

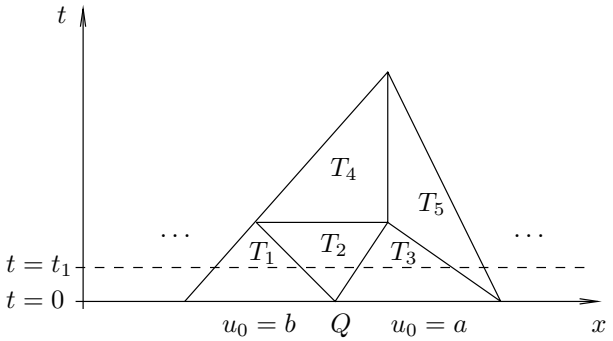


FIG. 5.

Acknowledgments. We are grateful to Robert B. Haber, Yangsuk Ko, and Jayandran Palaniappan for useful discussions, and to the Center for Process Simulation and Design (CPSD) at the University of Illinois for support.

REFERENCES

[1] C. ARVANITIS, C. MAKRIDAKIS, AND A. E. TZAVARAS, *Stability and convergence of a class of finite element schemes for hyperbolic systems of conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 1357–1393.

[2] P. BAITI AND A. BRESSAN, *The semigroup generated by a Temple class system with large data*, Differential Integral Equations, 10 (1997), pp. 401–418.

[3] B. COCKBURN AND P.-A. GREMAUD, *Error estimates for finite element methods for scalar conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 522–554.

[4] M. G. CRANDALL AND A. MAJDA, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[5] C. DAFERMOS, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, Berlin, 2000.

[6] A. HARTEN AND P. D. LAX, *A random choice finite difference scheme for hyperbolic conservation laws*, SIAM J. Numer. Anal., 18 (1981), pp. 289–315.

[7] A. Heibig, *Existence and uniqueness of solutions for some hyperbolic systems of conservation laws*, Arch. Ration. Mech. Anal., 126 (1994), pp. 79–101.
[8] J. Jaffre, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.
[9] K. Jegdic, *Analysis of a Spacetime Discontinuous Galerkin Method for Systems of Conservation Laws*, Ph.D. thesis, Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 2004.
[10] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.
[11] C. Johnson and J. Pitkaranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.
[12] N. Kuznecov and S. A. Volosin, *On monotone difference approximations for a first-order quasi-linear equation*, Soviet Math. Dokl., 17 (1976), pp. 1203–1206.
[13] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMSNSF Reg. Conf. Ser. Appl. Math. 11, SIAM, Philadelphia, 1973.
[14] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, Mathematical Aspects of Finite Elements in PDEs, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.
[15] R. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.
[16] R. LeVeque and B. Temple, *Stability of Godunov's method for a class of* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc., 288 (1985), pp. 115–123.
[17] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes* I, SIAM J. Numer. Anal., 33 (1996), pp. 760–779.
[18] R. B. Lowrie, *Compact Higher Order Numerical Methods for Hyperbolic Conservation Laws*, Ph.D. thesis, Department of Aerospace Engineering and Scientific Computing, University of Michigan, Ann Arbor, MI, 1996.
[19] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.
[20] J. Palaniappan, R. Haber, and R. Jerrard, *A spacetime discontinuous Galerkin method for scalar conservation laws*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 549–557.
[21] H. Rhee, R. Aris, and N. R. Amundson, *On the theory of multicomponent chromatography*, Philos. Trans. Roy. Soc. London Ser. A, 267 (1970), pp. 419–455.
[22] D. Serre, *Solutions a variations bornees pour certains systems hyperboliques de lois de conservation*, J. Differential Equations, 67 (1987), pp. 137–168.
[23] D. Serre, *Systems of Conservation Laws*, Cambridge University Press, Cambridge, UK, 1999.
[24] A. Szepessy, *Convergence of a shock-capturing streamline diffusion finite element method for a scalar conservation law in two space dimensions*, Math. Comp., 53 (1989), pp. 527–545.
[25] A. Szepessy, *Convergence of a streamline diffusion finite element method for scalar conservation laws with boundary conditions*, Math. Model. Numer. Anal., 25 (1991), pp. 749–782.
[26] E. Tadmor, *Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems*, Acta Numer., 25 (2004), pp. 451–512.
[27] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1954–1973

B-SPLINE LINEAR MULTISTEP METHODS AND THEIR CONTINUOUS EXTENSIONS*

FRANCESCA MAZZIA†, ALESSANDRA SESTINI‡, AND DONATO TRIGIANTE§

Abstract. In this paper, starting from a sequence of results which can be traced back to I. J. Schoenberg, we analyze a class of spline collocation methods for the numerical solution of ordinary differential equations (ODEs) with collocation points coinciding with the knots. Such collocation methods are naturally associated to a special class of linear multistep methods, here called B-spline (BS) methods, which are able to generate the spline values at the knots. We prove that, provided the additional conditions are appropriately chosen, such methods are all convergent and A-stable. The convergence property of the BS methods is naturally inherited by the related spline extensions, which, by the way, are easily and safely computable using their B-spline representation.

Key words. ordinary differential equations, linear multistep methods, boundary value methods, spline collocation, continuous extensions

AMS subject classifications. 65L06, 65D07, 41A15, 65M70

DOI. 10.1137/040614748

1. Introduction.

We analyze a class of spline collocation methods for the numerical solution of the general ordinary differential equation

$$\mathbf{y}'(x) = f(x, \mathbf{y}(x)),\ x \in [a, b], \tag{1.1}$$

that could be subject to either boundary ($\mathbf{g}(\mathbf{y}(a), \mathbf{y}(b)) = 0$) or initial ($\mathbf{y}(a) = \mathbf{y_a}$) conditions; $f : [a\,,\, b] \times \mathbb{R}^d \to \mathbb{R}^d$ is a sufficiently smooth function.

We started our research by reconsidering the classical papers [16] and [17], where a spline collocation approach at uniform knots is presented. In those papers, using $x_i = a + ih, i = 0, \ldots, N, h = (b-a)/N$ as knots, a spline of degree $k+1$, collocating the differential equation at the knots, is constructed with a forward approach, which is possible thanks to the use of $k-1$ additional initial conditions. The authors also prove that the knot restriction of the defined spline is the numerical solution of a suitable linear k-step method. It turned out that the approach was effective only for $k \leq 2$, that is, for the trapezoidal and the Simpson rules, because in this case it allowed the authors to define a convergent spline with uniform convergence order equal to 2 and 4, respectively. Unfortunately the convergence was lost for $k > 2$. This is clear considering that the resulting methods are not 0-stable, as proved in [16], for such values of k, and then not even convergence at knots can be obtained.

We first revisit this class of methods, showing how they can be implemented effectively as boundary value methods (BVMs) (see [5, 6] for a more detailed description of BVMs; also see section 4 below) and discussing in detail their stability properties.

*Received by the editors September 10, 2004; accepted for publication (in revised form) May 12, 2006; published electronically October 4, 2006. This work was supported by G.N.C.S. (INdAM) and COFIN-PRIN 2004 (project "Metodi numerici e software matematico per le applicazioni").
http://www.siam.org/journals/sinum/44-5/61474.html

†Dipartimento di Matematica, Università di Bari, Via Orabona 4, I-70125 Bari, Italy (mazzia@dm.uniba.it).

‡Dipartimento di Matematica, Università di Firenze, Viale Morgagni 67a, 50134 Firenze, Italy (sestini@math.unifi.it).

§Dipartimento di Energetica, Università di Firenze, Via Lombroso 6/17, 50134 Firenze, Italy (trigiant@unifi.it).

Then we explain how we have changed the approach introduced in [17] for defining the related spline extension, showing that convergence is recovered for all values of k. In our version, the spline is constructed in two successive steps: first, we compute a numerical solution of (1.1) by using the underlying multistep methods as BVMs, and then we construct a related spline extension by solving a linear system. The main features of this approach are that

1. it is a collocation method related to a class of linear multistep methods;
2. the continuous extension has the derivatives globally continuous up to order k;
3. we do not use extra collocation points which are different from the knots.

Other approaches devoted to obtaining spline collocation methods are well known in the literature; see, for example, those presented in [3, 4, 10, 25] that are related to Runge–Kutta methods, or the ones presented in [22, 23]. In this paper we rather prefer proving some mathematical properties of this approach than comparing it numerically with existing ones. Such an analysis allows us to gain insight into its deep relations with the BVM approach. As a matter of fact, it may be seen as an alternative way of constructing a BVM.

The importance of having a continuous extension of the solution provided by multistep methods is particularly relevant when we deal with a mesh refinement strategy, such as, for example, in the solution of boundary value problems, where it is necessary to work on the entire interval of integration $[a\,,\,b]$. In our case, we have implemented some BVMs in the code TOM [18, 20] (available at http://www.dm.uniba.it/~mazzia/bvp/index.html) and we have realized that, for many difficult problems, the interpolation needed by the mesh selection strategy may be a critical component. In the current version available on the Web, the solver uses cubic spline interpolation of the discrete computed solution as a tool to evaluate the approximation of the solution at off-mesh points.

The paper is organized as follows. In section 2 we introduce the special class of symmetric schemes we shall deal with, which will be called BS methods, since they are derived from B-splines. A detailed analysis of both their order of accuracy and their stability properties then follows. In sections 3 and 4 the problems of the computation of the related spline extensions and of choosing correctly the additional boundary conditions are considered, respectively. In section 5 we give the related global convergence analysis. As usual, all the properties are studied on a uniform mesh. Naturally, to have a class of methods that could be effectively implemented, it is necessary to extend them in order to use them on a nonuniform mesh. An extended treatment of the computation of the variable coefficients on a nonuniform mesh is presented in [19]. A preliminary implementation of the BS methods using a mesh selection strategy similar to the one described in [7, 8, 20], an error control strategy, and a technique for the solution of the nonlinear systems similar to the ones described in [18] has been done. As a tool to evaluate the approximation of the solution at off-mesh points, such an implementation uses the previously mentioned spline extension of the numerical solution. Some numerical results, using both uniform and nonuniform meshes, are reported in section 6 to confirm the features of the presented approach.

2. The BS linear multistep methods. Let $x_i = a + ih, i = 0, \ldots, N$, $h = (b-a)/N$, be a uniform partition of the integration interval $[a\,,\,b]$, and let us denote by $\mathbf{y}_i$ an approximation of $\mathbf{y}(x_i)$. For any $k \geq 1$, let us consider the classical B-spline $B_{k+2}(\cdot)$ of degree $(k+1)$ with uniform integer active knots $0, 1, \ldots, k+1, k+2$, which

is given, using the notation of truncated power [9], by

$$B_{k+2}(x) := \frac{1}{(k+1)!}\sum_{i=0}^{k+2}(-1)^i \binom{k+2}{i} (x-i)_+^{k+1}. \tag{2.1}$$

As proved in [16, Thm. 3], if a spline function $\mathbf{s}(x)$ of degree $k+1$ collocates the differential equation at the knots (that is, $\mathbf{s}'(x_i) = f(x_i, \mathbf{s}(x_i))$), then $\mathbf{s}(x)$ satisfies the following relation:

$$\sum_{i=0}^{k} \alpha_i^{(k)}\, \mathbf{s}(x_{n+i}) \;=\; h \sum_{i=0}^{k} \beta_i^{(k)}\, \mathbf{s}'(x_{n+i}), \quad n =, 0, \ldots, N-k, \tag{2.2}$$

with

$$\begin{cases} \alpha_i^{(k)} & := B'_{k+2}(k-i+1), \\ & \qquad\qquad i = 0, \ldots, k. \\ \beta_i^{(k)} & := B_{k+2}(k-i+1), \end{cases} \tag{2.3}$$

This result is a consequence of the previous theorem in the same paper, [16, Thm. 2], attributed to I. J. Schoenberg. In our notation it can be stated as follows,

THEOREM 1 (Schoenberg). *For any spline function* $\mathbf{s}$ *of degree* $k+1$ *and uniform knots* $0, h, \ldots, (n-1)h$, $n \geq k+1$, *the following relation holds:*

$$\sum_{i=0}^{k} \alpha_i^{(k)}\, \mathbf{s}(ih) \;=\; h \sum_{i=0}^{k} \beta_i^{(k)}\, \mathbf{s}'(ih), \tag{2.4}$$

where $\alpha_i^{(k)}$ *and* $\beta_i^{(k)}$ *are defined as in* (2.3).

The BS method with k steps is defined as follows:

$$\sum_{i=0}^{k} \alpha_i^{(k)}\, \mathbf{y}_{n+i} \;=\; h\sum_{i=0}^{k} \beta_i^{(k)}\, f(x_{n+i}, \mathbf{y}_{n+i}), \quad n = 0, \ldots, N-k. \tag{2.5}$$

Considering the symmetry properties of uniform B-splines, it can be shown that we are dealing with symmetric schemes, i.e.,

$$\alpha_i^{(k)} = -\alpha_{k-i}^{(k)}; \qquad \beta_i^{(k)} = \beta_{k-i}^{(k)}.$$

In addition, the following relations among the β coefficients come from the well-known recurrence relation for B-splines,

$$\beta_i^{(k)} \;=\; \frac{1}{k+1}[(k-i+1)\beta_{i-1}^{(k-1)} \;+\; (i+1)\beta_i^{(k-1)}], \;\; i = 0, \ldots, k, \tag{2.6}$$

where $\beta_{-1}^{(k-1)} = \beta_k^{(k-1)} = 0$. In Table 2.1 the $\alpha^{(k)}$ and $\beta^{(k)}$ coefficients of the BS methods with $k = 1, \ldots, 5$ are reported.

The related $\rho^{(k)}$ and $\sigma^{(k)}$ polynomials (that arise in the standard stability analysis of these formulas) are defined as

$$\rho^{(k)}(z) := \sum_{i=0}^{k} \alpha_i^{(k)}\, z^i, \qquad \sigma^{(k)}(z) := \sum_{i=0}^{k} \beta_i^{(k)}\, z^i.$$

TABLE 2.1
The $\alpha^{(k)}$ and $\beta^{(k)}$ coefficients of the BS methods with $k = 1, \ldots, 5$.

k	$\alpha^{(k)}$						$\beta^{(k)}$					
1	-1	1					$\frac{1}{2}$	$\frac{1}{2}$				
2	$-\frac{1}{2}$	0	$\frac{1}{2}$				$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$			
3	$-\frac{1}{6}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$			$\frac{1}{24}$	$\frac{11}{24}$	$\frac{11}{24}$	$\frac{1}{24}$		
4	$-\frac{1}{24}$	$-\frac{5}{12}$	0	$\frac{5}{12}$	$\frac{1}{24}$		$\frac{1}{120}$	$\frac{13}{60}$	$\frac{11}{20}$	$\frac{13}{60}$	$\frac{1}{120}$	
5	$-\frac{1}{120}$	$-\frac{5}{24}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{5}{24}$	$\frac{1}{120}$	$\frac{1}{720}$	$\frac{19}{240}$	$\frac{151}{360}$	$\frac{151}{360}$	$\frac{19}{240}$	$\frac{1}{720}$

An investigation of some properties of this class of methods is carried out in the following subsections.

Remark 1. We have used here an approach slightly different from that presented in [17]. In that paper the collocation spline of degree $k+1$ and smoothness C^k is first defined by using a forward approach based on the use of $k-1$ additional initial conditions. Then it is shown that its restriction to the knots verifies (2.5) with coefficients defined as in (2.3). Here we first compute the values generated by the linear multistep method at the knots, and then we compute the spline continuous extension. On the other hand, we observe that, if we had used the same approach, then the C^k continuity conditions at the knots would have generated exactly the conditions expressed in (2.5) (this assertion can be proved by considering the result of Theorem 6). The converse is also true; that is, the choice (2.3) of the coefficients will permit the definition of a spline extension of smoothness C^k collocating the differential equation at the knots (see section 3).

2.1. Order of the BS methods. The order of the k-step BS method can be proved by the following theorem.

THEOREM 2. *A k-step BS method is of order $p = k+1$ if k is odd and of order $p = k+2$ if k is even.*

Proof. The proof that p is at least $k+1$ can be obtained easily by using Theorem 1, by considering that any polynomial of degree less than or equal to $k+1$ is a spline of degree $k+1$, and that the relations defined in (2.4) are, if $\mathbf{s}$ is a polynomial, nothing but the order conditions. A direct proof that $p \geq k+1$, along the lines of the usual methodology used in the ordinary differential equation (ODE) setting, is reported in the appendix.

The fact that $p = k+2$ when k is even can be proved by using the symmetry of the BS methods. If $\mathbf{s}(x)$ is an even function with respect to the midpoint $x = kh/2$ of the interval $[0\,,\, kh]$, we have that $\sum_{i=0}^{k} \alpha_i^{(k)}\, \mathbf{s}(ih) = h \sum_{i=0}^{k} \beta_i^{(k)}\, \mathbf{s}'(ih) = 0$. Now, since for $k = 2\nu$ the polynomial $(x-\nu)^{k+2}$ is an even function with respect to ν, we can conclude that (2.4) still holds for $\mathbf{s}(x) = (x-\nu)^{k+2}$. ∎

2.2. Behavior of the polynomials $\boldsymbol{\rho^{(k)}(z)}$ and $\boldsymbol{\sigma^{(k)}(z)}$. Concerning the polynomial $\rho^{(k)}(z)$, by using the derivative formula for B-splines (see for instance [24,

p. 141], it can be observed that

$$\rho^{(k)}(z) = (z-1)\,\sigma^{(k-1)}(z). \tag{2.7}$$

Then its behavior depends on that of the σ polynomials.

Considering the symmetry of the B-splines, it is also easy to verify that $\sigma^{(k)}(z)$ has the root $z_1 = -1$ if k is odd and that in all the cases, if z is one of its roots, then $w = 1/z$ is also one of them. In addition, it will be proved that all the roots of $\sigma^{(k)}(z)$, except $z_1 = -1$ for k odd, do not belong to the boundary of the unit circle of the complex plane. This statement is important in proving the stability properties of the methods [5, 11], and its proof needs some preliminary considerations.

To begin with, it is almost trivial to recognize that, for any symmetric k-step linear multistep method, $\sigma^{(k)}(z)$ has the following form when evaluated at $z = e^{\mathbf{i}\theta}$:

$$\sigma^{(k)}(e^{\mathbf{i}\theta}) = \begin{cases} 2\,f_k(\theta)\cos(\frac{\theta}{2})\,e^{\mathbf{i}\,k\frac{\theta}{2}} & \text{if } k = 2\nu-1, \\ 2\,f_k(\theta)\,e^{\mathbf{i}\,k\frac{\theta}{2}} & \text{if } k = 2\nu, \end{cases} \tag{2.8}$$

where ν is any positive integer and $f_k(\theta)$ is a trigonometric polynomial defined as

$$\begin{cases} f_k(\theta) := \beta^{(k)}_{\nu-1} + \dfrac{1}{\cos(\frac{\theta}{2})}\displaystyle\sum_{j=2}^{\nu}\beta^{(k)}_{\nu-j}\cos\left((2j-1)\,\dfrac{\theta}{2}\right) & \text{if } k = 2\nu-1, \\ f_k(\theta) := \dfrac{\beta^{(k)}_{\nu}}{2} + \displaystyle\sum_{j=1}^{\nu}\beta^{(k)}_{\nu-j}\cos(j\,\theta) & \text{if } k = 2\nu \end{cases} \tag{2.9}$$

(in fact, for all $j \in \mathbb{N}^+$, $\cos\frac{\theta}{2}$ divides $\cos\,(2j-1)\frac{\theta}{2}$ exactly; see also below).

From (2.8), it follows that the proof that the polynomials $\sigma^{(k)}(z)$ have no roots of unit modulus, except for $z = -1$ for odd values of k, is equivalent to the statement that the trigonometric polynomials $f_k(\theta)$ are of constant sign in $[0\,,\,2\pi]$. Now, the problem of establishing whether a trigonometric polynomial of the form (2.9) does not change sign is well known in the literature. Results may be found in [12, 21]; more recent results can be found in [2]. Unfortunately, it seems that our coefficients $\beta^{(k)}_i, i = 0,\ldots,k$, do not satisfy the conditions established in the above mentioned references, and we need to look for new conditions valid for our case. In order to consider simultaneously both the odd and even cases, we introduce the following vector $\mathbf{b}^{(k)} \in \mathbb{R}^{k+1}$:

$$\mathbf{b}^{(k)} := \begin{cases} (0, \beta^{(k)}_{\nu-1}, 0, \beta^{(k)}_{\nu-2}, 0, \ldots, \beta^{(k)}_0)^T & \text{if } k = 2\nu-1, \\ (\beta^{(k)}_{\nu}/2, 0, \beta^{(k)}_{\nu-1}, \ldots, \beta^{(k)}_0)^T & \text{if } k = 2\nu, \end{cases} \tag{2.10}$$

and the vector function $\mathbf{v}^{(k)}(\theta) : [0\,,\,2\pi] \to \mathbb{R}^{k+1}$,

$$\mathbf{v}^{(k)}(\theta) := \left(1, \cos\left(\frac{\theta}{2}\right), \cos\left(2\frac{\theta}{2}\right), \cos\left(3\frac{\theta}{2}\right), \ldots, \cos\left(k\frac{\theta}{2}\right)\right)^T. \tag{2.11}$$

Using this notation, we can write

$$\begin{cases} \cos\left(\dfrac{\theta}{2}\right) f_k(\theta) = [\mathbf{b}^{(k)}]^T\,\mathbf{v}^{(k)}(\theta) & \text{if } k = 2\nu-1, \\ f_k(\theta) = [\mathbf{b}^{(k)}]^T\,\mathbf{v}^{(k)}(\theta) & \text{if } k = 2\nu. \end{cases}$$

Now, considering that $\cos(j\theta/2) = T_j(\cos(\theta/2))$, where T_j are the Chebyshev polynomials of first kind, an infinite matrix C can be found such that, for all $k \in \mathbb{N}^+$, the following formula holds:

$$\mathbf{v}^{(k)}(\theta) = C_{k+1}\, \mathbf{w}^{(k)}(\theta), \tag{2.12}$$

where C_{k+1} is the principal submatrix of order $k+1$ of C and the vector function $\mathbf{w}^{(k)}(\theta)$ is defined as

$$\mathbf{w}^{(k)}(\theta) := (1, \cos(\theta/2), \cos^2(\theta/2), \ldots, \cos^k(\theta/2))^T.$$

The infinite matrix C is lower triangular and its principal submatrix C_2 is coincident with the identity matrix I_2. The entries of the other rows are defined by the following recurrence relations (derived from the recurrence relation for Chebyshev polynomials):

$$\begin{cases} C_{i,j} & := -\,C_{i-2,j} & \text{if } j = 1,\ i \geq 3, \\ C_{i,j} & := 2\,C_{i-1,j-1} - C_{i-2,j} & \text{if } j \geq 2,\ i \geq 3. \end{cases}$$

For example, its principal submatrix C_8 is

$$C_8 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 4 & 0 & 0 & 0 & 0 \\ 1 & 0 & -8 & 0 & 8 & 0 & 0 & 0 \\ 0 & 5 & 0 & -20 & 0 & 16 & 0 & 0 \\ -1 & 0 & 18 & 0 & -48 & 0 & 32 & 0 \\ 0 & -7 & 0 & 56 & 0 & -112 & 0 & 64 \end{pmatrix}.$$

We remark that through (2.12) the cosine of the jth multiple of $\theta/2$ is expressed as a linear combination of powers of $\cos(\theta/2)$ of degree $\leq j$. In particular, because of the checkerboard structure of C, cosines of even multiples of $\theta/2$ become linear combinations of even powers, while cosines of odd multiples are transformed into linear combinations of odd powers. We are now in the position to prove the following theorem.

THEOREM 3. *For all $k \in \mathbb{N}^+$, if*

$$\hat{\mathbf{b}}^{(k)} := C_{k+1}^T\, \mathbf{b}^{(k)}$$

is nonnegative with at least one positive entry, then the trigonometric polynomials f_k defined in (2.9) *are positive.*

Proof. The proof is almost trivial considering that the entries of $\hat{\mathbf{b}}^{(k)}$ are the coefficients of the trigonometric polynomials in the new basis of powers of $\cos(\theta/2)$. In fact, as already observed, if k is even only even powers are involved in the expression of $f_k(\theta)$. If k is odd, only odd powers are involved in the expression of $\cos(\theta/2)\, f_k(\theta)$. Then, again, only even powers are involved in the expression of $f_k(\theta)$. ∎

We remark that the theorem holds true for any class of symmetric linear multistep methods and it allows us to assert that the polynomial $\sigma^{(k)}(z)$ has no roots of unit modulus, except $z_1 = -1$ when k is odd, provided that the vector $\hat{\mathbf{b}}^{(k)}$ is nonnegative with at least one positive entry. In the case of BS methods, such a question is settled by the following theorem.

THEOREM 4. *If BS methods are considered, then for all $k \in \mathbb{N}^+$ the vector $\hat{\mathbf{b}}^{(k)}$ is nonnegative with at least one positive entry.*

Proof. This can be proved by taking into account (2.6). In fact, by using the same notation as before, such a relation can be posed in matrix form as

$$\mathbf{b}_k = \frac{1}{k+1} A^{(k)} \mathbf{b}_{k-1}, \tag{2.13}$$

where $A^{(k)}$ is a bidiagonal rectangular matrix ($A_{i,j} \neq 0$ only if $j = i-1$ or $j = i+1$) of size $(k+1) \times k$ whose nonzero entries are as follows, assuming that $\nu = \lceil k/2 \rceil$:

$$\begin{cases} A^{(k)}_{2,1} & = 2\nu, \\ A^{(k)}_{3,2} & = \nu, \\ A^{(k)}_{2i,2i-1} & = \nu + 1 - i, \quad i = 2, \ldots, \nu - 1, \\ A^{(k)}_{2i+1,2i} & = \nu + 1 - i, \quad i = 2, \ldots, \nu - 1, \\ A^{(k)}_{2\nu,2\nu-1} & = 1, \\ A^{(k)}_{2i-1,2i} & = \nu + i, \qquad i = 1, \ldots, \nu - 1, \\ A^{(k)}_{2i,2i+1} & = \nu + i, \qquad i = 1, \ldots, \nu - 1. \end{cases}$$

These relations need to be completed by the following two nonzero entries when k is even ($k = 2\nu$):

$$\begin{cases} A^{(2\nu)}_{2\nu+1,2\nu} & = 1, \\ A^{(2\nu)}_{2\nu-1,2\nu} & = 2\nu. \end{cases}$$

Consequently, the vectors $\hat{\mathbf{b}}_k$ satisfy the recurrence relation

$$\hat{\mathbf{b}}_k = \frac{1}{k+1} R^{(k)} \hat{\mathbf{b}}_{k-1},$$

where $R^{(k)} = C^T_{k+1} A^{(k)} (C^{-T}_k)$. Such a matrix can be constructed in explicit form considering the checkerboard and bidiagonal structures, respectively, of C_k and $A^{(k)}$. It is not difficult to see that $R^{(k)}$ has the same structure of $A^{(k)}$. Moreover, its nonzero entries are, still assuming that $\nu = \lceil k/2 \rceil$,

$$\begin{cases} R^{(k)}_{2i,2i-1} & = 2\,(\nu + 1 - i), \quad i = 1, \ldots, \nu - 1, \\ R^{(k)}_{2i+1,2i} & = 2\,(\nu + 1 - i), \quad i = 1, \ldots, \nu - 1, \\ R^{(k)}_{2\nu,2\nu-1} & = 2, \\ R^{(k)}_{j,j+1} & = j, \qquad\qquad j = 1, \ldots, k - 1, \end{cases} \tag{2.14}$$

which must be completed, if $k = 2\nu$, by

$$R^{(2\nu)}_{2\nu+1,2\nu} = 2.$$

These relations can be proved by checking that $C^T_{k+1} A^{(k)} = R^{(k)} C^T_k$. Thus, noting that $R^{(k)}$ is a nonnegative matrix and that $\hat{\mathbf{b}}_1 = (0, 1/2)^T$, the theorem is proved by induction. ∎

Before stating Corollary 1, we need the following definition.

DEFINITION 2.1. *A k-degree polynomial is of type* $(r, k-r-s, s)$*, with* $r+s \le k$*,* $r, s \ge 0$ *if it has* r *roots inside the open unit disk,* $k-r-s$ *roots on the unit circle, and* s *roots outside the closed unit disk.*

We can now state the following corollary.

COROLLARY 1. *For all* $k \in \mathbb{N}^+$*, the trigonometric polynomial* $f_k(\theta)$ *defined in* (2.9) *and associated with the kth step BS method is positive in* $[0\,,\,2\pi]$*. Consequently, if* $k_1 = \lceil \frac{k}{2} \rceil$ *and* $k_2 = \lfloor \frac{k}{2} \rfloor$*, the corresponding* $\sigma^{(k)}(z)$ *polynomial is of type* $(k_1, 0, k_2)$ *if* k *is even and of type* $(k_1 - 1, 1, k_2)$ *if* k *is odd.*

Proof. The proof trivially follows from Theorems 3 and 4. ∎

2.3. Stability of the BS methods. Taking into account (2.7), Corollary 1 implies that the BS methods with $k \ge 3$ are not 0-stable and, as a consequence, they cannot be used as initial value methods because convergence is lost, as already pointed out by Loscalzo [17]. However, convergence can be recovered if they are used as BVMs with $k_1 = \lceil \frac{k}{2} \rceil$ initial and $k_2 = \lfloor \frac{k}{2} \rfloor$ final conditions. This can be proved by noting that such methods are $0_{k1,k2}$-stable, which is a concept generalizing the classical 0-stability. A deeper discussion about the generalization of the concept of 0-stability and A-stability (see A_{k_1,k_2}-stability below) can be found in [5, 6]. In fact (2.7) implies that $\rho^{(k)}(z)$ has k_1 and k_2 roots whose modulus is less than or equal to 1 and greater than 1, respectively, and that the roots on the unit circle (1 and, for k even, -1) are simple. Together with the fact that these methods have order $p \ge k+1$, this leads to convergence, provided that the additional conditions are appropriately chosen.

On the other hand, we are not only interested in convergence, but also in a good behavior of the methods for fixed h. More explicitly, we would like these methods, for fixed h, to generate well-conditioned discrete problems when applied to well-conditioned continuous ones. In this regard, a BVM using k_1 initial and k_2 final conditions is said to be A_{k_1,k_2}-stable if it generates a well-conditioned discrete problem for each well-conditioned continuous linear problem, thus generalizing the well-known concept of A-stability. We say that a method is *precisely* A_{k_1,k_2}*-stable* if the associated Dahlquist polynomial $\pi^{(k)}(z,q) := \rho^{(k)}(z) - q\sigma^{(k)}(z)$ has constant type $(k_1, 0, k_2)$ only for all $q \in \mathbf{C}^-$. This implies that the boundary of the stability region is the imaginary axis. Precisely A_{k_1,k_2}-stable methods are the safest in the case of nondissipative problems. In fact it can be proved that A_{k_1,k_2}-stable methods which are not precisely stable can give wrong results for such problems [5].

The connection between the type of the Dahlquist polynomial and the position of q in the complex plane can be investigated by looking at the *boundary locus* (see (2.15) below) associated to the method considered. Essentially, such a set is the locus in the complex plane, where $\pi^{(k)}(z,q)$ changes its type, that is, one root reaches the unit circle. For the BS method, we have the following result.

THEOREM 5. *The boundary locus related to the k-step BS method is the imaginary axis of the complex plane if* k *is odd and is just a segment of the imaginary axis of the complex plane if* k *is even.*

Proof. The boundary locus Γ_k of a k-step linear multistep method is defined as

$$\Gamma_k := \left\{ q_k(\theta) \in \mathbf{C} \mid q_k(\theta) = \frac{\rho^{(k)}(e^{\mathbf{i}\theta})}{\sigma^{(k)}(e^{\mathbf{i}\theta})},\ \theta \in [0\,,\,2\pi] \right\}. \tag{2.15}$$

Considering (2.7), we have now that for the k-step BS method the boundary locus is

$$q_k(\theta) = k\,(e^{\mathbf{i}\theta} - 1)\,\frac{\sigma^{(k-1)}(e^{\mathbf{i}\theta})}{\sigma^{(k)}(e^{\mathbf{i}\theta})}. \tag{2.16}$$

On the other hand, (2.8) implies that we can also write

$$q_k(\theta) = \begin{cases} k\,\mathbf{i}\,\sin(\theta)\,\frac{f_{k-1}(\theta)}{f_k(\theta)} & \text{if } k = 2\nu, \\ 2k\,\mathbf{i}\,\tan(\frac{\theta}{2})\,\frac{f_{k-1}(\theta)}{f_k(\theta)} & \text{if } k = 2\nu+1, \end{cases} \tag{2.17}$$

where $\mathbf{i}$ denotes the imaginary unit. The statement of the theorem is proved by using the result proved in Corollary 1. □

Noting that Γ_k is the locus where the Dahlquist polynomial changes its type it then follows that such a type is constantly equal to $(k_1, 0, k_2)$ in $\mathbf{C}^-$ for all k. Since Γ_k coincides with the imaginary axis for k odd, in this case $\pi^{(k)}(z, q)$ changes its type, moving from $\mathbf{C}^-$ to $\mathbf{C}^+$. Consequently, we can conclude that the BS methods are always A_{k_1,k_2}-stable and that they are precisely A_{k_1,k_2}-stable if k is odd (see [5]).

We postpone until section 4, after the definition of the spline extension, the discussion about possible strategies for choosing the necessary additional conditions. At the moment, we assume that each method, together with the given initial or boundary condition and with the $k-1$ additional conditions, generates a unique discrete solution.

3. The spline extension. Let us now assume that a numerical solution $\{\mathbf{y}_i, i = 0, \ldots, N\}$ has been computed by the k-step BS method used as a BVM. Thus, if we put $\mathbf{y}'_i := f(x_i, \mathbf{y}_i), i = 0, \ldots, N$, we are interested in determining a vector spline function $\mathbf{s}_k(x) = (s_k^1(x), \ldots, s_k^d(x))^T$ with $s_k^j \in S_{k,N}, j = 1, \ldots, d$, such that

$$\begin{cases} \mathbf{s}_k(x_i) &= \mathbf{y}_i, \quad i = 0, \ldots, N, \\ \mathbf{s}'_k(x_i) &= \mathbf{y}'_i, \quad i = 0, \ldots, N \end{cases} \tag{3.1}$$

(i.e., a spline solution of the Hermite interpolation problem at the mesh points), where $S_{k,N}$ denotes the classical linear functional space of piecewise polynomials of degree $k+1$ and knots x_i, $i = 0, \ldots, N$, with global smoothness $C^k([a\,,\,b])$. Now, after defining $k+1$ auxiliary left knots $x_i = a + ih, i = -(1+k), \ldots, -1$, and $k+1$ auxiliary right knots $x_i = a + ih, i = N+1, \ldots, N+k+1$, we can write

$$S_{k,N} := \langle\, B_{-(1+k),k+2}, \ldots, B_{N-1,k+2} \,\rangle,$$

where

$$B_{j,k+2}(x) := B_{k+2}\left(\frac{x - x_j}{h}\right),\ j = -(1+k), \ldots, N-1. \tag{3.2}$$

Thus, using the B-spline representation of any $s_k^j \in S_{k,N}$, we can write

$$\mathbf{s}_k(x) = \sum_{i=-(1+k)}^{N-1} \mathbf{c}_i B_{i,k+2}(x),\ x \in [a, b], \tag{3.3}$$

where $\mathbf{c}_i \in \mathbb{R}^d, i = -(1+k), \ldots, N-1$, are the vector spline coefficients we are looking for. Consequently, (3.1) can be rewritten in the compact form

$$(A \otimes I_d)\,\mathbf{c} \;=\; (\mathbf{y}_0^T, \ldots, \mathbf{y}_N^T, h(\mathbf{y}'_0)^T, \ldots, h(\mathbf{y}'_N)^T)^T, \tag{3.4}$$

where $\mathbf{c} = (\mathbf{c}_{-(1+k)}^T, \ldots, \mathbf{c}_{N-1}^T)^T \in \mathbb{R}^{d\,(N+k+1)}$, I_d is the identity matrix of size $d \times d$, and A is the block matrix

$$A := \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}. \tag{3.5}$$

Considering (3.2) and (2.3), it is easy to show that both the block matrices A_1, A_2 in (3.5) are Toeplitz matrices of size $(N+1) \times (N+k+1)$,

$$A_1 := \begin{pmatrix} \beta_0^{(k)} & \cdots & \beta_k^{(k)} & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \beta_0^{(k)} & \cdots & \beta_k^{(k)} \end{pmatrix}, \tag{3.6}$$

$$A_2 := \begin{pmatrix} \alpha_0^{(k)} & \cdots & \alpha_k^{(k)} & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \alpha_0^{(k)} & \cdots & \alpha_k^{(k)} \end{pmatrix}. \tag{3.7}$$

The spline solution of (3.1), $\mathbf{s}_k(x)$, is uniquely determined, as proved in the following theorem.

THEOREM 6. *The rectangular linear system* (3.4) *has only one solution if the entries of the vector on the right-hand side satisfy* (2.5) *with* α *and* β *coefficients as defined in* (2.3).

Proof. From the classical theory of spline interpolation, we know that the coefficient matrix defined in (3.5) is a full rank matrix. In fact, this is a consequence of the general results about osculatory spline interpolation first presented in the classic paper [15] and reported also in [9, Thm. XIII.4, p. 228]. On the other hand, we know that the right-hand side of (3.4), $(\mathbf{y}_0^T, \dots, \mathbf{y}_N^T, h(\mathbf{y}_0')^T, \dots, h(\mathbf{y}_N')^T)^T$, satisfies the $N-k+1$ linear conditions given in (2.5), because the numerical solution is computed using just these relations. Now it is also easy to verify that, if $\mathbf{a}_1^{(j)}, \dots, \mathbf{a}_{N+1}^{(j)}$ denote the rows of $A_j, j = 1, 2$, then

$$\sum_{i=0}^{k} \alpha_i^{(k)} \, \mathbf{a}_{n+i}^{(1)} = \sum_{i=0}^{k} \beta_i^{(k)} \, \mathbf{a}_{n+i}^{(2)}, \ n = 0, \dots, N-k.$$

Consequently, in (3.4) there are $N-k+1$ redundancies, and this implies that there exists a unique solution of it since A is a full rank matrix. ∎

3.1. Procedure for constructing the spline extension. Once we have proved that there exists a unique solution, $\mathbf{s}_k(x)$, of (3.1), we need to choose an algorithm for the computation of its coefficient vectors $\mathbf{c}_i, i = -(1+k), \dots, N-1$. Clearly, we could solve (3.4) in the least squares sense, but this is too expensive, considering we know that there exists an exact solution of it. From spline theory (see again [9, Thm.

TABLE 3.1
The condition numbers of M_1 in the Euclidean norm.

N	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
5	10.9	4.1	49.8	260.2	4708.7
10	20.2	3.9	71.0	226.0	5515.6
20	38.9	3.9	104.0	223.9	7503.7
40	76.3	3.9	156.3	223.7	10478.3
80	151.2	3.9	245.1	223.6	14747.3

XIII.4, p. 228]), we know that the square submatrix M_1 of A, given by

$$M_1 := \begin{pmatrix} \begin{pmatrix} \mathbf{a}_1^{(2)} \\ \vdots \\ \mathbf{a}_{k_1}^{(2)} \end{pmatrix} \\ A_1 \\ \begin{pmatrix} \mathbf{a}_{N-k_2+1}^{(2)} \\ \vdots \\ \mathbf{a}_N^{(2)} \end{pmatrix} \end{pmatrix}, \tag{3.8}$$

is nonsingular, where we recall that $k_1 = \lceil \frac{k}{2} \rceil$ and $k_2 = \lfloor \frac{k}{2} \rfloor$. So we could solve the equivalent square linear system

$$\begin{aligned} &(M_1 \otimes I_d)\,\mathbf{c} \\ &\quad = (h(\mathbf{y}'_0)^T, \ldots, h(\mathbf{y}'_{k_1})^T, \mathbf{y}_0^T, \ldots, \mathbf{y}_N^T, h(\mathbf{y}'_{N-k_2+1})^T, \ldots, h(\mathbf{y}'_N)^T)^T. \end{aligned} \tag{3.9}$$

In the case of odd degree splines, interpolants at the knots are often determined by solving such a system. Nevertheless, this is a good idea only if k is even. Looking at Table 3.1, it is possible to realize that the condition number of M_1 does not depend on N (i.e., M_1 is *well conditioned*) only if k is even. If k is odd, it grows with N in a linear fashion (i.e., M_1 is *weakly well conditioned*). This behavior is explained by Theorem 7 and Corollary 2.

Unfortunately it is the case of k odd that we are more interested in because of the precise stability feature of the corresponding BS methods (see Theorem 5). Since we have the freedom of choosing from among the equations in (3.4), we have replaced (3.9) with the equivalent linear system

$$\begin{aligned} &(M_2 \times I_d)\,\mathbf{c} \\ &\quad = (\mathbf{y}_0^T, \ldots, \mathbf{y}_{k_1}^T, (\mathbf{y}_0 + h\mathbf{y}'_0)^T, \ldots, (\mathbf{y}_N + h\mathbf{y}'_N)^T, \mathbf{y}_{N-k_2+1}^T, \ldots, \mathbf{y}_N^T)^T, \end{aligned} \tag{3.10}$$

where

$$M_2 := \begin{pmatrix} \begin{pmatrix} \mathbf{a}_1^{(1)} \\ \vdots \\ \mathbf{a}_{k_1}^{(1)} \end{pmatrix} \\ A_1 + A_2 \\ \begin{pmatrix} \mathbf{a}_{N-k_2+1}^{(1)} \\ \vdots \\ \mathbf{a}_N^{(1)} \end{pmatrix} \end{pmatrix}. \tag{3.11}$$

TABLE 3.2
The condition numbers of M_2 in the Euclidean norm.

N	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
5	4.6	10.5	46.6	493.4	6179.7
10	4.6	10.3	46.3	432.6	4861.2
20	4.6	10.3	46.3	416.5	4850.7
40	4.6	10.3	46.3	416.1	4850.2
80	4.6	10.3	46.3	416.1	4850.2

Now the condition number of the new matrix M_2 does not depend on N for any value of k, as confirmed by Table 3.2 and explicitly proved in Theorem 7 and Corollary 3. Before introducing Theorem 7, we recall that the *polynomial* and the *symbol* associated to a $(k+1)$-banded Toeplitz matrix $T = (t_{i,j}), t_{i,j} = \tau_{j-i}$, with k_1 nonzero lower diagonals are $p(z) = \sum_{i=0}^{k} \tau_{-k_1+i}\, z^i$ and $z^{-k_1}p(z)$, respectively (see, e.g., [13, 14]).

THEOREM 7. *Assume that*

1. *k, k_1, and k_2 are three positive integers, with $1 \le k_1 < k$ and $k_2 = k - k_1$;*
2. *$M = T + R$, where T is an $n \times n, n > 2k$, $(k+1)$-banded Toeplitz matrix with k_1 nonzero lower diagonals;*
3. *R is a matrix having nonzero entries $R_{i,j}$ only if $i = 1, \ldots, k_1$ and $j = 1, \ldots, k + k_1$ or if $i = n - k_2 + 1, \ldots, n$ and $j = n - k - k_2 + 1, \ldots, n$;*
4. *the polynomial $p(z)$ associated to T is of type $(k_1 - s, s, k_2)$, $s = 0, 1$.*

Then, constants $\eta > 0$ and $0 < \zeta < 1$ independent of n exist such that the following two statements hold:

(a) *The matrix $|M^{-1}|$ whose entries are the absolute values of the corresponding ones in M^{-1} satisfies the componentwise bound*

$$\begin{array}{ll} |M^{-1}| \le \eta(I_n + \Delta_n + \Delta_n^T) & \text{for } s = 0, \\ |M^{-1}| \le \eta(I_n + \Omega_n + \Delta_n^T) & \text{for } s = 1, \end{array} \tag{3.12}$$

where

$$\Delta_n := \begin{pmatrix} 0 & 0 & \ldots & 0 \\ \zeta & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \zeta^{n-1} & \ddots & \zeta & 0 \end{pmatrix}_{n\times n} \quad \text{and} \quad \Omega_n = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \ddots & 1 & 0 \end{pmatrix}_{n\times n};$$

(b) $||M^{-1}||_\infty,\ ||M^{-1}||_1,\ ||M^{-1}||_2 \le \frac{\eta}{1-\zeta}$ *for $s = 0$,*
$||M^{-1}||_\infty,\ ||M^{-1}||_1,\ ||M^{-1}||_2 \le \eta n$ *for $s = 1$.*

Proof. If the perturbation matrix R is not taken into account, the proof follows immediately from a more general result about conditioning of banded Toeplitz matrices given in [1] and in [5, p. 74]. Then, since the matrix M^{-1} can be recast as

$$M^{-1} = (I_n + T^{-1}R)^{-1}T^{-1}, \tag{3.13}$$

the proof is reduced to showing that the matrix $(I_n + T^{-1}R)^{-1}$ exists and has entries bounded with respect to n. This has been done in [14], and we omit the long manipulation for brevity. ∎

COROLLARY 2. *The matrix M_1 defined in* (3.8) *is well conditioned in the case of k even and weakly well conditioned in the case of k odd.*

Proof. M_1 can be split as $M_1 = T_1 + R_1$, where T_1 is a $(k+1)$-banded Toeplitz matrix whose associated symbol is $z^{-k_1}\sigma^{(k)}(z)$ and R_1 satisfies the related hypothesis assumed in Theorem 7. Having proved in section 2.2 that $\sigma^{(k)}(z)$ is of type $(k_1, 0, k_2)$ if k is even and of type $(k_1 - 1, 1, k_2)$ if k is odd, the proof follows from the previous theorem. ∎

COROLLARY 3. *The matrix M_2 defined in* (3.11) *is well conditioned for all values of k.*

Proof. M_2 can be split as $M_2 = T_2 + R_2$, where T_2 is a k-banded Toeplitz matrix whose associated symbol is $z^{-k_1}(\rho^{(k)}(z) + \sigma^{(k)}(z))$. In section 2.3 it has been proved that the k-step BS method is A_{k_1,k_2}-stable. This implies that the polynomial $\rho^{(k)}(z) - q\sigma^{(k)}(z)$ always has the same type $(k_1, 0, k_2)$ for all $q \in \mathbf{C}^-$. By taking $q = -1$, the corollary follows from the previous theorem. ∎

4. The additional conditions. The BVM usage of a k-step linear multistep formula does not impose all the $k-1$ additional conditions at the beginning of the interval, but it carefully splits such conditions at the beginning and at the end of the interval. It is interesting to note that the construction of interpolating splines has exactly the same degrees of freedom, and usually the required additional conditions are split similarly. In order to show the deep similarity between the spline interpolation process and the BVM approach, we consider the classical cubic spline interpolation; see [9, p. 53]. If the free slopes denoted there by $s_i, i = 2, n-1$ are taken to be equal to $f(x_i, y_i)$ defined in (1.1), then the continuity conditions give rise to the 2-step BS method, i.e., Simpson's method. On the other hand, it is known that the boundary conditions suggested in [9] are imposed one at the beginning and the other at the end of the interval, as suggested, using different reasoning, in [5]. The use of Simpson's method with both conditions at the beginning gives rise to a very unstable method.

Therefore, in order to use the k-step BS method correctly as a BVM we need to define the $k_1 = \lceil \frac{k}{2} \rceil$ initial and the $k_2 = \lfloor \frac{k}{2} \rfloor$ final conditions. One condition is naturally given by problem (1.1); the others need to be defined. There are many possible choices for the additional boundary conditions, precisely as in the classical definition of spline interpolating functions. Usually, they are obtained by using different linear multistep methods of order $\geq p - 1$. Here, alternatively, we can take advantage of the methodology used in the spline setting. One possibility is to use the so called "not-a-knot" condition which implies that some knots of the spline near the extremes of the integration interval are removed. The not-a-knot condition at the generic point x_j is $\mathbf{s}_k^{(k+1)}(x_j^-) = \mathbf{s}_k^{(k+1)}(x_j^+)$. Such a condition can be translated into a linear condition on the functional and derivative values assumed by the spline at the knots. Each translated condition will become an additional condition. For example, it is known (see [9]) that for the Simpson method the not-a-knot condition at x_{N-1} corresponds to the equation

$$-\mathbf{y}_{N-2} - 4\mathbf{y}_{N-1} + 5\,\mathbf{y}_N = h\,(4\mathbf{y}'_{N-1} + 2\mathbf{y}'_N). \tag{4.1}$$

For brevity, we avoid presenting the manipulation needed to obtain such conditions for $k > 2$ (it is fully introduced in the general case of a nonuniform mesh in [19]).

Different techniques are also possible. For example, one can compute an estimate of some derivatives of $\mathbf{y}(x)$ at the boundary points using finite differences. In the above example, a good choice (which does not affect the fourth order convergence) could be to compute the derivative using the backward differentiation formula of the

same order, that is,

$$\mathbf{y}'_N = \frac{-2\mathbf{y}_{N-3} + 9\mathbf{y}_{N-2} - 18\mathbf{y}_{N-1} + 11\mathbf{y}_N}{6h}.$$

This means that $\mathbf{y}'_N$ is taken to be equal to the derivative value at x_N of the cubic polynomial passing through $(x_i, \mathbf{y}_i), i = N-3, \ldots, N$.

5. Convergence of the spline extension. An important feature of the class of BS methods is the convergence in the uniform norm of the related continuous extension $\mathbf{s}_k(x)$ to the solution $\mathbf{y}(x)$ of (1.1). We now prove the following theorem, where for the sake of brevity we restrict ourselves to $d = 1$.

THEOREM 8. *Let us assume that $y \in C^{k+2}[a, b]$. Then*

$$\|s_k - y\|_\infty \leq C_{k,y}\, h^p,$$

where $C_{k,y}$ is a suitable constant depending on k and y and

$$p = \begin{cases} k+2 & \text{if } k \text{ is even,} \\ k+1 & \text{if } k \text{ is odd.} \end{cases} \tag{5.1}$$

Proof. Clearly, we can write

$$\|s_k - y\|_\infty \leq \|s_k - \hat{s}_k\|_\infty + \|\hat{s}_k - y\|_\infty, \tag{5.2}$$

where $\hat{s}_k(x) \in S_{k,N}$ is the best approximation of $y(x)$ in $S_{k,N}$. The second term on the right-hand side of (5.2) can be bounded using the general result proved in the Jackson-type theorem reported in [9, Thm. XII.1, p. 170],

$$\|\hat{s}_k - y\|_\infty \leq C_k^{(1)}\, h^{k+2}\, \|y^{(k+2)}\|_\infty, \tag{5.3}$$

where $C_k^{(1)}$ is a constant depending only on k and $y^{(k+2)}$ denotes the $(k+2)$-derivative of y.

So let us consider the first term on the right-hand side of (5.2). If $s_k(x) = \sum_{i=-(1+k)}^{N-1} c_i\, B_{i,k+2}(x)$, and $\hat{s}_k(x) = \sum_{i=-(1+k)}^{N-1} \hat{c}_i\, B_{i,k+2}(x)$, since the B-splines $B_{i,k+2}(\cdot), i = (-1+k), \ldots, N-1$, are nonnegative and satisfy the unit partition property in $[a\,, b]$, we can write

$$\|s_k - \hat{s}_k\|_\infty = \left\| \sum_{i=-(1+k)}^{n-1} (c_i - \hat{c}_i)\, B_{i,k+2}(x) \right\|_\infty \leq \|\mathbf{c} - \hat{\mathbf{c}}\|_\infty.$$

Now, as mentioned in section 4.1, the coefficient vector $\mathbf{c}$ can be safely computed for all k by solving linear system (3.10). Thus, considering that p defined in (5.1) is the order of the k-step BS method, we can write

$$M_2\, \mathbf{c} = \mathbf{b}_y + O\,(h^p), \tag{5.4}$$

where M_2 is the square matrix defined in (3.11) and $\mathbf{b}_y = (y(x_0), \ldots, y(x_{k_1}), y(x_0) + hy'(x_0), \ldots, y(x_N) + hy'(x_N), y(x_{n-k_2+1}), \ldots, y(x_N))^T$. On the other hand, since M_2 is nonsingular, we can also write

$$M_2\, \hat{\mathbf{c}} = \mathbf{b}_{\hat{s}_k},$$

where $\mathbf{b}_{\hat{s}_k}$ is defined similarly to $\mathbf{b}_y$. Now, (5.3) implies that $\mathbf{b}_{\hat{s}_k} = \mathbf{b}_y + O(h^{k+2})$. Thus,

$$M_2(\mathbf{c} - \hat{\mathbf{c}}) = O(h^p).$$

Recalling from section 4.1 that the condition number of M_2 does not depend on N, we can conclude that $\|\mathbf{c} - \hat{\mathbf{c}}\|_\infty = O(h^p)$, and this implies that there exists a suitable constant $\bar{C}_{k,y}$ depending on k and y such that

$$\|s_k - \hat{s}_k\|_\infty \leq \bar{C}_{k,y}\, h^p. \tag{5.5}$$

Now, using (5.2), the statement of the theorem is proved. ∎

6. Numerical results. In order to test the features of the BS methods and of the related spline extensions, we have applied the methods for the numerical solution of classical test problems for which we know the exact solutions. We presents numerical results using both uniform (Problem 1 and 2) and nonuniform (Problem 3) meshes.

The additional $(k_1 - 1) = \lceil \frac{k}{2} \rceil - 1$ left and $k_2 = \lfloor \frac{k}{2} \rfloor$ right conditions have been chosen by requiring that the associated spline extension $\mathbf{s}_k$ verifies the not-a-knot condition at the knots $x_1, \ldots, x_{k_1-1}$ and $x_{N-k_2}, \ldots, x_{N-1}$. As pointed out in section 4, alternative choices are possible, but this one guarantees a quite natural additional request to the spline extension.

In the case of a nonuniform mesh we use a preliminary implementation of the BS methods written in MATLAB that is very similar to the one presented in [18, 20] for the code TOM; they are especially similar concerning the stepsize variation strategy and the solution of the nonlinear equations. The continuous extension is used to compute the new approximation of the solution if the mesh is changed. This approximation is also used in the stopping criteria for the quasi-linearization procedure. We do not give details of this implementation because here we are interested only in showing that these methods can be effectively used for the numerical solution of boundary value problems. The extension of the BS methods to the general case of a nonuniform mesh requires the computation of the variable coefficients of the numerical scheme. Details on this topic are presented in [19].

Problem 1. The first problem is the second order boundary value problem

$$\begin{cases} \epsilon y''(x) = y'(x), \ x \in [0\,,\,1], \\ y(0) = 1, \qquad y(1) = 0. \end{cases} \tag{6.1}$$

with $\epsilon = 0.05$, which corresponds to a moderately and poorly conditioned continuous problem. The results related to this test problem are presented in Table 6.1. For each value of N and k two relative errors are reported, one related to the numerical solution and the other to the spline extension, defined by $Em_N^{(k)} := \max_{i=0,\ldots,N} |s_k(x_i) - y(x_i)| / \max(1, |y(x_i)|)$ and $Es_N^{(k)} := \|(s_k - y)/\max(1, |y|)\|_\infty$, respectively. The latter has always been evaluated sampling both s_k and y on 3000 uniformly spaced points of the integration interval.

Table 6.1 shows both $Em_N^{(k)}$ and $Es_N^{(k)}$ for different values of k. More precisely, each row contains the results for two successive BS methods of the same order (i.e., $Em_N^{(k)}, Es_N^{(k)}$ and $Em_N^{(k+1)}, Es_N^{(k+1)}$ for a fixed even value of k). The columns labeled L_N contain the values of the numerically computed order of convergence for the spline extension. It is evident that for k even the error in the solution is greater with respect to the method of the same order with odd k.

TABLE 6.1
Test problem 1.

k	N	$Em_N^{(k+1)}$	$Es_N^{(k+1)}$	$L_N^{(k+1)}$	$Em_N^{(k)}$	$Es_N^{(k)}$	$L_N^{(k)}$
2	40	$2.36\,10^{-4}$	$2.42\,10^{-4}$	3.96	$1.54\,10^{0}$	$1.55\,10^{0}$	
2	80	$1.11\,10^{-5}$	$1.13\,10^{-5}$	4.42	$8.87\,10^{-2}$	$8.88\,10^{-2}$	4.12
2	160	$4.32\,10^{-7}$	$4.33\,10^{-7}$	4.71	$6.59\,10^{-3}$	$6.60\,10^{-3}$	3.75
2	320	$1.50\,10^{-8}$	$1.50\,10^{-8}$	4.85	$4.54\,10^{-4}$	$4.51\,10^{-4}$	3.87
2	640	$4.96\,10^{-10}$	$4.92\,10^{-10}$	4.93	$2.98\,10^{-5}$	$2.95\,10^{-5}$	3.93
4	40	$1.75\,10^{-5}$	$1.83\,10^{-5}$	5.39	$2.60\,10^{-3}$	$2.60\,10^{-3}$	4.11
4	80	$2.59\,10^{-7}$	$2.63\,10^{-7}$	6.12	$7.87\,10^{-5}$	$7.87\,10^{-5}$	5.04
4	160	$2.80\,10^{-9}$	$2.81\,10^{-9}$	6.54	$1.70\,10^{-6}$	$1.70\,10^{-6}$	5.53
4	320	$2.58\,10^{-11}$	$2.58\,10^{-11}$	6.76	$3.13\,10^{-8}$	$3.13\,10^{-8}$	5.76
4	640	$2.19\,10^{-13}$	$2.17\,10^{-13}$	6.89	$5.30\,10^{-10}$	$5.26\,10^{-10}$	5.90

We can thus conclude that, in general, BS methods with odd k are preferable.

Problem 2. The second test considered is the harmonic oscillator equation with boundary conditions described by

$$\begin{cases} \mathbf{y}'(x) &= A\,\mathbf{y}(x),\ x \in [0\,,\,2\pi], \\ y_1(0) &= 3, \\ y_2(2\pi) &= 3, \end{cases} \tag{6.2}$$

where $A = \begin{pmatrix} 0 & 5 \\ -5 & 0 \end{pmatrix}$. As is known, the solution in the phase plane is a circle (with radius $3\sqrt{2}$). Figures 6.1 and 6.2 show the obtained results. The values used for k are 3 and 5. In both cases the spline extension does not introduce undesirable additional perturbations, but for relatively small mesh sizes it nicely reproduces the shape of the continuous solution.

Problem 3. The third test problem is

$$\begin{cases} \epsilon y''(x) = y(x) + y^2(x) - \exp(-2x/\sqrt{\epsilon}),\ x \in [0\,,\,1], \\ y(0) = 1, \qquad y(1) = \exp(-1/\sqrt{\epsilon}), \end{cases} \tag{6.3}$$

whose exact solution is $y(x) = \exp(-x/\sqrt{\epsilon})$. This is a nonlinear problem that has been solved using a variable stepsize implementation of the numerical schemes with k odd. This problem has been chosen because we know the exact solution, so we can compute the error. Moreover, the numerical solution presents a boundary layer at $x = 0$, and the use on a nonuniform mesh is preferable.

Table 6.2 reports the numerical results obtained for different values of ϵ and different values of the relative and absolute tolerances used to compute the solution. In the example we use $tol = rtol = atol$ and we accept the solution if

$$\max_{i=1,N} |y_i - \hat{y}_i| / \max(atol, rtol|\hat{y}_i|) < 1,$$

where y_i is the computed solution at the mesh points, and $\hat{y}_i$ is an approximation of the exact solution computed using a higher order method in the same class by a deferred correction like procedure. We compare the order 4, 6, and 8 ($k = 3, 5, 7$) schemes. Their behavior seems to be consistent with the order of the methods. For lower values of the tolerances the order 8 scheme reaches the solution using the minimum number of mesh points ($N_{\max}$ in the table). The error at the mesh points and the error in the spline are very similar, so we report only the error at the mesh points (Em in the table).

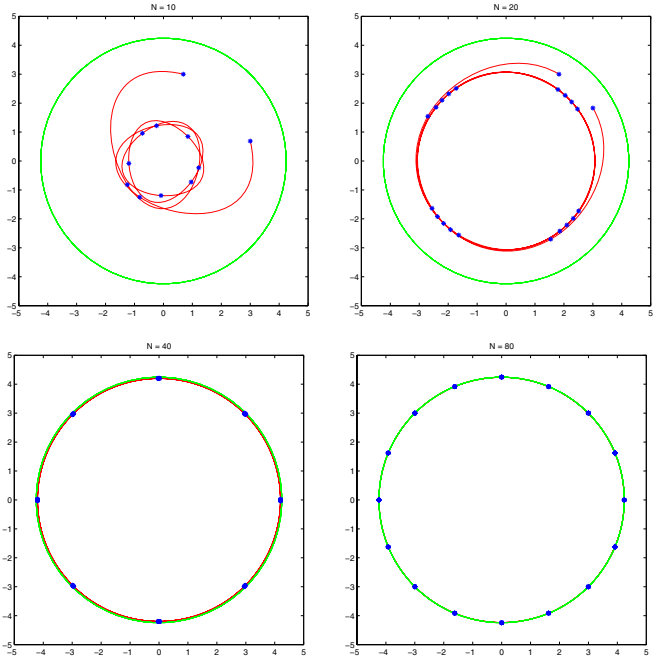


FIG. 6.1. *Test problem* (6.2). *Results for the case when* $k = 3$ *('-' exact solution; '-*' computed solution and mesh points).*

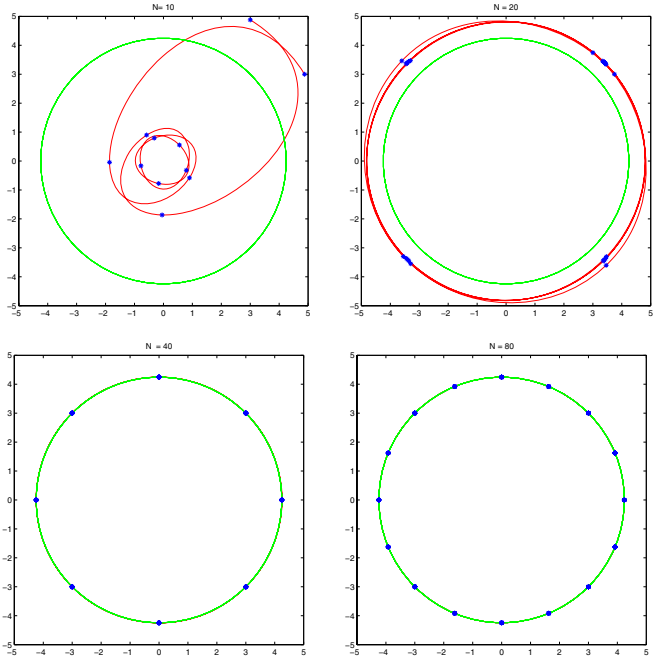


FIG. 6.2. *Test problem* (6.2). *Results for the case when* $k = 5$ *('-' exact solution; '-*' computed solution and mesh points).*

TABLE 6.2
Test problem 3.

	$k=3$			$k=5$			$k=7$		
ϵ	N_{max}	$\frac{h_{\max}}{h_{min}}$	E_m	$N_{\max}$	$\frac{h_{\max}}{h_{\min}}$	E_m	$N_{\max}$	$\frac{h_{\max}}{h_{\min}}$	E_m
	tol = 1e-4								
10^{-2}	21	1.0e0	2.0e-4	21	1.0e0	1.1e-4	21	1.0e0	1.1e-04
10^{-4}	97	2.7e1	3.8e-6	105	2.8e1	3.7e-6	99	2.0e1	3.7e-06
10^{-6}	131	1.3e3	5.7e-7	133	1.3e3	2.9e-7	156	5.1e2	7.1e-08
	tol = 1e-6								
10^{-2}	87	2.5e0	6.4e-8	21	1.0e0	1.5e-5	21	1.0e0	1.9e-06
10^{-4}	169	1.3e1	1.8e-7	105	2.8e1	9.3e-8	99	2.0e1	9.1e-08
10^{-6}	331	6.4e2	2.3e-7	233	3e+2	3.6e-9	192	2.5e2	1.2e-09
	tol = 1e-8								
10^{-2}	173	4.5e0	4.8e-09	41	1.0e0	2.3e-07	41	1.0e0	6e-09
10^{-4}	631	2.7e1	3.5e-09	185	1.4e1	6.8e-10	99	2.0e1	2.7e-09
10^{-6}	1085	2.6e3	1.2e-08	249	3e2	6.0e-10	284	2.1e2	3.4e-11

The behavior of E_m is consistent with the order of the methods. For lower values of the tolerances higher order methods reach the solution using a smaller number of mesh points. As expected, the ratio $h_{\max}/h_{\min}$ increases as ϵ decreases, showing that the BS methods could also be used with the nonuniform mesh. More details and examples about the BS methods on nonuniform grids are presented in [19].

Appendix (alternative proof of the convergence order). We give here an alternative proof that $p \geq k+1$, starting from the following identity, which holds for arbitrary $t, x \in \mathbb{R}$, (see for instance [24, Thm. 4.21, p. 125]):

$$(t - x)^{k+1} = \sum_{i=-\infty}^{+\infty} \phi_{i,k+1}(t)\, B_{k+2}(x-i), \tag{A.1}$$

where $\phi_{i,k+1} \in \Pi_{k+1}$ is the factorial power

$$\phi_{i,k+1}(t) := (t-i-1)^{(k+1)}.$$

Deriving (A.1) successively with respect to t, we get

$$(k+1)^{(j)}\,(t-x)^{k+1-j} = \sum_{i=-\infty}^{+\infty} \frac{d^j\,\phi_{i,k+1}}{d^j\,t}(t)\; B_{k+2}(x-i),\; j=0,\ldots,k+1. \tag{A.2}$$

Assuming that $t = x$, these identities allow us to rederive the well-known partition of unity property of B-splines (obtained for $j = k+1$),

$$\sum_{i=-\infty}^{+\infty} B_{k+2}(x-i) = 1, \tag{A.3}$$

and the identities

$$\sum_{i=-\infty}^{+\infty} \frac{d^j\,\phi_{i,k+1}}{d^j\,x}(x)\; B_{k+2}(x-i) = 0,\; j=0,\ldots,k. \tag{A.4}$$

If we consider $x \in [k+1\,,\,k+2]$, the range for the index i in both (A.3) and (A.4) can be restricted to $i = 0, \ldots, k+1$ and so these identities can be rewritten in matrix form using finite matrices as

$$Z(x)\,\mathbf{B}(x) \;=\; (1, 0, \ldots, 0)^T, \tag{A.5}$$

where $\mathbf{B}(x) := (B_{k+2}(x), \ldots, B_{k+2}(x-k-1))^T$ and where $Z(x)$ is a square matrix of order $k+2$ defined as

$$Z(x) \;:=\; \begin{pmatrix} 1 & \cdots & 1 \\ \frac{d^k\,\phi_{0,k+1}}{d^k\,x}(x) & \cdots & \frac{d^k\,\phi_{k+1,k+1}}{d^k\,x}(x) \\ \vdots & \vdots & \vdots \\ \phi_{0,k+1}(x) & \cdots & \phi_{k+1,k+1}(x) \end{pmatrix}.$$

Now each $\phi_{i,k+1}(x)$ is a polynomial of degree $k+1$ and can be expressed in terms of the classical power basis as

$$\phi_{i,k+1}(x) = \sum_{r=0}^{k+1} S_r^{(k+1)}\,(x-i-1)^r,$$

where the coefficients $S_r^{(k+1)}, r = 0, \ldots, k+1$, are the Stirling numbers of first kind. Thus, there exists a nonsingular constant lower triangular matrix L such that

$$Z(x) \;=\; L\,W(x), \tag{A.6}$$

where $W(x)$ is the Vandermonde matrix

$$W(x) \;:=\; \begin{pmatrix} 1 & \cdots & 1 \\ x-1 & \cdots & x-k-2 \\ \vdots & \vdots & \vdots \\ (x-1)^{k+1} & \cdots & (x-k-2)^{k+1} \end{pmatrix},$$

and where the nonzero elements of L are $L_{1,1} \;=\; 1$ and $L_{i,j} \;=\; S_{j+k-i+1}^{(k+1)}\,(j+k-i+1)^{(k+2-i)}, j \le i\,,\; i = 2, \ldots, k+2$.

Replacing (A.6) in (A.5), we get $W(x)\,\mathbf{B}(x) \;=\; L^{-1}(1, 0, \ldots, 0)^T$ and, deriving with respect to x, we obtain the formula

$$W'(x)\,\mathbf{B}(x) \;+\; W(x)\,\mathbf{B}'(x) = 0, \tag{A.7}$$

where the meanings of $W'(x)$ and $\mathbf{B}'(x)$ are obvious. Considering that

$$W'(x) \;=\; \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & \cdots & \vdots \\ 0 & 2 & \ddots & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & k+1 & 0 \end{pmatrix} W(x)\,,$$

and that

$$\begin{cases} \mathbf{B}(k+2) & = (0, \beta_0^{(k)}, \ldots, \beta_k^{(k)})^T \;=\; \quad (0, \beta_k^{(k)}, \ldots, \beta_0^{(k)})^T, \\ \mathbf{B}'(k+2) & = (0, \alpha_0^{(k)}, \ldots, \alpha_k^{(k)})^T \;=\; -\quad (0, \alpha_k^{(k)}, \ldots, \alpha_0^{(k)})^T, \end{cases}$$

evaluating (A.7) in $x = k + 2$, we get the usual $k + 1$ order conditions (expressed in matrix form, see, e.g., [5]).

Acknowledgments. The authors are grateful to the referees for several helpful comments.

REFERENCES

[1] P. Amodio and L. Brugnano, *The conditioning of Toeplitz band matrices*, Math. Comput. Modelling, 23 (1996), pp. 29–42.
[2] A. Gluchoff and F. Hartmann, *Univalent polynomials and non-negative trigonometric sums*, Amer. Math. Monthly, 105 (1998), pp. 508–522.
[3] U. M. Ascher, R. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for ODEs*, Classics Appl. Math. B, SIAM, Philadelphia, 1995.
[4] U. M. Ascher and R. J. Spiteri, *Collocation software for boundary value differential-algebraic equations*, SIAM J. Sci. Comput., 15 (1994), pp. 938–952.
[5] L. Brugnano and D. Trigiante, *Solving Differential Problems by Multistep Initial and Boundary Value Methods*, Gordon and Breach Science Publishers, Amsterdam, 1998.
[6] L. Brugnano and D. Trigiante, *Convergence and stability of boundary value methods for ordinary differential equations*, J. Comput. Appl. Math., 66 (1996), pp. 97–109.
[7] J. R. Cash and F. Mazzia, *A new mesh selection algorithm, based on conditioning, for two-point boundary value codes*, J. Comput. Appl. Math., 184 (2005), pp. 362–381.
[8] J. Cash, F. Mazzia, N. Sumarti, and D. Trigiante, *The role of conditioning in mesh selection algorithms for first order systems of linear two-point boundary value problems*, J. Comput. Appl. Math., 185 (2006), pp. 212–224.
[9] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
[10] W. H. Enright, K. R. Jackson, S. P. Nørsett, and P. G. Thomsen, *Interpolants for Runge–Kutta formulas*, ACM Trans. Math. Software, 12 (1986), pp. 193–218.
[11] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations*, Springer-Verlag, Berlin, Heidelberg, 1993.
[12] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 1, Wiley Classics Library, John Wiley and Sons, New York, 1988.
[13] F. Iavernaro, F. Mazzia, and D. Trigiante, *Eigenvalues and quasi-eigenvalues of banded Toeplitz matrices: Some properties and applications*, Numer. Algorithms, 31 (2002), pp. 157–170.
[14] F. Iavernaro and D. Trigiante, *Preconditioning and conditioning of systems arising from boundary value methods*, Nonlinear Dyn. Syst. Theory, 1 (2001), pp. 59–79.
[15] S. Karlin and Z. Ziegler, *Chebyshevian spline functions*, SIAM J. Numer. Anal., 3 (1966), pp. 514–543.
[16] F. R. Loscalzo and T. D. Talbot, *Spline function approximations for solutions of ordinary differential equations*, SIAM J. Numer. Anal., 4 (1967), pp. 433–445.
[17] F. R. Loscalzo, *An introduction to the application of spline functions to initial value problems*, in Theory and Applications of Spline Functions, T. N. E. Greville, ed., Academic Press, New York, 1969, pp. 37–64.
[18] F. Mazzia and I. Sgura, *Numerical approximation of nonlinear BVPs by means of BVMs*, Appl. Numer. Math., 42 (2002), pp. 337–352.
[19] F. Mazzia, A. Sestini, and D. Trigiante, *BS linear multistep methods on non-uniform meshes*, Journal of Numerical Analysis, Industrial and Applied Mathematics, 1 (2006), pp. 129–142.
[20] F. Mazzia, D. Trigiante, *A hybrid mesh selection strategy based on conditioning for boundary value ODE problems*, Numer. Algorithms, 36 (2004), pp. 169–187.
[21] D. S. Mitrinović, J. E. Pečarić, and A. M. Fink, *Classical and New Inequalities in Analysis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
[22] S. Sallam and M. Naim Anwar, *Stabilized cubic C^1 spline collocation method for solving first-order ordinary initial value problems*, Int. J. Comput. Math., 74 (2000), pp. 87–96.
[23] S. Sallam, *Stable quartic spline integration method for solving stiff ordinary differential equations*, Appl. Math. Comput., 116 (2000), pp. 245–255.
[24] L. L. Schumaker, *Spline Functions: Basic Theory*, J. Wiley and Sons, New York, 1981.
[25] J. H. Verner and M. Zennaro, *Continuous explicit Runge–Kutta methods of order* 5, Math. Comp., 64 (1995), pp. 1123–1146.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1974–1996

WEIGHTED-NORM FIRST-ORDER SYSTEM LEAST SQUARES (FOSLS) FOR PROBLEMS WITH CORNER SINGULARITIES*

E. LEE†, T. A. MANTEUFFEL†, AND C. R. WESTPHAL‡

Abstract. A weighted-norm least-squares method is considered for the numerical approximation of solutions that have singularities at the boundary. While many methods suffer from a global loss of accuracy due to boundary singularities, the least-squares method can be particularly sensitive to a loss of regularity. The method we describe here requires only a rough lower bound on the power of the singularity and can be applied to a wide range of elliptic equations. Optimal order discretization accuracy is achieved in weighted H^1, and functional norms and L^2 accuracy are retained for boundary value problems with a dominant div/curl operator. Our analysis, including interpolation bounds and several Poincaré-type inequalities, are carried out in appropriately weighted Sobolev spaces. Numerical results confirm the error bounds predicted in the analysis.

Key words. least squares, finite element method, singularities, weighted norm, weighted Sobolev space

AMS subject classifications. 65N30, 65N12, 65F10, 35F15

DOI. 10.1137/050636279

1. Introduction.

In this paper, we develop a method for treating div/curl systems with reduced regularity. While motivated by first-order systems that arise from second-order elliptic boundary value problems, div/curl systems appear in many contexts, such as for example, Maxwell's equations and the vorticity form of Stokes equations. These problems have the fortunate property of a guaranteed smooth solution as long as the data and domain are smooth. However, many problems of interest are posed in nonsmooth domains and, as a consequence, lose this property at a finite number of points on the boundary in two dimensions or along curves on the boundary in three dimensions. In the present paper, we study two-dimensional problems that have nonsmooth solutions at *irregular boundary points*, that is, points that are corners of polygonal domains, locations of changing boundary condition type, or both. Similar behavior occurs in problems with discontinuous material coefficients, and the methods presented here can easily be extended to that situation.

Standard solution techniques that attempt to approximately solve a div/curl system with reduced regularity using H^1-conforming finite elements will, in general, fail to converge. This phenomenon can be explained by noting that the Sobolev space $(H^1)^2$ is a closed subspace of $H(\mathrm{div})\cap H(\mathrm{curl})$ (see section 2), which implies either that $(H^1)^2 = H(\mathrm{div}) \cap H(\mathrm{curl})$, the case of full regularity, or $(H^1)^2$ is a proper subspace. In this case, the codimension is finite and is spanned by so-called singular functions. In the presence of reduced regularity, the solution will, in general, not be in $(H^1)^2$. A standard finite element method using H^1-conforming elements will converge to the

*Received by the editors July 18, 2005; accepted for publication (in revised form) May 25, 2006; published electronically October 4, 2006.

http://www.siam.org/journals/sinum/44-5/63627.html

†Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO 80309-0526 (eunjung@colorado.edu, tmanteuf@colorado.edu). The work of the second author was sponsored by the National Science Foundation under grant DMS–0420873 and by the Department of Energy under grants DE–FC02-01ER25479 and DE–FG02-03ER25574.

‡Department of Mathematics and Computer Science, Wabash College, P.O. Box 352, Crawfordsville, IN 47933 (westphac@wabash.edu). The work of this author was sponsored by the National Science Foundation under grant DMS–9810751.

element of $(H^1)^2$ closest to the true solution in the $H(\mathrm{div}) \cap H(\mathrm{curl})$ norm. Local mesh refinement will not alter this outcome.

If a basis for the singular functions is known, it can be incorporated directly into the finite element space (cf. [10, 16, 24]). In [3, 4], this approach is applied to div/curl systems and shown to restore optimal convergence throughout the domain at a minimal additional cost. For some two-dimensional problems, the singular basis functions are known and can be included in the finite element space. For the other problems, or in three dimensions, the exact character of the singular functions is less understood, which makes this approach more difficult to implement.

As a different type of remedy for this so-called pollution effect, least-squares methods based on inverse norms can be effective for problems with irregular boundary points, discontinuous coefficients, and data in H^{-1}. For example, in [5, 8, 11, 12, 13, 19], the functional is posed in terms of H^{-1}-norms rather than L^2-norms, resulting in optimal L^2-approximations to the solution. A more recent approach, called FOSLL*, uses an inverse norm induced by the equations and is shown in [14, 22] to be more efficient than the H^{-1}-norm methods.

Graded mesh refinement in weighted Sobolev spaces has been shown to be effective in restoring optimal convergence, in L^2- and H^1-norms, in the context of a Galerkin formulation of second-order elliptic problems with reduced regularity (cf. [2, 1]). However, in that context, the solution is in H^1. Convergence would occur, although more slowly, without weighting and mesh refinement.

In [15], a weighted norm and a sequence of graded meshes is used in an $H(\mathrm{div})$ least-squares functional, arising from a second-order elliptic problem, to restore optimal convergence for the primal variable, in both L^2- and H^1-norms, if the flux variable is approximated in a finite element space satisfying the grid decomposition property, for example, Raviart–Thomas elements (cf. [6]). The flux variable converges in a weighted $H(\mathrm{div})$-norm.

In this paper we examine div/curl systems that lack regularity within the first-order system least-squares (FOSLS) framework. The basic FOSLS approach is to recast the original system as an appropriate first-order system and apply an L^2 minimization principle over the residual of the equations. If possible, this reformulation is done by minimizing a functional whose quadratic part is equivalent to the product H^1-norm, indicating that the process is similar to solving a weakly coupled system of Poisson-like equations. This equivalence also guarantees optimal H^1-accuracy for standard discretizations. For div/curl systems with reduced regularity, as briefly mentioned above, the L^2-based functional fails to be H^1-equivalent and, as a consequence, standard discretizations suffer from the pollution effect. Here, a weighted-norm least-squares method is developed that restores optimal convergence using H^1-conforming finite elements without graded mesh refinement. It replaces the L^2-norms in the FOSLS functional with weighted L^2-norms, making the functional norm equivalent to a weighted H^1-norm. With an appropriate weighting function, this method recovers optimal order accuracy in the weighted L^2- and H^1-norms and retains optimal L^2 convergence even near the singularity. Our method requires only the power of the singularity (not the actual singular solution) to be known a priori and, in practice, can be used with only a rough estimate of the power of the singularity, which can be adaptively determined if unknown (cf. [4]).

The method developed here has some similarity to [15] but considers an $H(\mathrm{div}) \cap H(\mathrm{curl})$ functional and considers more general boundary conditions. Most important, our analysis admits a more aggressive weighting, resulting in optimal order accuracy in the weighted norms without mesh refinement. In addition, we prove several Poincaré-

type inequalities in weighted Sobolev spaces under a variety of boundary conditions that, in addition to being necessary for our main result, may be of independent interest.

We use Poisson's equation on a domain with a reentrant corner as a model problem and as the formal setting for analysis. The resulting div/curl system is the focus. The analysis in this paper is restricted to two dimensions. However, the approach suggests a natural generalization to problems with reduced regularity due to discontinuous coefficients and to problems in three dimensions.

This paper is organized as follows. Section 2 contains notation and a preliminary discussion. In section 3, the weighted FOSLS functionals are described. Section 4 contains Poincaré inequalities and regularity results in weighted Sobolev spaces. Error bounds for the weighted FOSLS functional are presented in section 5, and section 6 contains computational results.

2. Singular solutions and preliminaries. For vector function $\mathbf{u} = (u_1, u_2)^t$, let the divergence and curl of $\mathbf{u}$ be defined in the standard way: $\nabla \cdot \mathbf{u} = \partial_x u_1 + \partial_y u_2$ and $\nabla \times \mathbf{u} = \partial_x u_2 - \partial_y u_1$. Further, define the formal adjoint of the curl operator by

$$\nabla^{\perp} q = \begin{pmatrix} \partial_y q \\ -\partial_x q \end{pmatrix}.$$

We use standard notation for Sobolev spaces $H^k(\Omega)^d$, corresponding inner product $(\cdot,\cdot)_{k,\Omega}$, and norm $\|\cdot\|_{k,\Omega}$, for $k \geq 0$. We drop subscript Ω and superscript d when the domain and dimension are clear by context. Since $H^0(\Omega)$ coincides with $L^2(\Omega)$ we often denote $\|\cdot\|_0$ by $\|\cdot\|$. Define the subspaces of $L^2(\Omega)$ induced by the divergence and curl of $\mathbf{u}$ by

$$H(\mathrm{div}) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \cdot \mathbf{u}\| < \infty\},$$
$$H(\mathrm{curl}) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \times \mathbf{u}\| < \infty\}.$$

We also make use of the following general inequalities for nonnegative a and b:

$$|a|^2 + |b|^2 \leq |a+b|^2 \leq 2(|a|^2 + |b|^2). \tag{2.1}$$

Consider the function $f(r,\theta) = r^a$ in two-dimensional polar coordinates. Assume that the origin lies on the boundary of domain

$$\Omega_w = \{(r,\theta) : 0 < r < R,\ 0 < \theta < \omega < 2\pi\},$$

as pictured in Figure 2.1. By a direct computation it is clear that $f \in H^k(\Omega)$ only for $k < a + 1$.

Now, consider Poisson's equation on a domain in $\mathbb{R}^2$ with a corner of interior angle ω. It is well known that, for the case of Dirichlet or Neumann boundary conditions, the solutions of this boundary value problem may include those with radial part of the form $p \sim r^{\frac{\pi}{\omega}}$ in a local polar coordinate system centered at the corner. Thus, for the case of reentrant corners, $\omega > \pi$, the solution fails to be in $H^2(\Omega)$ and we say that the problem has a singularity (or singular solution). For problems with Dirichlet and Neumann boundary conditions meeting at the corner, solutions may have components of the form $p \sim r^{\frac{\pi}{(2\omega)}}$. Thus, for mixed boundary conditions, singularities may occur at corners with $\omega > \pi/2$. We now explore this issue in more detail.

Define the power of the singularity to be $\alpha = \pi/\omega$ for Dirichlet or Neumann boundary conditions and $\alpha = \pi/(2\omega)$ for mixed boundary conditions. The solution

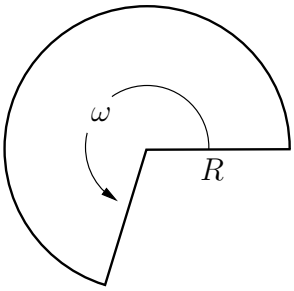


FIG. 2.1. *Simple wedge-shaped domain, Ω_w.*

to Poisson's equation may be written as

$$p(r,\theta) = p_0(r,\theta) + s(r,\theta),$$

where $p_0(r,\theta) \in H^2(\Omega)$ and $s(r,\theta) \in H^{1+m}(\Omega)$ for $m < \alpha$. The singular part of the solution has the form

$$s(r,\theta) = r^\alpha(\kappa_1 \sin(\alpha\theta) + \kappa_2 \cos(\alpha\theta)),$$

where the values of κ_1 and κ_2 depend on boundary conditions (see [17, 18]).

For the FOSLS formulation of this problem, we may similarly decompose unknown $\mathbf{u} = \nabla p$ as

$$\mathbf{u}(r,\theta) = \mathbf{u}_0(r,\theta) + \nabla s(r,\theta),$$

where $\nabla s(r,\theta)$ has the form

$$\nabla s(r,\theta) = \alpha r^{\alpha-1} \begin{pmatrix} \kappa_1 \sin(\alpha-1)\theta + \kappa_2 \cos(\alpha-1)\theta \\ \kappa_1 \cos(\alpha-1)\theta - \kappa_2 \sin(\alpha-1)\theta \end{pmatrix}.$$

Thus, the unknown $\mathbf{u}(r,\theta)$ is in $H^k(\Omega)$ only for $k < \alpha$.

For example, consider Poisson's equation posed on the simple domain in Figure 2.1. Let the solution to this boundary value problem in polar coordinates be $p = \chi(r) r^{\frac{2}{3}} \sin(2\theta/3)$, where $\chi(r)$ is a smooth transition function that is 1 on a platform near the origin and vanishes at the boundaries not adjacent to the origin. Then, $p = 0$ on $\partial\Omega$ and

$$\begin{aligned} \Delta p &= \frac{1}{r}\partial_r(r\partial_r p) + \frac{1}{r^2}\partial^2_{\theta\theta} p \\ &= \left(r^{\frac{2}{3}}\chi''(r) + \frac{7}{3} r^{\frac{-1}{3}} \chi'(r) \right) \sin(2\theta/3), \end{aligned}$$

and, thus, it is clear that $\Delta p \in L^2(\Omega)$, but $p \notin H^2(\Omega)$. We say this problem fails to provide *full lifting* of the data (from $L^2(\Omega)$ to $H^2(\Omega)$, for example). The solution, $\mathbf{u} = \nabla p$, is thus not in $H^1(\Omega)$.

3. Weighted-norm least squares. As before, let Ω be a domain with a corner of interior angle ω at the origin, and we may, without loss of generality, further assume $diam(\Omega) \leq 1$. For $f \in L^2(\Omega)$, let p satisfy

$$\begin{cases} -\Delta p = f & \text{in} \quad \Omega, \\ p = 0 & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot \nabla p = 0 & \text{on} \quad \Gamma_N, \end{cases} \tag{3.1}$$

where $\mathbf{n}$ is the outward unit normal to Ω and $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$. When this problem is H^2 regular, the normal FOSLS methodology is to introduce the new unknown, $\mathbf{u} = \nabla p$, and rewrite system (3.1) as

$$\begin{cases} \mathbf{u} - \nabla p = 0 & \text{in} \quad \Omega, \\ -\nabla \cdot \mathbf{u} = f & \text{in} \quad \Omega, \\ \nabla \times \mathbf{u} = 0 & \text{in} \quad \Omega, \\ \boldsymbol{\tau} \cdot \mathbf{u} = 0 & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot \mathbf{u} = 0 & \text{on} \quad \Gamma_N, \\ p = 0 & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot \nabla p = 0 & \text{on} \quad \Gamma_N. \end{cases} \tag{3.2}$$

Here, $\boldsymbol{\tau}$ is the counterclockwise unit tangential vector to Ω. Since this system can be posed completely in terms of $\mathbf{u}$, we may decouple the equations in (3.2), solve for $\mathbf{u}$ first, and then recover p from $\mathbf{u}$. To this end, define the two L^2-norm functionals,

$$\begin{aligned} G(\mathbf{u}; f) &= \|\nabla \cdot \mathbf{u} + f\|^2 + \|\nabla \times \mathbf{u}\|^2, \\ G_2(p; \mathbf{u}) &= \|\mathbf{u} - \nabla p\|^2, \end{aligned}$$

and the spaces,

$$\begin{aligned} \mathcal{V} &= \{\mathbf{v} \in H^1(\Omega) : \ \boldsymbol{\tau} \cdot \mathbf{v} = 0 \ \text{ on } \ \Gamma_D, \ \ \mathbf{n} \cdot \mathbf{v} = 0 \ \text{ on } \ \Gamma_N\}, \\ \mathcal{W} &= \{q \in H^1(\Omega) : \ q = 0 \ \text{ on } \ \Gamma_D\}. \end{aligned}$$

Thus, the two-stage solution process is to minimize $G(\mathbf{v}; f)$ over $\mathcal{V}$ and then, given the minimizer $\mathbf{u}$, minimize G_2 over $\mathcal{W}$:

$$G(\mathbf{u}; f) = \inf_{\mathbf{v} \in \mathcal{V}} G(\mathbf{v}; f), \tag{3.3a}$$

$$G_2(p; \mathbf{u}) = \inf_{q \in \mathcal{W}} G_2(q; \mathbf{u}). \tag{3.3b}$$

The goal of the FOSLS methodology is, generally, to formulate functionals whose quadratic part is equivalent to the H^1-norm whenever possible. The second stage functional is H^1-equivalent and the solution we seek is always in H^1. The first stage functional, however, is not always H^1-equivalent. For domains with reentrant corners, there is no H^1 sequence of functions that converges to the solution in the $H(\text{div}) \cap H(\text{curl})$ norm. To see an illustration, consider the example above, where $p = \chi(r) r^{\frac{2}{3}} \sin(2\theta/3)$ and $\mathbf{u} = \nabla p$. A simple computation reveals that $\nabla \cdot \mathbf{u}, \nabla \times \mathbf{u} \in L^2(\Omega)$, but $\mathbf{u} \notin H^1(\Omega)$.

Define the weighted functional by

$$G_w(\mathbf{u}; f) = \|w(\nabla \cdot \mathbf{u} + f)\|^2 + \|w \nabla \times \mathbf{u}\|^2, \tag{3.4}$$

where the weight function has the form $w = r^\beta$ for some $\beta > 0$.

Define the weighted Sobolev norm, $\|\cdot\|_{k,\beta}$, on Ω in terms of the standard L^2 norm, $\|\cdot\|_0$, by

$$\|q\|_{k,\beta} = \left(\sum_{|j| \le k} \|r^{\beta - k + |j|} D^j q\|_0^2 \right)^{\frac{1}{2}}, \tag{3.5}$$

where D^j is the standard distributional derivative of order j. Similarly, define the weighted seminorm by

$$|q|_{k,\beta} = \left(\sum_{|j|=k} \|r^{\beta-k+|j|} D^j q\|_0^2 \right)^{\frac{1}{2}} \tag{3.6}$$

and the associated weighted Sobolev space by

$$H^k_\beta(\Omega) = \{q : \|q\|_{k,\beta} < \infty\}. \tag{3.7}$$

Define the div/curl operator, L, and vector $\mathbf{f}$ by $L = \begin{pmatrix} \nabla \cdot \\ \nabla \times \end{pmatrix}$ and $\mathbf{f} = \begin{pmatrix} f \\ 0 \end{pmatrix}$. We may now write the weighted functional from (3.4) as

$$G_w(\mathbf{u}; f) = \|L\mathbf{u} - \mathbf{f}\|^2_{0,\beta}.$$

The weighted-norm least-squares minimization problem for the first-stage solution is then the following: Find $\mathbf{u} \in \mathcal{V}$ such that

$$G_w(\mathbf{u}; f) = \inf_{\mathbf{v} \in \mathcal{V}} G_w(\mathbf{v}; f).$$

The second-stage solution for p remains as described above. We seek values of β that make $H^1(\Omega)$ dense in $H(\text{div}) \cap H(\text{curl})$ in the weighted functional norm and result in the most accurate discretizations possible.

For the discrete problem, we may choose any finite-dimensional subset of H^1 over which to minimize the weighted functional. Let $\mathcal{P}^h$ denote the space of C^0 piecewise polynomial (or tensor product) elements on triangles (or quadrilaterals) of meshsize h, and let $\mathcal{V}^h$ denote the subspace of $\mathcal{P}^h$ that satisfies the appropriate boundary conditions on Ω:

$$\mathcal{V}^h = \{\mathbf{v}^h \in \mathcal{P}^h : \ \boldsymbol{\tau} \cdot \mathbf{v}^h = 0 \ \text{ on } \ \Gamma_D, \ \ \mathbf{n} \cdot \mathbf{v}^h = 0 \ \text{ on } \ \Gamma_N\}.$$

The discrete weighted-norm least-squares minimization problem is, then, to minimize the discrete functional as follows: Find $\mathbf{u}^h \in \mathcal{V}^h$ such that

$$G_w(\mathbf{u}^h; f) = \min_{\mathbf{v}^h \in \mathcal{V}^h} G_w(\mathbf{v}^h; f). \tag{3.8}$$

By unweighting the equations near the singularity, the functional is freed from trying to approximate the solution (which is not in $H^1(\Omega)$) in the H^1 sense near the singularity. But, away from the singularity, the weighted functional retains the same character as the normal nonweighted functional. We now consider the choice of weight parameter β and its relation to weighted and nonweighted a priori error bounds on the approximated solution.

4. Poincaré bounds and regularity estimates. In this section, we establish several theoretical results in weighted Sobolev spaces and error bounds for the weighted-norm method.

Here, we establish several Poincaré bounds in the domain Ω_w. We prove first a result for the scalar pure Neumann and pure Dirichlet problems and then for the scalar mixed boundary condition problem. These results lead to a Poincaré inequality for the vector case.

LEMMA 4.1. *Take $\Omega = \Omega_w$ and let $\beta > 0$, $\epsilon > 0$, and $\gamma = \beta - 3/2 - \epsilon$. Further, assume that $\gamma \neq -2$. For functions $q \in H^1_\beta(\Omega)$, with $r^{\beta-\epsilon}\nabla q, r^{\gamma+1}\nabla q \in L^2(\Omega)$, that can be chosen to satisfy*

$$\iint_\Omega r^\gamma \, q(r,\theta)\, r \, dr \, d\theta = 0, \tag{4.1}$$

we have the bound

$$\|q\|_{0,\beta-1} \leq C(\epsilon)\|\nabla q\|_{0,\beta-\epsilon}, \tag{4.2}$$

where $C(\epsilon)$ depends on ϵ, β, and Ω and $C(\epsilon) \to \infty$ as $\epsilon \to 0$.

Proof. For any two points (r,θ) and (r_0,θ_0) in Ω, write $q(r,\theta)$ as

$$q(r,\theta) = q(r_0,\theta_0) + \int_{r_0}^{r} \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, d\hat{r} + \int_{\theta_0}^{\theta} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\hat{\theta}.$$

Multiplying both sides of the equation by $r_0^{\gamma+1}$, integrating with respect to r_0 and θ_0 over Ω, and using Fubini's theorem yield

$$\begin{aligned}
&\frac{R^{\gamma+2}\omega}{\gamma+2} q(r,\theta) \\
&= \int_0^\omega \int_0^R \int_{r_0}^r r_0^{\gamma+1} \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, d\hat{r}\, dr_0\, d\theta_0 + \int_0^\omega \int_0^R \int_{\theta_0}^\theta r_0^{\gamma+1} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\hat{\theta}\, dr_0\, d\theta_0 \\
&= \omega \int_0^r \int_0^{\hat{r}} r_0^{\gamma+1} \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, dr_0\, d\hat{r} - \omega \int_r^R \int_{\hat{r}}^R r_0^{\gamma+1} \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, dr_0\, d\hat{r} \\
&\quad + \int_0^R \int_0^\theta \int_0^{\hat{\theta}} r_0^{\gamma+1} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\theta_0\, d\hat{\theta}\, dr_0 - \int_0^R \int_\theta^\omega \int_{\hat{\theta}}^\omega r_0^{\gamma+1} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\theta_0\, d\hat{\theta}\, dr_0 \\
&= \frac{\omega}{\gamma+2} \int_0^R \hat{r}^{\gamma+2} \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, d\hat{r} - \frac{\omega R^{\gamma+2}}{\gamma+2} \int_r^R \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\, d\hat{r} \\
&\quad + \int_0^R \int_0^\omega \hat{\theta}\, r_0^{\gamma+1} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\hat{\theta}\, dr_0 - \omega \int_0^R \int_0^\omega r_0^{\gamma+1} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}}\, d\hat{\theta}\, dr_0.
\end{aligned}$$

Note from above that q is, by assumption, sufficiently smooth to apply Fubini's theorem. By the triangle inequality, we have that

$$\begin{aligned}
&\left| \frac{R^{\gamma+2}\omega}{\gamma+2} q(r,\theta) \right| \\
&\leq \frac{\omega}{\gamma+2} \int_0^R \hat{r}^{\gamma+2} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| d\hat{r} + \frac{\omega R^{\gamma+2}}{\gamma+2} \int_r^R \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| d\hat{r} \\
&\quad + 2\omega \int_0^R \int_0^\omega r_0^{\gamma+1} \left| \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}} \right| d\hat{\theta}\, dr_0.
\end{aligned}$$

Now, squaring each side and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we get

$$
\begin{aligned}
|q(r,\theta)|^2 \le{}& \frac{3}{R^{2(\gamma+2)}} \left(\int_0^R \hat{r}^{\gamma+2} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 + 3 \left(\int_r^R \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 \\
&+ \frac{12(\gamma+2)^2}{R^{2(\gamma+2)}} \left(\int_0^R \int_0^\omega r_0^{\gamma+1} \left| \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}} \right| \, d\hat{\theta} \, dr_0 \right)^2 .
\end{aligned} \tag{4.3}
$$

Multiply each side of (4.3) by $r^{2\beta-1}$ and integrate with respect to r and θ over Ω. We consider each of the terms on the resulting right-hand side separately. First,

$$
\begin{aligned}
&\int_0^\omega \int_0^R r^{2\beta-1} \left(\int_0^R \hat{r}^{\gamma+2} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 dr \, d\theta \\
&\quad \le R \int_0^\omega \int_0^R \int_0^R r^{2\beta-1} \, \hat{r}^{2\gamma+4} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r} \, dr \, d\theta \\
&\quad = \frac{R^{2\beta+1}}{2\beta} \int_0^\omega \int_0^R \hat{r}^{2\gamma+3} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 \hat{r} \, d\hat{r} \, d\theta \\
&\quad \le \frac{R^{2\beta+1}}{2\beta} \|\nabla q\|^2_{0,\gamma+\frac{3}{2}}.
\end{aligned}
$$

We now consider the second term in (4.3). Since by the Schwarz inequality,

$$
\begin{aligned}
&\left(\int_r^R \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 \\
&\quad = \left(\int_r^R \hat{r}^{(\epsilon-1)/2} \hat{r}^{(1-\epsilon)/2} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 \\
&\quad \le \int_r^R \hat{r}^{\epsilon-1} \, d\hat{r} \int_r^R \hat{r}^{1-\epsilon} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r} \\
&\quad = \frac{R^\epsilon}{\epsilon} \int_r^R \hat{r}^{1-\epsilon} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r},
\end{aligned}
$$

we can apply Fubini's theorem to bound the second term:

$$
\begin{aligned}
&\int_0^\omega \int_0^R r^{2\beta-1} \left(\int_r^R \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right| \, d\hat{r} \right)^2 dr \, d\theta \\
&\quad \le \frac{R^\epsilon}{\epsilon} \int_0^\omega \int_0^R \int_r^R r^{2\beta-1} \, \hat{r}^{1-\epsilon} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r} \, dr \, d\theta \\
&\quad = \frac{R^\epsilon}{\epsilon} \int_0^\omega \int_0^R \int_0^{\hat{r}} r^{2\beta-1} \, \hat{r}^{1-\epsilon} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 dr \, d\hat{r} \, d\theta \\
&\quad = \frac{R^\epsilon}{2\epsilon\beta} \int_0^\omega \int_0^R \hat{r}^{2\beta-\epsilon} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 \hat{r} \, d\hat{r} \, d\theta \\
&\quad = \frac{R^\epsilon}{2\epsilon\beta} \|\nabla q\|^2_{0,\beta-\frac{\epsilon}{2}} \\
&\quad \le \frac{R^\epsilon}{2\epsilon\beta} \|\nabla q\|^2_{0,\beta-\epsilon},
\end{aligned}
$$

The third term can be bounded similarly:

$$\begin{aligned}
&\int_0^\omega \int_0^R r^{2\beta-1} \left(\int_0^R \int_0^\omega r_0^{\gamma+1} \left| \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}} \right| \, d\hat{\theta}\, dr_0 \right)^2 dr\, d\theta \\
&\quad \leq \left(\int_0^R \int_0^\omega r^{2\beta-1} \, dr\, d\theta \right) R\omega \int_0^\omega \int_0^R r_0^{2\gamma+2} \left| \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}} \right|^2 \, dr_0\, d\hat{\theta} \\
&\quad = \frac{R^{2\beta+1}\omega^2}{2\beta} \int_0^\omega \int_0^R r_0^{2\gamma+3} \left| \frac{1}{r_0} \frac{\partial q(r_0,\hat{\theta})}{\partial \hat{\theta}} \right|^2 r_0 \, dr_0\, d\hat{\theta} \\
&\quad = \frac{R^{2\beta+1}\omega^2}{2\beta} \|\nabla q\|^2_{0,\gamma+\frac{3}{2}}.
\end{aligned}$$

Putting the three terms together and substituting $\gamma = \beta - 3/2 - \epsilon$, we may now write the bound

$$\begin{aligned}
&\int_0^\omega \int_0^R r^{2\beta-2} \left| q(r,\theta) \right|^2 r \, dr\, d\theta \\
&\quad \leq \left(\frac{3R^{-2\epsilon}}{2\beta} + \frac{3R^{\epsilon}}{2\epsilon\beta} + \frac{6(\beta+\frac{1}{2}-\epsilon)^2 R^{-2\epsilon}\omega^2}{\beta} \right) \|\nabla q\|^2_{0,\beta-\epsilon},
\end{aligned}$$

and the lemma follows by taking the square root of both sides. ∎

In what follows, let χ be a smooth function of r, where $\chi = 1$ for $r < \eta$ and $\chi = 0$ for $r > 2\eta$. We take η to be sufficiently small to ensure that $\mathrm{supp}(\chi) \subset \Omega$.

LEMMA 4.2. *Take $\Omega = \Omega_w$ and let q be a scalar function in $H^1_\beta(\Omega)$, where $\beta > 0$. The following bound holds for χ as defined above:*

$$\|\chi q\|_{0,\beta-1} \leq \frac{1}{\beta} \|\nabla(\chi q)\|_{0,\beta}. \tag{4.4}$$

Proof. Hardy's inequality for $f(t)$ defined for $t > 0$ with $\lim_{t\to 0} f(t) = 0$ gives (see [20])

$$\int_0^\infty \frac{f^2}{t^2} \, dt \leq 4 \int_0^\infty |f'|^2 \, dt. \tag{4.5}$$

The lemma follows after a change of variables, $t = r^{-2\beta}$, a substitution $f(r) = \chi q(r,\theta)$ for fixed θ, and an integration on both sides with respect to θ. ∎

LEMMA 4.3. *Take $\Omega = \Omega_w$ and let either $q \in H^1_\beta(\Omega)$ with $q = 0$ on $\partial\Omega$ or $q \in H^1_\beta(\Omega)/\mathbb{R}$ with $n \cdot \nabla q = 0$ on $\partial\Omega$. Then*

$$\|q\|_{0,\beta-1} \leq C \|\nabla q\|_{0,\beta} \tag{4.6}$$

for $\beta > 0$, where C depends only on Ω and β.

Proof. First, if $q = 0$ on $\partial\Omega$, write $q = q(r,\theta)$ as

$$q(r,\theta) = \int_0^\theta \frac{\partial q(r,\hat{\theta})}{\partial \hat{\theta}} \, d\hat{\theta}.$$

Square both sides and multiply by $r^{2\beta-1}$:

$$r^{2\beta-1}\,|q(r,\theta)|^2 = r^{2\beta-1}\left|\int_0^\theta \frac{\partial q(r,\hat{\theta})}{\partial\hat{\theta}}\,d\hat{\theta}\right|^2 \le r^{2\beta+1}\omega\int_0^\omega\left|\frac{1}{r}\frac{\partial q(r,\hat{\theta})}{\partial\hat{\theta}}\right|^2 d\hat{\theta}.$$

Integrate both sides with respect to r and θ over Ω:

$$\int_0^\omega\int_0^R r^{2\beta-2}\,|q(r,\theta)|^2\;r\,dr\,d\theta \le \omega\int_0^\omega\int_0^{R_0} r^{2\beta+1}\int_0^\omega\left|\frac{1}{r}\frac{\partial q(r,\hat{\theta})}{\partial\hat{\theta}}\right|^2 d\hat{\theta}\;dr\,d\theta \le \omega^2\int_0^\omega\int_0^{R_0} r^{2\beta}\left|\frac{1}{r}\frac{\partial q(r,\hat{\theta})}{\partial\hat{\theta}}\right|^2 r\,dr\,d\hat{\theta}.$$

The lemma follows since the right-hand side is bounded by $C\|\nabla q\|^2_{0,\beta}$.

Now, if $q \in H^1_\beta(\Omega)/\mathbb{R}$, then it may be chosen to satisfy

$$\iint_\Omega r^\gamma\, q\, r\,dr\,d\theta = 0 \tag{4.7}$$

for γ chosen as in Lemma 4.1. By the triangle inequality and Lemma 4.2, we get

$$\begin{aligned}\|q\|_{0,\beta-1} &\le \|\chi\, q\|_{0,\beta-1} + \|(1-\chi)q\|_{0,\beta-1}\\ &\le C\|\nabla(\chi q)\|_{0,\beta} + \left(\int_0^\omega\int_\eta^R\left(\frac{r}{\eta}\right)^2 r^{2\beta-2}(1-\chi)^2\,q^2\,r\,dr\,d\theta\right)^{\frac{1}{2}}\\ &\le C\left(\|\nabla(q)\|_{0,\beta} + \|q\|_{0,\beta}\right).\end{aligned}$$

Apply Lemma 4.1 with $\epsilon = 1$ to the $\|q\|_{0,\beta}$ term on the right-hand side, and the Lemma follows. ∎

For the problem with mixed boundary conditions, consider Ω_w partitioned into subdomains $\Omega_0 = \{(r,\theta) : r \le \frac{1}{2}R_0, 0 \le \theta \le \omega\}$ and $\Omega_1 = \Omega\backslash\Omega_0$, as shown in Figure 4.1.

LEMMA 4.4. *Consider domain $\Omega = \Omega_w$, as pictured in Figure* 4.1*, and let $q \in H^1_\beta(\Omega)$ vanish on the line segment of $\partial\Omega$ corresponding to $\theta = 0$ and $r < R_0$. Then there is a constant, C, dependent only on Ω, β, and R_0, such that*

$$\|q\|_{0,\beta-1} \le C\|\nabla q\|_{0,\beta} \tag{4.8}$$

for $\beta > 0$.

Proof. For points (r,θ) in Ω_0 we may derive the bound,

$$\|q\|_{0,\beta-1,\Omega_0} \le C\|\nabla q\|_{0,\beta,\Omega_0}, \tag{4.9}$$

completely analogous to the proof of Lemma 4.3. Now, consider points (r,θ) in Ω_1. We may write $q = q(r,\theta)$ as

$$q(r,\theta) = \int_{\tilde{r}}^r \frac{\partial q(\hat{r},\theta)}{\partial\hat{r}}\,d\hat{r} + \int_0^\theta \frac{\partial q(\tilde{r},\hat{\theta})}{\partial\hat{\theta}}\,d\hat{\theta},$$

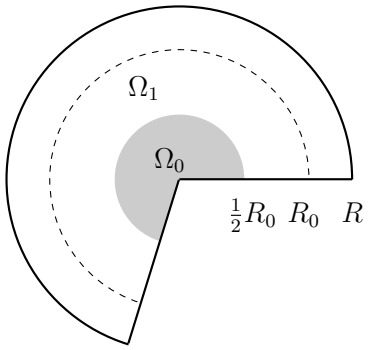


FIG. 4.1. *Wedge-shaped domain, Ω_w, partitioned into subdomains Ω_0 and Ω_1.*

where the point $(\tilde{r}, 0)$ is on the part of $\partial\Omega_1$ where q vanishes. By the Schwarz inequality, the triangle inequality, and inequality (2.1), we have the bound

$$|q(r,\theta)|^2 \leq 2(R - \frac{1}{2}R_0) \int_{\tilde{r}}^{r} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r} + 2\omega \int_0^{\theta} \left| \frac{\partial q(\tilde{r},\hat{\theta})}{\partial \hat{\theta}} \right|^2 d\hat{\theta}. \tag{4.10}$$

We now expand the limits in the integrals, multiply each side by $r^{2\beta-1}$, integrate with respect to r over $(\frac{1}{2}R_0, R)$, integrate with respect to θ over $(0, \omega)$, and integrate with respect to $\tilde{r}$ over $(\frac{1}{2}R_0, R_0)$, and apply Fubini's theorem to get

$$\begin{aligned} &\left(\frac{1}{2}R_0\right) \int_0^{\omega} \int_{\frac{1}{2}R_0}^{R} r^{2\beta-1} |q(r,\theta)|^2 \, dr \, d\theta \\ &\quad \leq 2\left(R - \frac{1}{2}R_0\right) \int_{\frac{1}{2}R_0}^{R_0} \int_0^{\omega} \int_{\frac{1}{2}R_0}^{R} \int_{\frac{1}{2}R_0}^{R} r^{2\beta-1} \left| \frac{\partial q(\hat{r},\theta)}{\partial \hat{r}} \right|^2 d\hat{r} \, dr \, d\theta \, d\tilde{r} \\ &\quad + 2\omega \int_{\frac{1}{2}R_0}^{R_0} \int_0^{\omega} \int_{\frac{1}{2}R_0}^{R} \int_0^{\theta} \left| \frac{\partial q(\tilde{r},\hat{\theta})}{\partial \hat{\theta}} \right|^2 d\hat{\theta} \, dr \, d\theta \, d\tilde{r}. \end{aligned} \tag{4.11}$$

We use the inequalities

$$\frac{1}{2}R_0 \leq \tilde{r} \leq R_0, \quad \tilde{r} \leq \hat{r} \leq r \leq R$$

to derive the following simple bounds:

$$1 \leq \left(\frac{2\hat{r}}{R_0}\right), \quad r \leq \left(\frac{2R}{R_0}\right)\hat{r}, \quad r \leq \left(\frac{2R}{R_0}\right)\tilde{r}. \tag{4.12}$$

By applying the bounds in (4.12) and by Fubini's theorem, we may now write (4.11) as

$$\begin{aligned}
&\left(\frac{1}{2}R_0\right)\int_0^{\omega}\int_{\frac{1}{2}R_0}^{R} r^{2\beta-2}|q(r,\theta)|^2 r\; dr\, d\theta \\
&\quad\le \left(R-\frac{1}{2}R_0\right)^2 R_0 \int_0^{\omega}\int_{\frac{1}{2}R_0}^{R}\left(\frac{2\hat{r}}{R_0}\right)^2\left(\frac{2R}{R_0}\right)^{2\beta-1}\hat{r}^{2\beta-1}\left|\frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\right|^2 d\hat{r}\, d\theta \\
&\qquad + 2\omega^2\left(R-\frac{1}{2}R_0\right)\int_{\frac{1}{2}R_0}^{R_0}\int_0^{\omega}\left(\frac{2R}{R_0}\right)^{2\beta-1}\tilde{r}^{2\beta-1}\left|\frac{\partial q(\tilde{r},\hat{\theta})}{\partial \hat{\theta}}\right|^2 d\hat{\theta}\, d\tilde{r} \\
&\quad\le 2^{2\beta+1}\left(R-\frac{1}{2}R_0\right)^2 R_0^{-2\beta}R^{2\beta-1}\int_0^{\omega}\int_{\frac{1}{2}R_0}^{R}\hat{r}^{2\beta}\left|\frac{\partial q(\hat{r},\theta)}{\partial \hat{r}}\right|^2 \hat{r}\; d\hat{r}\, d\theta \\
&\qquad + 2^{2\beta}\omega^2\left(R-\frac{1}{2}R_0\right)\left(\frac{R}{R_0}\right)^{2\beta-1}\int_{\frac{1}{2}R_0}^{R}\int_0^{\omega}\tilde{r}^{2\beta}\left|\frac{1}{\tilde{r}}\frac{\partial q(\tilde{r},\hat{\theta})}{\partial \hat{\theta}}\right|^2 \tilde{r}\; d\hat{\theta}\, d\tilde{r},
\end{aligned}$$

which directly implies

$$\|q\|_{0,\beta-1,\Omega_1} \le C\|\nabla q\|_{0,\beta,\Omega_1}, \tag{4.13}$$

where

$$C = 2^{2\beta+1}\left(R-\frac{1}{2}R_0\right)R^{2\beta-1}R_0^{-2\beta-1}\left(2(R-\frac{1}{2}R_0)+\frac{\omega^2}{R_0}\right).$$

Combining inequalities (4.9) and (4.13) completes the lemma. ∎

We now consider a similar Poincaré inequality for the vector case. Again, consider $\Omega = \Omega_w$, where $\partial\Omega$ is partitioned into Dirichlet and Neumann boundaries Γ_D and Γ_N, respectively. The following lemma is valid for the pure Dirichlet and Neumann cases and for the mixed boundary condition cases when Γ_D includes a part of the boundary adjacent to the origin and $\omega \neq \frac{3\pi}{2}$.

LEMMA 4.5. *Take $\Omega = \Omega_w$ and let $\mathbf{u} \in H^1_\beta(\Omega)^2$ satisfy $\boldsymbol{\tau}\cdot\mathbf{u} = 0$ on Γ_D and $\mathbf{n}\cdot\mathbf{u} = 0$ on Γ_N. Assume for the mixed boundary condition case that $\omega \neq 3\pi/2$. Then there is a constant, C, dependent only on Ω, β, and the length of the segments of Γ_D and Γ_N adjacent to the origin, such that*

$$\|\mathbf{u}\|_{0,\beta-1} \le C\|\nabla\mathbf{u}\|_{0,\beta} \tag{4.14}$$

for $\beta > 0$.

Proof. First, consider the case when $\boldsymbol{\tau}\cdot\mathbf{u} = 0$ on $\partial\Omega$. Denote the part of $\partial\Omega$ aligned with $\theta = 0$ as Γ_1 and the part of $\partial\Omega$ aligned with $\theta = \omega$ as Γ_2. Thus, $u_1 = 0$ on Γ_1 and $\tau_x u_1 + \tau_y u_2 = 0$ on Γ_2. Since u_1 and $\tau_x u_1 + \tau_y u_2$ satisfy the conditions in Lemma 4.4, we may use

$$\|u_1\|_{0,\beta-1} \le C\|\nabla u_1\|_{0,\beta}$$

and

$$\|\tau_x u_1 + \tau_y u_2\|_{0,\beta-1} \le C\|\nabla(\tau_x u_1 + \tau_y u_2)\|_{0,\beta}.$$

Further, take $\tau_y \neq 0$, since $\tau_y = 0$ corresponds to either $\omega = \pi$, for which the result holds trivially since the boundary is smooth, or $\omega = 2\pi$, which we do not consider. Now,

$$\begin{aligned}
\|\mathbf{u}\|^2_{0,\beta-1} &= \|u_1\|^2_{0,\beta-1} + \|u_2\|^2_{0,\beta-1} \\
&= \|u_1\|^2_{0,\beta-1} + \frac{1}{\tau_y^2}\|\tau_x u_1 - \tau_x u_1 + \tau_y u_2\|^2_{0,\beta-1} \\
&\leq \left(1 + 2\frac{\tau_x^2}{\tau_y^2}\right)\|u_1\|^2_{0,\beta-1} + \frac{2}{\tau_y^2}\|\tau_x u_1 + \tau_y u_2\|^2_{0,\beta-1} \\
&\leq C(\|\nabla u_1\|^2_{0,\beta} + \|\nabla(\tau_x u_1 + \tau_y u_2)\|^2_{0,\beta}) \\
&\leq C\|\nabla \mathbf{u}\|^2_{0,\beta}.
\end{aligned}$$

The case when $\mathbf{n}\cdot\mathbf{u} = 0$ on $\partial\Omega$ is analogous since $u_2 = 0$ on Γ_1 and $n_x u_1 + n_y u_2 = 0$ on Γ_2. Also, when $\omega \neq \frac{\pi}{2}, \frac{3\pi}{2}$, the case for mixed boundary conditions follows similarly using the result of Lemma 4.4. The case for mixed boundary conditions when $\omega = \pi/2$ follows from appealing to symmetry in pure Dirichlet problem for $\omega = \pi$. ∎

Remark 4.6. Lemma 4.5 can be directly extended to more generally shaped domains. The proof of the scalar Poincaré bounds in Lemmas 4.1, 4.2, 4.3, and 4.4 is simplified when the domain has the shape of Ω_w with only one irregular boundary point. Since we are primarily interested in a local result, proving Lemma 4.5 in the simple domain is sufficient for our purposes.

Consider the following scalar Poisson problem in Ω_w:

$$\begin{cases} \Delta p = f & \text{in} \quad \Omega, \\ p = 0 & \text{on} \quad \Gamma_D, \\ \mathbf{n}\cdot\nabla p = 0 & \text{on} \quad \Gamma_N. \end{cases} \tag{4.15}$$

We refer to system (4.15) as the pure Dirichlet problem when $\partial\Omega = \Gamma_D$; the pure Neumann problem when $\partial\Omega = \Gamma_N$; and the mixed boundary condition problem when Γ_D includes the part of $\partial\Omega$ coinciding with one of either $\theta = 0$ or $\theta = \omega$, and $\Gamma_N = \partial\Omega\backslash\Gamma_D$ with $\Gamma_N \neq \emptyset$.

The following regularity results can be found in [23] and [21].

LEMMA 4.7. *Assume $|1-\beta| < \pi/\omega$ for the pure Dirichlet problem, $0 < |1-\beta| < \pi/\omega$ for the pure Neumann problem, and $|1-\beta| < \pi/2\omega$ for the mixed boundary condition problem. Then, for every $f \in H^0_\beta(\Omega)$, there exists a unique solution to* (4.15)*, $p \in H^2_\beta(\Omega)$ for the pure Dirichlet and mixed boundary condition cases, and $p \in H^2_\beta(\Omega)/\mathbb{R}$ for the pure Neumann problem. Moreover, there exists a constant, C, independent of p, such that*

$$\|p\|_{2,\beta} \leq C\|f\|_{0,\beta}. \tag{4.16}$$

Proof. See Chapter 1 of [23] for the Dirichlet and Neumann problems and Chapter 2 of [21] for the mixed boundary problem. ∎

Define the subspace of functions in $H^1_\beta(\Omega)$ satisfying the appropriate boundary conditions by

$$\mathcal{V}_\beta = \{\mathbf{v} \in H^1_\beta(\Omega) : \boldsymbol{\tau}\cdot\mathbf{v} = 0 \text{ on } \Gamma_D, \ \mathbf{n}\cdot\mathbf{v} = 0 \text{ on } \Gamma_N\}.$$

We now prove a regularity result for functions in $\mathcal{V}_\beta$. Recall that the power of the singularity is defined as $\alpha = \pi/\omega$ for Dirichlet or Neumann boundary conditions and $\alpha = \pi/(2\omega)$ for mixed boundary conditions.

LEMMA 4.8. *Consider domain $\Omega = \Omega_w$. Then there is a positive constant, C, independent of $\mathbf{u}$, such that, for $|1-\beta| < \alpha$, the following bound holds for all $\mathbf{u} \in \mathcal{V}_\beta$:*

$$\|\mathbf{u}\|_{1,\beta} \leq C\|L\mathbf{u}\|_{0,\beta}. \tag{4.17}$$

Proof. From Lemma 4.7 we know that any $\mathbf{u} \in \mathcal{V}_\beta$ has the decomposition

$$\mathbf{u} = \nabla\phi + \nabla^\perp\psi, \tag{4.18}$$

where $\phi, \psi \in H^2_\beta(\Omega)$ satisfy

$$\begin{cases} \Delta\phi = \nabla \cdot \mathbf{u} & \text{in} \quad \Omega, \\ \phi = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{4.19}$$

and

$$\begin{cases} \Delta\psi = \nabla \times \mathbf{u} & \text{in} \quad \Omega, \\ \mathbf{n} \cdot \nabla\psi = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{4.20}$$

Then, by applying Lemma 4.7 to problems (4.19) and (4.20) we have

$$\begin{aligned} \|\mathbf{u}\|_{1,\beta} &= \|\nabla\phi + \nabla^\perp\psi\|_{1,\beta} \leq \|\nabla\phi\|_{1,\beta} + \|\nabla^\perp\psi\|_{1,\beta} \\ &\leq \|\phi\|_{2,\beta} + \|\psi\|_{2,\beta} \leq C(\|\nabla \cdot \mathbf{u}\|_{0,\beta} + \|\nabla \times \mathbf{u}\|_{0,\beta}) \leq C\|L\mathbf{u}\|_{0,\beta}, \end{aligned}$$

which completes the proof. ∎

5. Error bounds. Let $\mathcal{T}^h = \cup_{i=1}^N \tau_i$ be a quasi-uniform triangulation of polygonal domain Ω. Let $\mathcal{I}^h$ represent standard interpolation onto a piecewise polynomial finite element space of degree k. From finite element theory, we have the following interpolation bounds.

LEMMA 5.1. *Let Ω be a polygonal domain. There exists a constant, C, independent of v, such that, for all $v \in H^m(\Omega)$,*

$$\left(\sum_{\tau \in \mathcal{T}^h} \|v - \mathcal{I}^h v\|^2_{s,\tau} \right)^{1/2} \leq Ch^{m-s}|v|_m \tag{5.1}$$

for $0 \leq s \leq m$ and $1 < m$. Here, $\mathcal{I}^h$ denotes interpolation by a piecewise polynomial of degree $k = m-1$. (Note that here the norm $\|\cdot\|_{s,\tau}$ is the standard $H^s(\tau)$ norm.)

Proof. See [9] or [7]. ∎

We now consider a weighted interpolation bound for functions on domains with a polygonal corner at the origin. Define the modified interpolation operator, $\mathcal{I}^h_0$, by

$$\mathcal{I}^h_0 u|_\tau \;=\; \begin{cases} \mathcal{I}^h u = \sum_{i=0}^n u(a_i)\phi_i & \text{if } \overline{\tau} \text{ does not intersect the origin,} \\ \sum_{i=1}^n u(a_i)\phi_i & \text{if } \overline{\tau} \text{ intersects the origin,} \end{cases}$$

where $\mathcal{I}^h$ is a standard polynomial interpolation operator, ϕ_i are basis functions corresponding to the $n+1$ nodal points, a_i, and a_0 is the origin, $(0,0)$. Thus, the

modified interpolation has a value of zero at the origin and resembles $\mathcal{I}^h$ away from the origin.

LEMMA 5.2. *Let Ω be a polygonal domain. There exists a constant, C, independent of u, such that, for all $u \in H^m_\beta(\Omega)$ satisfying* (4.8),

$$\left(\sum_{\tau \in \mathcal{T}^h} \|u - \mathcal{I}_0^h u\|^2_{1,\beta,\tau} \right)^{1/2} \leq Ch^{m-1} \|u\|_{m,\beta}, \tag{5.2}$$

for $1 < m$ and $\beta > 0$, where $\mathcal{I}_0^h$ is the modified interpolation operator onto piecewise polynomials of degree $k = m - 1$ defined above.

Proof. Define $K_0 = \{\tau \mid \bar{\tau} \cap (0,0) \neq \emptyset\}$ as the set of elements adjacent to the origin. On $\mathcal{T}^h \backslash K_0$, we have $h \leq r_{min} \leq r = \sqrt{x^2 + y^2} \leq r_{max} \leq r_{min} + \sqrt{2}h$ with $r_{min} = \inf\{r | (x,y) \in \tau\}$ and $r_{max} = \sup\{r | (x,y) \in \tau\}$ in τ, and

$$\begin{aligned}
\|u - \mathcal{I}_0^h u\|^2_{1,\beta,\tau} &= \|u - \mathcal{I}^h u\|^2_{1,\beta,\tau} \\
&= \int_\tau r^{2\beta} |\nabla(u - \mathcal{I}^h u)|^2 + r^{2(\beta-1)} |u - \mathcal{I}^h u|^2 d\tau \\
&\leq r_{max}^{2\beta} \int_\tau |\nabla(u - \mathcal{I}^h u)|^2 d\tau + r_{max}^{2\beta} r_{min}^{-2} \int_\tau |u - \mathcal{I}^h u|^2 d\tau \\
&\leq C r_{max}^{2\beta} h^{2(m-1)} |u|^2_{m,0,\tau} + C r_{max}^{2\beta} r_{min}^{-2} h^{2m} |u|^2_{m,0,\tau} \\
&= C r_{max}^{2\beta} h^{2(m-1)} (1 + r_{min}^{-2} h^2) |u|^2_{m,0,\tau} \leq C r_{max}^{2\beta} h^{2(m-1)} |u|^2_{m,0,\tau} \\
&\leq C h^{2(m-1)} r_{max}^{2\beta} r_{min}^{-2\beta} \int_\tau r^{2\beta} |D^m u|^2 d\tau \\
&\leq C h^{2(m-1)} \left(\frac{r_{min} + \sqrt{2}h}{r_{min}} \right)^{2\beta} \int_\tau r^{2\beta} |D^m u|^2 d\tau \\
&\leq C h^{2(m-1)} \int_\tau r^{2\beta} |D^m u|^2 d\tau.
\end{aligned}$$

We now consider the case for which $\tau \in K_0$. Let $\delta \in \mathcal{C}^\infty$ be a cut-off function defined by

$$\delta(r) = \begin{cases} 1 & \text{if} \quad r \leq h/3, \\ 0 & \text{if} \quad r > 2h/3, \end{cases}$$

with $|\delta^{(m)}| \leq ch^{-m}$, where $\delta^{(m)}$ is the mth derivative of δ. By the triangle inequality,

$$\|u - \mathcal{I}_0^h u\|_{1,\beta,\tau} \leq \|\delta u - \mathcal{I}_0^h(\delta u)\|_{1,\beta,\tau} + \|(1-\delta)u - \mathcal{I}_0^h((1-\delta)u)\|_{1,\beta,\tau}. \tag{5.3}$$

By the definition of δ we have $\mathcal{I}_0^h((1-\delta)u) = \mathcal{I}^h((1-\delta)u)$ and $\mathcal{I}_0^h(\delta u) = 0$. For the second term in (5.3), we apply Lemmas 4.4 and 5.1, the properties in δ, and Fubini's

theorem to obtain

$$\begin{aligned}
&\|(1-\delta)u-\mathcal{I}_0^h((1-\delta)u)\|_{1,\beta,\tau}^2=\|(1-\delta)u-\mathcal{I}^h((1-\delta)u)\|_{1,\beta,\tau}^2\\
&\le C\|\nabla((1-\delta)u-\mathcal{I}^h((1-\delta)u))\|_{0,\beta,\tau}^2\le Ch^{2\beta}\int_\tau|\nabla((1-\delta)u-\mathcal{I}^h((1-\delta)u))|^2d\tau\\
&\le Ch^{2(\beta+m-1)}\int_\tau|D^m((1-\delta)u)|^2d\tau\le Ch^{2(\beta+m-1)}\int_\tau\sum_{|j|\le m}|D^{m-j}(1-\delta)D^ju|^2d\tau\\
&\le Ch^{2(\beta+m-1)}\left(\iint_{\frac{h}{3}}^{\frac{2h}{3}}\sum_{|j|\le m-1}|h^{|j|-m}D^ju|^2d\tau+\iint_{\frac{h}{3}}^{r(\theta)}|(1-\delta)D^mu|^2d\tau\right)\\
&\le Ch^{2(\beta+m-1)}\left(\sum_{|j|\le m-1}\iint_{\frac{h}{3}}^{\frac{2h}{3}}h^{-2\beta}r^{2(\beta+|j|-m)}|D^ju|^2d\tau+\iint_{\frac{h}{3}}^{r(\theta)}\Big(\frac{r}{h}\Big)^{2\beta}|D^mu|^2d\tau\right)\\
&=Ch^{2(m-1)}\sum_{|j|\le m}\int_\tau r^{2(\beta+|j|-m)}|D^ju|^2d\tau=Ch^{2(m-1)}\|u\|_{m,\beta,\tau}^2.
\end{aligned}$$

Using the properties of δ and Fubini's theorem results in a similar bound for the first term in (5.3):

$$\begin{aligned}
&\|\delta u-\mathcal{I}_0^h(\delta u)\|_{1,\beta,\tau}^2=\|\delta u\|_{1,\beta,\tau}^2=\int_\tau r^{2\beta}|\nabla(\delta u)|^2+r^{2(\beta-1)}|\delta u|^2d\tau\\
&\le C\int_\tau r^{2\beta}(|\nabla\delta\cdot u|^2+|\delta\nabla u|^2)+r^{2(\beta-1)}|\delta u|^2d\tau\\
&\le C\iint_{\frac{h}{3}}^{\frac{2h}{3}}r^{2\beta}h^{-2}|u|^2d\tau+C\iint_0^{\frac{2h}{3}}r^{2\beta}|\nabla u|^2+r^{2(\beta-1)}|u|^2d\tau\\
&\le C\iint_{\frac{h}{3}}^{\frac{2h}{3}}r^{2(\beta-1)}|u|^2d\tau+C\iint_0^{\frac{2h}{3}}r^{2\beta}|\nabla u|^2+r^{2(\beta-1)}|u|^2d\tau\\
&\le C\int_\tau r^{2(m-1)}(r^{2(\beta-m+1)}|\nabla u|^2+r^{2(\beta-m)}|u|^2)d\tau\le Ch^{2(m-1)}\|u\|_{m,\beta,\tau}^2.
\end{aligned}$$

Thus we have

$$\sum_{\tau\in\mathcal{T}_h}\|u-\mathcal{I}_0^hu\|_{1,\beta,\tau}^2\le Ch^{2(m-1)}\sum_{\tau\in\mathcal{T}_h}\|u\|_{m,\beta,\tau}^2\le Ch^{2(m-1)}\|u\|_{m,\beta}^2,$$

and the lemma follows. ∎

LEMMA 5.3. *Assume* (4.14) *holds in* Ω. *Then, for all* $\mathbf{u}^h\in\mathcal{V}^h$,

$$\|\mathbf{u}^h\|_{0,\beta}\le Ch^{-\eta}\|\mathbf{u}^h\|_{0,\beta+\eta} \tag{5.4}$$

for $\beta>-1$ *and* $\eta>0$.

Proof. Using Lemma 4.5 and an inverse inequality, we may write

$$\|\mathbf{u}^h\|_{0,\beta}\le C\|\nabla\mathbf{u}^h\|_{0,\beta+1}\le Ch^{-1}\|\mathbf{u}^h\|_{0,\beta+1},$$

which establishes (5.4) for $\eta=1$. Repeated application of this inequality thus validates (5.4) for any positive integer. Now consider

$$\begin{aligned}
\|\mathbf{u}^h\|_{0,\beta}^2&=\langle r^\beta\mathbf{u}^h,r^\beta\mathbf{u}^h\rangle=\langle r^{\beta-1/2}\mathbf{u}^h,r^{\beta+1/2}\mathbf{u}^h\rangle\\
&\le\|\mathbf{u}^h\|_{0,\beta-1/2}\|\mathbf{u}^h\|_{0,\beta+1/2}\le Ch^{-1}\|\mathbf{u}^h\|_{0,\beta+1/2}^2.
\end{aligned}$$

Taking the square root establishes (5.4) for $\eta = 1/2$. Repeating these steps leads to (5.4) for all $\eta = \eta_n = k_n/2^{\ell_n}$ for any nonnegative integers k_n and ℓ_n. For any $\eta > 0$, choose a monotonically decreasing sequence, $\{\eta_n\}$, such that $\eta_n > \eta$ and $\lim_{n\to\infty} |\eta_n - \eta| = 0$. Now, $g_n = (r^{\beta+\eta_n}\mathbf{u}^h)^2$ is a monotonically increasing function that converges to $g = (r^{\beta+\eta}\mathbf{u}^h)^2$ pointwise everywhere. Thus, by the Lebesgue monotone convergence theorem, we have

$$\|\mathbf{u}^h\|^2_{0,\beta+\eta} = \int g dx = \lim_n \int g_n dx = \lim_n \|\mathbf{u}^h\|^2_{0,\beta+\eta_n}$$

and, therefore,

$$\begin{aligned}\|\mathbf{u}^h\|_{0,\beta} &= \lim_n \|\mathbf{u}^h\|_{0,\beta} \\ &\le \lim_n Ch^{-\eta_n}\|\mathbf{u}^h\|_{0,\beta+\eta_n} \\ &= (\lim_n Ch^{-\eta_n})(\lim_n \|\mathbf{u}^h\|_{0,\beta+\eta_n}) \\ &= Ch^{-\eta}\|\mathbf{u}^h\|_{0,\beta+\eta},\end{aligned}$$

which completes the proof. ∎

Define an *irregular boundary point* of polygonal domain Ω to be a point on $\partial\Omega$, where interior angle ω satisfies $\omega > \pi$ when Dirichlet or Neumann boundary conditions are applied on both sides of the point or $\omega > \pi/2$ when one Dirichlet boundary condition and one Neumann boundary meet at the corner. We now present error bounds for the numerical solution in weighted and unweighted norms.

THEOREM 5.4. *Let Ω be a polygonal domain with one irregular boundary point of interior angle ω and let $\mathbf{f} \in L^2(\Omega)$. Suppose $\mathbf{u} \in \mathcal{V}$ satisfy $L\mathbf{u} = \mathbf{f}$. If $\mathbf{u}^h \in \mathcal{V}^h$ is chosen to minimize the weighted functional,*

$$G_w(\mathbf{u}^h;\mathbf{f}) = \|L\mathbf{u}^h - \mathbf{f}\|^2_{0,\beta} = \inf_{\mathbf{v}^h\in\mathcal{V}^h} \|L\mathbf{v}^h - \mathbf{f}\|^2_{0,\beta},$$

for $|1-\beta| < \alpha$, then the approximation error, $\mathbf{u} - \mathbf{u}^h$, satisfies the following bounds:

$$\|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} \le Ch^{\alpha+\beta-1}\|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{5.5}$$

$$G_w(\mathbf{u} - \mathbf{u}^h;\mathbf{0})^{\frac{1}{2}} \le Ch^{\alpha+\beta-1}\|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{5.6}$$

$$\|\mathbf{u} - \mathbf{u}^h\|_{0,\beta} \le Ch^{s+\beta}\|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{5.7}$$

$$\|\mathbf{u} - \mathbf{u}^h\|_0 \le Ch^s\|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{5.8}$$

where $s < \alpha$, for $\alpha + \beta \le k+1$ with k the degree of the piecewise polynomial elements in $\mathcal{V}^h$.

Proof. By Lemmas 4.8 and 5.2, we have

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} &\le C\|L(\mathbf{u} - \mathbf{u}^h)\|_{0,\beta} \le C\|L(\mathbf{u} - \mathcal{I}^h_0\mathbf{u})\|_{0,\beta} \\ &\le C\|\mathbf{u} - \mathcal{I}^h_0\mathbf{u}\|_{1,\beta} \le Ch^{\alpha+\beta-1}\|\mathbf{u}\|_{\alpha+\beta,\beta},\end{aligned}$$

which establishes both (5.5) and (5.6) since we may write

$$\|L(\mathbf{u} - \mathbf{u}^h)\|_{0,\beta} = G_w(\mathbf{u} - \mathbf{u}^h;\mathbf{0})^{\frac{1}{2}}.$$

Note that Lemmas 4.8 and 5.2 are satisfied for $|1-\beta| < \alpha$ and $\alpha + \beta \le 2$.

For the weighted L^2-norm, we write

$$\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta}^2 = \sum_{\tau\in\mathcal{T}^h} \|\mathbf{u}-\mathbf{u}^h\|_{0,\beta,\tau}^2.$$

The Cauchy inequality yields, for any $\epsilon \in (0,1)$,

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta,\tau}^2 &= \int_\tau r^{2\beta}|\mathbf{u}-\mathbf{u}^h|^2 d\tau \le \left(\int_\tau 1 d\tau\right)^{1-\epsilon}\left(\int_\tau \left(r^{2\beta}|\mathbf{u}-\mathbf{u}^h|^2\right)^{\frac{1}{\epsilon}} d\tau\right)^{\epsilon}\\ &\le Ch^{2-2\epsilon}\left(\int_\tau \left(r^\beta|\mathbf{u}-\mathbf{u}^h|\right)^{\frac{2}{\epsilon}} d\tau\right)^{\frac{\epsilon}{2}\cdot 2} = Ch^{2-2\epsilon}\|r^\beta(\mathbf{u}-\mathbf{u}^h)\|_{L^{\frac{2}{\epsilon}}(\tau)}^2.\end{aligned}$$

Since $H^1(\Omega)$ is continuously imbedded into $L^q(\Omega)$ for all $q\in[1,\infty)$,

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta}^2 &= \sum_{\tau\in\mathcal{T}^h}\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta,\tau}^2 \le Ch^{2-2\epsilon}\sum_{\tau\in\mathcal{T}^h}\|r^\beta(\mathbf{u}-\mathbf{u}^h)\|_{L^{\frac{2}{\epsilon}}(\tau)}^2\\ &\le Ch^{2-2\epsilon}\|r^\beta(\mathbf{u}-\mathbf{u}^h)\|_{L^{\frac{2}{\epsilon}}(\Omega)}^2 \le Ch^{2-2\epsilon}\|r^\beta(\mathbf{u}-\mathbf{u}^h)\|_{H^1(\Omega)}^2\\ &\le Ch^{2-2\epsilon}\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta}^2.\end{aligned}$$

Thus, by (5.5), we have

$$\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta} \le Ch^{1-\epsilon}\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta} \le Ch^{s+\beta}\|\mathbf{u}\|_{\alpha+\beta,\beta},$$

where any $s<\alpha$.

We now consider the bound on $\|\mathbf{u}-\mathbf{u}^h\|_0$. Let $K_0=\{\tau \mid \bar\tau\cap(0,0)\neq\emptyset\}$ and $K_1=\mathcal{T}^h\backslash K_0$. First, we consider the case $\beta<1$. If $\tau\in K_0$, then $r\le Ch$ and $r^{1-\beta}\le Ch^{1-\beta}$. Thus, we have

$$\|\mathbf{u}-\mathbf{u}^h\|_{0,\tau}^2 \le Ch^{2(1-\beta)}\|\mathbf{u}-\mathbf{u}^h\|_{0,\beta-1,\tau}^2 \le Ch^{2(1-\beta)}\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta,\tau}^2. \tag{5.9}$$

If $\tau\in K_1$, we use the technique above to get

$$\|\mathbf{u}-\mathbf{u}^h\|_{0,\tau}^2 \le Ch^{2(1-\epsilon)}\|\mathbf{u}-\mathbf{u}^h\|_{L^{\frac{2}{\epsilon}}(\tau)}^2.$$

Again, since H^1 is continuously imbedded into L^q for all $q\in[1,\infty)$, we have

$$\begin{aligned}&\sum_{\tau\in K_1}\|\mathbf{u}-\mathbf{u}^h\|_{0,\tau}^2 \le Ch^{2-2\epsilon}\sum_{\tau\in K_1}\|\mathbf{u}-\mathbf{u}^h\|_{L^{\frac{2}{\epsilon}}(\tau)}^2\\ &\le Ch^{2-2\epsilon}\|\mathbf{u}-\mathbf{u}^h\|_{L^{\frac{2}{\epsilon}}(K_1)}^2 \le Ch^{2-2\epsilon}\|\mathbf{u}-\mathbf{u}^h\|_{H^1(K_1)}^2\\ &= Ch^{2-2\epsilon}\int_{K_1} r^{-2\beta}r^{2\beta}\left(|\mathbf{u}-\mathbf{u}^h|^2+|\nabla(\mathbf{u}-\mathbf{u}^h)|^2\right)d\Omega\\ &\le Ch^{2(1-\beta-\epsilon)}\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta,K_1}^2. \qquad (5.10)\end{aligned}$$

Hence by (5.9), (5.10), and (5.5) we have

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}^h\|_0^2 &= \sum_{\tau\in K_0}\|\mathbf{u}-\mathbf{u}^h\|_{0,\tau}^2 + \sum_{\tau\in K_1}\|\mathbf{u}-\mathbf{u}^h\|_{0,\tau}^2 \le h^{2(1-\beta-\epsilon)}\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta}^2\\ &\le Ch^{2s}\,\|\mathbf{u}\|_{\alpha+\beta,\beta}^2,\end{aligned}$$

where $s<\alpha$. The proof for $\beta\ge 1$ follows analogously. ∎

Remark 5.5. For the optimal finite element convergence of $O(h)$ with respect to the weighted functional and H^1-norms, we select $\beta = 2 - \alpha$. But Theorem 5.4 also requires that $\beta < 1 + \alpha$. Thus, when $\alpha \in [1/2, 1)$, we may use a weighting with $\beta = 2 - \alpha$ and expect optimal rates, but when $\alpha \in (0, 1/2)$, our theory guarantees only at best $O(2\alpha)$ convergence using $\beta = 1 + \alpha$. Numerical results, however, indicate that values of β larger than the theory allows can be used to recover optimal rates. We explore this in the next section.

6. Computational results. In this section, we present some numerical examples of the weighted-norm procedure to validate the error bounds in the previous section.

As a test problem, we minimize the weighted functional on the following L-shaped domain: $\Omega = (-0.5, 0.5)^2 \setminus [0, 0.5) \times (-0.5, 0]$, which yields $\alpha = \pi/\omega = 2/3$. Function $\mathbf{f}$ is chosen so that the solution of this test problem is $\mathbf{u} = \nabla(\chi(r) r^{\frac{2}{3}} \sin(2\theta/3))$, where $\chi(r) = 1$ for $r < 1/8$, $\chi(r) = 0$ for $r > 3/8$, and $\chi(r)$ is C^2 smooth. Again, note that $\mathbf{f} \in L^2(\Omega)$ but $\mathbf{u} \notin H^1(\Omega)$.

Define the following measures of the accuracy of the computed solution, $\mathbf{u}^h$:

nonweighted functional norm	$G^{1/2} = (\|\nabla \cdot \mathbf{u}^h - f\|_0^2 + \|\nabla \times \mathbf{u}^h\|_0^2)^{1/2}$,
nonweighted L^2-norm of the error	$\epsilon^0 = \|\mathbf{u} - \mathbf{u}^h\|_0$,
nonweighted H^1 seminorm of the error	$\epsilon^1 = \|\mathbf{u} - \mathbf{u}^h\|_1$,
weighted functional norm	$G_w^{1/2} = G_w(\mathbf{u}^h; f)^{1/2}$,
weighted L^2-norm of the error	$\epsilon_w^0 = \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta}$,
weighted H^1 seminorm of the error	$\epsilon_w^1 = \|\mathbf{u} - \mathbf{u}^h\|_{1,\beta}$.

Since $\alpha = 2/3$, we choose the optimal weight parameter, $\beta = 2 - \alpha = 4/3$, for our computations. Table 6.1 summarizes discretization error and convergence rates for $\beta = 4/3$.

TABLE 6.1
Convergence of discretization error for weighted-norm FOSLS.

h	$G_w^{1/2}$	Ratio	Rate	ϵ_w^1	Ratio	Rate
8^{-1}	5.52			3.81		
16^{-1}	4.34	1.27	0.35	1.47	2.59	1.37
32^{-1}	2.34	1.85	0.89	6.66e-01	2.21	1.14
64^{-1}	1.19	1.97	0.98	2.97e-01	2.24	1.16
128^{-1}	5.98e-01	1.99	0.99	1.41e-01	2.11	1.08
256^{-1}	3.00e-01	1.99	0.99	6.74e-02	2.09	1.06
512^{-1}	1.50e-01	2.00	1.00	3.31e-02	2.04	1.03

h	ϵ_w^0	Ratio	Rate	ϵ^0	Ratio	Rate
8^{-1}	3.08e-01			3.72e-01		
16^{-1}	1.35e-01	2.28	1.19	1.93e-01	1.93	0.95
32^{-1}	4.07e-02	3.32	1.73	8.93e-02	2.16	1.11
64^{-1}	1.11e-02	3.67	1.88	4.93e-02	1.81	0.86
128^{-1}	2.98e-03	3.72	1.90	3.00e-02	1.64	0.71
256^{-1}	7.84e-04	3.80	1.93	1.87e-02	1.60	0.68
512^{-1}	2.06e-04	3.81	1.93	1.18e-02	1.58	0.66

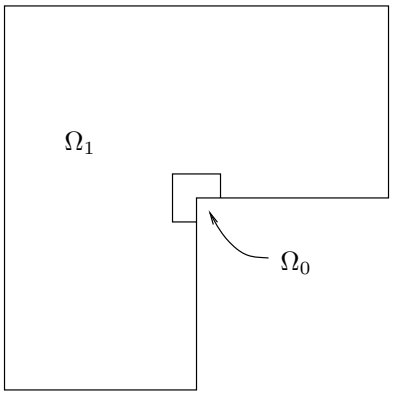


FIG. 6.1. *L-shaped domain Ω and subdomains Ω_0 and Ω_1.*

TABLE 6.2
Accuracy in Ω_0, Ω_1, and Ω with $\beta = 2 - \alpha$.

	$G_w^{1/2}$	$G^{1/2}$	ϵ_w^1	ϵ^1	ϵ_w^0	ϵ^0
Ω_1	$O(h)$	$O(h)$	$O(h)$	$O(h)$	$O(h^2)$	$O(h^2)$
Ω_0	$O(h)$	$O(1)$	$O(h)$	$O(1)$	$O(h^2)$	$O(h^{\frac{2}{3}})$
Ω	$O(h)$	$O(1)$	$O(h)$	$O(1)$	$O(h^2)$	$O(h^{\frac{2}{3}})$

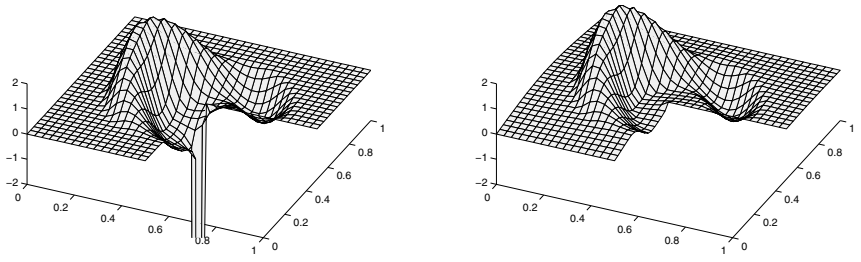
Asymptotic convergence rates in Ω are found to be approximately $O(h)$ for $G_w^{\frac{1}{2}}$ and ϵ_w^1, $O(h^2)$ for ϵ_w^0, and $O(h^{\frac{2}{3}})$ for ϵ_0. The approximation does not converge in either the ϵ_1 or $G^{\frac{1}{2}}$ measures since $\mathbf{u} \notin H^1(\Omega)$.

To distinguish between behavior near to and away from the singularity, we consider the error of the solution above on a partitioning of Ω. Define $\Omega_0 = \Omega \cap (\frac{3}{8}, \frac{5}{8})^2$ and $\Omega_1 = \Omega \backslash \Omega_0$; see Figure 6.1.

Table 6.2 summarizes the asymptotic discretization accuracy obtained at the finest mesh size in subdomains Ω_0 and Ω_1. Away from the singularity we observe optimal accuracy in all measures. As expected, near the singularity, the solution fails to converge in the nonweighted functional and H^1-norms. The nonweighted L^2 error achieves accuracy of approximately $O(h^{\frac{2}{3}})$ near the singularity.

Figure 6.2 shows the first component of the exact solution, u_1, and the standard FOSLS approximation u_1^h. Figure 6.3 shows the error of the first component of the approximated solution for the standard FOSLS and the weighted-norm FOSLS methods. We see that the error in the approximation in standard FOSLS is highest near the singularity but remains large even away from the corner point. In the weighted-norm FOSLS implementation, the error remains large near the singularity, as we expect, but is now concentrated only near the corner point. The pollution effect is removed by the weighting procedure.

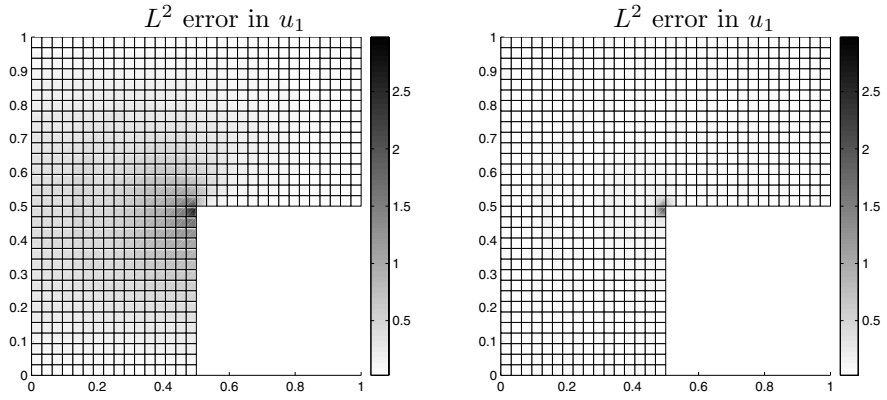
There are many boundary value problems not directly covered by the theory presented here that are of interest. For example, Poisson's equation with mixed boundary conditions on the domain used above has a value of $\alpha = 1/3$. To recover optimal convergence for this problem, the weighted-norm method requires a value of



(a) Exact component u_1.

(b) Component u_1^h computed by standard FOSLS.

FIG. 6.2. *Exact solution component u_1 and solution component u_1^h approximated by standard FOSLS on the $h = 32^{-1}$ mesh.*



(a) Standard FOSLS.

(b) Weighted-norm FOSLS, $\beta = 2 - \alpha$.

FIG. 6.3. *Reduction of the pollution effect by the weighted-norm procedure. Each plot is the error of solution component u_1^h on the $h = 32^{-1}$ mesh.*

β larger than Theorem 5.4 allows. In other elliptic equations (e.g., Stokes or the linear elasticity equations), the value of α is generally smaller than for Poisson's equation for the same domain and boundary condition type. In each of these cases, a larger β value is necessary for optimal convergence. This leads us to consider using larger β than the theory allows.

Consider the same example problem as above on uniform mesh sizes of $h = 1/8, 1/16, \ldots, 1/512$, and values of β ranging from $1/3$ to $23/6$.

Figure 6.4 plots the convergence rate at the finest level for the weighted functional norm, $G_w^{1/2}$; the weighted L^2-norm, ϵ_w^0; and the L^2-norm, ϵ^0. While the functional norm retains optimal accuracy for large values of β, the solution fails to converge in the weighted and nonweighted L^2 measures for $\beta \gtrsim 3$. This indicates that, although the weighted-norm approach seems to be more robust than the theory allows, large values of β should still be used with caution.

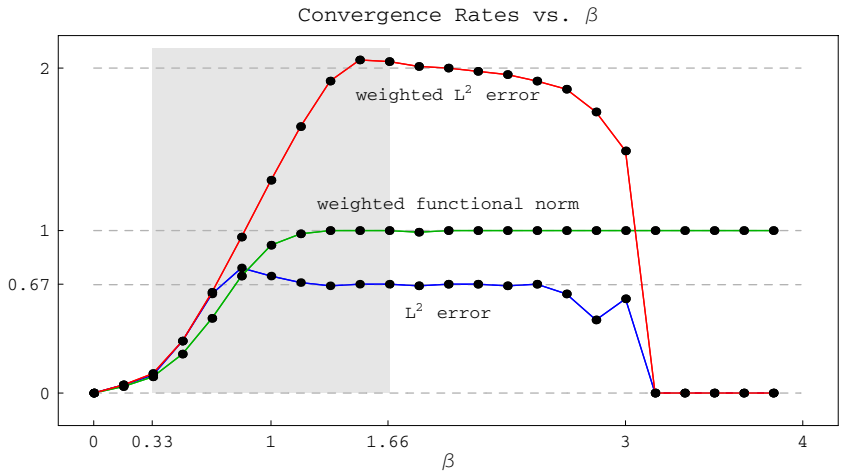


FIG. 6.4. *Convergence rates versus β. The shaded region indicates values of β for which the assumptions of Theorem* 5.4 *are satisfied.*

The method presented here is applicable to a wide range of problems and provides an efficient alternative to more specialized techniques for treating singularities in boundary value problems.

REFERENCES

[1] TH. APEL AND B. HEINRICH, *Mesh refinement and windowing near edges for some elliptic problem*, SIAM J. Numer. Anal., 31 (1994), pp. 695–708.

[2] I. BABUSKA, R.B. KELLOGG, AND J. PITKARANTA, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[3] M. BERNDT, T.A. MANTEUFFEL, S.F. MCCORMICK, AND G. STARKE, *Analysis of first-order system least squares (fosls) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., 43 (2005), pp. 386–408.

[4] M. BERNDT, T.A. MANTEUFFEL, AND S.F. MCCORMICK, *Analysis of first-order system least squares (fosls) for elliptic problems with discontinuous coefficients: Part* II, SIAM J. Numer. Anal., 43 (2005), pp. 409–436.

[5] P.B. BOCHEV AND M.D. GUNZBURGER, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[6] P. BOCHEV AND M. GUNZBURGER, *On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal., 43 (2005), pp. 340–362.

[7] D. BRAESS, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 2001.

[8] J. BRAMBLE, R. LAZAROV, AND J. PASCIAK, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[9] S.C. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[10] Z. CAI AND S. KIM, *A finite element method using singular functions for the Poisson equation: Corner singularities*, SIAM J. Numer. Anal., 39 (2001), pp. 286–299.

[11] Z. CAI, T.A. MANTEUFFEL, AND S.F. MCCORMICK, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[12] Z. CAI, T.A. MANTEUFFEL, AND S.F. MCCORMICK, *First-order system least squares for velocity-vorticity-pressure form of the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1997), pp. 150–159.

[13] Z. CAI, T.A. MANTEUFFEL, S.F. MCCORMICK, AND S.V. PARTER, *First-order system least squares (*FOSLS*) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[14] Z. Cai, T.A. Manteuffel, S.F. McCormick, and J. Ruge, *First-order system $\mathcal{LL}^*$ (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[15] C. Cox and G. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[16] G.J Fix, S. Gulati, and G.I. Wakoff, *On the use of singular functions with finite element approximations*, J. Comput. Phys., 13 (1973), pp. 209–228.

[17] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[18] P. Grisvard, *Singularities in Boundary Value Problems*, Springer-Verlag, Berlin, 1992.

[19] S.D. Kim, T.A. Manteuffel, and S.F. McCormick, *First-order system least squares (fosls) for spatial linear elasticity: Pure traction*, SIAM J. Numer. Anal., 38 (2001), pp. 1454–1482.

[20] V.A. Kondratiev, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[21] V.A. Kozlov, V.G. Maz'ya, and J. Rossmann, *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Equations*, Math. Surveys Monogr. 85, American Mathematical Society, Providence, RI, 2001.

[22] T.A. Manteuffel, S.F. McCormick, J. Ruge, and J.G. Schmidt, *First-order system $\mathcal{LL}^*$ (FOSLL*) for general scalar elliptic problems in the plane*, SIAM J. Numer. Anal., 43 (2005), pp. 2098–2120.

[23] V.G. Maz'ya, S. Nazarov, and B. Plamenevskij, *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains, Vol.* I, Oper. Theory Adv. Appl. 111, Birkhäuser Verlag, Basel, 2000. Operator Theory Advances and Applications.

[24] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 1997–2024

CONVERGENCE OF A COMPACT SCHEME FOR THE PURE STREAMFUNCTION FORMULATION OF THE UNSTEADY NAVIER–STOKES SYSTEM*

MATANIA BEN-ARTZI†, JEAN-PIERRE CROISILLE‡, AND DALIA FISHELOV§

Abstract. This paper is devoted to the analysis of a new compact scheme for the Navier–Stokes equations in pure streamfunction formulation. Numerical results using that scheme have been reported in [M. Ben-Artzi et al., *J. Comput. Phys.*, 205 (2005), pp. 640–664]. The scheme discussed here combines the Stephenson scheme for the biharmonic operator and ideas from box-scheme methodology. Consistency and convergence are proved for the full nonlinear system. Instead of customary periodic conditions, the case of boundary conditions is addressed. It is shown that in one dimension the truncation error for the biharmonic operator is $O(h^4)$ at interior points and $O(h)$ at near-boundary points. In two dimensions the truncation error is $O(h^2)$ at interior points (due to the cross-terms) and $O(h)$ at near-boundary points. Hence the scheme is globally of order four in the one-dimensional periodic case and of order two in the two-dimensional periodic case, but of order 3/2 for one- and two-dimensional nonperiodic boundary conditions. We emphasize in particular that there is no special treatment of the boundary, thus allowing robust use of the scheme. The finite element analogy of the finite difference schemes is invoked at several stages of the proofs in order to simplify their verifications.

Key words. finite difference compact schemes, Stephenson scheme, box schemes, finite elements, Navier–Stokes equations, streamfunction formulation, biharmonic problem, fourth order problem

AMS subject classifications. 65L20, 65L70, 65M06, 65M12, 65M70, 35Q30, 76D05, 78M10, 78M20

DOI. 10.1137/05062915X

1. Introduction.

In a recent paper [3] we presented a fourth-order compact scheme for the pure streamfunction formulation of the two-dimensional (incompressible) Navier–Stokes equations. We have given there a convergence analysis for the linearized model. In this paper we prove the convergence of the nonlinear scheme, without any further assumptions. Recall that the pure streamfunction formulation of the (two-dimensional) Navier–Stokes equations is classical [15]. It has the advantage of reducing the system to a single evolution equation for the scalar streamfunction having the form

$$\frac{\partial \Delta\psi}{\partial t} + \nabla^{\perp}\psi \cdot \nabla\Delta\psi - \nu\Delta^2\psi = 0. \tag{1}$$

The velocity field is $(u, v) = \nabla^{\perp}\psi = (-\frac{\partial\psi}{\partial y}, \frac{\partial\psi}{\partial x})$, and the vorticity is $\omega = \Delta\psi$. The price paid for reducing the system to a single equation is that one must now deal with the biharmonic Δ^2 operator. There are therefore two boundary conditions imposed

*Received by the editors April 13, 2005; accepted for publication (in revised form) March 28, 2006; published electronically October 16, 2006. This work was partially supported by the High Council for Scientific and Technological Cooperation between France and Israel.

http://www.siam.org/journals/sinum/44-5/62915.html

†Institute of Mathematics, The Hebrew University, Jerusalem 91904, Israel (mbartzi@math.huji.ac.il).

‡Laboratoire Mathématiques & Applications de Metz, UMR CNRS 7122, Université de Metz, F-57045 Metz, France (jean-pierre.croisille@univ-metz.fr).

§Afeka-Tel-Aviv Academic College of Engineering, 218 Bnei-Efraim St., Tel-Aviv 69107, Israel, and School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel (daliaf@post.tau.ac.il).

on ψ. For the typical "no-leak no-slip" conditions (vanishing velocity on the fixed boundary) we have

$$\nabla\psi = 0 \qquad \text{on the boundary.} \tag{2}$$

Since the function ψ is only determined up to a constant, condition (2) is equivalent to

$$\psi = \frac{\partial\psi}{\partial n} = 0, \tag{3}$$

which, for simplicity, will be the case treated in this paper. Clearly (2) is equivalent to the assumption $\psi \in H_0^2$, the closure of smooth compactly supported test functions in the Sobolev space of functions having square-summable derivatives up to second order.

Our scheme can be described as follows (see [3] for details). At each time step the scheme solves a time implicit version of (1). This leads to a fourth-order biharmonic problem of the form

$$\Delta\psi - \nu\Delta^2\psi = f, \tag{4}$$

subject to the boundary conditions (2).

The spatial discretization of (4) makes use of the Stephenson scheme for the the biharmonic operator introduced in [19], [12]. See also [2]. This scheme can be interpreted as a mixed scheme in $(\psi, \nabla\psi)$, similar in form to a version of a box scheme [14], [7]. More specifically, its design is obtained by a spline collocation procedure on a nine-point stencil, which we recall in section 3 below.

The streamline-vorticity formulation has been extensively used for the simulation of the two-dimensional Navier–Stokes system. As representative references we mention [17], [8], [5], [9], [13], and the references therein. One difficult point is that "... the $\psi - \omega$ system is inextricably coupled; BC's and solution methods must contend with this fact..." [10, p. 431]. Indeed, one must cope with the vorticity boundary values, resulting from the fact that the relation $\Delta\psi = \omega$ is overdetermined under condition (2). An attempt to avoid this difficulty has been made in [4], where the need to determine these values was circumvented by switching to the biharmonic equation (at each time step), exploiting the natural condition (2). The scheme presented in [3], whose convergence is proved here, has avoided all explicit mention of the vorticity by using a pure streamfunction formulation. We mention that recently in [11] a very similar algorithm has been proposed, but it deals only with the steady-state Navier–Stokes system.

The paper is organized as follows. First, we introduce in section 2 our notation and the setup for our discrete spaces. Then we establish in sections 3 and 4 the necessary analytic properties of the scheme in one and two dimensions. In particular, in analogy with the coercivity of Δ^2 in H_0^2, we prove the coercivity of the discretized biharmonic operator in a suitable discrete analogue of H_0^2. We prove that the truncation error of the biharmonic scheme is of order four in one dimension and of order two in two dimensions, at all interior points and of first order at near-boundary points, giving a 3/2 order of convergence rate in the natural discrete L^2 norm. Note that in the periodic case all points are interior. Then in section 5, we prove that the same order of convergence extends to the spatial semidiscrete version of the full nonlinear scheme. We emphasize the fact that we do not need any special treatment of boundary points, and the boundary condition (2) is naturally incorporated here. As mentioned above,

this causes a reduced (from four to one) order of local truncation error at the boundary, and is reflected in the fact that our result yields a 3/2 convergence rate in the discrete L^2 norm. The present convergence result can be compared to the convergence results obtained in [9], [13]. In both papers, the time evolution is performed on the vorticity, and hence a very careful treatment of the vorticity boundary conditions is required, either by "ghost-points" [9] or by replacing condition (2) on the normal derivative of the streamfunction by boundary conditions on the vorticity [13] (which, as these authors observe, amounts to an algorithm for vorticity generation on the boundary).

2. Discrete spaces and basic inequalities. Let $0 \le i, j \le N$. We denote by (ih, jh) a finite difference mesh on the square $[0,1]^2$, with equal mesh size $h = 1/N$ in the x and y directions. We denote by $u_{i,j}$ a grid function on $[0,1]^2$, with $0 \le i, j \le N$. The centered and upwind derivative operators δ_x, $\delta_x^{\pm}$ are defined as usual in each direction by

$$\delta_x u_{i,j} = \frac{u_{i+1,j} - u_{i-1,j}}{2h}, \quad \delta_x^+ u_{i,j} = \frac{u_{i+1,j} - u_{i,j}}{h}, \quad \delta_x^- u_{i,j} = \frac{u_{i,j} - u_{i-1,j}}{h}, \tag{5}$$

and similarly in the y direction:

$$\delta_y u_{i,j} = \frac{u_{i,j+1} - u_{i,j-1}}{2h}, \quad \delta_y^+ u_{i,j} = \frac{u_{i,j+1} - u_{i,j}}{h}, \quad \delta_y^- u_{i,j} = \frac{u_{i,j} - u_{i,j-1}}{h}. \tag{6}$$

The centered second-order derivatives are

$$\delta_x^2 u_{i,j} = \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2}, \quad \delta_y^2 u_{i,j} = \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{h^2}. \tag{7}$$

The five-point Laplacian is

$$\Delta_h u_{i,j} = \delta_x^2 u_{i,j} + \delta_y^2 u_{i,j} = \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{h^2}. \tag{8}$$

The crossed derivative operators δ_{xy}^+, δ_{xy}^-, δ_{xy} are

$$\delta_{xy}^+ u_{i,j} = \delta_x^+ \delta_y^+ u_{i,j} = \frac{u_{i+1,j+1} - u_{i+1,j} - u_{i,j+1} + u_{i,j}}{h^2}, \tag{9}$$

$$\delta_{xy}^- u_{i,j} = \delta_x^- \delta_y^- u_{i,j} = \frac{u_{i,j} - u_{i,j-1} - u_{i-1,j} + u_{i-1,j-1}}{h^2}, \tag{10}$$

$$\delta_{xy} u_{i,j} = \delta_x \delta_y u_{i,j} = \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4h^2}. \tag{11}$$

It is easy to check that

$$\delta_x^2 \delta_y^2 u_{i,j} = \delta_{xy}^+ \delta_{xy}^- u_{i,j}. \tag{12}$$

The L_h^2 space is the space of sequences $u_{i,j}$, $0 \le i, j \le N$. $L_{h,0}^2$ is the subspace of $u_{i,j}$ with zero boundary conditions $u_{i,j} = 0$ for $i \in \{0, N\}$ or $j \in \{0, N\}$. The scalar product on $L_{h,0}^2$ is

$$(u, v)_h = h^2 \sum_{i,j=1}^{N-1} u_{i,j} v_{i,j}, \tag{13}$$

with the corresponding norm

$$|u|_h = \left\{ h^2 \sum_{i,j=1}^{N-1} (u_{i,j})^2 \right\}^{1/2}. \tag{14}$$

Furthermore, we denote by l_h^2 the space of sequences u_i, $0 \le i \le N$, and by $l_{h,0}^2$ the subspace of sequences with zero boundary conditions. The scalar product and the norm on $l_{h,0}^2$ are

$$(u,v)_h = h \sum_{i=1}^{N-1} u_i v_i, \quad |u|_h^2 = \left\{ h \sum_{i=1}^{N-1} u_i^2 \right\}^{1/2}. \tag{15}$$

We also define the discrete infinity norm

$$|u|_{\infty,h} = \max_i |u_i|. \tag{16}$$

We skip the proof of the following lemma, which states the discrete integration by parts in $L_{h,0}^2$ for the operators $\delta_x^{\pm}$, δ_x^2. For each grid function $u \in L_{h,0}^2$, we denote the one-dimensional column vector $u^j = [u_{1,j}, u_{2,j}, \ldots, u_{N-1,j}]^T$, $1 \le j \le N-1$.

LEMMA 2.1 (discrete integration by parts). *For any $u, v \in L_{h,0}^2$, we have*

$$\text{(i)} \quad (\delta_x^+ u, v)_h = -(u, \delta_x^- v)_h; \tag{17}$$

$$\text{(ii)} \quad (\delta_x^2 u, v)_h = -(\delta_x^+ u, \delta_x^+ v)_h = -(\delta_x^- u, \delta_x^- v)_h. \tag{18}$$

Note that in (17), (18), the finite difference operators are extended to the points $i = 0$, $i = N$ by

$$\delta_x^{\pm} u_0 = \delta_x^{\pm} u_N = 0, \quad \delta_x^2 u_0 = \delta_x^2 u_N = 0. \tag{19}$$

Observe that this assumption is only for notational convenience, in order to have formally $\delta_x^{\pm} u, \delta_x^2 u \in L_{h,0}^2$. Results similar to (17), (18) in the y direction are obtained by substituting the subscript y to the subscript x. The following lemma is the counterpart of the Poincaré inequality at the discrete level.

LEMMA 2.2 (discrete Poincaré inequality). *For all $u \in L_{h,0}^2$ and any $1 \le j \le N-1$,*

$$|u^j|_h \le 2 |\delta_x^+ u^j|_h. \tag{20}$$

COROLLARY 2.1. *For all $u \in L_{h,0}^2$,*

$$|u|_h \le \sqrt{2} \left[|\delta_x^+ u|_h^2 + |\delta_y^+ u|_h^2 \right]^{1/2}. \tag{21}$$

Proof. For all $u \in l_{h,0}^2$, we have

$$|u|_h^2 = h \sum_{i_0=1}^{N-1} u_{i_0}^2. \tag{22}$$

For all $1 \le i_0 \le N-1$,

$$\begin{aligned} u_{i_0}^2 &= \sum_{i=0}^{i_0-1} (u_{i+1} - u_i)(u_{i+1} + u_i) = \sum_{i=0}^{i_0-1} h \delta_x^+ u_i (u_i + (Su)_i) \\ &\le 2 |\delta_x^+ u|_h |u|_h, \end{aligned}$$

where $(Su)_j = u_{j+1}$, $j = 0, \dots, N-1$. Therefore,

$$|u|_h^2 = h \sum_{i_0=1}^{N-1} u_{i_0}^2 \le 2|\delta_x^+ u|_h |u|_h, \tag{23}$$

which gives (20).

Now for all $u \in L_{h,0}^2$, we have

$$\begin{aligned} |u|_h^2 = h \sum_{j_0=1}^{N-1} |u^{j_0}|_h^2 &\le 2h \sum_{j_0=1}^{N-1} |\delta_x^+ u^{j_0}|_h |u^{j_0}|_h \\ &\le 2\Big(\sum_{j_0=1}^{N-1} h|\delta_x^+ u^{j_0}|^2 \Big)^{1/2} \Big(\sum_{j_0=1}^{N-1} h|u^{j_0}|^2 \Big)^{1/2} \\ &\le 2|\delta_x^+ u|_h |u|_h. \end{aligned} \tag{24}$$

In a similar way, we obtain in the y direction

$$|u|_h^2 \le 2|\delta_y^+ u|_h |u|_h. \tag{25}$$

Summing (24) and (25), we obtain (21). ∎

3. The Stephenson scheme in one dimension.

3.1. Design by collocation. Consider the one-dimensional biharmonic equation

$$\begin{cases} u^{(4)}(x) = f(x), \quad 0 < x < 1, \\ u(0) = u(1) = u_x(0) = u_x(1) = 0. \end{cases} \tag{26}$$

Suppose that at each node $x_j = jh$, $0 \le j \le N$, of a finite difference grid, there are two unknowns u_j and $u_{x,j}$ approximating, respectively, $u(x_j)$ and $u_x(x_j)$, which is referred to as a "mixed scheme." The values u_j, $u_{x,j}$ are solutions of the linear system, designed by the following Galerkin collocation method. At each interior node j, $1 \le j \le N-1$, we consider a fourth-order polynomial, with domain $[x_{j-1}, x_{j+1}]$

$$Q(x) = a_0 + a_1(x - x_j) + a_2(x - x_j)^2 + a_3(x - x_j)^3 + a_4(x - x_j)^4. \tag{27}$$

The five coefficients a_k, $k \in \{0, 1, 2, 3, 4\}$, are defined by the five collocation conditions on the compact stencil $\{x_{j-1}, x_j, x_{j+1}\}$ (see Figure 1):

$$\begin{cases} Q(x_{j-1}) = u_{j-1}, \quad Q(x_j) = u_j, \quad Q(x_{j+1}) = u_{j+1}, \\ Q'(x_{j-1}) = u_{x,j-1}, \quad Q'(x_{j+1}) = u_{x,j+1}. \end{cases} \tag{28}$$

The five coefficients of the unique polynomial (27), solution of (28), are given by

$$\begin{cases} a_0 = u_j, \\ a_1 = \dfrac{3}{2}\delta_x u_j - \dfrac{1}{4}(u_{x,j+1} + u_{x,j-1}), \\ a_2 = \delta_x^2 u_j - \dfrac{1}{2}(\delta_x u_x)_j, \\ a_3 = \dfrac{1}{h^2}(\delta_x u_j - u_{x,j}) = \dfrac{1}{6}(\delta_x^2 u_x)_j, \\ a_4 = \dfrac{1}{2h^2}\big[(\delta_x u_x)_j - \delta_x^2 u_j\big]. \end{cases} \tag{29}$$

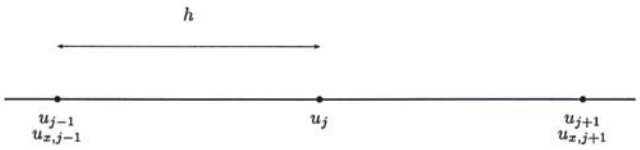


FIG. 1. *Stephenson's scheme for $u^{(4)} = f$: The finite difference operator $\delta_x^4 u_j$ at point j is $Q^{(4)}(x_j)$, where $Q(x) \in P^4[x_{j-1}, x_{j+1}]$ is defined by the five collocated values for u_{j-1}, u_j, u_{j+1}, $u_{x,j-1}$, $u_{x,j+1}$.*

Now, since $Q'(x_j) = a_1$ and $Q''''(x_j) = 24a_4$, it is natural to define the following compact scheme: find $[u_0, u_1, \ldots, u_{N-1}, u_N]$, $[u_{x,0}, u_{x,1}, \ldots, u_{x,N-1}, u_{x,N}] \in l^2_{h,0}$, which solve

$$\begin{cases} \text{(a)} \ \ (P_x u_x)_j = \delta_x u_j, \quad 1 \le j \le N-1, \\ \text{(b)} \ \ \delta_x^4 u_j = f(x_j), \quad 1 \le j \le N-1, \\ \text{(c)} \ \ u_0 = u_1 = u_{x,0} = u_{x,N} = 0, \end{cases} \tag{30}$$

where the operators P_x, δ_x^4 are, respectively, defined in (31), (34).

For $u \in l^2_{h,0}$, the operator P_x is defined by

$$(P_x u)_j = \frac{1}{6} u_{j-1} + \frac{2}{3} u_j + \frac{1}{6} u_{j+1}, \quad 1 \le j \le N-1. \tag{31}$$

P_x will be referred to as the *Simpson operator* in the x direction, because the coefficients in (30) are those of the Simpson quadrature formula over $[x_{j-1}, x_{j+1}]$. Note also that

$$P_x = I + \frac{h^2}{6} \delta_x^2. \tag{32}$$

We also note that the connection (30)(a) is already given in the classical book by Collatz [6, Chap. III, Eq. 2.9]. We call $\mathcal{S}$ the discrete space of grid functions $(u, u_x) \in l^2_{h,0} \times l^2_{h,0}$,

$$\mathcal{S} = \big\{ (u, u_x) \in l^2_{h,0} \text{ such that } P_x u_x = \delta_x u \big\}. \tag{33}$$

In (30), we define the *Stephenson discrete biharmonic* to be the compact difference operator given on $\mathcal{S}$ by

$$\delta_x^4 u_j = \frac{12}{h^2} \big\{ (\delta_x u_x)_j - \delta_x^2 u_j \big\}, \quad 1 \le j \le N-1. \tag{34}$$

This is a one-dimensional version of the original scheme proposed by Stephenson in [19]. Note that for simplicity, we will refer in what follows to a grid function in $\mathcal{S}$ by $u \in \mathcal{S}$, meaning that it is the first component of a pair $(u, u_x) \in \mathcal{S}$.

Remark. We note that the implicit scheme (30)(a) defining the grid function u_x as a function of u is exactly the one obtained in the piecewise cubic spline interpolation; see, e.g., [18]. The classical question that occurs in spline interpolation about fixing the two degrees of freedom $u_{x,0}$, $u_{x,N}$ at end points is here pointless, since they are precisely given in (30)(c).

3.2. Consistency. On a periodic grid, the order of consistency can be obtained by a simple Taylor expansion at point x_j. Equivalently, one can compute the symbol of the operators. Recall that in the context of finite difference operators, we have to use the semidiscrete Fourier transform; see, e.g., [20]. In practice, if the values of the periodic grid function (u_j) are represented by $e^{ij\xi h}$, then the symbol of the linear operator L_h is $l_h(\xi)$ defined by

$$L_h u_j = l_h(\xi) u_j. \tag{35}$$

Furthermore, if $l(\xi)$ is the symbol of L, then the order of consistency is given by the greatest value $p > 0$ such that (see [20])

$$l_h(\xi) - l(\xi) = O(h^p). \tag{36}$$

Doing so, it is quite easy to verify that the Stephenson gradient is fourth-order accurate and that the biharmonic operator (34) is as well. Indeed, we verify the following:

- The symbol of the discrete operator u_x in (30)(a) is

$$g_h(\xi) = i\xi - \frac{1}{180} i\xi^5 h^4 + O(h^6), \tag{37}$$

so that the order of accuracy with respect to the operator ∂_x, whose symbol is $i\xi$, is

$$g_h(\xi) - i\xi = O(h^4). \tag{38}$$

- The symbol of the discrete operator $\delta_x^4 u$ in (34) is

$$d_h(\xi) = \xi^4 - \frac{1}{720} \xi^8 h^4 + O(h^6), \tag{39}$$

so that the order of accuracy with respect to ∂_x^4 is

$$d_h(\xi) - (i\xi)^4 = O(h^4). \tag{40}$$

On a finite grid with homogeneous boundary conditions at the two ends, we have to perform a more careful analysis, because the symbolic computation no longer holds in this case.

LEMMA 3.1. *Suppose that $u(x)$ is a regular function on $[0,1]$. Then the finite difference gradient u_x defined from the values $u(x_j)$, $0 \le j \le N$, by $(P_x u_x)_j = \delta_x u(x_j)$ has a truncation error $(u_x)_j - u'(x_j)$ of order four at each point x_j. More precisely,*

$$|(u_x)_j - u'(x_j)| \le C h^4 |u^{(5)}|_{\infty,[0,1]}. \tag{41}$$

Proof. The Stephenson gradient u_x is defined in the space $l^2_{h,0}$ by

$$(P_x u_x)_j = (\delta_x u)_j, \quad 1 \le j \le N-1, \tag{42}$$

where P_x is the $N-1 \times N-1$ matrix-operator acting on $l^2_{h,0}$ as defined in (31), that is,

$$P_x = \begin{pmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \ddots & \\ \vdots & & \ddots & \ddots & \\ 0 & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & \dots & \frac{1}{6} & \frac{2}{3} \end{pmatrix}. \tag{43}$$

Consider a regular function $u(x)$, differentiable as much as needed, and denote by u', $u'', \ldots, u^{(p)}$, its derivatives. At each point x_j, $1 \le j \le N-1$, the Taylor formula gives (we note $u_j^{(m)} = u^{(m)}(x_j)$)

$$(\delta_x u)(x_j) = u'_j + \frac{h^2}{6} u_j^{(3)} + \frac{h^4}{2\,5!}\big[u^{(5)}(\xi_{1,j}^-) + u^{(5)}(\xi_{1,j}^+)\big], \tag{44}$$

where $\xi_{1,j}^- \in]x_{j-1}, x_j[$ and $\xi_{1,j}^+ \in]x_j, x_{j+1}[$. Similarly, there exist $\xi_{2,j}^- \in]x_{j-1}, x_j[$, $\xi_{2,j}^+ \in]x_j, x_{j+1}[$ such that

$$(\delta_x^2 u)(x_j) = u''_j + \frac{h^2}{4!}\big[u^{(4)}(\xi_{2,j}^-) + u^{(4)}(\xi_{2,j}^+)\big]. \tag{45}$$

We deduce that, applying (45) to u',

$$\begin{aligned}
\delta_x u(x_j) - P_x u'(x_j) &= \delta_x u(x_j) - \left[u'(x_j) + \frac{h^2}{6}\delta_x^2 u'(x_j)\right] \\
&= u'_j + \frac{h^2}{6} u_j^{(3)} + \frac{h^4}{2.5!}\left(u^{(5)}(\xi_{1,j}^-) + u^{(5)}(\xi_{1,j}^+)\right) \\
&\quad - \left[u'_j + \frac{h^2}{6}\left(u_j^{(3)} + \frac{h^2}{4!}\big[u^{(5)}(\xi_{2,j}^-) + u^{(5)}(\xi_{2,j}^+)\big]\right)\right] \\
&= h^4 v_j,
\end{aligned}$$

where the grid function v_j is defined by

$$v_j = \frac{1}{2.5!}\big(u^{(5)}(\xi_{1,j}^+) + u^{(5)}(\xi_{1,j}^-)\big) - \frac{1}{6.4!}\big(u^{(5)}(\xi_{2,j}^-) + u^{(5)}(\xi_{2,j}^+)\big). \tag{46}$$

Therefore, the grid function $u \in l^2_{h,0}$ verifies the identity

$$\delta_x u(x_j) - P_x u'(x_j) = h^4 v_j. \tag{47}$$

On the other hand, $u_x \in l^2_{h,0}$ is defined by

$$\delta_x u - P_x u_x = 0. \tag{48}$$

Subtracting (48) from (47), we obtain the identity in $l^2_{h,0}$,

$$u' - u_x = h^4 P_x^{-1} v, \tag{49}$$

where $u' = [u'(x_1), \ldots, u'(x_{N-1})]$. Writing $P_x = I + \frac{h^2}{6}\delta_x^2$, the inverse of P_x is obtained by the Neumann series

$$P_x^{-1} = \sum_{k=0}^{\infty} \left(-\frac{h^2}{6}\delta_x^2\right)^k, \tag{50}$$

which gives the estimate of $|P_x^{-1}|_{\infty,h}$,

$$|P_x^{-1}|_{\infty,h} \le \sum_{k=0}^{\infty} \frac{h^{2k}}{6^k} |\delta_x^2|_{\infty,h}^k \le \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^k = 3. \tag{51}$$

Observe that the matrix-operator δ_x^2 above is defined at the near-boundary points $j = 1$, $j = N - 1$ by

$$\delta_x^2 u_1 = \frac{u_2 - 2u_1}{h^2}, \quad \delta_x^2 u_{N-1} = \frac{u_{N-2} - 2u_{N-1}}{h^2}. \tag{52}$$

We deduce now from (49) and (51) that

$$|u' - u_x|_{\infty,h} \le h^4 |P_x^{-1}|_{\infty,h} |v|_{\infty,h} \le C h^4 |u^{(5)}|_{\infty,[0,1]}. \qquad \square \tag{53}$$

LEMMA 3.2. *Suppose that $u(x)$ is a regular function on $[0,1]$. Then the Stephenson biharmonic operator δ_x^4 defined by* (34) *has a truncation error $\delta_x^4 u - u^{(4)}$ of order $3/2$ in the $l_{h,0}^2$ norm,*

$$|\delta_x^4 u - u^{(4)}|_h \le C h^{3/2} \big(|u^{(6)}|_{\infty,[0,1]} + |u^{(5)}|_{\infty,[0,1]} \big), \tag{54}$$

where the notation $u^{(4)}$ stands for

$$u^{(4)} = [u^{(4)}(x_1), \dots, u^{(4)}(x_{N-1})] \in l_{h,0}^2. \tag{55}$$

Remark. The difference in accuracy between the periodic case and the nonperiodic case is only due to the near-boundary points 1 and $N - 1$.

Proof. Recall that the finite difference biharmonic operator δ_x^4 is the three-points compact operator, expressed in terms of u and u_x by

$$\delta_x^4 u_j = \frac{12}{h^2} \big[\delta_x u_x - \delta_x^2 u \big]. \tag{56}$$

Here, we handle the finite difference operators acting on one-dimensional grid functions $u = [u_1, \dots, u_{N-1}]$, as $N - 1 \times N - 1$ matrices; see [3]. We can rewrite (30)(a) as

$$P_x u_x = \frac{1}{2h} K u = \delta_x u \in l_{h,0}^2, \tag{57}$$

where the antisymmetric matrix $K = \{K_{i,m}\}_{1 \le i,m \le N-1}$ is given by

$$K_{i,m} = \begin{cases} \operatorname{sgn}(m - i), & |m - i| = 1, \\ 0, & |m - i| \ne 1, \end{cases} \tag{58}$$

and the operator δ_x is expressed as

$$\delta_x = \frac{1}{2h} K. \tag{59}$$

In matrix form, (57) is simply written as

$$P_x u_x = \delta_x u \quad \text{or} \quad u_x = P_x^{-1} \delta_x u. \tag{60}$$

Using (34), the operator δ_x^4 can be rewritten in matrix form

$$\begin{aligned} \delta_x^4 &= \frac{12}{h^2} \big[\delta_x P_x^{-1} \delta_x - \delta_x^2 \big] \\ &= \frac{12}{h^2} \big[P_x^{-1} (\delta_x)^2 + \big[\delta_x P_x^{-1} - P_x^{-1} \delta_x \big] \delta_x - \delta_x^2 u \big]. \end{aligned}$$

Applying the operator P_x, we obtain, for all $u \in l^2_{h,0}$,

$$P_x\left[\delta_x^4 u - u^{(4)}\right] = \frac{12}{h^2}\left[(\delta_x)^2 u + [P_x\delta_x - \delta_x P_x]P_x^{-1}\delta_x u - P_x\delta_x^2 u\right] - P_x u^{(4)} := v. \tag{61}$$

Note that in (60)–(61), we refer to P_x as the symmetric positive definite matrix (see (32)–(43)),

$$(P_x)_{i,m} = \begin{cases} \frac{2}{3}, & m = i, \\ \frac{1}{6}, & |m-i| = 1, \\ 0, & |m-i| \geq 2. \end{cases} \tag{62}$$

Clearly the commutator $[P_x, K] = P_x K - K P_x$ is

$$(P_x K - K P_x)_{i,j} = \begin{cases} -\frac{1}{3}, & i = j = 1, \\ \frac{1}{3}, & i = j = N-1, \\ 0 & \text{otherwise}, \end{cases} \tag{63}$$

so that the commutator $[P_x, \delta_x] = \frac{1}{2h}[P_x, K]$ is

$$P_x\delta_x - \delta_x P_x = \begin{cases} -\frac{1}{6h}, & i = j = 1, \\ \frac{1}{6h}, & i = j = N-1, \\ 0 & \text{otherwise}. \end{cases} \tag{64}$$

This means that the operators P_x and δ_x do not commute and that the nonzero commutator values are restricted to points $j = 1$ and $j = N-1$.

Let us first evaluate (61) at points $j = 2, 3, \ldots, N-2$.

$$\begin{aligned} \frac{12}{h^2}\left[(\delta_x)^2 u_j - P_x\delta_x^2 u_j\right] - P_x u_j^{(4)} = \frac{12}{h^2}\Bigg\{ & (\delta_x)^2 u_j \\ & - \left[\frac{2}{3}\delta_x^2 u_j + \frac{1}{6}\delta_x^2 u_{j+1} + \frac{1}{6}\delta_x^2 u_{j-1}\right]\Bigg\} \\ & - \left[\frac{2}{3}u_j^{(4)} + \frac{1}{6}u_{j-1}^{(4)} + \frac{1}{6}u_{j+1}^{(4)}\right]. \end{aligned} \tag{65}$$

The first term on the right-hand side of (65) is

$$(\delta_x)^2 u_j = u_j'' + \frac{h^2}{3}u_j^{(4)} + \frac{32}{6!}h^4 u_j^{(6)} + \frac{128}{8!}h^6 u_j^{(8)} + Ch^8 u^{(10)}(\xi_j). \tag{66}$$

Using (45) for evaluating $\delta_x^2 u_m$ at $m = j-1, j, j+1$, we find that $P_x\delta_x^2 u_j$ in (65) is

$$\frac{2}{3}\delta_x^2 u_j + \frac{1}{6}\delta_x^2 u_{j+1} + \frac{1}{6}\delta_x^2 u_{j-1} = u_j'' + \frac{1}{4}h^2 u_j^{(4)} + \frac{22}{6!}h^4 u_j^{(6)} + \frac{86}{8!}h^6 u_j^{(8)} + h^8 w_j, \tag{67}$$

where $|w_j| \leq C|u^{(10)}|_{\infty,[0,1]}$. In addition, we have that the third line of the right-hand side in (65) is

$$\left[\frac{2}{3}u_j^{(4)} + \frac{1}{6}u_{j-1}^{(4)} + \frac{1}{6}u_{j+1}^{(4)}\right] = u_j^{(4)} + \frac{1}{6}h^2 u_j^{(6)} + Ch^4 z_j, \tag{68}$$

where $|z_j| \le C|u^{(8)}|_{\infty,[0,1]}$. Therefore, we have, for $2 \le j \le N-2$,

$$\left| \frac{12}{h^2}\big[(\delta_x)^2 u - P_x \delta_x^2 u_j\big] - P_x u_j^{(4)} \right| \le Ch^4 |u^{(8)}|_{\infty,[0,1]}, \tag{69}$$

and this order is optimal. Consider now the truncation term for $j = 1$ (the computation is the same for $j = N-1$). We have

$$(\delta_x^4 u)_1 = \frac{12}{h^2}\big[(\delta_x u_x)_1 - \delta_x^2 u_1\big]. \tag{70}$$

Since $|u_{x,j} - u'_j| \le Ch^4 |u^{(5)}|_{\infty,[0,1]}$, we have

$$\begin{aligned} (\delta_x u_x)_1 &= \frac{u_{x,2}}{2h} = \frac{u_{x,2} - u_{x,0}}{2h} \\ &= \frac{u'(x_2) - u'(x_0)}{2h} + \tilde{v} \\ &= u''(x_1) + \frac{h^2}{6} u^{(4)}(x_1) + \tilde{v}, \end{aligned} \tag{71}$$

where $\tilde{v}$ stands for a generic term such that $|\tilde{v}| \le Ch^3 |u^{(5)}|_{\infty,[0,1]}$. In addition, we have

$$(\delta_x^2 u)_1 = u''(x_1) + \frac{h^2}{12} u^{(4)}(x_1) + w, \tag{72}$$

where

$$|w| \le Ch^4 |u^{(6)}|_{\infty,[0,1]}. \tag{73}$$

Therefore (71), (73) show that the truncation error at the near-boundary point x_1 is

$$\frac{12}{h^2}\big[(\delta_x u_x)_1 - (\delta_x^2 u)_1\big] - u^{(4)}(x_1) = t_1, \text{ with } |t_1| \le Ch |u^{(5)}|_{\infty,[0,1]}. \tag{74}$$

We deduce from (61), (69), (74) that the truncation error $e = \delta_x^4 u - u^{(4)}$ is the solution of the linear system

$$\overline{P}_x e = v, \quad v \in l^2_{h,0}, e \in l^2_{h,0}, \tag{75}$$

where $\overline{P}_x$ is the matrix

$$\overline{P}_x = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix}, \tag{76}$$

and v is such that

$$|v_1|, |v_{N-1}| \le Ch |u^{(5)}|_{\infty,[0,1]}; \quad |v_j| \le Ch^4 |u^{(8)}|_{\infty,[0,1]}, \ \ j = 2, \dots, N-2. \tag{77}$$

By Gerschgorin's theorem, $\overline{P}_x^{-1}$ is a bounded matrix independent of h; therefore $e = \overline{P}_x^{-1} v$ is such that

$$|e|_h \le C|v|_h, \tag{78}$$

where

$$|v|_h^2 \le Ch\left(2h^2 + \sum_{j=2}^{N-2} h^8\right) \le Ch^3. \tag{79}$$

Taking the square root in (79), we obtain (54) (using the weaker estimate $|v_j| \le Ch^2|u^{(6)}|_{\infty,[0,1]}$ at interior points). ∎

Remark. Note that the error at the interior points is fourth order and that the $h^{3/2}$ error is fully due to the loss of accuracy at the two boundary points $j = 1$, $j = N - 1$.

3.3. Interpretation with finite elements. In this section, we establish the finite element counterpart of scheme (30). This allows us to obtain in a simple way the stability of the Stephenson finite difference operator δ_x^4. To each grid function $v \in l_{h,0}^2$, we match the function $v_h(x)$ defined by $v_h(x_j) = v_j$, in the finite element space $P_{c,0}^1$, the space of continuous functions, piecewise linear in each interval $[x_j, x_{j+1}]$, $j = 0, \ldots, N-1$, and such that $v_h(x_0) = v_h(x_N) = 0$. Clearly, it is an isomorphism between $l_{h,0}^2$ and $P_{c,0}^1$. In addition, starting with $v \in l_{h,0}^2$, we introduce the two piecewise constant functions $\bar{v}_h$ and $v_{h,x}$, defined in each interval $K_{j+1/2} =]x_j, x_{j+1}[$ by

$$\bar{v}_{h,j+1/2} = \frac{v_j + v_{j+1}}{2}, \quad v_{h,x,j+1/2} = \frac{v_{j+1} - v_j}{h}. \tag{80}$$

An important aspect of using $P_{c,0}^1$ in the study of finite difference schemes is that it allows one to streamline analytic operations like integration by parts or averaged quantities over intervals $K_{j+1/2} = [x_j, x_{j+1}]$. The $L^2[0,1]$ scalar product is denoted by

$$(\varphi, \psi) = \int_0^1 \varphi(x)\psi(x)dx. \tag{81}$$

Writing the representation of $u_h(x)$ in $K_{j+1/2}$ as ($x_{j+1/2} = \frac{1}{2}(x_{j+1} + x_j)$),

$$u_h(x)_{|K_{j+1/2}} = \bar{u}_{h,j+1/2} + u_{h,x,j+1/2}(x - x_{j+1/2}), \tag{82}$$

we can compare different scalar products for $(.,.)_h$ and in $L^2(0,1)$ as follows.

LEMMA 3.3. *For any $u, v \in l_{h,0}^2$, let $u_h(x), v_h(x) \in P_{c,0}^1$ be the corresponding finite element functions. Then we have*

$$\text{(i)} \quad (u,v)_h = (u_h, v_h) + \frac{h^2}{6}(u_{h,x}, v_{h,x}) = (\bar{u}_h, \bar{v}_h) + \frac{h^2}{4}(u_{h,x}, v_{h,x}); \tag{83}$$

$$\text{(ii)} \quad (\delta_x u, v)_h = (u_{h,x}, v_h); \tag{84}$$

$$\text{(iii)} \quad (\delta_x^2 u, v)_h = -(\delta_x^+ u, \delta_x^+ v)_h = -(\delta_x^- u, \delta_x^- v)_h = -(u_{h,x}, v_{h,x}) \text{ (see (18))}. \tag{85}$$

Proof. The proof is an elementary computation resulting from the piecewise linearity of $u_h(x)$ in each $K_{j+1/2} = [x_j, x_{j+1}]$ given by (82). In fact, it clearly suffices to check that (83), (84), (85) hold for $u_h = \varphi_k$, $v_h = \varphi_m$, where (φ_k) is a basis of $P_{c,0}^1$. ∎

Let $(u, u_x) \in \mathcal{S}$. Since $u_x \in l_{h,0}^2$, it has a matching function $p_h \in P_{c,0}^1$. On the other hand, we have the piecewise constant function $u_{h,x}$. The connection between these two functions is given by the following lemma.

LEMMA 3.4. (i) *Let $u \in \mathcal{S}$ with grid gradient $u_x \in l^2_{h,0}$. Then the finite element function $p_h(x) \in P^1_{c,0}$ corresponding to u_x is the orthogonal projection of the piecewise constant function $u_{h,x}$ onto $P^1_{c,0}$. In other words, it is the unique solution $p_h \in P^1_{c,0}$ of*

$$(p_h, q_h) = (u_{h,x}, q_h) \qquad \forall q_h \in P^1_{c,0}. \tag{86}$$

In addition, we have, with $q_h \in P^1_{c,0}$ corresponding to $q \in l^2_{h,0}$,

$$(P_x u_x, q)_h = (p_h, q_h) = (u_x, P_x q)_h. \tag{87}$$

(ii) *Let $u, v \in \mathcal{S}$ and let $(u_h, p_h), (v_h, q_h) \in P^1_{c,0} \times P^1_{c,0}$ be the matching finite element functions. Then the bilinear form $\langle .;. \rangle_h$ defined on $\mathcal{S} \times \mathcal{S}$ by*

$$\langle u, v \rangle_h = (\delta_x^4 u, v)_h = \frac{12}{h^2}\big(u_{h,x} - p_h, v_{h,x} - q_h\big) = (u, \delta_x^4 v)_h \tag{88}$$

is a scalar product on $\mathcal{S} \times \mathcal{S}$.

(iii) *Translated in terms of finite difference operators,* (88) *is*

$$\begin{aligned}\langle u, v \rangle_h = &\sum_{j=0}^{N-1} h \frac{u_{x,j+1} - u_{x,j}}{h} \frac{v_{x,j+1} - v_{x,j}}{h} \\ &+ \frac{12}{h^2} \sum_{j=0}^{N-1} h \left[\frac{u_{j+1} - u_j}{h} - \frac{1}{2}(u_{x,j} + u_{x,j+1})\right]\left[\frac{v_{j+1} - v_j}{h} - \frac{1}{2}(v_{x,j} + v_{x,j+1})\right].\end{aligned} \tag{89}$$

Proof. (i) The discrete gradient $u_x \in l^2_{h,0}$ is defined by

$$\big[P_x u_x\big]_j = \delta_x u_j, \quad 1 \le j \le N-1, \tag{90}$$

where P_x is the Simpson operator given in (31). Equation (90) is equivalent to

$$(u_x, q)_h + \frac{1}{6} h^2 (\delta_x^2 u_x, q)_h = (\delta_x u, q)_h \qquad \forall q \in l^2_{h,0}. \tag{91}$$

Taking any $q \in l^2_{h,0}$ and the p_h corresponding to $u_x \in l^2_{h,0}$, and using (83), (84), and (85), we can rewrite (91) as

$$\begin{aligned}(u_{h,x}, q_h) = (\delta_x u, q)_h &= (u_x, q)_h + \frac{h^2}{6}(\delta_x^2 u_x, q)_h \\ &= (p_h, q_h) + \frac{h^2}{6}(p_{h,x}, q_{h,x}) - \frac{h^2}{6}(p_{h,x}, q_{h,x}) \\ &= (p_h, q_h),\end{aligned}$$

which gives (86). The symmetry of P_x is clear from the definition; see (31), (62). In addition, we have

$$(P_x u_x, q)_h = (\delta_x u, q)_h = (u_{h,x}, q_h) = (p_h, q_h), \tag{92}$$

which proves (87).

(ii) The Stephenson biharmonic operator is (see (34))

$$\delta_x^4 u_j = \frac{12}{h^2}\Big\{(\delta_x u_x)_j - \delta_x^2 u_j\Big\}. \tag{93}$$

We have

$$(\delta_x^4 u, v)_h = \frac{12}{h^2}\big[(p_{h,x}, v_h) + (u_{h,x}, v_{h,x})\big] = \frac{12}{h^2}\big(v_{h,x}, u_{h,x} - p_h\big). \tag{94}$$

Subtracting $(q_h, u_{h,x} - p_h) = 0$ from (94), we deduce

$$\langle u, v\rangle_h = (\delta_x^4 u, v)_h = \frac{12}{h^2}\big(u_{h,x} - p_h, v_{h,x} - q_h\big). \tag{95}$$

We verify now that $\langle u, u\rangle_h^{1/2}$ is a norm on $\mathcal{S}$. $\langle u, u\rangle_h = 0$ is equivalent to $|u_{h,x} - p_h| = 0$. Therefore the piecewise affine function $p_h \in P_{c,0}^1$ is actually piecewise constant. Since it vanishes at $x = 0$ and is continuous at any x_j, we have $p_h \equiv 0$, which is $u_{h,x} \equiv 0$. Therefore u_h is piecewise constant as well. Since $u_h(0) = 0$ we have also $u_h \equiv 0$.

Finally, we prove (89). Recall that for any $q_h \in P_{c,0}^1$, the difference $q_h - \overline{q}_h$ is orthogonal to piecewise constant functions. Thus, replacing in (95) p_h, q_h by $\overline{p}_h$, $\overline{q}_h$, respectively, and noting (see (83)) that

$$(p_h, q_h) = (\overline{p}_h, \overline{q}_h) + \frac{h^2}{12}(p_{h,x}, q_{h,x}), \tag{96}$$

we get

$$\langle u, v\rangle_h = (p_{h,x}, q_{h,x}) + \frac{12}{h^2}(u_{h,x} - \bar{p}_h, v_{h,x} - \bar{q}_h), \tag{97}$$

which gives (89) using (80). ∎

Remarks. The result of Lemma 3.4(ii) gives the uniqueness of the discrete solution of scheme (30).

The following lemma states the discrete counterpart of the equivalence of

(i) $|u_x|$ and $\|u\|_{H_1}$ for $u \in H_0^1$;

(ii) $|u_{xx}|$ and $\|u\|_{H_2}$ for $u \in H_0^2$.

LEMMA 3.5. *There exist constants C, C', C'' independent of h such that for any grid function $u \in \mathcal{S}$,*

$$\text{(i) } |u_h| \le |u|_h \le C|\delta_x^+ u|_h = C|u_{h,x}| \quad (\textit{Poincaré inequality}); \tag{98}$$

$$\text{(ii) } |\delta_x^+ u|_h \le C'\langle u, u\rangle_h^{1/2}; \tag{99}$$

$$\text{(iii) } |\delta_x^+ u_x|_h \le C''\langle u, u\rangle_h^{1/2}. \tag{100}$$

Proof. Inequality (i) is simply the Poincaré inequality (21) in the one-dimensional setting, reformulated with the finite element notation. Inequality (iii) follows directly from (97) since $\delta_x^+ u_x = p_{h,x}$ as piecewise constant functions.

For (ii), we use the notation p for the grid function u_x and, as before, denote by u_h, p_h the $P_{c,0}^1$ functions associated with u, p, respectively. In view of (86), we have

$$\begin{aligned} |\delta_x^+ u|_h^2 = |u_{h,x}|^2 &= (u_{h,x} - p_h, u_{h,x} - p_h) + (p_h, p_h) \\ &= \frac{h^2}{12}\langle u, u\rangle_h + |p_h|^2, \end{aligned} \tag{101}$$

where in the second equality we have used (95). Now, applying the Poincaré inequality (98) to p instead of u, we get

$$|p_h|^2 \le C^2 |\delta_x^+ p|_h^2 \le C^2 (C'')^2 \langle u, u\rangle_h, \tag{102}$$

where in the last inequality we have used (100). Inserting this inequality in (101), we obtain (99) with $C' = CC''$. ∎

Remarks. 1. We know that $|u_{xx}|_{0,[0,1]}$ is a norm on the Sobolev space H_0^2. We may wonder if, at the discrete level, $|\delta_x^+ u_x|_h = |p_{h,x}|_{0,[0,1]}$ is a norm on $\mathcal{S}$. Actually it is a norm only if the number of points N is an even integer. We have that $p_{h,x} = 0$ implies $p_h = 0$. But the relation $P_x u_x = \delta_x u$ implies only $\delta_x u = 0$, which gives $u = 0$ only if N is an even integer.

2. For other finite difference schemes for the biharmonic problem and their link with the finite element method, we refer to the book by Li, Chen, and Wu [16].

3.4. Convergence of the Stephenson scheme. We derive now the following convergence result

PROPOSITION 3.1. *Let U be the $P_{c,0}^1$ Lagrange interpolate of the exact solution $u(x)$ of (26) and $\tilde{u}$ the discrete solution of (30). Then the following error estimate holds in the mesh dependent norm $\langle \tilde{v}, \tilde{v}\rangle_h^{1/2}$,*

$$\langle U - \tilde{u}, U - \tilde{u}\rangle_h^{1/2} \le Ch^{3/2}\big(|f''|_{\infty,[0,1]} + |f'|_{\infty,[0,1]}\big), \tag{103}$$

where the constant C is independent of h.

Proof. We estimate as usual the error by the sum of the approximation error and of the consistency error. Here, we work with the discrete norm $\langle ., .\rangle_h^{1/2}$, so that there is no approximation error. We have

$$\langle U - \tilde{u}, U - \tilde{u}\rangle_h^{1/2} = \sup_{\tilde{v}\in\mathcal{S}, \tilde{v}\neq 0} \frac{\langle U - \tilde{u}, \tilde{v}\rangle_h}{\langle \tilde{v}, \tilde{v}\rangle_h^{1/2}}. \tag{104}$$

For the numerator on the right-hand side of (104),

$$\langle U - \tilde{u}, \tilde{v}\rangle_h = (\delta_x^4 (U - \tilde{u}), \tilde{v})_h = h \sum_{j=1}^{N-1} (\delta_x^4 U_j - f_j)\tilde{v}_j. \tag{105}$$

Therefore, in view of Lemma 3.2,

$$\begin{aligned} |\langle U - \tilde{u}, \tilde{v}\rangle_h| &\le |\delta_x^4 U - f|_h |\tilde{v}|_h \\ &\le Ch^{3/2} |\tilde{v}|_h \big(|f''|_{\infty,[0,1]} + |f'|_{\infty,[0,1]}\big). \end{aligned} \tag{106}$$

Using the fact that $|\tilde{v}|_h \le C\langle \tilde{v}, \tilde{v}\rangle_h^{1/2}$ (see (99), (100)), we find that

$$|\langle U - \tilde{u}, \tilde{v}\rangle_h| \le Ch^{3/2} \langle \tilde{v}, \tilde{v}\rangle_h^{1/2} \big(|f''|_{\infty,[0,1]} + |f'|_{\infty,[0,1]}\big), \tag{107}$$

which gives the result. ∎

4. The Stephenson scheme in two dimensions.

4.1. The compact biharmonic scheme of Stephenson. We consider in this section the biharmonic problem in a square $\Omega =]0,1[^2$:

$$\begin{cases} \Delta^2 u(x,y) = \partial_x^4 u(x,y) + \partial_y^4 u(x,y) + 2\partial_{xy}^2 u(x,y) = f(x,y), \quad (x,y) \in \Omega, \\ u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{108}$$

For any $f \in L^2(\Omega)$, problem (108) has a unique solution $u \in H_0^2(\Omega)$. Its discrete version, using the Stephenson scheme, is to find a solution $u_{i,j} \in L_{h,0}^2$ to the equation

$$\begin{cases} \Delta_h^2 u_{i,j} = f(x_i, y_j), \quad 1 \le i,j \le N-1, \\ u_{i,j} = u_{x,i,j} = u_{y,i,j} = 0 \quad \text{for } \{i,j\} \in \{0,N\}. \end{cases} \tag{109}$$

The Stephenson biharmonic operator Δ_h^2 is defined by

$$\Delta_h^2 u_{i,j} = \delta_x^4 u_{i,j} + \delta_y^4 u_{i,j} + 2\delta_x^2 \delta_y^2 u_{i,j}. \tag{110}$$

For any $u \in L_{h,0}^2$, the grid gradient $(u_x, u_y) \in (L_{h,0}^2)^2$ is defined by

$$\begin{cases} P_x u_{x,i,j} = \delta_x u_{i,j}, \quad 1 \le i,j \le N-1, \\ P_y u_{y,i,j} = \delta_y u_{i,j}, \quad 1 \le i,j \le N-1, \end{cases} \tag{111}$$

where P_x, P_y are the Simpson operators (see (31)),

$$\begin{cases} P_x = Id + \dfrac{1}{6} h^2 \delta_x^2, \\ P_y = Id + \dfrac{1}{6} h^2 \delta_y^2. \end{cases} \tag{112}$$

The one-dimensional operators $\delta_x^4 u_{i,j}$, $\delta_y^4 u_{i,j}$ are given as functions of u, u_x, u_y by

$$\delta_x^4 u_{i,j} = \frac{12}{h^2}\big[(\delta_x u_x)_{i,j} - (\delta_x^2 u)_{i,j}\big], \;\; \delta_y^4 u_{i,j} = \frac{12}{h^2}\big[(\delta_y u_y)_{i,j} - (\delta_y^2 u)_{i,j}\big]. \tag{113}$$

For the convenience of the reader, we recall briefly how the operator Δ_h^2 has been originally derived by Stephenson [19]. At each point (x_i, y_j) of the grid, $0 \le i,j \le N$, are attached the three unknowns $u_{i,j}$, $u_{x,i,j}$, $u_{y,i,j}$ as well as a fourth-order polynomial $P_{i,j}$, simply denoted $P(x,y)$,

$$P(x,y) = \sum_{x^l y^m \in \mathcal{V}} a_{l,m} x^l y^m, \tag{114}$$

where the monomial set $\mathcal{V}$ is

$$\mathcal{V} = \{1, x, y, x^2, y^2, xy, x^3, x^2 y, xy^2, y^3, x^4, x^2 y^2, y^4\}, \quad \#\mathcal{V} = 13. \tag{115}$$

The 13 coefficients $a_{l,m}$ are uniquely determined by the following collocation conditions (see Figure 2):

$$\begin{cases} \bullet \text{ 9 collocations for } u_{l,m} \text{ at points } (x_l, y_m) \text{ for } l \in \{i-1, i, i+1\}, \\ \quad m \in \{j-1, j, j+1\}. \\ \bullet \text{ 2 collocations for } u_{x,l,m} \text{ at points } (x_{i-1,j}, y_{i,j}),\ (x_{i+1,j}, y_{i,j}). \\ \bullet \text{ 2 collocations for } u_{y,l,m} \text{ at points } (x_{i,j}, y_{i,j+1}),\ (x_{i,j}, y_{i,j-1}). \end{cases} \tag{116}$$

The collocation system gives a 13×13 linear system which can be solved explicitly. The result is given by [19].

LEMMA 4.1. *Denoting by* $\diamond$, $\square$, *and* $\diamond'$ *the finite difference operators*

$$\begin{cases} \diamond u_{i,j} = u_{i-1,j} + u_{i+1,j} + u_{i,j+1} + u_{i,j-1}, \\ \square u_{i,j} = u_{i+1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} + u_{i-1,j+1}, \\ \diamond' u_{i,j} = u_{x,i+1,j} - u_{x,i-1,j} + u_{y,i,j+1} - u_{y,i,j-1}, \end{cases} \tag{117}$$

the 13 *coefficients* $a_{l,m}$ *of* $P(x,y)$ *at point* (x_i, y_j) *uniquely determined by the* 13 *conditions* (116) *are*

(118)

$$\begin{cases} a_{0,0} = u_{i,j}, \\ a_{1,0} = \dfrac{3}{2}\delta_x u_{i,j} - \dfrac{1}{4}(u_{x,i+1,j} + u_{x,i-1,j}), \quad a_{0,1} = \dfrac{3}{2}\delta_y u_{i,j} - \dfrac{1}{4}(u_{y,i,j+1} + u_{y,i,j-1}), \\ a_{2,0} = \delta_x^2 u_{i,j} - \dfrac{1}{2}(\delta_x u_x)_{i,j}, \quad a_{0,2} = \delta_y^2 u_{i,j} - \dfrac{1}{2}(\delta_y u_y)_{i,j}, \quad a_{1,1} = \delta_{xy} u_{i,j}, \\ a_{3,0} = \dfrac{1}{6}(\delta_x^2 u_x)_{i,j}, \quad a_{0,3} = \dfrac{1}{6}(\delta_y^2 u_y)_{i,j}, \\ a_{2,1} = \dfrac{1}{2}(\delta_x^2 \delta_y u)_{i,j}, \quad a_{1,2} = \dfrac{1}{2}(\delta_y^2 \delta_x u)_{i,j}, \\ a_{4,0} = \dfrac{1}{2h^2}\big[(\delta_x u_x)_{i,j} - \delta_x^2 u_{i,j}\big], \quad a_{0,4} = \dfrac{1}{2h^2}\big[(\delta_y u_y)_{i,j} - \delta_y^2 u_{i,j}\big], \\ a_{2,2} = \dfrac{1}{4}(\delta_x^2 \delta_y^2 u)_{i,j}. \end{cases}$$

The gradient of $P(x,y)$ at (x_i, y_j) is $(\partial_x P(x_i,y_j), \partial_y P(x_i,y_j)) = (a_{1,0}, a_{0,1})$. Defining $u_{x,i,j} = P_x(x_i,y_j)$, $u_{y,i,j} = P_y(x_i,y_j)$, we obtain (111). Furthermore the operators δ_x^4, δ_y^4 are defined by

$$\begin{cases} \delta_x^4 u_{i,j} = \partial_x^4 P(x_i,y_j) = 24 a_{4,0}, \\ \delta_y^4 u_{i,j} = \partial_y^4 P(x_i,y_j) = 24 a_{0,4}, \end{cases} \tag{119}$$

which is (113). Finally the operator $\Delta_h^2 u_{i,j}$ is defined by $\Delta_h^2 u_{i,j} = \Delta^2 P(x_i,y_j) = 24a_{4,0} + 8a_{2,2} + 24a_{0,4}$, which is (110). Furthermore, by expanding the finite difference operators, we find the following expression for the biharmonic operator Δ_h^2:

$$\begin{aligned} \Delta_h^2 u_{i,j} = \frac{1}{h^4}\Big\{ & 56 u_{i,j} - 16\big[u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1}\big] \\ & + 2\big[u_{i+1,j+1} + u_{i-1,j+1} + u_{i-1,j-1} + u_{i+1,j-1}\big] \\ & + 6h\big[(u_x)_{i+1,j} - (u_x)_{i-1,j} + (u_y)_{i,j+1} - (u_y)_{i,j-1}\big] \Big\}. \end{aligned}$$

For alternative schemes for (108), see [19, 1].

4.2. Consistency and convergence for the elliptic operator. The order of consistency is deduced from the consistency in the one-dimensional case.

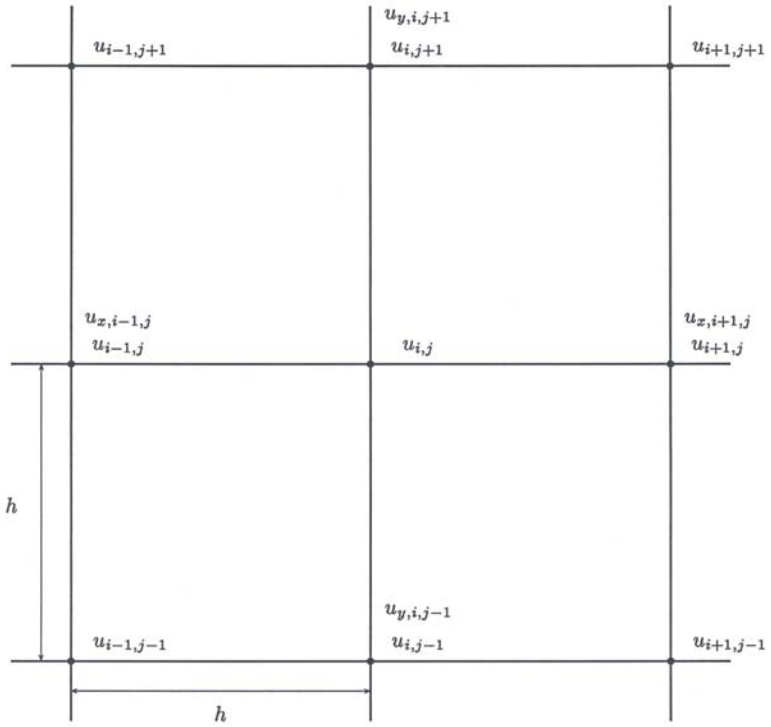


FIG. 2. *Stephenson's scheme for $\Delta^2 u = f$: The finite difference operator $\Delta_h^2 u_{i,j}$ at point (i,j) is $\Delta_h^2 u_{i,j} = \Delta^2 Q(x_i, y_j)$, where $Q(x,y) \in P^{3.5}([x_{i-1}, x_{i+1}] \times [y_{j-1}, y_{j+1}])$ is defined by the 13 collocated values on the picture.*

LEMMA 4.2. *Let u be continuously differentiable up to sixth order in Ω and suppose that it vanishes, along with its gradient on $\partial\Omega$. Then the truncation grid function $e = \Delta_h^2 u(x_i, y_j) - \Delta^2 u(x_i, y_j) \in L_{h,0}^2$ satisfies*

$$|e|_h \le Ch^{3/2} \|u\|_{6,\infty}, \tag{120}$$

where $\|u\|_{6,\infty}$ is

$$|u|_{6,\infty} = \sum_{0 \le \alpha_1 + \alpha_2 \le 6} |\partial_x^{\alpha_1} \partial_y^{\alpha_2} u|_{\infty,[0,1]^2}.$$

Proof. We have

$$|\Delta_h^2 u - \Delta^2 u|_h \le |\delta_x^4 u - \partial_x^4 u|_h + |\delta_y^4 u - \partial_y^4 u|_h + 2|\delta_x^2 \delta_y^2 u - \partial_x^2 \partial_y^2 u|_h. \tag{121}$$

Using the consistency result (54) row by row and column by column we obtain

$$|\delta_x^4 u - \partial_x^4 u|_h \le Ch^{3/2} \big(|\partial_x^6 u|_{\infty,[0,1]^2} + |\partial_x^5 u|_{\infty,[0,1]^2} \big), \tag{122}$$

$$|\delta_y^4 u - \partial_y^4 u|_h \le Ch^{3/2} \big(|\partial_y^6 u|_{\infty,[0,1]^2} + |\partial_y^5 u|_{\infty,[0,1]^2} \big). \tag{123}$$

The consistency for the mixed term is deduced from (45):

$$|\delta_x^2\delta_y^2 u - \partial_x^2\partial_y^2 u|_h \le Ch^2\Big(\sum_{\alpha_1+\alpha_2=6} |\partial_x^{\alpha_1}\partial_y^{\alpha_2} u|_{\infty,[0,1]^2}\Big). \qquad \square \tag{124}$$

In order to carry out convergence analysis, we need to develop discrete analogues of the basic differential estimates, as in the one-dimensional case of section 3. We do this in the framework of a suitable "finite element" space, namely, the Q_c^1 space of continuous functions in Ω satisfying the following condition: In every cell $K_{i+1/2,j+1/2} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$, they are linear (separately) in x, y. Otherwise stated, it is (in every cell) in $\mathrm{Span}(1, x, y, xy)$. The subspace of interest to us is $Q_{c,0}^1$, consisting of functions (in Q_c^1) vanishing on $\partial\Omega$. It is clear how to match an element $u_h \in Q_{c,0}^1$ to a given $u \in L_{h,0}^2$: we simply take the function $a_0 + a_1 x + a_2 y + a_3 xy$, which interpolates the four values $u_{i,j}$, $u_{i+1,j}$, $u_{i,j+1}$, $u_{i+1,j+1}$. Since $u_h(x, y)$ is linear in x (resp., in y) for every fixed value of y (resp., of x), we can in particular treat the function $u(x_i, y_j)$, for every fixed j, as a function of x_i in $l_{h,0}^2$ and then associate with it the functions u_x in $l_{h,0}^2$ (see (30)) and u_h, p_h, their associated $P_{c,0}^1$ functions.

Note that these functions are determined for each fixed value of y_j. In the same way, we define the piecewise constant in $[x_j, x_{j+1}]$ function $u_{h,x}(., y_j)$. We define also the analogous functions in the y direction. Finally, $u_{h,xy}$ is the piecewise (in cells) constant function given by the coefficient a_3 above. We now equip $Q_{c,0}^1$ with two scalar products. Each of them corresponds to an $L^2(0,1)$ product in one direction (i.e., the function is regarded as an element of $P_{c,0}^1$ in that direction), followed by an $l_{h,0}^2$ product in the other direction. They are given by

$$\begin{cases} (u_h, v_h)^x = h\sum_{j=1}^{N-1} (u_h(., y_j), v_h(., y_j))_{L^2(0,1)}, \\ (u_h, v_h)^y = h\sum_{i=1}^{N-1} (u_h(x_i, .), v_h(x_i, .))_{L^2(0,1)}. \end{cases} \tag{125}$$

The link between the grid scalar product $(u, v)_h$ on $L_{h,0}^2$ and the two scalar products $(u_h, v_h)^x$, $(u_h, v_h)^y$ is given by (see (83))

$$(u, v)_h = (u_h, v_h)^x + \frac{h^2}{6}(u_{h,x}, v_{h,x})^x, \tag{126}$$

$$(u, v)_h = (u_h, v_h)^y + \frac{h^2}{6}(u_{h,y}, v_{h,y})^y. \tag{127}$$

As in the one-dimensional case (see (33)), we introduce here a space $\mathcal{S}$ consisting of triples $(u, u_x, u_y) \in L_{h,0}^2$, where u_x, u_y are related to u by (111). For brevity, we shall sometimes refer to the triple simply by $u \in \mathcal{S}$. As in the one-dimensional case (see Lemma 3.4), we have the following result.

LEMMA 4.3. *Let $u \in \mathcal{S}$. Let $p_h, q_h \in Q_{c,0}^1$ correspond to u_x, u_y, respectively. Then they are the projections of $u_{h,x}$, $u_{h,y}$ in the following sense:*

$$(p_h, v_h)^x = (u_{h,x}, v_h)^x,\ (q_h, v_h)^y = (u_{h,y}, v_h)^y \qquad \forall v_h \in Q_{c,0}^1. \tag{128}$$

Proof. For each $1 \le j_0 \le N-1$, it results from (86) that

$$\begin{aligned}(p_h, v_h)^x &= h \sum_{j=1}^{N-1} (p_h(.,y_j), v_h(.,y_j))_{L^2(0,1)} \\ &= h \sum_{j=1}^{N-1} (u_{h,x}(.,y_j), v_h(.,y_j))_{L^2(0,1)} \\ &= (u_{h,x}, v_h)^x.\end{aligned}$$

Therefore, the function $p_h \in Q^1_{c,0}$ matching $u_x \in L^2_{h,0}$ is the unique solution of

$$(p_h, v_h)^x = (u_{h,x}, v_h)^x \qquad \forall v_h \in Q^1_{c,0}. \tag{129}$$

The proof is the same for $u_{h,y}$. ∎

We summarize in the following proposition the basic properties of the discrete operator Δ_h^2. As in the one-dimensional case, that operator gives rise to a *positive definite* bilinear form.

PROPOSITION 4.1. (i) *Let* $(u, u_x, u_y), (v, v_x, v_y) \in \mathcal{S}$, *and let* (u_h, p_h, q_h), (v_h, r_h, z_h) *be their matches, respectively, in* $Q^1_{c,0}$. *Then the discrete biharmonic operator* Δ_h^2 *defined by*

$$\Delta_h^2 u_{i,j} = \delta_x^4 u_{i,j} + \delta_y^4 u_{i,j} + 2\delta_x^2 \delta_y^2 u_{i,j}, \quad 1 \le i,j \le N-1, \tag{130}$$

induces a scalar product $\langle u, v\rangle_h = (\Delta_h^2 u, v)_h$ *on* $\mathcal{S} \times \mathcal{S}$ *defined by*

$$\begin{aligned}\langle u, v\rangle_h = (\Delta_h^2 u, v)_h = {} & \frac{12}{h^2}(u_{h,x} - p_h, v_{h,x} - r_h)^x + \frac{12}{h^2}(u_{h,y} - q_h, v_{h,y} - z_h)^y \\ & + 2(u_{h,xy}, v_{h,xy}).\end{aligned} \tag{131}$$

In particular, the discrete operator Δ_h^2 *is symmetric positive definite on* $\mathcal{S}$.

(ii) *In terms of the basic finite difference operators, the product* $\langle u, v\rangle_h$ *is given by*

$$\begin{aligned}(\Delta_h^2 u, v)_h = {} & (\delta_x^+ u_x, \delta_x^+ v_x)_h + (\delta_y^+ u_y, \delta_y^+ v_y)_h + 2(\delta_x^+ \delta_y^+ u, \delta_x^+ \delta_y^+ v)_h \\ & + \frac{12}{h^2}\left(\delta_x^+ u - \frac{1}{2}(u_x + u_{x,i+1,j}), \delta_x^+ v - \frac{1}{2}(v_x + v_{x,i+1,j})\right)_h \\ & + \frac{12}{h^2}\left(\delta_y^+ v - \frac{1}{2}(u_y + u_{y,i,j+1}), \delta_y^+ v - \frac{1}{2}(v_y + v_{y,i,j+1})\right)_h.\end{aligned} \tag{132}$$

(iii) *We have the two following coercivity properties of the norm* $\langle u, u\rangle_h = (\Delta_h^2 u, u)_h$*:*

$$\langle u, u\rangle_h \ge C\left[|\delta_x^+ u_x|_h^2 + |\delta_y^+ u_y|_h^2 + |\delta_x^+ u_y|_h^2 + |\delta_y^+ u_x|_h^2\right] \tag{133}$$

and

$$\langle u, u\rangle_h^{1/2} \ge C' |u|_h, \tag{134}$$

where C, C' *are constants independent of* h.

Proof. (i) By (130), we have

$$(\Delta_h^2 u, v)_h = \underbrace{(\delta_x^4 u, v)_h}_{\text{(I)}} + \underbrace{(\delta_y^4 u, v)_h}_{\text{(II)}} + 2\underbrace{(\delta_x^2\delta_y^2 u, v)_h}_{\text{(III)}}. \tag{135}$$

We consider separately each term (I), (II), (III). For the term (I), we have

$$\begin{aligned}(\delta_x^4 u, v)_h &= h\sum_{j=1}^{N}\Big(\delta_x^4 u(\cdot, y_j), v(\cdot, y_j)\Big)_h \\ &= h\sum_{j=1}^{N}\left\{\frac{12}{h^2}(u_{h,x}(\cdot, y_j) - p_h, v_{h,x}(\cdot, y_j) - r_h(\cdot, y_j)\right\} \\ &= \frac{12}{h^2}(u_{h,x} - p_h, v_{h,x} - r_h)^x.\end{aligned}$$

In the same way

$$(\delta_y^4 u, v)_h = \frac{12}{h^2}(u_{h,y} - q_h, v_{h,y} - z_h)^y. \tag{136}$$

For (III), we just note that

$$(\delta_x^2\delta_y^2 u, v)_h = (\delta_x^+\delta_y^+ u, \delta_x^+\delta_y^+ u)_h = (u_{h,xy}, v_{h,xy}). \tag{137}$$

Consider now the positive-definiteness of (131). Suppose that $(\Delta_h^2 u, u) = 0$. Then $p_h(., y_j)$ is constant and continuous and is zero at the end points; therefore $p_h = 0$. The same result holds for q_h and u_h. We conclude that $\langle u, u\rangle_h^{1/2} = (\Delta_h^2 u, u)_h^{1/2}$ is a norm in $\mathcal{S}$.

(ii) Translating (131) in term of finite difference operators, we obtain (132), as in (89).

(iii) It results from (132) that

$$(\Delta_h^2 u, u)_h \geq |\delta_x^+ u_x|_h^2 + |\delta_y^+ u_y|_h^2 + 2|\delta_x^+\delta_y^+ u|_h^2. \tag{138}$$

For the mixed term $\delta_x^+\delta_y^+ u$, we will show next that

$$|\delta_x^+\delta_y^+ u|_h \geq \frac{1}{6}|\delta_x^+ u_y|_h. \tag{139}$$

Indeed

$$\delta_x^+\delta_y^+ u_{i,j} = \frac{\delta_y^+ u_{i+1,j} - \delta_y^+ u_{i,j}}{h}. \tag{140}$$

Using $\delta_y^+ u_{i,j} = \delta_y u_{i,j} + \frac{h}{2}\delta_y^2 u_{i,j}$ and the definition of P_y (see (112)), we deduce

$$\begin{aligned}\delta_x^+\delta_y^+ u_{i,j} &= \frac{\delta_y u_{i+1,j} - \delta_y u_{i,j}}{h} + \frac{1}{2}\big[\delta_y^2 u_{i+1,j} - \delta_y^2 u_{i,j}\big] \\ &= \frac{1}{h}\big[u_{y,i+1,j} - u_{y,i,j}\big] + \frac{h}{6}\big[\delta_y^2 u_{y,i+1,j} - \delta_y^2 u_{y,i,j}\big] + \frac{1}{2}\big[\delta_y^2 u_{i+1,j} - \delta_y^2 u_{i,j}\big] \\ &= \delta_x^+ u_{y,i,j} + \frac{h^2}{6}\delta_y^2\delta_x^+ u_{y,i,j} + \frac{1}{2}h\delta_y^2\delta_x^+ u_{i,j}.\end{aligned}$$

In addition, using the definition of δ_y^2 we have

$$|\delta_y^2 \delta_x^+ u_y| \le \frac{4}{h^2} |\delta_x^+ u_y|_h \tag{141}$$

and

$$|\delta_y^2 \delta_x^+ u|_h \le \frac{2}{h} |\delta_y^+ \delta_x^+ u|_h. \tag{142}$$

Therefore, we have

$$\begin{aligned} |\delta_x^+ \delta_y^+ u|_h &\ge |\delta_x^+ u_y|_h - \frac{h^2}{6} |\delta_y^2 \delta_x^+ u_y|_h - \frac{h}{2} |\delta_y^2 \delta_x^+ u|_h \\ &\ge |\delta_x^+ u_y|_h - \frac{2}{3} |\delta_x^+ u_y|_h - |\delta_x^+ \delta_y^+ u|_h, \end{aligned}$$

which gives finally $2|\delta_x^+ \delta_y^+ u|_h \ge \frac{1}{3} |\delta_x^+ u_y|_h$, or equivalently (139). We proceed in the same way in proving the symmetric estimate

$$|\delta_x^+ \delta_y^+ u|_h \ge \frac{1}{6} |\delta_y^+ u_x|_h. \tag{143}$$

Finally, the last coercivity inequality (134) is obtained starting from

$$|\delta_x^+ u|_h^2 = (|u_{h,x}|^x)^2 \tag{144}$$

and following along the same lines as in the proof of (99) in Lemma 3.5. ∎

We conclude this section with the following error estimate.

PROPOSITION 4.2. *Let U be the $Q_{c,0}^1$ Lagrange interpolation of the exact solution $u(x)$ of* (108) *and $\tilde{u}$ the discrete solution of* (109). *Then there exists a constant C independent of h such that*

$$\langle U - \tilde{u}, U - \tilde{u} \rangle_h^{1/2} \le C h^{3/2} \sum_{\alpha_1 + \alpha_2 \le 6} |\partial_x^{\alpha_1} \partial_y^{\alpha_1} u|_{\infty, [0,1]^2}. \tag{145}$$

Proof. The proof follows along the same lines as the one of Proposition 3.1. We use in particular (134). ∎

5. A Stephenson-based compact scheme for the streamfunction formulation of the Navier–Stokes equations. The pure streamfunction form of the Navier–Stokes equation is

$$\partial_t \Delta \psi = -\nabla^\perp \psi \cdot \nabla(\Delta \psi) + \nu \Delta^2 \psi. \tag{146}$$

The streamfunction was introduced already by Lagrange; see [15, Chap. IV]. For simplicity, we deal only with the "no-slip" boundary condition, namely, the velocity vanishes on the boundary. This implies that we seek the streamfunction $\psi \in H_{h,0}^2$ (see [3] for a full discussion of the functional space for ψ). The notation is as follows. We denote by $\psi_{i,j} \in L_{h,0}^2$ a grid function and by $\psi_{x,i,j}, \psi_{y,i,j} \in L_{h,0}^2$ the Stephenson gradient defined by

$$P_x \psi_x = \delta_x \psi, \qquad P_y \psi_y = \delta_y \psi, \tag{147}$$

where the interpolation operators P_x, P_y are (see (112))

$$P_x\psi_{|i,j} = \frac{1}{6}\psi_{i-1,j} + \frac{2}{3}\psi_{i,j} + \frac{1}{6}\psi_{i+1,j}, \quad P_y\psi_{|i,j} = \frac{1}{6}\psi_{i,j-1} + \frac{2}{3}\psi_{i,j} + \frac{1}{6}\psi_{i,j+1}. \tag{148}$$

The discrete gradient $\nabla_h\psi$ is defined as the pair of the discrete functions (ψ_x, ψ_y) and the discrete velocity is defined as the discrete curl of the streamfunction in the sense

$$\nabla_h^\perp\psi_{i,j} = U_{i,j} = \big[u_{i,j}, v_{i,j}\big] = \big[-\psi_{y,i,j}, \psi_{x,i,j}\big]. \tag{149}$$

The discrete Laplacian is defined by the standard five-points formula

$$\Delta_h\psi_{i,j} = \delta_x^2\psi_{i,j} + \delta_y^2\psi_{i,j}. \tag{150}$$

The discrete Stephenson biharmonic Δ_h^2 introduced in (109) is

$$\Delta_h^2 u_{i,j} = \delta_x^4 u_{i,j} + \delta_y^4 u_{i,j} + 2\delta_x^2\delta_y^2 u_{i,j}, \quad 1 \le i, j \le N-1. \tag{151}$$

Δ_h^2 is a nine point operator acting at every point (i, j) interior to the domain. The semidiscrete scheme associated with (146) consists in finding $\widetilde{\psi}(t) \in L^2_{h,0}$, which satisfies the evolution equation

$$\partial_t\Delta_h\widetilde{\psi} = -\nabla_h^\perp\widetilde{\psi}\cdot(\Delta_h\nabla_h\widetilde{\psi}) + \nu\Delta_h^2\widetilde{\psi}, \tag{152}$$

with initial condition

$$\widetilde{\psi}_{i,j}(0) = (\psi_0)(x_i, y_j). \tag{153}$$

Note that in (152) and in what follows we use pointwise multiplication of functions in $L^2_{h,0}$, i.e., $(u\cdot v)_{i,j} = u_{i,j}v_{i,j}$. We denote by $e(t) = \widetilde{\psi}(t) - \psi(t)$ the difference between the computed and exact solutions. The exact solution verifies

$$\partial_t\Delta_h\psi = -\nabla_h^\perp\psi\cdot\big[\Delta_h\nabla_h(\psi)\big] + \nu\Delta_h^2\psi + F, \tag{154}$$

where F is the truncation error of the scheme depending on the regularity of the exact solution. We call U and $\widetilde{U}$ the discrete velocities associated to ψ, $\widetilde{\psi}$ by

$$U = (-\psi_y, \psi_x), \quad \widetilde{U} = (-\widetilde{\psi}_y, \widetilde{\psi}_x). \tag{155}$$

Recall that in (155), the x and y subscripts stand for the discrete derivatives defined in (147). In particular, ψ_x, ψ_y are not the values of the exact derivatives of ψ. The error $e(t)$ evolves according to

$$\partial_t\Delta_h e - \nu\Delta_h^2 e = -\big[\widetilde{U}\cdot\Delta_h(\widetilde{\psi}_x, \widetilde{\psi}_y) - U\cdot\Delta_h(\psi_x, \psi_y)\big] - F. \tag{156}$$

The right-hand side of (156) is decomposed into four terms:

$$\begin{aligned}\big[(\widetilde{U}\cdot\Delta_h(\widetilde{\psi}_x, \widetilde{\psi}_y) - U\cdot\Delta_h(\psi_x, \psi_y)\big] + F = {}& (\widetilde{U} - U)\cdot\Delta_h\big[(\widetilde{\psi} - \psi)_x, (\widetilde{\psi} - \psi)_y\big] \\ & + (\widetilde{U} - U)\cdot\Delta_h\big[(\psi_x, \psi_y\big] \\ & + U\cdot\Delta_h\big[(\widetilde{\psi} - \psi)_x, (\widetilde{\psi} - \psi)_y\big] + F.\end{aligned}$$

Taking the h scalar product with $e(t)$, we obtain

$$
\begin{aligned}
(\partial_t \Delta_h e_h - \nu \Delta_h^2 e, e)_h = &-\Big((\widetilde{U} - U) \cdot \Delta_h \big[(\widetilde{\psi} - \psi)_x, (\widetilde{\psi} - \psi)_y \big], e \Big)_h \\
&- \big((\widetilde{U} - U) \cdot \Delta_h (\psi_x, \psi_y), e \big)_h \\
&- \Big(U \cdot \Delta_h \big[(\widetilde{\psi} - \psi)_x, \widetilde{\psi} - \psi)_y \big], e \Big)_h \\
&- \big(F, e \big)_h .
\end{aligned} \tag{157}
$$

We denote the four terms of the right-hand side by J_1, J_2, J_3, J_4:

$$
\begin{aligned}
J_1 &= \big((\widetilde{U} - U) \cdot \Delta_h (\widetilde{\psi} - \psi)_x, (\widetilde{\psi} - \psi)_y, e \big)_h , \\
J_2 &= \big((\widetilde{U} - U) \cdot \Delta_h (\psi_x, \psi_y), e \big)_h , \\
J_3 &= \big(U \cdot \Delta_h (\widetilde{\psi} - \psi)_x, (\widetilde{\psi} - \psi)_y, e \big)_h , \\
J_4 &= (F, e)_h .
\end{aligned}
$$

We estimate separately the four terms J_1, J_2, J_3, J_4.

Term J_1. The term J_1 is

$$
J_1 = \big((\widetilde{U} - U) \cdot \Delta_h (e_x, e_y), e \big)_h . \tag{158}
$$

We have

$$
\widetilde{U} - U = \big[- (\widetilde{\psi} - \psi)_y, (\widetilde{\psi} - \psi)_x \big] = (-e_y, e_x), \tag{159}
$$

where the subscripts x and y are the Stephenson derivation operators. Therefore

$$
\begin{aligned}
J_1 &= \big((\widetilde{U} - U) \cdot \Delta_h (e_x, e_y), e \big)_h = \big(- e_y (\delta_x^2 e_x + \delta_y^2 e_x) + e_x (\delta_x^2 e_y + \delta_y^2 e_y), e \big)_h \\
&= \big(- e_y (\delta_x^2 e_x + \delta_y^2 e_x), e \big)_h + \big(e_x (\delta_x^2 e_y + \delta_y^2 e_y), e \big)_h \\
&= - \big(\delta_x^2 e_x, e e_y \big)_h - \big(\delta_y^2 e_x, e e_y \big)_h + \big(\delta_x^2 e_y, e e_x \big)_h + \big(\delta_y^2 e_y, e e_x \big)_h \\
&= \big(\delta_x^+ e_x, \delta_x^+ (e e_y) \big)_h + \big(\delta_y^+ e_x, \delta_y^+ (e e_y) \big)_h \\
&\quad - \big(\delta_x^+ e_y, \delta_x^+ (e e_x) \big)_h - \big(\delta_y^+ e_y, \delta_y^+ (e e_x) \big)_h .
\end{aligned}
$$

In order to formulate a discrete Leibniz rule for $w, z \in L^2_{h,0}$ we use the "shift operators" $(S_x w)_{i,j} = w_{i+1,j}, (S_y z)_{i,j} = z_{i,j+1}$. In terms of these operators we have

$$
\delta_x^+ (wz) = (S_x w)_{i,j} \delta_x^+ z + z \delta_x^+ w, \tag{160}
$$

which is quite easy to verify. Using (160), we expand J_1 in the sum of eight terms:

$$
\begin{aligned}
J_1 = &\big(\delta_x^+ e_x, (S_x e_y)_{i,j} \delta_x^+ e \big)_h + \big(\delta_x^+ e_x, e \delta_x^+ e_y \big)_h \\
&+ \big(\delta_y^+ e_x, (S_y e_y)_{i,j} \delta_y^+ e \big)_h + \big(\delta_y^+ e_x, e \delta_y^+ e_y \big)_h \\
&- \big(\delta_x^+ e_y, (S_x e_x)_{i,j} \delta_x^+ e \big)_h - \big(\delta_x^+ e_y, e \delta_x^+ e_x \big)_h \\
&- \big(\delta_y^+ e_y, (S_y e_x)_{i,j} \delta_y^+ e \big)_h - \big(\delta_y^+ e_y, e \delta_y^+ e_x \big)_h .
\end{aligned}
$$

There is a cancellation of terms 2 and 6 on one hand, and 4 and 8 on the other hand, so that

$$J_1 = \big(\delta_x^+ e_x, (S_x e_y)\delta_x^+ e\big)_h + \big(\delta_y^+ e_x, (S_y e_y)\delta_y^+ e\big)_h \\ + \big(\delta_x^+ e_y, (S_x e_x)\delta_x^+ e\big)_h + \big(\delta_y^+ e_y, (S_y e_x)\delta_y^+ e\big)_h.$$

We now observe that if $\theta \in L^2_{h,0}$, then $|\theta|_{\infty,h} \le \frac{1}{h}|\theta|_h$. We can therefore estimate J_1 as follows:

$$\begin{aligned} |J_1| &= \big|\big((\widetilde{U} - U)\cdot \Delta_h(e_x, e_y), e\big)_h\big| \\ &\le \varepsilon\big[|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_x|_h^2 + |\delta_x^+ e_y|_h^2 + |\delta_y^+ e_y|_h^2\big] + \frac{1}{4\varepsilon}\big[|(e_x, e_y)|_{\infty,h}^2\big(|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big)\big] \\ &\le \varepsilon\big[|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_x|_h^2 + |\delta_x^+ e_y|_h^2 + |\delta_y^+ e_y|_h^2\big] + \frac{C}{\varepsilon h^2}\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big]^2, \end{aligned}$$

where in the last step we have used (51) to estimate $|e_x|_{\infty,h} \le C|\delta_x^+ e|_{\infty,h}$ and $|e_y|_{\infty,h} \le C|\delta_y^+ e|_{\infty,h}$ with a constant independent of h. The factor $\varepsilon > 0$ will be specified later.

Term J_2. The term J_2 is estimated by (C is a generic constant)

$$|J_2| = |\big((\widetilde{U} - U)\cdot \Delta_h(\psi_x, \psi_y), e\big)_h| \le C\big[|\widetilde{U} - U|_h^2 + |e|_h^2\big]. \tag{161}$$

We have used that $\Delta_h(\psi_x, \psi_y)$ is the discrete operator Δ_h composed by the Stephenson gradient applied to the exact solution, and is bounded if the exact solution is sufficiently regular. In addition, using the fact that $\widetilde{U} - U = \big[-(\widetilde{\psi}_y - \psi_y), \widetilde{\psi}_x - \psi_x\big]$, we have

$$|\widetilde{U} - U|_h^2 = |e_x|_h^2 + |e_y|_h^2. \tag{162}$$

Furthermore, we have, in view of (60), (78),

$$|e_x|_h \le C|\delta_x^+ e|_h, \quad |e_y|_h \le C|\delta_y^+ e|_h, \tag{163}$$

and, due to the Poincaré inequality (21), we deduce

$$|J_2| \le C\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big]. \tag{164}$$

Term J_3. We have

$$J_3 = \big[U\cdot \Delta_h(e_x, e_y), e\big]_h = \underbrace{(u\delta_x^2 e_x, e)_h}_{J_{3,1}} + \underbrace{(u\delta_y^2 e_x, e)_h}_{J_{3,2}} \\ + \underbrace{(v\delta_x^2 e_y, e)_h}_{J_{3,3}} + \underbrace{(v\delta_y^2 e_y, e)_h}_{J_{3,4}}.$$

We have

$$J_{3,1} = (u\delta_x^2 e_x, e)_h = (\delta_x^2 e_x, ue)_h = -\big[\delta_x^+ e_x, \delta_x^+(ue)\big]_h. \tag{165}$$

Using (160), the term $J_{3,1}$ is estimated by

$$\begin{aligned} |J_{3,1}| = |\big[\delta_x^+ e_x, \delta_x^+(ue)\big]_h &\le |\delta_x^+ e_x|_h |\delta_x^+(ue)|_h \\ &\le |\delta_x^+ e_x|_h\big[|(S_x u)_{i,j}\delta_x^+ e|_h + |e\delta_x^+ u|_h\big] \\ &\le |\delta_x^+ e_x|_h\big[|u|_{\infty,h}|\delta_x^+ e|_h + |\delta_x^+ u|_{\infty,h}|e|_h\big]. \end{aligned}$$

Therefore, using the Poincaré inequality (21), the term $J_{3,1}$ is estimated by

$$|J_{3,1}| \le \max\big[|u|_{\infty,h}, |\delta_x^+ u|_{\infty,h}\big]\left[\varepsilon|\delta_x^+ e_x|_h^2 + \frac{1}{4\varepsilon}(|\delta_x^+ e|_h + |e|_h)^2\right]$$
$$\le \max(|u|_{\infty,h}, |\delta_x^+ u|_{\infty,h})\left[\varepsilon|\delta_x^+ e_x|_h^2 + \frac{C}{\varepsilon}(|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2)\right].$$

Using the same principle in the y direction, we obtain for the term $J_{3,2}$

(166)
$$|J_{3,2}| = |(u\delta_y^2 e_x, e)_h| \le \max(|u|_{\infty,h}, |\delta_y^+ u|_{\infty,h})\left[\varepsilon|\delta_y^+ e_x|_h^2 + \frac{C}{\varepsilon}(|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2)\right].$$

Therefore, with $m(u) = \max\big[|u|_{\infty,h}, |\delta_x^+ u|_{\infty,h}, |\delta_y^+ u|_{\infty,h}\big]$, the estimate for the term $J_{3,1} + J_{3,2}$ is

(167)
$$|J_{3,1} + J_{3,2}| \le |J_{3,1}| + |J_{3,2}| \le m(u)\left[\varepsilon\big\{|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_x|_h^2\big\} + \frac{C}{\varepsilon}\big\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big\}\right].$$

Treating the term $J_{3,3} + J_{3,4}$ in the same way, we obtain

(168)
$$|J_{3,3} + J_{3,4}| \le |J_{3,3}| + |J_{3,4}| \le m(v)\left[\varepsilon\{|\delta_x^+ e_y|_h^2 + |\delta_y^+ e_y|_h^2\} + \frac{C}{\varepsilon}\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\}\right].$$

The estimate for the term J_3 is finally, with $M(u,v) = \max(m(u), m(v))$,

(169)
$$|J_3| \le M(u,v)\left[\varepsilon\big\{|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_x|_h^2 + |\delta_x^+ e_y|_h^2 + |\delta_y^+ e_y|_h^2\big\} + \frac{2C}{\varepsilon}\big\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big\}\right].$$

Term J_4. The term J_4 is the truncation error and is of order 3/2 (in the $|\cdot|_h$ norm) in view of Lemmas 3.1 and 4.2. For any time $T > 0$, the term J_4 is estimated by

$$|J_4| \le C(T)|e|_h h^{3/2} \le C(T)\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2 + h^3\big], \tag{170}$$

where $C(T)$ is a constant depending only on $T > 0$ and on the regularity of the exact solution $\psi(t)$ on $[0,T]$.

Turning back to (157), we have, on $[0, T_0]$,

$$\left(\frac{\partial}{\partial t}\Delta_h e, e\right)_h - \nu(\Delta_h^2 e, e)_h = -\frac{1}{2}\frac{d}{dt}\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\} - \nu(\Delta_h^2 e, e)_h$$
$$= -J_1 - J_2 - J_3 - J_4,$$

or

$$\frac{1}{2}\frac{d}{dt}\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\} = J_1 + J_2 + J_3 + J_4 - \nu(\Delta_h^2 e, e)_h$$
$$\le |J_1| + |J_2| + |J_3| + |J_4| - \nu(\Delta_h^2 e, e)_h$$
$$\le |J_1| + |J_2| + |J_3| + |J_4|$$
$$- C\nu\big[|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_y|_h^2 + |\delta_x^+ e_y|_h^2 + |\delta_y^+ e_x|_h^2\big],$$

where in the last inequality we have used the coercivity property (133). Collecting the terms of the form $|\delta_x^+ e_x|_h^2 + |\delta_y^+ e_y|_h^2 + |\delta_x^+ e_y|_h^2 + |\delta_y^+ e_x|_h^2$, which appear in the estimates for J_1, J_2, J_3, J_4, and selecting $\varepsilon > 0$ sufficiently small, we find that these terms are absorbed in the right-hand side of the last inequality. We are therefore left with the estimate

$$\frac{d}{dt}\big\{|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big\} \le C\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big]\left[1 + \frac{1}{h^2}(|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2)\right] + C'h^3, \tag{171}$$

where C, C' depend on the exact solution ψ and on the viscosity coefficient ν but not on h.

In order to prove convergence of the approximate solution $\widetilde{\psi}$ to the exact solution ψ using (171), we proceed as follows. We use the fact that at $t = 0$ the error $e = 0$ to prove an estimate for $|\delta_x^+ e|_h + |\delta_y^+ e|_h$ up to any given time $T > 0$.

THEOREM 5.1. *Let $T > 0$. Then there exist constants $C, h_0 > 0$, depending possibly on T, ν, and the exact solution ψ, such that, for all $0 \le t \le T$,*

$$|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2 \le Ch^3, \quad 0 < h \le h_0. \tag{172}$$

Using Corollary 2.1, we obtain a 3/2 convergence rate in the discrete L^2 norm.

Proof. Fix some $K > 0$. Observe that at $t = 0$ we have $e = 0$; hence also $\delta_x^+ e = \delta_y^+ e = 0$ (at $t = 0$). Thus, taking $h > 0$, there exists a time $\tau > 0$ (in general depending on h) such that

$$\sup_{0 \le t \le \tau} \big\{|\delta_x^+ e|_h + |\delta_y^+ e|_h\big\} \le Kh. \tag{173}$$

Inserting (173) in (171) we have for $t \le \tau$

$$\frac{d}{dt}\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big] \le C(1 + K^2)\big[|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2\big] + C'h^3, \quad 0 < h \le h_0; \tag{174}$$

hence by Gronwall's inequality (174) gives

$$|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2 \le C_1 e^{C(1+K^2)t} h^3, \quad t \le \tau, \tag{175}$$

with a suitable constant $C_1 > 0$. Observe that in (175) τ depends on h, and define $\tau_0 = \tau_0(h)$ by

$$\tau_0 = \sup\{t > 0 \text{ such that } |\delta_x^+ e|_h + |\delta_y^+ e|_h \le Kh\}. \tag{176}$$

We have $\tau_0 \ge \tau$ and, as in (175), we obtain

$$|\delta_x^+ e|_h^2 + |\delta_y^+ e|_h^2 \le C_1 e^{C(1+K^2)t} h^3, \quad t \le \tau_0. \tag{177}$$

We can now select h_0 so small that

$$C_1 e^{C(1+K^2)T} h_0 < K^2. \tag{178}$$

Now the definition of τ_0 and (177)–(178) imply that, for any $0 < h \le h_0$, we have $\tau_0(h) \ge T$ and, in particular, for such h, the estimate (175) holds true for all $t \le T$. This concludes the proof of the theorem. ∎

REFERENCES

[1] I. Altas, J. Dym, M. M. Gupta, and R. P. Manohar, *Multigrid solution of automatically generated high-order discretizations for the biharmonic equation*, SIAM J. Sci. Comput., 19 (1998), pp. 1575–1585.

[2] M. Arad, A. Yakhot, and G. Ben-Dor, *A highly accurate numerical solution of a biharmonic equation*, Numer. Methods Partial Differential Equations, 13 (1997), pp. 375–393.

[3] M. Ben-Artzi, J.-P. Croisille, D. Fishelov, and S. Trachtenberg, *A pure-compact scheme for the streamfunction formulation of Navier-Stokes equations*, J. Comput. Phys., 205 (2005), pp. 640–664.

[4] M. Ben-Artzi, D. Fishelov, and S. Trachtenberg, *Vorticity dynamics and numerical resolution of Navier-Stokes equations*, Math. Model. Numer. Anal., 35 (2001), pp. 313–330.

[5] D. Calhoun, *A Cartesian grid method for solving the two-dimensional streamfunction-vorticity equations in irregular regions*, J. Comput. Phys., 176 (2002), pp. 231–275.

[6] L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1960.

[7] J.-P. Croisille, *Keller's box-scheme for the one-dimensional stationary convection-diffusion equation*, Computing, 68 (2002), pp. 37–63.

[8] E. J. Dean, R. Glowinski, and O. Pironneau, *Iterative solution of the stream function-vorticity formulation of the Stokes problem. Applications to the numerical simulation of incompressible viscous flow*, Comput. Methods Appl. Mech. Engrg., 87 (1991), pp. 117–155.

[9] W. E and J.-G. Liu, *Essentially compact schemes for unsteady viscous incompressible flows*, J. Comput. Phys., 126 (1996), pp. 122–138.

[10] P. M. Gresho, *Incompressible fluid dynamics: Some fundamental formulation issues*, Annu. Rev. Fluid Mech., 23 (1991), pp. 413–453.

[11] M. M. Gupta and J. C. Kalita, *A new paradigm for solving Navier-Stokes equations: Stream-function-velocity formulation*, J. Comput. Phys., 207 (2005), pp. 52–68.

[12] M. M. Gupta, R. P. Manohar, and J. W. Stephenson, *Single cell high order scheme for the convection-diffusion equation with variable coefficients*, Internat. J. Numer. Methods Fluids, 4 (1984), pp. 641–651.

[13] T. Y. Hou and B. T. R. Wetton, *Convergence of a finite difference scheme for the Navier–Stokes equations using vorticity boundary conditions*, SIAM J. Numer. Anal., 29 (1992), pp. 615–639.

[14] H. B. Keller, *A new difference scheme for parabolic problems*, in Numerical Solutions of Partial Differential Equations, II, Academic Press, New York, 1971, pp. 327–350.

[15] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1993.

[16] R. Li, Z. Chen, and W. Wu, *Generalized Difference Methods for Differential Equations*, Marcel Dekker, New York, 2000. puncuate

[17] S. A. Orszag and M. Israeli, *Numerical simulation of viscous incompressible flows*, in Annual Review of Fluid Mechanics, Vol. 6, M. Van Dyke, W. A. Vincenti, and J. V. Wehausen, eds., Annual Reviews, Palo Alto, CA, 1974, pp. 281–318.

[18] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, UK, 1981.

[19] J. W. Stephenson, *Single cell discretizations of order two and four for biharmonic problems*, J. Comput. Phys., 55 (1984), pp. 65–80.

[20] J. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2025–2056

NUMERICAL ANALYSIS OF VANKA-TYPE SOLVERS FOR STEADY STOKES AND NAVIER–STOKES FLOWS*

S. MANSERVISI†

Abstract. We consider Vanka-type smoothers for solving Stokes and Navier–Stokes problems. In each iteration step, this smoother requires the solution of several small local subproblems over finite element blocks. It is shown that for particular choices for the blocks, the algorithm always converges to the solution of the Stokes problem and, under suitable conditions, to the solution of the Navier–Stokes problem. The convergence properties are analyzed and numerical examples are presented.

Key words. Navier–Stokes equations, Vanka solvers, finite elements

AMS subject classifications. 76D05, 65N55

DOI. 10.1137/060655407

1. Introduction.

In recent years, a new multigrid method has been proposed for solving the Navier–Stokes equations based on the iterative solution of several problems over small overlapping domains; for examples, see [12, 13, 14, 24, 25]. This smoothing procedure can be considered a block Gauss–Seidel algorithm whose iteration step consists of solving local problems for each block of unknowns involving pressure and velocity over small subdomains. In this multigrid smoothing step the solution must be computed and updated subdomain by subdomain as in a multiplicative Schwarz-type iterative algorithm. This multigrid technique shows excellent convergence and is naturally suitable for domain decomposition problems and parallel computing [2, 4, 5, 12, 13, 14, 24]. It is not restricted to the Navier–Stokes equations since it can be easily extended to elliptic operators and mixed variational problems if the appropriate local problems for the smoothing procedure are identified. The finite element method introduces a natural decomposition of the problem in subdomains which turns out to be ideal for block subdivisions. For particular blocks of elements some Vanka-type smoothers are well known to be computationally among the best laminar solvers of the stationary Stokes problem [12, 24].

Despite these excellent properties very little is known so far about convergence and smoothing properties of this Gauss–Seidel block iterative method. There is substantial literature on Schwarz alternating methods for the incompressible Navier–Stokes equations for domain decomposition (see, for example, [6, 9, 10, 15, 16, 17, 22, 26] and the references therein), but nothing at all on Vanka-type smoothers. An attempt to investigate this method can be found in [19, 21]. In the first paper a Fourier analysis for a simple model for the Poisson equation in one dimension is presented. In [21] an additive Schwarz-type version of the iterative algorithm for the Stokes problem is presented and transformed into an inexact Uzawa methods under suitable condi-

*Received by the editors September 30, 2004; accepted for publication (in revised form) April 17, 2006; published electronically October 20, 2006.
http://www.siam.org/journals/sinum/44-5/65540.html
†Department of Energy, Nuclear and Environmental Control Engineering, University of Bologna, Via dei Colli 16, 40136 Bologna, Italy (sandro.manservisi@mail.ing.unibo.it), and Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409-1042.

tions. These conditions cannot in general be satisfied and the additive version of the algorithm does not match the real multiplicative Schwarz-type nature of the iterative method.

The aim of this paper is to prove that, for a particular choice of the block, this method leads to a monotonic convergent algorithm for the steady Stokes problem and that the same algorithm applied to the nonlinear Navier–Stokes equations converges for relative small Reynolds numbers. In this paper we consider conforming finite elements and construct the unknown block directly from the finite element discretization. We investigate the convergence for two Vanka-type smoothers which we call type A and type B blocks. In type A we consider finite element blocks of unknowns whose solution leads to divergence-free approximate solutions, i.e., all the approximates satisfy the global divergence-free constraint. In type B we consider finite element blocks of unknowns whose solution does not satisfy the global divergence-free constraint but only a local one. In type B the global constraint is satisfied only when convergence is reached. We will prove that this local problem for the Stokes system can be obtained as a minimization problem of a suitable functional of the residuals. This leads to a monotonic convergence of the residual norm and the smoothing property for the multigrid algorithm. The same algorithm for the steady Navier–Stokes system cannot converge unconditionally but will be proved to be convergent for small Reynolds numbers. This iterative algorithm has been applied successfully to domain decomposition and optimal flow control problems [2, 3, 18, 4, 5]. We leave the numerical analysis of these problems to future papers.

The paper is organized as follows. In section 2 we introduce the variational problem and the continuous domain decomposition problem. In section 3, we give a precise definition of the discrete finite element problem with Vanka-type smoothers. We prove convergence and some properties. Issues related to the numerical implementation of the fully discrete algorithms and some computational experiments are discussed in section 4.

2. Formulation of the variational problem.

2.1. Notation. We introduce the following standard notation over a bounded, connected, open set Ω with polygonal boundary Γ. We shall use the standard notation for the vector-valued Sobolev spaces $H^s(\Omega)$ with its norm $\|\cdot\|_s$ ($H^0(\Omega) = L^2(\Omega)$ and $\|\cdot\|_0 = \|\cdot\|$). Let $H_0^1(\Omega)$ denote the closure of $C_0^\infty(\Omega)$ under the norm $\|\cdot\|_1$ and let $H^{-1}(\Omega)$ be the dual space of $H_0^1(\Omega)$. Also, we define

$$L_0^2(\Omega) = \left\{ p \in L^2(\Omega) : \int_\Omega p \, d\vec{x} = 0 \right\}.$$

In the remainder of the paper $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ and $(\cdot, \cdot)_0$ the usual scalar product in $L^2(\Omega)$. The scalar product in $H_0^1(\Omega)$ is denoted by $(\cdot, \cdot)_1$, which defines the seminorm $|\cdot|_1$ in the same space. Let $\mathbf{V}_0(\Omega)$ be the space of divergence-free vectors that vanish on the boundary. We define $\mathbf{V}_0^\perp(\Omega)$, the orthogonal complement of $\mathbf{V}_0(\Omega)$ in $\mathbf{H}_0(\Omega)$ with respect to the scalar product $(\cdot, \cdot)_1$, by

$$\mathbf{V}_0^\perp(\Omega) = \{ \vec{u} \in \mathbf{H}_0^1(\Omega) : (\vec{u}, \vec{v})_1 = 0 \ \forall \, \vec{v} \in \mathbf{V}_0(\Omega) \},$$

which can also be identified with the following set [11]:

$$\{ \vec{v} \in \mathbf{H}_0^1(\Omega) : \vec{v} = -A^{-1} B^* p; p \in L_0^2(\Omega) \}, \tag{2.1}$$

where A^{-1} is the Green's function which solves the Laplacian problem with homogeneous boundary conditions; see (2.5) for the definition of B. For details concerning these spaces, see [1, 11, 23]. In order to define the weak form of the Stokes and Navier–Stokes equations, we introduce two continuous bilinear forms,

$$a(\vec{u},\vec{v}) = 2\nu \sum_{i,j=1}^{n} \int_\Omega D_{ij}(\vec{u}) D_{ij}(\vec{v})\, d\vec{x} \qquad \forall\, \vec{u},\vec{v} \in H^1(\Omega),$$

$$b(\vec{v},q) = -\int_\Omega q \vec{\nabla}\cdot\vec{v}\, d\vec{x} \qquad \forall q \in L^2_0, \quad \forall\, \vec{v} \in H^1(\Omega),$$

and the trilinear form,

$$c(\vec{w};\vec{u},\vec{v}) = \sum_{i,j=1}^{n} \int_\Omega w_j \left(\frac{\partial u_i}{\partial x_j}\right) v_i\, d\vec{x} \qquad \forall\, \vec{w},\vec{u},\vec{v} \in H^1(\Omega),$$

where $D(\vec{u}) = \frac{1}{2}(\nabla\vec{u} + \nabla\vec{u}^T)$. It is well known that (see, e.g., [11, 18, 23])

$$\begin{aligned} &c(\vec{u};\vec{v},\vec{w}) = -c(\vec{u};\vec{w},\vec{v}) \qquad \forall\, \vec{u} \in V(\Omega), \quad \forall\, \vec{v},\vec{w} \in H^1(\Omega),\\ &c(\vec{u};\vec{v},\vec{v}) = 0 \quad \forall\, \vec{u} \in V(\Omega), \qquad \forall\, \vec{v} \in H^1(\Omega),\\ &|c(\vec{u};\vec{v},\vec{w})| \le C|\vec{u}|_1\,|\vec{v}|_1\,|\vec{w}|_1 \qquad \forall\, \vec{u},\vec{v},\vec{w} \in H^1_0(\Omega), \end{aligned} \tag{2.2}$$

where C is independent of the functions $\vec{u}, \vec{w}$, and $\vec{v}$. For details concerning notation employed and properties of the forms, one may consult, e.g., [11, 23].

We will also make use the following operators:

$$A : H^1(\Omega) \to H^{-1}(\Omega)$$
$$\langle A\vec{u},\vec{v}\rangle = a(\vec{u},\vec{v}) \qquad \forall\, \vec{u} \in H^1(\Omega), \quad \forall\, \vec{v} \in H^1_0(\Omega), \tag{2.3}$$

$$C : H^1(\Omega) \times H^1(\Omega) \to H^{-1}(\Omega)$$
$$\langle C(\vec{w})\vec{u},\vec{v}\rangle = c(\vec{w};\vec{u},\vec{v}) \qquad \forall\, \vec{w},\vec{u} \in H^1(\Omega), \quad \forall\, \vec{v} \in H^1_0(\Omega), \tag{2.4}$$

$$B : H^1(\Omega) \to L^2_0(\Omega)$$
$$\langle B\vec{u}, p\rangle = b(\vec{u},p) \qquad \forall\, p \in L^2_0(\Omega), \quad \forall\, \vec{u} \in H^1(\Omega), \tag{2.5}$$

$$B^* : L^2_0(\Omega) \to H^{-1}(\Omega)$$
$$\langle \vec{u}, B^*p\rangle = b(\vec{u},p) \qquad \forall\, p \in L^2_0(\Omega), \quad \forall\, \vec{u} \in H^1_0(\Omega). \tag{2.6}$$

With this notation we can introduce the Stokes problem from different points of view. In the rest of the paper we assume homogeneous Dirichlet boundary conditions, but a generalization to nonhomogeneous Dirichlet and Neumann boundary conditions can be extended in a straightforward manner.

THEOREM 2.1. *Let Ω be a Lipschitz open bounded domain and $\vec{f} \in \mathbf{H}^{-1}(\Omega)$. These formulations of the Stokes problems are equivalent:*

(a) *Find $(\vec{u},p) \in \mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$, solution of the system*

$$\begin{cases} \nu a(\vec{u},\vec{v}) + b(\vec{v},p) = \langle \vec{f},\vec{v}\rangle \qquad \forall\, \vec{v} \in \mathbf{H}^1_0(\Omega),\\ b(\vec{u},q) = 0 \qquad \forall\, q \in L^2_0(\Omega). \end{cases} \tag{2.7}$$

(b) *Find $(\vec{u},p) \in \mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$, which solves the saddle point problem*

$$\min_{\widetilde{u} \in \mathbf{H}^1_0(\Omega)} \Big(\sup_{\widetilde{p} \in L^2_0(\Omega)} L(\widetilde{u},\widetilde{p}) \Big), \tag{2.8}$$

where

$$L(\widetilde{u},\widetilde{p}) = \frac{1}{2}\nu a(\widetilde{u},\widetilde{u}) + b(\widetilde{u},\widetilde{p}) - \langle \vec{f},\widetilde{u}\rangle. \tag{2.9}$$

(c) *Find* $\vec{u} \in \mathbf{V}_0(\Omega)$*, which solves*

$$\nu a(\vec{u},\vec{v}) = \langle \vec{f},\vec{v}\rangle \qquad \forall\, \vec{v} \in \mathbf{V}_0(\Omega). \tag{2.10}$$

(d) *Find* $\vec{u} \in \mathbf{V}_0(\Omega)$*, which minimizes the functional*

$$L(\widetilde{u}) = \frac{1}{2}\nu a(\widetilde{u},\widetilde{u}) - \langle \vec{f},\widetilde{u}\rangle \tag{2.11}$$

for all $\widetilde{u} \in \mathbf{V}_0^1(\Omega)$.

(e) *Find* $\vec{u}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega)$*, which solves*

$$\nu a(\vec{u}^{\perp},\vec{v}^{\perp}) + \langle \vec{f},\vec{v}^{\perp}\rangle = 0 \qquad \forall\, \vec{v}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega). \tag{2.12}$$

(f) *Find* $\vec{u}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega)$*, which minimizes the functional*

$$L(\widetilde{u}^{\perp}) = \frac{1}{2}\nu a(\widetilde{u}^{\perp},\widetilde{u}^{\perp}) + \langle \vec{f},\widetilde{u}^{\perp}\rangle \tag{2.13}$$

for all $\widetilde{u}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega)$.

Proof. The proof can be found in [11]. □

The formulations in (2.12) and (2.13) take into account the decomposition $\mathbf{V}_0(\Omega)+\mathbf{V}^{\perp}(\Omega)$ of the space $\mathbf{H}_0^1(\Omega)$. We also remark that the formulation in (2.12) is the Schur complement formulation when the equation is written in pressure terms. The equations in (2.12) and (2.10) can be solved separately and yield the uncoupled form of the Stokes problem. However, this uncoupled form is not very useful from the numerical point of view since the construction of test functions in $\mathbf{V}_0^{\perp}(\Omega)$ involves the construction of a scalar function as in the original formulation.

Now we can introduce the Navier–Stokes problem. Let $\vec{f} \in H^{-1}(\Omega)$ denote the steady distributed force and $\vec{u} \in H_0^1(\Omega)$ and $p \in L_0^2(\Omega)$ the state variables, i.e., the velocity and pressure fields, respectively. The state variables are constrained to satisfy the weak form of the following Navier–Stokes equations:

$$\begin{cases} a(\vec{u},\vec{v}) + c(\vec{u};\vec{u},\vec{v}) + b(\vec{u},p) = \langle \vec{f},\vec{v}\rangle \qquad \forall\, \vec{v} \in H_0^1(\Omega), \\ b(\vec{u},p) = 0 \qquad \forall\, q \in L_0^2(\Omega), \end{cases} \tag{2.14}$$

with Dirichlet boundary conditions $\vec{u} = 0$ over Γ.

Existence and uniqueness results for solutions of the system (2.14) are contained in the following theorem; see, e.g., [23].

THEOREM 2.2. *Let* Ω *be an open, bounded set with polygonal boundary* Γ*. Let* $\vec{f} \in \mathbf{H}^{-1}(\Omega)$*. Then*

(i) *there exists at least one solution* $(\vec{u},p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ *of* (2.14)*;*

(ii) *the set of velocity fields that are solutions of* (2.14) *is closed in* $\mathbf{H}_0^1(\Omega)$ *and is compact in* $\mathbf{L}_0^2(\Omega)$*;*

(iii) *if* $\nu > \nu_0(\Omega,\vec{f})$ *for some positive* ν_0 *whose value is determined by the given data, then the set of solutions of* (2.14) *consists of a single element.*

Note that solutions of (2.14) exist for any value of the Reynolds number. However, (iii) implies that uniqueness can be guaranteed only for "large enough" values of ν or for "small enough" data $\vec{f}$.

2.2. Vanka-type smoothers and domain decomposition. Let Ω be an open, bounded, and simply connected domain with polygonal boundary such that $\Omega = \cup_i^m \Omega^i$, where the subdomains Ω^i have smooth boundary Γ^i and are overlapping in the sense that $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2) + \cdots + \mathbf{H}_0^1(\Omega^m)$. The idea behind the Vanka-type smoother is that it is possible to solve the problem over the domain Ω simply by updating and solving a sequence of local problems over all the overlapping subdomains. In the discrete case the subdomain is taken as a small block of finite elements, but at the continuous level this method can be seen as a domain decomposition method over a countable number of overlapping smooth subdomains. We will show that in both the discrete and the continuous cases convergence can be proved. In the rest of the paper we use "hat" notation to denote functions defined over the whole domain Ω and standard notation for local variables defined over the subregions Ω^i.

For example, we can define the local problem i over the domain Ω^i in the following way. Given the velocity field $\widehat{u}^{\,i-1} \in \mathbf{V}_0(\Omega)$ at step $i-1$, we define the local problems for the state $(\vec{u}^{\,i}, p^{\,i}) \in \mathbf{H}_0^1(\Omega^i) \times L_0^2(\Omega^i)$ over the subregion Ω^i by

$$\begin{cases} \nu a(\vec{u}^{\,i}, \vec{v}^{\,i}) + c(\vec{u}^{\,i}, \vec{u}^{\,i}, \vec{v}^{\,i}) + b(p^{\,i}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in H_0^1(\Omega^i), \\ b(r^i, \vec{v}^{\,i}) = 0 \qquad \forall r^i \in L_0^2(\Omega^i), \end{cases} \tag{2.15}$$

with boundary condition $\vec{u}^{\,i} = \widehat{u}^{\,i-1}$ over Γ^i. At each iteration the global solution $(\widehat{u}^{\,i}, \widehat{p}^{\,i})$ is determined by the local solution as $\widehat{u}^{\,i} = \vec{u}^{\,i}$, $\widehat{p}^{\,i} = p^{\,i}$ over Ω^i and $\widehat{u}^{\,i} = \widehat{u}^{\,i-1}$, $\widehat{p}^{\,i} = \widehat{p}^{\,i-1}$ over $\Omega - \Omega^i$. We remark that the local pressure $p^{\,i}$ is the solution of the system (2.15) and a constant must be determined such that the global solution $\widehat{p}^{\,i}$ is in $L_0^2(\Omega)$.

The problem over this subdomain Ω^i can be rewritten in a more appropriate way. As in the standard Schwarz alternating method we can define $\vec{w}^{\,i}$ by $\vec{u}^{\,i} - \widehat{u}^{\,i-1}$ over Ω^i and zero over $\Omega - \Omega^i$ and solve the residual equation instead of the original problem (2.15). The function $\vec{w}^{\,i}$ is in $\mathbf{V}_0(\Omega^i)$ and the residual equation takes the form

$$\begin{cases} \nu a(\vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{v}^{\,i}) + c(\vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i), \\ b(r^i, \vec{w}^{\,i}) = 0 \qquad \forall r^i \in L_0^2(\Omega^i), \end{cases}$$

with boundary condition $\vec{w}^{\,i} = 0$ over Γ^i. Now the solution of the Navier–Stokes equations over different domains is in the standard space $\mathbf{H}_0^1(\Omega^i)$ and the old velocity field $\widehat{u}^{\,i-1}$ over Ω^i can be seen as the projection from the space $\mathbf{V}_0(\Omega)$ to the space $\mathbf{V}_0(\Omega^i)$ with respect to the scalar product $(\cdot,\cdot)_1$. It is clear that the convergence of the Vanka-type smoother is related to the properties of these projection operators, and the starting point is to investigate the decomposition of the global solution into the sum of many local solutions. We start to state a space decomposition theorem which was partially shown in [15, 16, 17].

THEOREM 2.3. *Let* $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2) + \mathbf{H}_0^1(\Omega^3) \cdots \mathbf{H}_0^1(\Omega^m)$ *and* $\mathbf{H}_0^1(\Omega^i) = \mathbf{V}_0^{\perp}(\Omega^i) + \mathbf{V}_0(\Omega^i)$ *for all* $i = 1, \ldots, m$, *where* $\mathbf{V}_0(\Omega^i)$ *and* $\mathbf{V}_0^{\perp}(\Omega^i)$ *are the divergence-free function space and its orthogonal complement in* $\mathbf{H}_0^1(\Omega^i)$ *with respect to the inner product* $(\cdot,\cdot)_1$ *over* Ω^i. *Then we have*

(a) $\mathbf{V}_0(\Omega) = \mathbf{V}_0(\Omega^1) + \mathbf{V}_0(\Omega^2) + \mathbf{V}_0(\Omega^3) \cdots \mathbf{V}_0(\Omega^m)$;

(b) $\mathbf{V}_0^{\perp}(\Omega) = \mathbf{V}_0^{\perp}(\Omega^1) + \mathbf{V}_0^{\perp}(\Omega^2) + \mathbf{V}_0^{\perp}(\Omega^3) \cdots \mathbf{V}_0^{\perp}(\Omega^m)$.

Proof. (a) In [15] it is proved that if $\Omega = \Omega^1 \cup \Omega^2$ and $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2)$, then $\mathbf{V}_0(\Omega) = \mathbf{V}_0(\Omega^1) + \mathbf{V}_0(\Omega^2)$. In the case of m domains we can write

$\Omega^* = \Omega^1 \cup \Omega^2 \cup \Omega^3 \cup \cdots \cup \Omega^{m-1}$ and prove the assertion by induction. By induction hypothesis $\mathbf{H}_0^1(\Omega^*) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2) + \mathbf{H}_0^1(\Omega^3) \cdots \mathbf{H}_0^1(\Omega^{m-1})$ and $\mathbf{V}_0(\Omega^*) = \mathbf{V}_0(\Omega^1) + \mathbf{V}_0(\Omega^2) + \mathbf{V}_0(\Omega^3) \cdots \mathbf{V}_0(\Omega^{m-1})$. Combining this and the result in [15] we have $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^*) + \mathbf{H}_0^1(\Omega^m)$ and $\mathbf{V}_0(\Omega) = \mathbf{V}_0(\Omega^*) + \mathbf{V}_0(\Omega^m)$.

(b) First consider the case for $m = 2$. Let $\vec{u}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega_h)$. Since the domains are overlapping in the space $\mathbf{H}_0^1(\Omega)$ we can write $\vec{u}^{\perp} = \vec{u}_1 + \vec{u}_2$ with $\vec{u}_i \in \mathbf{H}_0^1(\Omega^i)$, $i = 1, 2$. The vector $\vec{u}_1$ can be decomposed in $\mathbf{H}_0^1(\Omega^1)$ as $\vec{u}_1 = \vec{u}_1^{\perp} + \vec{u}_1^{o}$ with $\vec{u}_1^{\perp} \in \mathbf{V}_0^{\perp}(\Omega^1)$ and $\vec{u}_1^{o} \in \mathbf{V}_0(\Omega^1)$.

Consider the zero extension $\vec{w}_1^{\perp}$ and $\vec{w}_1^{o}$ to Ω of $\vec{u}_1^{\perp}$ and $\vec{u}_1^{o}$. The vectors $\vec{w}_1^{\perp}$ and $\vec{w}_1^{o}$ are in $\mathbf{V}_0^{\perp}(\Omega)$ and $\mathbf{V}_0(\Omega)$, respectively. The extension $\vec{w}_1$ of $\vec{u}_1$ to Ω can be written as $\vec{w}_1 = \vec{w}_1^{\perp} + \vec{w}_1^{o}$.

In a similar way we can decompose $\vec{u}_2 = \vec{u}_2^{\perp} + \vec{u}_2^{o}$ with $\vec{u}_2^{\perp} \in \mathbf{V}_0^{\perp}(\Omega^2)$ and $\vec{u}_1^{o} \in \mathbf{V}_0(\Omega^1)$. The extension $\vec{w}_2$ of $\vec{u}_2$ to Ω can be written as $\vec{w}_2 = \vec{w}_2^{\perp} + \vec{w}_2^{o}$, where $\vec{w}_2^{\perp}$ and $\vec{w}_2^{o}$ are the corresponding extensions. From the hypothesis we have that $\vec{u}^{\perp} = \vec{u}_1 + \vec{u}_2$ and therefore $\vec{u}_2 = \vec{u}^{\perp} - \vec{u}_1$ and $\vec{w}_2 = \vec{u}^{\perp} - \vec{w}_1$. The unique decomposition of $\vec{w}_2$ over the subspaces $\mathbf{V}_0^{\perp}(\Omega)$ and $\mathbf{V}_0(\Omega)$ implies that $\vec{u}_2^{o} = -\vec{u}_1^{o}$ and that $\vec{u}_1^{o} = \vec{w}_1^{o}$ is zero over $\Omega - (\Omega^1 \cap \Omega^2)$. Therefore $\vec{u}_1^{\perp} = \vec{u}_1 - \vec{u}_1^{o} \in \mathbf{V}_0^{h\perp}(\Omega^1)$ and $\vec{u}_2^{\perp} = \vec{u}^{\perp} - \vec{u}_1 + \vec{u}_1^{o} \in \mathbf{V}_0^{\perp}(\Omega^2)$ give the desired decomposition $\vec{u}^{\perp} = \vec{u}_1^{\perp} + \vec{u}_2^{\perp}$. For any m the theorem can be proved by induction using the case $m = 2$ and standard techniques [15, 16, 11]. In the case of m domains we can write $\Omega^* = \Omega^1 \cup \Omega^2 \cup \Omega^3 \cup \cdots \cup \Omega^{m-1}$ and prove the assertion by induction. By induction hypothesis $\mathbf{H}_0^1(\Omega^*) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2) + \mathbf{H}_0^1(\Omega^3) \cdots \mathbf{H}_0^1(\Omega^{m-1})$ and $\mathbf{V}_0^{\perp}(\Omega^*) = \mathbf{V}_0^{\perp}(\Omega^1) + \mathbf{V}_0^{\perp}(\Omega^2) + \mathbf{V}_0^{\perp}(\Omega^3) \cdots \mathbf{V}_0^{\perp}(\Omega^{m-1})$. Combining this and the above results for $m = 2$ we have $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^*) + \mathbf{H}_0^1(\Omega^m)$ and $\mathbf{V}_0^{\perp}(\Omega) = \mathbf{V}_0^{\perp}(\Omega^*) + \mathbf{V}_0^{\perp}(\Omega^m)$. □

Let T_i, P_i, and Π_i denote the orthogonal projections from $\mathbf{H}_0^1(\Omega), \mathbf{V}_0(\Omega), \mathbf{V}_0^{\perp}(\Omega)$ onto $\mathbf{H}_0^1(\Omega^i), \mathbf{V}_0^i(\Omega^i)$, and $\mathbf{V}_0^{\perp i}(\Omega^i)$ with respect to the inner product $(\cdot, \cdot)_1$ defined by the operator $a(\vec{u}, \vec{v})$ for all $\vec{u}$ and $\vec{v}$ in $\mathbf{H}_0^1(\Omega)$. The projections $T_i\,\widehat{u}, P_i\,\widehat{u}, \Pi_i\,\widehat{u}$ of $\widehat{u}$ from $\mathbf{H}_0^1(\Omega), \mathbf{V}_0(\Omega), \mathbf{V}_0^{\perp}(\Omega)$ onto $\mathbf{H}_0^1(\Omega^i), \mathbf{V}_0(\Omega^i), \mathbf{V}_0^{\perp}(\Omega^i)$ are defined by the solutions of the equations

$$
\begin{aligned}
&a(T_i\widehat{u}, \vec{v}^{\,i}) = a(\widehat{u}, \vec{v}^{\,i}) \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i),\\
&a(P_i\widehat{u}, \vec{v}^{\,i}) = a(\widehat{u}, \vec{v}^{\,i}) \qquad \forall \vec{v}^{\,i} \in \mathbf{V}_0(\Omega^i),\\
&a(\Pi_i\widehat{u}, \vec{v}^{\,\perp i}) = a(\widehat{u}, \vec{v}^{\,\perp i}) \qquad \forall \vec{v}^{\,\perp i} \in \mathbf{V}_0^{\perp i}(\Omega^i),
\end{aligned}
$$

respectively. We note that $\Pi_i\widehat{u}$ is in general different from $\widehat{u}$ due to the homogeneous boundary conditions on the boundary Γ^i. The orthogonal projections T_i, P_i, and Π_i have the following properties [15, 6].

THEOREM 2.4. *Let Ω be a bounded simply connected domain with smooth boundary and let Ω^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains with smooth boundary such that $\Omega = \Omega^1 \cup \Omega^2 \cdots \Omega^m$. Let T_i (P_i or Π_i) be the orthogonal projection from $\mathbf{H}_0^1(\Omega)$ ($\mathbf{V}_0(\Omega)$ or $\mathbf{V}_0^{\perp}(\Omega)$) onto $\mathbf{H}_0^1(\Omega^i)$ ($\mathbf{V}_0(\Omega^i)$ or $\mathbf{V}_0^{\perp}(\Omega^i)$) for $i = 1, 2, \ldots, m$. Then*

(a) $|I - T_i|_1 \le 1$ ($|I - P_i|_1 \le 1$ *or* $|I - \Pi_i|_1 \le 1$) *for* $i = 1, 2, \ldots, m$;

(b) $|\Pi_{i=1}^m (I - T_i)|_1 < 1$ ($|\Pi_{i=1}^m (I - P_i)|_1 < 1$ *or* $|\Pi_{i=1}^m (I - \Pi_i)|_1 < 1$).

Proof. We use a standard minimization approach. Let $\mathbf{V}(\Omega)$ be a subspace of $\mathbf{H}_0^1(\Omega)$ such that $\mathbf{V}(\Omega) = \mathbf{V}(\Omega^1) + \mathbf{V}(\Omega^2) + \cdots + \mathbf{V}(\Omega^m)$ with $\mathbf{V}(\Omega^i) \subseteq \mathbf{H}_0^1(\Omega^i)$ for $i = 1, 2, \ldots, m$. Here $\mathbf{V}(\mathbf{\Omega})$ could be $\mathbf{H}_0^1(\Omega)$, $\mathbf{V}_0(\Omega)$, or $\mathbf{V}_0^{\perp}(\Omega)$. Consider the Laplace operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ defined by

$$
\langle A\vec{u}, \vec{v} \rangle = (\vec{u}, \vec{v})_1 \qquad \forall\, \vec{u} \in H_0^1(\Omega), \quad \forall\, \vec{v} \in H_0^1(\Omega)
$$

and the following equation for $\vec{u} \in \mathbf{V}(\Omega)$ with homogeneous Dirichlet boundary condition:

$$\langle A\vec{u}, \vec{v}\rangle = \langle \vec{f}, \vec{v}\rangle \qquad \forall \vec{v} \in \mathbf{V}(\Omega). \tag{2.16}$$

We note that the operator A is elliptic, and therefore the natural approach is the minimization of the associated functional

$$L = \left(\frac{1}{2}(A\widetilde{u}, \widetilde{u}) - \langle \vec{f}, \widetilde{u}\rangle + \frac{1}{2}\langle A^{-1}\vec{f}, \vec{f}\rangle\right) \tag{2.17}$$

over the sequence of spaces $\mathbf{V}(\Omega^i)$ for $i = 1, 2, \ldots, m$. The constant $\frac{1}{2}\langle A^{-1}\vec{f}, \vec{f}\rangle$ is used to vanish the functional at the critical point, and the operator A^{-1} from $\mathbf{H}^{-1}(\Omega)$ to $\mathbf{H}_0^1(\Omega)$ denotes the Green's operator of the homogeneous problem with the operator A. We remark that (2.17) can also be written as

$$L = \left(\frac{1}{2}(A\,\widetilde{u}, \widetilde{u}) - \langle A\,\vec{u}, \widetilde{u}\rangle + \frac{1}{2}\langle A\vec{u}, \vec{u}\rangle\right) = \frac{1}{2}(A\,\widetilde{u} - \vec{u}, \widetilde{u} - \vec{u}),$$

where $\vec{u}$ is the unique critical point which solves (2.16). Let $\widehat{u}^{\,i-1}$ be the solution at step $i-1$ over Ω and let R_i be the projector from $\mathbf{V}(\Omega)$ to $\mathbf{V}(\Omega^i)$. The operator $\Pi_{i=1}^m(I - R_i)$ is defined from the solution sequence of the minimization problem over $\mathbf{V}(\Omega^i)$ for $i = 1, 2, \ldots, m$. In fact the minimization over Ω^i gives

$$(A\vec{u}^{\,i}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i}\rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{V}(\Omega^i). \tag{2.18}$$

If we set $\vec{w}^{\,i} = \vec{u}^{\,i} - \widehat{u}^{\,i-1} \in \mathbf{V}(\Omega^i)$, then (2.18) gives

$$(A\vec{w}^{\,i}, \vec{v}^{\,i}) + (A\widehat{u}^{\,i-1}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i}\rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{V}(\Omega^i). \tag{2.19}$$

Let $\vec{u}$ be the solution of (2.16) and let $\vec{e}^{\,i} = \vec{u}^{\,i} - \vec{u}$ over Ω^i and $\widehat{e}^{\,i-1} = \widehat{u}^{\,i-1} - \vec{u}$ over Ω. We have $\vec{w}^{\,i} = \vec{e}^{\,i} - \widehat{e}^{\,i-1}$ and $\vec{w}^{\,i}$ and $\vec{e}^{\,i}$ can be seen as a local error over Ω^i or as a global error over Ω if the corresponding extensions are considered. We use "hat" notation for functions defined over the domain Ω or the corresponding extensions. We remark that $\widehat{w}^{\,i}$ is the zero extension to Ω of $\vec{w}^{\,i}$. With this notation the above equation becomes

$$(A\vec{w}^{\,i}, \vec{v}^{\,i}) + (A\widehat{e}^{\,i-1}, \vec{v}^{\,i}) = 0 \qquad \forall \vec{v}^{\,i} \in \mathbf{V}(\Omega^i),$$

which is solved by

$$\vec{u}^{\,i} = (I - R_i)\,\widehat{u}^{\,i-1} + A^{-1}\vec{f} \quad \text{or} \quad \widehat{e}^{\,i} = (I - R_i)\,\widehat{e}^{\,i-1}. \tag{2.20}$$

Since this estimate is true for all $i = 1, 2, \ldots, m$ we can combine them in the following solution:

$$\widehat{e}^{\,m} = \Pi_{i=1}^m(I - R_i)\,\widehat{e}^{\,0}, \tag{2.21}$$

which gives the global error over Ω needed to estimate the functional L^i and the operator norm over the domain Ω.

Now we show that the functional L^i, which is the norm of the error $\widehat{e}^{\,i}$, is monotonically decreasing for increasing i for all initial guesses $\widehat{e}^{\,0} \in \mathbf{V}(\Omega^i)$. In fact

$$\begin{aligned} L^i &= \frac{1}{2}(A\widehat{u}^{\,i}, \widehat{u}^{\,i}) - \langle \vec{f}, \widehat{u}^{\,i}\rangle + \frac{1}{2}\langle A^{-1}\vec{f}, \vec{f}\rangle \\ &= \frac{1}{2}(A(\vec{w}^{\,i} + \widehat{u}^{\,i-1}), \vec{w}^{\,i} + \widehat{u}^{\,i-1}) - \langle \vec{f}, \vec{w}^{\,i} + \widehat{u}^{\,i-1}\rangle + \frac{1}{2}\langle A^{-1}\vec{f}, \vec{f}\rangle \\ &= L^{i-1} + \frac{1}{2}(A\vec{w}^{\,i}, \vec{w}^{\,i}) + (A\widehat{u}^{\,i-1}, \vec{w}^{\,i}) - \langle \vec{f}, \vec{w}^{\,i}\rangle, \end{aligned}$$

and by using the optimality condition (2.19) we have

$$L^i = L^{i-1} - \frac{1}{2}(A\vec{w}^{\,i}, \vec{w}^{\,i}), \tag{2.22}$$

which implies $L^i \le L^{i-1}$ or $L^i < L^{i-1}$ if we assume $\vec{w} \neq 0$. Now

$$L^i = \frac{1}{2}(\widehat{u}^{\,i} - A^{-1}\vec{f}, \widehat{u}^{\,i} - A^{-1}\vec{f}) = \frac{1}{2}(\widehat{e}^{\,i}, \widehat{e}^{\,i})_1 < L^{i-1} = \frac{1}{2}(\widehat{e}^{\,i-1}, \widehat{e}^{\,i-1})_1,$$

and therefore

$$|\widehat{e}^{\,i}|_1^2 = (\widehat{e}^{\,i}, \widehat{e}^{\,i})_1 \le (\widehat{e}^{\,i-1}, \widehat{e}^{\,i-1})_1 = |\widehat{e}^{\,i-1}|_1^2. \tag{2.23}$$

Since $|\widehat{e}^{\,i}|_1^2 \le |\widehat{e}^{\,i-1}|_1^2$ for all initial vectors $\widehat{e}^{\,i-1} \in \mathbf{V}(\Omega)$ this implies that $|(I - R_i)|_1 \le 1$ for all $i = 1, \ldots, m$, which proves (a).

(b) We use the same notation as in (a). By starting with an initial guess $\widehat{u}^{\,0} \in \mathbf{V}(\Omega)$ after a smoothing step over m domains, from (2.22), we have

$$L^m = L^0 - \frac{1}{2}\sum_{i=1}^{m} |\widehat{w}^{\,i}|_1^2,$$

with $\widehat{w}^{\,i} = -R_i\widehat{e}^{\,i-1}$, $L^m = |\widehat{e}^{\,m}|_1^2/2$, and $L^0 = |\widehat{e}^{\,0}|_1^2/2$. Also we have $\widehat{e}^{\,m} = \Pi_{i=1}^m (I - R_i)\,\widehat{e}^{\,0}$ for all $\widehat{e}^{\,0} \in \mathbf{V}(\Omega)$. From the definition of norm we have

$$|\Pi_{i=1}^m (I - R_i)|_1 = \sup |\widehat{e}^{\,m}|_1 = \sqrt{1 - \inf \sum_{i=1}^{m} |\widehat{w}^{\,i}|_1^2}, \tag{2.24}$$

where the sup and inf are taken over the set of functions $\widehat{e}^{\,0} \in \mathbf{V}(\Omega)$ with $|\widehat{e}^{\,0}|_1 = 1$. The infimum is the minimizer of the minimization problem

$$\min \sum_{i=1}^{m} |\widehat{w}^{\,i}|_1^2,$$

where the minimum is taken over all the functions $\widehat{e}^{\,0}$ in $\mathbf{V}(\Omega)$ with $|\widehat{e}^{\,0}|_1 = 1$. We remark that $\sum_{i=1}^m |\widehat{w}^{\,i}|_1^2 = |\widehat{e}^{\,m} - \widehat{e}^{\,0}|_1^2$. Standard theory can be applied to prove existence, well-posedness, and uniqueness of the solution of the problem over the set $\mathbf{V}(\Omega)$. Now by contradiction we prove that the minimum cannot be zero and the corresponding norm $|\Pi_{i=1}^m (I - R_i)|_1 < 1$. Suppose that the minimum is zero; then this implies, from the definition of the norm itself, that there exists an $\widehat{e}^{\,0}$ in $\mathbf{V}(\Omega)$ such that $\widehat{e}^{\,m} = \widehat{e}^{\,0}$, and therefore $\widehat{w}^{\,i} = -R_i\widehat{e}^{\,0} = 0$ for all $i = 1, 2, \ldots, m$. Such a function $\widehat{e}^{\,0}$ has norm 1 but zero projections over all $\mathbf{V}(\Omega^i)$, $i = 1, 2, \ldots, m$, and this contradicts the fact that $\mathbf{V}(\Omega) = \mathbf{V}(\Omega^1) + \mathbf{V}(\Omega^2) + \cdots + \mathbf{V}(\Omega^m)$; i.e., the domains are overlapping in the space $\mathbf{V}(\Omega)$. Therefore the functional must be different from zero and must assume a well-defined positive value at the critical point. Combining this result with (3.15), we obtain the inequality $|\Pi_{i=1}^m (I - R_i)|_1 < 1$. ∎

2.3. Domain decomposition of type A. The projections introduced above can be used to solve the Stokes and the Navier–Stokes equations with different types of domain decompositions. There are several different possible domain decompositions, but we are interested only in the discussion of domain decompositions whose discretization leads to the Vanka-type smoothers with elements of our interest.

Consider the following domain decomposition algorithm, which we can call type A domain decomposition since all the approximates satisfy globally the divergence-free constraint. Let Ω be an open, bounded, and simply connected domain with polygonal boundary. Let $\Omega = \cup_i^m \Omega^i$, where the subdomains Ω^i have smooth boundary Γ^i and are overlapping in the sense that $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0^1(\Omega^1) + \mathbf{H}_0^1(\Omega^2) + \cdots + \mathbf{H}_0^1(\Omega^m)$. Given the velocity field $\widehat{u}^{\,i-1} \in \mathbf{V}_0(\Omega)$ at step $i-1$ we define the local problem over the domain Ω^i for the unknown $\vec{w}^{\,i} = \vec{u}^{\,i} - \widehat{u}^{\,i-1} \in \mathbf{V}_0(\Omega^i)$ by

$$a(\vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{v}^{\,i}) + c(\vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{w}^{\,i} + \widehat{u}^{\,i-1}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{V}_0(\Omega^i),$$

with boundary condition $\vec{w}^{\,i} = 0$. We remark that $\vec{w}^{\,i}$ is solved over Ω^i and can be extended to Ω such that $\widehat{u}^{\,i} = \widehat{w}^{\,i} + \widehat{u}^{\,i-1}$ is defined over the whole domain Ω. Also we note that the vector solution $\vec{w}^{\,i}$ is in the standard space $\mathbf{V}_0(\Omega^i)$, but the old velocity field $\widehat{u}^{\,i-1}$ is in $\mathbf{V}_0(\Omega^i)$, and therefore it can be seen as a projection from the space $\mathbf{V}_0(\Omega)$ to the space $\mathbf{V}_0(\Omega^i)$ with respect to the scalar product $(\cdot,\cdot)_1$. At the ith iteration the global solution $\widehat{u}^{\,i}$ is determined by the local solution as $\widehat{u}^{\,i} = \vec{u}^{\,i}$ over the subdomain Ω^i and $\widehat{u}^{\,i} = \widehat{u}^{\,i-1}$ over $\Omega - \Omega^i$. We remark that if $\widehat{u}^{\,0}$ is in $\mathbf{V}_0(\Omega)$, then the sequence $\widehat{u}^{\,i}$ is in $\mathbf{V}_0(\Omega)$ for all $i = 1, 2, \ldots, m$.

This domain decomposition algorithm can be proved to be convergent for the Stokes problem.

Details on the Stokes problem for finite element discretizations are reported in the next sections. The Navier–Stokes case requires more restrictions since the uniqueness of the solution can be guaranteed only for small Reynolds numbers, as stated in Theorem 2.2. Therefore we state a convergence theorem only for small Reynolds numbers [16, 17].

THEOREM 2.5. *Let Ω be a bounded simply connected domain with smooth boundary and let Ω^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains with smooth boundary such that $\Omega = \Omega^1 \cup \Omega^2 \cup \cdots \cup \Omega^m$. Let the jth smoothing step $\widehat{u}^{\,j,m} \in \mathbf{V}_0(\Omega)$ be defined by solving iteratively the local system over the domain Ω^i for $i = 1, \ldots, m$ with $\widehat{u}^{\,j,0} = \widehat{u}^{\,j-1,m} \in \mathbf{V}_0(\Omega)$. The global sequence is defined by $\widehat{u}^{\,j,i} = \vec{w}^{\,i} + \widehat{u}^{\,j,i-1}$ over the subdomain Ω^i and $\widehat{u}^{\,j,i} = \widehat{u}^{\,j,i-1}$ over $\Omega - \Omega^i$ $(i = 1, \ldots, m)$, where $\vec{w}^{\,i} \in \mathbf{V}_0(\Omega^i)$ solves the local problem*

(2.25)
$$\nu a(\vec{w}^{\,i} + \widehat{u}^{\,j,i-1}, \vec{v}^{\,i}) + c(\vec{w}^{\,i} + \widehat{u}^{\,j,i-1}, \vec{w}^{\,i} + \widehat{u}^{\,j,i-1}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{V}_0(\Omega^i),$$

with boundary condition $\vec{w}^{\,i} = 0$.

Let $\widehat{u}^{\,0,0} \in \mathbf{V}_0(\Omega)$. Then the global sequence $\widehat{u}^{\,j,m} \in \mathbf{V}_0(\Omega)$ converges $(j \to \infty)$ for sufficiently small Reynolds number to the solution of the Navier–Stokes problem in (2.14).

Proof. At the smoothing step j over the subdomain Ω^i in (2.25) we have

$$\nu a(\vec{w}^{\,i}, \vec{v}^{\,i}) + c(\vec{u}^{\,i}, \vec{u}^{\,i}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle - \nu a(\widehat{u}^{\,j,i-1}, \vec{v}^{\,i}), \tag{2.26}$$

where $\vec{u}^{\,i} = \vec{w}^{\,i} + \widehat{u}^{\,j,i-1}$ is the local unknown velocity field defined over Ω^i. Let $\vec{u} \in \mathbf{V}_0(\Omega)$ be the solution of the Navier–Stokes problem in (2.14), i.e., $\vec{u}$ is the solution of the equation

$$\nu a(\vec{u}^{\,j}, \vec{v}^{\,i}) + \Delta t\, c(\vec{u}^{\,j}, \vec{u}^{\,j}, \vec{v}^{\,i}) = (\vec{u}^{\,j-1}, \vec{v}^{\,i}) + \Delta t\, \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{V}_0(\Omega^i),$$

with the appropriate boundary conditions over Γ^i. The appropriate boundary conditions are Dirichlet boundary conditions that match the global solution $\vec{u}$ itself over Γ_i.

We note that $\vec{w}^{\,i}$ is in $\mathbf{V}_0(\Omega^i)$ but that $\widehat{u}^{\,j,i-1}$ and $\vec{u}^{\,i}$ do not vanish on the boundary Γ^i, and therefore they are not in $\mathbf{V}_0(\Omega^i)$. Equation (2.26) becomes

$$\begin{aligned} &\nu a(\vec{w}^{\,i} + P_i\,(\widehat{u}^{\,j,i-1} - \vec{u}), \vec{v}^{\,i}) + c(\vec{u}^{\,i} - \vec{u}, \vec{u}^{\,i} - \vec{u}, \vec{v}^{\,i}) \\ &+ c(\vec{u}, \vec{u}^{\,i} - \vec{u}, \vec{v}^{\,i}) + c(\vec{u}^{\,i} - \vec{u}, \vec{u}, \vec{v}^{\,i}) = 0, \end{aligned} \tag{2.27}$$

where P_i is the projector operator from $\mathbf{V}_0(\Omega)$ to $\mathbf{V}_0(\Omega^i)$. Let $\vec{e}^{\,j,i} = \vec{u}^{\,i} - \vec{u}$ and $\vec{e}^{\,j,i-1} = \widehat{u}^{\,j,i-1} - \vec{u}$ be the errors at smoothing steps i and $i-1$, respectively. The error functions $\vec{e}^{\,j,i}$ and $\vec{e}^{\,j,i-1}$ defined over Ω^i can be extended over Ω naturally and will be denoted by $\widehat{e}^{\,j,i}$ and $\widehat{e}^{\,j,i-1}$, respectively. Then (2.27) can be written in operator form as

$$A(\vec{w}^{\,i} + P_i\,\widehat{e}^{\,j,i-1}) + \frac{1}{\nu}\Big(C(\vec{e}^{\,j,i})\vec{e}^{\,j,i} + C(\vec{u})\,\vec{e}^{\,j,i} + C(\vec{e}^{\,j,i})\vec{u}\Big) = 0 \quad \text{on } \Omega^i \tag{2.28}$$

or

$$\vec{w}^{\,i} + P_i\,\widehat{e}^{\,j,i-1} = -\frac{1}{\nu}\,A_i^{-1}\Big(C(\vec{e}^{\,j,i})\vec{e}^{\,j,i} + C(\vec{u})\,\vec{e}^{\,j,i} + C(\vec{e}^{\,j,i})\vec{u}\Big) \text{ over } \Omega^i, \tag{2.29}$$

where A_i^{-1} is the inverse operator of A over the domain Ω^i with homogeneous boundary condition on Γ^i. If we define $\widehat{w}^{\,j,i}$ as the zero extension over Ω of $\vec{w}^{\,i}$, then (2.29) gives a global estimate in $\widehat{e}^{\,j,i}$ which satisfies

$$\widehat{e}^{\,j,i} = (I - P_i)\,\widehat{e}^{\,j,i-1} - \frac{1}{\nu}\,A_i^{-1}\Big(C(\vec{e}^{\,j,i})\vec{e}^{\,j,i} - C(\vec{u})\,\vec{e}^{\,j,i} - C(\vec{e}^{\,j,i})\vec{u}\Big). \tag{2.30}$$

Now we can prove that the error $\widehat{e}^{\,j,i}$ is bounded for large values of ν. By using the Schwarz inequality and (2.2) in (2.30) we have

$$|\widehat{e}^{\,j,i}|_1 \le |(I - P_i)\,\widehat{e}^{\,j,i-1}|_1 + \frac{1}{\nu}\,K\,(|\widehat{e}^{\,j,i}|_1^2 + |\vec{u}|_1\,|\widehat{e}^{\,j,i}|_1) \tag{2.31}$$

for some K constant. Let the initial error $|\widehat{e}^{\,j,0}|_1$ be bounded by M and let λ be a positive number such that $\lambda < 1 - |\Pi_{i=1}^m (I - P_i)|_1^{1/m}$. Theorem 2.4 assures us that $\lambda > 0$ since $|\Pi_{i=1}^m (I - P_i)|_1 < 1$. Since $|\vec{u}|_1$ is bounded, then from a simple geometric consideration there exists a ν^* such that if $\nu > \nu^*$, then

$$\lambda x > \frac{K}{\nu}\,(x^2 + x|\vec{u}^{\,j}|_1) \tag{2.32}$$

for $0 \le x \le M/(1-\lambda)^m = M'$. It is easy to see that if $|\widehat{e}^{\,j,0}|_1 \le M$, then $|\widehat{e}^{\,j,i}|_1 \le M'$ for all $i = 1, 2, \ldots, m$. In fact for $i = m$ in (3.27) and $\nu > \nu^*$ we have

$$(1 - \lambda)|\widehat{e}^{\,j,m}|_1 \le |\widehat{e}^{\,j,m-1}|_1,$$

and for $i = m - 1$ and $\nu > \nu^*$ we can write

$$(1 - \lambda)^2|\widehat{e}^{\,j,m}|_1 \le |\,\widehat{e}^{\,j,m-2}|_1$$

since $|(I - P_m)|_1 \le 1$. We can obtain similar estimates for $i = m-2, m-3, \ldots, 1$ and combine them to obtain

$$|\widehat{e}^{\,j,i}|_1 \le \frac{|\,\widehat{e}^{\,j,0}|_1}{(1-\lambda)^i} \le M' \tag{2.33}$$

for all $i = 1, \ldots, m$ and $\nu > \nu^*$. However, this estimate is not sharp enough since M' can be much larger than M and the global sequence in j may diverge. In order to improve the bound we need to take larger values of ν and prove that the sequence $|\widehat{e}^{\,j,m}|$ is monotone decreasing. Consider (3.27) and the fact that $|\widehat{e}^{\,j,m}|_1 \leq |\widehat{e}^{\,j,0}|_1/(1-\lambda)^m$ and $|\widehat{e}^{\,j,0}|_1 \leq M$ for some large $\nu > \nu^*$. We can write, for $i = m$,

$$|\widehat{e}^{\,j,m}|_1 \leq |(I - P_m)\,\widehat{e}^{\,j,m-1}|_1 + \frac{1}{\nu}\,C\,|\widehat{e}^{\,j,0}|_1 \tag{2.34}$$

for some constant C and $\nu > \nu^*$. By substituting (2.30) for $i = m-1$ in (2.34) and using the same estimates, we have

$$|\widehat{e}^{\,j,m}|_1 \leq |(I - P_m)(I - P_{m-1})\,\widehat{e}^{\,j,m-2}|_1 + 2\frac{1}{\nu}\,C\,|\widehat{e}^{\,j,0}|_1\,.$$

Again by using (2.30) for $i = m-2, \ldots, 1$ we finally obtain

$$|\widehat{e}^{\,j,m}|_1 \leq |\Pi_{i=1}^{m}(I - P_i)\,\widehat{e}^{\,j,0}|_1 + m\frac{1}{\nu}\,C\,|\widehat{e}^{\,j,0}|_1\,.$$

We remark that this is not a very sharp bound and can be improved, i.e., for example, the integer number m in the bound can be substituted by the maximum number of overlapping domains. If ν is taken greater than $\nu_1 = (1 - |\Pi_{i=1}^{m}(I - P_i)|_1)\frac{1}{m\,C}$, then

$$|\widehat{e}^{\,j,m}|_1 < |\widehat{e}^{\,j,0}|_1.$$

We can apply this technique to each smoothing step for $j = 1, 2, \ldots$ and claim that, for $\nu > \nu_1 > \nu^*$, the sequence $|\vec{e}^{\,j,m}|_1$ is bounded by the initial error M. Also the sequence $|\widehat{e}^{\,j,m}|_1$ converges to zero and therefore $\widehat{u}^{\,j,m}$ to solution of (2.14). ∎

2.4. Domain decomposition of type B. The domain decomposition algorithm of type A keeps the global solution divergence-free at each step. This could be computationally very expensive since the divergence-free constraint must be satisfied everywhere and at each iteration. This global constraint could be relaxed and imposed only locally if we use a different form of the Navier–Stokes system. Since any vector $\vec{u}^{\,'} \in \mathbf{H}_0^1(\Omega)$ can be written in a unique way as $\vec{u} + \vec{u}^{\,\perp}$ where $\vec{u} \in \mathbf{V}_0(\Omega)$ and $\vec{u}^{\,\perp} \in \mathbf{V}_0^{\perp}(\Omega)$ we can set $\vec{u} = \vec{u}^{\,'} - \vec{u}^{\,\perp}$ in the Navier–Stokes system (2.14) and write

$$\begin{cases} \nu a(\vec{u}^{\,'}, \vec{v}) - a(\vec{u}^{\,\perp}, \vec{v}) + c(\vec{u}, \vec{u}, \vec{v}) + b(p, \vec{v}) = \langle \vec{f}, \vec{v} \rangle \quad \forall \vec{v} \in \mathbf{H}_0^1(\Omega), \\ b(r, \vec{u}^{\,\perp}) = b(r, \vec{u}^{\,'}) \qquad \forall r \in L_0^2(\Omega), \\ \vec{u} = \vec{u}^{\,'} - \vec{u}^{\,\perp}. \end{cases} \tag{2.35}$$

Since the Stokes operator in the continuous form satisfies the LBB condition from the definition of $\mathbf{V}_0^{\perp}(\Omega)$ in (2.1) for each $p \in L_0^2(\Omega)$ there is a unique $\vec{u}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega)$ such that [11]

$$a(\vec{u}^{\,\perp}, \vec{v}) = b(p, \vec{v}) \qquad \forall \vec{v} \in \mathbf{H}_0^1(\Omega).$$

We identify $\vec{u}^{\,\perp}$ with p, and this implies that the system (2.35) can be written as

$$\begin{cases} \nu a(\vec{u}^{\,'}, \vec{v}) + c(\vec{u}, \vec{u}, \vec{v}) = \langle \vec{f}, \vec{v} \rangle \qquad \forall \vec{v} \in \mathbf{H}_0^1(\Omega), \\ a(\vec{v}^{\perp}, \vec{u}^{\,\perp}) = a(\vec{v}^{\perp}, \vec{u}^{\,'}) \qquad \forall \vec{v}^{\perp} \in \mathbf{V}_0^{\perp}(\Omega), \\ \vec{u}^{\,'} = \vec{u} + \vec{u}^{\,\perp}. \end{cases} \tag{2.36}$$

The second equation in (2.36) can be seen as a simple projection of $\vec{u}'$ from the space $\mathbf{H}_0^1(\Omega)$ to $\mathbf{V}_0^\perp(\Omega)$ or as an equation formulated in terms of the Schur complement operator. The Navier–Stokes system is reduced to solving its Schur complement equation, and this form of the Navier–Stokes system allows us to relax the divergence-free constraint. A new domain decomposition theorem can be proved for this different class of problems.

THEOREM 2.6. *Let Ω be a bounded simply connected domain with smooth boundary and let Ω^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains with smooth boundary such that $\Omega = \Omega^1 \cup \Omega^2 \cup \cdots \cup \Omega^m$. Let the jth smoothing step $(\widehat{u}'^{j,m}, \widehat{u}^{j,m\perp}) \in \mathbf{H}_0^1(\Omega) \times \mathbf{V}_0^\perp(\Omega)$ be defined by solving iteratively the local system over the domain Ω^i for $i = 1, \ldots, m$ with $\widehat{u}'^{j,0} = \widehat{u}'^{j-1,m}$ and $\widehat{u}^{\perp j,0} = \widehat{u}^{\perp j-1,m}$. The global sequence is defined by $\widehat{u}'^{j,i} = \vec{w}'^{\,i} + \widehat{u}'^{j,i-1}$, $\widehat{u}^{\perp j,i} = \vec{z}^{\perp i} + \widehat{u}^{\perp j,i-1}$ over Ω^i and $\widehat{u}'^{j,i} = \widehat{u}'^{j,i-1}$, $\widehat{u}^{\perp j,i} = \widehat{u}^{\perp j,i-1}$ over $\Omega - \Omega^i$ $(i = 1, \ldots, m)$, where $(\vec{w}'^{\,i}, \vec{z}^{\perp i})$ solves the local problem*

$$
\begin{cases}
\nu a(\vec{w}'^{\,i} + \widehat{u}'^{j,i-1}, \vec{v}^{\,i}) + c(\vec{u}^{\,i}, \vec{u}^{\,i}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i), \\
a(\vec{z}^{\perp i} + \widehat{u}^{\perp j,i-1}, \vec{v}^{\perp i}) = a(\vec{w}'^{\,i} + \widehat{u}'^{j,i-1}, \vec{v}^{\perp i}) \qquad \forall \vec{v}^{\perp i} \in \mathbf{V}_0^\perp(\Omega^i), \\
\widehat{u}'^{j,i} = \vec{u}^{\,i} + \widehat{u}^{\perp j,i} \qquad \text{on} \quad \Omega^i.
\end{cases} \tag{2.37}
$$

Let $(\widehat{u}'^{0,0}, \widehat{u}^{\perp 0,0})$ be in $\mathbf{H}_0^1(\Omega) \times \mathbf{V}_0^\perp(\Omega)$. Then the global sequence $(\widehat{u}'^{j,m}, \widehat{u}^{\perp j,m})$ converges $(j \to \infty)$ for sufficiently small Reynolds numbers to the solution of the Navier–Stokes problem in (2.36).

Proof. Let $(\vec{u}', \vec{u}^\perp) \in \mathbf{H}_0^1(\Omega) \times \mathbf{V}_0^\perp(\Omega)$ be the solution of the Navier–Stokes problem in (2.36). Then $(\vec{u}', \vec{u}^\perp)$ satisfies

$$
\nu a(\vec{u}', \vec{v}^{\,i}) + c(\vec{u}, \vec{u}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i), \tag{2.38}
$$

$$
a(\vec{u}^{\,\perp}, \vec{v}^{\perp i}) = a(\vec{u}', \vec{v}^{\perp i}) \qquad \forall \vec{v}^{\perp i} \in \mathbf{V}_0^\perp(\Omega^i), \tag{2.39}
$$

with $\vec{u} = \vec{u}' - \vec{u}^\perp$ and appropriate Dirichlet boundary conditions, i.e., the solution $(\vec{u}', \vec{u}^\perp)$ matches the boundary condition over Γ^i. With this notation at the smoothing step j and over Ω^i, the function $\vec{w}'^{\,i}$ satisfies

$$
\begin{aligned}
&\nu a(\vec{w}'^{\,i}, \vec{v}^{\,i}) + c(\vec{u}^{\,i} - \vec{u}, \vec{u}^{\,i} - \vec{u}, \vec{v}^{\,i}) + c(\vec{u}, \vec{u}^{\,i} - \vec{u}, \vec{v}^{\,i}) \\
&+ c(\vec{u}^{\,i} - \vec{u}, \vec{u}, \vec{v}^{\,i}) + \nu a(\widehat{u}'^{j,i-1} - \vec{u}', \vec{v}^{\,i}) = 0 \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i).
\end{aligned} \tag{2.40}
$$

We define $\vec{e}^{\,j,i} = \vec{u}^{\,j,i} - \vec{u}$, $\vec{e}^{\,j,i-1} = \widehat{u}^{\,j,i-1} - \vec{u}$, $\vec{e}^{\,\prime j,i} = \vec{u}'^{j,i} - \vec{u}'$, $\vec{e}'^{j,i-1} = \widehat{u}'^{j,i-1} - \vec{u}'$ over Ω^i and the corresponding extensions as $\widehat{e}^{\,j,i}$, $\widehat{e}^{\,j,i-1}$, $\widehat{e}^{\,\prime j,i}$, $\widehat{e}^{\,\prime j,i-1}$ over Ω in the usual way. Then (2.40) gives

$$
\begin{aligned}
&\nu a(\vec{w}'^{\,i}, \vec{v}^{\,i}) + c(\vec{e}^{\,j,i}, \vec{e}^{\,j,i}, \vec{v}^{\,i}) \\
&+ c(\vec{u}, \vec{e}^{\,j,i}, \vec{v}^{\,i}) + c(\vec{e}^{\,j,i}, \vec{u}, \vec{v}^{\,i}) + \nu a(\vec{e}'^{\,j,i-1}, \vec{v}^{\,i}) = 0.
\end{aligned} \tag{2.41}
$$

Again we note that $\vec{w}'^{\,i}$ is in $\mathbf{H}_0^1(\Omega^i)$ and that $\widehat{u}'^{j,i-1}$, $\vec{u}'^{\,i}$ do not vanish on the boundary Γ^i. For this reason we need to introduce T_i, the projector operator from $\mathbf{H}_0^1(\Omega)$ to $\mathbf{H}_0^1(\Omega^i)$, and write (2.41) in the form

$$
A(\vec{w}'^{\,i} + T_i \widehat{e}'^{j,i-1}) + \frac{1}{\nu}\left(C(\vec{e}^{\,j,i})\,\vec{e}^{\,j,i} + C(\vec{u})\,\vec{e}^{\,j,i} + C(\vec{e}^{\,j,i})\vec{u}^{\,j} \right) = 0 \quad \text{over } \Omega^i \tag{2.42}
$$

or

$$\widehat{e}^{\,\prime j,i} = (I - T_i)\,\widehat{e}^{\,\prime j,i-1} + \frac{1}{\nu}\,A_{*i}^{-1}\left(C(\vec{e}^{\,j,i})\vec{e}^{\,j,i} + C(\vec{u})\,\vec{e}^{\,j,i} + C(\vec{e}^{\,j,i})\vec{u}\right) \tag{2.43}$$

by introducing A_{*i}^{-1}, the inverse operator of A_* over the domain Ω^i with homogeneous boundary condition on Γ^i and the global estimate $\widehat{e}^{\,j,i}$ over Ω.

Now it is possible to prove that $|\widehat{e}^{\,j,i}|_1$ is bounded for $i = 1, 2, \ldots, m$ and $|\widehat{e}^{\,j,m}|_1$ monotonically decreases as j increases for large values of ν. In fact from (2.43) by using the Schwarz inequality and the inequalities in (2.2), since $|\widehat{e}^{\,j,i}|_1 \le |\widehat{e}^{\,\prime j,i}|_1$, we have

$$|\widehat{e}^{\,\prime j,i}|_1 \le |(I - T_i)\,\widehat{e}^{\,\prime j,i-1}|_1 + \frac{K}{\nu}\,(|\widehat{e}^{\,\prime j,i}|_1^2 + |\vec{u}|_1\,|\widehat{e}^{\,\prime j,i}|_1) \tag{2.44}$$

for some K constant. Let the initial error $|\vec{e}^{\,\prime j,0}|_1$ be bounded by M and therefore by $M' = M/(1-\lambda)^m$, where $\lambda = 1 - |\Pi(I - T_i)|_1^{1/m} > 0$. By following the same steps as in the previous theorem we can claim that for ν sufficiently large, the error $|\vec{e}^{\,\prime j,i}|_1$ is bounded by M' for $i = 1, 2, \ldots, m$ and $|\vec{e}^{\,\prime j,m}|_1 < |\vec{e}^{\,\prime j,0}|_1$ for $j = 1, 2, \ldots$.

Now consider (2.37) in the form

$$a(\vec{z}^{\,\perp j,i} - \vec{w}^{\,\prime i}, \vec{v}^{\,\perp i}) + a(\widehat{u}^{\,\perp j,i-1} - \widehat{u}^{\,\prime j,i-1}, \vec{v}^{\,\perp i}) = 0$$

or

$$a(\vec{z}^{\,\perp j,i} - S\,\vec{w}^{\,\prime j,i} + \Pi_i(\widehat{u}^{\,\perp\prime j,i-1} - S\,\widehat{u}^{\,\prime j,i-1}), \vec{v}^{\,\perp i}) = 0, \tag{2.45}$$

where Π_i is the projector from $\mathbf{V}_0^{\perp}(\Omega)$ to $\mathbf{V}_0^{\perp}(\Omega^i)$ and S is the projection operator from $\mathbf{H}_0^1(\Omega)$ to $\mathbf{V}_0^{\perp}(\Omega)$. Let $\vec{r}^{\,j,i} = \widehat{u}^{\,\perp j,i} - S\,\vec{u}^{\,\prime j,i}$ and $\vec{r}^{\,j,i-1} = \widehat{u}^{\,\perp j,i-1} - S\,\vec{u}^{\,\prime j,i-1}$ over Ω^i and let the corresponding extension over Ω be denoted by $\widehat{r}^{\,j,i}$ and $\widehat{r}^{\,j,i-1}$, respectively. With this notation (2.45) becomes

$$a(\vec{z}^{\,\perp j,i} - S\,\vec{w}^{\,\prime j,i} + \Pi_i\widehat{r}^{\,j,i-1}, \vec{v}^{\,\perp i}) = 0 \tag{2.46}$$

and its solution $\vec{z}^{\,\perp j,i} - S\,\vec{w}^{\,\prime j,i} = -\Pi_i\,\widehat{r}^{\,j,i-1}$. If we consider the zero extension of $\vec{z}^{\,\perp j,i}$ and $\vec{w}^{\,\prime j,i}$ and the corresponding extension $\widehat{r}^{\,j,i}$ over Ω, then the global solution $\widehat{r}^{\,j,i}$ satisfies

$$\widehat{r}^{\,j,i} = (I - \Pi_i)\,\widehat{r}^{\,j,i-1}$$

for $i = 1, 2, \ldots, m$, which implies

$$\widehat{r}^{\,j,m} = \Pi_{i=1}^{m}(I - \Pi_i)\,\widehat{r}^{\,j,0}.$$

Since $|\Pi_{i=1}^{m}(I - \Pi_i)|_1 < 1$ we have $|\widehat{r}^{\,j,m}|_1 < |\widehat{r}^{\,j,0}|_1$. Applying the same procedure for each smoothing step j we have that the sequence $|\widehat{r}^{\,j,m}|_1$ monotonically decreases for increasing j. However, from the previous estimate, the sequence $|\widehat{e}^{\,\prime j,m}|_1$ is monotonically decreasing and therefore $\widehat{u}^{\,\prime j,m}$ tends to $\widehat{u}^{\,\prime}$. The sequence $\widehat{u}^{\,\perp j,m}$ tends to $S\,\vec{u}^{\,\prime j,m}$ and therefore to $S\vec{u}^{\,\prime} = \vec{u}^{\perp}$, which is the solution of (2.36) as j tends to infinity. ∎

We can reformulate the above algorithm with standard notation in velocity and pressure. Given $(\widehat{u}^{\,i-1}, \widehat{p}^{\,i-1}) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ at the local step $i-1$ we define the local problem for $\vec{w}^{\,i} = \vec{u}^{\,i} - \widehat{u}^{\,i-1} \in \mathbf{H}_0^1(\Omega^i)$ and $z^{\,i} = p^{\,i} - p^{\,i-1} \in L_0^2(\Omega^i)$ by

$$\begin{cases} \nu a(\vec{w}^{\,i}, \vec{v}^{\,i}) + b(z^{\,i}, \vec{v}^{\,i}) + c(\vec{w}^{\,i}, \vec{w}^{\,i}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i}\rangle - \nu a(\widehat{u}^{\,i-1}, \vec{v}^{\,i}) - b(\widehat{p}^{\,i}, \vec{v}^{\,i}) \\ \qquad - c(\widehat{u}^{\,i-1}, \vec{w}^{\,i}, \vec{v}^{\,i}) - c(\vec{w}^{\,i}, \widehat{u}^{\,i-1}, \vec{v}^{\,i}) \qquad \forall \vec{v}^{\,i} \in \mathbf{H}_0^1(\Omega^i), \\ b(r^i, \vec{w}^{\,i}) + b(r^i, \widehat{u}^{\,i-1}) = 0 \qquad \forall r^i \in L_0^2(\Omega^i), \end{cases}$$

with boundary condition $\vec{w}^{\,i} = 0$. The global solution $(\widehat{u}^{\,i}, \widehat{p}^{\,i})$ is determined by the local solution as $\widehat{u}^{\,i} = \vec{w}^{\,i} + \widehat{u}^{\,i-1}$, $\widehat{p}^{\,i} = z^{\,i} + \widehat{p}^{\,i-1}$ over the subdomain Ω^i and $\widehat{u}^{\,i} = \widehat{u}^{\,i-1}$, $\widehat{p}^{\,i} = \widehat{p}^{\,i-1}$ over $\Omega - \Omega^i$. We remark that the restriction of $\widehat{u}^{\,i}$ over Ω^i is in $\mathbf{V}_0(\Omega^i)$, but the global function $\widehat{u}^{\,i}$ is not in $\mathbf{V}_0(\Omega)$. The divergence-free constraint is relaxed and is matched only in the limit of the algorithm when convergence is reached.

3. Formulation of the discrete problem.

3.1. Introduction. In this section we investigate the numerical behavior of the Vanka-type smoothers by using the finite element method. The finite element method is particularly suitable for multigrid and Vanka-type smoothers since the domain geometry is well defined by the finite element structure and the local problems can be solved at the element level, where the topology is already available. In this paper we deal only with the convergence issue and leave the multigrid discussion to successive papers. In particular we would like to discuss the convergence of the Vanka smoothers when the local solution is obtained over overlapping subdomains that consist of a cluster of a few finite elements. It is possible to show that smoothers based on these blocks lead to the solution of the Stokes system for any value of viscosity. The same smoothers can be used to solve the Navier–Stokes system but only at low Reynolds numbers. The issue of the high Reynolds is not treated here but can be studied in the analysis of the time-dependent Navier–Stokes formulation.

In order to avoid technicalities in this approach we use rectangular conforming finite elements (standard Taylor–Hood). It is possible to generalize to triangular elements in a very straightforward manner. It is also possible to use different elements and generalize the analysis to nonstandard Taylor–Hood elements. Under these assumptions we consider two cases. In case A the local subdomain Ω^i consists of one quadrilateral finite element and all its neighboring elements. In order to solve the local problem we have to solve for all the velocity and pressure unknowns. The boundary conditions are imposed along the boundary Γ_h^i of the subdomain Ω^i. For the two-dimensional case we can see such an element in Figure 3.1 on the left. In the case of Taylor–Hood finite elements we have sixteen unknowns for each velocity component and pressure.

Case B consists of one quadrilateral finite element and all its neighboring elements, but we do not solve for the pressure on the boundary Γ_h^i. The two-dimensional case can be seen in Figure 3.1 on the right. The two-dimensional case for the Taylor–Hood rectangular finite element has four pressure unknowns and sixteen nodes for each velocity component.

Other elements have been used with Vanka-type smoothers and many, especially nonconforming elements, can be found in the literature; see, for example, [13, 14, 24] and the references therein.

Let Ω_h be a polygonal domain with boundary Γ_h. Then we subdivide the domain Ω_h into the rectangle elements by using unstructured families of meshes, T_h^{i,l_0}. In this analysis we consider only conforming finite element approximations. Let $X^h \subset H^1(\Omega)$ and $P^h \subset L^2(\Omega)$ be two families of finite-dimensional subspaces parameterized by h, which tends to zero. We also denote $X_0^h = X^h \cap H_0^1(\Omega)$ and $P_0^h = P^h \cap L_0^2(\Omega)$. We make the following assumptions on X^h and P^h:

(a) *The approximation hypotheses:* there exists an integer l and a constant C, independent of h, $\vec{u}$, and p, such that for $1 \le k \le l$ we have

$$\inf_{\vec{u}_h \in X^h} \|\vec{u}_h - \vec{u}\|_1 \le C h^k \|\vec{u}\|_{k+1} \qquad \forall\, \vec{u} \in H^{k+1}(\Omega) \cap H_0^1(\Omega), \tag{3.1}$$

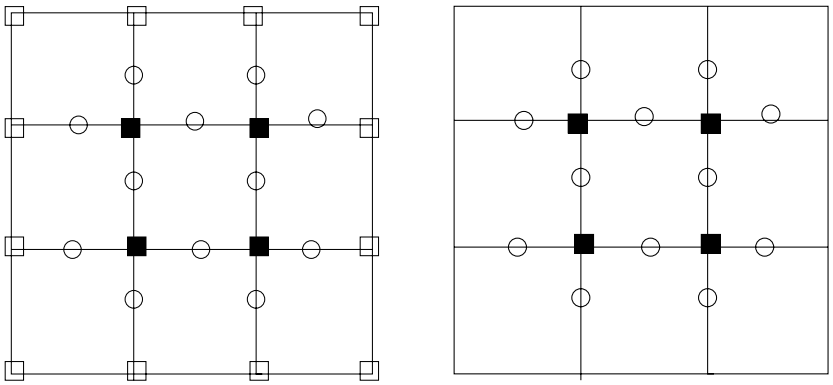


FIG. 3.1. *A block of type A (on the left) and a block of type B (on the right) with velocity (circle), pressure (square), and velocity-pressure (black square) nodes for rectangular Taylor–Hood elements in the two-dimensional geometry.*

$$\inf_{p_h \in P^h} \|p - p_h\| \le C h^k \|p\|_k \qquad \forall\, p \in H^k(\Omega) \cap L_0^2(\Omega). \tag{3.2}$$

(b) *The inf-sup condition or LBB condition:* there exists a constant C', independent of h, such that

$$\inf_{0 \neq q_h \in P^h} \sup_{0 \neq \vec{u}_h \in X^h} \frac{\int_\Omega q_h div \vec{u}_h}{\|\vec{u}_h\|_1 \|q_h\|} \ge C' > 0. \tag{3.3}$$

This condition assures the stability of the discrete Navier–Stokes solutions and plays a key role in the case of blocks of type B.

3.2. Stokes problem with element blocks of type A. Let Ω_h be a bounded simply connected domain with polygonal boundary Γ_h and rectangular triangulations. Let $\Omega_h = \cup_i^m \Omega_h^i$, where the subdomains Ω_h^i are blocks of finite elements of type A with boundary Γ_h^i. This family of subdomains is overlapping in the sense that $\mathbf{X}_0^h(\Omega_h) = \mathbf{X}_0^h(\Omega_h^1) + \mathbf{X}_0^h(\Omega_h^2) + \cdots + \mathbf{X}_0^h(\Omega_h^m)$.

We introduce the divergence-free function space as

$$\mathbf{V}_0^h(\Omega_h) = \{\vec{u}_h \in \mathbf{X}_0^h(\Omega_h) :\, b(\vec{u}_h, r_h) = 0 \quad \forall r_h \in P_0^h(\Omega_h)\},$$

which is also $\mathbf{V}_0^h(\Omega_h) = \mathrm{Ker}\,(B_h) \cap \mathbf{X}_0^h(\Omega_h)$, and note that this is not a subspace of $\mathbf{V}_0(\Omega_h)$. We equip $\mathbf{X}_0^h(\Omega_h)$ and $\mathbf{V}_0^h(\Omega_h)$ with the scalar product $(\cdot,\cdot)_1$ and the norm $|\cdot|_1$ defined by $(\vec{v}_h, \vec{u}_h)_1 = a(\vec{v}_h, \vec{u}_h)$ and $|\vec{u}_h|_1 = a_h(\vec{u}_h, \vec{u}_h)$ for all $\vec{v}_h$, $\vec{u}_h$ in $\mathbf{X}_0^h(\Omega_h)$, respectively.

With this notation we can introduce the Stokes problem over the domain Ω_h.

Given $\vec{f} \in \mathbf{H}^{-1}(\Omega_h)$, find the pair $(\vec{u}_h, p_h) \in \mathbf{X}_0^h(\Omega_h) \times P_0^h(\Omega_h)$ solution of

$$\begin{cases} a(\vec{u}_h, \vec{v}_h) + b(p_h, \vec{v}_h) = \langle \vec{f}, \vec{v}_h \rangle \qquad \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h), \\ b(r_h, \vec{v}_h) = 0 \qquad \forall r_h \in P_0^h(\Omega_h), \end{cases} \tag{3.4}$$

with boundary condition $\vec{v}_h^i = 0$ over Γ_h.

The Vanka-type smoother of type A is an iterative method which computes the solution of the Stokes problem in (3.4) by solving local Stokes problems over the unknowns which are located in a finite element block of type A. The main characteristic

of this block is that the velocity solution $\vec{v}_h$ is in $\mathbf{V}_0^h(\Omega_h)$ at each step and is determined by the solution of a Stokes local problem. More precisely, given the state solution $(\vec{v}_h^{\,i-1}, p_h^{\,i-1}) \in \mathbf{V}_0^h(\Omega_h) \times \mathbf{P}_0^h(\Omega_h)$ at step $i-1$, we define the ith local problem for the pair $(\vec{v}_h^{\,i}, p_h^{\,i}) \in \mathbf{X}^h(\Omega_h^i) \cap \mathbf{V}^h(\Omega_h^i) \times \mathbf{P}_0^h(\Omega_h^i)$ by

$$\begin{cases} a(\vec{u}_h^{\,i}, \vec{v}_h^{\,i}) + b(p_h^i, \vec{v}_h^{\,i}) = \langle \vec{f}, \vec{v}_h^{\,i} \rangle \qquad \forall \vec{v}_h^{\,i} \in \mathbf{X}_0^h(\Omega_h^i), \\ b(r_h^i, \vec{v}_h^{\,i}) = 0 \qquad \forall r_h^i \in P_0^h(\Omega_h^i), \end{cases} \tag{3.5}$$

with boundary conditions $\vec{v}_h^{\,i} = \vec{v}_h^{\,i-1}$ over Γ_h^i and $\vec{v}_h^{\,i} = 0$ over $\Gamma_h^i \cap \Gamma_h$. At the ith iteration the global solution $(\widehat{u}^{\,i}, \widehat{p}^{\,i})$ is determined by the local solution as $\widehat{u}^{\,i} = \vec{u}^{\,i}$, $\widehat{p}^{\,i} = p^{\,i}$ over Ω_h^i and by the old solution as $\widehat{u}^{\,i} = \widehat{u}^{\,i-1}$, $\widehat{p}^{\,i} = \widehat{p}^{\,i-1}$ over $\Omega_h - \Omega_h^i$.

The local problem in (3.5) can be written in the natural way as a projection of the initial Stokes problem over the domain Ω_h^i if we use the variable $\vec{w}^{\,i} = \vec{u}^{\,i} - \widehat{u}^{\,i-1} \in \mathbf{X}_0^h(\Omega_h^i) \cap \mathbf{V}_0^h(\Omega_h^i)$ and rewrite the system as

$$\begin{cases} a(\vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + b(p_h^i, \vec{v}_h^{\,i}) + a(\widehat{u}_h^{\,i-1}, \vec{v}_h^{\,i}) = \langle \vec{f}, \vec{v}_h^{\,i} \rangle \qquad \forall \vec{v}_h^{\,i} \in \mathbf{X}_0^h(\Omega_h^i), \\ b(r_h^i, \vec{w}_h^{\,i}) = 0 \qquad \forall r_h^i \in P_0^h(\Omega_h^i), \end{cases}$$

with boundary condition $\vec{w}_h^{\,i} = 0$ over Γ_h^i.

If the projection P_h^i from $\mathbf{V}_0^h(\Omega_h^i)$ onto $\mathbf{V}_0^h(\Omega_h^i)$ with respect to the inner product $(\cdot,\cdot)_1$ is introduced, then we can write the local problem for $\vec{w}_h^{\,i} \in \mathbf{X}_0^h$ as

$$a(\vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + a(P_h^i \widehat{u}_h^{\,i-1}, \vec{v}_h^{\,i}) = \langle \vec{f}, \vec{v}_h^{\,i} \rangle \qquad \forall \vec{v}_h^{\,i} \in \mathbf{V}_0^h(\Omega_h^i), \tag{3.6}$$

which matches exactly the projection of the Stokes problem over $\mathbf{V}_0^h(\Omega_h^i)$. The global solution $\widehat{u}_h^i$ can be computed by $\vec{w}_h^{\,i} + \widehat{u}_h^{\,i-1}$. The solution of (3.6) is clearly $\vec{w}_h^{\,i} = -P_i^h \widehat{u}_h^{\,i-1} + A_{ih}^{-1} \vec{f}$ and the global solution simply $\widehat{u}_h^i = (I - P_i^h)\widehat{u}_h^{\,i-1} + A_i^{-1} \vec{f}$, where A_{ih}^{-1} is the green operator of the Laplacian operator over Ω_h^i with homogeneous boundary conditions on Γ_h^i. Therefore we start to discuss convergence for these Vanka smoothers by investigating the properties of the operators $(I - P_h^i)$.

In order to study the behavior of the projection operator P_i^h we state a space decomposition theorem which allows us to search the solution of the original Stokes problem in the projection spaces.

THEOREM 3.1. *Let Ω_h be a bounded simply connected domain with polygonal boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains obtained by clustering blocks of elements of type A such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cdots \Omega_h^m$ and $\mathbf{X}_0^1(\Omega_h) = \mathbf{X}_0^h(\Omega_h^1) + \mathbf{X}_0^h(\Omega_h^2) + \mathbf{X}_0^h(\Omega_h^3) \cdots \mathbf{X}_0^h(\Omega_h^m)$. Let $\mathbf{V}_0^h(\Omega_h^i)$ be the divergence-free function space in $\mathbf{X}_0^h(\Omega_h^i)$ with respect to the inner product $(\cdot,\cdot)_1$ for $i = 1, 2, \ldots, m$. Then we have that $\mathbf{V}_0^h(\Omega_h) = \mathbf{V}_0^h(\Omega_h^1) + \mathbf{V}_0^h(\Omega_h^2) + \mathbf{V}_0^h(\Omega_h^3) \cdots \mathbf{V}_0^h(\Omega^m)$.*

Proof. First consider the case for $m = 2$. Let $\vec{v}_h \in \mathbf{V}_0^h(\Omega_h)$. Since the domains are overlapping in the space $\mathbf{X}_0^h(\Omega_h)$ we can write $\vec{v}_h = \vec{v}_{1h} + \vec{v}_{2h}$ with $\vec{v}_{ih} \in \mathbf{X}_0^h(\Omega_h^i)$, $i = 1, 2$. Let $\Omega_{12}^h = \Omega_h^1 \cap \Omega_2^h$. Then there exists an $\vec{s}_h$ over Ω_{12}^h with zero extension such that $B_h(\vec{s}_h - \vec{v}_{1h}) = 0$. Then $\vec{v}_{1h}^{\,'} = \vec{v}_{1h} - \vec{s}_h$ and $\vec{v}_{2h}^{\,'} = \vec{v}_{2h} - \vec{s}_h$ gives the desired decomposition. For any m the theorem can be proved by induction by clustering the subdomains into two groups and using the result for $m = 2$ [15, 16, 11]. ∎

In order to prove convergence of the algorithm we need to use the following properties of the projection P_h^i.

THEOREM 3.2. *Let Ω_h be a bounded simply connected domain with polygonal boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains obtained*

by clustering blocks of elements of type A such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cdots \Omega_h^m$. *Let* P_i^h *be the orthogonal projections from* $\mathbf{V}_0^h(\Omega_h)$ *onto* $\mathbf{V}_0^h(\Omega_h^i)$ *for* $i = 1, 2, \ldots, m$. *Then* $|\Pi_{i=1}^m (I - P_i^h)|_1 < 1$.

Proof. As in the continuous case we use a standard minimization approach. Given $\vec{f} \in L^2(\Omega_h)$ consider the solution $\widehat{u}_h \in \mathbf{V}_0^h(\Omega_h)$ of the following Laplace problem:

$$a(\widehat{u}_h, \widehat{v}_h) = \langle \vec{f}, \widehat{v}_h \rangle \qquad \forall \widehat{v}_h \in \mathbf{V}_0^h(\Omega_h), \tag{3.7}$$

with homogeneous Dirichlet boundary condition or the equivalent minimization problem

$$\widehat{u}_h = \min_{\widetilde{u} \in \mathbf{V}_0^h(\Omega_h)} \left(\frac{1}{2}(A_h \widetilde{u}_h, \widetilde{u}_h) - \langle \vec{f}, \widetilde{u}_h \rangle + \frac{1}{2} \langle A_h^{-1} \vec{f}, \vec{f} \rangle \right). \tag{3.8}$$

The constant $\frac{1}{2}\langle A_h^{-1}\vec{f}, \vec{f}\rangle$ is added to the functional in order to make the functional vanish at the critical point $\widetilde{u}_h = \widehat{u}_h$, and the operator A_h^{-1} from $\mathbf{H}^{-1}(\Omega_h)$ to $\mathbf{X}_0^h(\Omega_h)$ denotes the Green's operator of the homogeneous problem with the operator A_h.

Given $\widehat{u}_h \in \mathbf{V}_0^h(\Omega_h)$, then the projection $\vec{u}_h^i = P_h^i \widehat{u}_h$ from $\mathbf{V}_0^h(\Omega_h)$ to $\mathbf{V}_0^h(\Omega_h^i)$ can be seen as a solution of the following problem:

$$(A_h \vec{u}_h^i, \vec{v}_h^i) = (A_h \widehat{u}_h, \vec{v}_h^i) \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i), \tag{3.9}$$

with homogeneous boundary conditions over Ω_h^i. The residual vector can be written as $\vec{w}_h^i = \widehat{u}_h - \vec{u}_h^i$ over Ω_h^i or $\widehat{w}_h^i = \widehat{u}_h - P_h^i \widehat{u}_h^i = (I - P_h^i)\widehat{u}_h^i$ over Ω_h. The natural approach of studying the operator $\Pi_{i=1}^m (I - P_h^i)$ is to minimize (3.8) over the sequence of domains Ω_h^i in the space $\mathbf{V}_0^h(\Omega_h^i)$ for $i = 1, 2, \ldots, m$ and solve for the residual vector. Let $\widehat{u}_h^{i-1}$ be the iterative solution over the whole domain Ω_h at step $i-1$ and let P_i^h be the projector from $\mathbf{V}_0^h(\Omega_h)$ to $\mathbf{V}_h^i(\Omega_h^i)$. Minimization with respect to the unknowns $\widetilde{u}_h^i$ related to the domain Ω_h^i gives

$$(A_h \vec{u}_h^i, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i), \tag{3.10}$$

where $\vec{v}_h^i$ is set to be equal to the variation $\delta \widetilde{u}_h^i$ and $\vec{u}_h^i$ is the minimizer over $\mathbf{V}_0^h(\Omega_h^i)$. If we set $\vec{w}_h^i = \vec{u}_h^i - \widehat{u}_h^{i-1}$ over the domain Ω_h^i, then (3.10) gives

$$(A_h \vec{w}_h^i, \vec{v}_h^i) + (A_h \widehat{u}_h^{i-1}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i).$$

It is clear that in the above equation the information on the boundary is taken from the old solution $\widehat{u}_h^{i-1}$. If we use the exact solution $\vec{u}_h$ which satisfies

$$(A_h \vec{u}_h, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \quad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i),$$

then we have

$$(A_h \vec{w}_h^i, \vec{v}_h^i) + (A_h(\widehat{u}_h^{i-1} - \vec{u}_h), \vec{v}_h^i) = 0 \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i),$$

which is solved by

$$\vec{w}_h^i = -P_i^h(\widehat{u}_h^{i-1} - \vec{u}_h) \tag{3.11}$$

over Ω_h^i or

$$\widehat{u}_h^i = (I - P_i^h)\, \widehat{u}_h^{i-1} + P_i^h \vec{u}_h \tag{3.12}$$

over Ω_h, where the extension $\widehat{u}_h^i$ of $\vec{u}_h^i$ over $\Omega_h - \Omega_h^i$ is taken to be equal to $\widehat{u}_h^{i-1}$. If we define the error as the difference between the iterative and the finite element solution,

$$\widehat{e}_h^{\,i-1} = \widehat{u}_h^{i-1} - \vec{u}_h \quad \text{and} \quad \widehat{e}_h^{\,i} = \widehat{u}_h^{i} - \vec{u}_h,$$

then the solution error of the Laplace problem satisfies

$$\widehat{e}_h^{\,i} = (I - P_i^h)\,\widehat{e}_h^{\,i-1}. \tag{3.13}$$

Since these estimates are true for all $i = 1, 2, \ldots, m$, we can combine them in the following solution:

$$\widehat{e}_h^{\,m} = \Pi_{i=1}^m (I - P_i^h)\,\widehat{e}_h^{\,0}, \tag{3.14}$$

which gives the global error over Ω_h needed to estimate the functional L_h^i and the operator norm over the domain Ω_h.

Now we show that the functional L_h^i, which is the norm of the error $\widehat{e}_h^{\,i}$, is monotonically decreasing for increasing i for all initial guesses $\widehat{e}_h^{\,0} \in \mathbf{V}_0^h(\Omega_h^i)$. In fact

$$\begin{aligned} L_h^i &= \frac{1}{2}(A_h \widehat{u}_h^i, \widehat{u}_h^i) - \langle \vec{f}, \widehat{u}_h^i \rangle + \frac{1}{2}\langle A_h^{-1}\vec{f}, \vec{f}\rangle \\ &= \frac{1}{2}(A_h(\vec{w}_h^i + \widehat{u}_h^{i-1}), \vec{w}_h^i + \widehat{u}_h^{i-1}) - \langle \vec{f}, \vec{w}_h^i + \widehat{u}_h^{i-1}\rangle + \frac{1}{2}\langle A_h^{-1}\vec{f}, \vec{f}\rangle \\ &= L_h^{i-1} + \frac{1}{2}(A_h \vec{w}_h^i, \vec{w}_h^i) + (A_h \widehat{u}_h^{i-1}, \vec{w}_h^i) - \langle \vec{f}, \vec{w}_h^i\rangle, \end{aligned}$$

and by using the optimality condition (3.10),

$$L_h^i = L_h^{i-1} - \frac{1}{2}(A_h \vec{w}_h^i, \vec{w}_h^i).$$

Therefore, by starting with an initial guess $\widehat{u}_h^0 \in \mathbf{V}_0^h(\Omega_h)$ after a smoothing step over m domains, we have

$$L^m = L^0 - \frac{1}{2}\sum_{i=1}^m |\widehat{w}_h^i|_1^2,$$

with $\widehat{w}_h^i = -P_i^h \widehat{e}_h^{\,i-1}$, $L_h^m = |\widehat{e}_h^{\,m}|_1^2/2$, and $L_h^0 = |\widehat{e}_h^{\,0}|_1^2/2$. Also we have $\widehat{e}_h^{\,m} = \Pi_{i=1}^m (I - P_i^h)\,\widehat{e}_h^{\,0}$ for all $\widehat{e}_h^{\,0} \in \mathbf{V}_0^h(\Omega_h)$. From the definition of norm we have

$$|\Pi_{i=1}^m (I - P_i^h)|_1 = \sup \ |\widehat{e}_h^{\,m}|_1 = \sqrt{1 - \inf \sum_{i=1}^m |\widehat{w}_h^i|_1^2}, \tag{3.15}$$

where the sup and inf are taken over the set of functions $\widehat{e}_h^{\,0} \in \mathbf{V}_0^h(\Omega_h)$ with $|\widehat{e}_h^{\,0}|_1 = 1$. The infimum is the minimizer of the minimization problem

$$\min \ \sum_{i=1}^m |\widehat{w}_h^i|_1^2,$$

where the minimum is taken over all the functions $\widehat{e}_h^{\,0}$ in $\mathbf{V}_0^h(\Omega_h)$ with $|\widehat{e}_h^{\,0}|_1 = 1$. Standard theory can be applied to prove existence, well-posedness, and uniqueness of

the solution of the problem over the set $\mathbf{V}_0^h(\Omega_h)$. Now we prove that the minimum cannot be zero and the corresponding norm $|\Pi_{i=1}^m (I - P_i^h)|_1 < 1$. Suppose that the minimum is zero. Then this implies, from the definition of the norm itself, that there exists an $\widehat{e}_h^0$ in $\mathbf{V}_0^h(\Omega_h)$ such that $\widehat{e}_h^n = \widehat{e}_h^0$, and therefore $P_i^h \widehat{e}_h^0 = 0$ for all $i = 1, 2, \ldots, m$. Such a function $\widehat{e}_h^0$ has norm 1 but zero projections over all $\mathbf{V}_0^h(\Omega_h^i)$ $i = 1, 2, \ldots, m$, and this contradicts the fact that $\mathbf{V}_0^h(\Omega_h) = \mathbf{V}_0^h(\Omega_h^1) + \mathbf{V}_0^h(\Omega_h^2) + \cdots + \mathbf{V}_0^h(\Omega_h^m)$, i.e., the domains are overlapping in the space $\mathbf{V}_0^h(\Omega_h)$. Therefore the functional must be different from zero and must assume a well-defined positive value at the critical point. Combining this result with (3.15), the inequality $|\Pi_{i=1}^m (I - P_i^h)|_1 < 1$ is obtained. ∎

The projections introduced above can be used to solve the Stokes and the Navier–Stokes equations with different types of domain decompositions.

Now we are ready to state a convergence theorem for the Stokes problem.

THEOREM 3.3. *Let Ω_h be a bounded simply connected domain with polygonal boundary Γ_h and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type A such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$. Let the jth smoothing step $\widehat{u}_h^{j,m} \in \mathbf{V}_0^h(\Omega_h)$ be defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$ with $\widehat{u}^{j,0} = \widehat{u}^{j-1,m} \in \mathbf{V}_0^h(\Omega_h)$. The global sequence is defined by $\widehat{u}_h^{j,i} = \vec{w}_h^i + \widehat{u}_h^{i-1}$ over Ω_h^i and $\widehat{u}_h^{j,i} = \widehat{u}_h^{i-1}$ over $\Omega_h - \Omega_h^i$ $(i = 1, \ldots, m)$, where $\vec{w}_h^i \in \mathbf{V}_0^h(\Omega_h^i)$ solves the local problem*

$$\nu a(\vec{w}_h^i, \vec{v}_h^i) + \nu a(\widehat{u}_h^{i-1}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i), \tag{3.16}$$

with boundary condition $\vec{w}_h^i = 0$ over Γ^i.

Let $\widehat{u}^{0,0}$ be in $\mathbf{V}_0^h(\Omega_h)$. Then global sequence $\widehat{u}^{j,m}$ converges $(j \to \infty)$ to the solution of the Stokes problem in (3.4).

Proof. Let $\widehat{u}_h^{j,i-1}$ be in $\mathbf{V}_0^h(\Omega_h)$. In (3.16) we have

$$\nu a(\vec{w}_h^{j,i}, \vec{v}_h^i) + \nu a(\widehat{u}_h^{j,i-1}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i).$$

Let $\vec{u}_h$ be the solution of the Stokes problem. Then

$$\nu a(P_i \vec{u}_h, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{V}_0^h(\Omega_h^i),$$

and therefore

$$\nu a(\vec{w}_h^{j,i}, \vec{v}_h^i) + \nu a(P_i^h(\widehat{u}_h^{j,i-1} - \vec{u}_h), \vec{v}_h^i) = 0. \tag{3.17}$$

If we set $\vec{e}_h^{j,i} = \widehat{u}_h^{j,i} - \vec{u}_h$ over Ω_h^i and $\widehat{e}_h^{j,i-1} = \widehat{u}_h^{j,i-1} - \vec{u}_h$ over Ω_h, the solution of (3.17) gives

$$\vec{e}_h^{j,i} = (I - P_i^h)\, \widehat{e}_h^{j,i-1}.$$

We can define $\widehat{e}_h^{j,i}$ as the extension of $\widehat{e}_h^{j,i}$ over Ω_h and write the solution as

$$\widehat{e}_h^{j,i} = (I - P_i^h)\, \widehat{e}_h^{j,i-1}. \tag{3.18}$$

Since (3.18) is true for all i and j, then

$$\widehat{e}_h^{j,m} = (I - P_m^h)(I - P_{m-1}^h)(I - P_{m-2}^h) \cdots (I - P_2^h)(I - P_1^h)\, \widehat{e}_h^{j,0}, \tag{3.19}$$

and from Theorem 3.2 we have that

$$|\widehat{e}_h^{j,m}|_1 = \lambda_j\, |\widehat{e}_h^{j,0}|_1,$$

with $\lambda_j < 1$. Since $\lambda_j < 1$ for all j the sequence $\widehat{e}_h^{\,j,m}$ converges to zero as $j \to \infty$, and therefore $\widehat{u}_h^{j,m}$ to $\vec{u}_h$. ∎

Now we analyze the algorithm with blocks of type A for the Navier–Stokes system.

Given $\vec{f} \in H^{-1}(\Omega)$, then $(\vec{u}_h, p_h)$ is called a *generalized solution* of the fully discrete approximate Navier–Stokes equations if $\vec{u}_h \in \mathbf{X}_0^h(\Omega_h)$, $p_h \in S_0^h(\Omega_h)$ and satisfies the following system of equations:

$$\begin{cases} \nu a(\vec{u}_h, \vec{v}_h) + c(\vec{u}_h; \vec{u}_h, \vec{v}_h) + b(\vec{v}_h, p_h) = 0 \qquad \forall\, \vec{v}_h \in \mathbf{X}_0^h(\Omega_h), \\ b(\vec{v}_h, q_h) = 0 \qquad \forall\, q_h \in S_0^h(\Omega_h). \end{cases} \tag{3.20}$$

The Vanka smoother of type A for the steady Navier–Stokes problem can be proved to be convergent only for small Reynolds numbers.

THEOREM 3.4. *Let Ω_h be a bounded simply connected domain with smooth boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type A such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$. Let the jth smoothing step $\widehat{u}_h^{j,m} \in \mathbf{V}_0^h(\Omega_h)$ be defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$ with $\widehat{u}_h^{j,0} = \widehat{u}_h^{j-1,m} \in \mathbf{V}_0^h(\Omega_h)$. The global sequence is defined by $\widehat{u}_h^{j,i} = \vec{w}_h^{\,i} + \widehat{u}_h^{j,i}$ over Ω_h^i and $\widehat{u}_h^{j,i} = \widehat{u}_h^{i-1}$ over $\Omega_h - \Omega_h^i$ $(i = 1, \ldots, m)$, where $\vec{w}_h^{\,i} \in \mathbf{V}_0^h(\Omega_h^i)$ solves the local problem*

$$\nu a(\vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + c(\vec{w}_h^{\,i} + \widehat{u}^{\,i-1}, \vec{w}_h^{\,i} + \widehat{u}^{\,i-1}, \vec{v}_h^{\,i}) = \langle \vec{f}, \vec{v}_h^{\,i} \rangle - \nu a(\widehat{u}_h^{\,i-1}, \vec{v}_h^{\,i}) \tag{3.21}$$

for all $\vec{v}_h^{\,i} \in \mathbf{V}_0^1(\Omega_h^i)$ with boundary condition $\vec{w}_h^{\,i} = 0$ over Γ_h^i.

Let $\widehat{u}^{\,0,0}$ be in $\mathbf{V}_0(\Omega_h)$. Then the global sequence $\widehat{u}_h^{j,m}$ converges $(j \to \infty)$ for sufficiently small Reynolds numbers to the solution of the Navier–Stokes problem in (3.20).

Proof. The proof follows the Stokes approach. From (3.21) we have

$$\nu a(\vec{w}^{\,i}, \vec{v}_h^{\,i}) + c(\widehat{w}^{\,i} + \widehat{u}^{\,j,i-1}, \widehat{w}^{\,i} + \widehat{u}^{\,j,i-1}, \vec{v}_h^{\,i}) + \nu a(\widehat{u}^{\,j,i-1}, \vec{v}^{\,i}) = \langle \vec{f}, \vec{v}^{\,i} \rangle \tag{3.22}$$

for all $\vec{v}_h^{\,i} \in \mathbf{V}_0^1(\Omega_h^i)$. Let $\vec{u}_h$ be the solution of the Navier–Stokes problem in (3.20) with the appropriate boundary condition. Then we have

$$\nu a(\vec{u}_h, \vec{v}_h^{\,i}) + c(\vec{u}_h, \vec{u}_h, \vec{v}_h^{\,i}) = \langle \vec{f}, \vec{v}_h^{\,i} \rangle \qquad \forall \vec{v}_h^{\,i} \in \mathbf{V}_0^h(\Omega^h),$$

and we can rewrite (3.22) in the following form:

$$\nu a(\vec{w}_h^{\,j,i} + \widehat{u}_h^{\,j,i-1}, \vec{v}^{\,i}) + c(\vec{w}_h^{\,j,i} + \widehat{u}_h^{\,j,i-1}, \vec{w}_h^{\,j,i} + \widehat{u}_h^{\,j,i-1}, \vec{v}^{\,i}) = c(\vec{u}_h, \vec{u}_h, \vec{v}_h^{\,i}). \tag{3.23}$$

If we set $\vec{e}_h^{\,j,i} = \vec{w}_h^{\,j,i} + \widehat{u}_h^{\,j,i-1} - \vec{u}_h$ over Ω_h^i and $\widehat{e}_h^{\,j,i-1} = \widehat{u}_h^{\,j,i-1} - \vec{u}_h$ over Ω_h, then (3.23) gives

$$\nu a(\vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + c(\vec{u}_h, \widehat{e}_h^{\,j,i}, \vec{v}_h^{\,i}) + c(\widehat{e}^{\,j,i}, \vec{u}_h, \vec{v}_h^{\,i}) + \nu a(\widehat{e}_h^{\,j,i-1}, \vec{v}_h^{\,i}) = 0. \tag{3.24}$$

We note that $\vec{w}_h^{\,i} = \vec{e}_h^{\,j,i} - \widehat{e}_h^{\,j,i-1}$ is in $\mathbf{V}_0^1(\Omega_h^i)$. By using the zero extension of $\vec{w}_h^{\,i}$ we can define the extension $\widehat{e}_h^{\,j,i} = \widehat{e}_h^{\,j,i-1} + \vec{w}_h^{\,i}$ of $\vec{e}_h^{\,j,i}$ to Ω_h. Then (3.24) can be written in operator form as

$$A(\vec{w}_h^{\,i} + P_i^h\, \widehat{e}_h^{\,j,i-1}) + \frac{1}{\nu} \Big(C(\vec{e}_h^{\,j,i}) \vec{e}_h^{\,j,i} + C(\vec{u}_h)\, \vec{e}_h^{\,j,i} + C(\vec{e}_h^{\,j,i}) \vec{u} \Big) = 0 \quad \text{on } \Omega_h^i \tag{3.25}$$

or

$$\widehat{e}_h^{\,j,i} = (I - P_i^h)\,\widehat{e}_h^{\,j,i-1} + \frac{1}{\nu}\,A_i^{-1}\Big(\,C(\vec{e}_h^{\,j,i})\vec{e}_h^{\,j,i} + C(\vec{u}_h)\,\vec{e}_h^{\,j,i} + C(\vec{e}_h^{\,j,i})\vec{u}_h\Big) \tag{3.26}$$

over Ω_h^i, where A_i^{-1} is the inverse operator of A over the domain Ω_h^i with homogeneous boundary condition on Γ_h^i. By using the discrete analogous inequalities (2.2) we estimate

$$|\widehat{e}_h^{\,j,i}|_1 \le |(I - P_i^h)\,\widehat{e}_h^{\,j,i-1}|_1 + \frac{1}{\nu}\,K\,(|\widehat{e}_h^{\,j,i}|_1^2 + |\vec{u}_h|_1\,|\widehat{e}_h^{\,j,i}|_1) \tag{3.27}$$

for some K constant.

By using the same approach as in the continuous case we can prove that the error $\widehat{e}_h^{\,j,i}$ is bounded for large values of ν; i.e., if the initial error $|\widehat{e}_h^{\,0,0}|_1$ is bounded by M, then for some large $\nu > \nu^*$, the error in the smoothing step $|\widehat{e}_h^{\,j,i}|_1$ can be bounded for $i = 1, 2, \ldots, m$. Also we can prove, with the same approach used in the continuous case, that $|\widehat{e}_h^{\,j,m}| < |\widehat{e}_h^{\,j,0}|_1$ for large $\nu > \nu_1 > \nu^*$. Since ν_1 and ν^* are independent of j, then we can claim that the sequence $|\vec{e}_h^{\,j,m}|_1$ is bounded by the initial error M and converges to zero. Therefore $\widehat{u}_h^{\,j,m}$ tends to the solution of (3.20) for small Reynolds numbers $1/\nu$. ∎

In primitive variables $\widehat{u}_h^{\,j,i} = \vec{w}_h^{\,i} + \widehat{u}_h^{\,j,i-1}$ and p_h^i, the local problem, corresponds to solve

$$\begin{cases} \nu a(\vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + c(\vec{w}_h^{\,i}, \vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + c(\vec{w}_h^{\,i}, \widehat{u}_h^{\,i-1}, \vec{v}_h^{\,i}) + c(\widehat{u}_h^{\,i-1}, \vec{w}_h^{\,i}, \vec{v}_h^{\,i}) + b(p_h^i, \vec{v}_h^{\,i}) \\ \quad = \langle \vec{f}, \vec{v}_h^{\,i}\rangle - a(\vec{u}_h^{\,i-1}, \vec{v}_h^{\,i}) - c(\widehat{u}_h^{\,i-1}, \widehat{u}_h^{\,i-1}, \vec{v}_h^{\,i}) \qquad \forall \vec{v}_h^{\,i} \in \mathbf{V}_0^h(\Omega_h^i), \\ b(r_h^i, \vec{w}_h^{\,i}) = -b(r_h^i, \vec{u}_h^{\,i-1}) \qquad \forall r_h^i \in P_0^h(\Omega_h^i), \end{cases} \tag{3.28}$$

with homogeneous boundary condition over Γ_h^i. In order to solve the system (3.28) we must solve the pressure at all points of the block and keep the global solution divergence-free. If parabolic or linear finite elements are used in order to satisfy the divergence-free constraint, all the equations that contain the block variables must be solved. Sometimes this is too expensive and the block of type B seems more attractive.

3.3. Stokes problem with block of type B. The Vanka solver with blocks of type A keeps the solution divergence-free at each iteration but has a high number of equations to be solved. The divergence-free constraint on the boundary of the block can be relaxed and the pressure on the border taken from the previous iteration, which leads to a block of type B. In order to prove convergence for this type of block we need to write the Stokes problem without using the divergence-free functions or the divergence-free constraints.

Given $P_0^h(\Omega_h)$ and $\mathbf{X}_0^h(\Omega_h)$ we can define the set $\mathbf{V}_0^{\,h}(\Omega_h)$ of free-divergence vectors and its orthogonal complement with respect to the scalar product $(\cdot,\cdot)_1$ by

$$\mathbf{V}_0^{h\perp}(\Omega_h) = \{\vec{u}_h \in \mathbf{X}_0^h(\Omega_h) : (\vec{u}_h, \vec{v}_h)_1 = 0 \quad \forall \vec{v}_h \in \mathbf{V}_0^h(\Omega_h)\}.$$

Suppose that the functions in the space $\mathbf{X}_0^h(\Omega_h)$ and $\mathbf{V}_0^h(\Omega_h)$ satisfy the approximation properties and in particular the LLB condition. Then the orthogonal $\mathbf{V}_0^{h\perp}(\Omega_h)$ of $\mathbf{V}_0^h(\Omega_h)$ in $\mathbf{X}_0^h(\Omega_h)$ is a well-defined space (see, for example, [11] for a more abstract setting).

For each $\vec{u}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$ there is an associated pressure p_h which is the solution of the following Stokes problem:

$$\begin{cases} a(\vec{u}_h, \vec{v}_h) + b(p_h, \vec{v}_h) = a(\vec{u}_h^{\perp}, \vec{v}_h) & \forall \vec{v}_h \in \mathbf{X}_0^h, \\ b(\vec{u}_h, r_h) = 0 \qquad \forall r_h \in P_0^h(\Omega_h). \end{cases} \tag{3.29}$$

The solution p_h is unique in $P_0^h(\Omega_h)$ and $\vec{u}_h$ is zero. If $\vec{u}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$, the set of all solutions p_h is a subspace of $P_0^h(\Omega_h)$. We introduce the natural space for the discrete pressure by

$$S_0^h(\Omega_h) = \{p_h \in P_0^h : p_h \text{ is a solution of (3.29)} \quad \forall \vec{u}_h \in \mathbf{V}_0^h(\Omega_h)\}.$$

The set $S_0^h(\Omega_h) \subseteq P_0^h(\Omega_h) \subseteq L_0^2(\Omega_h)$ is a space equipped with the norm $\|\cdot\|_{S_h}$ generated by the Schur complement operator $S_h = B_h A_h^{-1} B_h^T$. The Schur complement norm is the natural norm for the set $S_0^h(\Omega_h)$ which is defined by (3.29). The LBB condition assures that $\|\cdot\|_{S_h}$ is a norm (see, for example, [20, 7]). The solution of (3.29) is unique and therefore for each $\vec{u}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$ there is a unique $p_h \in S_0^h(\Omega_h)$. Also for each $p_h \in S_0^h(\Omega_h)$ there is an element $\vec{u}_h^{\perp} \in \mathbf{V}_0^{\perp h}(\Omega_h)$ such that $\vec{u}_h^{\perp}$ is the unique solution of

$$a(\vec{u}_h^{\perp}, \vec{v}_h) = b(p_h, \vec{v}_h) \qquad \forall \vec{v}_h \in \mathbf{X}_0^h. \tag{3.30}$$

Clearly for all $\vec{v}_h \in \mathbf{V}_0^h(\Omega_h)$ we have $(\vec{u}_h^{\perp}, \vec{v}_h)_1 = a(\vec{u}_h^{\perp}, \vec{v}_h) = b(\vec{v}_h, p_h) = 0$, which implies that the $\vec{u}_h^{\perp} \in \mathbf{X}_0^h(\Omega_h)$ is indeed in $\mathbf{V}_0^{h\perp}(\Omega_h)$. Therefore for all $p_h \in S^h(\Omega_h)$, (3.30) associates a vector $\vec{u}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$.

With these spaces we can write the Stokes problem in a different equivalent way; i.e., given $\vec{f} \in H^{-1}(\Omega_h)$, find $(\vec{u}_h^{'}, \vec{u}_h, \vec{u}_h^{\perp}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$, which solves the following system:

$$\begin{cases} \nu a(\vec{u}_h^{'}, \vec{v}_h) = (\vec{f}, \vec{v}_h) \qquad \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h), \\ a(\vec{u}_h^{\perp}, \vec{v}_h^{\perp}) = (\vec{u}_h^{'}, \vec{v}_h^{\perp})_1 \qquad \forall \vec{v}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h), \\ \vec{u}_h^{'} = \vec{u}_h + \vec{u}_h^{\perp}. \end{cases} \tag{3.31}$$

The second equation in (3.31) can be seen as a projection of $\vec{u}_h^{'}$ over the space $\mathbf{V}_0^{h\perp}$ or the Schur complement equation for the pressure. The system in (3.31) is an uncoupled system of equations in the sense that we can solve first for the variable $\vec{u}_h^{'}$, then for the variable $\vec{u}_h^{\perp}$, and finally for $\vec{u}_h$. It is clear that $\vec{u}_h$ is in $\mathbf{V}_0^h(\Omega_h)$ since the second equation in (3.31) implies that

$$(\vec{u}_h, \vec{v}_h^{\perp})_1 = (\vec{u}_h^{'} - \vec{u}_h^{\perp}, \vec{v}_h^{\perp})_1 = (\vec{u}_h^{'} \vec{v}_h^{\perp})_1 - (\vec{u}_h^{\perp}, \vec{v}_h^{\perp})_1 = 0$$

for all $\vec{v}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$. It is also clear that the solution of the second equation cannot be done easily since this involves the construction of basis functions in the space $\mathbf{V}_0^{h\perp}(\Omega_h)$. However, the formulation in (3.31) is very useful from the theoretical point of view, especially in the analysis of algorithms that relax the incompressibility constraints. The equivalence between these two forms of the Stokes system is a consequence of the LBB inf-sup hypothesis [11] and is stated in the following theorem.

THEOREM 3.5. *The solution* $(\vec{u}_h, \vec{u}_h^{\perp}) \in \times \mathbf{V}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$ *in* (3.31) *solves*

$$\begin{cases} \nu a(\vec{u}_h, \vec{v}_h) + b(p_h, \vec{v}_h) = (\vec{f}_h, \vec{v}_h) & \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h), \\ b(r_h, \vec{u}_h) = 0 & \forall r_h \in P_0^h(\Omega_h), \end{cases} \tag{3.32}$$

with p_h *defined by* (3.29)*, and conversely, if* $(\vec{u}_h, p_h) \in \mathbf{V}_0^h(\Omega_h) \times S_0^h(\Omega_h)$ *is a solution of* (3.32)*, then it solves* (3.31) *with* $\vec{u}_h^{\perp}$ *defined by* (3.30).

Proof. If $(\vec{u}_h, \vec{u}_h^{\perp}) \in \mathbf{V}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$ solves (3.31) with p_h defined by (3.29), we have $\vec{u}_h \in \mathbf{V}_0^h$, $b(r_h, \vec{u}_h) = 0$ for all $r_h \in P_0^h(\Omega_h)$, and $\vec{u}_h^{'} = \vec{u}_h + \vec{u}_h^{\perp}$, namely,

$$\nu a(\vec{u}_h^{'}, \vec{v}_h) = \nu a(\vec{u}_h + \vec{u}_h^{\perp}, \vec{v}_h) = \nu a(\vec{u}_h, \vec{v}_h) + \nu a(\vec{u}_h^{\perp}, \vec{v}_h) = \nu a(\vec{u}_h, \vec{v}_h) + b(p_h, \vec{u}_h)$$

which implies that $(\vec{u}_h, p_h)$ solves (3.32).

Conversely, if $(\vec{u}_h, p_h)$ solves (3.32) with (3.30) and $\vec{u}_h^{'} = \vec{u}_h + \vec{u}_h^{\perp}$, then from the first equation we obtain

$$a(\vec{u}_h^{'}, \vec{v}_h) = (\vec{f}, \vec{v}_h) \qquad \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h). \tag{3.33}$$

If we set $\vec{v}_h = \vec{v}_h^{\perp} \in \mathbf{V}_0^{\perp}(\Omega_h)$, the first equation in (3.32) and the above equation give

$$a(\vec{u}_h^{\perp}, \vec{v}_h^{\perp}) = (\vec{u}_h, \vec{v}_h^{\perp})_1 \qquad \forall \vec{v}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h), \tag{3.34}$$

which proves the theorem. ∎

The formulation in (3.31) can be seen as a simple change of the variable p_h in the Stokes formulation. There is no advantage in this formalism from the numerical point of view, but in the formulation (3.31) the system can now be solved in sequence and the divergence-free constraint can be relaxed during the iteration.

In order to prove convergence for our algorithm we should be able to decompose the global problem into a sequence of local subproblems over the projection subspaces. First we state a space decomposition theorem which allows us to search the solution of the original Stokes problem in the projection spaces.

THEOREM 3.6. *Let* Ω_h *be a bounded simply connected domain with polygonal boundary and let* Ω_h^i, $i = 1, 2, \ldots, m$, *be a sequence of overlapping subdomains obtained by clustering blocks of elements of type A or B such that* $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cdots \Omega_h^m$ *and* $\mathbf{X}_0^1(\Omega_h) = \mathbf{X}_0^h(\Omega_h^1) + \mathbf{X}_0^h(\Omega_h^2) + \mathbf{X}_0^h(\Omega_h^3) \cdots \mathbf{X}_0^h(\Omega_h^m)$. *Let* $\mathbf{V}_0^{h\perp}(\Omega_h^i)$ *be the orthogonal space to the divergence-free function space* $\mathbf{V}_0^h(\Omega_h^i)$ *in* $\mathbf{X}_0^h(\Omega_h^i)$ *with respect to the inner product* $(\cdot, \cdot)_1$ *over* Ω_h^i. *Then we have that* $\mathbf{V}_0^{h\perp}(\Omega_h) = \mathbf{V}_0^{h\perp}(\Omega_h^1) + \mathbf{V}_0^{h\perp}(\Omega_h^2) + \mathbf{V}_0^{h\perp}(\Omega_h^3) + \cdots + \mathbf{V}_0^{h\perp}(\Omega^m)$.

Proof. First consider the case for $m = 2$. Let $\vec{u}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h)$. Since the domains are overlapping in the space $\mathbf{X}_0^h(\Omega_h)$ we can write $\vec{u}_h^{\perp} = \vec{u}_1 + \vec{u}_2$ with $\vec{u}_i \in \mathbf{X}_0^h(\Omega_h^i)$, $i = 1, 2$. The vector $\vec{u}_1$ can be decomposed in $\mathbf{X}_0^h(\Omega_h^1)$ as $\vec{u}_1 = \vec{u}_1^{\perp} + \vec{u}_1^o$, with $\vec{u}_1^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h^1)$ and $\vec{u}_1^o \in \mathbf{V}_0^h(\Omega_h^1)$.

Consider the zero extension $\vec{w}_1^{\perp}$ and $\vec{w}_1^o$ to Ω^h of $\vec{u}_1^{\perp}$ and $\vec{u}_1^o$. The vectors $\vec{w}_1^{\perp}$ and $\vec{w}_1^o$ are in $\mathbf{V}_0^{h\perp}(\Omega_h)$ and $\mathbf{V}_0^h(\Omega_h)$, respectively. The extension of $\vec{w}_1$ of $\vec{u}_1$ to Ω_h can be written as $\vec{w}_1 = \vec{w}_1^{\perp} + \vec{w}_1^o$.

In a similar way we can decompose $\vec{u}_2 = \vec{u}_2^{\perp} + \vec{u}_2^o$ with $\vec{u}_2^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h^2)$ and $\vec{u}_1^o \in \mathbf{V}_0^h(\Omega_h^1)$. The extension of $\vec{w}_2$ of $\vec{u}_2$ to Ω_h can be written as $\vec{w}_2 = \vec{w}_2^{\perp} + \vec{w}_2^o$, where $\vec{w}_2^{\perp}$ and $\vec{w}_2^o$ are the corresponding extensions. From the hypothesis we have that $\vec{u}_h^{\perp} = \vec{u}_1 + \vec{u}_2$, and therefore $\vec{u}_2 = \vec{u}_h^{\perp} - \vec{u}_1$ and $\vec{w}_2 = \vec{u}_h^{\perp} - \vec{w}_1$. The unique decomposition of $\vec{w}_2$ over the subspaces $\mathbf{V}_0^{h\perp}(\Omega_h)$ and $\mathbf{V}_0^h(\Omega_h)$ implies that $\vec{u}_2^o = -\vec{u}_1^o$ and that $\vec{u}_1^o = \vec{w}_1^o$ is zero over $\Omega_h - (\Omega_h^1 \cap \Omega_h^2)$. Therefore $\vec{u}_1^{\perp} = \vec{u}_1 - \vec{u}_1^o \in \mathbf{V}_0^{h\perp}(\Omega_h^1)$

and $\vec{u}_2^{\perp} = \vec{u}_h^{\perp} - \vec{u}_1 + \vec{u}_1^{\,o} \in \mathbf{V}_0^{h\perp}(\Omega_h^2)$ gives the desired decomposition $\vec{u}_h^{\perp} = \vec{u}_1^{\perp} + \vec{u}_2^{\perp}$. For any m the theorem can be proved by induction using the case $m = 2$ and standard techniques [15, 16, 11]. ∎

Let T_i^h, P_i^h, and Π_i^h denote the orthogonal projections from $\mathbf{X}_0^h(\Omega_h)$, $\mathbf{V}_0^h(\Omega_h)$, and $\mathbf{V}_0^{h\perp}(\Omega_h)$ onto $\mathbf{X}_0^h(\Omega_h^i)$, $\mathbf{V}_0^h(\Omega_h^i)$, and $\mathbf{V}_0^{h\perp}(\Omega_h^i)$ with respect to the inner product $(\cdot,\cdot)_1$ defined by the operator $a(\cdot,\cdot)$. The projection $T_i^h\vec{u}_h$, $P_i^h\vec{u}_h$, and $\Pi_i^h\vec{u}_h$ of $\vec{u}_h$ onto $\mathbf{X}_0^h(\Omega_h^i)$, $\mathbf{V}_0^h(\Omega_h^i)$, and $\mathbf{V}_0^{h\perp}(\Omega_h^i)$ is defined by

$$\begin{aligned} &a(T_i^h\vec{u}_h, \vec{v}_h) = a(\vec{u}_h, \vec{v}_h) \qquad \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h^i),\\ &a(P_i^h\vec{u}_h, \vec{v}_h^{\perp}) = a(\vec{u}, \vec{v}_h^{\perp}) \qquad \forall \vec{v}_h^{\perp} \in \mathbf{V}_0^h(\Omega_h^i),\\ &a(\Pi_i^h\vec{u}_h, \vec{v}_h^{\perp}) = a(\vec{u}, \vec{v}_h^{\perp}) \qquad \forall \vec{v}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h^i), \end{aligned}$$

respectively.

It is possible to prove that the iterative application of these operators leads to a contraction in norm $|\cdot|_1$.

THEOREM 3.7. *Let Ω_h be a bounded simply connected domain with polygonal boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type B with boundary such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$. Let T_i^h, P_i^h, and Π_i^h be the orthogonal projections from $\mathbf{X}_0^h(\Omega_h), \mathbf{V}_0^h(\Omega_h)$, and $\mathbf{V}_0^{h\perp}(\Omega_h)$ onto $\mathbf{X}_0^h(\Omega_h^i), \mathbf{V}_0^h(\Omega_h^i)$, and $\mathbf{V}_0^{h\perp}(\Omega_h^i)$, respectively, for $i = 1, 2, \ldots, m$. Then $|\Pi_{i=1}^m(I - T_i)|_1 < 1$, $|\Pi_{i=1}^m(I - P_i)|_1 < 1$ and $|\Pi_{i=1}^m(I - \Pi_i)|_1 < 1$.*

Proof. The proof can follow the approach used in Theorem 2.4 or 3.2. Let $\mathbf{V}^h(\Omega_h)$ be a closed subspace of $\mathbf{X}_0^h(\Omega_h)$ such that $\mathbf{V}^h(\Omega_h) = \mathbf{V}^h(\Omega_h^1) + \mathbf{V}^h(\Omega_h^2) + \cdots + \mathbf{V}^h(\Omega_h^m)$ with $\mathbf{V}^h(\Omega_h^i) \subseteq \mathbf{X}_0^h(\Omega_h^i)$ for $i = 1, 2, \ldots, m$. Here $\mathbf{V}^h$ could be $\mathbf{X}_0^h$, $\mathbf{V}_0^h$, or $\mathbf{V}_0^{h\perp}$. Given $\vec{f} \in H^{-1}(\Omega_h)$, consider the solution $\widehat{u}_h \in \mathbf{V}^h(\Omega_h)$ of the discrete Laplace equation

$$a(\widehat{u}_h, \widehat{v}_h) = \langle \vec{f}, \widehat{v}_h \rangle \qquad \forall \widehat{v}_h \in \mathbf{V}(\Omega_h), \tag{3.35}$$

with homogeneous Dirichlet boundary condition or the equivalent minimization problem

$$\widehat{u}_h = \min_{\widetilde{u} \in \mathbf{V}^h(\Omega_h)} \left(\frac{1}{2}(A\widetilde{u}_h, \widetilde{u}_h) - \langle \vec{f}, \widetilde{u}_h \rangle + \frac{1}{2}\langle A^{-1}\vec{f}, \vec{f} \rangle \right), \tag{3.36}$$

where the constant $\frac{1}{2}\langle A^{-1}\vec{f}, \vec{f}\rangle$ is introduced in order to vanish the functional at the critical point $\widetilde{u}_h = \widehat{u}_h$. The natural approach is to minimize (3.36) over the sequence of domains Ω_h^i in the space $\mathbf{V}^h(\Omega_h^i)$ for $i = 1, 2, \ldots, m$. Let $\widehat{u}_h^{i-1}$ be the solution of the minimization problem over the whole domain Ω_h at step $i - 1$ and let R_i^h be the projector from $\mathbf{V}^h(\Omega_h)$ onto $\mathbf{V}^h(\Omega_h^i)$. By proceeding as in Theorem 2.4 we have $|\Pi_{i=1}^m (I - R_i^h)|_1 < 1$. ∎

Now we are ready to state a convergence theorem for the Stokes problem. If a subdomain of type B is used, the classical Stokes formulation must be modified since the divergence-free constraint is not satisfied at each iteration. The formulation in (3.32) can be used where the solution is written in the variables $\vec{u}_h^{\,'}$ and $\vec{u}_h^{\perp}$ (which represents the pressure distribution).

THEOREM 3.8. *Let Ω_h be a bounded simply connected domain with polygon boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type B such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$. Let the jth smoothing step $(\widehat{u}_h^{\,'j,m}, \widehat{u}_h^{\perp j,m}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$ be defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$ with $\widehat{u}_h^{\,'j,0} = \widehat{u}_h^{\,'j-1,m} \in \mathbf{X}_0^h(\Omega_h)$ and $\widehat{u}_h^{\perp j,0} = \widehat{u}_h^{\perp j-1,m} \in \mathbf{V}_0^{h\perp}(\Omega_h)$.*

Let the global sequence be defined by $\widehat{u}_h^{'j,i} = \vec{w}_h^{'i} + \widehat{u}_h^{'j,i-1}$, $\widehat{u}_h^{\perp j,i} = \vec{z}_h^{\perp i} + \widehat{u}_h^{\perp j,i-1}$ *over* Ω_h^i *and* $\widehat{u}_h^{'j,i} = \widehat{u}_h^{'j,i-1}$, $\widehat{u}_h^{\perp j,i} = \widehat{u}_h^{\perp j,i-1}$ *over* $\Omega_h - \Omega_h^i$ $(i = 1, \ldots, m)$, *where* $(\vec{w}_h^{'i}, \vec{z}_h^{\perp i}) \in \mathbf{X}_0^h(\Omega_h^i) \times \mathbf{V}_0^{h\perp}(\Omega_h^i)$ *solves the local problem*

$$(3.37)\quad \begin{cases} \nu a(\vec{w}_h^{'i}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}^i \rangle - \nu a(\vec{u}_h^{'j,i-1}, \vec{v}_h^i) \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i), \\ a(\vec{z}_h^{\perp i}, \vec{v}_h^{\perp i}) = a(\vec{w}_h^{'i}, \vec{v}_h^{\perp i}) + a(\widehat{u}_h^{'j,i-1}, \vec{v}_h^{\perp i}) - a(\widehat{u}_h^{\perp j,i-1}, \vec{v}_h^{\perp i}) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall \vec{v}_h^{\perp i} \in \mathbf{V}_0^{h\perp}(\Omega_h^i), \end{cases}$$

with boundary condition $\vec{w}_0^{'i} = 0$ *over* Γ_h^i.

Given $(\widehat{u}_h^{'0,0}, \widehat{u}_h^{\perp 0,0}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$, *then the global sequence* $(\widehat{u}_h^{'j,m}, \widehat{u}^{\perp j,m})$ *converges* $(j \to \infty)$ *to the solution of the Stokes problem in* (3.32).

Proof. At the smoothing step j and iteration over Ω^i in (3.37) we have

$$\nu a(\vec{w}_h^{'i}, \vec{v}_h^i) + \nu a(\widehat{u}_h^{'j,i-1}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i).$$

Let $\vec{u}_h^{'}$ be the solution of the Stokes problem in (3.31). Then

$$\nu a(\vec{u}_h^{'}, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i),$$

and therefore

$$(3.38)\qquad \nu a(\vec{w}_h^{'i}, \vec{v}_h^i) + \nu a(\widehat{u}_h^{'j,i-1} - \vec{u}_h^{'}, \vec{v}_h^i) = 0.$$

If we set $\vec{e}_h^{'j,i} = \vec{w}_h^{'i} + \widehat{u}_h^{'j,i-1} - \vec{u}_h^{'}$ over Ω_h^i and $\widehat{e}_h^{'j,i-1} = \widehat{u}_h^{'j,i-1} - \vec{u}_h^{'}$ over Ω_h, the solution of (3.38) is

$$\vec{e}^{\,'j,i} = (I - T_i^h)\, \widehat{e}_h^{'j,i-1}.$$

The global error $\widehat{e}_h^{'j,i}$ over Ω_h satisfies

$$(3.39)\qquad \widehat{e}^{\,'j,i} = (I - T_i^h)\, \widehat{e}_h^{'j,i-1}.$$

Now consider the second equation in (3.37). We have

$$a(\vec{z}^{\perp i} - \vec{w}^{\,'i}, \vec{v}_h^{\perp i}) + a(\widehat{u}^{\perp j,i-1} - \widehat{u}^{\,'j,i-1}, \vec{v}_h^{\perp i}) = 0.$$

Let $\vec{r}_h^{j,i} = \vec{z}^{\perp i} - \vec{w}^{\,'i} \in \mathbf{X}_0^h(\Omega_h^i)$ and $\widehat{r}_h^{j,i-1} = \widehat{u}^{\perp j,i-1} - \widehat{u}^{\,'j,i-1}$ over Ω_h. Then the solution error for $\vec{r}_h^{j,i}$ is given by

$$\vec{r}_h^{j,i} = (I - \Pi_i^h)\, \widehat{r}_h^{j,i-1}.$$

If we extend the function $\vec{r}_h^{j,i}$ to Ω_h^i, then the extension $\widehat{r}_h^{j,i}$ satisfies

$$(3.40)\qquad \widehat{r}_h^{j,i} = (I - \Pi_i^h)\, \widehat{r}_h^{j,i-1}.$$

Since (3.39)–(3.40) are true for all i and j, then

$$(3.41)\qquad \widehat{e}_h^{'j,m} = (I - T_m^h)(I - T_{m-1}^h)(I - T_{m-2}^h) \cdots (I - T_2^h)(I - T_1^h)\, \widehat{e}_h^{'j,0},$$

$$(3.42)\qquad \widehat{r}_h^{j,m} = (I - \Pi_m^h)(I - \Pi_{m-1}^h)(I - \Pi_{m-2}^h) \cdots (I - \Pi_2^h)(I - \Pi_1^h)\, \widehat{r}_h^{j,0}.$$

By using Theorem 3.7 we have $|\widehat{e}_h^{\,'j,m}|_1 < |\widehat{e}_h^{\,'j,0}|_1$ and $|\widehat{r}_h^{j,m}|_1 < |\widehat{r}_h^{j,0}|_1$. Since this is true for all j, then the sequences $|\widehat{e}_h^{\,'j,m}|_1$ and $|\widehat{r}_h^{j,m}|_1$ converge to zero as $j \to \infty$. The sequence $(\widehat{u}_h^{\,'j,m}, \widehat{u}_h^{\perp j,m})$ tends to $(\vec{u}_h^{\,'}, \vec{u}_h^{\perp})$, which is the solution of (3.31). ∎

For elements of type B the Navier–Stokes system must be written in a different form. If we proceed as in the Stokes case, then the problem is reduced to finding $(\vec{u}_h^{\,'}, \vec{u}_h, \vec{u}_h^{\perp}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$, which solves the following system:

$$\begin{cases} \nu a(\vec{u}_h^{\,'}, \vec{v}_h) + c(\vec{u}_h, \vec{u}_h, \vec{v}_h) = (\vec{f}_h, \vec{v}_h) \qquad \forall \vec{v}_h \in \mathbf{X}_0^h(\Omega_h), \\ a(\vec{u}_h^{\perp}, \vec{v}_h^{\perp}) = (\vec{u}_h^{\,'}, \vec{v}_h^{\perp})_1 \qquad \forall \vec{v}_h^{\perp} \in \mathbf{V}_0^{h\perp}(\Omega_h), \\ \vec{u}_h = \vec{u}_h^{\,'} - \vec{u}_h^{\perp}. \end{cases} \tag{3.43}$$

We can solve the Navier–Stokes system above as a sequence of local problems over subdomains of type B. The following theorem states the convergence for small Reynolds numbers.

THEOREM 3.9. *Let Ω_h be a bounded simply connected domain with polygon boundary and let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type B such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$. Let the jth smoothing step $(\widehat{u}_h^{\,'j,m}, \widehat{u}_h^{\perp j,m}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$ be defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$ with $\widehat{u}_h^{\,'j,0} = \widehat{u}_h^{\,'j-1,m} \in \mathbf{X}_0^h(\Omega_h)$ and $\widehat{u}_h^{\perp j,0} = \widehat{u}_h^{\perp j-1,m} \in \mathbf{V}_0^{h\perp}(\Omega_h)$. Let the global sequence be defined by $\widehat{u}_h^{\,'j,i} = \vec{w}_h^{\,'i} + \widehat{u}_h^{\,'j,i-1}$, $\widehat{u}_h^{\perp j,i} = \vec{z}_h^{\perp i} + \widehat{u}_h^{\perp j,i-1}$ over Ω_h^i and $\widehat{u}_h^{\,'j,i} = \widehat{u}_h^{\,'j,i-1}$, $\widehat{u}_h^{\perp j,i} = \widehat{u}_h^{\perp j,i-1}$ over $\Omega_h - \Omega_h^i$ $(i = 1, \ldots, m)$, where $(\vec{w}_h^{\,'i}, \vec{z}_h^{\perp i}) \in \mathbf{X}_0^h(\Omega_h^i) \times \mathbf{V}_0^{h\perp}(\Omega_h^i)$ solves the local problem*

$$\begin{cases} \nu a(\vec{w}_h^{\,'i}, \vec{v}_h^i) + c(\vec{u}_h^i, \vec{u}_h^i, \vec{v}_h^i) = \langle \vec{f}, \vec{v}^{\,i} \rangle - \nu a(\widehat{u}_h^{\,'j,i-1}, \vec{v}_h^i) \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i), \\ a(\vec{z}_h^{\perp i}, \vec{v}_h^{\perp i}) + a(\widehat{u}_h^{\perp j,i-1}, \vec{v}_h^{\perp i}) = (\vec{w}_h^{\,'i}, \vec{v}_h^{\perp i})_1 + (\widehat{u}_h^{\,'j,i-1}, \vec{v}_h^{\perp i})_1 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall \vec{v}_h^{\perp i} \in \mathbf{V}_0^{h\perp}(\Omega_h^i), \\ \vec{u}_h^i = \vec{u}_h^{\,'i} - \vec{u}_h^{\perp i}, \end{cases} \tag{3.44}$$

with boundary condition $\vec{w}_0^{\,'i} = 0$ over Γ_h^i.

Given $(\widehat{u}_h^{\,'0,0}, \widehat{u}^{\perp 0,0})_h \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{V}_0^{h\perp}(\Omega_h)$, then the global sequence $(\widehat{u}_h^{\,'j,m}, \widehat{u}^{\perp j,m})$ converges $(j \to \infty)$ to the solution of the Navier–Stokes problem in (3.43).

Proof. For the first equation the estimate can be obtained by using the same techniques as in the continuous case.

Let $\vec{e}_h^{\,'j,i} = \vec{w}_h^{\,'i} + \widehat{u}_h^{\,'j,i-1} - \vec{u}_h^{\,'}$ in Ω_h^i and $\widehat{e}_h^{\,'j,i-1} = \widehat{u}_h^{\,'j,i-1} - \vec{u}_h^{\,'}$ over Ω_h, where $\vec{u}_h^{\,'}$ is the solution of the problem in (3.43). Also let $\widehat{e}_h^{\,'j,i}$ be the extension of $\vec{e}_h^{\,'j,i}$ to Ω_h when the corresponding zero extension of $\vec{w}_h^{\,'i}$ is considered. Following the same steps as in the continuous case we obtain

$$|\widehat{e}_h^{\,'j,m}|_1 < |\widehat{e}_h^{\,'j,0}|_1 \tag{3.45}$$

for all $j = 1, 2, \ldots$ and for small Reynolds numbers. Consider the second equation in the following form:

$$a(\vec{z}_h^{\perp i} - \vec{w}_h^{\,'i}, \vec{v}_h^{\perp i}) + a(\widehat{u}_h^{\perp j,i-1} - \widehat{u}_h^{\,'j,i-1}, \vec{v}_h^{\perp i}) = 0 \qquad \forall \vec{v}_h^{\perp i} \in \mathbf{V}_0^{h\perp}(\Omega_h^i).$$

Let $\vec{r}_h^{\,j,i} = \vec{z}_h^{\perp i} - \vec{w}_h^{\,'i} + \widehat{u}_h^{\perp j,i-1} - \widehat{u}_h^{\,'j,i-1}$ over Ω_h^i and $\widehat{r}_h^{\,j,i-1} = \widehat{u}_h^{\perp j,i-1} - \widehat{u}_h^{\,'j,i-1}$ over Ω_h. If $\widehat{r}_h^{\,j,i}$ is the extension of $\vec{r}_h^{\,j,i}$ obtained with the zero extension of $\vec{z}_h^{\perp i} - \vec{w}_h^{\,'i}$ over

Ω_h, then, following the same steps as in the continuous case, we obtain

$$|\widehat{r}_h^{\,j,m}|_1 < |\widehat{r}_h^{\,j,0}|_1$$

for small Reynolds numbers. By using standard arguments we can conclude that the errors vanish as j tends to infinity. ∎

The Vanka-type smoother of type B in the standard velocity and pressure variable solves the following local problem over the subregion Ω_h^i of type B.

Given the global solution $(\widehat{u}_h^{i-1}, \widehat{p}_h^{i-1}) \in \mathbf{X}_0^h(\Omega_h) \times P_0^h(\Omega_h)$ obtained after solving and updating the solution over the block Ω_h^{i-1}, we solve for the variable $(\vec{w}_h^i, z_h^i) \in \mathbf{X}_0^h(\Omega_h^i) \times P_0^h(\Omega_h^i)$ the following local problem:

$$\left\{\begin{array}{l} \nu a(\vec{w}_h^i, \vec{v}_h^i) + b(\vec{z}_h^i, \vec{v}_h^i) + c(\vec{w}_h^i + \widehat{u}_h^{i-1}, \vec{w}_h^i + \widehat{u}_h^{i-1}, \vec{v}_h^i) \\ \qquad\qquad = \langle \vec{f}, \vec{v}_h^i \rangle - \nu a(\widehat{u}_h^{i-1}, \vec{v}_h^i) - b(\widehat{p}_h^{i-1}, \vec{v}_h^i) \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i), \\ b(\vec{w}_h^i, r_h) = -b(\widehat{u}_h^{i-1}, r_h) \qquad \forall r_h \in P_0^h(\Omega_h^i), \end{array}\right. \tag{3.46}$$

with boundary condition $\vec{w}_h^i = 0$ over Γ_h^i. The update for the solution $(\widehat{u}_h^{j,i}, \widehat{p}_h^{j,i})$ is defined by $(\vec{w}_h^i + \widehat{u}_h^{i-1}, z_h^i + \widehat{p}_h^{i-1})$ over Ω_h^i and by $(\widehat{u}_h^{i-1}, \widehat{p}_h^{i-1})$ over $\Omega_h - \Omega_h^i$.

We remark that only the node inside the block of element of type B must be solved, and therefore the divergence-free constraint is enforced only inside the domain.

The area close to the boundary Γ_h^i is not divergence-free and there the pressure can take the value of the previous iteration. Only when convergence is reached does the value of the pressure match the global solution and the divergence-free constraint is completely enforced.

4. Computational examples. In order to compute the solution of the Navier–Stokes equation we combine standard finite element techniques in conjunction with multigrid methods. It is well known that Vanka smoothers converge slowly, but the combination of these smoothers with multigrid methods shows that they are extremely competitive in both CPU time and accuracy (see [13, 24]). We would like to use rectangular conforming finite elements (standard Taylor–Hood) over the domain Ω_h and a Vanka-type smoother with blocks of type B. By starting at the multigrid coarse level l_0, we subdivide Ω_h into unstructured families of rectangular meshes, T_h^{i,l_0}. Based on the simple element midpoint refinement, different multigrid levels can be constructed to reach the finest multigrid level l. The unique representations of $\vec{u}_{h_l}$ and p_{h_l} as a function of the nodal point values $\vec{u}_l(k_1, n)$ and $p_l(k_2, n)$ ($k_1 = 1, 2, \ldots, nvt$, with nvt = number of velocity nodal points, and $k_2 = 1, 2, \ldots, npt$, with npt = number of pressure nodal points) define the finite element isomorphisms $\Phi_l : U_l \to X^{h_l}$, $\Psi_l : \Pi_l \to S^{h_l}$ between the vector spaces U_l, Π_l of nvt-dimension and npt-dimension vectors and the finite element spaces X^{h_l}, S^{h_l} at the multigrid level l.

Essential elements of a multigrid algorithm are the velocity and pressure prolongation maps

$$P_{l,l-1}(u) : U_{l-1} \to U_l, \qquad P_{l,l-1}(p) : \Pi_{l-1} \to \Pi_l \tag{4.1}$$

and the velocity and pressure restriction operators

$$R_{l-1,l}(u) = P_{l,l-1}^*(u) : U_l \to U_{l-1}, \qquad R_{l-1,l}(u) = P_{l,l-1}^*(u) : \Pi_l \to \Pi_{l-1}. \tag{4.2}$$

Since we would like to use conforming Taylor–Hood finite element approximation spaces we have the nested finite element hierarchies $X^{h_0} \subseteq X^{h_1} \subseteq \cdots \subseteq X^{h_l}$ and

$S^{h_0} \subseteq S^{h_1} \subseteq \cdots \subseteq S^{h_l}$, and the canonical prolongation maps $P_{l,l-1}(u)$, $P_{l,l-1}(p)$ can be obtained simply by

$$P_{l,l-1}(u) = \Phi_{l-1}(\Phi_l^{-1}(u)), \tag{4.3}$$

$$P_{l,l-1}(p) = \Psi_{l-1}(\Psi_l^{-1}(p)). \tag{4.4}$$

Let Ω_h^i, $i = 1, 2, \ldots, m$, be a sequence of overlapping subdomains of type B such that $\Omega_h = \Omega_h^1 \cup \Omega_h^2 \cup \cdots \cup \Omega_h^m$ and the jth smoothing step be defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$. Given the global solution $(\widehat{u}_h^{i-1}, \widehat{p}_h^{i-1}) \in \mathbf{X}_0^h(\Omega_h) \times P_0^h(\Omega_h)$, obtained after solving and updating the solution over the block Ω_h^{i-1}, we solve for $(\vec{w}_h^i, z_h^i) \in \mathbf{X}_0^h(\Omega_h^i) \times P_0^h(\Omega_h^i)$ the following local problem:

$$\begin{cases} \nu a(\vec{w}_h^i, \vec{v}_h^i) + b(\vec{z}_h^i, \vec{v}_h^i) + c(\vec{w}_h^i, \vec{w}_h^i, \vec{v}_h^i) + c(\vec{w}_h^i, \widehat{u}_h^{i-1}, \vec{v}_h^i) \\ \qquad + c(\widehat{u}_h^{i-1}, \vec{w}_h^i, \vec{v}_h^i) = \langle \vec{f}, \vec{v}_h^i \rangle - \nu a(\widehat{u}_h^{i-1}, \vec{v}_h^i) \\ \qquad - b(\widehat{p}_h^{i-1}, \vec{v}_h^i) - c(\widehat{u}_h^{i-1}, \widehat{u}_h^{i-1}, \vec{v}_h^i) \qquad \forall \vec{v}_h^i \in \mathbf{X}_0^h(\Omega_h^i), \\ b(\vec{w}_h^i, r_h) + b(\widehat{u}_h^{i-1}, r_h) = 0 \qquad \forall r_h \in P_0^h(\Omega_h^i), \end{cases} \tag{4.5}$$

with boundary condition $\vec{w}_h^i = 0$ over Γ_h^i. Then the global sequence is defined by $\widehat{u}_h^{j,i} = \vec{w}_h^i + \widehat{u}_h^{i-1}$, $\widehat{p}_h^{j,i} = z_h^i + \widehat{p}_h^{i-1}$ over Ω_h^i and $\widehat{u}_h^{j,i} = \widehat{u}_h^{i-1}$, $\widehat{p}_h^{j,i} = p_h^{i-1}$ over $\Omega_h - \Omega_h^i$ $(i = 1, \ldots, m)$. The jth smoothing step $(\widehat{u}_h^{j,m}, \widehat{p}_h^{j,m}) \in \mathbf{X}_0^h(\Omega_h) \times \mathbf{S}_0^h(\Omega_h)$ is defined by solving iteratively the local system over the domain Ω_h^i for $i = 1, \ldots, m$, with $\widehat{u}_h^{j,0} = \widehat{u}_h^{j-1,m} \in \mathbf{X}_0^h(\Omega_h)$ and $\widehat{p}_h^{j,0} = \widehat{p}_h^{j-1,m} \in \mathbf{S}_0^h(\Omega_h)$.

We solve the coupled system (4.5) exactly, but preconditioner or other iterative methods can be used. However, we remark that the exact solution or the coupled in velocity-pressure solution of (4.5) allows us to solve unconditionally the Stokes problem. In the remainder of the paper we perform two tests to show the stability of the convergence and the capability for parallel computations.

4.1. Convergence test. In this test we compute the solution of the Navier–Stokes system with different Reynolds numbers and by using the iterative method proposed in the previous section when we impose a velocity and pressure test field defined by

$$\begin{cases} \vec{u} = \sin(\pi x)^2 \, \sin(2\pi y), \\ \vec{v} = -\sin(\pi y)^2 \, \sin(2\pi x), \\ p = xy. \end{cases} \tag{4.6}$$

The results are shown in Tables 4.1 and 4.2. As expected this Vanka-type smoother cannot converge around 1000 Reynolds $(1/\nu)$ and therefore the results are given in the range 0–500 Reynolds. However, with this resolution it is possible to compute a solution beyond 10000 Reynolds if the time-dependent Navier–Stokes equation is solved with reasonable small time steps. Also the range of Reynolds numbers can be substantially increased if a regularization term is included in the formulation. In Tables 4.1 and 4.2 the norms of the error between the computed solution and the true solution in (4.6) are shown. The error for the velocity is given in the L^2 and H^1 norms as a function of different Reynolds numbers and different resolutions. The error in pressure is shown only in the L^2 norm. All the results are obtained by using a Vanka-type smoother with block of type B and a V multigrid cycle. The convergence is considered achieved when the multigrid residual norm reaches 10^{-13}.

TABLE 4.1
Error norm for velocity and pressure fields for 1–10 *Reynolds and different resolutions.*

Re	*Grid*	$L^2(vel)$	$H^1(vel)$	$L^2(p)$
1	8×8	1.65e-3	1.04e-1	2.96e-3
	16×16	2.06e-4	2.56e-2	2.41e-4
	64×64	2.57e-5	6.38e-3	1.88e-5
	128×128	3.22e-6	1.59e-3	1.52e-6
	256×256	4.02e-7	3.99e-4	1.27e-7
10	8×8	1.66e-3	1.04e-1	5.6e-4
	16×16	2.06e-4	2.56e-2	3.92e-5
	64×64	2.57e-5	6.38e-3	2.68e-6
	128×128	3.22e-6	1.59e-3	1.93e-7
	256×256	4.02e-7	3.99e-4	1.48e-8

TABLE 4.2
Error norm for velocity and pressure fields for 100–500 *Reynolds and different resolutions.*

Re	*Grid*	$L^2(vel)$	$H^1(vel)$	$L^2(p)$
100	8×8	2.04e-3	1.23e-1	7.38e-4
	16×16	2.22e-4	2.75e-2	4.26e-5
	64×64	2.62e-5	6.52e-3	2.02e-6
	128×128	3.24e-6	1.60e-3	1.15e-7
	256×256	4.03e-7	3.99e-4	7.11e-9
500	8×8	3.84e-3	2.06e-1	2.57e-3
	16×16	4.23e-4	4.81e-2	2.00e-4
	64×64	3.67e-5	9.01e-3	6.49e-6
	128×128	3.62e-6	1.79e-3	1.85e-7
	256×256	4.16e-7	4.12e-4	1.06e-8

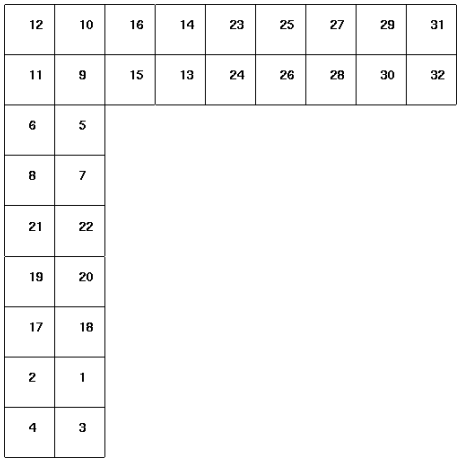


FIG. 4.1. *L-shaped domain configuration at the multigrid low level* l_0.

4.2. L-shaped channel. In this second numerical experiment parallel computations of a flow through an L-shaped channel is presented. The first multigrid level l_0 is the coarse mesh designed to contains all relevant information such as boundary conditions and geometric details and is shown in Figure 4.1. The mesh is an unstructured coarse mesh of rectangular finite elements for P_2/P_1 velocity/pressure representation. The other levels l_i $(i = 1, 2, 3)$ are generated by an unstructured grid generator by

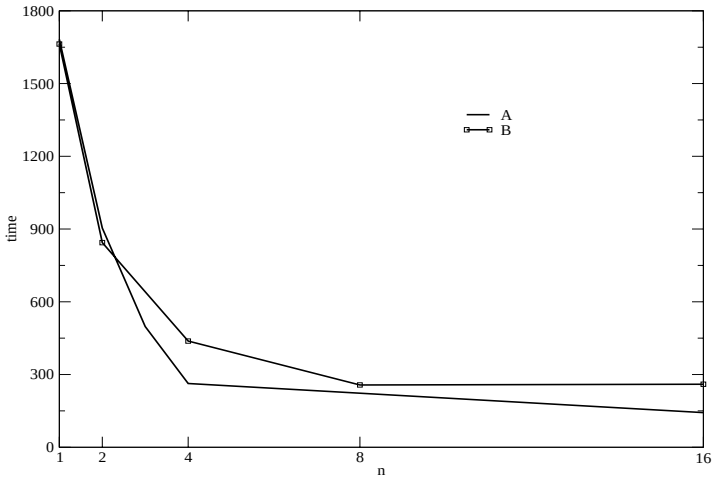


FIG. 4.2. *CPU time (on the left) with exchange at the end of the block relaxation (case A) and at the end of the grid relaxation (case B) as a function of the number n of the CPUs used.*

midpoint refinements. With Vanka-like solvers we can partition the domain and the processor load at the element block over the level l_0 and therefore in a very efficient and flexible way. The processors are distributed in a different way for different cases in order to balance the load and speed up the computations. In this case all the processors are distributed uniformly over the 32 elements of level l_0. Since we have 32 elements in the coarse grid, no more than 32 processors can be used. However, since the communication time should be optimized it is reasonable to use only 16 processors and assign the pair elements along the L-shape to only one processor. We compute and compare the solutions of the problem obtained by using 1, 2, 4, 8, and 16 processors, respectively.

The boundary conditions for this problem are inflow boundary conditions on the bottom with parabolic profile and outflow boundary conditions on the right side. Dirichlet boundary conditions are applied at the rest of the boundary. In the configuration proposed the solution is obtained at the level l_3 by a standard V-cycle multigrid and is stopped when the residual of the linear system is 10^{-13} for the velocity. The reference velocity is 1 m/s, which is the maximum velocity of the parabolic inflow profile at the inlet.

In the Vanka relaxation approach the solution of the multigrid algebraic system requires the solution, block by block, of several small algebraic systems and the iterative update of the solution. We call the solution of this small algebraic system a block relaxation since this operation gives the solutions and the new update for the block unknowns. Since our solving block is based on an element the update between different processor regions can be performed in many ways. In order to minimize the communication among processors the necessary data exchange for the update during the global relaxation can be performed after a fixed number of block relaxations. In this paper we explore the different possibilities by computing the two limiting cases: the data exchange is performed after every element block relaxation (case A) or the data exchange is performed at the end of a global relaxation (grid relaxation) (case B). In Figure 4.2 we show the CPU time and in Figure 4.3 the relative speedup for the parallel computations for these two different communication configurations (cases A

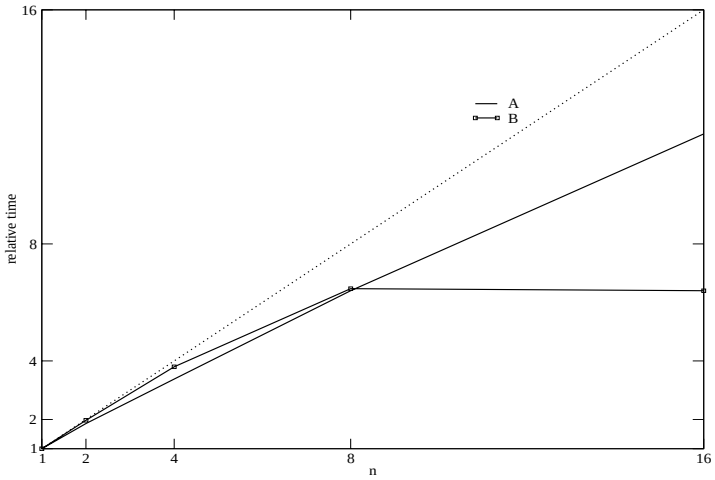


FIG. 4.3. *CPU relative time with exchange at the end of the block relaxation (case A) and at the end of the grid relaxation (case B) as a function of the number n of the CPUs used.*

and B). We note that the use of parallel computing can reduce the time of computation enormously. This test is not intended to show the performance, which can be improved with a real full parallel Vanka smoother, but rather the natural way in which the Vanka-type smoothers can be implemented in the finite element framework. In Figure 4.3 we have the relative speedup as a function of the number of CPUs for cases A and B. All the results are relative to the computation with one single CPU. We note that the relative speedup scales with the number of processors in case A but not in case B. In this case it converges very slowly with this data exchange and the speedup reaches a saturation value. The updating of the solution after a single block relaxation appears to be important for a very fast and regular convergence of the multigrid. However, this could be an effect of the geometry since the domain is divided into very narrow regions. Also we remark that the absolute CPU time is not significantly reduced if the communication is performed at the end of the grid relaxation instead of the end of a single block relaxation. This suggests that the communication time is negligible.

All the computations are performed at 150 Reynolds number in the steady laminar regime. We have performed computation at different Reynolds numbers with similar results.

5. Conclusions. In this paper we have investigated the numerical convergence of a Vanka-type multigrid solver for the Navier–Stokes equations based on the iterative solution of several problems over small overlapping domains. In each iteration step, this smoother requires the solution of several small local subproblems over finite element blocks. We prove that for a particular choice of the block, this method leads to a monotonic convergent algorithm for the steady Stokes problem and that the same algorithm applied to the nonlinear Navier–Stokes equations converges for relative small Reynolds numbers. It is shown that for particular choices for the blocks, the algorithm always converges to the solution of the Stokes problem and, under suitable conditions, to the solution of the Navier–Stokes problem.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] E. Aulisa, S. Manservisi, and P. Seshaiyer, *A non-conforming finite element method for fluid simulations*, in Proceedings of ECCOMAS 2004, Vol. 2, P. Neittaanmaki, T. Rossi, S. Korotov, E. Onate, J. Periaux, and D. Knorzer, eds., University of Jyväskylä, Jyväskylä, Finland, 2004, pp. 2–21.

[3] E. Aulisa and S. Manservisi, *A multigrid approach to optimal control computations for Navier-Stokes flows*, in Robust Optimization-Directed Design, Nonconvex Optim. Appl. 81, Springer-Verlag, New York, 2006, pp. 3–23.

[4] E. Aulisa, S. Manservisi, and P. Seshaiyer, *A computational multilevel approach for solving* 2*D Navier-Stokes equations over non-matching grids*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 6239–6257.

[5] E. Aulisa, S. Manservisi, and P. Seshaiyer, *A non-conforming computational methodology for modeling coupled problems*, Nonlinear Anal., 63 (2005), pp. 1445–1454.

[6] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp., 57 (1991), pp. 1–21.

[7] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.

[8] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978; reprinted, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[9] M. Dryja and Hackbusch, *On the nonlinear domain decomposition method*, BIT, 37 (1997), pp. 296–311.

[10] M. Fatone, P. Gervasio, and A. Quarteroni, *Multimodels for incompressible flows*, J. Math. Fluid Mech., 2 (2000), pp. 126–150.

[11] V. Girault and P. A. Raviart, *The Finite Element Method for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.

[12] V. John, *A comparison of parallel solvers for the incompressible Navier–Stokes equations*, Comput. Visual. Sci., 1 (1999), pp. 193–200.

[13] V. John and L. Tobiska, *Numerical performance of smoothers in coupled multigrid methods for the parallel solution of the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 33 (2000), pp. 453–473.

[14] V. John and L. Tobiska, *A coupled multigrid method for nonconforming finite element discretizations of the* 2*D-Stokes equation*, Computing, 64 (2000), pp. 307–321.

[15] P.-L. Lions, *On the Schwarz alternating method.* I, in Proceedings of the 1st International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, 1988, pp. 1–42.

[16] S. H. Lui, *On Schwarz alternating methods for the incompressible Navier–Stokes equations*, SIAM J. Sci. Comput., 22 (2001), pp. 1974–1986.

[17] S. H. Lui, *On Schwarz alternating methods for the incompressible Navier-Stokes equations in N dimensions*, in Proceedings of the 11th International Conference on Domain Decomposition Methods, C.-H. Lai, P. E. Bjørstad, M. Cross, and B. Widlund, eds., 1999, pp. 65–72.

[18] S. Manservisi, *An extended domain method for optimal boundary control for Navier-Stokes equations*, Internat. J. Numer. Anal. Modeling, to appear.

[19] J. Molenaar, *A two-grid analysis of the combination of mixed finite elements and Vanka-type relaxation,* in Multigrid Methods III, Internat. Ser. Numer. Math. 98, W. Hackbusch and U. Trottenberg, eds., Birkhäuser, Basel, 1991, pp. 313–323.

[20] M. A. Olshanskii, *An Iterative Solver for Oseen Problem and Numerical solution of Incompressible Navier-Stokes Equations*, Report 9827, University of Nijmegem, The Netherlands, 1998.

[21] J. Schoberl and W. Zulehner, *On Additive Schwarz-Type Smoothers for Saddle Point Problems*, SFB-Report 01-20, Johannes Kepler Universität Linz, Linz, Austria, 2001.

[22] X.-C. Tai and M. Espedal, *Rate of convergence of some space decomposition methods for linear and nonlinear problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1558–1570.

[23] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1979.

[24] S. Turek, *Multilevel Pressure Schur Complement Techniques for the Numerical Solution of the Incompressible Navier–Stokes Equations*, Habilitation Thesis, University of Heidelberg, Heidelberg, Germany, 1997.

[25] S. Vanka, *Block-implicit multigrid solution of Navier-Stokes equations in primitive variables*, J. Comput. Phys., 65 (1986), pp. 138–158.

[26] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Society, 15 (2002), pp. 573–597.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2057–2081

FIRST-ORDER SYSTEM LEAST SQUARES FOR GEOMETRICALLY NONLINEAR ELASTICITY*

T. A. MANTEUFFEL†, S. F. MCCORMICK†, J. G. SCHMIDT‡, AND C. R. WESTPHAL§

Abstract. We present a first-order system least-squares (FOSLS) method to approximate the solution to the equations of geometrically nonlinear elasticity in two dimensions. With assumptions of regularity on the problem, we show H^1 equivalence of the norm induced by the FOSLS functional in the case of pure displacement boundary conditions as well as local convergence of Newton's method in a nested iteration setting. Theoretical results hold for deformations satisfying a small strain assumption, a set we show to be largely coincident with the set of deformations allowed by the model. Numerical results confirm optimal multigrid performance and finite element approximation rates of the discrete functional with a total work bounded by about 25 fine-grid relaxation sweeps.

Key words. least squares, elasticity, finite element, nonlinear, multigrid

AMS subject classifications. 74B20, 65N12, 65N30, 65F10

DOI. 10.1137/050628027

1. Introduction.

The primary goal in the study of elasticity is to model the deformation of an elastic body under applied forces, including both internal body forces, such as gravity, and applied surface tractions. For simplicity, we consider forces whose associated density per unit volume is independent of the deformation. Under these applied forces, the elastic body is said to occupy the deformed configuration and, in the absence of forces, the reference configuration. With this in mind, we may think of the central problem as one of finding the mapping from the reference configuration to the deformed configuration. We refer to this mapping function as the deformation and to the Jacobian of the map as the deformation gradient. Two tensor-valued physical quantities are also of interest: strain and stress. The strain tensor, a completely geometrical quantity, is purely a measure of deviation from the reference configuration, while the stress tensor is directly related to the internal force density across the deformed configuration. While the deformation itself is usually the primary unknown in the study of elasticity, the resulting stress and strain are often of interest as well. In this case, the solution methodology we describe in this paper has a distinct advantage over more traditional approaches.

The partial differential equations that are commonly used to govern the deformation are composed of two main components: the equilibrium equation and a constitutive equation. The equilibrium equation and associated boundary conditions relate a balance of forces in the deformed configuration. But, since the deformed configuration is unknown, the equation is mapped back to the reference configuration. The necessity of this mapping introduces a source of nonlinearity into the equations of elasticity.

*Received by the editors March 29, 2005; accepted for publication (in revised form) April 24, 2006; published electronically October 24, 2006.
http://www.siam.org/journals/sinum/44-5/62802.html
†Department of Applied Mathematics, University of Colorado, Campus Box 526, Boulder, CO 80309-0526 (tmanteuf@colorado.edu, stevem@colorado.edu).
‡C&C Research Laboratories, NEC Europe Ltd., Rathausallee 10, 53757 Sankt Augustin, Germany (schmidt@ccrl-nece.de).
§Department of Mathematics and Computer Science, Wabash College, P.O. Box 352, Crawfordsville, IN 47933 (westphac@wabash.edu).

The constitutive equation, or material law, as it is sometimes called, relates the stress to the strain, taking the material properties into account. In general, a material law may be designed for a specific material in a specific range of deformations, as is often the case in applications. There can be as many material laws as materials, but we focus here on a general two-parameter linear relationship between the stress and strain. When this approximation is valid for homogeneous, isotropic materials, we call them St. Venant–Kirchhoff materials. To understand the general behavior of the elasticity system, such materials are considered exclusively.

The model we have described here is both three-dimensional and nonlinear. In this paper, we consider the plane strain model of two-dimensional elasticity, which retains the same character as the full three-dimensional problem both physically and mathematically. It is common to linearize this problem about the reference configuration. However, inherent in the linearization of this naturally nonlinear model is the additional assumption that the displacement is small. There are many applications in which this is a valid assumption and the resulting solution remains sufficiently accurate. For example, a structure whose displacement is magnitudes of order smaller than the structure itself may be accurately modeled by this linear approximation. However, when the small displacement assumption is unreasonable, the partial differential equations of linear elasticity should be used with caution. For this reason, we choose to study a more realistic problem.

In [5, 6, 8, 16], the first-order system least-squares (FOSLS) method is applied to the equations of linear elasticity using the displacement gradient as a new variable. A suitable least-squares functional is minimized over finite element subspaces of H^1. This method allows for the displacement gradient and displacement to be approximated in a two-stage algorithm, with full H^1 control on all variables when the solution is sufficiently smooth. More recent methods, developed in [4, 9, 10], use the stress and displacement as primary unknowns for linear elasticity. The stress, which for linear elasticity is naturally in $H(\mathrm{div})$, is approximated in an $H(\mathrm{div})$ conforming space, thereby avoiding the need to consider effects of boundary singularities. Results from these studies show that a least-squares formulation can be effective for elasticity problems.

This leads us to consider a least-squares method for the geometrically nonlinear model of elasticity that relaxes the small displacement assumption while retaining a linear material law, thus widening the scope of problems that can be effectively treated by least-squares methods. In this model, a linear stress-strain relationship is assumed, but the full nonlinear strain-displacement relationship is preserved. Such a formulation is accurate for the "large displacement, small strain" cases. While not necessarily the best model to use for a given material or for configurations with large strain, this is a common model for elastic materials, and certainly more accurate than linear elasticity. See [11] for further background on elasticity theory.

Our general approach is to linearize the equations of elasticity about a current approximation by Newton's method, to reformulate the resulting linear problem as a well-posed least-squares minimization problem, and to let its minimizer become the new approximation. The reference configuration (i.e., zero displacement) is always taken to be the initial approximation. Thus, the first Newton step reduces to the equations of linear elasticity and subsequent steps are corrections thereof. Since the constitutive equation involves products of the unknowns, we focus on using the displacement gradient as the new dependent variable. The stress and strain tensors are then just simple combinations of this new dependent variable and can be computed

in a postprocessing stage with no loss of accuracy. Each Newton step is cast as an appropriate first-order system, and the associated least-squares functional is minimized over an appropriate finite element subspace of $H^1(\Omega)$.

Our approach also employs a two-stage solution process. The first stage solves for the displacement gradients, while the second stage recovers the actual displacement vector. This decoupling of the unknowns in stages is desirable for several reasons. First, when the primary interest is in the stress or strain, the second stage does not need to be performed. Second, if the problem requires several Newton steps, the deformations can be retrieved after the first stage converges. Third, this approach obviates the need to determine relative weights for the stages if they are incorporated into a single functional. Finally, decoupling the variables is somewhat more efficient than solving for them simultaneously.

We define the term H^1 ellipticity to mean H^1 equivalence with the norm induced by the homogeneous FOSLS functional. The focus of much of this paper is on the formulation and efficiency of the first-stage algorithm by establishing H^1 ellipticity of the FOSLS functional for a general linearization step. The second stage is essentially a coupled Poisson problem that is ideally suited for FOSLS and already discussed in some detail in [6, 8].

2. Notation. Throughout this paper, we refer to our Newton-FOSLS algorithm as *linearized elasticity* (linearized about a current approximation) and to the first Newton step as *linear elasticity* (linearized about the reference configuration). This is not strictly standard convention, but one we find convenient in what follows.

Vector $\mathbf{u}$ and matrix $\mathbf{U}$ are represented componentwise by

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}.$$

The gradient of scalar p and vector $\mathbf{u}$ are given by

$$\nabla p = \begin{pmatrix} \partial_x p \\ \partial_y p \end{pmatrix} \quad \text{and} \quad \nabla \mathbf{u} = \begin{pmatrix} \partial_x u_1 & \partial_y u_1 \\ \partial_x u_2 & \partial_y u_2 \end{pmatrix}.$$

Define the respective divergence, curl, and trace operators by

$$\nabla \cdot \mathbf{u} = \partial_x u_1 + \partial_y u_2, \quad \nabla \cdot \mathbf{U} = \begin{pmatrix} \partial_x U_{11} + \partial_y U_{12} \\ \partial_x U_{21} + \partial_y U_{22} \end{pmatrix},$$
$$\nabla \times \mathbf{U} = \begin{pmatrix} \partial_x U_{12} - \partial_y U_{11} \\ \partial_x U_{22} - \partial_y U_{21} \end{pmatrix}, \quad \text{and} \quad \operatorname{tr}(\mathbf{U}) = U_{11} + U_{22}.$$

Also, denoting the formal adjoint of the curl operator by $\nabla^\perp$, we define

$$\nabla^\perp p = \begin{pmatrix} \partial_y p \\ -\partial_x p \end{pmatrix} \quad \text{and} \quad \nabla^\perp \mathbf{u} = \begin{pmatrix} \partial_y u_1 & -\partial_x u_1 \\ \partial_y u_2 & -\partial_x u_2 \end{pmatrix}.$$

We extend the respective outward unit normal and counterclockwise unit tangential operators, $\mathbf{n}\cdot$ and $\boldsymbol{\tau}\cdot$, componentwise to block column vectors and matrices in the natural way:

$$\mathbf{n} \cdot \mathbf{U} = \begin{pmatrix} n_x U_{11} + n_y U_{12} \\ n_x U_{21} + n_y U_{22} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\tau} \cdot \mathbf{U} = \begin{pmatrix} \tau_x U_{11} + \tau_y U_{12} \\ \tau_x U_{21} + \tau_y U_{22} \end{pmatrix}.$$

We also note that $n_x = \tau_y$ and $n_y = -\tau_x$, and that $\mathbf{n} \cdot \nabla = -\boldsymbol{\tau} \cdot \nabla^\perp$.

We use standard notation for Sobolev spaces $H^k(\Omega)^d$, corresponding inner product $(\cdot,\cdot)_{k,\Omega}$, and norm $\|\cdot\|_{k,\Omega}$ for $k \geq 0$. We drop subscript Ω and superscript d when the domain and dimension are clear by context. For noninteger k, $H^k(\Omega)$ is the interpolation space between $H^{\lfloor k \rfloor}(\Omega)$ and $H^{\lceil k \rceil}(\Omega)$ as in [17]. The case of $k = 0$ corresponds to the Lebesgue measurable space, $L^2(\Omega)$, in which case we generally denote the norm and inner product by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, respectively. Define the subspaces of $L^2(\Omega)$ induced by the divergence and the curl of vector $\mathbf{u}$ by

$$H(\mathrm{div}) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \cdot \mathbf{u}\| < \infty\},$$
$$H(\mathrm{curl}) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \times \mathbf{u}\| < \infty\},$$

with norms

$$\|\mathbf{u}\|^2_{H(\mathrm{div})} = \|\mathbf{u}\|^2 + \|\nabla \cdot \mathbf{u}\|^2,$$
$$\|\mathbf{u}\|^2_{H(\mathrm{curl})} = \|\mathbf{u}\|^2 + \|\nabla \times \mathbf{u}\|^2.$$

Denote by $C^k(\Omega)$ the space of k times continuously differentiable functions on Ω, an open set in $\mathbb{R}^2$. The boundary of Ω, denoted by $\partial\Omega$, is of class C^k if it satisfies the conditions of a Lipschitz boundary (see [20]) and is the union of the graphs of a finite number of C^k functions. We say that $\partial\Omega$ is a $C^{k,l}$ boundary when it is Lipschitz and is the graph of the union of a finite number of Hölder continuous $C^{k,l}$ functions.

We also make use of the following general inequalities:

$$|a|^2 + |b|^2 \leq |a+b|^2 \leq 2(|a|^2 + |b|^2). \tag{2.1}$$

3. The nonlinear problem. Let Ω be a bounded open connected subset of $\mathbb{R}^2$ with boundary $\partial\Omega$, which is partitioned into displacement, Γ_D, and traction, Γ_T, segments ($\bar{\Gamma}_D \cup \bar{\Gamma}_T = \partial\Omega$ and $\Gamma_D \cap \Gamma_T = \emptyset$). For simplicity, we assume that the displacements vanish on Γ_D, as is often the case in practice. The geometrically nonlinear elasticity equations may be written as

$$\begin{cases} \nabla \cdot [(\mathbf{I} + \nabla \mathbf{u})\Sigma] = \mathbf{f} & \text{in} \quad \Omega, \\ \mathbf{n} \cdot [(\mathbf{I} + \nabla \mathbf{u})\Sigma] = \mathbf{g} & \text{on} \quad \Gamma_T, \\ \mathbf{u} = \mathbf{0} & \text{on} \quad \Gamma_D, \end{cases} \tag{3.1}$$

where the material law,

$$\Sigma = \Sigma(\mathbf{E}) = \lambda \mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E}, \tag{3.2}$$

is the second Piola–Kirchhoff stress tensor and

$$\mathbf{E} = \mathbf{E}(\nabla \mathbf{u}) = \frac{\mathbf{1}}{\mathbf{2}}(\nabla \mathbf{u} + \nabla \mathbf{u}^{\mathbf{t}} + \nabla \mathbf{u}^{\mathbf{t}} \nabla \mathbf{u}) \tag{3.3}$$

is the Green–St. Venant strain tensor. This problem is often referred to as one for a St. Venant–Kirchhoff material. Again, this describes materials in configurations in which the "large displacement, small strain" assumption is valid.

We may also separate the linear and nonlinear parts of the first equation in (3.1) and write it as

$$\mu \Delta \mathbf{u} + (\lambda + \mu) \nabla \nabla \cdot \mathbf{u} + \nabla \cdot \mathbf{P_3}(\nabla \mathbf{u}) = \mathbf{f}, \tag{3.4}$$

where $\Delta\mathbf{u} = \nabla \cdot \nabla\mathbf{u}$ is the vector Laplacian of $\mathbf{u}$ and $\mathbf{P_3}(\nabla\mathbf{u})$ is the following matrix of degree 3 polynomials of the components of $\nabla\mathbf{u}$:

$$\mathbf{P_3}(\mathbf{X}) = \frac{1}{2}\lambda\left(\mathrm{tr}(\mathbf{X}^t\mathbf{X})\mathbf{I} + \mathrm{tr}(\mathbf{X} + \mathbf{X}^t + \mathbf{X}^t\mathbf{X})\mathbf{X}\right) + \mu\left(\mathbf{X}^2 + \mathbf{X}^t\mathbf{X} + \mathbf{X}\mathbf{X}^t + \mathbf{X}\mathbf{X}^t\mathbf{X}\right).$$

The linear part of the left-side operator in (3.4) is simply the linear elasticity equations, and the nonlinear part can be thought of as a perturbation that begins to dominate as $\nabla\mathbf{u}$ becomes large compared to $\mathbf{u}$.

The unknown, $\mathbf{u}$, is the usual displacement vector. We assume that the Lamé constants, λ and μ, are bounded by satisfying $0 < \mu_0 < \mu < \mu_1$ and $0 < \lambda_0 < \lambda < \lambda_1$, for appropriate positive bounds. Physically, this corresponds to an assumption of compressibility of the material. The more difficult problem of incompressible materials is considered for linear elasticity in [5, 6, 8, 16]. A complete study of the geometrically nonlinear elasticity problem in a least-squares context in the incompressible limit remains an open problem. Without loss of generality, we scale the problem so that $\mu = 1$ and let λ determine the level of compressibility. See section 11 for examples of Lamé constants for different materials.

The case where $\Gamma_T = \emptyset$ corresponds to a pure displacement problem, $\Gamma_D = \emptyset$ a pure traction problem, and otherwise a mixed boundary condition problem.

4. Existence and uniqueness of solutions. In this section, we establish existence and uniqueness results that confirm well-posedness of system (3.1). We restrict ourselves here to the pure displacement problem on domains with sufficiently smooth data and boundaries (see Remark 4.3 at the end of this section).

Let ∂ represent either first partial derivative, ∂_x or ∂_y, and suppose $\delta > 0$ and $k \geq 0$. The following lemma addresses smoothness of products of functions in $H^{1+\delta}(\Omega)$ and $H^{1+k}(\Omega)$.

LEMMA 4.1. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^2$. Then there exists a constant, C, depending only on Ω, such that, for $u \in H^{1+\delta}(\Omega)$ and $v \in H^{1+k}(\Omega)$, the product uv satisfies*

$$\|uv\|_{1+k} \leq C\|u\|_{1+\delta}\|v\|_{1+k},$$
$$\|\partial(uv)\|_k \leq C\|u\|_{1+\delta}\|v\|_{1+k}.$$

Proof. This is a consequence of the Sobolev imbedding theorem, and a proof can be seen in Chapter 1 of [14]. ∎

The following theorem establishes criteria for existence and uniqueness of solutions to problem (3.1).

THEOREM 4.2. *Let Ω be a domain in $\mathbb{R}^2$ with boundary of class C^{2+m} for some $m > 0$. Then there exists a neighborhood, $\mathcal{Q}_0^m$, of the origin in $H^m(\Omega)$ and a neighborhood, $\mathcal{U}_0^{1+m}$, of the origin in $\mathcal{U}^{1+m} = \{\nabla\mathbf{v} : \mathbf{v} \in H^{2+m}(\Omega), \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\} \subset H^{1+m}(\Omega)$ such that for each $\mathbf{f} \in \mathcal{Q}_0^m$, the boundary value problem*

$$\mathcal{L}(\nabla\mathbf{u}) := \nabla \cdot [(\mathbf{I} + \nabla\mathbf{u})\Sigma(\mathbf{E}(\nabla\mathbf{u}))] = \mathbf{f} \tag{4.1}$$

has exactly one solution, $\nabla\mathbf{u}^*$, in $\mathcal{U}_0^{1+m}$.

Proof. We observe that nonlinear operator $\mathcal{L}$ maps $\nabla\mathbf{u} \in H^{1+m}(\Omega)$ into $H^m(\Omega)$ by applying Lemma 4.1, and that $\mathcal{L}$ is differentiable between these spaces (in fact, all derivatives of order ≥ 4 are zero).

Since $\mathcal{L}(\mathbf{0}) = \mathbf{0}$, we can then apply the implicit function theorem in a neighborhood of the origin in $\mathcal{U}^{1+m} \times H^m(\Omega)$. Thus, we now only need to check that the

derivative of $\mathcal{L}$ at the origin, $\mathcal{L}'(\mathbf{0})$, is bijective between $\mathcal{U}^{1+m}$ and $H^m(\Omega)$ and has continuous inverse.

But $\mathcal{L}'(\mathbf{0})$ is exactly the operator of linear elasticity. It is known that if $\partial\Omega$ is a C^{2+m} boundary and $\mathbf{f} \in H^m(\Omega)$, then there is a unique weak solution to the linear pure displacement problem, $\mathbf{u} \in H^{2+m}(\Omega)$ (see [11]). This immediately implies $\nabla\mathbf{u} \in H^{1+m}(\Omega)$. Thus, we have shown that continuous operator $\mathcal{L}'(\mathbf{0})$ is bijective. Now since $\mathcal{L}'(\mathbf{0})$ is a continuous, bijective, linear map between two Banach spaces, by the closed graph theorem, it must have a continuous inverse.

By the implicit function theorem there is, therefore, a neighborhood, $\mathcal{Q}_0^m$, of the origin in $H^m(\Omega)$ and a neighborhood, $\mathcal{U}_0^{1+m}$, of the origin in $\mathcal{U}^{1+m}$ such that there is a unique solution, $\nabla\mathbf{u}^* \in \mathcal{U}_0^{1+m}$, for any function $\mathbf{f} \in \mathcal{Q}_0^m$. □

Thus, the pure displacement problem with sufficiently smooth data and domain is well-posed, and the solution, $\nabla\mathbf{u}$, remains small in the H^{1+m} norm, with no direct restriction on $\mathbf{u}$ itself. This is consistent with the small strains assumption in the geometrically nonlinear elastic model.

Remark 4.3. We are ultimately interested in nonhomogeneous problems on polygonal domains, which are known to have solutions less smooth than described above. At corner points and/or points of changing boundary condition type on the boundary, a locally weighted norm can be used to remove the effect of the nonsmooth solution. In [19] a weighted-norm least-squares method is developed for problems with boundary singularities. For simplicity, we choose here to focus on the formulation and analysis of the linearized problem since, even for problems lacking full global regularity, we may expect the regularity predicted in Theorem 4.2 away from abruptly changing material interfaces in the interior of Ω for sufficiently smooth data $\mathbf{f}$.

5. Least-squares formulation. We want to replace the nonlinear elasticity problem with a series of linear problems, which we then reformulate as a first-order system. Introducing the deformation, $\boldsymbol{\phi} = \mathbf{x} + \mathbf{u}$, the deformation gradient, $\boldsymbol{\Phi} = \nabla\boldsymbol{\phi}$, and the displacement gradient, $\mathbf{U} = \nabla\mathbf{u}$, we see that problem (3.1) becomes one of finding the zero of

$$\mathcal{F}(\mathbf{U}) = \nabla \cdot \left[\frac{1}{2}\lambda \mathrm{tr}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) + (\mathbf{I} + \mathbf{U})(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})\right] - \mathbf{f} \tag{5.1}$$

subject to the constraint

$$\nabla \times \mathbf{U} = \mathbf{0} \tag{5.2}$$

for $\mathbf{U}$ satisfying appropriate boundary conditions (recall also that we assume $\mu = 1$).

The Fréchet derivative of $\mathcal{F}(\mathbf{U})$ in the direction of $\mathbf{V}$ is

$$\begin{aligned}\mathcal{F}'(\mathbf{U})[\mathbf{V}] = \nabla \cdot \Bigg[&\lambda \mathrm{tr}\left(\mathbf{U} + \frac{1}{2}\mathbf{U}^t\mathbf{U}\right)\mathbf{V} + \lambda \mathrm{tr}(\mathbf{V} + \mathbf{V}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) \\ &+ (\mathbf{I} + \mathbf{U})(\mathbf{V} + \mathbf{V}^t + \mathbf{U}^t\mathbf{V} + \mathbf{V}^t\mathbf{U}) + \mathbf{V}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})\Bigg].\end{aligned} \tag{5.3}$$

Thus, Newton's method for approximating the solution of (5.1) is given by iteratively solving the linear problem

$$\begin{cases} \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}_{n+1}] = \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}_n] - \mathcal{F}(\mathbf{U}_n), \\ \quad \nabla \times \mathbf{U}_{n+1} = \mathbf{0} \end{cases} \tag{5.4}$$

for $\mathbf{U}_{n+1}$, with initial approximation $\mathbf{U}_0 = \mathbf{0}$.

It is convenient to view 2×2 matrices as 4×1 vectors so that general linear operators on such quantities can be written as 4×4 matrices. Thus, define operator $\mathcal{K} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{4\times 1}$ by

$$\mathcal{K}\begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} = (U_{11}, U_{12}, U_{21}, U_{22})^t$$

for any 2×2 matrix, $(\mathbf{U})_{ij} = U_{ij}$. If A is a 4×4 matrix, then the quantity $A\mathbf{U}$ should be interpreted as $A\mathbf{U} = \mathcal{K}^{-1}(A\mathcal{K}\mathbf{U})$.

With the relation $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$, we define the following linear operators:

$$M_1(\mathbf{\Phi}) = \begin{pmatrix} \Phi_{11}^2 & \Phi_{11}\Phi_{12} & \Phi_{11}\Phi_{21} & \Phi_{11}\Phi_{22} \\ \Phi_{12}\Phi_{11} & \Phi_{12}^2 & \Phi_{12}\Phi_{21} & \Phi_{12}\Phi_{22} \\ \Phi_{21}\Phi_{11} & \Phi_{21}\Phi_{12} & \Phi_{21}^2 & \Phi_{21}\Phi_{22} \\ \Phi_{22}\Phi_{11} & \Phi_{22}\Phi_{12} & \Phi_{22}\Phi_{21} & \Phi_{22}^2 \end{pmatrix},$$

$$M_2(\mathbf{\Phi}) = (\Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 + \Phi_{22}^2 - 2)I,$$

$$M_3(\mathbf{\Phi}) = \begin{pmatrix} 3\Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 & 2\Phi_{11}\Phi_{12} + \Phi_{21}\Phi_{22} & 2\Phi_{11}\Phi_{21} + \Phi_{12}\Phi_{22} & \Phi_{12}\Phi_{21} \\ 2\Phi_{11}\Phi_{12} + \Phi_{21}\Phi_{22} & \Phi_{11}^2 + 3\Phi_{12}^2 + \Phi_{22}^2 & \Phi_{11}\Phi_{22} & \Phi_{11}\Phi_{21} + 2\Phi_{12}\Phi_{22} \\ 2\Phi_{11}\Phi_{21} + \Phi_{12}\Phi_{22} & \Phi_{11}\Phi_{22} & \Phi_{11}^2 + 3\Phi_{21}^2 + \Phi_{22}^2 & \Phi_{11}\Phi_{12} + 2\Phi_{21}\Phi_{22} \\ \Phi_{12}\Phi_{21} & \Phi_{11}\Phi_{21} + 2\Phi_{12}\Phi_{22} & \Phi_{11}\Phi_{12} + 2\Phi_{21}\Phi_{22} & \Phi_{12}^2 + \Phi_{21}^2 + 3\Phi_{22}^2 \end{pmatrix}.$$

Using the relation $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$ as a change of variables, define the system matrix, A, as a function of $\mathbf{U}$ by

$$A(\mathbf{U}) = \lambda M_1(\mathbf{I} + \mathbf{U}) + \frac{1}{2}\lambda M_2(\mathbf{I} + \mathbf{U}) + M_3(\mathbf{I} + \mathbf{U}) - \mathbf{I}.$$

In this way, we may denote the linear operator in (5.4) as

$$\mathcal{F}'(\mathbf{U})[\mathbf{V}] = \nabla \cdot A(\mathbf{U})\mathbf{V}.$$

Denoting $A_n = A(\mathbf{U}_n)$ and $\mathcal{F}_n = \mathcal{F}(\mathbf{U}_n)$, the Newton step for the $(n+1)$st iterate $\mathbf{U}$ (dropping the subscript) may now be written as

$$\begin{cases} \nabla \cdot A_n\mathbf{U} = \nabla \cdot A_n\mathbf{U}_n - \mathcal{F}_n, \\ \quad \nabla \times \mathbf{U} = \mathbf{0}. \end{cases} \tag{5.5}$$

We may apply an analogous linearization technique to the traction boundary conditions by defining

$$\mathcal{T}(\mathbf{U}) = \mathbf{n} \cdot \left[\frac{1}{2}\lambda \mathrm{tr}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) + (\mathbf{I} + \mathbf{U})(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})\right] - \mathbf{g}$$

and letting $\mathcal{T}_n = \mathcal{T}(\mathbf{U}_n)$. The corresponding Newton step for the traction boundaries then becomes

$$\mathbf{n} \cdot A_n\mathbf{U} = \mathbf{n} \cdot A_n\mathbf{U}_n - \mathcal{T}_n \quad \text{on} \quad \Gamma_T. \tag{5.6}$$

Since $\mathbf{u} = \mathbf{0}$ on the displacement boundaries, we may enforce the derivative of $\mathbf{u}$ along those boundaries to be zero:

$$\boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0} \quad \text{on} \quad \Gamma_D. \tag{5.7}$$

Thus, we may completely decouple the unknowns in $\mathbf{u}$ from the unknowns in $\mathbf{U}$. We concentrate here on the first-stage solution of $\mathbf{U}$, that is, solving the problem for $\mathbf{U}$ and later recovering $\mathbf{u}$, if necessary.

We take the initial approximation for Newton's method to be the reference configuration, $\mathbf{U}_0 = \mathbf{0}$; the system matrix for the first Newton step is

$$A_0 = \begin{pmatrix} \lambda+2 & 0 & 0 & \lambda \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ \lambda & 0 & 0 & \lambda+2 \end{pmatrix};$$

and we can write $\nabla \cdot (A_0 \mathbf{U}_0) - \mathcal{F}_0 = \mathbf{f}$ and $\mathbf{n} \cdot (A_0 \mathbf{U}_0) - \mathcal{T}_0 = \mathbf{g}$. Thus, we may write the first step of Newton's method as

$$\begin{cases} \nabla \cdot (A_0 \mathbf{U}) = \mathbf{f} & \text{in} \quad \Omega, \\ \nabla \times \mathbf{U} = \mathbf{0} & \text{in} \quad \Omega, \\ \boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0} & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot (A_0 \mathbf{U}) = \mathbf{g} & \text{on} \quad \Gamma_T. \end{cases} \tag{5.8}$$

This is the form of the linear elasticity equations studied in [5, 8].

System (5.5) depends explicitly on the current approximation to the solution. Specifically, matrix A_n deviates from A_0 as $\mathbf{\Phi}$ deviates from the identity (or, as $\mathbf{U}$ deviates from $\mathbf{0}$). Much is known about the first Newton step because it is exactly the linear elasticity case. For example, assuming sufficient smoothness of the solution, a least-squares functional associated with system (5.8) can be shown to be H^1 elliptic with the aid of Korn's inequality. In fact, this ellipticity property is even retained for a modification of system (5.8) in the incompressible limit in [8]. Existence, uniqueness, and optimal finite element approximation bounds immediately follow (see [5, 8]). For the linearized problem, however, the literature reflects relatively little theory in $W^{k,2}$ Sobolev spaces, and a thorough study of these equations in a least-squares context has, to our knowledge, not been explored. Thus, we are led to develop new theory that establishes well-posedness of, and a fast solution technique for, the linearized equations consisting of (5.5), (5.6), and (5.7).

6. Problem modification. One goal of the least-squares methodology is to develop a functional that is H^1 elliptic whenever possible. It is well known that such systems admit uniform and optimal H^1 approximations when using standard finite elements for the discretization and standard multigrid solvers for the resulting linear system (see [7]). For system (5.5), this poses a challenge because the system matrix, A_n, is generally pointwise indefinite. In this section, we introduce a modification to (5.5) that overcomes this difficulty, and we make a reasonable physical assumption that guarantees positive-definiteness of the modified system matrix. To this end, consider modifying A_n by adding to it a matrix of the form

$$B(c) = \begin{pmatrix} 0 & 0 & 0 & c \\ 0 & 0 & -c & 0 \\ 0 & -c & 0 & 0 \\ c & 0 & 0 & 0 \end{pmatrix},$$

where c is any fixed constant. It is easy to see that $\nabla \cdot B(c) \nabla \mathbf{p} = \mathbf{0}$ for any function $\mathbf{p}$, so the solution to (5.5) is unaffected by replacing A_n with $A_n + B(c)$. (We note,

however, that this modification cannot be applied to the traction boundary conditions given in (5.6).) In [5, 8], this idea is applied with $c = \mu = 1$ in conjunction with a rotation of the unknowns so that the equations of linear elasticity in the incompressible limit mirror the Stokes equations. We apply the same idea here, not to transform the equations to a more well known form, but rather to shift the spectrum to be positive. Indeed, in the linear case, the spectrum of A_0, which is $\{0, 2, 2, 2\lambda+2\}$, can be shifted by $B(1)$ so that the spectrum of $A_0 + B(1)$ becomes $\{1, 1, 1, 2\lambda + 3\}$. Numerical experiments on the spectrum of A_n indicate that a choice of $c = 1$ is also most effective for shifting the spectrum to be positive for general deformations. We now study this question analytically.

Matrix $\tilde{A}_n = A_n + B(1)$ seems to depend on the four linearly independent components of $\mathbf{\Phi}_n$. However, under an appropriate change of variables, the eigenvalues can be exactly expressed in terms of just two scalar functions over Ω:

$$\begin{aligned} \sigma &= \Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 + \Phi_{22}^2, \\ \delta &= \Phi_{11}\Phi_{22} - \Phi_{12}\Phi_{21}. \end{aligned} \tag{6.1}$$

In fact, the eigenvalues of $\tilde{A}_n$ are as follows:

$$\begin{aligned} \mathbf{\Lambda}_1 &= \frac{1}{2}(\lambda+2)\sigma - \delta - \lambda, \\ \mathbf{\Lambda}_2 &= \frac{1}{2}(\lambda+2)\sigma + \delta - \lambda - 2, \\ \mathbf{\Lambda}_3 &= \left(\lambda + \frac{3}{2}\right)\sigma - (\lambda+1) - \sqrt{\frac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1}, \\ \mathbf{\Lambda}_4 &= \left(\lambda + \frac{3}{2}\right)\sigma - (\lambda+1) + \sqrt{\frac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1}. \end{aligned} \tag{6.2}$$

That the spectrum can be represented by only two independent quantities is surprising, but that the two quantities have such an obvious physical meaning is remarkable. For example, δ, the determinant of the Jacobian of the mapping of the current approximation, is a local measure of change in volume: $\delta > 1$ indicates areas under tension and $\delta < 1$ indicates areas under compression. Similarly, $\sigma < 2$ when there is significant local compression. In general, we know that in the small strains regime $\sigma \approx 2$ and $0 < \delta \approx 1$.

Since the model for the geometrically nonlinear elasticity equations assumes a deformed configuration with small strains, we may assume small strains of the solution. We show in section 8 that for an initial guess sufficiently close to the solution, each iterate remains bounded near the solution and Newton's method converges. Under these constraints, we take each iterate to satisfy some small strain condition of the form $\|\mathbf{E}\| \ll 1$. We now choose the norm to enforce this condition.

Define the following *Frobenius* norm for tensor-valued quantities:

$$\|\mathbf{X}\|_{Fr}^2 = \sup_{\Omega} \sum_{ij} \left(\mathbf{X}_{ij}\right)^2.$$

Thus, we may write $\|\mathbf{\Phi}\|_{Fr} = \|\sigma\|_\infty$. We can also express the Frobenius norm of the strain tensor exactly in terms of variables σ and δ. We now establish bounds on the strain that guarantee that the modified system matrix is uniformly symmetric positive definite.

Recall that the strain tensor is given by $\mathbf{E}(\mathbf{U}) = \frac{1}{2}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})$. Define

$$\mathcal{S}_\lambda = \left\{ \mathbf{U} : \|\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U}\|_{Fr} < \frac{\sqrt{2}}{\lambda + 3} \right\}$$

as the set of all displacement gradients corresponding to deformations with "small strains." We may choose $\mathcal{Q}_0^m$ small enough to ensure that $\mathbf{f} \in \mathcal{Q}_0^m$ guarantees $\mathbf{U} \in \mathcal{S}_\lambda$. Thus, the condition of small strains follows from the assumptions in Theorem 4.2. We explore the regime of small strains in more detail in section 11.

THEOREM 6.1. *For all* $\mathbf{U} \in \mathcal{S}_\lambda$, *matrix* $\tilde{A} = A(\mathbf{U}) + B(1)$ *is uniformly positive definite over* Ω.

Proof. We directly compute positive lower bounds on each eigenvalue of $\tilde{A}$. For convenience, we work with $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$, where $\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U} = \mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}$. Let $\varepsilon = \|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr}$. By direct computation, we write

$$\varepsilon^2 = (\sigma - 1)^2 - 2\delta^2 + 1. \tag{6.3}$$

We also have

$$\sigma \geq 2\delta \tag{6.4}$$

because $\sigma - 2\delta = (\Phi_1 - \Phi_{22})^2 + (\Phi_{12} + \Phi_{21})^2 \geq 0$. Using (6.4), we can also establish upper and lower bounds on σ in terms of ε. Specifically, $\varepsilon^2 = (\sigma - 1)^2 - 2\delta^2 + 1 \geq (\sigma - 1)^2 - \frac{1}{2}\sigma^2 - 1 = \frac{1}{2}(\sigma - 2)^2$, so

$$2 - \sqrt{2}\varepsilon \leq \sigma \leq 2 + \sqrt{2}\varepsilon. \tag{6.5}$$

Expressions for the eigenvalues of $\tilde{A}$ are given in (6.2). Starting with $\mathbf{\Lambda}_1$ and using (6.4) and (6.5), we obtain

$$\begin{aligned}
\mathbf{\Lambda}_1 &= \frac{1}{2}(\lambda + 2)(\sigma - 2) - \delta + 2 \\
&\geq \frac{1}{2}(\lambda + 2)(\sigma - 2) - \tfrac{1}{2}\sigma + 2 \\
&= \frac{1}{2}(\lambda + 1)\sigma - \lambda \\
&\geq \frac{1}{2}(\lambda + 1)(2 - \sqrt{2}\varepsilon) - \lambda \\
&= 1 - \frac{\sqrt{2}}{2}(\lambda + 1)\varepsilon,
\end{aligned}$$

which is strictly positive when $\varepsilon < \frac{\sqrt{2}}{\lambda+1}$.

Again, using (6.4) and (6.5) along with (6.3), the second eigenvalue satisfies

$$\begin{aligned}
\mathbf{\Lambda}_2 &= \frac{1}{2}(\lambda + 2)(\sigma - 2) + \delta \\
&\geq \frac{1}{2}(\lambda + 2)(2\delta - 2) + \delta \\
&= \delta(\lambda + 3) - (\lambda + 2) \\
&= \frac{\sqrt{2}}{2}((\sigma - 1)^2 + 1 - \varepsilon^2)^{\frac{1}{2}}(\lambda + 3) - (\lambda + 2)
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{\sqrt{2}}{2}((1-\sqrt{2}\varepsilon)^2+1-\varepsilon^2)^{\frac{1}{2}}(\lambda+3)-(\lambda+2)\\
&= \left(1-\sqrt{2}\varepsilon+\frac{1}{2}\varepsilon^2\right)^{\frac{1}{2}}(\lambda+3)-(\lambda+2),
\end{aligned}$$

which is strictly positive when $f(\varepsilon)=\frac{1}{2}\varepsilon^2-\sqrt{2}\varepsilon+1-(\frac{\lambda+2}{\lambda+3})^2>0$. Solving for the roots of $f(\varepsilon)$, we see that $f(\varepsilon)$ is positive for $\varepsilon<\frac{\sqrt{2}}{\lambda+3}$.

The third eigenvalue is more cumbersome to treat and requires a bit more care than the first two. Write $\mathbf{\Lambda}_3 = R-\sqrt{Z}$, where $R=(\lambda+\frac{3}{2})\sigma-(\lambda+1)$ and $Z=\frac{1}{4}(\lambda+3)^2\sigma^2-(6\lambda+9)\delta^2+2\lambda\delta+1$. It can be seen that Z must be nonnegative for $\lambda>0$ by writing

$$\begin{aligned}
Z &= \frac{1}{4}(\lambda+3)^2\sigma^2-(6\lambda+9)\delta^2+2\lambda\delta+1\\
&= \frac{1}{4}\lambda^2\sigma^2+\tfrac{1}{4}(6\lambda+9)(\sigma+2\delta)(\sigma-2\delta)+2\lambda\delta+1\\
&> 0,
\end{aligned}$$

since $\sigma\geq 2\delta$. From the bound on $\mathbf{\Lambda}_2$ and (6.5), we know, for $\lambda>0$, that

$$\sigma \;\geq\; 2-\sqrt{2}\varepsilon \;>\; 2-\sqrt{2}\left(\frac{\sqrt{2}}{\lambda+3}\right) \;>\; \frac{4}{3}$$

and, thus, $R>0$. Therefore, $\mathbf{\Lambda}_3$ is positive when R^2-Z is positive. But we may write

$$\begin{aligned}
R^2-Z &= \left(\lambda+\frac{3}{2}\right)^2\sigma^2-2(\lambda+1)\left(\lambda+\frac{3}{2}\right)\sigma+(\lambda+1)^2\\
&\quad -\frac{1}{4}(\lambda+3)^2\sigma^2+(6\lambda+9)\delta^2-2\lambda\delta-1\\
&\geq \left(\lambda+\frac{3}{2}\right)^2\sigma^2-2(\lambda+1)\left(\lambda+\frac{3}{2}\right)\sigma+(\lambda+1)^2\\
&\quad -\frac{1}{4}(\lambda+3)^2\sigma^2+(6\lambda+9)\delta^2-\lambda\sigma-1\\
&= \left(\lambda+\frac{3}{2}\right)^2\sigma^2-2(\lambda+1)\left(\lambda+\frac{3}{2}\right)\sigma+(\lambda+1)^2\\
&\quad -\frac{1}{4}(\lambda+3)^2\sigma^2+\frac{1}{2}(6\lambda+9)((\sigma-1)^2-\varepsilon^2+1)-\lambda\sigma-1\\
&= \frac{1}{4}(\lambda^2+6\lambda+6)(3\sigma-2)(\sigma-2)+(2\lambda+3)\left(1-\frac{3}{2}\varepsilon\right).
\end{aligned}$$

Since $\sigma>\frac{4}{3}$ implies that the quadratic term in σ, $(3\sigma-2)(\sigma-2)$, is monotonically increasing, we can apply the lower bound in (6.5) to get

$$\begin{aligned}
R^2-Z &\geq \frac{1}{4}(\lambda^2+6\lambda+6)(3\sigma-2)(\sigma-2)+(2\lambda+3)\left(1-\frac{3}{2}\varepsilon\right)\\
&\geq \frac{1}{4}(\lambda^2+6\lambda+6)\left(-\sqrt{2}\varepsilon+\frac{3}{2}\varepsilon^2\right)+(2\lambda+3)\left(1-\frac{3}{2}\varepsilon\right)\\
&= \frac{3}{2}(\lambda^2+4\lambda+3)\varepsilon^2-\sqrt{2}(\lambda^2+6\lambda+6)\varepsilon+2\lambda+3.
\end{aligned}$$

Again, solving for the roots of this quadratic equation in ε, we see that $R^2 - Z$ is positive when $\varepsilon < \frac{\sqrt{2}}{\lambda+3}$.

Finally, the fourth eigenvalue, $\mathbf{\Lambda}_4$, is bounded below by $\mathbf{\Lambda}_3$ and the proof is complete. ∎

It is interesting to note that the bounds for the first three eigenvalues are of the same order (the second and third are even the exact same bound). This suggests that the modification to matrix A_n is optimally balanced with $B(c)$ for $c = 1$.

The full, modified, linearized system may now be written as

$$\begin{cases} \nabla \cdot (\tilde{A}_n \mathbf{U}) = \mathbf{f}_n & \text{in} \quad \Omega, \\ \nabla \times \mathbf{U} = \mathbf{0} & \text{in} \quad \Omega, \\ \boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0} & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot (A_n \mathbf{U}) = \mathbf{g}_n & \text{on} \quad \Gamma_T, \end{cases}$$

where $\mathbf{f}_n = \nabla \cdot (\tilde{A}_n \mathbf{U}_n) - \mathcal{F}_n$ and $\mathbf{g}_n = \mathbf{n} \cdot (A_n \mathbf{U}) - \mathcal{T}_n$.

Define the L^2 functional

$$G(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \|\nabla \cdot (\tilde{A}_n \mathbf{U}) - \mathbf{f}_n\|^2 + \|\nabla \times \mathbf{U}\|^2, \tag{6.6}$$

and define, for any $m > 0$, the space

$$\mathcal{V}^m = \{\mathbf{V} \in H^m(\Omega)^4 : \ \mathbf{n} \cdot (A_n \mathbf{V}) = \mathbf{g}_n \text{ on } \Gamma_T, \ \boldsymbol{\tau} \cdot \mathbf{V} = \mathbf{0} \text{ on } \Gamma_D\}.$$

In the case of pure displacement boundary conditions ($\Gamma_N = \emptyset$), we denote the space by $\mathcal{V}_D^m$.

The least-squares minimization problem for each Newton step is as follows: given $\mathbf{U}_n$, $\mathbf{f}_n$, and $\mathbf{g}_n$, find $\mathbf{U} \in \mathcal{V}^1$ such that

$$G(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \inf_{\mathbf{V} \in \mathcal{V}^1} G(\mathbf{V}; \mathbf{U}_n, \mathbf{f}_n).$$

7. Ellipticity. To use the L^2-based functional in (6.6) on each Newton step, we must assume that the previous iterate is in $H^{1+\delta}(\Omega)$ for some $\delta > 0$ because (6.6) is composed of derivatives of products of the unknown and the previous solution and, in $\mathbb{R}^2$, the space $H^{1+\delta}(\Omega)$ is closed under multiplication only for $\delta > 0$ (see Lemma 4.1). Thus, showing only H^1 ellipticity of (6.6) is not sufficient to establish a well-defined Newton iteration; we must show that each iterate remains in $H^{1+\delta}(\Omega)$. In this section, we establish H^{1+k} ellipticity of an H^k-based functional for $k \geq 0$ and show that minimizing the L^2-based functional is sufficient to guarantee the required smoothness of each iterate. Our theoretical results hold for the pure displacement problem.

For clarity, we use the following conventions: $\delta > 0$ and $k \geq 0$ (our final results require the cases $k = 0$ and $k = \delta$).

In Theorem 6.1, matrix $\tilde{A} = A(\mathbf{U}) + B(1)$ is uniformly symmetric positive definite over Ω when the strain of $\mathbf{U}$ is sufficiently small, that is, for $\mathbf{U} \in \mathcal{S}_\lambda$. In this section, we assume this property holds and consider the solution of a general Newton step of the pure displacement problem by minimizing the more general H^k-based functional

$$G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \|\nabla \cdot (\tilde{A}_n \mathbf{U}) - \mathbf{f}_n\|_k^2 + \|\nabla \times \mathbf{U}\|_k^2. \tag{7.1}$$

Its associated minimization problem is as follows: given $\mathbf{U}_n \in H^{1+\delta}(\Omega)$ and $\mathbf{f}_n \in H^k(\Omega)$, find $\mathbf{U} \in \mathcal{V}_D^{1+k}$ such that

$$G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \inf_{\mathbf{V} \in \mathcal{V}_D^{1+k}} G_k(\mathbf{V}; \mathbf{U}_n, \mathbf{f}_n). \tag{7.2}$$

By Lemma 4.1, it is clear that $\mathbf{U} \in H^{1+k}(\Omega)$ and $\mathbf{U}_n \in H^{1+\delta}(\Omega)$ are sufficient to ensure that $\nabla \cdot (\tilde{A}_n \mathbf{U}) \in H^k(\Omega)$.

The following series of lemmas leads to establishing equivalence of $G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{0})^{\frac{1}{2}}$ to the H^{1+k} norm.

LEMMA 7.1. *Let Ω be a simply connected domain in $\mathbb{R}^2$ and suppose $\mathbf{V} \in L^2(\Omega)^4$. Then $\nabla \cdot \mathbf{V} = \mathbf{0}$ and $\int_{\partial\Omega} \mathbf{n} \cdot \mathbf{V} = \mathbf{0}$ if and only if there exists a function $\mathbf{r} \in H^1(\Omega)^2$ such that $\mathbf{V} = \nabla^\perp \mathbf{r}$. Furthermore, $\mathbf{r} \in H^1(\Omega)^2$ is unique up to an additive constant vector in $\mathbb{R}^2$.*

Proof. The result follows by applying Theorem 3.1 in Chapter I of [14] to each block component of $\mathbf{V}$. ∎

LEMMA 7.2. *Let Ω be a simply connected domain in $\mathbb{R}^2$. Every $\mathbf{V} \in L^2(\Omega)^4$ has the orthogonal decomposition $\mathbf{V} = \nabla \mathbf{p} + \nabla^\perp \mathbf{q}$ for $\mathbf{p} \in H^1(\Omega)^2, \mathbf{q} \in H_0^1(\Omega)^2$. Furthermore, $\mathbf{q}$ is unique in $H_0^1(\Omega)^2$ and $\mathbf{p}$ is unique in $H^1(\Omega)^2$ up to an additive constant vector in $\mathbb{R}^2$.*

Proof. The result follows by applying Theorem 3.2 in Chapter I of [14] to each block component of $\mathbf{V}$. ∎

LEMMA 7.3. *Assume that $\mathbf{U} \in \mathcal{S}_\lambda$ and denote $\tilde{A} = A + B$, with $A = A(\mathbf{U})$ and $B = B(1)$ as defined in section 6. Also assume that $A\mathbf{Z}$ and $B\mathbf{Z}$ are in $L^2(\Omega)^4$. If $\mathbf{Z} \in \mathcal{V}_D^1$ satisfies the system*

$$\begin{cases} \nabla \cdot \tilde{A}\mathbf{Z} = \mathbf{0} & \text{in} \quad \Omega, \\ \nabla \times \mathbf{Z} = \mathbf{0} & \text{in} \quad \Omega, \end{cases} \tag{7.3}$$

then it must be the trivial solution, $\mathbf{Z} = \mathbf{0}$.

Proof. By Lemma 7.2, $\mathbf{Z} = \nabla \mathbf{p} + \nabla^\perp \mathbf{q}$ for $\mathbf{p} \in H^1(\Omega)^2, \mathbf{q} \in H_0^1(\Omega)^2$. The second equation in (7.3) implies

$$\mathbf{0} = \nabla \times \mathbf{Z} = \nabla \times \nabla \mathbf{p} + \nabla \times \nabla^\perp \mathbf{q} = -\Delta \mathbf{q},$$

and, since $\mathbf{q} \in H^1(\Omega)_0^2$, we must have $\mathbf{q} = \mathbf{0}$. Thus, $\mathbf{Z} = \nabla \mathbf{p}$.

Now, using Green's formula with $\mathbf{1} = (1, 1)^t$, we get

$$\mathbf{0} = \langle \nabla \cdot A\mathbf{Z}, \mathbf{1} \rangle + \langle A\mathbf{Z}, \nabla \mathbf{1} \rangle = \int_{\partial\Omega} \mathbf{n} \cdot A\mathbf{Z}.$$

Applying Lemma 7.1 to $A\mathbf{Z}$ yields $A\mathbf{Z} = \nabla^\perp \mathbf{r}$ for $\mathbf{r} \in H^1(\Omega)^2$. Since $\mathbf{0} = \boldsymbol{\tau} \cdot \mathbf{Z} = \boldsymbol{\tau} \cdot \nabla \mathbf{p}$ on $\partial\Omega$, we know that $\mathbf{p} = \mathbf{p}_0$ is constant on $\partial\Omega$. We thus have

$$\begin{aligned} \langle A\mathbf{Z}, \mathbf{Z} \rangle &= \langle \nabla^\perp \mathbf{r}, \nabla \mathbf{p} \rangle \\ &= \langle -\nabla \cdot \nabla^\perp \mathbf{r}, \mathbf{p} \rangle + \int_{\partial\Omega} (\mathbf{n} \cdot \nabla^\perp \mathbf{r}) \mathbf{p} \\ &= \int_{\partial\Omega} (\mathbf{n} \cdot \nabla^\perp \mathbf{r}) \mathbf{p} \\ &= \mathbf{p}_0 \int_{\partial\Omega} \mathbf{n} \cdot A\mathbf{Z} \\ &= \mathbf{0}. \end{aligned}$$

Since $\mathbf{Z} = \nabla \mathbf{p}$, we may then write $B\mathbf{Z} = \nabla^{\perp}\mathbf{s}$, where $\mathbf{s} = \left(\begin{smallmatrix} p_2 \\ -p_1 \end{smallmatrix}\right)$. Thus,

$$\begin{aligned}\langle B\mathbf{Z}, \mathbf{Z}\rangle &= \langle \nabla^{\perp}\mathbf{s}, \nabla \mathbf{p}\rangle \\ &= \langle \mathbf{s}, \nabla \times \nabla \mathbf{p}\rangle - \int_{\partial\Omega} (\boldsymbol{\tau} \cdot \nabla \mathbf{p})\mathbf{s} \\ &= -\int_{\partial\Omega} (\boldsymbol{\tau} \cdot \nabla \mathbf{p}_0)\mathbf{s} \\ &= \mathbf{0},\end{aligned}$$

which implies $\langle \tilde{A}\mathbf{Z}, \mathbf{Z}\rangle = \mathbf{0}$. Since $\mathbf{U} \in \mathcal{S}_{\lambda}$, matrix $\tilde{A}$ is positive definite, and we must have $\mathbf{Z} = \mathbf{0}$. ∎

Now consider the following elliptic boundary value problem:

$$\nabla \cdot M \nabla \mathbf{p} = \mathbf{f} \quad \text{in} \quad \Omega, \tag{7.4}$$

satisfying either $\mathbf{p} = \mathbf{0}$ or $\mathbf{n} \cdot M\nabla \mathbf{p} = \mathbf{0}$ on $\partial\Omega$. When Ω has $C^{1+k,1}$ boundary and M is uniformly positive definite over Ω with coefficients in $C^{k,1}(\bar{\Omega})$, problem (7.4) admits the regularity bound,

$$\|\mathbf{p}\|_{2+k} \leq C\|\mathbf{f}\|_k, \tag{7.5}$$

for $\mathbf{p} \in H^{2+k}(\Omega)$. Chapter 2 of [15] establishes this for integer values of k. For noninteger k, we may appeal to interpolation in Sobolev norms as in [17]. Similar regularity results, with different assumptions than here, are given in [1, 2, 12].

LEMMA 7.4. *Assume a solution to nonlinear problem* (5.1)*:* $\mathbf{U}^* \in \mathcal{V}_D^{2+k} \cap \mathcal{S}_{\lambda}$. *Assume also that* Ω *is smooth enough to admit* (7.5) *for* $k \geq 0$. *Let* $\tilde{A}_* = A(\mathbf{U}^*) + B(1)$. *Then there exists a positive constant,* c_**, independent of* $\mathbf{U}$*, such that*

$$\|\mathbf{U}\|_{1+k} \leq c_*(\|\nabla \cdot \tilde{A}_*\mathbf{U}\|_k + \|\nabla \times \mathbf{U}\|_k)$$

for all $\mathbf{U} \in \mathcal{V}_D^{1+k}$.

Proof. Consider the skew-symmetric orthogonal matrix

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{7.6}$$

The following relations are easily derived:

$$\begin{aligned} \nabla\times &= \nabla \cdot Q, \\ \nabla\cdot &= \nabla \times Q^t, \\ \nabla^{\perp} &= Q\nabla, \\ \nabla &= Q^t\nabla^{\perp}, \\ \mathbf{n}\cdot &= -\boldsymbol{\tau} \cdot Q, \\ \boldsymbol{\tau}\cdot &= \mathbf{n} \cdot Q^t. \end{aligned} \tag{7.7}$$

Since $\tilde{A}_*$ is uniformly positive definite over Ω, there are constants, λ_1, $\lambda_2 > 0$, such that

$$\lambda_1 \boldsymbol{\xi}^t\boldsymbol{\xi} \leq \boldsymbol{\xi}^t\tilde{A}_*\boldsymbol{\xi} \leq \lambda_2\boldsymbol{\xi}^t\boldsymbol{\xi} \tag{7.8}$$

and

$$\frac{1}{\lambda_2}\boldsymbol{\xi}^t\boldsymbol{\xi} \le \boldsymbol{\xi}^t\tilde{A}_*^{-1}\boldsymbol{\xi} \le \frac{1}{\lambda_1}\boldsymbol{\xi}^t\boldsymbol{\xi} \tag{7.9}$$

for any $\boldsymbol{\xi} \in \mathbb{R}^4$. Define

$$\mathcal{C} = Q^t\tilde{A}_*^{-1}Q,$$

and note that

$$\boldsymbol{\xi}^t\boldsymbol{\xi} = \boldsymbol{\xi}^tQ^tQ\boldsymbol{\xi} = (Q\boldsymbol{\xi})^t(Q\boldsymbol{\xi})$$

and

$$\boldsymbol{\xi}^t\mathcal{C}\boldsymbol{\xi} = \boldsymbol{\xi}^tQ^t\tilde{A}_*^{-1}Q\boldsymbol{\xi} = (Q\boldsymbol{\xi})^t\tilde{A}_*^{-1}(Q\boldsymbol{\xi}).$$

Now, it can easily be seen that $\mathcal{C}$ is symmetric and uniformly positive definite over Ω:

$$\frac{1}{\lambda_2}\boldsymbol{\xi}^t\boldsymbol{\xi} \le \boldsymbol{\xi}^t\mathcal{C}\boldsymbol{\xi} \le \frac{1}{\lambda_1}\boldsymbol{\xi}^t\boldsymbol{\xi}. \tag{7.10}$$

We also note that

$$\nabla \times \tilde{A}_*^{-1}\nabla^\perp = \nabla \cdot Q\tilde{A}_*^{-1}Q\nabla = -\nabla \cdot \mathcal{C}\nabla.$$

With $\mathbf{U}^* \in \mathcal{V}_D^{2+k}$ and $\mathbf{U} \in \mathcal{V}_D^{1+k}$, we have that $\nabla \cdot \tilde{A}_*\mathbf{U} \in H^k(\Omega)$, and thus, for any $\mathbf{U} \in \mathcal{V}_D^{1+k}$, there is a unique $\mathbf{p} \in H^{2+k}(\Omega)$ that satisfies

$$\begin{cases} \nabla \cdot \tilde{A}_*\nabla\mathbf{p} = \nabla \cdot \tilde{A}_*\mathbf{U} & \text{in} \quad \Omega, \\ \qquad\quad \mathbf{p} = \mathbf{0} & \text{on} \quad \partial\Omega \end{cases} \tag{7.11}$$

and a $\mathbf{q} \in H^{2+k}(\Omega)$ that satisfies

$$\begin{cases} -\nabla \cdot \mathcal{C}\nabla\mathbf{q} = \nabla \times \mathbf{U} & \text{in} \quad \Omega, \\ \mathbf{n} \cdot \mathcal{C}\nabla\mathbf{q} = \mathbf{0} & \text{on} \quad \partial\Omega, \end{cases} \tag{7.12}$$

and $\int_\Omega \mathbf{q}\, dx = \mathbf{0}$. Now define $\mathbf{Z} = \mathbf{U} - \nabla\mathbf{p} - \tilde{A}_*^{-1}\nabla^\perp\mathbf{q}$. Note that $\mathbf{Z} \in H^{1+k}(\Omega)$ and, on $\partial\Omega$, that

$$\begin{aligned} \boldsymbol{\tau} \cdot \mathbf{Z} &= \boldsymbol{\tau} \cdot \mathbf{U} - \boldsymbol{\tau} \cdot \nabla\mathbf{p} - \boldsymbol{\tau} \cdot \tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= -\boldsymbol{\tau} \cdot Q\mathcal{C}Q^t\nabla^\perp\mathbf{q} \\ &= \mathbf{n} \cdot \mathcal{C}\nabla\mathbf{q} \\ &= \mathbf{0}. \end{aligned}$$

Thus, $\mathbf{Z} \in \mathcal{V}_D^{1+k}$. We further see that

$$\begin{aligned} \nabla \cdot \tilde{A}_*\mathbf{Z} &= \nabla \cdot \tilde{A}_*\mathbf{U} - \nabla \cdot \tilde{A}_*\nabla\mathbf{p} - \nabla \cdot \tilde{A}_*\tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= \mathbf{0} \end{aligned}$$

and

$$\begin{aligned} \nabla \times \mathbf{Z} &= \nabla \times \mathbf{U} - \nabla \times \nabla\mathbf{p} - \nabla \times \tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= \nabla \times \mathbf{U} + \nabla \cdot \mathcal{C}\nabla\mathbf{q} \\ &= \mathbf{0}. \end{aligned}$$

By Lemma 7.3, we therefore conclude that $\mathbf{Z} = \mathbf{0}$. Since $\mathbf{U}^* \in H^{2+k}(\Omega) \subseteq C^{k,1}(\bar{\Omega})$, we may apply (7.5) to problems (7.11) and (7.12). Combining these bounds with the triangle inequality and (7.9), we may write

$$\begin{aligned}\|\mathbf{U}\|_{1+k} &\leq \|\nabla \mathbf{p}\|_{1+k} + \|\tilde{A}_*^{-1}\nabla^{\perp}\mathbf{q}\|_{1+k} \\ &\leq \|\mathbf{p}\|_{2+k} + 1/\lambda_1\|\mathbf{q}\|_{2+k} \\ &\leq C(\|\nabla \cdot \tilde{A}_*\mathbf{U}\|_k + \|\nabla \times \mathbf{U}\|_k).\end{aligned}$$

Application of (2.1) completes the proof. ∎

Recall that we cast the solution of nonlinear problem (3.1) as the zero of $\mathcal{F}(\mathbf{U})$, where $\mathcal{F}'(\mathbf{U})[\mathbf{V}]$ is given in (5.3). We consider the boundedness of the second Fréchet derivative of $\mathcal{F}(\mathbf{U})$ in directions $\mathbf{V}$ and $\mathbf{W}$ in the following lemma.

LEMMA 7.5. *For all $\mathbf{U}, \mathbf{V} \in H^{1+\delta}(\Omega)$ and $\mathbf{W} \in H^{1+k}(\Omega)$, there exists a positive constant, c_2, such that*

$$\|\mathcal{F}''(\mathbf{U})[\mathbf{V}, \mathbf{W}]\|_k \leq c_2\|\mathbf{U}\|_{1+\delta}\|\mathbf{V}\|_{1+\delta}\|\mathbf{W}\|_{1+k}. \tag{7.13}$$

Proof. Writing the second Fréchet derivative of $\mathcal{F}$ in the directions $\mathbf{V}$ and $\mathbf{W}$ as

$$\begin{aligned}\mathcal{F}''(\mathbf{U})[\mathbf{V}, \mathbf{W}] = \nabla \cdot \big[&\lambda \mathrm{tr}(\mathbf{W}^t\mathbf{V})(\mathbf{I}+\mathbf{U}) + \lambda \mathrm{tr}(\mathbf{V}^t(\mathbf{I}+\mathbf{U}))\mathbf{W} + \lambda|\mathrm{tr}(\mathbf{W}^t(\mathbf{I}+\mathbf{U}))\mathbf{V} \\ &+ (\mathbf{I}+\mathbf{U})(\mathbf{W}^t\mathbf{V} + \mathbf{V}^t\mathbf{W}) + \mathbf{V}(\mathbf{W} + \mathbf{W}^t + W^t\mathbf{U} + \mathbf{U}^t\mathbf{W}) \\ &+ \mathbf{W}(\mathbf{V} + \mathbf{V}^t + V^t\mathbf{U} + \mathbf{U}^t\mathbf{V})\big],\end{aligned}$$

we see that each component may be written as a linear combination of terms of the form $\partial(W_iV_jU_k)$ or $\partial(W_iV_j)$, $i, j, k = 1, 2, 3, 4$, where ∂ again represents either ∂_x or ∂_y. The lemma then follows by the triangle inequality and by applying Lemma 4.1 to each term once or twice. ∎

Define the $H^{1+\delta}$ neighborhood of the solution by

$$\mathcal{B}_r = \{\mathbf{U} \in H^{1+\delta}(\Omega) : \|\mathbf{U} - \mathbf{U}^*\|_{1+\delta} < r\}.$$

We now are able to state the main result of this section.

THEOREM 7.6. *Assume Ω has C^{3+k} boundary and that $\mathbf{f} \in H^{1+k}(\Omega)$ is small enough to guarantee that problem (5.1) has solution $\mathbf{U}^* \in H^{2+k}(\Omega) \cap \mathcal{S}_\lambda$ by Theorem 4.2. Then there exist some $r > 0$ and constants $c_0, c_1 > 0$, depending only on f and Ω, such that, for all $\mathbf{U}_n \in \mathcal{V}_D^{1+\delta} \cap \mathcal{B}_r$,*

$$c_0\|\mathbf{U}\|_{1+k}^2 \;\leq\; G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{0}) \;\leq\; c_1\|\mathbf{U}\|_{1+k}^2 \tag{7.14}$$

for every $\mathbf{U} \in \mathcal{V}_D^{1+k}$ for $k \geq 0$.

Proof. The upper bound follows from the triangle inequality and Lemma 4.1. With $\mathbf{U}_n \in H^{1+\delta}(\Omega)$, Lemma 4.1 guarantees that $A\mathbf{U}, B\mathbf{U} \in H^{1+k}(\Omega)$ when $\mathbf{U} \in H^{1+k}(\Omega)$. By Theorem 4.2 with $m = 1 + k$, there is a solution to problem (5.1), $\mathbf{U}^* \in H^{2+k}(\Omega) \cap \mathcal{S}_\lambda$. Thus, by Lemma 7.4, we know there exists a positive constant, c_*, independent of $\mathbf{U}$, such that

$$\|\mathbf{U}\|_{1+k} \leq c_*(\|\nabla \cdot \tilde{A}(\mathbf{U}^*)\mathbf{U}\|_k + \|\nabla \times \mathbf{U}\|_k) \tag{7.15}$$

for all $\mathbf{U} \in \mathcal{V}_D^{1+k}$. We now need only to extend this result to the operator linearized about $\mathbf{U}_n$ rather than $\mathbf{U}^*$. Recall that we may denote $\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}] = \nabla \cdot A(\mathbf{U}_n)\mathbf{U}$ and $\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] = \nabla \cdot A(\mathbf{U}^*)\mathbf{U}$. By the mean value theorem, we may write

$$\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}] - \mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] = \mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U}, \mathbf{U}_n - \mathbf{U}^*] \tag{7.16}$$

for some $\hat{\mathbf{U}} = \theta\mathbf{U}_n + (1-\theta)\mathbf{U}^*$ with $\theta \in [0,1]$. Since $\mathbf{U}_n \in \mathcal{B}_r$, $\hat{\mathbf{U}}$ can be bounded in the $H^{1+\delta}$ norm in the following way:

$$\begin{aligned}\|\hat{\mathbf{U}}\|_{1+\delta} &= \|\theta\mathbf{U}_n + (1-\theta)\mathbf{U}^*\|_{1+\delta} \\ &\le \|\theta(\mathbf{U}_n - \mathbf{U}^*)\|_{1+\delta} + \|\mathbf{U}^*\|_{1+\delta} \\ &\le r + \|\mathbf{U}^*\|_{1+\delta}.\end{aligned} \tag{7.17}$$

So, by (7.16), the triangle inequality, (7.13), and (7.15), we have

$$\begin{aligned}&\|\nabla\cdot\tilde{A}(\mathbf{U}_n)\mathbf{U}\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad= \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad= \|\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] + \mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U},\mathbf{U}_n-\mathbf{U}^*]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad\ge \|\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}]\|_k - \|\mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U},\mathbf{U}_n-\mathbf{U}^*]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad\ge \|\nabla\cdot A(\mathbf{U}^*)\mathbf{U}\|_k + \|\nabla\times\mathbf{U}\|_k - c_2\|\hat{\mathbf{U}}\|_{1+\delta}\|\mathbf{U}\|_{1+k}\|\mathbf{U}_n-\mathbf{U}^*\|_{1+\delta} \\ &\quad\ge c_*^{-1}\|\mathbf{U}\|_{1+k} - c_2 r(r+\|\mathbf{U}^*\|_{1+\delta})\|\mathbf{U}\|_{1+k} \\ &\quad= (c_*^{-1} - c_2 r(r+\|\mathbf{U}^*\|_{1+\delta}))\|\mathbf{U}\|_{1+k} \\ &\quad\ge C\|\mathbf{U}\|_{1+k},\end{aligned} \tag{7.18}$$

where C is guaranteed to be positive for r sufficiently small. Application of (2.1) completes the proof. ∎

COROLLARY 7.7. *Assume that Ω, $\mathbf{f}$, $\mathbf{U}^*$, and $\mathbf{U}_n \in \mathcal{V}_D^{1+\delta} \cap \mathcal{B}_r$ satisfy the assumptions of Theorem* 7.6. *Then, for some r sufficiently small, the unique $\mathbf{U}$ that satisfies*

$$G_0(\mathbf{U};\mathbf{U}_n,\mathbf{f}_n) = \inf_{\mathbf{V}\in\mathcal{V}_D^1} G_0(\mathbf{V};\mathbf{U}_n,\mathbf{f}_n) \tag{7.19}$$

also satisfies

$$G_\delta(\mathbf{U};\mathbf{U}_n,\mathbf{f}_n) = \inf_{\mathbf{V}\in\mathcal{V}_D^{1+\delta}} G_\delta(\mathbf{V};\mathbf{U}_n,\mathbf{f}_n). \tag{7.20}$$

Proof. From the Riesz representation theorem and Theorem 7.6 with $k=0$, we have a unique minimizer, $\mathbf{U}$, of the L^2-based functional in (7.19) in $H^1(\Omega)$. Similarly, for $k=\delta>0$, we also have a unique minimizer, $\mathbf{U}'$, of the H^δ-based functional in (7.20) in $H^{1+\delta}(\Omega)$. Since these functionals both have zero minimum, $\mathbf{U}'$ must also minimize the functional in (7.19). Thus, $\mathbf{U} = \mathbf{U}' \in H^{1+\delta}(\Omega)$. ∎

Therefore, we are able to conclude that under sufficient smoothness requirements, minimizing the L^2-based functional is sufficient to guarantee that each Newton iterate, $\mathbf{U} = \mathbf{U}_{n+1}$, remains in $H^{1+\delta}(\Omega)$.

8. Convergence of Newton's method. We now consider the sequence of iterates arising from the minimization of each linearized functional under the assumptions of Theorem 7.6. This section details the theory and assumptions for the convergence of Newton's method. As in Theorem 7.6, we assume the solution to the previous Newton step to be in $\mathcal{B}_r$. Here, we show convergence of the iterates in the $H^{1+\delta}$ norm and that each iterate remains in $\mathcal{B}_r$.

Consider the Taylor expansion of $\mathcal{F}(\mathbf{U}^*)$ about the current approximation $\mathbf{U}_n$:

$$\mathbf{0} = \mathcal{F}(\mathbf{U}^*) = \mathcal{F}(\mathbf{U}_n) + \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^*-\mathbf{U}_n] + \frac{1}{2}\mathcal{F}''(\tilde{\mathbf{U}})[\mathbf{U}^*-\mathbf{U}_n,\mathbf{U}^*-\mathbf{U}_n] \tag{8.1}$$

for $\tilde{\mathbf{U}} = \omega \mathbf{U}_n + (1-\omega)\mathbf{U}^*$ with $\omega \in [0,1]$. As in (7.17), if $\mathbf{U}_n \in \mathcal{B}_r$, then $\tilde{\mathbf{U}}$ satisfies

$$\|\tilde{\mathbf{U}}\|_{1+\delta} \leq r + \|\mathbf{U}^*\|_{1+\delta}. \tag{8.2}$$

Recall that we may write the Newton iterate, $\mathbf{U}$, as the solution to problem (7.2) and, thus,

$$\mathcal{F}'(\mathbf{U}_n)[\mathbf{U} - \mathbf{U}_n] = -\mathcal{F}(\mathbf{U}_n), \tag{8.3}$$

with $\nabla \times \mathbf{U} = \nabla \times \mathbf{U}_n = \nabla \times \mathbf{U}^* = 0$.

Applying (8.1), (8.3), (7.13), and (8.2) to the bound in (7.18), and recalling that $\mathbf{U}_n \in \mathcal{B}_r$, we get

$$\begin{aligned}
\|\mathbf{U}^* - \mathbf{U}\|_{1+\delta} &\leq \frac{1}{\sqrt{c_0}} \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^* - \mathbf{U}]\|_\delta + \|\nabla \times (\mathbf{U}^* - \mathbf{U})\|_\delta \\
&= \frac{1}{\sqrt{c_0}} \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^* - \mathbf{U}_n] - \mathcal{F}'(\mathbf{U}_n)[\mathbf{U} - \mathbf{U}_n]\|_\delta \\
&= \frac{1}{2\sqrt{c_0}} \|\mathcal{F}''(\tilde{\mathbf{U}})[\mathbf{U}^* - \mathbf{U}_n, \mathbf{U}^* - \mathbf{U}_n]\|_\delta \\
&\leq \frac{c_2}{2\sqrt{c_0}} \|\tilde{\mathbf{U}}\|_{1+\delta} \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&\leq \frac{c_2}{2\sqrt{c_0}} (r + \|\mathbf{U}^*\|_{1+\delta}) r \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&:= c_3 \; r \; \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta},
\end{aligned} \tag{8.4}$$

which proves that Newton's method converges for r sufficiently small. Again noting that $\mathbf{U}_n \in \mathcal{B}_r$, we further note that

$$\begin{aligned}
\|\mathbf{U}^* - \mathbf{U}\|_{1+\delta} &\leq c_3 r \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&\leq c_3 r^2.
\end{aligned}$$

To verify that $\mathbf{U} \in \mathcal{B}_r$, we only need to show that $c_3 r^2 < r$. Substituting the definition of c_3, we see that this is satisfied for

$$r \; < \; \frac{1}{2} \left(\sqrt{\|\mathbf{U}^*\|_{1+\delta}^2 + \eta} - \|\mathbf{U}^*\|_{1+\delta} \right) \; < \; \frac{\eta}{4\|\mathbf{U}^*\|_{1+\delta}},$$

where $\eta = \frac{8\sqrt{c_0}}{c_2}$. This shows that for guaranteed convergence, larger solutions require better initial guesses than smaller solutions (as measured in the $H^{1+\delta}$ norm). We now consider the issue of finding an appropriate "good" initial guess.

9. Multilevel solution. As described above, the solution to nonlinear system (3.1) is generally comprised of several Newton iterations. The first few iterations are crude approximations to the true solution of the nonlinear problem. It is therefore appropriate to represent the early approximations on a mesh with fewer degrees of freedom. As the Newton iterates remove more of the error due to the nonlinearity, the approximations can be represented on increasingly finer meshes. In other words, we wish to eliminate as much of the nonlinear error as possible on coarse grids where it is less expensive.

The approach in [13] uses this multilevel nested iteration Newton idea with a FOSLS finite element discretization and a multigrid solver to achieve a robust solution strategy for a certain class of nonlinear problems. Under particular assumptions

on the form of the nonlinearity, the finite element spaces used, the smoothness of the solution, and the ellipticity of the linearized equations, convergence to the solution is established with accuracy comparable to discretization error on the finest level at a cost proportional to the degrees of freedom on the finest level. We briefly summarize this nested iteration-Newton-FOSLS-multigrid (NI-Newton-FOSLS-MG) algorithm and detail the additional assumptions we must make for application to the geometrically nonlinear elasticity system.

Define the hierarchy of discrete nested subspaces,

$$\mathcal{V}^{h_0} \subset \mathcal{V}^{h_1} \cdots \subset \mathcal{V}^{h_J} \subset \mathcal{V}_D^{1+\delta}. \tag{9.1}$$

The following algorithm describes the NI-Newton-FOSLS-MG method:

1. Begin with a zero approximation, $\mathbf{U}_0$, on coarsest level $\mathcal{V}^{h_0}$.
2. Linearize the equations about the current approximation and form the discrete least-squares minimization problem.
3. Apply m multigrid cycles to the resulting matrix equations.
4. Repeat steps 2 and 3 n times on the current level.
5. Interpolate the current approximation to the next finer level, $\mathcal{V}^{h_i}$.
6. Repeat steps 2–5 until desired accuracy is achieved.

To apply the results of [13] to the nonlinear elasticity system, we must make the following series of assumptions.

A1. Assume the existence of a solution, $\mathbf{U}^* \in H^{2+\delta}(\Omega)$, to problem (3.1). For our problem, this is justified in section 4. Theorem 4.2 with $m = 1 + \delta$ requires that the boundary of Ω be $C^{3+\delta}$ smooth and $\mathbf{f} \in H^{1+\delta}(\Omega)$ in order to guarantee $\mathbf{U}^* \in H^{2+\delta}(\Omega)$. In the context of elasticity, the internal forcing function, $\mathbf{f}$, is generally at least this smooth for a wide range of practical problems. Assuming a very smooth domain, however, is a stronger restriction than we generally wish to adhere to in practice. We do find that in practice this can be relaxed in some cases, but in many cases we must consider complimentary methods for dealing with nonsmooth domains.

A2. Assume the operator of linearized elasticity maps $\mathcal{V}_D^{1+k}$ into $H^k(\Omega)$. This is established in section 7.

A3. Assume H^{1+k} ellipticity of the functional as in (7.1). Theorem 7.6 establishes this for the pure displacement problem under the small strains assumption of Theorem 6.1.

A4. Assume boundedness of the second Fréchet derivative of $\mathcal{F}$ as in (7.13). Justification of this is established in Lemma 7.5.

A5. Assume the finite element spaces in (9.1) guarantee the following approximation properties and inverse estimate. Let I_ν^h be a bounded H^ν projection onto finite element space $\mathcal{V}^h$. We assume interpolation bounds of the form

$$\|\mathbf{U} - I_{1+\delta}^h \mathbf{U}\|_\gamma \leq C h^{2+\delta-\gamma} \|\mathbf{U}\|_{2+\delta} \quad \forall \gamma \in [0, 1+\delta]$$

and the inverse estimate

$$\|\mathbf{U}\|_\beta \leq \frac{C}{h^{\beta-\gamma}} \|\mathbf{U}\|_\gamma \quad \forall \mathbf{U} \in \mathcal{V}^h,\ \beta \in [0, 1+\delta],\ \gamma \in [0, \beta].$$

We concentrate on standard finite element subspaces of H^1 (for example, bilinears on rectangles) which exhibit these properties; see [3] for details.

A6. Assume a sufficiently fine coarsest level by insisting that $\mathcal{B}_r \cap \mathcal{V}^{h_0} \neq \emptyset$ and that the initial guess is sufficiently close to the solution by choosing $\mathbf{U}_0 \in \mathcal{B}_r \cap \mathcal{V}^{h_0}$. Bounds on r can be found in the full theory in [13].

Under these assumptions, we may directly apply the theory developed in [13]. By this theory, there are values of m and n, independent of h, in the multilevel algorithm described above that result in an approximation on the finest level that is accurate to the level of discretization error at a cost proportional to the degrees of freedom on the finest level.

There are many contributions to the error in each approximation in the NI-Newton-FOSLS-MG solution process. In the innermost iteration, the multigrid solver reduces the algebraic error by performing a number of multigrid cycles before relinearizing. On each grid level, there is discretization error associated with the finite element space used. A sufficient number of Newton steps must be performed on each level to eliminate the error associated with the nonlinearity. For a truly optimal algorithm, we must consider the sources of error that contribute to the total error in the current approximation, and make decisions on how to proceed in the algorithm in order to efficiently reduce the total error to an acceptable level.

10. Computational results. To validate the theory presented above, consider the numerical approximation to the solution of a pure displacement problem on domain $\Omega = [0,1]^2$, with Lamé constants $\lambda = 2.15$, $\mu = 1$. As a test problem, we choose the solution to nonlinear problem (3.1) to be

$$\mathbf{u}^* = \begin{pmatrix} x(1-x)y^2(1-y)^2\sin(\pi x) \\ x^2(1-x)^2y^2(1-y)^2\cos(\pi y) \end{pmatrix},$$

and let $\mathbf{U}^* = \nabla\mathbf{u}^*$ be the exact solution for the first-stage problem. The right-side function, $\mathbf{f}$, is computed accordingly.

Denote by $\mathcal{V}^h$ the space of continuous piecewise bilinear finite elements on a uniform grid of mesh size h. For convenience, we use this space for all test problems. Each step of the pure displacement problem is found by minimizing the discrete functional,

$$G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f}_n) = \|\nabla\cdot(\tilde{A}_n\mathbf{U}^h) - \mathbf{f}_n\|^2 + \|\nabla\times\mathbf{U}^h\|^2, \tag{10.1}$$

over the space

$$\mathcal{V}_D^h = \{\mathbf{V}^h \in \mathcal{V}^h : \boldsymbol{\tau}\cdot\mathbf{V}^h = \mathbf{0} \text{ on } \partial\Omega\}.$$

We begin with an initial guess of $\mathbf{U}_0 = \mathbf{0}$ so that the first Newton step corresponds to the linear elasticity case. Recall that we seek the solution to the original nonlinear problem as well as each linearized step. Define the following nonlinear functional to measure the convergence to nonlinear problem (3.1):

$$\mathcal{G}(\mathbf{U};\mathbf{f}) = \|\mathcal{F}(\mathbf{U})\|^2 + \|\nabla\times\mathbf{U}\|^2. \tag{10.2}$$

In an $H^{1+\delta}$ neighborhood near the solution, a simple computation on a Taylor series of $\mathcal{F}$ about $\mathbf{U}^*$ (invoking Lemma (7.5)) shows that $\mathcal{G}(\mathbf{U};\mathbf{f})$ is equivalent to $G(\mathbf{U};\mathbf{U}^*,\mathbf{0})$, indicating that the H^1 norm of the error to the nonlinear problem can be effectively monitored by $\mathcal{G}(\mathbf{U};\mathbf{f})$. Near convergence of Newton's method, the nonlinear and linearized functionals tend to take on the same values. Thus, a practical measure of how much of the error in the approximation is due to the nonlinearity can be obtained by the difference in the linearized and nonlinear functional values.

For the test problem summarized in Table 10.1, we ensure that essentially all algebraic error is removed from each system by reducing the residual by a factor of

10^6 using $V(1,1)$ cycles. Numerical results are reported for the following: grid level, N ($h = (N+1)^{-1}$); Newton step, m; linearized functional norm, $G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f})^{\frac{1}{2}}$; nonlinear functional norm, $\mathcal{G}(\mathbf{U}^h; \mathbf{f})^{\frac{1}{2}}$; L^2 error of the solution, $\|\mathbf{U}^* - \mathbf{U}^h\|$; and asymptotic multigrid convergence factor, ρ. On each mesh size, the Newton iterations were started with initial guess $\mathbf{U}_0^h = \mathbf{0}$.

TABLE 10.1
Numerical results for the pure displacement problem with known smooth solution, without using nested iteration, using $V(1,1)$ cycles.

N	m	$G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f})^{\frac{1}{2}}$	$\mathcal{G}(\mathbf{U}^h; \mathbf{f})^{\frac{1}{2}}$	$\|\mathbf{U}^* - \mathbf{U}^h\|$	ρ
8	1	4.73e-02	4.73e-02	2.16e-03	0.70
8	2	2.58e-02	2.58e-02	1.91e-03	0.67
8	3	2.58e-02	2.58e-02	1.91e-03	0.61
16	1	1.32e-02	4.44e-02	1.31e-03	0.70
16	2	1.29e-02	1.29e-02	4.84e-04	0.69
16	3	1.29e-02	1.29e-02	4.84e-04	0.46
32	1	6.66e-03	4.38e-02	1.26e-03	0.73
32	2	6.44e-03	6.44e-03	1.22e-04	0.73
32	3	6.44e-03	6.44e-03	1.22e-04	0.70
64	1	3.38e-03	4.37e-02	1.26e-03	0.77
64	2	3.22e-03	3.22e-03	3.02e-05	0.75
64	3	3.22e-03	3.22e-03	3.04e-05	0.72
128	1	1.73e-03	4.36e-02	1.27e-03	0.80
128	2	1.61e-03	1.62e-03	7.63e-06	0.79
128	3	1.61e-03	1.61e-03	7.53e-06	0.76

By comparing the functional norm and L^2 error values after three Newton steps on a sequence of levels in Table 10.1, we see that the method achieves the optimal discretization accuracy of $O(h^2)$ with respect to the L^2 error norm, and $O(h)$ with respect to the linearized and nonlinear functional norms. Newton's method essentially converges by the second iteration independent of the mesh size. But, even with such fast convergence, we see that the nonlinear functional values of the first Newton step on each level essentially stall, indicating that, even for this relatively simple problem, the linear elasticity approximation is a poor approximation to the geometrically nonlinear approximation.

We see that, for this problem, the multigrid convergence factors based on $V(1,1)$ cycles are bounded above by about 0.8. While these are acceptable convergence factors, in the remainder of the numerical test problems, we use an AMG $V(1,1)$ preconditioned conjugate gradient cycle to improve performance. We denote these accelerated cycles by $V(1,1)$-*pcg*, and because these cycles generally do not reduce the error by a consistent amount, we report the average convergence factor, $\bar{\rho}$, rather than the asymptotic convergence factor. Refer to [18] for complete details on such cycles.

The convergence factor does not take into account the amount of work done per cycle. For an appropriate measure of the work expended by a multigrid cycle, we define the cycle complexity as the total work per cycle relative to one fine grid relaxation sweep. To obtain a numerical estimate of the cycle complexity, we compute the total number of nonzero matrix entries on each level, multiplied by the number of relaxation sweeps on that level, divided by the number of nonzero matrix entries of the finest level operator. Define the work per Newton step as the work per cycle multiplied by the number of cycles per step, and the total work, W_T, as the cumulative amount of work expended relative to the current level. One such work unit is equivalent to one relaxation sweep on the finest level.

We now wish to solve the same problem as above, but in the most efficient way possible. To this end, we implement the nested iteration strategy described in section 9. Instead of reducing the residual of each linear system by a given amount, we take only three $V(1,1)$-*pcg* cycles per Newton step and one Newton step per level. Table 10.2 summarizes these results.

TABLE 10.2
Numerical results for the pure displacement problem with known smooth solution, using nested iteration and three $V(1,1)$-pcg cycles per step.

N	m	$\mathcal{G}(\mathbf{U}^h;\mathbf{f})^{\frac{1}{2}}$	$\bar{\rho}$	W_T	time (s)
8	2	2.64e-02	0.29	12.3	1
16	3	1.31e-02	0.25	16.0	4
32	4	6.61e-03	0.24	19.0	15
64	5	3.32e-03	0.24	20.8	60
128	6	1.66e-03	0.23	21.6	242

As Tables 10.1 and 10.2 show, the nested iteration method achieves optimal discretization accuracy, and the nonlinear functional on the finest grid is within 5% of discretization error. The average convergence factors for the $V(1,1)$-*pcg* cycles remain bounded and of very reasonable size for this problem. The total amount of work required for the solution at each level is essentially bounded at less than 25 work units, and the time to solution for each level scales almost exactly with the number of degrees of freedom of the problem.

The numerical results presented here are for the pure displacement problem with small strains. In practice, we find that the method performs similarly to the results shown here for mixed boundary conditions and for somewhat larger strains than the theory allows. In the next section, we show that the small strains assumption admits a large class of interesting problems.

11. Validating the small strains assumption. According to Ciarlet in [11], for any homogeneous, isotropic, elastic material, the stress and strain tensors satisfy the relation given by

$$\Sigma(\mathbf{E}) = \lambda \mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E} + o(\mathbf{E}). \tag{11.1}$$

But the model of geometrically nonlinear elasticity uses the linear stress-strain relation given in (3.2), that is, we drop the higher-order terms, $o(\mathbf{E})$, under an assumption of small strains. Thus, in analysis of the geometrically nonlinear elasticity system, we are free to impose reasonable restrictions on the size of $\|\mathbf{E}\|$ without limiting the scope of the model.

In Theorem 6.1, we assume that the strain associated with the solution of each Newton iterate satisfies

$$\|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr} < \frac{\sqrt{2}}{\lambda + 3}, \tag{11.2}$$

where we have scaled the problem such that $\mu = 1$. In this section, we investigate this restriction and provide examples of different configurations and their relation to (11.2) and material constant λ.

Since physically we must have $\lambda > 0$, we first see that an upper bound on the allowed strain is at $\|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr} = \sqrt{2}/3 \approx 0.471$. We further notice that bound (11.2) is always violated by any nonzero strain in the limit as $\lambda \to \infty$. Thus, our

TABLE 11.1
Material constants of homogeneous isotropic materials.

	ν	λ
Rubber	0.49	33.3
Lead	0.44	7.30
Aluminum	0.34	2.15
Nickel	0.30	1.56
Steel	0.28	1.22
Glass	0.25	1.00

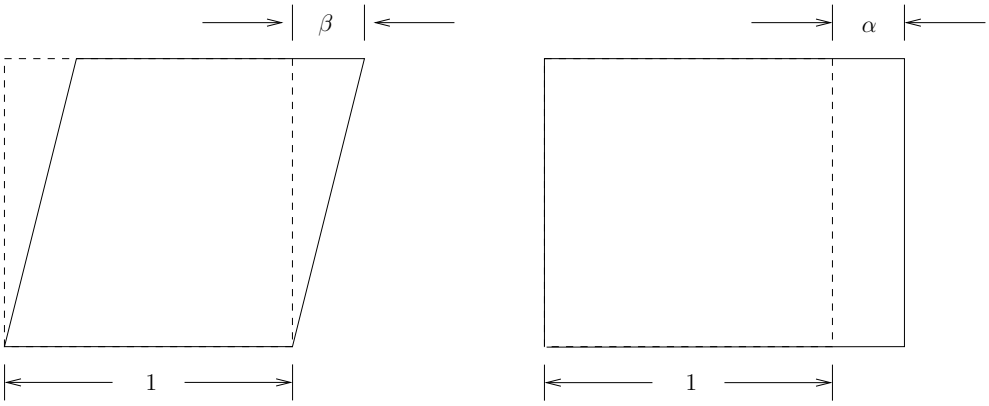


FIG. 11.1. *Pure shear and pure tensile strains.*

notion of "small strains" is coupled to the assumption of compressibility. The Poisson ratio of an elastic material is given by

$$\nu = \frac{\lambda}{2(\lambda + \mu)},$$

and we may think of the incompressible limit as $\lambda \to \infty$ (for bounded μ) or $\nu \to 0.5$. In Table 11.1, we provide a few examples of common materials and their material properties. Because we are chiefly concerned with the value of λ relative to μ, we report the unitless $\lambda \leftarrow \lambda/\mu$. For unscaled constants with meaningful physical units, consult [11].

For the numerical test problems in this paper, we uniformly choose to use $\lambda = 2.15$, that of aluminum, as the level of compressibility.

Consider the two basic modes of strain: shear and tensile strain. A unit square domain under either uniform shear or uniform tensile strain has corresponding displacements of the form

$$\mathbf{u}_{shear} = \begin{pmatrix} \beta y \\ 0 \end{pmatrix} \text{ or } \mathbf{u}_{tensile} = \begin{pmatrix} \alpha x \\ 0 \end{pmatrix}.$$

Parameters β and α determine the extent of deformation as pictured in Figure 11.1.

Under these deformations, we may apply (6.1) and (6.3) to satisfy (11.2) for these two cases. For pure shear strain, we require β and λ to satisfy

$$(\lambda + 3)^2 \beta^2 (\beta^2 + 2) - 2 < 0,$$

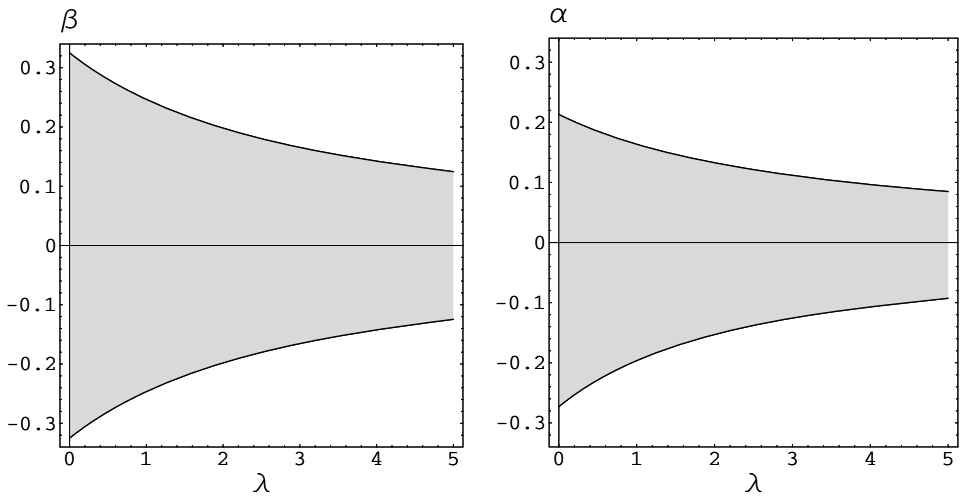


FIG. 11.2. *Shear and tensile strain limits for small strains.*

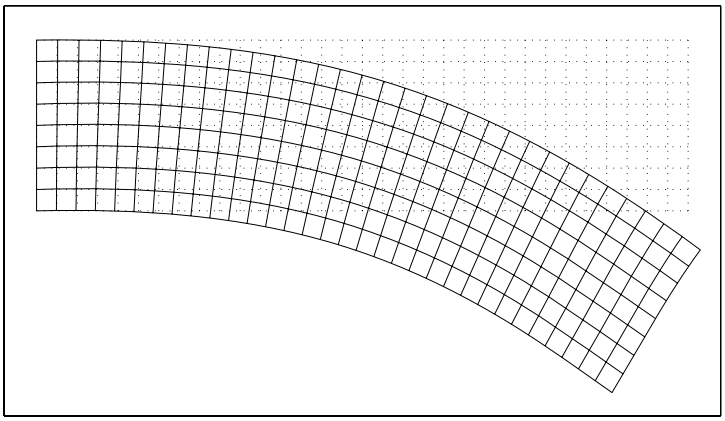


FIG. 11.3. *Displacement plot for cantilever beam displaying "large displacement and small strains."*

and, for pure tensile strain, we require α and λ to satisfy

$$(\lambda + 3)^2\alpha^2(\alpha + 2)^2 - 2 < 0.$$

These relations are satisfied for the parameters in the shaded regions shown in Figure 11.2.

Now consider the following example of a deformed configuration with large displacements but small strains. The strain of the discrete approximation is computed pointwise from (6.3) for mesh size $h = 1/16$. The deformation is from a simple cantilever beam under a constant gravitational force. The max pointwise strain is 0.241 and, for this configuration to satisfy (11.2), the largest allowable λ is approximately 2.88, which corresponds to a Poisson ratio of $\nu = 0.37$. Figure 11.3 shows a plot of the deformed configuration.

REFERENCES

[1] C. Bacuta, J. Bramble, and J. Xu, *Regularity estimates for elliptic boundary value problems with smooth data on polygonal domains*, J. Numer. Math., 11 (2003), pp. 75–94.

[2] C. Bacuta, J. Bramble, and J. Xu, *Regularity estimates for elliptic boundary value problems in Besov spaces*, Math. Comp., 72 (2003), pp. 1577–1599.

[3] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[4] Z. Cai, J. Korsawe, and G. Starke, *An adaptive least squares mixed finite element method for the stress-displacement formulation of linear elasticity*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 132–148.

[5] Z. Cai, C.-O. Lee, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for linear elasticity: Numerical results*, SIAM J. Sci. Comput., 21 (2000), pp. 1706–1727.

[6] Z. Cai, C.-O. Lee, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes linear elasticity equations: Further results*, SIAM J. Sci. Comput., 21 (2000), pp. 1728–1739.

[7] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[8] Z. Cai, T. A. Manteuffel, S. F. McCormick, and S. V. Parter, *First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[9] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: Linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.

[10] Z. Cai and G. Starke, *Least squares methods for linear elasticity*, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.

[11] P. G. Ciarlet, *Mathematical Elasticity, Volume* 1: *Three Dimensional Elasticity*, North-Holland, Amsterdam, 1988.

[12] S. Clain, *Elliptic operators of divergence type with Hölder coefficients in fractional Sobolev spaces*, Rend. Mat. Appl., 17 (1997), pp. 207–236.

[13] A. L. Codd, T. A. Manteuffel, and S. F. McCormick, *Multilevel first-order system least squares for nonlinear elliptic partial differential equations*, SIAM J. Numer. Anal., 41 (2003), pp. 2197–2209.

[14] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[15] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[16] S. D. Kim, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares (FOSLS) for spatial linear elasticity: Pure traction*, SIAM J. Numer. Anal., 38 (2000), pp. 1454–1482.

[17] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, I, Springer-Verlag, New York, 1972.

[18] U. Trottenberg, C. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, San Diego, CA, 2001.

[19] C. Westphal, *First-Order System Least Squares (FOSLS) for Geometrically Nonlinear Elasticity in Nonsmooth Domains*, Ph.D. thesis, University of Colorado, Boulder, 2004.

[20] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1980.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2082–2106

A MULTIPOINT FLUX MIXED FINITE ELEMENT METHOD*

MARY F. WHEELER† AND IVAN YOTOV‡

Abstract. We develop a mixed finite element method for single phase flow in porous media that reduces to cell-centered finite differences on quadrilateral and simplicial grids and performs well for discontinuous full tensor coefficients. Motivated by the multipoint flux approximation method where subedge fluxes are introduced, we consider the lowest order Brezzi–Douglas–Marini (BDM) mixed finite element method. A special quadrature rule is employed that allows for local velocity elimination and leads to a symmetric and positive definite cell-centered system for the pressures. Theoretical and numerical results indicate second-order convergence for pressures at the cell centers and first-order convergence for subedge fluxes. Second-order convergence for edge fluxes is also observed computationally if the grids are sufficiently regular.

Key words. mixed finite element, multipoint flux approximation, cell-centered finite difference, tensor coefficient, error estimates

AMS subject classifications. 65N06, 65N12, 65N15, 65N30, 76S05

DOI. 10.1137/050638473

1. Introduction.

Mixed finite element (MFE) methods have been widely used for modeling flow in porous media due to their local mass conservation, accurate approximation of the velocity, and proper treatment of discontinuous coefficients. A computational drawback of these methods is the need to solve an algebraic system of saddle point type. One possible approach to address this issue is to use the hybrid form of the MFE method [9, 15]. In this case the method can be reduced to a symmetric positive definite system for the pressure Lagrange multipliers on the element faces. Alternatively, it was established in [29] that, in the case of diagonal tensor coefficients and rectangular grids, MFE methods can be reduced to cell-centered finite differences (CCFD) for the pressure through the use of a quadrature rule for the velocity mass matrix. This relationship was explored in [33] to obtain convergence of CCFD on rectangular grids. This result was extended to full tensor coefficients and logically rectangular grids in [7, 6], where the expanded mixed finite element (EMFE) method was introduced. The EMFE method is very accurate for smooth grids and coefficients, but loses accuracy near discontinuities. This is due to the arithmetic averaging of discontinuous coefficients. Higher order accuracy can be recovered if pressure Lagrange multipliers are introduced along discontinuous interfaces [6], but then the cell-centered structure is lost.

Several other methods have been introduced that handle well rough grids and coefficients. The control volume mixed finite element (CVMFE) method [16] is based on discretizing Darcy's law on specially constructed control volumes. Mimetic finite difference (MFD) methods [23] are designed to mimic on the discrete level critical

*Received by the editors August 19, 2005; accepted for publication (in revised form) April 6, 2006; published electronically October 30, 2006.

http://www.siam.org/journals/sinum/44-5/63847.html

†Institute for Computational Engineering and Sciences (ICES), Department of Aerospace Engineering & Engineering Mechanics and Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712 (mfw@ices.utexas.edu). The research of this author was partially supported by NSF grant DMS 0411413 and DOE grant DE-FGO2-04ER25617.

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (yotov@math.pitt.edu). The research of this author was supported in part by DOE grant DE-FG02-04ER25618 and NSF grant DMS 0411694.

properties of the differential operators. The approximating spaces in both methods are closely related to RT_0, the lowest order Raviart–Thomas MFE spaces [27]. These relationships have been explored in [17, 30] and [10, 12] to establish convergence of the CVMFE methods and the MFD methods, respectively. However, as in the case of MFE methods, both methods lead to an algebraic saddle point problem. The multipoint flux approximation (MPFA) method [1, 2, 19, 20] has been developed as a finite volume method and combines the advantages of the above mentioned methods; i.e., it is accurate for rough grids and coefficients and reduces to a cell-centered stencil for the pressures. However, due to the nonvariational formulation of the MPFA, there exist only limited theoretical results in the literature for the well posedness and convergence of this method [24].

In this paper we design a MFE method that reduces to accurate CCFD for full tensors and irregular grids and performs well for discontinuous coefficients. Motivated by the MPFA [2, 20], where subedge fluxes are introduced, we consider the lowest order Brezzi–Douglas–Marini (BDM) MFE method [14, 15]. In two dimensions, for example, there are two velocity degrees of freedom per edge. A special quadrature rule is employed that allows for local velocity elimination and leads to a cell-centered stencil for the pressures. The resulting algebraic system is symmetric and positive definite. We call our method a multipoint flux mixed finite element (MFMFE) method, due to its close relationship with the MPFA method.

We emphasize that the formulation of the MFMFE method involves K^{-1}; see (2.41)–(2.42). For diagonal discontinuous K, the resulting coefficient is a harmonic average. This explains the superior performance of the MFMFE method for problems with rough grids and coefficients, compared to the EMFE method.

The MFMFE method results in a smaller algebraic system than the hybrid MFE method does, since finite element partitions have fewer elements than edges or faces. Moreover, many existing petroleum simulators are based on cell-centered discretizations and their data structures are more compatible with the MFMFE method than with the hybrid MFE method.

The variational framework allows for MFE analysis tools to be combined with quadrature error analysis to establish well posedness and accuracy of the MFMFE method. We formulate and analyze the method on simplicial grids in two and three dimensions as well as on quadrilateral grids. We obtain first order convergence for the pressure in the L^2-norm and for the velocity in the $H(\mathrm{div})$-norm. A duality argument is employed to establish second order convergence for the pressure in a discrete L^2-norm involving the centers of mass of the elements.

The analysis in the quadrilateral case is more involved, since it requires mapping to a reference element. As a result a restriction needs to be imposed on the geometry of each quadrilateral, namely, that it is an $O(h^2)$-perturbation of a parallelogram; see (3.1). We have verified numerically that this restriction is not just an artifact of the analysis, but is needed in practice as well. We also note that second order convergence is observed numerically for the velocities at the midpoints of the edges on h^2-parallelogram grids.

The techniques used in this paper can be employed to formulate and analyze extensions of the MFMFE method to nonmatching multiblock grids via mortar finite elements in the spirit of [5], multiscale MFMFE methods in the spirit of [4], and adaptive mortar MFMFE methods in the spirit of [34].

The rest of the paper is organized as follows. The method is developed in section 2. Sections 3 and 4 are devoted to the error analysis of the velocity and the pressure, respectively. Numerical experiments are presented in section 5. We end with some

conclusions in section 6.

2. Definition of the method.

2.1. Preliminaries. We consider the second order elliptic problem written as a system of two first order equations,

$$\mathbf{u} = -K\nabla p \quad \text{in } \Omega, \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = f \quad \text{in } \Omega, \tag{2.2}$$

$$p = g \quad \text{on } \Gamma_D, \tag{2.3}$$

$$\mathbf{u} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N, \tag{2.4}$$

where the domain $\Omega \subset \mathbf{R}^d$, $d = 2$ or 3, has a boundary $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, measure$(\Gamma_D) > 0$, $\mathbf{n}$ is the outward unit normal on $\partial\Omega$, and K is a symmetric, uniformly positive definite tensor satisfying, for some $0 < k_0 \le k_1 < \infty$,

$$k_0\xi^T\xi \le \xi^T K(\mathbf{x})\xi \le k_1\xi^T\xi \quad \forall \mathbf{x} \in \Omega \ \forall \xi \in \mathbf{R}^d. \tag{2.5}$$

In flow in porous media modeling, p is the pressure, $\mathbf{u}$ is the Darcy velocity, and K represents the permeability divided by the viscosity. The choice of boundary conditions is made for the sake of simplicity. More general boundary conditions, including nonhomogeneous full Neumann problems, can also be treated.

Throughout this paper, C denotes a generic positive constant that is independent of the discretization parameter h. We will also use the following standard notation. For a domain $G \subset \mathbf{R}^d$, the $L^2(G)$ inner product and norm for scalar and vector valued functions are denoted $(\cdot,\cdot)_G$ and $\|\cdot\|_G$, respectively. The norms and seminorms of the Sobolev spaces $W^{k,p}(G)$, $k \in \mathbf{R}$, $p > 0$ are denoted by $\|\cdot\|_{k,p,G}$ and $|\cdot|_{k,p,G}$, respectively. The norms and seminorms of the Hilbert spaces $H^k(G)$ are denoted by $\|\cdot\|_{k,G}$ and $|\cdot|_{k,G}$, respectively. We omit G in the subscript if $G = \Omega$. For a section of the domain or element boundary $S \subset \mathbf{R}^{d-1}$ we write $\langle\cdot,\cdot\rangle_S$ and $\|\cdot\|_S$ for the $L^2(S)$ inner product (or duality pairing) and norm, respectively. For a tensor-valued function M, let $\|M\|_\alpha = \max_{i,j}\|M_{ij}\|_\alpha$ for any norm $\|\cdot\|_\alpha$. We will also use the space

$$H(\mathrm{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$$

equipped with the norm

$$\|\mathbf{v}\|_{\mathrm{div}} = (\|\mathbf{v}\|^2 + \|\nabla \cdot \mathbf{v}\|^2)^{1/2}.$$

The weak formulation of (2.1)–(2.4) is the following: find $\mathbf{u} \in \mathbf{V}$ and $p \in W$ such that

$$(K^{-1}\mathbf{u}, \mathbf{v}) = (p, \nabla \cdot \mathbf{v}) - \langle g, \mathbf{v} \cdot \mathbf{n}\rangle_{\Gamma_D}, \quad \mathbf{v} \in \mathbf{V}, \tag{2.6}$$

$$(\nabla \cdot \mathbf{u}, w) = (f, w), \quad w \in W, \tag{2.7}$$

where

$$\mathbf{V} = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma_N\}, \quad W = L^2(\Omega).$$

It is well known [15, 28] that (2.6)–(2.7) has a unique solution.

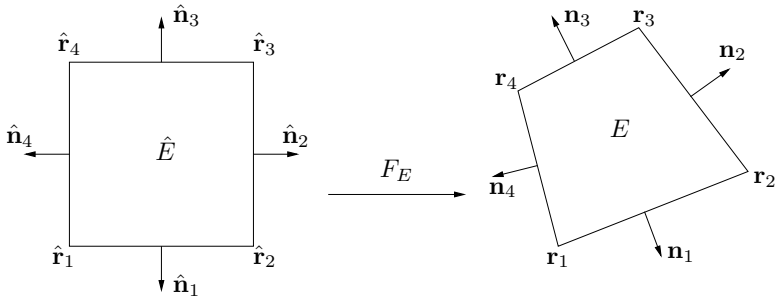


FIG. 2.1. *Mapping in the case of a quadrilateral.*

2.2. Finite element mappings. Consider a polygonal domain $\Omega \in \mathbf{R}^d$ and let $\mathcal{T}_h$ be a finite element partition of Ω consisting of triangles and/or convex quadrilaterals in two dimensions and tetrahedra in three dimensions, where $h = \max_{E\in\mathcal{T}_h} \operatorname{diam}(E)$. We assume that $\mathcal{T}_h$ is shape regular and quasi-uniform [18]. For any element $E \in \mathcal{T}_h$ there exists a bijection mapping $F_E : \hat{E} \to E$ where $\hat{E}$ is the reference element. Denote the Jacobian matrix by DF_E and let $J_E = |\det(DF_E)|$. Denote the inverse mapping by F_E^{-1}, its Jacobian matrix by DF_E^{-1}, and let $J_{F_E^{-1}} = |\det(DF_E^{-1})|$. We have that

$$DF_E^{-1}(x) = (DF_E)^{-1}(\hat{x}), \qquad J_{F_E^{-1}}(x) = \frac{1}{J_E(\hat{x})}.$$

In the case of convex quadrilaterals, $\hat{E}$ is the unit square with vertices $\hat{\mathbf{r}}_1 = (0,\,0)^T$, $\hat{\mathbf{r}}_2 = (1,\,0)^T$, $\hat{\mathbf{r}}_3 = (1,\,1)^T$, and $\hat{\mathbf{r}}_4 = (0,\,1)^T$. Denote by $\mathbf{r}_i = (x_i,\,y_i)^T$, $i = 1,\ldots,4$, the four corresponding vertices of element E as shown in Figure 2.1. The outward unit normal vectors to the edges of E and $\hat{E}$ are denoted by $\mathbf{n}_i$ and $\hat{\mathbf{n}}_i$, $i = 1,\ldots,4$, respectively. In this case F_E is the bilinear mapping given by

$$\begin{aligned} F_E(\hat{\mathbf{r}}) &= \mathbf{r}_1\,(1-\hat{x})(1-\hat{y}) + \mathbf{r}_2\,\hat{x}(1-\hat{y}) + \mathbf{r}_3\,\hat{x}\hat{y} + \mathbf{r}_4\,(1-\hat{x})\hat{y} \\ &= \mathbf{r}_1 + \mathbf{r}_{21}\hat{x} + \mathbf{r}_{41}\hat{y} + (\mathbf{r}_{34} - \mathbf{r}_{21})\hat{x}\hat{y}, \end{aligned} \tag{2.8}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$. It is easy to see that DF_E and J_E are linear functions of $\hat{x}$ and $\hat{y}$:

$$\begin{aligned} DF_E &= [(1-\hat{y})\,\mathbf{r}_{21} + \hat{y}\,\mathbf{r}_{34},\ (1-\hat{x})\,\mathbf{r}_{41} + \hat{x}\,\mathbf{r}_{32}] \\ &= [\mathbf{r}_{21}, \mathbf{r}_{41}] + [(\mathbf{r}_{34} - \mathbf{r}_{21})\hat{y}, (\mathbf{r}_{34} - \mathbf{r}_{21})\hat{x}], \end{aligned} \tag{2.9}$$

$$J_E = 2|T_1| + 2(|T_2| - |T_1|)\hat{x} + 2(|T_4| - |T_1|)\hat{y}, \tag{2.10}$$

where $|T_i|$ is the area of the triangle formed by the two edges sharing $\mathbf{r}_i$. Since E is convex, the Jacobian determinant J_E is uniformly positive, i.e., $J_E(\hat{x},\hat{y}) > 0$.

In the case of triangles, $\hat{E}$ is the reference right triangle with vertices $\hat{\mathbf{r}}_1 = (0,\,0)^T$, $\hat{\mathbf{r}}_2 = (1,\,0)^T$, and $\hat{\mathbf{r}}_3 = (0,\,1)^T$. Let $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$ be the corresponding vertices of E, oriented in a counterclockwise direction. The linear mapping for triangles has the form

$$F_E(\hat{\mathbf{r}}) = \mathbf{r}_1(1-\hat{x}-\hat{y}) + \mathbf{r}_2\hat{x} + \mathbf{r}_3\hat{y}, \tag{2.11}$$

with respective Jacobian matrix and Jacobian determinant

$$DF_E = [\mathbf{r}_{21}, \mathbf{r}_{31}]^T \quad \text{and} \quad J_E = 2|E|. \tag{2.12}$$

The mapping in the case of tetrahedra is described similarly to the triangular case. Note that in the case of simplicial elements the mapping is affine and the Jacobian matrix and its determinant are constants.

Using the mapping definitions (2.8)–(2.12), it is easy to check that for any edge (face) $e_i \subset \partial E$

$$\mathbf{n}_i = \frac{1}{|e_i|} J_E (DF_E^{-1})^T \hat{\mathbf{n}}_i. \tag{2.13}$$

It is also easy to see that, for all element types, the mapping definitions and the shape-regularity and quasiuniformity of the grids imply that

$$\|DF_E\|_{0,\infty,\hat{E}} \sim h, \quad \|J_E\|_{0,\infty,\hat{E}} \sim h^d, \quad \text{and} \quad \|J_{F_E^{-1}}\|_{0,\infty,\hat{E}} \sim h^{-d} \quad \forall\, E \in \mathcal{T}_h, \tag{2.14}$$

where the notation $a \sim b$ means that there exist positive constants c_0 and c_1 independent of h such that $c_0 b \le a \le c_1 b$.

2.3. Mixed finite element spaces. Let $\mathbf{V}_h \times W_h$ be the lowest order BDM_1 MFE spaces [14, 15]. On the reference unit square these spaces are defined as

$$\begin{aligned} \hat{\mathbf{V}}(\hat{E}) &= P_1(\hat{E})^2 + r\ \mathrm{curl}(\hat{x}^2\hat{y}) + s\ \mathrm{curl}(\hat{x}\hat{y}^2) \\ &= \begin{pmatrix} \alpha_1\hat{x} + \beta_1\hat{y} + \gamma_1 + r\hat{x}^2 + 2s\hat{x}\hat{y} \\ \alpha_2\hat{x} + \beta_2\hat{y} + \gamma_2 - 2r\hat{x}\hat{y} - s\hat{y}^2 \end{pmatrix}, \quad \hat{W}(\hat{E}) = P_0(\hat{E}), \end{aligned} \tag{2.15}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2, s, r$ are real constants and P_k denotes the space of polynomials of degree $\le k$. In the case where the reference element $\hat{E}$ is the unit triangle or tetrahedron, the BDM_1 spaces are defined as

$$\hat{\mathbf{V}}(\hat{E}) = P_1(\hat{E})^d, \quad \hat{W}(\hat{E}) = P_0(\hat{E}). \tag{2.16}$$

Note that in all three cases $\hat{\nabla} \cdot \hat{\mathbf{V}}(\hat{E}) = \hat{W}(\hat{E})$ and that for all $\hat{\mathbf{v}} \in \hat{\mathbf{V}}(\hat{E})$ and for any edge (or face) $\hat{e}$ of $\hat{E}$,

$$\hat{\mathbf{v}} \cdot \hat{\mathbf{n}}_{\hat{e}} \in P_1(\hat{e}).$$

It is well known [14, 15] that the degrees of freedom for $\hat{\mathbf{V}}(\hat{E})$ can be chosen to be the values of $\hat{\mathbf{v}} \cdot \hat{\mathbf{n}}_{\hat{e}}$ at any two points on each edge $\hat{e}$ if $\hat{E}$ is the unit triangle or the unit square, or any three points on each face $\hat{e}$ if $\hat{E}$ is the unit tetrahedron. *We choose these points to be the vertices of* $\hat{e}$; see Figure 2.2 for the quadrilateral case. This choice is motivated by the requirement of accuracy and certain orthogonalities for the quadrature rule introduced in the next section.

The BDM_1 spaces on any element $E \in \mathcal{T}_h$ are defined via the transformations

$$\mathbf{v} \leftrightarrow \hat{\mathbf{v}} : \mathbf{v} = \frac{1}{J_E} DF_E \hat{\mathbf{v}} \circ F_E^{-1}, \quad w \leftrightarrow \hat{w} : w = \hat{w} \circ F_E^{-1}.$$

The vector transformation is known as the Piola transformation. It is designed to preserve the normal components of the velocity vectors on the edges (faces) and satisfies the important properties [15]

$$(\nabla \cdot \mathbf{v}, w)_E = (\hat{\nabla} \cdot \hat{\mathbf{v}}, \hat{w})_{\hat{E}} \qquad \text{and} \qquad \langle \mathbf{v} \cdot \mathbf{n}_e, w \rangle_e = \langle \hat{\mathbf{v}} \cdot \hat{\mathbf{n}}_{\hat{e}}, \hat{w} \rangle_{\hat{e}}. \tag{2.17}$$

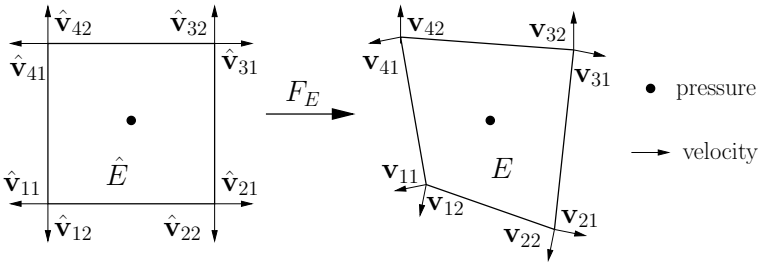


FIG. 2.2. *Degrees of freedom and basis functions for the BDM_1 spaces on quadrilaterals.*

Moreover, (2.13) implies

$$\mathbf{v}\cdot\mathbf{n}_e = \frac{1}{J_E} DF_E\hat{\mathbf{v}}\cdot\frac{1}{|e|}J_E(DF_E^{-1})^T\hat{\mathbf{n}}_{\hat{e}} = \frac{1}{|e|}\hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_{\hat{e}}. \tag{2.18}$$

Also note that the first equation in (2.17) and $(\nabla\cdot\mathbf{v}, w)_E = (\widehat{\nabla\cdot\mathbf{v}}, \hat{w}J_E)_{\hat{E}}$ imply

$$\nabla\cdot\mathbf{v} = \left(\frac{1}{J_E}\hat{\nabla}\cdot\hat{\mathbf{v}}\right)\circ F_E^{-1}(\mathbf{x}). \tag{2.19}$$

Therefore on quadrilaterals $\nabla\cdot\mathbf{v}|_E \neq$ constant.

The BDM_1 spaces on $\mathcal{T}_h$ are given by

$$\begin{aligned} \mathbf{V}_h &= \{\mathbf{v}\in\mathbf{V}\colon \quad \mathbf{v}|_E \leftrightarrow \hat{\mathbf{v}},\ \hat{\mathbf{v}}\in\hat{\mathbf{V}}(\hat{E}) \quad \forall E\in\mathcal{T}_h\},\\ W_h &= \{w\in W\colon \quad w|_E \leftrightarrow \hat{w},\ \hat{w}\in\hat{W}(\hat{E}) \quad \forall E\in\mathcal{T}_h\}. \end{aligned} \tag{2.20}$$

It is known [14, 15, 32] that there exists a projection operator Π from $\mathbf{V}\cap(H^1(\Omega))^d$ onto $\mathbf{V}_h$ satisfying

$$(\nabla\cdot(\Pi\mathbf{q}-\mathbf{q}), w) = 0 \qquad \forall\, w\in W_h. \tag{2.21}$$

The operator Π is defined locally on each element E by

$$\Pi\mathbf{q}\leftrightarrow\widehat{\Pi\mathbf{q}}, \quad \widehat{\Pi\mathbf{q}} = \hat{\Pi}\hat{\mathbf{q}}, \tag{2.22}$$

where $\hat{\Pi}: (H^1(\hat{E}))^d \to \hat{\mathbf{V}}(\hat{E})$ is the reference element projection operator satisfying

$$\forall\,\hat{e}\subset\partial\hat{E}, \qquad \langle(\hat{\Pi}\hat{\mathbf{q}}-\hat{\mathbf{q}})\cdot\hat{\mathbf{n}}, \hat{p}_1\rangle_{\hat{e}} = 0 \qquad \forall\,\hat{p}_1\in P_1(\hat{e}). \tag{2.23}$$

To see that $\Pi\mathbf{q}\cdot\mathbf{n} = 0$ on Γ_N if $\mathbf{q}\cdot\mathbf{n} = 0$ on Γ_N, note that for any $e\in\Gamma_N$ and for all $p_1\leftrightarrow\hat{p}_1\in P_1(\hat{e})$,

$$\langle\Pi\mathbf{q}\cdot\mathbf{n}, p_1\rangle_e = \langle\widehat{\Pi\mathbf{q}}\cdot\hat{\mathbf{n}}, \hat{p}_1\rangle_{\hat{e}} = \langle\hat{\Pi}\hat{\mathbf{q}}\cdot\hat{\mathbf{n}}, \hat{p}_1\rangle_{\hat{e}} = \langle\hat{\mathbf{q}}\cdot\hat{\mathbf{n}}, \hat{p}_1\rangle_{\hat{e}} = 0,$$

implying $\Pi\mathbf{q}\cdot\mathbf{n} = 0$, where we have used (2.17), (2.22), and (2.23).

In addition to the mixed projection operator Π onto $\mathbf{V}_h$, we will use a similar projection operator onto the lowest order Raviart–Thomas spaces [27, 15]. The RT_0 spaces are defined on the unit square as

$$\hat{\mathbf{V}}^0(\hat{E}) = \begin{pmatrix}\alpha_1+\beta_1\hat{x}\\ \alpha_2+\beta_2\hat{y}\end{pmatrix}, \quad \hat{W}^0(\hat{E}) = P_0(\hat{E}), \tag{2.24}$$

and on the unit triangle as

$$\hat{\mathbf{V}}^0(\hat{E}) = \begin{pmatrix} \alpha_1 + \beta\hat{x} \\ \alpha_2 + \beta\hat{y} \end{pmatrix}, \quad \hat{W}^0(\hat{E}) = P_0(\hat{E}). \tag{2.25}$$

On the unit tetrahedron $\hat{\mathbf{V}}^0(\hat{E})$ has an additional component $\alpha_3 + \beta\hat{z}$. In all cases $\hat{\nabla}\cdot\hat{\mathbf{V}}^0(\hat{E}) = \hat{W}^0(\hat{E})$ and $\hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_{\hat{e}} \in P_0(\hat{e})$. The degrees of freedom of $\hat{\mathbf{V}}^0(\hat{E})$ are the values of $\hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_{\hat{e}}$ at the midpoints of all edges (faces) $\hat{e}$. The projection operator $\hat{\Pi}_0 : (H^1(\hat{E}))^d \to \hat{\mathbf{V}}^0(\hat{E})$ satisfies

$$\forall\, \hat{e} \subset \partial\hat{E}, \qquad \langle(\hat{\Pi}_0\hat{\mathbf{q}} - \hat{\mathbf{q}})\cdot\hat{\mathbf{n}}, \hat{p}_0\rangle_{\hat{e}} = 0 \qquad \forall\, \hat{p}_0 \in P_0(\hat{e}). \tag{2.26}$$

The spaces $\mathbf{V}_h^0$ and W_h^0 on $\mathcal{T}_h$ and the projection operator $\Pi_0 : (H^1(\Omega))^d \to \mathbf{V}_h^0$ are defined similarly to the case of BDM_1 spaces. Note that $\mathbf{V}_h^0 \subset \mathbf{V}_h$ and $W_h^0 = W_h$. It follows immediately from the definition of Π_0 that

$$\nabla\cdot\mathbf{v} = \nabla\cdot\Pi_0\mathbf{v} \quad \forall\, \mathbf{v} \in \mathbf{V}_h \tag{2.27}$$

and

$$\|\Pi_0\mathbf{v}\| \le C\|\mathbf{v}\| \quad \forall\, \mathbf{v} \in \mathbf{V}_h. \tag{2.28}$$

2.4. The BDM_1 method. The BDM_1 mixed finite element method is based on approximating the variational formulation (2.6)–(2.7) in the discrete spaces $\mathbf{V}_h \times W_h$: find $\mathbf{u}_h^{bdm} \in \mathbf{V}_h$ and $p_h^{bdm} \in W_h$ such that

$$(K^{-1}\mathbf{u}_h^{bdm}, \mathbf{v}) = (p_h^{bdm}, \nabla\cdot\mathbf{v}) - \langle g, \mathbf{v}\cdot\mathbf{n}\rangle_{\Gamma_D}, \quad \mathbf{v} \in \mathbf{V}_h, \tag{2.29}$$

$$(\nabla\cdot\mathbf{u}_h^{bdm}, w) = (f, w), \quad w \in W_h. \tag{2.30}$$

The method has a unique solution and is second order accurate for the velocity and first order accurate for the pressure in L^2-norms on affine grids [14, 32]. It handles well discontinuous coefficients due to the presence of K^{-1} in the mass matrix. A drawback is that the resulting algebraic system is a large coupled velocity-pressure system of a saddle point problem type. In the next section we develop a quadrature rule that allows for local elimination of the velocities and results in a positive definite cell-centered pressure matrix.

2.5. A quadrature rule. For $\mathbf{q}, \mathbf{v} \in \mathbf{V}_h$, define the global quadrature rule

$$(K^{-1}\mathbf{q}, \mathbf{v})_Q \equiv \sum_{E\in\mathcal{T}_h} (K^{-1}\mathbf{q}, \mathbf{v})_{Q,E}.$$

The integration on any element E is performed by mapping to the reference element $\hat{E}$. The quadrature rule is defined on $\hat{E}$. Using the definition (2.20) of the finite element spaces and omitting the subscript E, we have

$$\begin{aligned}\int_E K^{-1}\mathbf{q}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} &= \int_{\hat{E}} \hat{K}^{-1}\frac{1}{J}DF\hat{\mathbf{q}}\cdot\frac{1}{J}DF\hat{\mathbf{v}}\,J\,\mathrm{d}\hat{\mathbf{x}} \\ &= \int_{\hat{E}} \frac{1}{J}DF^T\hat{K}^{-1}DF\hat{\mathbf{q}}\cdot\hat{\mathbf{v}}\,\mathrm{d}\hat{\mathbf{x}} \equiv \int_{\hat{E}} \mathcal{K}^{-1}\hat{\mathbf{q}}\cdot\hat{\mathbf{v}}\,\mathrm{d}\hat{\mathbf{x}},\end{aligned}$$

where

$$\mathcal{K} = JDF^{-1}\hat{K}(DF^{-1})^T. \tag{2.31}$$

Clearly, due to (2.14),

$$\|\mathcal{K}\|_{0,\infty,\hat{E}} \sim h^{d-2}\|K\|_{0,\infty,E} \quad \text{and} \quad \|\mathcal{K}^{-1}\|_{0,\infty,\hat{E}} \sim h^{2-d}\|K^{-1}\|_{0,\infty,E}. \tag{2.32}$$

The quadrature rule on an element E is defined as

$$(K^{-1}\mathbf{q},\mathbf{v})_{Q,E} \equiv (\mathcal{K}^{-1}\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{Q},\hat{E}} \equiv \frac{|\hat{E}|}{s}\sum_{i=1}^{s}\mathcal{K}^{-1}(\hat{\mathbf{r}}_i)\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)\cdot\hat{\mathbf{v}}(\hat{\mathbf{r}}_i), \tag{2.33}$$

where $s = 3$ for the unit triangle and $s = 4$ for the unit square or the unit tetrahedron. Note that on the unit square this is the trapezoidal quadrature rule.

The corner vector $\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)$ is uniquely determined by its normal components to the two edges (or three faces) that share that vertex. Recall that we chose the velocity degrees of freedom on any edge (face) $\hat{e}$ to be the normal components at the vertices of $\hat{e}$. Therefore, there are two (three) degrees of freedom associated with each corner $\hat{\mathbf{r}}_i$ and they uniquely determine the corner vector $\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)$. More precisely,

$$\hat{\mathbf{q}}(\hat{\mathbf{r}}_i) = \sum_{j=1}^{d}\hat{\mathbf{q}}\cdot\hat{\mathbf{n}}_{ij}(\hat{\mathbf{r}}_i)\hat{\mathbf{n}}_{ij},$$

where $\hat{\mathbf{n}}_{ij}$, $j = 1,\dots,d$, are the outward unit normal vectors to the two edges (three faces) intersecting at $\hat{\mathbf{r}}_i$, and $\hat{\mathbf{q}}\cdot\hat{\mathbf{n}}_{ij}(\hat{\mathbf{r}}_i)$ are the velocity degrees of freedom associated with this corner. Let us denote the basis functions associated with $\hat{\mathbf{r}}_i$ by $\hat{\mathbf{v}}_{ij}$, $j = 1,\dots,d$; see Figure 2.2, i.e.,

$$\hat{\mathbf{v}}_{ij}\cdot\hat{\mathbf{n}}_{ij}(\hat{\mathbf{r}}_i) = 1,\ \ \hat{\mathbf{v}}_{ij}\cdot\hat{\mathbf{n}}_{ik}(\hat{\mathbf{r}}_i) = 0,\ k \neq j,\ \text{and}\ \hat{\mathbf{v}}_{ij}\cdot\hat{\mathbf{n}}_{lk}(\hat{\mathbf{r}}_l) = 0,\ l \neq i,\ k = 1,\dots,d.$$

Clearly the quadrature rule (2.33) couples only the two (or three) basis functions associated with a corner. On the unit square, for example,

$$(\mathcal{K}^{-1}\hat{\mathbf{v}}_{11},\hat{\mathbf{v}}_{11})_{\hat{Q},\hat{E}} = \frac{\mathcal{K}_{11}^{-1}(\hat{\mathbf{r}}_1)}{4}, \quad (\mathcal{K}^{-1}\hat{\mathbf{v}}_{11},\hat{\mathbf{v}}_{12})_{\hat{Q},\hat{E}} = \frac{\mathcal{K}_{12}^{-1}(\hat{\mathbf{r}}_1)}{4}, \tag{2.34}$$

and

$$(\mathcal{K}^{-1}\hat{\mathbf{v}}_{11},\hat{\mathbf{v}}_{ij})_{\hat{Q},\hat{E}} = 0 \quad \forall\, ij \neq 11, 12. \tag{2.35}$$

Remark 2.1. The quadrature rule can be defined directly on an element E. It is easy to see from (2.10) and (2.12) that on simplicial elements

$$(K^{-1}\mathbf{q},\mathbf{v})_{Q,E} = \frac{|E|}{s}\sum_{i=1}^{s}K^{-1}(\mathbf{r}_i)\mathbf{q}(\mathbf{r}_i)\cdot\mathbf{v}(\mathbf{r}_i), \tag{2.36}$$

and on quadrilaterals

$$(K^{-1}\mathbf{q},\mathbf{v})_{Q,E} = \frac{1}{2}\sum_{i=1}^{4}|T_i|K^{-1}(\mathbf{r}_i)\mathbf{q}(\mathbf{r}_i)\cdot\mathbf{v}(\mathbf{r}_i). \tag{2.37}$$

The above quadrature rules are closely related to some inner products used in the mimetic finite difference methods [23]. We note that in the case of quadrilaterals, it is simpler to evaluate the quadrature rule on the reference element $\hat{E}$.

Denote the element quadrature error by

$$\sigma_E(K^{-1}\mathbf{q},\mathbf{v}) \equiv (K^{-1}\mathbf{q},\mathbf{v})_E - (K^{-1}\mathbf{q},\mathbf{v})_{Q,E} \tag{2.38}$$

and define the global quadrature error by $\sigma(K^{-1}\mathbf{q},\mathbf{v})|_E = \sigma_E(K^{-1}\mathbf{q},\mathbf{v})$. Similarly, denote the quadrature error on the reference element by

$$\hat{\sigma}_{\hat{E}}(\mathcal{K}^{-1}\hat{\mathbf{q}},\hat{\mathbf{v}}) \equiv (\mathcal{K}^{-1}\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{E}} - (\mathcal{K}^{-1}\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{Q},\hat{E}}. \tag{2.39}$$

The next two lemmas will be used in the analysis.

LEMMA 2.1. *On simplicial elements, if* $\mathbf{q} \in \mathbf{V}_h(E)$, *then*

$$\sigma_E(\mathbf{q},\mathbf{v}_0) = 0 \quad \textit{for all constant vectors } \mathbf{v}_0.$$

Proof. It is enough to consider $\mathbf{v}_0 = (1,0)^T$ or $\mathbf{v}_0 = (1,0,0)^T$; the arguments for the other cases are similar. We have

$$(\mathbf{q},\mathbf{v}_0)_{Q,E} = \frac{|E|}{s}\sum_{i=1}^{s} q_1(\mathbf{r}_i) = \int_E \mathbf{q}\cdot\mathbf{v}_0\,d\mathbf{x},$$

using that the quadrature rule $(\varphi)_E = \frac{|E|}{s}\sum_{i=1}^{s}\varphi(\mathbf{r}_i)$ is exact for linear functions. ∎

LEMMA 2.2. *On the reference square, for any* $\hat{\mathbf{q}} \in \hat{\mathbf{V}}(\hat{E})$,

$$(\hat{\mathbf{q}} - \hat{\Pi}_0\hat{\mathbf{q}},\hat{\mathbf{v}}_0)_{\hat{Q},\hat{E}} = 0 \quad \textit{for all constant vectors } \hat{\mathbf{v}}_0. \tag{2.40}$$

Proof. On any edge $\hat{e}$, if the degrees of freedom of $\hat{\mathbf{q}}$ are $\hat{q}_{\hat{e},1}$ and $\hat{q}_{\hat{e},2}$, then (2.26) and an application of the trapezoidal quadrature rule imply that $\hat{\Pi}_0\hat{\mathbf{q}}|_{\hat{e}} = (\hat{q}_{\hat{e},1} + \hat{q}_{\hat{e},2})/2$. The assertion of the lemma follows from a simple calculation, using (2.33). ∎

2.6. The multipoint flux mixed finite element method. We are now ready to define our method. We seek $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in W_h$ such that

$$(K^{-1}\mathbf{u}_h,\mathbf{v})_Q = (p_h,\nabla\cdot\mathbf{v}) - \langle g,\mathbf{v}\cdot\mathbf{n}\rangle_{\Gamma_D}, \quad \mathbf{v}\in\mathbf{V}_h, \tag{2.41}$$

$$(\nabla\cdot\mathbf{u}_h,w) = (f,w), \quad w\in W_h. \tag{2.42}$$

Remark 2.2. We call the method (2.41)–(2.42) a MFMFE method, since it is related to the MPFA method.

To prove that (2.41)–(2.42) is well posed, we first show that the quadrature rule (2.33) produces a coercive bilinear form. We will need the following auxiliary result.

LEMMA 2.3. *If* $E \in \mathcal{T}_h$ *and* $\mathbf{q} \in (L^2(E))^d$, *then*

$$\|\mathbf{q}\|_E \sim h^{\frac{2-d}{2}}\|\hat{\mathbf{q}}\|_{\hat{E}}. \tag{2.43}$$

Proof. The assertion of the lemma follows from the relations

$$\int_E \mathbf{q}\cdot\mathbf{q}\,d\mathbf{x} = \int_{\hat{E}} \frac{1}{J}DF\hat{\mathbf{q}}\cdot\frac{1}{J}DF\hat{\mathbf{q}}\,J\,d\hat{\mathbf{x}},$$

$$\int_{\hat{E}} \hat{\mathbf{q}}\cdot\hat{\mathbf{q}}\,d\hat{\mathbf{x}} = \int_E \frac{1}{J_{F^{-1}}}DF^{-1}\mathbf{q}\cdot\frac{1}{J_{F^{-1}}}DF^{-1}\mathbf{q}J_{F^{-1}}\,d\mathbf{x},$$

and bounds (2.14). ∎

LEMMA 2.4. *There exists a positive constant C independent of h such that*

$$(K^{-1}\mathbf{q},\mathbf{q})_Q \geq C\|\mathbf{q}\|^2 \quad \forall \mathbf{q}\in \mathbf{V}_h. \tag{2.44}$$

Proof. Let $\mathbf{q}=\sum_{i=1}^{s}\sum_{j=1}^{d} q_{ij}\mathbf{v}_{ij}$ on an element E. Using (2.36)–(2.37) and (2.5) we obtain

$$(K^{-1}\mathbf{q},\mathbf{q})_{Q,E} \geq C\frac{|E|}{k_1}\sum_{i=1}^{s}\mathbf{q}(\mathbf{r}_i)\cdot\mathbf{q}(\mathbf{r}_i) \geq C\frac{|E|}{k_1}\sum_{i=1}^{s}\sum_{j=1}^{d} q_{ij}^2.$$

On the other hand,

$$\|\mathbf{q}\|_E^2 = \left(\sum_{i=1}^{s}\sum_{j=1}^{d} q_{ij}\mathbf{v}_{ij}, \sum_{k=1}^{s}\sum_{l=1}^{d} q_{kl}\mathbf{v}_{kl}\right) \leq C|E|\sum_{i=1}^{s}\sum_{j=1}^{d} q_{ij}^2.$$

A combination of the above two estimates implies the assertion of the lemma. ∎

COROLLARY 2.5. *The bilinear form $(K^{-1}\mathbf{q},\mathbf{v})_Q$ is an inner product in $\mathbf{V}_h$ and $(K^{-1}\mathbf{q},\mathbf{q})_Q^{1/2}$ is a norm in $\mathbf{V}_h$ equivalent to $\|\cdot\|$.*

Proof. Since $(K^{-1}\mathbf{q},\mathbf{v})_Q$ is linear and symmetric, Lemma 2.4 implies that it is an inner product and that $(K^{-1}\mathbf{q},\mathbf{q})_Q^{1/2}$ is a norm in $\mathbf{V}_h$. Let us denote this norm by $\|\cdot\|_{Q,K^{-1}}$. It remains to show that it is bounded above by $\|\cdot\|$. Using (2.32), (2.5), the equivalence of norms on reference element $\hat{E}$, and (2.43), we have that for all $\mathbf{q}\in\mathbf{V}_h$

$$(K^{-1}\mathbf{q},\mathbf{q})_{Q,E} = (\mathcal{K}^{-1}\hat{\mathbf{q}},\hat{\mathbf{q}})_{\hat{Q},\hat{E}} \leq C\frac{h^{2-d}}{k_0}\|\hat{\mathbf{q}}\|_{\hat{E}}^2 \leq C\|\mathbf{q}\|_E^2,$$

which, combined with (2.44), implies that

$$c_0\|\mathbf{q}\| \leq \|\mathbf{q}\|_{Q,K^{-1}} \leq c_1\|\mathbf{q}\| \tag{2.45}$$

for some positive constants c_0 and c_1. ∎

Remark 2.3. The results of Lemma 2.4 and Corollary 2.5 hold if K^{-1} is replaced by any symmetric and positive definite matrix M.

We are now ready to establish the solvability of (2.41)–(2.42).

LEMMA 2.6. *The MFMFE method* (2.41)–(2.42) *has a unique solution.*

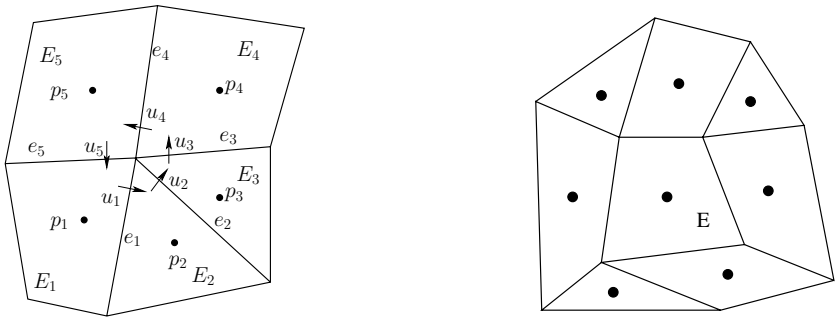
Proof. Since (2.41)–(2.42) is a square system, it is enough to show uniqueness. Let $f=0$, $g=0$, and take $\mathbf{v}=\mathbf{u}_h$ and $w=p_h$. This implies that $(K^{-1}\mathbf{u}_h,\mathbf{u}_h)_Q=0$, and therefore $\mathbf{u}_h=0$, due to (2.44). We now consider the auxiliary problem

$$\begin{aligned} -\nabla\cdot K\nabla\phi &= -p_h \quad \text{in} \quad \Omega,\\ \phi &= 0 \quad \text{on} \quad \Gamma_D,\\ -K\nabla\phi\cdot\mathbf{n} &= 0 \quad \text{on} \quad \Gamma_N. \end{aligned}$$

The choice $\mathbf{v}=\Pi K\nabla\phi\in\mathbf{V}_h$ in (2.41) gives

$$0=(p_h,\nabla\cdot\Pi K\nabla\phi)=(p_h,\nabla\cdot K\nabla\phi)=\|p_h\|^2,$$

therefore $p_h=0$. ∎



A. Five elements sharing a vertex. B. Pressure stencil.

FIG. 2.3. *Interactions of the degrees of freedom in MFMFE.*

2.7. Reduction to a cell-centered stencil. We next describe how the MFMFE method reduces to a system for the pressures at the cell centers. Let us consider any interior vertex $\mathbf{r}$ and suppose that it is shared by k elements $E_1, \ldots, E_k$; see Figure 2.3(A) for a specific example with 5 elements. We denote the edges (faces) that share the vertex by $e_1, \ldots, e_k$, the velocity basis functions on these edges (faces) that are associated with the vertex by $\mathbf{v}_1, \ldots, \mathbf{v}_k$, and the corresponding values of the normal components of $\mathbf{u}_h$ by $u_1, \ldots, u_k$. Note that for clarity the normal velocities on Figure 2.3(A) are drawn at a distance from the vertex.

Since the quadrature rule $(K^{-1}\cdot,\cdot)_Q$ localizes the basis functions interaction (see (2.34)–(2.35)), taking $\mathbf{v} = \mathbf{v}_1$ in (2.41), for example, will only lead to coupling u_1 with u_5 and u_2. Similarly, u_2 will only be coupled with u_1 and u_3, etc. Therefore, the k equations obtained from taking $\mathbf{v} = \mathbf{v}_1, \ldots, \mathbf{v}_k$ form a linear system for $u_1, \ldots, u_k$.

PROPOSITION 2.7. *The $k \times k$ local linear system described above is symmetric and positive definite.*

Proof. The system is obtained by taking $\mathbf{v} = \mathbf{v}_1, \ldots, \mathbf{v}_k$ in (2.41). On the left-hand side we have

$$(K^{-1}\mathbf{u}_h, \mathbf{v}_i)_Q = \sum_{j=1}^{k} u_j (K^{-1}\mathbf{v}_j, \mathbf{v}_i)_Q \equiv \sum_{j=1}^{k} a_{ij} u_j, \quad i = 1, \ldots, k.$$

Using Corollary 2.5 we conclude that the matrix $\bar{A} = \{a_{ij}\}$ is symmetric and positive definite. ∎

Solving the small $k \times k$ linear system allows us to express the velocities u_i in terms of the cell-centered pressures p_i, $i = 1, \ldots, k$. Substituting these expressions into the mass conservation equation (2.42) leads to a cell-centered stencil. The pressure in each element E is coupled with the pressures in the elements that share a vertex with E; see Figure 2.3(B).

For any vertex on the boundary $\partial\Omega$, the size of the local linear system equals the number of non-Neumann (interior or Dirichlet) edges/faces that share that vertex. Inverting the local system allows one to express the velocities in terms of the element pressures and the boundary data.

We use the example in Figure 2.3(A) to describe the CCFD equations obtained from the above procedure. Taking $\mathbf{v} = \mathbf{v}_1$ in (2.41), on the left-hand side we have

$$(K^{-1}\mathbf{u}_h, \mathbf{v}_1)_Q = (K^{-1}\mathbf{u}_h, \mathbf{v}_1)_{Q,E_1} + (K^{-1}\mathbf{u}_h, \mathbf{v}_1)_{Q,E_2}. \tag{2.46}$$

The first term on the right in (2.46) gives

$$\begin{aligned}(K^{-1}\mathbf{u}_h,\mathbf{v}_1)_{Q,E_1} &= (\mathcal{K}^{-1}\hat{\mathbf{u}}_h,\hat{\mathbf{v}}_1)_{\hat{Q},\hat{E}}\\ &= \frac{1}{4}(\mathcal{K}^{-1}_{11,E_1}\hat{u}_1\hat{v}_{1,1}+\mathcal{K}^{-1}_{12,E_1}\hat{u}_5\hat{v}_{1,1})\\ &= \frac{1}{4}(\mathcal{K}^{-1}_{11,E_1}|e_1|u_1+\mathcal{K}^{-1}_{12,E_1}|e_5|u_5)|e_1|,\end{aligned} \tag{2.47}$$

where we have used (2.18) for the last equality. Here $\mathcal{K}^{-1}_{ij,E_1}$ denotes a component of $\mathcal{K}^{-1}$ in E_1 and all functions are evaluated at the vertex of $\hat{E}$ corresponding to vertex $\mathbf{r}$ in the mapping F_{E_1}. Similarly,

$$(K^{-1}\mathbf{u}_h,\mathbf{v}_1)_{Q,E_2} = \frac{1}{6}(\mathcal{K}^{-1}_{11,E_2}|e_1|u_1+\mathcal{K}^{-1}_{12,E_2}|e_2|u_2)|e_1|. \tag{2.48}$$

For the right-hand side of (2.41) we write

$$\begin{aligned}(p_h,\nabla\cdot\mathbf{v}_1) &= (p_h,\nabla\cdot\mathbf{v}_1)_{E_1}+(p_h,\nabla\cdot\mathbf{v}_1)_{E_2}\\ &= \langle p_h,\mathbf{v}_1\cdot\mathbf{n}_{E_1}\rangle_{e_1}+\langle p_h,\mathbf{v}_1\cdot\mathbf{n}_{E_2}\rangle_{e_1}\\ &= \langle\hat{p}_h,\hat{\mathbf{v}}_1\cdot\hat{\mathbf{n}}_{E_1}\rangle_{\hat{e}_1}+\langle\hat{p}_h,\hat{\mathbf{v}}_1\cdot\hat{\mathbf{n}}_{E_2}\rangle_{\hat{e}_1}\\ &= \frac{1}{2}(p_1-p_2)|e_1|,\end{aligned} \tag{2.49}$$

where we have used the trapezoidal rule for the integrals on $\hat{e}_1$, which is exact since $\hat{p}_h$ is constant and $\hat{\mathbf{v}}_1\cdot\hat{\mathbf{n}}$ is linear. A combination of (2.46)–(2.49) gives the equation

$$\left(\frac{1}{2}\mathcal{K}^{-1}_{11,E_1}+\frac{1}{3}\mathcal{K}^{-1}_{11,E_2}\right)|e_1|u_1+\frac{1}{2}\mathcal{K}^{-1}_{12,E_1}|e_5|u_5+\frac{1}{3}\mathcal{K}^{-1}_{12,E_2}|e_2|u_2 = p_1-p_2.$$

The other four equations of the local system for $u_1,\dots,u_5$ are obtained similarly.

We end the section with a statement about an important property of the CCFD algebraic system.

PROPOSITION 2.8. *The CCFD system for the pressure obtained from* (2.41)–(2.42) *using the procedure described above is symmetric and positive definite.*

Proof. Let $\{\mathbf{v}_i\}$ and $\{w_j\}$ be the bases of $\mathbf{V}_h$ and W_h, respectively. The algebraic system that arises from (2.41)–(2.42) is of the form

$$\begin{pmatrix}A & B^T\\ B & 0\end{pmatrix}\begin{pmatrix}U\\ P\end{pmatrix}=\begin{pmatrix}G\\ F\end{pmatrix}, \tag{2.50}$$

where $A_{ij}=(K^{-1}\mathbf{v}_i,\mathbf{v}_j)_Q$ and $B_{ij}=-(\nabla\cdot\mathbf{v}_i,w_j)$. The matrix A is block-diagonal with symmetric and positive definite blocks, as noted in Proposition 2.7. The elimination of U leads to a system for P with a matrix

$$BA^{-1}B^T,$$

which is symmetric and positive semidefinite. In the proof of Lemma 2.6 we showed that $B^TP=0$ implies $P=0$. Therefore $BA^{-1}B^T$ is positive definite. ∎

3. Velocity error analysis. Although our method can be defined and is well posed on general quadrilaterals (see section 2), for the convergence analysis we need to impose a restriction on the element geometry. This is due to the reduced approximation properties of the MFE spaces on general quadrilaterals [8]. The restriction

is not needed for theoretical purpose only; deterioration of convergence is observed computationally as well [3].

For the remainder of the paper we will assume that the quadrilateral elements are $O(h^2)$-perturbations of parallelograms:

$$\|\mathbf{r}_{34} - \mathbf{r}_{21}\| \le Ch^2. \tag{3.1}$$

We call such elements h^2-parallelograms, following the terminology from [21]. Elements of this type are obtained by uniform refinements of a general quadrilateral grid. It is not difficult to check that in this case $||T_2| - |T_1|| \le Ch^3$, $||T_4| - |T_1|| \le Ch^3$, and

$$|DF_E|_{1,\infty,\hat{E}} \le Ch^2 \quad \text{and} \quad \left|\frac{1}{J_E} DF_E\right|_{j,\infty,\hat{E}} \le Ch^{j-1}, \quad j = 1, 2. \tag{3.2}$$

In this section we establish first-order convergence for the velocity. We start with several auxiliary results that will be used in the analysis.

In addition to the mixed projection operators defined earlier, we will also make use of the L^2-orthogonal projection onto W_h: for any $\phi \in L^2(\Omega)$, let $\mathcal{Q}_h\phi \in W_h$ satisfy

$$(\phi - \mathcal{Q}_h\phi, w) = 0 \qquad \forall w \in W_h.$$

We state several well-known approximation properties of the projection operators. On simplices and h^2-parallelograms,

$$\|\phi - \mathcal{Q}_h\phi\| \le C\|\phi\|_r h^r, \quad 0 \le r \le 1, \tag{3.3}$$

$$\|\mathbf{q} - \Pi\mathbf{q}\| \le C\|\mathbf{q}\|_r h^r, \quad 1 \le r \le 2, \tag{3.4}$$

$$\|\mathbf{q} - \Pi_0\mathbf{q}\| \le C\|\mathbf{q}\|_1 h, \tag{3.5}$$

$$\|\nabla \cdot (\mathbf{q} - \Pi\mathbf{q})\| + \|\nabla \cdot (\mathbf{q} - \Pi_0\mathbf{q})\| \le C\|\nabla \cdot \mathbf{q}\|_r h^r, \quad 0 \le r \le 1. \tag{3.6}$$

Bound (3.3) is a standard L^2-projection approximation results [18]; bounds (3.4), (3.5), and (3.6) can be found in [15, 28] for affine elements and [32, 8] for h^2-parallelograms. We note that on general quadrilaterals bounds (3.3) and (3.5) are also true, while bounds (3.4) and (3.6) are only valid for $r = 1$ and $r = 0$, respectively [8].

It was shown in [21, Lemma 5.5] that on h^2-parallelograms, for $\mathbf{u} \in H^j(E)$,

$$|\hat{\mathbf{u}}|_{j,\hat{E}} \le Ch^j \|\mathbf{u}\|_{j,E}, \quad j \ge 0. \tag{3.7}$$

We will make use of the following continuity bounds for Π and Π_0.

LEMMA 3.1. *For all elements E there exists a constant C independent of h such that*

$$\|\Pi\mathbf{q}\|_{j,E} \le C\|\mathbf{q}\|_{j,E} \quad \forall \mathbf{q} \in (H^j(E))^d, \quad j = 1, 2, \tag{3.8}$$

$$\|\Pi_0\mathbf{q}\|_{1,E} \le C\|\mathbf{q}\|_{1,E} \quad \forall \mathbf{q} \in (H^1(E))^d. \tag{3.9}$$

Proof. The proof uses the inverse inequality

$$\|\mathbf{v}\|_{j,E} \le Ch^{-1}\|\mathbf{v}\|_{j-1,E}, \quad j = 1, 2 \quad \forall\, E \in \mathcal{T}_h,\ \mathbf{v} \in \mathbf{V}_h(E), \tag{3.10}$$

which is well known for affine elements [18] and can be shown for quadrilaterals via mapping to the reference element $\hat{E}$ and using the standard inverse inequality on $\hat{E}$; see [11] for details.

Let $\bar{\mathbf{q}}$ be the $L^2(E)$-projection of $\mathbf{q}$ onto the space of constant vectors on E. Using (3.10), we have

$$\begin{aligned}|\Pi\mathbf{q}|_{1,E} &= |\Pi\mathbf{q} - \bar{\mathbf{q}}|_{1,E} \le Ch^{-1}\|\Pi\mathbf{q} - \bar{\mathbf{q}}\|_E \\ &\le Ch^{-1}(\|\Pi\mathbf{q} - \mathbf{q}\|_E + \|\mathbf{q} - \bar{\mathbf{q}}\|_E) \le C\|\mathbf{q}\|_{1,E},\end{aligned}$$

where we have used the approximation properties (3.3) and (3.4) for the last inequality.

Similarly, taking $\mathbf{q}_1$ to be the $L^2(E)$-projection of $\mathbf{q}$ onto the space of linear vectors on E, we obtain

$$\begin{aligned}|\Pi\mathbf{q}|_{2,E} &= |\Pi\mathbf{q} - \mathbf{q}_1|_{2,E} \le Ch^{-2}\|\Pi\mathbf{q} - \mathbf{q}_1\|_E \\ &\le Ch^{-2}(\|\Pi\mathbf{q} - \mathbf{q}\|_E + \|\mathbf{q} - \mathbf{q}_1\|_E) \le C\|\mathbf{q}\|_{2,E}.\end{aligned}$$

The bound $\|\Pi\mathbf{q}\|_E \le C\|\mathbf{q}\|_{1,E}$ follows from the approximation property (3.4). This completes the proof of (3.8). The proof of (3.9) is similar. ∎

The following two lemmas will also be used in the analysis.

LEMMA 3.2. *If E is an h^2-parallelogram, then there exists a constant C independent of h such that*

$$|\mathcal{K}^{-1}|_{j,\infty,\hat{E}} \le Ch^j\|K^{-1}\|_{j,\infty,E}, \quad j = 1, 2. \tag{3.11}$$

Proof. Using (3.2), we have

$$|\mathcal{K}^{-1}|_{1,\infty,\hat{E}} \le C(|\hat{K}^{-1}|_{1,\infty,\hat{E}} + h\|\hat{K}^{-1}\|_{0,\infty,\hat{E}}) \le Ch\|K^{-1}\|_{1,\infty,E},$$

where the last inequality follows from the use of the chain rule and (2.14). Similarly,

$$|\mathcal{K}^{-1}|_{2,\infty,\hat{E}} \le C(|\hat{K}^{-1}|_{2,\infty,\hat{E}} + h|\hat{K}^{-1}|_{1,\infty,\hat{E}} + h^2\|\hat{K}^{-1}\|_{0,\infty,\hat{E}}) \le Ch^2\|K^{-1}\|_{2,\infty,E},$$

where we have also used $|DF_E|_{2,\infty,\hat{E}} = 0$. ∎

Let $W^{\alpha}_{\mathcal{T}_h}$ consist of functions φ such that $\varphi|_E \in W^{\alpha}(E)$ for all $E \in \mathcal{T}_h$ and $\|\varphi\|_{\alpha,E}$ is uniformly bounded, independently of h. Let $\|\varphi\|_{\alpha} = \max_{E\in\mathcal{T}_h}\|\varphi\|_{\alpha,E}$.

LEMMA 3.3. *On h^2-parallelograms, if $K^{-1} \in W^{1,\infty}_{\mathcal{T}_h}$, then there exists a constant C independent of h such that for all $\mathbf{v} \in \mathbf{V}_h$*

$$|(K^{-1}\Pi\mathbf{u}, \mathbf{v} - \Pi_0\mathbf{v})_Q| \le Ch\|\mathbf{u}\|_1\|\mathbf{v}\|. \tag{3.12}$$

Proof. On any element E we have

$$\begin{aligned}(K^{-1}\Pi\mathbf{u}, \mathbf{v} - \Pi_0\mathbf{v})_{Q,E} &= (\mathcal{K}^{-1}\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}} \\ &= ((\mathcal{K}^{-1} - \overline{\mathcal{K}^{-1}})\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}} + (\overline{\mathcal{K}^{-1}}\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}},\end{aligned} \tag{3.13}$$

where $\overline{\mathcal{K}^{-1}}$ is the mean value of $\mathcal{K}^{-1}$ on $\hat{E}$. Using Taylor expansion and (2.45), we have for the first term on the right above

$$\begin{aligned}|((\mathcal{K}^{-1} - \overline{\mathcal{K}^{-1}})\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}}| &\le C|\mathcal{K}^{-1}|_{1,\infty,\hat{E}}\|\hat{\Pi}\hat{\mathbf{u}}\|_{\hat{E}}\|\hat{\mathbf{v}}\|_{\hat{E}} \\ &\le Ch\|K^{-1}\|_{1,\infty,E}\|\mathbf{u}\|_{1,E}\|\mathbf{v}\|_E,\end{aligned} \tag{3.14}$$

where we have used (3.11), (2.43), and (3.8) for the last inequality. Using (2.40) and letting $\overline{\hat{\Pi}\hat{\mathbf{u}}}$ be the L^2-projection of $\hat{\Pi}\hat{\mathbf{u}}$ onto the space of constant vectors on $\hat{E}$, we bound the last term in (3.13) as follows:

(3.15)
$$|(\overline{\mathcal{K}^{-1}}\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}}| = |(\overline{\mathcal{K}^{-1}}(\hat{\Pi}\hat{\mathbf{u}} - \overline{\hat{\Pi}\hat{\mathbf{u}}}), \hat{\mathbf{v}} - \hat{\Pi}_0\hat{\mathbf{v}})_{\hat{Q},\hat{E}}|$$
$$\leq C\|\mathcal{K}^{-1}\|_{0,\infty,\hat{E}}|\hat{\Pi}\hat{\mathbf{u}}|_{1,\hat{E}}\|\hat{\mathbf{v}}\|_{\hat{E}} \leq Ch\|K^{-1}\|_{0,\infty,E}\|\mathbf{u}\|_{1,E}\|\mathbf{v}\|_E,$$

where we have also used (2.32), (3.7), and (3.8). The proof is completed by combining (3.13)–(3.15). ∎

3.1. First-order convergence for the velocity. Subtracting the numerical method (2.41)–(2.42) from the variational formulation (2.6)–(2.7), we obtain the error equations

$$(K^{-1}(\Pi\mathbf{u} - \mathbf{u}_h), \mathbf{v})_Q = (Q_h p - p_h, \nabla\cdot\mathbf{v})$$
$$- (K^{-1}\mathbf{u}, \mathbf{v}) + (K^{-1}\Pi\mathbf{u}, \mathbf{v})_Q, \quad \mathbf{v} \in \mathbf{V}_h, \tag{3.16}$$
$$(\nabla\cdot(\Pi\mathbf{u} - \mathbf{u}_h), w) = 0, \quad w \in W_h. \tag{3.17}$$

The last two terms in (3.16) can be manipulated as follows:

$$\begin{aligned} -(K^{-1}\mathbf{u}, \mathbf{v}) + (K^{-1}\Pi\mathbf{u}, \mathbf{v})_Q &= -(K^{-1}\mathbf{u}, \mathbf{v} - \Pi_0\mathbf{v}) - (K^{-1}(\mathbf{u} - \Pi\mathbf{u}), \Pi_0\mathbf{v}) \\ &\quad - (K^{-1}\Pi\mathbf{u}, \Pi_0\mathbf{v}) + (K^{-1}\Pi\mathbf{u}, \Pi_0\mathbf{v})_Q + (K^{-1}\Pi\mathbf{u}, \mathbf{v} - \Pi_0\mathbf{v})_Q. \end{aligned} \tag{3.18}$$

For the first term on the right above we have

$$(K^{-1}\mathbf{u}, \mathbf{v} - \Pi_0\mathbf{v}) = 0, \tag{3.19}$$

which follows by taking $\mathbf{v} - \Pi_0\mathbf{v}$ as a test function in the variational formulation (2.6) and using (2.27). Using (3.4) and (2.28), the second term on the right in (3.18) can be bounded as

$$|(K^{-1}(\mathbf{u} - \Pi\mathbf{u}), \Pi_0\mathbf{v})| \leq Ch\|K^{-1}\|_{0,\infty}\|\mathbf{u}\|_1\|\mathbf{v}\|. \tag{3.20}$$

The third and fourth term on the right in (3.18) represent the quadrature error, which can be bounded by Lemma 3.5 as

$$|\sigma(K^{-1}\Pi\mathbf{u}, \Pi_0\mathbf{v})| \leq Ch|||K^{-1}|||_{1,\infty}\|\mathbf{u}\|_1\|\mathbf{v}\|, \tag{3.21}$$

using also (3.8) and (2.28). The last term on the right in (3.18) is bounded in Lemma 3.3.

We take $\mathbf{v} = \Pi\mathbf{u} - \mathbf{u}_h$ in the error equation (3.16) above. Note that

$$\nabla\cdot(\Pi\mathbf{u} - \mathbf{u}_h) = 0, \tag{3.22}$$

since, due to (2.19), we can choose $w = J_E\nabla\cdot(\Pi\mathbf{u} - \mathbf{u}_h) \in W_h$ on any element E in (3.17) and J_E is uniformly positive. Combining (3.18)–(3.21) with (2.44) and (3.12), we obtain

$$\|\Pi\mathbf{u} - \mathbf{u}_h\| \leq Ch|||K^{-1}|||_{1,\infty}\|\mathbf{u}\|_1. \tag{3.23}$$

The theorem below now follows from (3.23), (3.22), (3.4), and (3.6).

THEOREM 3.4. *If $K^{-1} \in W^{1,\infty}_{\mathcal{T}_h}$, then, for the velocity $\mathbf{u}_h$ of the MFMFE method (2.41)–(2.42), there exists a constant C independent of h such that*

$$\|\mathbf{u} - \mathbf{u}_h\| \le Ch\|\mathbf{u}\|_1, \tag{3.24}$$

$$\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\| \le Ch\|\nabla \cdot \mathbf{u}\|_1. \tag{3.25}$$

We now proceed with the analysis of the quadrature error.

LEMMA 3.5. *If $K^{-1} \in W^{1,\infty}_{\mathcal{T}_h}$, then there exists a constant C independent of h such that for all $\mathbf{q} \in \mathbf{V}_h$ and for all $\mathbf{v} \in \mathbf{V}_h^0$,*

$$|\sigma(K^{-1}\mathbf{q}, \mathbf{v})| \le C \sum_{E \in \mathcal{T}_h} h\|K^{-1}\|_{1,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_E. \tag{3.26}$$

Proof. We first consider the case of simplicial elements. We have on any element E

$$|\sigma_E(K^{-1}\mathbf{q}, \mathbf{v})| \le |\sigma_E((K^{-1} - \overline{K^{-1}})\mathbf{q}, \mathbf{v})| + |\sigma_E(\overline{K^{-1}}\mathbf{q}, \mathbf{v})|, \tag{3.27}$$

where $\overline{K^{-1}}$ is the mean value of K^{-1} on E. For the first term on the right we have

$$|\sigma_E((K^{-1} - \overline{K^{-1}})\mathbf{q}, \mathbf{v})| \le Ch|K^{-1}|_{1,\infty,E}\|\mathbf{q}\|_E\|\mathbf{v}\|_E, \tag{3.28}$$

where we have used Taylor expansion and (2.45). Let $\overline{\mathbf{q}}$ be the L^2-projection of $\mathbf{q}$ onto the space of constant vectors on E. For the second term on the right in (3.27), using Lemma 2.1, we have that

$$|\sigma_E(\overline{K^{-1}}\mathbf{q}, \mathbf{v})| = |\sigma_E(\overline{K^{-1}}(\mathbf{q} - \overline{\mathbf{q}}), \mathbf{v})| \le Ch\|K^{-1}\|_{0,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_E, \tag{3.29}$$

using (3.3). Combining (3.27)–(3.29), we obtain

$$|\sigma_E(K^{-1}\mathbf{q}, \mathbf{v})| \le Ch\|K^{-1}\|_{1,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_E, \tag{3.30}$$

completing the proof of (3.26) for simplicial elements.

Next, consider the quadrature error on h^2-parallelograms. We have

$$\sigma_E(K^{-1}\mathbf{q}, \mathbf{v}) = \hat{\sigma}_{\hat{E}}(\mathcal{K}^{-1}\hat{\mathbf{q}}, \hat{\mathbf{v}}) = \hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1} - \overline{\mathcal{K}^{-1}})\hat{\mathbf{q}}, \hat{\mathbf{v}}) + \hat{\sigma}_{\hat{E}}(\overline{\mathcal{K}^{-1}}\hat{\mathbf{q}}, \hat{\mathbf{v}}), \tag{3.31}$$

where $\overline{\mathcal{K}^{-1}}$ is the mean value of $\mathcal{K}^{-1}$ on $\hat{E}$. Using Taylor expansion, the first term on the right above can be bounded as

$$|\hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1} - \overline{\mathcal{K}^{-1}})\hat{\mathbf{q}}, \hat{\mathbf{v}})| \le C|\mathcal{K}^{-1}|_{1,\infty,\hat{E}}\|\hat{\mathbf{q}}\|_{\hat{E}}\|\hat{\mathbf{v}}\|_{\hat{E}} \le Ch\|K^{-1}\|_{1,\infty,E}\|\mathbf{q}\|_E\|\mathbf{v}\|_E, \tag{3.32}$$

where we used (3.11) and (2.43) for the last inequality. For the last term in (3.31) we have that $\hat{\sigma}_{\hat{E}}(\overline{\mathcal{K}^{-1}}\hat{\mathbf{q}}_0, \hat{\mathbf{v}}) = 0$ for any constant vector $\hat{\mathbf{q}}_0$, since the trapezoidal quadrature rule $(\cdot,\cdot)_{\hat{Q},\hat{E}}$ is exact for linear functions. Hence, the Bramble–Hilbert lemma [13] implies

$$|\hat{\sigma}_{\hat{E}}(\overline{\mathcal{K}^{-1}}\hat{\mathbf{q}}, \hat{\mathbf{v}})| \le C\|\mathcal{K}^{-1}\|_{0,\infty,\hat{E}}|\hat{\mathbf{q}}|_{1,\hat{E}}\|\hat{\mathbf{v}}\|_{\hat{E}}.$$

Using (3.7) and (2.32), we obtain

$$|\hat{\sigma}_{\hat{E}}(\overline{K^{-1}}\hat{\mathbf{q}},\hat{\mathbf{v}})| \le Ch\|K^{-1}\|_{0,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_E. \tag{3.33}$$

The above bound, together with (3.31)–(3.32), implies that

$$|\sigma_E(K^{-1}\mathbf{q},\mathbf{v})| \le Ch\|K^{-1}\|_{1,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_E.$$

The proof is completed by summing over all elements E. ∎

4. Error estimates for the pressure. In this section we use a standard inf-sup argument to prove optimal convergence for the pressure. We also employ a duality argument to establish superconvergence for the pressure at the element centers of mass.

4.1. First-order convergence for the pressure. We start with an optimal error bound for the pressure.

THEOREM 4.1. *If $K^{-1} \in W^{1,\infty}_{\mathcal{T}_h}$, then, for the pressure p_h of the MFMFE method* (2.41)–(2.42)*, there exists a constant C independent of h such that*

$$\|p - p_h\| \le Ch(\|\mathbf{u}\|_1 + \|p\|_1).$$

Proof. It is well known [27, 15, 32] that the RT_0 spaces $\mathbf{V}_h^0 \times W_h^0$ satisfy the inf-sup condition

$$\inf_{0\neq w\in W_h^0}\sup_{0\neq \mathbf{v}\in\mathbf{V}_h^0}\frac{(\nabla\cdot\mathbf{v},w)}{\|\mathbf{v}\|_{\mathrm{div}}\|w\|} \ge \beta, \tag{4.1}$$

where β is a positive constant independent of h. Using (4.1) and (3.16), we obtain

$$\begin{aligned}
&\|\mathcal{Q}_h p - p_h\| \\
&\quad\le \frac{1}{\beta}\sup_{0\neq\mathbf{v}\in\mathbf{V}_h^0}\frac{(\nabla\cdot\mathbf{v},\mathcal{Q}_h p - p_h)}{\|\mathbf{v}\|_{\mathrm{div}}} \\
&\quad= \frac{1}{\beta}\sup_{0\neq\mathbf{v}\in\mathbf{V}_h^0}\frac{(K^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\mathbf{v})_Q - (K^{-1}(\Pi\mathbf{u}-\mathbf{u}),\mathbf{v}) + \sigma(K^{-1}\Pi\mathbf{u},\mathbf{v})}{\|\mathbf{v}\|_{\mathrm{div}}} \\
&\quad\le \frac{C}{\beta}h|||K^{-1}|||_{1,\infty}\|\mathbf{u}\|_1,
\end{aligned}$$

where we have used the Cauchy–Schwarz inequality, (3.23), and (3.26) in the last inequality. The proof is completed by an application of the triangle inequality and (3.3). ∎

4.2. Second-order convergence for the pressure. We continue with the superconvergence estimate. We first present a bound on the quadrature error that will be used in the analysis.

LEMMA 4.2. *Let $K^{-1} \in W^{2,\infty}_{\mathcal{T}_h}$. On simplicial elements, for all $\mathbf{v},\mathbf{q} \in \mathbf{V}_h$, there exists a positive constant C independent of h such that*

$$|\sigma(K^{-1}\mathbf{q},\mathbf{v})| \le C\sum_{E\in\mathcal{T}_h} h^2\|K^{-1}\|_{2,\infty,E}\|\mathbf{q}\|_{1,E}\|\mathbf{v}\|_{1,E}. \tag{4.2}$$

On h^2-parallelograms, for all $\mathbf{q} \in \mathbf{V}_h$, $\mathbf{v} \in \mathbf{V}_h^0$, there exists a positive constant C independent of h such that

$$|\sigma(K^{-1}\mathbf{q}, \mathbf{v})| \leq C \sum_{E \in \mathcal{T}_h} h^2 \|K^{-1}\|_{2,\infty,E} \|\mathbf{q}\|_{2,E} \|\mathbf{v}\|_{1,E}. \tag{4.3}$$

Proof. We present first the proof for simplicial elements. For any element E, using Lemma 2.1, we have

$$\begin{aligned} \sigma_E(K^{-1}\mathbf{q}, \mathbf{v}) = {} & \sigma_E((K^{-1} - \overline{K^{-1}})(\mathbf{q} - \bar{\mathbf{q}}), \mathbf{v}) + \sigma_E((K^{-1} - \overline{K^{-1}})\bar{\mathbf{q}}, \mathbf{v} - \bar{\mathbf{v}}) \\ & + \sigma_E(K^{-1}\bar{\mathbf{q}}, \bar{\mathbf{v}}) + \sigma_E(\overline{K^{-1}}(\mathbf{q} - \bar{\mathbf{q}}), \mathbf{v} - \bar{\mathbf{v}}), \end{aligned} \tag{4.4}$$

where $\bar{\mathbf{q}}$ and $\bar{\mathbf{v}}$ are the $L^2(E)$-orthogonal projections of $\mathbf{q}$ and $\mathbf{v}$, respectively, onto the space of constant vectors, and $\overline{K^{-1}}$ is the mean value of K^{-1} on E. Using (2.45), the first, second, and fourth term on the right above are bounded by

$$Ch^2 \|K^{-1}\|_{1,\infty,E} \|\mathbf{q}\|_{1,E} \|\mathbf{v}\|_{1,E}. \tag{4.5}$$

For the third term on the right in (4.4) it is easy to check that the quadrature rule is exact for linear tensors. An application of the Bramble–Hilbert lemma [13] gives

$$|\sigma_E(K^{-1}\bar{\mathbf{q}}, \bar{\mathbf{v}})| \leq Ch^2 |K^{-1}\bar{\mathbf{q}}|_{2,E} \|\bar{\mathbf{v}}\|_E \leq Ch^2 |K^{-1}|_{2,\infty,E} \|\mathbf{q}\|_E \|\mathbf{v}\|_E. \tag{4.6}$$

A combination of (4.4)–(4.6) completes the proof for simplicial elements.

We proceed with the bound on the quadrature error in the case of h^2-parallelograms. We have

$$\sigma_E(K^{-1}\mathbf{q}, \mathbf{v}) = \hat{\sigma}_{\hat{E}}(\mathcal{K}^{-1}\hat{\mathbf{q}}, \hat{\mathbf{v}}) = \hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1}\hat{\mathbf{q}})_1, \hat{v}_1) + \hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1}\hat{\mathbf{q}})_2, \hat{v}_2). \tag{4.7}$$

Let us consider the first term on the right. Since the quadrature rule is exact for linear functions, the Peano kernel theorem [31, Theorem 5.2–3] implies

$$\begin{aligned} \hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1}\hat{\mathbf{q}})_1, \hat{v}_1) = {} & \int_0^1 \int_0^1 \varphi(\hat{x}) \frac{\partial^2}{\partial \hat{x}^2} ((\mathcal{K}^{-1}\hat{\mathbf{q}})_1 \hat{v}_1)(\hat{x}, 0) d\hat{x}\, d\hat{y} \\ & + \int_0^1 \int_0^1 \varphi(\hat{y}) \frac{\partial^2}{\partial \hat{y}^2} ((\mathcal{K}^{-1}\hat{\mathbf{q}})_1 \hat{v}_1)(0, \hat{y})\, d\hat{x} d\hat{y} \\ & + \int_0^1 \int_0^1 \psi(\hat{x}, \hat{y}) \frac{\partial^2}{\partial \hat{x} \partial \hat{y}} ((\mathcal{K}^{-1}\hat{\mathbf{q}})_1 \hat{v}_1)(\hat{x}, \hat{y}) d\hat{x}\, d\hat{y}, \end{aligned} \tag{4.8}$$

where $\varphi(s) = s(s-1)/2$ and $\psi(s,t) = (1-s)(1-t) - 1/4$. Therefore, using that $\hat{\mathbf{v}}$ is linear,

$$\begin{aligned} |\hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1}\hat{\mathbf{q}})_1, \hat{v}_1)| \leq {} & C((|\mathcal{K}^{-1}|_{1,\infty,\hat{E}} \|\hat{\mathbf{q}}\|_{\hat{E}} + \|\mathcal{K}^{-1}\|_{0,\infty,\hat{E}} |\hat{\mathbf{q}}|_{1,\hat{E}}) |\hat{\mathbf{v}}|_{1,\hat{E}} \\ & + (|\mathcal{K}^{-1}|_{2,\infty,\hat{E}} \|\hat{\mathbf{q}}\|_{\hat{E}} + |\mathcal{K}^{-1}|_{1,\infty,\hat{E}} |\hat{\mathbf{q}}|_{1,\hat{E}} + \|\mathcal{K}^{-1}\|_{0,\infty,\hat{E}} |\hat{\mathbf{q}}|_{2,\hat{E}}) \|\hat{\mathbf{v}}\|_{\hat{E}}). \end{aligned}$$

The term $\hat{\sigma}_{\hat{E}}((\mathcal{K}^{-1}\hat{\mathbf{q}})_2, \hat{v}_2)$ in (4.7) can be bounded similarly. Using (4.7), (2.32), (3.11), and (3.7), we obtain

$$|\sigma_E(K^{-1}\mathbf{q}, \mathbf{v})| \leq Ch^2 \|K^{-1}\|_{2,\infty,E} \|\mathbf{q}\|_{2,E} \|\mathbf{v}\|_{1,E}.$$

Summing over all elements completes the proof. ∎

We are now ready to establish superconvergence of the pressure at the cell centers.

THEOREM 4.3. *Assume that* $K \in W^{1,\infty}_{\mathcal{T}_h}$ *and* $K^{-1} \in W^{2,\infty}_{\mathcal{T}_h}$ *and the elliptic regularity* (4.11) *below holds. Then, for the pressure* p_h *of the MFMFE method* (2.41)–(2.42)*, there exists a constant* C *independent of* h *such that*

$$\|\mathcal{Q}_h p - p_h\| \le Ch^2(\|\mathbf{u}\|_1 + \|\nabla\cdot\mathbf{u}\|_1) \quad \text{on simplices} \tag{4.9}$$

and

$$\|\mathcal{Q}_h p - p_h\| \le Ch^2\|\mathbf{u}\|_2 \quad \text{on } h^2\text{-parallelograms.} \tag{4.10}$$

Proof. The proof is based on a duality argument. Let ϕ be the solution of

$$\begin{aligned} -\nabla\cdot K\nabla\phi &= -(\mathcal{Q}_h p - p_h) && \text{in } \Omega,\\ \phi &= 0 && \text{on } \Gamma_D,\\ -K\nabla\phi\cdot\mathbf{n} &= 0 && \text{on } \Gamma_N. \end{aligned}$$

We assume that this problem has H^2-elliptic regularity:

$$\|\phi\|_2 \le C\|\mathcal{Q}_h p - p_h\|_0. \tag{4.11}$$

Sufficient conditions for (4.11) can be found in [22, 26]. For example, (4.11) holds if the components of $K \in C^{0,1}(\overline{\Omega})$, $\partial\Omega$ is smooth enough, and either Γ_D or Γ_N is empty.

Let us consider first the case of simplicial elements. Here it is more convenient to rewrite the error equation (3.16) as

$$(K^{-1}(\mathbf{u}-\mathbf{u}_h),\mathbf{v}) = (Q_h p - p_h, \nabla\cdot\mathbf{v}) - \sigma(K^{-1}\mathbf{u}_h,\mathbf{v}). \tag{4.12}$$

Take $\mathbf{v} = \Pi K\nabla\phi \in \mathbf{V}_h$ in (4.12) to get

$$\begin{aligned} \|\mathcal{Q}_h p - p_h\|_0^2 &= (\mathcal{Q}_h p - p_h, \nabla\cdot\Pi K\nabla\phi)\\ &= (K^{-1}(\mathbf{u}-\mathbf{u}_h),\Pi K\nabla\phi) + \sigma(K^{-1}\mathbf{u}_h,\Pi K\nabla\phi). \end{aligned} \tag{4.13}$$

For the first term on the right above we have

$$\begin{aligned} &(K^{-1}(\mathbf{u}-\mathbf{u}_h),\Pi K\nabla\phi)\\ &\quad= (K^{-1}(\mathbf{u}-\mathbf{u}_h),\Pi K\nabla\phi - K\nabla\phi) + (\mathbf{u}-\mathbf{u}_h,\nabla\phi)\\ &\quad= (K^{-1}(\mathbf{u}-\mathbf{u}_h),\Pi K\nabla\phi - K\nabla\phi) - (\nabla\cdot(\mathbf{u}-\mathbf{u}_h),\phi-\mathcal{Q}_h\phi)\\ &\quad\le C(h\|\mathbf{u}-\mathbf{u}_h\|\|K\|_{1,\infty}\|\phi\|_2 + h\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|\|\phi\|_1)\\ &\quad\le Ch^2\|K\|_{1,\infty}(\|\mathbf{u}\|_1 + \|\nabla\cdot\mathbf{u}\|_1)\|\phi\|_2, \end{aligned} \tag{4.14}$$

where we have used (3.4) and (3.3) for the first inequality, and (3.24) and (3.25) for the second inequality.

Using (4.2), we bound the second term on the right in (4.13) as

$$\begin{aligned} &|\sigma(K^{-1}\mathbf{u}_h,\Pi K\nabla\phi)|\\ &\quad\le C\|K^{-1}\|_{2,\infty}\sum_{E\in\mathcal{T}_h} h^2\|\mathbf{u}_h\|_{1,E}\|\Pi K\nabla\phi\|_{1,E}\\ &\quad\le C\|K^{-1}\|_{2,\infty}\sum_{E\in\mathcal{T}_h} h^2(\|\mathbf{u}_h-\Pi\mathbf{u}\|_{1,E} + \|\Pi\mathbf{u}\|_{1,E})\|K\nabla\phi\|_{1,E}\\ &\quad\le C\|K^{-1}\|_{2,\infty}\sum_{E\in\mathcal{T}_h} h^2(h^{-1}\|\mathbf{u}_h-\Pi\mathbf{u}\|_E + \|\mathbf{u}\|_{1,E})\|K\|_{1,\infty,E}\|\phi\|_{2,E}\\ &\quad\le Ch^2\|K^{-1}\|_{2,\infty}\|K\|_{1,\infty}\|\mathbf{u}\|_1\|\phi\|_2, \end{aligned} \tag{4.15}$$

where we have used (3.8), the inverse inequality (3.10), and (3.23). Now (4.9) follows from (4.13)–(4.15) and (4.11).

For the analysis on h^2-parallelograms we rewrite the error equation (3.16) in the form

$$(K^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\mathbf{v})_Q=(Q_hp-p_h,\nabla\cdot\mathbf{v})+(K^{-1}(\Pi\mathbf{u}-\mathbf{u}),\mathbf{v})-\sigma(K^{-1}\Pi\mathbf{u},\mathbf{v}). \tag{4.16}$$

Take $\mathbf{v}=\Pi_0K\nabla\phi\in\mathbf{V}_h$ in (4.16) to get

$$\begin{aligned}\|\mathcal{Q}_hp-p_h\|_0^2&=(\mathcal{Q}_hp-p_h,\nabla\cdot\Pi_0K\nabla\phi)\\&=(K^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K\nabla\phi)_Q-(K^{-1}(\Pi\mathbf{u}-\mathbf{u}),\Pi_0K\nabla\phi)\\&\quad+\sigma(K^{-1}\Pi\mathbf{u},\Pi_0K\nabla\phi).\end{aligned} \tag{4.17}$$

Using (3.4) and (3.9), the second term on the right above can be bounded as

$$|(K^{-1}(\Pi\mathbf{u}-\mathbf{u}),\Pi_0K\nabla\phi)|\le Ch^2\|K^{-1}\|_{0,\infty}|\!|\!|K|\!|\!|_{1,\infty}\|\mathbf{u}\|_2\|\phi\|_2. \tag{4.18}$$

For the last term on the right in (4.17), bounds (4.3), (3.8), and (3.9) imply that

$$\sigma(K^{-1}\Pi\mathbf{u},\Pi_0K\nabla\phi)\le Ch^2|\!|\!|K^{-1}|\!|\!|_{2,\infty}|\!|\!|K|\!|\!|_{1,\infty}\|\mathbf{u}\|_2\|\phi\|_2. \tag{4.19}$$

The first term on the right in (4.17) can be manipulated as follows:

$$\begin{aligned}&(K^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K\nabla\phi)_{Q,E}\\&\quad=((K^{-1}-K_0^{-1})(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K\nabla\phi)_{Q,E}+(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0(K-K_0)\nabla\phi)_{Q,E}\\&\qquad+(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K_0(\nabla\phi-\nabla\phi_1))_{Q,E}+(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K_0\nabla\phi_1)_{Q,E},\end{aligned} \tag{4.20}$$

where K_0 is the value of K at the center of E and ϕ_1 is a linear approximation to ϕ such that (see [13])

$$\|\phi-\phi_1\|_E\le Ch^2\|\phi\|_{2,E},\quad \|\phi-\phi_1\|_{1,E}\le Ch\|\phi\|_{2,E}. \tag{4.21}$$

Using (3.9), the first term on the right in (4.20) can be bounded as

$$|((K^{-1}-K_0^{-1})(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K\nabla\phi)_{Q,E}|\le Ch\|K^{-1}\|_{1,\infty,E}\|K\|_{1,\infty,E}\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi\|_{2,E}. \tag{4.22}$$

For the second and third terms on the right in (4.20) we use that for any $\psi\in(H^1(E))^2$

$$\|\Pi_0\psi\|_E\le\|\Pi_0\psi-\psi\|_E+\|\psi\|_E\le C(h\|\psi\|_{1,E}+\|\psi\|_E)$$

to obtain

$$|(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0(K-K_0)\nabla\phi)_{Q,E}|\le Ch\|K^{-1}\|_{0,\infty,E}\|K\|_{1,\infty,E}\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi\|_{2,E} \tag{4.23}$$

and

$$\begin{aligned}&|(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K_0(\nabla\phi-\nabla\phi_1))_{Q,E}|\\&\quad\le Ch\|K^{-1}\|_{0,\infty,E}\|K\|_{0,\infty,E}\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi\|_{2,E},\end{aligned} \tag{4.24}$$

having also used (4.21) in the last inequality. For the last term in (4.20) we have

(4.25)
$$(K_0^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K_0\nabla\phi_1)_{Q,E}=(\Pi\mathbf{u}-\mathbf{u}_h,\nabla\phi_1)_{Q,E}=(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}\hat{\phi}_1)_{\hat{Q},\hat{E}},$$

using $\nabla\phi_1=(DF^{-1})^T\hat{\nabla}\hat{\phi}_1$ in the second equality. Note that $\hat{\phi}(\hat{x},\hat{y})$ is a bilinear function. Let $\tilde{\phi}_1$ be the linear part of $\hat{\phi}_1$. We have

$$(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}\hat{\phi}_1)_{\hat{Q},\hat{E}}=(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}(\hat{\phi}_1-\tilde{\phi}_1))_{\hat{Q},\hat{E}}+(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}\tilde{\phi}_1)_{\hat{Q},\hat{E}}. \tag{4.26}$$

Since (see (2.8))

$$\hat{\nabla}(\hat{\phi}_1-\tilde{\phi}_1)=[(\mathbf{r}_{34}-\mathbf{r}_{21})\cdot\nabla\phi_1]\begin{pmatrix}\hat{y}\\ \hat{x}\end{pmatrix},$$

(3.1) implies

(4.27)
$$\begin{aligned}|(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}(\hat{\phi}_1-\tilde{\phi}_1))_{\hat{Q},\hat{E}}|&\le Ch^2\|\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h\|_{\hat{E}}\|\nabla\phi_1\|_{\hat{E}}\\ &\le Ch\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\nabla\phi_1\|_E\le Ch\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi\|_{2,E}.\end{aligned}$$

It remains to bound the last term in (4.26). Using (2.40) and the fact that the trapezoidal rule is exact for linear functions, we have

$$\begin{aligned}(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h,\hat{\nabla}\tilde{\phi}_1)_{\hat{Q},\hat{E}}&=(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}\tilde{\phi}_1)_{\hat{Q},\hat{E}}=(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}\tilde{\phi}_1)_{\hat{E}}\\ &=(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}(\tilde{\phi}_1-\hat{\phi}_1))_{\hat{E}}+(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}\hat{\phi}_1)_{\hat{E}}.\end{aligned} \tag{4.28}$$

The first term on the right in (4.28) is bounded similarly to (4.27):

$$|(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}(\tilde{\phi}_1-\hat{\phi}_1))_{\hat{E}}|\le Ch\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi\|_{2,E}. \tag{4.29}$$

For the last term in (4.28) we have

$$(\hat{\Pi}_0(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}}_h),\hat{\nabla}\hat{\phi}_1)_{\hat{E}}=(\Pi_0(\Pi\mathbf{u}-\mathbf{u}_h),\nabla\phi_1)_E. \tag{4.30}$$

Combining (4.20)–(4.30) and summing over all elements, we obtain

$$(K^{-1}(\Pi\mathbf{u}-\mathbf{u}_h),\Pi_0K\nabla\phi)_Q=R+\sum_{E\in\mathcal{T}_h}(\Pi_0(\Pi\mathbf{u}-\mathbf{u}_h),\nabla\phi_1)_E, \tag{4.31}$$

where

$$|R|\le Ch^2|||K^{-1}|||_{1,\infty}|||K|||_{1,\infty}\|\mathbf{u}\|_1\|\phi\|_2, \tag{4.32}$$

having also used (3.23). For the last term in (4.31), using the regularity of ϕ, (3.22), (2.27), and that $(\Pi\mathbf{u}-\mathbf{u}_h)\cdot\mathbf{n}=0$ on Γ_N and $\phi=0$ on Γ_D, we obtain

$$\begin{aligned}\left|\sum_{E\in\mathcal{T}_h}(\Pi_0(\Pi\mathbf{u}-\mathbf{u}_h),\nabla\phi_1)_E\right|&=\left|\sum_{E\in\mathcal{T}_h}(\Pi_0(\Pi\mathbf{u}-\mathbf{u}_h),\nabla(\phi_1-\phi))_E\right|\\ &\le C\sum_{E\in\mathcal{T}_h}\|\Pi\mathbf{u}-\mathbf{u}_h\|_E\|\phi_1-\phi\|_{1,E}\\ &\le Ch^2|||K^{-1}|||_{1,\infty}\|\mathbf{u}\|_1\|\phi\|_2,\end{aligned} \tag{4.33}$$

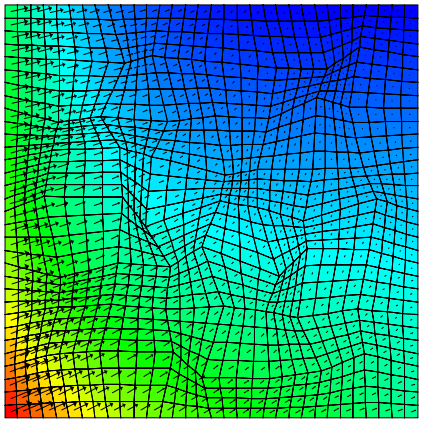


FIG. 5.1. *Computed solution on the second level of refinement in Example* 1.

where we have used (3.23) and (4.21). The proof of (4.10) is completed by combining (4.17)–(4.19) and (4.31)–(4.33), and using (4.11). ∎

Remark 4.1. Since $\mathcal{Q}_h p$ is $O(h^2)$-close to p at the center of mass of each element, the above theorem implies that

$$|||p - p_h||| \le Ch^2,$$

where $|||\cdot||| = \left(\sum_E |E|(p(m_E) - p_h)^2\right)^{1/2}$ and m_E is the center of mass of E.

5. Numerical experiments. In this section we present several numerical results on quadrilateral grids that confirm the theoretical results from the previous sections.

In the first example we test the method on a sequence of meshes obtained by a uniform refinement of an initial rough quadrilateral mesh. The boundary conditions are of Dirichlet type. The tensor coefficient and the true solution are

$$K = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}, \quad p(x,y) = (1-x)^4 + (1-y)^3(1-x) + \sin(1-y)\cos(1-x).$$

The initial 8×8 mesh is generated from a square mesh by randomly perturbing the location of each vertex within a disk centered at the vertex with a radius $h\sqrt{2}/3$. Due to (2.31), the nonsmoothness of the grid translates into a discontinuous computational permeability $\mathcal{K}$. The computed solution on the second level of refinement is shown in Figure 5.1. The colors represent the pressure values and the arrows represent the velocity vectors. The numerical errors and asymptotic convergence rates are obtained on a sequence of six mesh refinements and are reported in Table 5.1. Here, for scalar functions $|||w|||$ is the discrete L^2-norm defined in Remark 4.1 and for vectors $|||\mathbf{v}|||$ denotes a discrete vector L^2-norm that involves only the normal vector components at the midpoints of the edges. We note that the obtained convergence rates of $O(h^2)$ for $|||p-p_h|||$ and $O(h)$ for $\|\mathbf{u}-\mathbf{u}_h\|$ confirm the theoretical results. The $O(h^2)$ accuracy for $|||\mathbf{u} - \mathbf{u}_h|||$ and $|||\nabla \cdot (\mathbf{u} - \mathbf{u}_h)|||$ indicates superconvergence for the normal velocities at the midpoints of the edges and for the divergence at the cell-centers.

In the second example we consider an irregularly shaped domain consisting of two subdomains; see Figure 5.2. The grid is nonsmooth across the interface leading

TABLE 5.1
Discretization errors and convergence rates for Example 1.

$1/h$	$\|\|\|p-p_h\|\|\|$	$\|\|\mathbf{u}-\mathbf{u}_h\|\|$	$\|\|\|\mathbf{u}-\mathbf{u}_h\|\|\|$	$\|\|\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|\|\|$
8	0.123E-1	0.882E-1	0.281E-1	0.112E-1
16	0.372E-2	0.542E-1	0.129E-1	0.287E-2
32	0.103E-2	0.292E-1	0.411E-2	0.722E-3
64	0.270E-3	0.151E-1	0.114E-2	0.181E-3
128	0.692E-4	0.772E-2	0.307E-3	0.455E-4
256	0.175E-4	0.390E-2	0.817E-4	0.127E-4
Rate	1.98	0.99	1.91	1.84

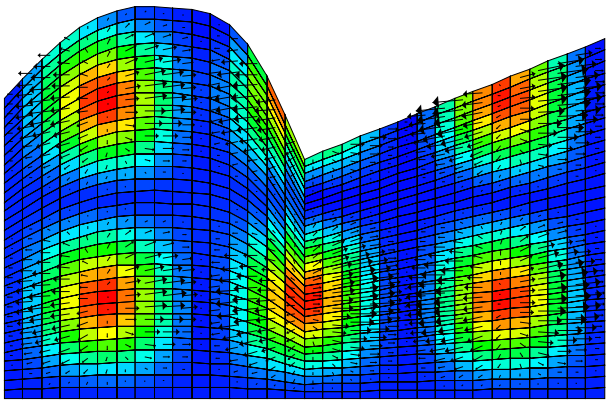


FIG. 5.2. *Computed solution on the second level of refinement in Example* 2.

to a discontinuous computational permeability $\mathcal{K}$. The permeability tensor and true solution are

$$K = \begin{pmatrix} 4+(x+2)^2+y^2 & 1+\sin(xy) \\ 1+\sin(xy) & 2 \end{pmatrix}, \quad p(x,y) = (\sin(3\pi x))^2(\sin(3\pi y))^2.$$

The boundary conditions are of Neumann type. The computed solution on the second refinement level is shown in Figure 5.2. The numerical errors and asymptotic convergence rates are presented in Table 5.2. As in the previous example, the numerical convergence rates confirm the theory.

6. Conclusions. We have presented a BDM_1-based MFE method with quadrature that reduces to CCFD for the pressure on simplicial and quadrilateral grids. The resulting algebraic system is symmetric and positive definite. The method is closely related to the MPFA method and it performs well on irregular grids and rough coefficients. The analysis is based on combining MFE techniques with quadrature error estimates. First order convergence is obtained for the pressure and the velocity in their natural norms. Second order convergence is obtained for the pressure and the element centers of mass. Computational results also indicate superconvergence for the velocity at the midpoints of the edges on h^2-parallelogram grids. We have also developed and analyzed the method on hexahedral elements that are $O(h^2)$-perturbations of parallelepipeds. These results will be presented in a forthcoming paper.

Remark 6.1. We recently learned of the concurrent and related work of Klausen and Winther [25]. They formulate the MPFA method from [1] as a MFE method

TABLE 5.2
Discretization errors and convergence rates for Example 2.

$1/h$	$\|\|\|p - p_h\|\|\|$	$\|\mathbf{u} - \mathbf{u}_h\|$	$\|\|\|\mathbf{u} - \mathbf{u}_h\|\|\|$	$\|\|\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|\|\|$
8	0.177E+2	0.492E0	0.512E0	0.764E-2
16	0.151E0	0.179E0	0.138E0	0.647E-4
32	0.653E-1	0.919E-1	0.513E-1	0.279E-4
64	0.185E-1	0.453E-1	0.132E-1	0.790E-5
128	0.460E-2	0.226E-1	0.334E-2	0.196E-5
256	0.116E-4	0.113E-1	0.838E-3	0.494E-6
Rate	1.99	0.99	1.99	1.99

using an enhanced Raviart–Thomas space and obtain convergence results on h^2-parallelogram grids.

REFERENCES

[1] I. AAVATSMARK, *An introduction to multipoint flux approximations for quadrilateral grids*, Comput. Geosci., 6 (2002), pp. 405–432.

[2] I. AAVATSMARK, T. BARKVE, Ø. BØE, AND T. MANNSETH, *Discretization on unstructured grids for inhomogeneous, anisotropic media.* I. *Derivation of the methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1700–1716.

[3] I. AAVATSMARK, G. EIGESTAD, R. KLAUSEN, M. F. WHEELER, AND I. YOTOV, *Convergence of a Symmetric MPFA Method on Quadrilateral Grids*, Technical report TR-MATH 05-14, Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, 2005.

[4] T. ARBOGAST, *Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase Darcy flow*, Comput. Geosci., 6 (2002), pp. 453–481.

[5] T. ARBOGAST, L. C. COWSAR, M. F. WHEELER, AND I. YOTOV, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[6] T. ARBOGAST, C. N. DAWSON, P. T. KEENAN, M. F. WHEELER, AND I. YOTOV, *Enhanced cell-centered finite differences for elliptic equations on general geometry*, SIAM J. Sci. Comput., 19 (1998), pp. 404–425.

[7] T. ARBOGAST, M. F. WHEELER, AND I. YOTOV, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[8] D. N. ARNOLD, D. BOFFI, AND R. S. FALX, *Quadrilateral H(div) finite elements*, SIAM J. Numer. Anal., 42 (2005), pp. 2429–2451.

[9] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, post-processing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[10] M. BERNDT, K. LIPNIKOV, J. D. MOULTON, AND M. SHASHKOV, *Convergence of mimetic finite difference discretizations of the diffusion equation*, East-West J. Numer. Math., 9 (2001), pp. 265–284.

[11] M. BERNDT, K. LIPNIKOV, M. SHASHKOV, M. F. WHEELER, AND I. YOTOV, *A mortar mimetic finite difference method on non-matching grids*, Numer. Math., 102 (2005), pp. 203–230.

[12] M. BERNDT, K. LIPNIKOV, M. SHASHKOV, M. F. WHEELER, AND I. YOTOV, *Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals*, SIAM J. Numer. Anal., 43 (2005), pp. 1728–1749.

[13] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 2002.

[14] F. BREZZI, J. DOUGLAS, JR., AND L. D. MARINI, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[15] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, Berlin, 1991.

[16] Z. CAI, J. E. JONES, S. F. MCCORMICK, AND T. F. RUSSELL, *Control-volume mixed finite element methods*, Comput. Geosci., 1 (1997), pp. 289–315 (1998).

[17] S.-H. CHOU, D. Y. KWAK, AND K. Y. KIM, *A general framework for constructing and analyzing mixed finite volume methods on quadrilateral grids: The overlapping covolume case*, SIAM J. Numer. Anal., 39 (2001), pp. 1170–1196.

[18] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.
[19] M. G. Edwards, *Unstructured, control-volume distributed, full-tensor finite-volume schemes with flow based grids*, Comput. Geosci., 6 (2002), pp. 433–452.
[20] M. G. Edwards and C. F. Rogers, *Finite volume discretization with imposed flux continuity for the general tensor pressure equation*, Comput. Geosci., 2 (1998), pp. 259–290 (1999).
[21] R. E. Ewing, M. Liu, and J. Wang, *Superconvergence of mixed finite element approximations over quadrilaterals*, SIAM J. Numer. Anal., 36 (1999), pp. 772–787.
[22] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.
[23] J. M. Hyman, M. Shashkov, and S. Steinberg, *The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials*, J. Comput. Phys., 132 (1997), pp. 130–148.
[24] R. A. Klausen and T. F. Russell, *Relationships among some locally conservative discretization methods which handle discontinuous coefficients*, Comput. Geosci., 8 (2004), pp. 341–377.
[25] R. A. Klausen and R. Winther, *Convergence of multi point flux approximations on quadrilateral grids*, Numer. Methods Partial Differential Equations, to appear.
[26] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.
[27] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977, pp. 292–315.
[28] J. E. Roberts and J. M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, P. Ciarlet and J. Lions, eds., North–Holland, Amsterdam, 1991, pp. 523–639.
[29] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, Frontiers Appl. Math. 1, R. E. Ewing, ed., SIAM, Philadelphia, 1984, pp. 35–106.
[30] T. F. Russell, M. F. Wheeler, and I. Yotov, *Superconvergence for control volume mixed finite element methods on rectangular grids*, SIAM J. Numer. Anal., to appear.
[31] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[32] J. Wang and T. P. Mathew, *Mixed finite element method over quadrilaterals*, in Conference on Advances in Numerical Methods and Applications, I. T. Dimov, B. Sendov, and P. Vassilevski, eds., World Scientific, River Edge, NJ, 1994, pp. 203–214.
[33] A. Weiser and M. F. Wheeler, *On convergence of block-centered finite differences for elliptic problems*, SIAM J. Numer. Anal., 25 (1988), pp. 351–375.
[34] M. F. Wheeler and I. Yotov, *A posteriori error estimates for the mortar mixed finite element method*, SIAM J. Numer. Anal., 43 (2005), pp. 1021–1042.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2107–2130

STABILIZED FEM-BEM COUPLING FOR HELMHOLTZ TRANSMISSION PROBLEMS*

R. HIPTMAIR† AND P. MEURY†

Abstract. We consider time-harmonic acoustic scattering at a nonsmooth penetrable object and coupled boundary element finite element schemes for its numerical simulation. Straightforward coupling approaches are haunted by instabilities at wave numbers related to interior resonances, the so-called spurious resonances. A remedy is offered by adopting the idea underlying the widely used combined field integral equations. We apply it in the form of modified trace operators. These will also feature regularizing operators to offset the lack of compactness of the double-layer potential integral operators on nonsmooth surfaces. Calderón projectors can be defined based on the modified trace operators. Thus, Costabel's approach to the symmetric coupling of domain variational formulations and boundary integral equations carries over. The modified traces guarantee uniqueness of solutions of the coupled problem, whereas regularization ensures coercivity. From this we immediately concludeasymptotic quasi-optimality of a combined finite element and boundary element Galerkin discretization for all frequencies.

Key words. acoustic scattering, boundary integral equations, combined field integral equations, Galerkin discretization, finite elements, boundary elements

AMS subject classifications. 65N38, 65N12, 65R20, 65N30

DOI. 10.1137/050639958

1. Introduction.

Let $\Omega^- \subset \mathbb{R}^3$ denote volume occupied by an inhomogeneous bounded object.[1] Plane time harmonic sound waves described by a pressure amplitude U^{i} propagate in the exterior homogeneous air region $\Omega^+ := \mathbb{R}^3 \setminus \bar{\Omega}^-$, hit the object, and get scattered.

As explained in [14, sect. 2.1], a suitably scaled pressure amplitude U of the resulting sound field will satisfy the homogeneous Helmholtz equation

$$-\Delta U - \kappa^2 n(\mathbf{x}) U = 0 \quad \text{in } \Omega^- \cup \Omega^+, \tag{1}$$

plus suitable radiation boundary conditions at ∞. The refractive index n belongs to $L^\infty\left(\mathbb{R}^3\right)$. It is allowed to vary spatially inside Ω^-, but is equal to 1 in Ω^+. Furthermore, we assume the *wave number* κ to be positive and real.

The numerical simulation of this acoustic scattering problem is faced with the unbounded domain Ω^+. Many different strategies have been devised to cope with this challenge: one could truncate Ω^+ and use standard finite elements in conjunction with absorbing boundary conditions [22]. An alternative is provided by infinite elements in Ω^+ [5, 3] or the method of fundamental solutions [20].

However, in this article we will restrict ourselves to another possibility, namely boundary integral equation methods, which reduce the problem in Ω^+ to equations on the bounded surface $\Gamma := \partial\Omega^-$. Boundary integral equations come in different varieties, among them direct and indirect methods [21, Ch. 8].

Useful integral equations remain elusive for boundary value problems with nonconstant coefficients. This is the case inside Ω^- and, therefore, we are forced to use a

*Received by the editors September 9, 2005; accepted for publication (in revised form) April 10, 2006; published electronically November 3, 2006.
http://www.siam.org/journals/sinum/44-5/63995.html
†SAM, ETH Zurich, Ch-8092 Zürich, Switzerland (hiptmair@sam.math.ethz.ch, meury@sam.math.ethz.ch). The work of the second author was supported by Schweizer Nationalfonds.
[1]We assume that Ω^- is a curvilinear Lipschitz-polyhedron in the sense of [17, sect. 1].

classical spatial discretization like the finite element method to discretize (1) in Ω^-. This entails linking the weak variational formulation of (1) with boundary integral equations on Γ.

In short, coupled problems are derived by expressing the Dirichlet-to-Neumann map of the exterior problem by means of boundary integral operators. This can be done in many ways. Yet, in many cases, in particular with so-called indirect formulations, the resulting operator lacks structural properties of the Dirichlet-to-Neumann map. This is blatantly obvious in the case of second order elliptic problems [25]. If structure is not preserved, theoretical analysis becomes much more difficult, and the linear systems of equations obtained through Galerkin boundary element discretization are adversely affected.

For second order elliptic problems, Costabel [15] discovered that the so-called direct boundary integral equations provide a remedy. The key concept is that of the Calderón projector acting on the Cauchy data of the problem. For details and theoretical examinations we refer to [12, sect. 4.5] and [18]. In short, the Calderón projector supplies two sets of boundary integral equations. Judiciously combining them yields a version of the Dirichlet-to-Neumann map that perfectly lends itself to a Galerkin discretization. The realization of Costabel's idea is called the "symmetric coupling approach" to marrying finite elements and boundary elements. It has been applied to a wide range of transmission problems; see, among many others, [11, 23, 27, 24].

Unfortunately, for the acoustic scattering problem the direct symmetric coupling approach invariably leads to equations vulnerable to spurious resonances [19, 31]: if κ^2 agrees with a Dirichlet or Neumann eigenvalue (resonant frequency) of the Laplacian in Ω^-, then the integral equations may fail to possess a unique solution, though the overall scattering problem remains well posed.

One way to deal with spurious resonances is the use of integral operators with modified kernels [35, 26]. Here we will restrict our attention to another remedy, namely the widely used combined field integral equations (CFIE). They owe their name to the typical complex linear combination of different boundary integral operators on the left-hand side of the final boundary integral equation. In the case of indirect schemes this trick has been discovered independently by Brakhage and Werner [6], Leis [28], and Panich [30] in 1965. In 1971 Burton and Miller used the same idea to obtain direct boundary integral equations without spurious resonances [10]. Meanwhile, CFIEs have become the foundation for numerous numerical methods in direct and inverse acoustic and electromagnetic scattering [14, Ch. 3 & 6].

We aim to pursue symmetric coupling based on CFIE. To do so we first have to identify related Calderón projectors. Second, we have to overcome a potential lack of coercivity of the coupled system due to the fact that the double-layer integral operators fail to be compact on nonsmooth surfaces. Both problems are tackled by introducing modified trace operators. These are motivated by the regularization approach to CFIE developed in [7, 8, 9] based on ideas by Panich [30]. We remark that introducing additional operators into boundary integral equations in order to improve their properties has also been successful in the case of high frequency scattering [1, 2].

Throughout this paper, we adopt a theoretical focus stressing complete and rigorous mathematical analysis of the resulting variational formulations. Assessing the practical relevance of the new approach is difficult, because gains from unconditional stability have to be weighed against increased computational effort.

2. The transmission problem. We depart from a formulation of the scattering problem as a transmission problem. To do so we have to rely on the following continuous and surjective trace mappings [16, Lemma 3.2]:

$$\begin{aligned} \text{Dirichlet trace} \quad & \gamma_0^{\pm} : H^1_{\mathrm{loc}}\left(\Omega^{\pm}\right) \to H^{\frac{1}{2}}\left(\Gamma\right), \\ \text{Neumann trace} \quad & \gamma_1^{\pm} : H_{\mathrm{loc}}\left(\Delta, \Omega^{\pm}\right) \to H^{-\frac{1}{2}}\left(\Gamma\right). \end{aligned}$$

We refer to [29, Ch. 3] for the definitions of function spaces $H^1_{\mathrm{loc}}\left(\Omega^{\pm}\right)$, $H_{\mathrm{loc}}\left(\Delta, \Omega^{\pm}\right)$, $H^{\frac{1}{2}}\left(\Gamma\right)$, and $H^{-\frac{1}{2}}\left(\Gamma\right)$. The trace operators generalize the following pointwise restrictions of smooth functions $V \in C^{\infty}\left(\overline{\Omega^{\pm}}\right)$:

$$(\gamma_0^{\pm} V)(\mathbf{x}) := V(\mathbf{x}) \quad \text{and} \quad (\gamma_1^{\pm} V)(\mathbf{x}) := \mathbf{grad}\, V(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}), \quad \mathbf{x} \in \Gamma.$$

Then the mathematical model for the acoustic scattering problem boils down to the following transmission problem for the Helmholtz equation; see [32, sect. 2.9]:

$$\begin{aligned} -\Delta U - \kappa^2 n(\mathbf{x}) U = f(\mathbf{x}) \quad \text{in } \Omega^-, \qquad & -\Delta U^{\mathrm{s}} - \kappa^2 U^{\mathrm{s}} = 0 \quad \text{in } \Omega^+, \\ \gamma_0^+ U^{\mathrm{s}} - \gamma_0^- U = g_0 \quad \text{on } \Gamma, \qquad & \gamma_1^+ U^{\mathrm{s}} - \gamma_1^- U = g_1 \quad \text{on } \Gamma, \\ \frac{\partial U^{\mathrm{s}}}{\partial r} - i\kappa U^{\mathrm{s}} = o(r^{-1}) \quad & \text{uniformly for } r := |\mathbf{x}| \to \infty, \end{aligned} \tag{2}$$

with the refractive index $n \in L^{\infty}\left(\Omega^-\right)$, the source term $f \in H^{-1}\left(\Omega^-\right)$, and the wave number $\kappa > 0$. In the case of excitation by an incident field U^{i} the generic jump data $g_0 \in H^{\frac{1}{2}}\left(\Gamma\right)$ and $g_1 \in H^{-\frac{1}{2}}\left(\Gamma\right)$ evaluate to the Dirichlet and Neumann data of U^{i} on the boundary Γ:

$$g_0 := -\gamma_0^+ U^{\mathrm{i}}, \quad g_1 := -\gamma_1^+ U^{\mathrm{i}}.$$

It is known that the transmission problem (2) has a unique solution $u \in H_{\mathrm{loc}}\left(\Delta, \mathbb{R}^3\right)$ [32, sect. 2.10].

Remark 2.1. Please note that inside Ω^- the field U in (2) refers to the total field, whereas in Ω^+ we write U^{s} for the scattered field. There the total field can be recovered through $U = U^{\mathrm{s}} + U^{\mathrm{i}}$.

3. Potentials and boundary integral operators. In this section we define relevant boundary integral operators and review some of their properties. Only sketches of proofs will be given and the reader is referred to [32, 29, 16] for details. To begin with, let us fix some notations and notions: jumps of traces across Γ will be designated by

$$[\gamma_0 V]_{\Gamma} := \gamma_0^+ V - \gamma_0^- V, \quad [\gamma_1 V]_{\Gamma} := \gamma_1^+ V - \gamma_1^- V,$$

and averages across Γ will be denoted by

$$\{\gamma_0 V\}_{\Gamma} := \tfrac{1}{2}\left(\gamma_0^+ V + \gamma_0^- V\right), \quad \{\gamma_1 V\}_{\Gamma} := \tfrac{1}{2}\left(\gamma_1^+ V + \gamma_1^- V\right).$$

For a fixed wave number $\kappa > 0$ a distribution U on $\mathbb{R}^3$ is called a *radiating Helmholtz solution* if

$$\Delta U + \kappa^2 U = 0 \quad \text{in } \Omega^- \cup \Omega^+, \qquad \lim_{r \to \infty} r \left(\frac{\partial U}{\partial r} - i\kappa U \right) = 0,$$

where the limit is assumed to hold uniformly in all directions. Based on the Helmholtz kernel

$$G_\kappa(z) := \frac{1}{4\pi}\frac{\exp(i\kappa z)}{z} \tag{3}$$

we can state the transmission formula for radiating Helmholtz solutions U [32, Thm. 3.1.6] as

$$U = -\Psi^\kappa_{\mathrm{SL}}([\gamma_1 U]_\Gamma) + \Psi^\kappa_{\mathrm{DL}}([\gamma_0 U]_\Gamma), \tag{4}$$

with the potentials

$$\text{single-layer potential}\quad \Psi^\kappa_{\mathrm{SL}}(\vartheta)(\mathbf{x}) := \int_\Gamma G_\kappa(|\mathbf{x}-\mathbf{y}|)\,\vartheta(\mathbf{y})\,\mathrm{d}S(\mathbf{y}),$$

$$\text{double-layer potential}\quad \Psi^\kappa_{\mathrm{DL}}(v)(\mathbf{x}) := \int_\Gamma \frac{\partial G_\kappa(|\mathbf{x}-\mathbf{y}|)}{\partial \mathbf{n}(\mathbf{y})} v(\mathbf{y})\,\mathrm{d}S(\mathbf{y}).$$

The potentials provide radiating Helmholtz solutions and continuous mappings [32, Thm. 3.1.16]

$$\Psi^\kappa_{\mathrm{SL}} : H^{-\frac{1}{2}}(\Gamma) \to H^1_{\mathrm{loc}}(\mathbb{R}^3) \cap H_{\mathrm{loc}}(\Delta, \Omega^- \cup \Omega^+),$$
$$\Psi^\kappa_{\mathrm{DL}} : H^{\frac{1}{2}}(\Gamma) \to H_{\mathrm{loc}}(\Delta, \Omega^- \cup \Omega^+).$$

Applying the trace mappings yields the following four continuous boundary integral operators:

$$\begin{aligned}
\mathsf{V}_\kappa &: H^{s-\frac{1}{2}}(\Gamma) \to H^{s+\frac{1}{2}}(\Gamma), & \mathsf{V}_\kappa &:= \{\gamma_0 \Psi^\kappa_{\mathrm{SL}}\}_\Gamma,\\
\mathsf{K}_\kappa &: H^{s+\frac{1}{2}}(\Gamma) \to H^{s+\frac{1}{2}}(\Gamma), & \mathsf{K}_\kappa &:= \{\gamma_0 \Psi^\kappa_{\mathrm{DL}}\}_\Gamma,\\
\mathsf{K}'_\kappa &: H^{s-\frac{1}{2}}(\Gamma) \to H^{s-\frac{1}{2}}(\Gamma), & \mathsf{K}'_\kappa &:= \{\gamma_1 \Psi^\kappa_{\mathrm{SL}}\}_\Gamma,\\
\mathsf{W}_\kappa &: H^{s+\frac{1}{2}}(\Gamma) \to H^{s-\frac{1}{2}}(\Gamma), & \mathsf{W}_\kappa &:= -\{\gamma_1 \Psi^\kappa_{\mathrm{DL}}\}_\Gamma,
\end{aligned}$$

for a scale of Sobolev spaces with $|s| < \frac{1}{2}$; see [16, Thm. 1]. From the *jump relations* [32, Thm. 3.3.1]

$$\begin{aligned}
[\gamma_0 \Psi^\kappa_{\mathrm{SL}}(\vartheta)]_\Gamma &= 0, & [\gamma_1 \Psi^\kappa_{\mathrm{SL}}(\vartheta)]_\Gamma &= -\vartheta & &\forall \vartheta \in H^{-\frac{1}{2}}(\Gamma),\\
[\gamma_0 \Psi^\kappa_{\mathrm{DL}}(\varphi)]_\Gamma &= \varphi, & [\gamma_1 \Psi^\kappa_{\mathrm{DL}}(\varphi)]_\Gamma &= 0 & &\forall \varphi \in H^{\frac{1}{2}}(\Gamma),
\end{aligned} \tag{5}$$

we can directly deduce the following four identities:

$$\begin{aligned}
\gamma_0^\pm \Psi^\kappa_{\mathrm{SL}} &= \mathsf{V}_\kappa, & \gamma_1^\pm \Psi^\kappa_{\mathrm{SL}} &= \mathsf{K}_\kappa \pm \tfrac{1}{2}\mathsf{Id},\\
\gamma_0^\pm \Psi^\kappa_{\mathrm{DL}} &= \mathsf{K}'_\kappa \mp \tfrac{1}{2}\mathsf{Id}, & \gamma_1^\pm \Psi^\kappa_{\mathrm{DL}} &= -\mathsf{W}_\kappa.
\end{aligned} \tag{6}$$

In what follows $(\cdot,\cdot)_\Gamma$ will stand for the $L^2(\Gamma)$-inner product

$$(\vartheta, \varphi)_\Gamma := \int_\Gamma \overline{\vartheta}\varphi\,\mathrm{d}S, \qquad \vartheta, \varphi \in L^2(\Gamma),$$

which can be extended to a duality pairing on $H^{-\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\Gamma)$. Adjoints of operators with respect to $(\cdot,\cdot)_\Gamma$ will be tagged by $*$.

Crucial for any variational formulation based on boundary integral operators will be the following three lemmata; see [32, Lemma 3.9.8], [16, Thm. 2].

LEMMA 3.1. *The following operators are compact:*

$$\begin{aligned}
\mathsf{V}_\kappa - \mathsf{V}_0 &: H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma),\\
\mathsf{K}_\kappa - \mathsf{K}_0 &: H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma),\\
\mathsf{K}'_\kappa - \mathsf{K}'_0 &: H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma),\\
\mathsf{W}_\kappa - \mathsf{W}_0 &: H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma).
\end{aligned}$$

The proof relies on the fact that both $\mathsf{V}_\kappa - \mathsf{V}_0$ and $\mathsf{K}_\kappa - \mathsf{K}_0$ turn out to be integral operators with continuous and bounded kernels, which ensures that they map into $H^1(\Gamma)$, which is compactly embedded in $H^{\frac{1}{2}}(\Gamma)$. Details can be found in [8, sect. 2]. This result combined with the ellipticity of both V_0 and W_0 in $H^{-\frac{1}{2}}(\Gamma)$ and $H^{\frac{1}{2}}(\Gamma)$, respectively, yields the next lemma.

LEMMA 3.2. *The operators V_κ and W_κ satisfy a generalized Gårding inequality in the sense that there exist a constant $\gamma > 0$ and compact operators*

$$\mathsf{T}_\mathsf{V} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma), \qquad \mathsf{T}_\mathsf{W} : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$$

such that

$$\begin{aligned}
\operatorname{Re}\left\{\left(\vartheta, \left(\mathsf{V}_\kappa + \mathsf{T}_\mathsf{V}\right)(\vartheta)\right)_\Gamma\right\} &\geq \gamma \left\|\vartheta\right\|^2_{H^{-\frac{1}{2}}(\Gamma)},\\
\operatorname{Re}\left\{\left(\left(\mathsf{W}_\kappa + \mathsf{T}_\mathsf{W}\right)(\varphi), \varphi\right)_\Gamma\right\} &\geq \gamma \left\|\varphi\right\|^2_{H^{\frac{1}{2}}(\Gamma)}
\end{aligned}$$

holds true for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ and $\varphi \in H^{\frac{1}{2}}(\Gamma)$.

Finally, K'_κ is the $(\cdot,\cdot)_\Gamma$-adjoint of K_κ up to a compact perturbation.

LEMMA 3.3. *There exists a compact operator $\mathsf{T}_\mathsf{K} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ such that*

$$\left(\mathsf{K}^*_\kappa(\vartheta), \varphi\right)_\Gamma = \left(\left(\mathsf{K}'_\kappa + \mathsf{T}_\mathsf{K}\right)(\vartheta), \varphi\right)_\Gamma$$

holds true for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ and $\varphi \in H^{\frac{1}{2}}(\Gamma)$, where K^*_κ denotes the $L^2(\Gamma)$-adjoint of K_κ.

Proof. Following [32, sect. 3.1] and [16], we recall the representations

$$\mathsf{K}_\kappa = \{\gamma_0\}_\Gamma \circ \mathcal{N}_\kappa \circ {\gamma_1}^*, \qquad \mathsf{K}'_\kappa = \{\gamma_1\}_\Gamma \circ \mathcal{N}_\kappa \circ {\gamma_0}^*,$$

where $\mathcal{N}_\kappa : H^{-1}_{\text{comp}}\left(\mathbb{R}^3\right) \to H^1_{\text{loc}}\left(\mathbb{R}^3\right)$ is the Newton potential for the Helmholtz kernel.

$${\mathsf{K}_\kappa}^* - \mathsf{K}'_\kappa = \{\gamma_1\}_\Gamma \circ \left(\mathcal{N}_\kappa - {\mathcal{N}_\kappa}^*\right) \circ {\gamma_0}^*.$$

Observe that

$$\left(\mathcal{N}_\kappa - {\mathcal{N}_\kappa}^*\right)(V)(\mathbf{x}) = \frac{i}{2\pi} \int_{\mathbb{R}^3} \frac{\sin(\kappa|\mathbf{x}-\mathbf{y}|)}{|\mathbf{x}-\mathbf{y}|} V(\mathbf{y})\, \mathrm{d}\mathbf{y}$$

is an integral operator with analytic kernel, which maps continuously $H^{-1}_{\text{comp}}\left(\mathbb{R}^3\right) \mapsto H^s_{\text{loc}}(\mathbb{R}^3)$ for any $s \in \mathbb{R}$. Thus, $\mathsf{K}^*_\kappa - \mathsf{K}'_\kappa : H^{-\frac{1}{2}}(\Gamma) \mapsto H^1(\Gamma)$ is continuous and the compact embedding $H^1(\Gamma) \hookrightarrow H^{-\frac{1}{2}}(\Gamma)$ finishes the proof. ∎

4. Calderón projectors. A crucial tool for the coupling of the variational equations on Ω^- and boundary integral equations on Γ are the two *Calderón projectors* [32, sect. 3.6]

$$\mathsf{P}_\pm := \begin{bmatrix} \frac{1}{2}\mathsf{Id} \pm \mathsf{K}_\kappa & \mp\mathsf{V}_\kappa \\ \mp\mathsf{W}_\kappa & \frac{1}{2}\mathsf{Id} \mp \mathsf{K}'_\kappa \end{bmatrix} : H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma) \mapsto H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma).$$

They arise from applying the trace operators $\gamma_0^\pm$ and $\gamma_1^\pm$ to (4) and using (6). The operators P_+ and P_- obviously satisfy the identity

$$\mathsf{P}_+ + \mathsf{P}_- = \mathsf{Id}. \tag{7}$$

The Calderón projectors can be used to characterize pairs of functions in $H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma)$ that are eligible as traces of Helmholtz solutions; see [36].

THEOREM 4.1. *If and only if* $(\varphi, \vartheta) \in H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma)$ *belongs to the range of* $\mathsf{P}_\pm$*, there is a Helmholtz solution* U *such that* $\varphi = \gamma_0^\pm U$ *and* $\vartheta = \gamma_1^\pm U$*.*

The theorem paves the way for establishing expressions for the exterior *Dirichlet-to-Neumann* map for the Helmholtz problem in Ω^+. This is the operator $\mathrm{DtN}_\kappa^+ : H^{\frac{1}{2}}(\Gamma) \mapsto H^{-\frac{1}{2}}(\Gamma)$ returning the Neumann traces of an exterior Helmholtz solution matching prescribed Dirichlet boundary conditions on Γ. Three different formulas can instantly be obtained from (4.1), at least formally, because the inverses of operators might not exist:

$$\mathrm{DtN}_\kappa^+ := {\mathsf{V}_\kappa}^{-1} \circ \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right), \tag{8}$$

$$\mathrm{DtN}_\kappa^+ := -\left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)^{-1} \circ \mathsf{W}_\kappa, \tag{9}$$

$$\mathrm{DtN}_\kappa^+ := -\mathsf{W}_\kappa + \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right) \circ {\mathsf{V}_\kappa}^{-1} \circ \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right). \tag{10}$$

Only the third formula reflects the essential symmetry of the boundary value problem in the case $\kappa = 0$. It will be the starting point for symmetric coupling.

Remark 4.2. If the incident wave U^{i} can be extended to an interior Helmholtz solution, which is evidently the case, when U^{i} is a plane wave or generated by a sound source compactly supported in Ω^+, then, by (4) and (5), its traces on Γ will fulfill

$$\begin{bmatrix} \gamma_0 U^{\mathrm{i}} \\ \gamma_1 U^{\mathrm{i}} \end{bmatrix} = \mathsf{P}_- \begin{bmatrix} \gamma_0 U^{\mathrm{i}} \\ \gamma_0 U^{\mathrm{i}} \end{bmatrix} \quad \Leftrightarrow \quad \mathsf{P}_+ \begin{bmatrix} \gamma_0 U^{\mathrm{i}} \\ \gamma_1 U^{\mathrm{i}} \end{bmatrix} = 0. \tag{11}$$

For the same reasons, the scattered field U^{s} satisfies

$$\begin{bmatrix} \gamma_0^+ U^{\mathrm{s}} \\ \gamma_1^+ U^{\mathrm{s}} \end{bmatrix} = \mathsf{P}_+ \begin{bmatrix} \gamma_0^+ U^{\mathrm{s}} \\ \gamma_1^+ U^{\mathrm{s}} \end{bmatrix}. \tag{12}$$

Since the total field in Ω^+ is given by $U = U^{\mathrm{s}} + U^{\mathrm{i}}$, we can eliminate U^{s} from (11) and (12) and end up with

$$\begin{bmatrix} \gamma_0^+ U \\ \gamma_1^+ U \end{bmatrix} = \mathsf{P}_+ \begin{bmatrix} \gamma_0^+ U \\ \gamma_1^+ U \end{bmatrix} - \begin{bmatrix} g_0 \\ g_1 \end{bmatrix}. \tag{13}$$

As above, Dirichlet-to-Neumann maps for the total field can be constructed from this relationship.

5. Classical symmetric coupling. For the sake of completeness we will review the classical approach to the coupling of boundary integral equations and variational formulation in Ω^- due to Costabel [15]. First, integration by parts shows that a solution U of problem (2) will fulfill

$$\mathsf{a}(U,V) - \left(\gamma_1^- U, \gamma_0^- V\right)_\Gamma = \mathsf{f}(v) \quad \forall v \in H^1\left(\Omega^-\right), \tag{14}$$

where we have used the abbreviations

$$\begin{aligned} \mathsf{a}(U,V) &:= \int_{\Omega^-} \mathbf{grad}\overline{U} \cdot \mathbf{grad} V - \kappa^2 n(\mathbf{x})\, \overline{U}\, V \,\mathrm{d}\mathbf{x}, && U, V \in H^1\left(\Omega^-\right), \\ \mathsf{f}(V) &:= \int_{\Omega^-} \overline{f}\, V \,\mathrm{d}\mathbf{x}, && V \in H^1\left(\Omega^-\right). \end{aligned}$$

LEMMA 5.1. *The sesquilinear form* a *satisfies a generalized Gårding inequality in the sense that there exists a constant* $\gamma > 0$ *and a compact sesquilinear form*

$$\mathsf{k} : H^1\left(\Omega^-\right) \times H^1\left(\Omega^-\right) \to \mathbb{C}$$

such that

$$\mathrm{Re}\left\{\mathsf{a}(U,U) + \mathsf{k}(U,U)\right\} \geq \gamma \left\|U\right\|^2_{H^1(\Omega^-)}$$

holds true for all $u \in H^1(\Omega^-)$.

Proof. The lemma is a straightforward consequence of the compact embedding $H^1(\Omega^-) \hookrightarrow L^2(\Omega^-)$. ∎

The variational problem associated with the classical symmetric coupling approach emerges by employing the transmission conditions of (2) and using the Dirichlet-to-Neumann map (10) to express $\gamma_1^- U$ in (14). In order to avoid the operator products occurring in (10) we also introduce $\gamma_1^+ U^{\mathrm{s}}$ as the new variable,

$$\vartheta := \left({\mathsf{V}_\kappa}^{-1} \circ \left(\mathsf{K}_\kappa - \tfrac{1}{2} Id\right)\right)\left(\gamma_0^- U + g_0\right) \in H^{-\frac{1}{2}}(\Gamma).$$

Thus, we end up with: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$ there holds

$$\begin{aligned} \mathsf{a}(U,V) + \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma - \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)(\vartheta), \gamma_0^- V\right)_\Gamma &= \widetilde{f}(V), \\ \left(\varphi, \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}_\kappa\right)\left(\gamma_0^- U\right)\right)_\Gamma + \left(\varphi, \mathsf{V}_\kappa(\vartheta)\right)_\Gamma &= \widetilde{g}(\varphi), \end{aligned} \tag{15}$$

where

$$\begin{aligned} \widetilde{f}(V) &:= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma - \left(\mathsf{W}_\kappa(g_0), \gamma_0^- V\right)_\Gamma, \\ \widetilde{g}(\varphi) &:= \left(\varphi, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)(g_0)\right)_\Gamma. \end{aligned}$$

Using the lemmata of the previous section it is not difficult to verify that the bilinear form associated with (15) satisfies a Gårding inequality. Unfortunately, this is no safeguard against spurious resonances.

Assume the resonance case, that is, κ^2, is a Dirichlet eigenvalue of $-\Delta$ in Ω^-. Then we can find $U \in H^1(\Omega^-) \setminus \{0\}$ such that

$$\Delta U + \kappa^2 U = 0 \quad \text{in } \Omega^- \quad \text{and} \quad U = 0 \quad \text{on } \Gamma.$$

Since $\gamma_0^- U = 0$, by Theorem 4.1 we have that

$$\begin{bmatrix} 0 \\ \gamma_1^- U \end{bmatrix} = \mathsf{P}_- \begin{bmatrix} 0 \\ \gamma_1^- U \end{bmatrix} = \begin{bmatrix} \mathsf{V}_\kappa \left(\gamma_1^- U\right) \\ \left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)\left(\gamma_1^- U\right) \end{bmatrix},$$

which means that $(0, \gamma_1^- U)$ provides a solution of (15) in the case $\widetilde{f} = \widetilde{g} = 0$.

Even in the resonance case, the right-hand side of (15) will be consistent and the variational problem still has solutions (U, ϑ), whose first component will still be unique. Alas, this is little comfort as far as numerical solution procedures are concerned: first, inevitable perturbations introduced by discretization will destroy the consistency of the right-hand side. Second, whenever κ^2 is merely close to an interior resonant frequency, the resulting linear systems of equations may not be useless, but will be extremely ill-conditioned; see the profound analysis of the impact of spurious resonances in the case of electromagnetic scattering given in [13].

So, from a numerical point of view suppressing spurious resonances is essential for the efficacy of methods based on boundary integral equations.

Remark 5.2. Under the assumptions made in Remark 4.2 we may use a symmetric Dirichlet-to-Neumann map derived from (13). This will lead to a coupled variational problem of the form (15) with much simpler right-hand sides $\widetilde{f}(V) = \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma$ and $\widetilde{g}(\varphi) = -\left(\varphi, g_0\right)_\Gamma$.

6. Transformed traces. As pointed out at the end of the previous section, the existence of spurious resonances is directly linked to the fact that for certain κ there are nontrivial interior Helmholtz solutions U that satisfy $\gamma_0^- U = 0$. We know that there are Robin-type (mixed) boundary conditions that ensure the unique solvability of the corresponding boundary value problem for $-\Delta U - \kappa^2 U = 0$ in Ω^-. Note that we can rely on two Robin-type boundary operators to state the transmission conditions of (2) as long as we can recover from them the conventional Dirichlet and Neumann trace. In fact, this idea can serve as the starting point for the derivation of all CFIEs. Here, it motivates the introduction of the following generic trace transformation operator:

$$\mathcal{T} := \begin{bmatrix} \mathsf{A} & \mathsf{B} \\ \mathsf{C} & \mathsf{D} \end{bmatrix} : H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma). \tag{16}$$

We demand that the interior homogeneous "Dirichlet problem" for $-\Delta U - \kappa^2 U = 0$ and the modified traces have a unique solution for every κ. In light of Theorem 4.1 this amounts to the following assumption.

ASSUMPTION 6.1. *The trace transformation operator $\mathcal{T}$ satisfies*

$$\operatorname{Range}\left(\mathcal{T} \circ \mathsf{P}_-\right) \cap \left(\{0\} \times H^{-\frac{1}{2}}(\Gamma)\right) = \{0\}.$$

Then one can use $\mathcal{T}$, build associated Calderón projectors for the modified traces, derive symmetrically coupled variational problems, and check their properties. Here, we would like to skip this tedious process of creative discovery and present the final finding on what is required for $\mathcal{T}$.

ASSUMPTION 6.2. *The blocks of the transformation operator $\mathcal{T}$ from* (16) *are assumed to possess the following properties:*

1. $\mathcal{T} : H^{-\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\Gamma)$ *is bijective,*
2. $\mathsf{A} : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ *is bounded and bijective,*
3. $\mathsf{B} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ *is compact,*

4. $\mathsf{C} : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ *is compact,*
5. $\mathsf{D} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ *is bounded and bijective.*

The first requirement enables us to retrieve the conventional Dirichlet and Neumann trace from their transformed counterparts. This is essential because it is these traces that will invariably occur in (14) so that we have to resort to them in one way or another when pursuing the coupling of (14) with boundary integral equations. Switching back and forth between conventional and transformed traces employs the following splitting of the trace transformation operator:

$$\mathcal{T} = \mathcal{R} + \mathcal{S}, \qquad \mathcal{R} := \begin{bmatrix} \mathsf{A} & 0 \\ 0 & \mathsf{D} \end{bmatrix}, \qquad \mathcal{S} := \begin{bmatrix} 0 & \mathsf{B} \\ \mathsf{C} & 0 \end{bmatrix}. \tag{17}$$

Based on the splitting above, we define the following generalized Calderón projectors:

$$\mathcal{P}_\pm := \mathcal{R}^{-1} \circ (\mathcal{T} \circ \mathsf{P}_\pm - \mathcal{S}). \tag{18}$$

Note that they are meant to act on conventional traces. Let us make the transformed exterior Calderón projector more explicit: an elementary computation yields

$$\mathcal{P}_+ = \begin{bmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{bmatrix}, \tag{19}$$

where the entries of the operator matrix are given by

$$\mathbb{A} := \tfrac{1}{2}\mathsf{Id} + \mathsf{K}_\kappa - \mathsf{A}^{-1} \circ \mathsf{B} \circ \mathsf{W}_\kappa, \tag{20}$$

$$\mathbb{B} := -\mathsf{A}^{-1} \circ \mathsf{B} \circ \left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right) - \mathsf{V}_\kappa, \tag{21}$$

$$\mathbb{C} := \mathsf{D}^{-1} \circ \mathsf{C} \circ \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right) - \mathsf{W}_\kappa, \tag{22}$$

$$\mathbb{D} := \tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa - \mathsf{D}^{-1} \circ \mathsf{C} \circ \mathsf{V}_\kappa. \tag{23}$$

An analogue of Theorem 4.1 still holds for the transformed Calderón projectors.

LEMMA 6.3. *If and only if U is an exterior/interior radiating Helmholtz solution we have*

$$\mathcal{P}_\pm \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} = \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix}.$$

Proof. As $\mathcal{T}$ is one-to-one, we immediately conclude from Theorem 4.1 that U is an exterior/interior radiating Helmholtz solution if and only if

$$\begin{array}{c} \mathsf{P}_\pm \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} = \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} \\ \Updownarrow \\ (\mathcal{T} \circ \mathsf{P}_\pm) \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} = \mathcal{T} \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} = (\mathcal{R} + \mathcal{S}) \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} \\ \Updownarrow \\ \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix} = \mathcal{R}^{-1} \circ (\mathcal{T} \circ \mathsf{P}_\pm - \mathcal{S}) \begin{bmatrix} \gamma_0^\pm U \\ \gamma_1^\pm U \end{bmatrix}. \end{array} \tag{24}$$

□

Now, the same formal manipulations as in section 4 yield the following operator expression for the Dirichlet-to-Neumann map:

$$\mathrm{DtN}_\kappa^+ := \mathbb{C} + \mathbb{D} \circ \mathbb{B}^{-1} \circ (\mathsf{Id} - \mathbb{A}), \tag{25}$$

which maps exterior Dirichlet traces of radiating Helmholtz solutions U to exterior Neumann traces.

Remark 6.4. Again, if the incident wave U^{i} can be extended to an interior Helholtz solution, then we can apply the trace transformation operator to (13) and end up with

$$\mathcal{T}\begin{bmatrix}\gamma_0^+U\\ \gamma_1^+U\end{bmatrix} = (\mathcal{T}\circ\mathsf{P}_+)\begin{bmatrix}\gamma_0^+U\\ \gamma_1^+U\end{bmatrix} - \mathcal{T}\begin{bmatrix}g_0\\ g_1\end{bmatrix}. \tag{26}$$

Using the operator splitting (17) and definition (18) of the generalized Calderón projector we can eliminate the trace transformation operator $\mathcal{T}$ from the left-hand side of (26) and obtain

$$\begin{bmatrix}\gamma_0^+U\\ \gamma_1^+U\end{bmatrix} = \mathcal{P}_+\begin{bmatrix}\gamma_0^+U\\ \gamma_1^+U\end{bmatrix} - \left(\mathcal{R}^{-1}\circ\mathcal{T}\right)\begin{bmatrix}g_0\\ g_1\end{bmatrix}. \tag{27}$$

As above this relationship can be used to construct new Dirchlet-to-Neumann maps for the total field.

We end this section with an easily verifiable criterion telling us when Assumption 6.1 is satisfied.

LEMMA 6.5. *If the following equivalence holds:*

$$\mathrm{Im}\left\{\left(\vartheta, \left(\mathsf{A}^{-1}\circ\mathsf{B}\right)(\vartheta)\right)_\Gamma\right\} = 0 \Leftrightarrow \vartheta = 0,$$

then

$$\mathrm{Range}\left(\mathcal{T}\circ\mathsf{P}_-\right)\cap\left(\{0\}\times H^{-\frac{1}{2}}(\Gamma)\right) = \{0\}.$$

Proof. If $\xi\in H^{-\frac{1}{2}}(\Gamma)$ satisfies

$$\begin{bmatrix}0\\ \xi\end{bmatrix}\in\mathrm{Range}\left(\mathcal{T}\circ\mathsf{P}_-\right),$$

then there exists $\vartheta\in H^{\frac{1}{2}}(\Gamma)$ and $\varphi\in H^{-\frac{1}{2}}(\Gamma)$ such that

$$\begin{bmatrix}0\\ \xi\end{bmatrix} = (\mathcal{T}\circ\mathsf{P}_-)\begin{bmatrix}\vartheta\\ \varphi\end{bmatrix}.$$

Taking the transformed interior traces of the function

$$U(\mathbf{x}) := -\Psi_{\mathrm{DL}}^\kappa(\vartheta)(\mathbf{x}) + \Psi_{\mathrm{SL}}^\kappa(\varphi)(\mathbf{x}), \qquad \mathbf{x}\in\Omega^-$$

gives us the following set of equations:

$$\mathcal{T}\begin{bmatrix}\gamma_0^-U\\ \gamma_1^-U\end{bmatrix} = (\mathcal{T}\circ\mathsf{P}_-)\begin{bmatrix}\vartheta\\ \varphi\end{bmatrix} = \begin{bmatrix}0\\ \xi\end{bmatrix}. \tag{28}$$

Thus U is a solution to the boundary value problem

$$\Delta U + \kappa^2 U = 0 \quad \text{in } \Omega^-, \tag{29}$$

$$\mathsf{A}\left(\gamma_0^-U\right) + \mathsf{B}\left(\gamma_1^-U\right) = 0 \quad \text{on } \Gamma. \tag{30}$$

Recalling (14) and using that A is bijective, we obtain

$$\mathsf{a}(U,U) - \left(\gamma_1^- U, \gamma_0^- U\right)_\Gamma = \mathsf{a}(U,U) + \left(\gamma_1^- U, (\mathsf{A}^{-1}\circ\mathsf{B})\left(\gamma_1^- U\right)\right)_\Gamma = 0.$$

Since, $\mathsf{a}(U,U) \in \mathbb{R}$, by taking the imaginary part we get

$$\operatorname{Im}\left\{\left(\gamma_1^- U, \left(\mathsf{A}^{-1}\circ\mathsf{B}\right)\left(\gamma_1^- U\right)\right)_\Gamma\right\} = 0.$$

Thus, the assumption of the lemma implies $\gamma_1^- U = 0$, and via (30) we conclude $\gamma_0^- U = 0$. Eventually, (28) shows that $\xi = 0$. ∎

7. Stabilized coupling. Parallel to the approach in section 5, we use (14) in combination with the transformed Dirichlet-to-Neumann map (25) and introduce the new variable

$$\vartheta := -\left(\mathbb{B}^{-1}\circ(\mathsf{Id}-\mathbb{A})\right)\left(\gamma_0^- U + g_0\right) \in H^{-\frac{1}{2}}(\Gamma). \tag{31}$$

If U solves the Helmholtz transmission problem (2), then $\gamma_0^- U + g_0 = \gamma_0^+ U^{\mathrm{s}}$, and we learn from Lemma 6.3 and (19) that actually $\vartheta = -\gamma_1^+ U^{\mathrm{s}}$. As in the case of classical coupling, ϑ will supply the exterior Neumann trace of the scattered field.

Thus we arrive at the following *regularized variational formulation*: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$ there holds

$$\begin{aligned} \mathsf{a}(U,V) - \left(\mathbb{C}\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma + \left(\mathbb{D}(\vartheta), \gamma_0^- V\right)_\Gamma &= \widehat{f}(V), \\ \left(\varphi, (\mathbb{A}-\mathsf{Id})\left(\gamma_0^- U\right)\right)_\Gamma - \left(\varphi, \mathbb{B}(\vartheta)\right)_\Gamma &= \widehat{g}(\varphi), \end{aligned} \tag{32}$$

where

$$\widehat{f}(V) := \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma + \left(\mathbb{C}(g_0), \gamma_0^- V\right)_\Gamma, \tag{33}$$

$$\widehat{g}(V) := \left(\varphi, (\mathsf{Id}-\mathbb{A})(g_0)\right)_\Gamma. \tag{34}$$

We first investigate the $H^1(\Omega^-)\times H^{-\frac{1}{2}}(\Gamma)$-coercivity of the sesquilinear form underlying (32). From Assumption 6.2 it is immediate that the operators $\mathsf{A}^{-1}\circ\mathsf{B}$, $\mathsf{D}^{-1}\circ\mathsf{C}$ are compact. This plays a key role in the proofs of the following two lemmata.

LEMMA 7.1. *There exists a constant $\gamma > 0$ and compact operators*

$$\mathsf{T}_{\mathbb{B}} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma), \qquad \mathsf{T}_{\mathbb{C}} : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$$

such that

$$\begin{aligned} -\operatorname{Re}\left\{\left(\vartheta, (\mathbb{B}+\mathsf{T}_{\mathbb{B}})(\vartheta)\right)_\Gamma\right\} &\geq \gamma \left\|\vartheta\right\|^2_{H^{-\frac{1}{2}}(\Gamma)}, \\ -\operatorname{Re}\left\{\left(\varphi, (\mathbb{C}+\mathsf{T}_{\mathbb{C}})(\varphi)\right)_\Gamma\right\} &\geq \gamma \left\|\varphi\right\|^2_{H^{\frac{1}{2}}(\Gamma)} \end{aligned}$$

for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ and $\varphi \in H^{\frac{1}{2}}(\Gamma)$.

Proof. Using (21) and (22) a straightforward application of Lemma 3.2 yields

$$\begin{aligned} &\operatorname{Re}\left\{-\left(\vartheta, \mathbb{B}(\vartheta)\right)_\Gamma - \left(\vartheta, \left(\mathsf{A}^{-1}\circ\mathsf{B}\circ\left(\tfrac{1}{2}\mathsf{Id}+\mathsf{K}'_\kappa\right)\right)(\vartheta)\right)_\Gamma + \left(\vartheta, \mathsf{T}_{\mathsf{V}}(\vartheta)\right)_\Gamma\right\} \\ &\quad = \operatorname{Re}\left\{\left(\vartheta, (\mathsf{V}_\kappa + \mathsf{T}_{\mathsf{V}})(\vartheta)\right)_\Gamma\right\} \geq \gamma \left\|\vartheta\right\|^2_{H^{-\frac{1}{2}}(\Gamma)}, \\ &\operatorname{Re}\left\{-\left(\mathbb{C}(\varphi), \varphi\right)_\Gamma + \left(\left(\mathsf{D}^{-1}\circ\mathsf{C}\circ\left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)\right)(\varphi), \varphi\right)_\Gamma + \left(\mathsf{T}_{\mathsf{W}}(\varphi), \varphi\right)_\Gamma\right\} \\ &\quad = \operatorname{Re}\left\{\left((\mathsf{W}_\kappa + \mathsf{T}_{\mathsf{W}})(\varphi), \varphi\right)_\Gamma\right\} \geq \gamma \left\|\varphi\right\|^2_{H^{\frac{1}{2}}(\Gamma)} \end{aligned}$$

for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, $\varphi \in H^{\frac{1}{2}}(\Gamma)$. ∎

LEMMA 7.2. *There exist compact operators*

$$\mathsf{T}_{\mathbb{A}} : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma), \qquad \mathsf{T}_{\mathbb{D}} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$$

such that

$$(\vartheta, (\mathbb{A} - \mathsf{Id} + \mathsf{T}_{\mathbb{A}})(\varphi))_\Gamma + ((\mathbb{D} + \mathsf{T}_{\mathbb{D}})(\vartheta), \varphi)_\Gamma = 0$$

for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, $\varphi \in H^{\frac{1}{2}}(\Gamma)$.

Proof. We begin with an application of Lemma 3.3 and obtain

$$\begin{aligned}&\left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)(\vartheta), \varphi\right)_\Gamma + \left(\vartheta, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)(\varphi)\right)_\Gamma \\ &\quad = \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}^*_\kappa + \mathsf{T}_{\mathsf{K}}\right)(\vartheta), \varphi\right)_\Gamma - \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}^*_\kappa\right)(\vartheta), \varphi\right)_\Gamma\end{aligned}$$

for all $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, $\varphi \in H^{\frac{1}{2}}(\Gamma)$. Using this result we finally arrive at the following equation:

$$\begin{aligned}&(\vartheta, (\mathbb{A} - \mathsf{Id})(\varphi))_\Gamma + (\mathbb{D}(\vartheta), \varphi)_\Gamma \\ &\quad = \left(\left(\mathsf{T}_{\mathsf{K}} - \mathsf{D}^{-1} \circ \mathsf{C} \circ \mathsf{K}_\kappa\right)(\vartheta), \varphi\right)_\Gamma - \left(\vartheta, \left(\mathsf{A}^{-1} \circ \mathsf{B} \circ \mathsf{W}_\kappa\right)(\varphi)\right)_\Gamma,\end{aligned}$$

which holds for arbitrary $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, $\varphi \in H^{\frac{1}{2}}(\Gamma)$. ∎

Summing up, from the previous lemmata and Lemma 5.1 we conclude that the sesquilinear form of the regularized variational problem (32) satisfies a Gårding inequality in $H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma)$. It remains to establish uniqueness of solutions, which amounts to confirming that (32) is really immune to spurious resonances.

THEOREM 7.3. *Solutions to the regularized variational problem* (32) *are unique.*

Proof. In order to establish uniqueness of solutions of (32) we consider the case $\widehat{f} = \widehat{g} = 0$: seek $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$ there holds

$$\mathsf{a}(U, V) - \left(\mathbb{C}\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma + \left(\mathbb{D}(\vartheta), \gamma_0^- V\right)_\Gamma = 0, \tag{35}$$

$$\left(\varphi, (\mathbb{A} - \mathsf{Id})\left(\gamma_0^- U\right)\right)_\Gamma - (\varphi, \mathbb{B}(\vartheta))_\Gamma = 0. \tag{36}$$

Using integration by parts, we obtain

$$\Delta U + \kappa^2 n(\mathbf{x}) U = 0 \qquad \text{in } \Omega^-.$$

As a consequence, $\mathsf{a}(U, V) = \left(\gamma_1^- U, \gamma_0^- V\right)_\Gamma$. Plugging this identity into (35) and using the definition of $\mathcal{P}_+$ from (18), the identity

$$\begin{bmatrix} \gamma_0^- U \\ \gamma_1^- U \end{bmatrix} = \mathcal{P}_+ \begin{bmatrix} \gamma_0^- U \\ -\vartheta \end{bmatrix} \tag{37}$$

is immediate. By the definition of $\mathcal{P}_+$ and (7)

$$\begin{aligned}\mathcal{P}_+ &= \mathcal{R}^{-1} \circ (\mathcal{T} \circ \mathsf{P}_+ - \mathcal{S}) \\ &= \mathcal{R}^{-1} \circ (\mathcal{T} \circ (\mathsf{Id} - \mathsf{P}_-) - \mathcal{S}) \\ &= \mathcal{R}^{-1} \circ (\mathcal{R} + \mathcal{S} - \mathcal{T} \circ \mathsf{P}_- - \mathcal{S}) \\ &= \mathsf{Id} - \mathcal{R}^{-1} \circ \mathcal{T} \circ \mathsf{P}_-,\end{aligned}$$

and we infer

$$\mathcal{T} \circ \mathsf{P}_- = \mathcal{R} \circ (\mathsf{Id} - \mathcal{P}_+).$$

Together with (37) this identity confirms

$$(\mathcal{T} \circ \mathsf{P}_-) \begin{bmatrix} \gamma_0^- U \\ -\vartheta \end{bmatrix} = -\mathcal{R} \begin{bmatrix} 0 \\ \gamma_1^- U + \vartheta \end{bmatrix} \in \left(\{0\} \times H^{\frac{1}{2}}(\Gamma) \right),$$

and, by Assumption 6.1,

$$\mathcal{R} \begin{bmatrix} 0 \\ \gamma_1^- U + \vartheta \end{bmatrix} = 0 \quad \Rightarrow \quad \mathsf{D}(\gamma_1^- U + \vartheta) = 0.$$

Next, from Assumption 6.2, 5., we conclude that

$$\gamma_1^- U = -\vartheta. \tag{38}$$

From this and (37) we directly obtain, as in (24) in the proof of Lemma 6.3,

$$\begin{bmatrix} \gamma_0^- U \\ \gamma_1^- U \end{bmatrix} = \mathsf{P}_+ \begin{bmatrix} \gamma_0^- U \\ -\vartheta \end{bmatrix}.$$

Hence, by virtue of Theorem 4.1, setting

$$W(\mathbf{x}) := \begin{cases} U(\mathbf{x}), & \mathbf{x} \in \Omega^-, \\ \Psi_{\mathrm{DL}}^\kappa \left(\gamma_0^- U \right)(\mathbf{x}) - \Psi_{\mathrm{SL}}^\kappa (-\vartheta)(\mathbf{x}), & \mathbf{x} \in \Omega^+ \end{cases}$$

provides us with a solution to the Helmholtz transmission problem with zero right-hand side, and uniqueness of solutions to the Helmholtz transmission problem ensures $U = 0$ and, by (38), $\vartheta = 0$. We have demonstrated that (35) and (36) only possess the trivial solution and this finishes the proof. ∎

Eventually, the existence of solutions to the variational problem (32) follows from Theorem 7.3 and a Fredholm argument; see, for instance, [29, Thm. 2.33].

Finally, the arguments in the proof of Theorem 7.3 have also confirmed that we really get information about the solution of the Helmholtz transmission problem from (32).

COROLLARY 7.4. *If* $(W, \vartheta) \in H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma)$ *solves* (32)*, then* $W = U$ *and* $\vartheta = -\gamma_1^+ U^s$ *with* (U, U^s) *solving* (2)*.*

8. Regularization operators. In this section we present a rather simple specimen of a trace transformation operator $\mathcal{T}$, which satisfies both Assumptions 6.2 and 6.1. Its main ingredient is a *regularizing operator*

$$\mathsf{M} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma),$$

which satisfies the following assumption.

ASSUMPTION 8.1. *We suppose that*

1. $\mathsf{M} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ *is compact, and*
2. $(\vartheta, \mathsf{M}(\vartheta))_\Gamma > 0$ *for all* $\vartheta \in H^{-\frac{1}{2}}(\Gamma) \setminus \{0\}$.

Various examples of such operators are discussed in [9]. Below we will present a concrete representative. Then, for $\eta \in \mathbb{R} \setminus \{0\}$ we choose the following trace transformation operators:

$$\mathcal{T}_1 := \begin{bmatrix} \mathsf{Id} & i\eta\mathsf{M} \\ i\eta & \mathsf{Id} \end{bmatrix}, \qquad \mathcal{T}_2 := \begin{bmatrix} \mathsf{Id} & i\eta\mathsf{M} \\ 0 & \mathsf{Id} \end{bmatrix}. \tag{39}$$

Now, we have to verify Assumptions 6.1 and 6.2. We note that Assumption 6.1 can instantly be concluded from Assumption 8.1, 2., and Lemma 6.5. Items 2. through 5. of Assumption 6.2 are evidently appealing to 1. in Assumption 8.1. It is also obvious that $\mathcal{T}_2$ is bijective with

$$\mathcal{T}_2{}^{-1} = \begin{bmatrix} \mathsf{Id} & -i\eta\mathsf{M} \\ 0 & \mathsf{Id} \end{bmatrix}.$$

It remains to be established whether $\mathcal{T}_1$ is bijective, too. The key will be the following lemma.

LEMMA 8.2. *For $\zeta \in \mathbb{R}_+$ or $\zeta \in i\mathbb{R}$ the following operators are bijective:*

$$\mathsf{Id} + \zeta\mathsf{M} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma), \qquad \mathsf{Id} + \zeta\mathsf{M} : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma).$$

Proof. We verify that the operators have trivial kernel. In the first case we find that $(\mathsf{Id} + \zeta\mathsf{M})(\vartheta) = 0$ implies

$$(\vartheta, \varphi)_\Gamma + \overline{\zeta}\,(\mathsf{M}(\vartheta), \varphi)_\Gamma = 0,$$

which holds true for all $\varphi \in H^{\frac{1}{2}}(\Gamma)$. We choose $\varphi := \mathsf{M}(\vartheta)$ and we obtain

$$(\vartheta, \mathsf{M}(\vartheta))_\Gamma + \overline{\zeta}\,\|\mathsf{M}(\vartheta)\|^2_{L^2(\Gamma)} = 0.$$

For either $\zeta > 0$ or $\zeta \in i\mathbb{R}$ Assumption 8.1, 2., implies

$$(\vartheta, \mathsf{M}(\vartheta))_\Gamma = 0 \Leftrightarrow \vartheta = 0.$$

Thanks to Assumption 8.1, 1., we have a Fredholm alternative argument [29, Thm. 2.27] at our disposal and conclude that the operator $\mathsf{Id} + \zeta\mathsf{M} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ is surjective from the fact that it is injective.

In the $H^{\frac{1}{2}}(\Gamma)$-setting $(\mathsf{Id} + \zeta\mathsf{M})(\varphi) = 0$ is equivalent to

$$(\vartheta, \varphi)_\Gamma + \zeta\,(\vartheta, \mathsf{M}(\varphi))_\Gamma = 0 \quad \forall \vartheta \in H^{-\frac{1}{2}}(\Gamma).$$

The same reasoning as above also settles this case. ∎

The lemma tells us that the formal inverse

$$\mathcal{T}_1{}^{-1} = \left(\mathsf{Id} + \eta^2\mathsf{M}\right)^{-1} \circ \begin{bmatrix} \mathsf{Id} & -i\eta\mathsf{M} \\ -i\eta & \mathsf{Id} \end{bmatrix}$$

is well defined, which implies Assumption 6.2, 1., for $\mathcal{T}_1$.

A particularly convenient regularizing operator has been presented in [8]: there, $\mathsf{M} : H^{-1}(\Gamma) \to H^1(\Gamma)$ is implicitly defined by

$$(\mathbf{grad}_\Gamma \mathsf{M}(p), \mathbf{grad}_\Gamma q)_\Gamma + (\mathsf{M}(p), q)_\Gamma = (p, q)_\Gamma \tag{40}$$

for all $q \in H^1(\Gamma)$. It is an easy exercise to verify Assumption 8.1 for this M; see [8, sect. 4.2]. For later use we define the following sesquilinear form:

$$\mathsf{b}(p, q) := (\mathbf{grad}_\Gamma p, \mathbf{grad}_\Gamma q)_\Gamma + (p, q)_\Gamma, \qquad p, q \in H^1(\Gamma), \tag{41}$$

which allows us to restate definition (40) as

$$\mathsf{b}(\mathsf{M}(p), q) = (p, q)_\Gamma \quad \forall q \in H^1(\Gamma). \tag{42}$$

9. Mixed regularized variational formulations. Using the two trace transformation operators we obtain two variational formulations which are free from spurious resonances. However, from the point of view of boundary element discretization, they are not yet useful, because they still contain products of (nonlocal) operators that elude a straightforward Galerkin discretization. To get rid of the operator products, we rely on the usual trick and introduce extra unknown functions. We discuss the resulting variational problems for the trace transformation operators $\mathcal{T}_1$ and $\mathcal{T}_2$ from (39) and M given by (40).

Case $\mathcal{T} = \mathcal{T}_1$: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$

(43)
$$\begin{aligned} \mathsf{a}(U,V) - \left(\left(i\eta\left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right) - \mathsf{W}_\kappa\right)\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma \\ + \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa - i\eta\mathsf{V}_\kappa\right)(\vartheta), \gamma_0^- V\right)_\Gamma = f_1(V), \\ \left(\varphi, \left(i\eta\mathsf{M}\circ\left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right) + \mathsf{V}_\kappa\right)(\vartheta)\right)_\Gamma + \left(\varphi, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id} - i\eta\mathsf{M}\circ\mathsf{W}_\kappa\right)\left(\gamma_0^- U\right)\right)_\Gamma = g_1(\varphi), \end{aligned}$$

where the right-hand sides are given by

$$\begin{aligned} f_1(V) &:= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma + \left(\left(i\eta\left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right) - \mathsf{W}_\kappa\right)(g_0), \gamma_0^- V\right)_\Gamma, \\ g_1(\varphi) &:= \left(\varphi, \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}_\kappa + i\eta\mathsf{M}\circ\mathsf{W}_\kappa\right)(g_0)\right)_\Gamma. \end{aligned}$$

Case $\mathcal{T} = \mathcal{T}_2$: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$

(44)
$$\begin{aligned} \mathsf{a}(U,V) + \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma + \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma = f_2(V), \\ \left(\varphi, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id} - i\eta\mathsf{M}\circ\mathsf{W}_\kappa\right)\left(\gamma_0^- U\right)\right)_\Gamma + \left(\varphi, \left(i\eta\mathsf{M}\circ\left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right) + \mathsf{V}_\kappa\right)(\vartheta)\right)_\Gamma = g_2(\varphi), \end{aligned}$$

where the right-hand sides are given by

$$\begin{aligned} f_2(V) &:= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma - \left(\mathsf{W}_\kappa(g_0), \gamma_0^- V\right)_\Gamma, \\ g_2(\varphi) &:= \left(\varphi, \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}_\kappa + i\eta\mathsf{M}\circ\mathsf{W}_\kappa\right)(g_0)\right)_\Gamma. \end{aligned}$$

Both regularized variational formulations contain the same operator products, namely

$$\begin{aligned} -\mathbb{B} &= \mathsf{V}_\kappa + i\eta\mathsf{M}\circ\left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right), \\ \mathbb{A} - \mathsf{Id} &= \mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id} - i\eta\mathsf{M}\circ\mathsf{W}_\kappa. \end{aligned}$$

This suggests that we introduce the new variable

$$p := \left(\mathsf{M}\circ\left(\tfrac{1}{2} + \mathsf{K}'_\kappa\right)\right)(\vartheta) - (\mathsf{M}\circ\mathsf{W}_\kappa)\left(\gamma_0^- U + g_0\right) \in H^1(\Gamma), \tag{45}$$

which converts (32) into the following two variational problems. The first arises from using $\mathcal{T}_1$: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, and $p \in H^1(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$, and $q \in H^1(\Gamma)$ there holds

(46)
$$\begin{aligned} \mathsf{a}(U,V) + i\eta\left(\left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma + \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma \\ + \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)(\vartheta), \gamma_0^- V\right)_\Gamma + i\eta\left(\mathsf{V}_\kappa(\vartheta), \gamma_0^- V\right)_\Gamma = f_1(V), \\ \left(\varphi, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)\left(\gamma_0^- U\right)\right)_\Gamma + (\varphi, \mathsf{V}_\kappa(\vartheta))_\Gamma + i\eta(\varphi, p)_\Gamma = g_1(\varphi), \\ \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), q\right)_\Gamma - \left(\left(\mathsf{K}'_\kappa + \tfrac{1}{2}\mathsf{Id}\right)(\vartheta), q\right)_\Gamma + \mathsf{b}(p,q) = h_1(q), \end{aligned}$$

with right-hand sides

$$\begin{aligned} f_1(V) &:= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma - i\eta \left(\left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)(g_0), \gamma_0^- V\right)_\Gamma - \left(\mathsf{W}_\kappa(g_0), \gamma_0^- V\right)_\Gamma, \\ g_1(\varphi) &:= \left(\varphi, \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}_\kappa\right)(g_0)\right)_\Gamma, \\ h_1(q) &:= -\left(\mathsf{W}_\kappa(g_0), q\right)_\Gamma. \end{aligned}$$

The second arises from using $\mathcal{T}_2$: find $U \in H^1(\Omega^-)$, $\vartheta \in H^{-\frac{1}{2}}(\Gamma)$, and $p \in H^1(\Gamma)$ such that for all $V \in H^1(\Omega^-)$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$, and $q \in H^1(\Gamma)$ there holds

$$\begin{aligned} \mathsf{a}(U,V) + \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), \gamma_0^- V\right)_\Gamma + \left(\left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)(\vartheta), \gamma_0^- V\right)_\Gamma &= f_2(V), \\ \left(\varphi, \left(\mathsf{K}_\kappa - \tfrac{1}{2}\mathsf{Id}\right)\left(\gamma_0^- U\right)\right)_\Gamma + (\varphi, \mathsf{V}_\kappa(\vartheta))_\Gamma + i\eta(\varphi, p)_\Gamma &= g_2(V), \\ \left(\mathsf{W}_\kappa\left(\gamma_0^- U\right), q\right)_\Gamma - \left(\left(\mathsf{K}'_\kappa + \tfrac{1}{2}\mathsf{Id}\right)(\vartheta), q\right)_\Gamma + \mathsf{b}(p,q) &= h_2(q), \end{aligned} \tag{47}$$

with right-hand sides

$$\begin{aligned} f_2(V) &:= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma - \left(\mathsf{W}_\kappa(g_0), \gamma_0^- V\right)_\Gamma, \\ g_2(\varphi) &:= \left(\varphi, \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{K}_\kappa\right)(g_0)\right)_\Gamma, \\ h_2(q) &:= -\left(\mathsf{W}_\kappa(g_0), q\right)_\Gamma. \end{aligned}$$

In order to settle the issue of existence and uniqueness of solutions of (46) and (47) we first observe that by the very definition of M in (40) and (45) the first two components of any solution (U, ϑ, p) of (46) and (47) will also solve (43) and (44), respectively. Since these are special cases of (32) and both $\mathcal{T}_1$ and $\mathcal{T}_2$ are valid trace transformation operators, Theorem 7.3 yields uniqueness.

Next, it follows directly from the compact embeddings $H^1(\Gamma) \hookrightarrow H^{\frac{1}{2}}(\Gamma)$ and $H^{-\frac{1}{2}}(\Gamma) \hookrightarrow H^{-1}(\Gamma)$ that all new off-diagonal terms are compact sesquilinear forms. Since b is $H^1(\Gamma)$-elliptic, we obtain that the sesquilinear forms for both variational formulations satisfy a generalized Gårding inequality.

Again, a Fredholm argument ensures the existence of solutions from the uniqueness result. The statement of Corollary 7.4 directly carries over to the (U, ϑ)-components of (43) and (44). Thus we have obtained two well-posed variational formulations which yield weak solutions to the Helmholtz transmission problem 2 and which are also amenable to standard Galerkin discretizations.

We finish this section by making an important observation: (45) can be recast into

$$p = \left(\mathsf{M} \circ \left(\tfrac{1}{2} + \mathsf{K}'_\kappa\right)\right)(\vartheta) - (\mathsf{M} \circ \mathsf{W}_\kappa)\left(\gamma_0^- U + g_0\right).$$

At second glance, we realize that $p = 0$, if (U, ϑ) solves (43) and (44), respectively. This directly follows from Corollary 7.4, Theorem 4.1, and the definition of the exterior Calderón projector P_+. In short, p is a "dummy variable."

Remark 9.1. Under the assumptions made in Remark 4.2 we can derive a Dirichlet-to-Neumann map from (27) to obtain coupled variational problems of the form (46) and (47) with much simpler right-hand sides:

$$\begin{aligned} f_1(V) &= \mathsf{f}(V) + i\eta\left(g_0, \gamma_0^- V\right)_\Gamma - \left(g_1, \gamma_0^- V\right)_\Gamma, & f_2(V) &= \mathsf{f}(V) - \left(g_1, \gamma_0^- V\right)_\Gamma, \\ g_1(\varphi) &= (\varphi, g_0)_\Gamma, & g_2(\varphi) &= +(\varphi, g_0)_\Gamma, \\ h_1(q) &= -(g_1, q)_\Gamma, & h_2(q) &= -(g_1, q)_\Gamma. \end{aligned} \tag{48}$$

The solution U in Ω^- will remain the same.

10. Galerkin discretization. With operator products removed, the Galerkin discretization of the variational problems (43) and (44) is easily achieved by restricting them to finite element subspaces $\mathcal{V}_h$ of $H^1(\Omega^-)$ and boundary element subspaces Θ_h and $\mathcal{Q}_h$ of $H^{-\frac{1}{2}}(\Gamma)$ and $H^1(\Gamma)$, respectively. A powerful theorem about the Galerkin approximation of coercive variational problems, see [33] and [37], will then yield the *asymptotic quasi-optimality* of the Galerkin solutions: assuming a minimal resolution of $\mathcal{V}_h$, Θ_h, and $\mathcal{Q}_h$, existence and uniqueness of discrete solutions $(U_h, \vartheta_h, p_h) \in \mathcal{V}_h \times \Theta_h \times \mathcal{Q}_h$ of (43) and (44) is guaranteed and we have the a priori error estimate

$$\begin{aligned}(49)\quad &\|U-U_h\|_{H^1(\Omega^-)} + \|\vartheta-\vartheta_h\|_{H^{-\frac{1}{2}}(\Gamma)} \\ &\qquad \leq \gamma\left(\inf_{V_h\in\mathcal{V}}\|U-V_h\|_{H^1(\Omega^-)} + \inf_{\varphi_h\in\Theta_h}\|\vartheta-\varphi_h\|_{H^{-\frac{1}{2}}(\Gamma)}\right),\end{aligned}$$

where the constant $\gamma > 0$ does not depend on the discrete trial spaces.

The standard choices for $\mathcal{V}_h$, Θ_h, and $\mathcal{Q}_h$ are based on a tetrahedral or quadrilateral mesh $\mathcal{M}$ of Ω^-, which yields a mesh $\mathcal{M}_\Gamma$ of Γ by plain restriction to Γ. Then we may pick

$$\begin{aligned}(50)\quad V_h &:= \{V \in C^0(\Omega^-) :\ V_{|K} \in \mathcal{P}_k(K)\ \forall K \in \mathcal{M}\},\\ \Theta_h &:= \{\varphi \in L^2(\Gamma) :\ \varphi_{|K} \in \mathcal{P}_{k-1}(K)\ \forall K \in \mathcal{M}_\Gamma\},\\ \mathcal{Q}_h &:= \{q \in C^0(\Gamma) :\ q_{|K} \in \mathcal{P}_k(K)\ \forall K \in \mathcal{M}_\Gamma\}.\end{aligned}$$

Here, $\mathcal{P}_k(K)$ stands for the space of polynomials of degree $\leq k$ on the cell K. This refers to the total degree in the case of tetrahedra and the degree in each variable in the case of hexahedra.

Then, the usual best approximation estimates [32] for the h-version of finite elements and boundary elements give us

$$\begin{aligned}\inf_{V_h\in\mathcal{V}_h}\|U-V_h\|_{H^1(\Omega^-)} &\leq \gamma h^{\min\{s-1,k\}}\|U\|_{H^s(\Omega^-)},\\ \inf_{\varphi_h\in\Theta_h}\|\vartheta-\varphi_h\|_{H^{-\frac{1}{2}}(\Gamma)} &\leq \gamma h^{\min\{s+1/2,k\}}\|\vartheta\|_{H^s(\Gamma)},\end{aligned}$$

with constants depending on the shape regularity of $\mathcal{M}$ and $h > 0$ denoting the meshwidth of $\mathcal{M}$.

Remark 10.1. Why do we have to approximate the dummy variable p at all, since it vanishes and apparently the choice of $\mathcal{Q}_h$ does not affect the convergence of Galerkin solutions? The reason is that (49) is an asymptotic statement, whose proof also hinges on sufficiently good approximation properties of $\mathcal{Q}_h$. In the context of the h-version of finite elements and boundary elements, this means that the mesh has to be sufficiently fine to make (49) hold.

11. Numerical experiments. Limited computational resources allow the numerical exploration of asymptotic convergence rates only in two dimensions. Fortunately, the above theoretical developments carry over to two dimensions verbatim, when replacing the kernel G_κ by

$$(51)\qquad G_\kappa(z) := \frac{i}{4}H_0^{(1)}(kz),$$

where $H_0^{(1)}$ is the Hankel function of the first kind of order zero.

For the numerical experiments we considered the following:

- The unit circle $\Omega_\circ^- := \{\mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| < 1\}$ as a specimen of a domain with smooth boundary. The two smallest interior resonant frequencies are $\kappa_1 = 5.5201$ and $\kappa_2 = 11.7915$, which correspond to the second and fourth zero of the Bessel function $J_0(x)$.
- The unit square $\Omega_\square^- := \{\mathbf{x} \in \mathbb{R}^2 : -1/2 < x_1, x_2 < 1/2\}$, as representative of polygonal domains. The associated two lowest resonant frequencies are $\kappa_3 = 2\pi/\sqrt{2}$ and $\kappa_4 = 5\pi/\sqrt{2}$.

On each domain finite element meshes $\mathcal{M}_l$, $l \in \mathbb{N}$, consisting of quadrilaterals with straight edges were used. In the case of $\Omega_\circ^-$ the triangulation $\mathcal{M}_l$ is created by inscribing $\Omega_\circ^-$ a regular 2^{l+3}-gon and a centered unit square. The portions of the line segments from the center to the corners of the polygon are split into 2^l equal parts, whose endpoints are connected to form a quadrilateral mesh outside the unit square. This is extended by an orthogonal tensor product mesh inside the unit square. The mesh $\mathcal{M}_1$ is drawn in Figure 1. The family of meshes arising from this construction will be quasi-uniform and shape-regular with meshwidth h of $\mathcal{M}_l$ being proportional to $2^{-(l+1)}$.

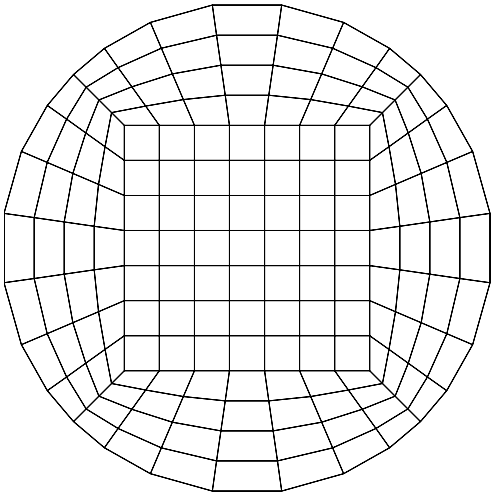


FIG. 1. *Quadrilateral mesh of the unit circle.*

On $\Omega_\square^-$ the mesh $\mathcal{M}_l$ is a plain uniform orthogonal tensor product grid with meshwidth $h = 2^{-(l+1)}$.

We used mapped bilinear Lagrangian finite elements to build $\mathcal{V}_h$, piecewise constants on $\mathcal{M}_\Gamma$ for Θ_h, and linear surface elements for $\mathcal{Q}_h$, that is, the case $k = 1$ of (50). The finite element stiffness matrix was assembled using a four-point Gaussian quadrature rule on the reference element. The dense matrices of the discrete boundary integral operators were computed using Duffy's trick and highly accurate adaptive composite Gauss–Legendre quadrature as proposed in [32, Example 5.1.9] and [34]. All computations were done in MATLAB and a direct solver was used whenever we aimed to study discretization errors.

In all the experiments we used $n(\mathbf{x}) = 1$ in Ω^- and excitation by incident plane waves. These will also provide the exact solutions. Please note that in this setting $\vartheta = -\gamma_1^+ U$, because there is no scattered field. As far as the stable regularized

coupled schemes are concerned we consistently used the second regularized variational formulation (47) together with the simple right-hand sides (48).

When analytic solutions are known, we measure the discretization error in the interior total field in either the $H^1(\Omega^-)$ or the $L^2(\Omega^-)$-norm and the error in ϑ in either the $H^{-\frac{1}{2}}(\Gamma)$ or the $L^2(\Gamma)$-norm. Integer Sobolev norms are calculated by means of four-point Gaussian quadrature. The $H^{-\frac{1}{2}}$-norm is evaluated by means of the discrete single layer potential operator on the current mesh after the exact solution for ϑ has been projected onto Θ_h.

Experiment 1. A plane incident wave $U^{\mathrm{i}}(\mathbf{x}) = \exp(-i\kappa\mathbf{d}\cdot\mathbf{x})$, $|\mathbf{d}| = 1$, is used, where the incident angle between the propagation direction $\mathbf{d}$ and the x-axis is $\pi/4$. We measure the discretization errors in different norms on the domain $\Omega^-_{\square}$ for the two frequencies κ_3 and κ_4 on the series of shape-regular meshes and for a regularization parameter $\eta = 1$; see Figures 2 and 3 for results. Table 1 lists the observed convergence rates, which are very low because of the corner discontinuity of ϑ.

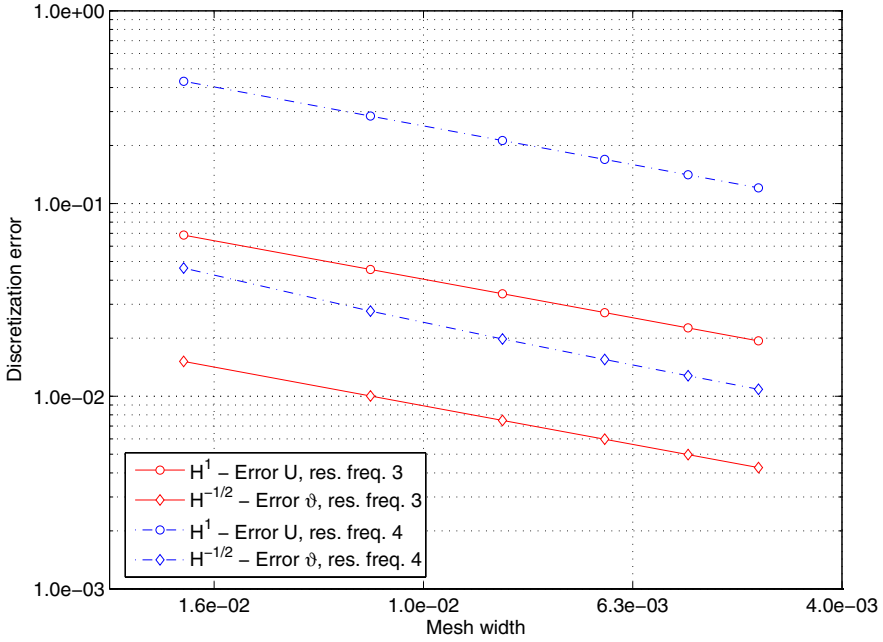


FIG. 2. *Discretization errors for* κ_3 $(-)$ *and* κ_4 $(-\cdot)$ *on the unit square* $\Omega^-_{\square}$.

TABLE 1
Observed convergence rates for different error norms.

	$\|U-U_h\|_{H^1(\Omega^-)}$	$\|U-U_h\|_{L^2(\Omega^-)}$	$\|\vartheta-\vartheta_h\|_{H^{-\frac{1}{2}}(\Gamma)}$	$\|\vartheta-\vartheta_h\|_{L^2(\Gamma)}$
Exp. 1	$O(h)$	$O(h)$	$O(h)$	$O(h^{\frac{1}{2}})$
Exp. 2	$O(h)$	$O(h^2)$	$O(h^2)$	$O(h)$

Experiment 2. Using the same excitation as before, we measure the discretization errors on the series of shape-regular meshes of the unit circle $\Omega^-_{\circ}$ for the two frequencies κ_1 and κ_2; see Figures 4 and 5. Now both ϑ and U are smooth, which translates into optimal convergence rates; see Table 1.

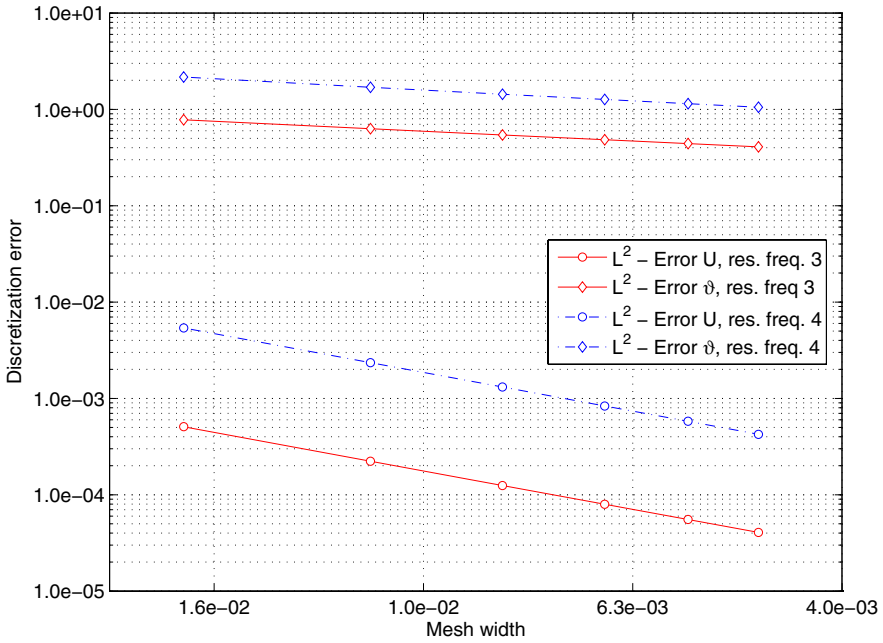


FIG. 3. *Discretization errors for* κ_3 $(-)$ *and* κ_4 $(-\cdot)$ *on the unit square* $\Omega_{\square}^{-}$.

Experiment 3. We examine the dependence of the discretization error, measured in the $H^1(\Omega^-)$, and $H^{-\frac{1}{2}}(\Gamma)$-norms, respectively, on the wave number for a mesh of the domain $\Omega_{\square}^{-}$ with 14161 elements. The results for conventional symmetric FEM-BEM coupling (15) are recorded in Figure 6. In Figure 7 the discretization errors are plotted for the second version of regularized FEM-BEM coupling (47). We note that the discretization errors for both methods are of exactly the same size. Moreover, they grow as κ increases. This is hardly surprising, because this is already observed for low order finite element discretizations of the Helmholtz equation [4].

The pronounced spikes in the discretization error graph for the ϑ-component in Figure 7 are due to the resonant frequencies which affect the conventional symmetric FEM-BEM coupling. In contrast, they are completely suppressed when using the second version of regularized FEM-BEM coupling (47) as we can see in Figure 6.

Experiment 4. We recorded the dependence of the spectral condition number of the entire system matrix on the wave number for

1. the symmetric FEM-BEM coupling (15) and
2. the second version of regularized FEM-BEM coupling (47)

in the neighborhood of the resonant frequency κ_3 for a mesh of the domain $\Omega_{\square}^{-}$ with 14161 elements; see Figure 8. In each case the extremal eigenvalues were computed by means of direct and inverse power iterations. Obviously, regularization manages to suppress the pronounced peak in the condition number in the case of the symmtrically coupled problem.

REFERENCES

[1] F. ALOUGES, S. BOREL, AND D. LEVADOUX, *A new well-conditioned integral formulation for*

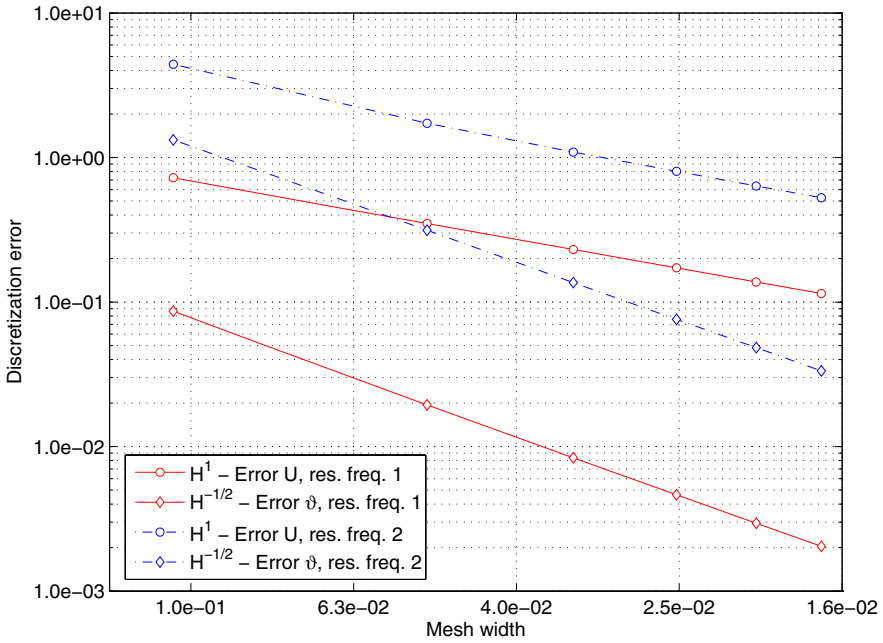


FIG. 4. *Discretization errors for* κ_1 $(-)$ *and* κ_2 $(-\cdot)$ *on the unit circle* $\Omega_\circ^-$.

Maxwell equations in three-dimensions, IEEE Trans. on Antennas and Propagation, 53 (2005), pp. 2995–3004.

[2] X. ANTOINE, A. BENDALI, AND M. DARBAS, *Analytic preconditioners for the electric field integral equation*, Internat. J. Numer. Methods Engrg., 61 (2004), pp. 1310–1331.

[3] R. ASTLEY, *Infinite elements for wave problems: A review of current formulations and an assessment of accuracy*, Int. J. Numer. Meth. Engr., 49 (2000), pp. 951–976.

[4] I. BABUŠKA AND S. SAUTER, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Review, 42 (2000), pp. 451–484.

[5] P. BETTESS, *Infinite elements*, Internat. J. Numer. Methods Engrg., 11 (1977), pp. 53–64.

[6] H. BRAKHAGE AND P. WERNER, *Ueber das Dirichletsche Außenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. der Math., 16 (1965), pp. 325–329.

[7] A. BUFFA AND R. HIPTMAIR, *A coercive combined field integral equation for electromagnetic scattering*, SIAM J. Numer. Anal., 42 (2004), pp. 621–640.

[8] A. BUFFA AND R. HIPTMAIR, *Regularized combined field integral equations*, Numer. Math., 100 (2005), pp. 1–19.

[9] A. BUFFA AND S. SAUTER, *Stabilisation of the acoustic single layer potential on nonsmooth domains*, preprint 19-2003, Institut für Mathematik, Universität Zürich, Zürich, Switzerland, 2003, to appear in SIAM J. Sci. Comput., 2006.

[10] A. BURTON AND G. MILLER, *The application of integral methods for the numerical solution of boundary value problems*, Proc. Roy. Soc. London, Ser. A, 232 (1971), pp. 201–210.

[11] C. CARSTENSEN AND P. WRIGGERS, *On the symmetric boundary element method and the symmetric coupling of boundary elements and finite elements*, IMA J. Numer. Anal., 17 (1997), pp. 201–238.

[12] G. CHEN AND J. ZHOU, *Boundary Element Methods*, Academic Press, New York, 1992.

[13] S. CHRISTIANSEN, *Discrete Fredholm properties and convergence estimates for the electric field integral equation*, Math. Comp., 73 (2004), pp. 143–167.

[14] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Applied Mathematical Sciences 93, Springer, Heidelberg, 1998.

[15] M. COSTABEL, *Symmetric methods for the coupling of finite elements and boundary elements*, in Boundary Elements IX, C. Brebbia, W. Wendland, and G. Kuhn, eds., Springer-Verlag, Berlin, 1987, pp. 411–420.

[16] M. COSTABEL, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM

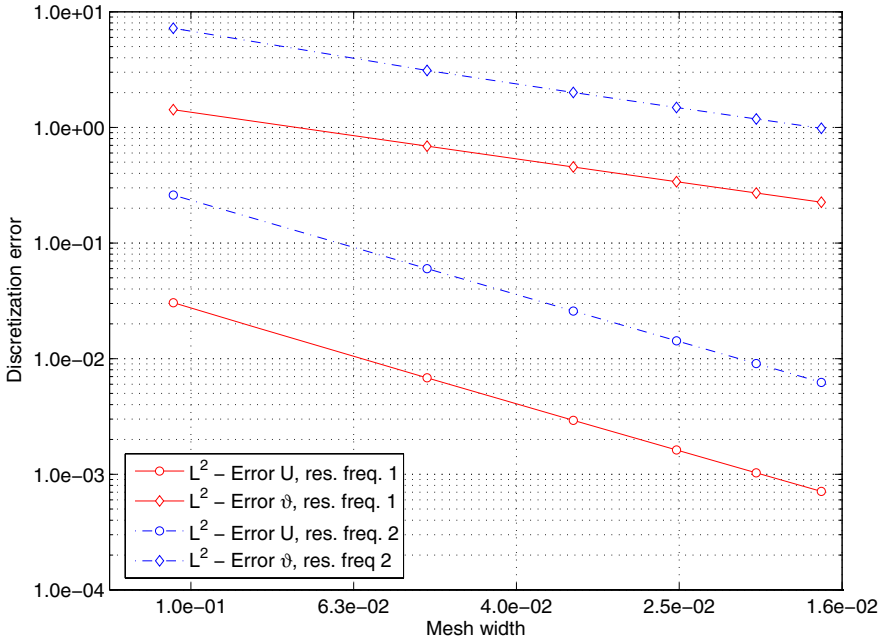


FIG. 5. *Discretization errors for* κ_1 $(-)$ *and* κ_2 $(-\cdot)$ *on the unit circle* $\Omega_\circ^-$.

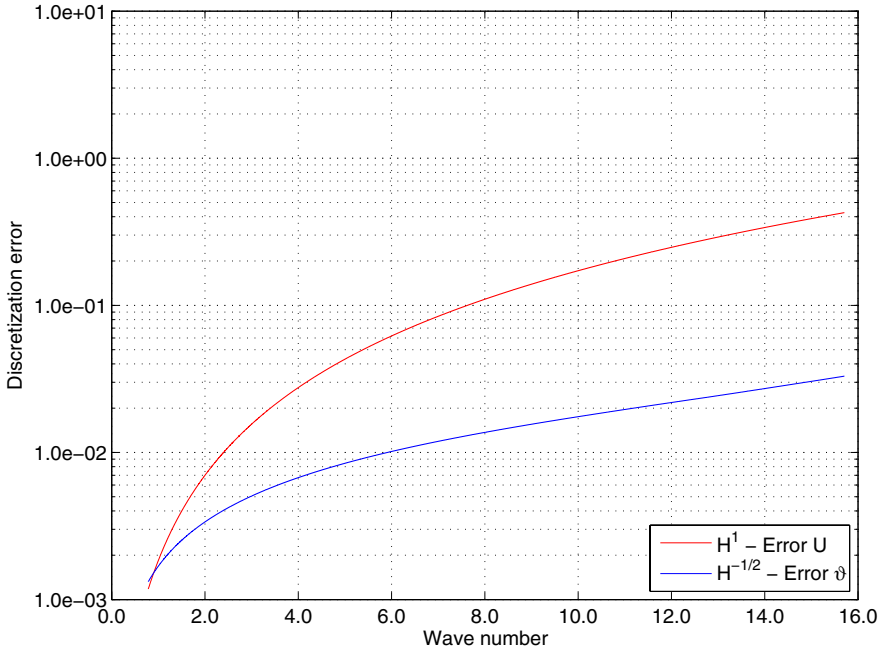


FIG. 6. *Discretization errors on the unit square* $\Omega_\square^-$.

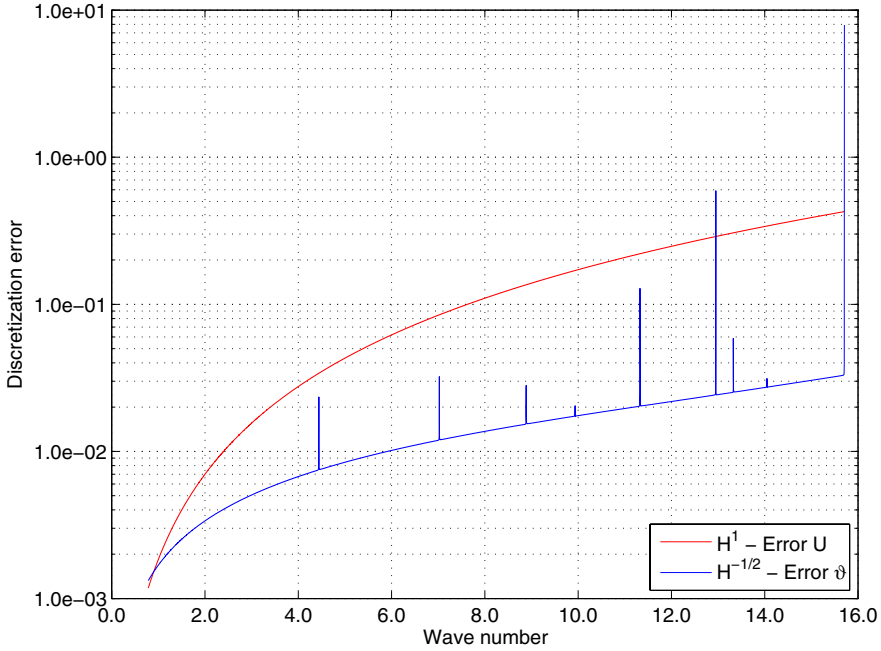


FIG. 7. *Discretization errors on the unit square* $\Omega_{\square}^{-}$.

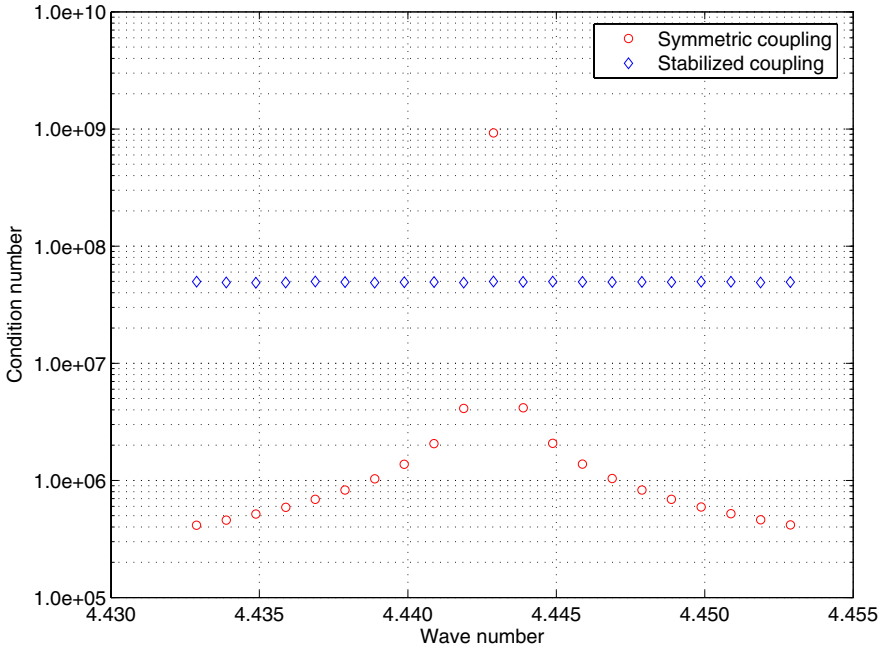


FIG. 8. *Spectral condition numbers on the unit square* $\Omega_{\square}^{-}$ *close to the resonant frequency* κ_3. *Condition numbers of the matrix underlying the classical variational formulation are labelled with* $\circ$, *whereas condition numbers related to the matrix underlying the regularized variational formulation are labelled with* $\diamond$.

J. Math. Anal., 19 (1988), pp. 613–626.
[17] M. Costabel and M. Dauge, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.
[18] M. Costabel and W. Wendland, *Strong ellipticity of boundary integral operators*, J. Reine Angew. Math., 372 (1986), pp. 39–63.
[19] L. Demkowicz, *Asymptotic convergence in finite and boundary element methods: Part 1, Theoretical results*, Comput. Math. Appl., 27 (1994), pp. 69–84.
[20] G. Fairweather, A. Karageorghis, and P. Martin, *The method of fundamental solutions for scattering and radiation problems*, Eng. Anal. Bound. Elem., 27 (2003), pp. 759–769.
[21] W. Hackbusch, *Integral Equations. Theory and Numerical Treatment*, International Series of Numerical Mathematics 120, Birkhäuser, Basel, 1995.
[22] T. Hagstrom, *Radiation boundary conditions for the numerical simulation of waves*, Acta Numer., 8 (1998), pp. 47–106.
[23] R. Hiptmair, *Symmetric coupling for eddy current problems*, SIAM J. Numer. Anal., 40 (2002), pp. 41–65.
[24] R. Hiptmair, *Coupling of finite elements and boundary elements in electromagnetic scattering*, SIAM J. Numer. Anal., 41 (2003), pp. 919–944.
[25] C. Johnson and J. Nédélec, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980), pp. 1063–1079.
[26] D. Jones, *Integral equations for the exterior acoustic problem*, Quart. J. Mech. Appl. Math., 27 (1974), pp. 129–142.
[27] M. Kuhn and O. Steinbach, *FEM-BEM coupling for 3d exterior magnetic field problems*, Math. Methods Appl. Sci., 25 (2002), pp. 357–371.
[28] R. Leis, *Zur Dirichletschen Randwertaufgabe des Aussenraumes der Schwingungsgleichung*, Math. Z., 90 (1965), pp. 205–211.
[29] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[30] O. Panich, *On the question of the solvability of the exterior boundary-value problems for the wave equation and maxwell's equations*, Usp. Mat. Nauk., 20A (1965), pp. 221–226. In Russian.
[31] A. Peterson, *The "interior resonance" problem associated with surface integral equations of electromagnetics: numerical consequences and a survey of remedies*, Electromagnetics, 10 (1990), pp. 293–312.
[32] S. Sauter and C. Schwab, *Randelementmethoden*, BG Teubner, Stuttgart, 2004.
[33] A. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[34] C. Schwab, *Variable order composite quadrature of singular and nearly singular integrals*, Computing, 53 (1994), pp. 173–194.
[35] F. Ursell, *On the exterior problem of acoustics*, Math. Proc. Cambridge Philos. Soc., 74 (1973), pp. 117–125.
[36] T. von Petersdorff, *Boundary integral equations for mixed Dirichlet, Neumann and transmission problems*, Math. Methods Appl. Sci., 11 (1989), pp. 185–213.
[37] W. Wendland, *Boundary element methods for elliptic problems*, in Mathematical Theory of Finite and Boundary Element Methods, A. Schatz, V. Thomée, and W. Wendland, eds., vol. 15 of DMV-Seminar, Birkhäuser, Basel, 1990, pp. 219–276.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2131–2158

OPTIMAL DISCONTINUOUS GALERKIN METHODS FOR WAVE PROPAGATION*

ERIC T. CHUNG† AND BJÖRN ENGQUIST‡

Abstract. We have developed and analyzed a new class of discontinuous Galerkin methods (DG) which can be seen as a compromise between standard DG and the finite element (FE) method in the way that it is explicit like standard DG and energy conserving like FE. In the literature there are many methods that achieve some of the goals of explicit time marching, unstructured grid, energy conservation, and optimal higher order accuracy, but as far as we know only our new algorithms satisfy all the conditions. We propose a new stability requirement for our DG. The stability analysis is based on the careful selection of the two FE spaces which verify the new stability condition. The convergence rate is optimal with respect to the order of the polynomials in the FE spaces. Moreover, the convergence is described by a series of numerical experiments.

Key words. discontinuous Galerkin, wave propagation, optimal rate of convergence

AMS subject classifications. 65M12, 65M15, 65M60, 78M10

DOI. 10.1137/050641193

1. Introduction.

Many applications involve the solution of wave equations. Examples are electromagnetic waves for radar and communication as well as acoustic and seismic wave propagation. Let $\Omega \subset \mathbb{R}^2$ be a two dimensional polygonal domain with outward normal vector n and let $T > 0$ be a fixed time. Given two positive constants $a_1 > 0$, $a_2 > 0$ and two given functions $F_1(x,t)$, $F_2(x,t)$, we will consider, for $(x,t) \in \Omega \times (0,T)$, the following wave propagation problem: find a function $u(x,t)$ and a vector field $v(x,t) \in \mathbb{R}^2$ such that

$$a_1 \frac{\partial u}{\partial t} + Bv = F_1, \tag{1.1}$$

$$a_2 \frac{\partial v}{\partial t} - B^* u = F_2, \tag{1.2}$$

where the two operators B and B^* satisfy

$$\int_{\Omega_0} (B^*\phi)\psi \, dx - \int_{\Omega_0} (B\psi)\phi \, dx = \int_{\partial\Omega_0} (L\psi)\phi \, d\sigma \tag{1.3}$$

for all subset $\Omega_0 \subset \Omega$. Here L is some operator depending on the two operators B, B^* and the subdomain Ω_0. We also denote by $L^\perp$ the operator such that $|\psi|^2 = |L\psi|^2 + |L^\perp\psi|^2$. Assume that there is an operator $B^\perp$ such that $BB^\perp p = 0$ for all p. Acting $(B^\perp)^*$ to (1.2) and using $(B^\perp)^* B^* = 0$, we have the following:

$$\frac{\partial}{\partial t}(a_2 (B^\perp)^* v) = (B^\perp)^* F_2. \tag{1.4}$$

*Received by the editors September 26, 2005; accepted for publication (in revised form) April 6, 2006; published electronically November 3, 2006.
http://www.siam.org/journals/sinum/44-5/64119.html
†Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125 (tschung@acm.caltech.edu).
‡Department of Mathematics, University of Texas at Austin, Austin, TX 78712 (engquist@math.utexas.edu).

The condition (1.4) usually has important physical significance. For example, in the case of electromagnetic wave propagation (see (E) in the following), v represents the electric field, $(B^\perp)^*$ is the divergence operator, and (1.4) is just the continuity equation expressing the conservation of charges. We supplement the system (1.1)–(1.2) with boundary condition

$$Lv = 0 \quad \forall x \in \partial\Omega \tag{1.5}$$

and initial conditions

$$u(x,0) = u_0(x) \quad \text{and} \quad v(x,0) = v_0(x) \quad \forall x \in \Omega, \tag{1.6}$$

where $u_0(x)$ and $v_0(x)$ are given. In particular, we are interested in the acoustic and the electromagnetic wave equations, which correspond to the following choice of B and B^*:

(A) Acoustic: $Bv = -\nabla \cdot v$, $B^*u = \nabla u$, and $Lv = v \cdot n$.

(E) Electromagnetic: $Bv = \nabla \times v$, $B^*u = \nabla \times u$, and $Lv = v \times n$.

Notice that, in (1.1)–(1.2), $u(x,t)$ is a function while $v(x,t) = (v^1(x,t), v^2(x,t))$ is a vector having two components. So the operator $\nabla\times$ is defined as $\nabla \times v = \partial_1 v^2 - \partial_2 v^1$ for any vector field v and $\nabla \times u = (\partial_2 u, -\partial_1 u)$ for any function u. Here $v \times n = v^1 n_2 - v^2 n_1$ with $n = (n_1, n_2)$. In (A) and (E), we use n to denote generically the unit normal vector of the corresponding subdomain which defines L and $L^\perp$. Moreover, we have $L^\perp\psi = \psi \times n$ for (A) while $L^\perp\psi = \psi \cdot n$ for (E). Furthermore, we have $B^\perp p = \nabla \times p$ and $(B^\perp)^* p = \nabla \times p$ for (A) while $B^\perp p = \nabla p$ and $(B^\perp)^* p = -\nabla \cdot p$ for (E). Wave propagation problems can be solved by partial differential equation (PDE) techniques, integral equation techniques, and asymptotic techniques. Among PDE techniques, finite difference (FD) method, finite volume (FV) method, finite element (FE) method, and discontinuous Galerkin (DG) method are the most popular choices. The FD method provides a simple way to solve wave propagation problems, but it is typically low order and applies only to structured grids. The FV method can be seen as a generalization of the FD method to unstructured grids, but it is still low order. The FE and DG methods provide high order solvers for the time dependent wave equations on unstructured grids.

Nédélec [15] introduces a curl-conforming FE method for solving Maxwell's equations. Geveci [8] proposed a mixed FE method for the scalar wave equation. The inversion of the mass matrix at each time step causes some possible drawback in the efficiency of those methods. Mass lumping techniques can be used to avoid solving linear systems. In Cohen and Monk [6], a mass lumping method for rectangular grids is developed. In Bécache, Joly, and Tsogka [1], a new class of mixed FE method, which is suitable for mass lumping, is developed for the scalar wave equation. Cohen, Joly, Torjman, and Roberts [5] design a mass lumping technique for triangular grids for polynomial order up to five. Discontinuous Galerkin methods provide explicit schemes in the sense that only block diagonal mass matrices have to be inverted. Hesthaven and Warburton [10] proposed a DG method based on upwind flux and Cockburn, Li, and Shu [4] proposed a DG method based on locally divergence free basis and upwind flux. While the schemes are successful, energy is not conserved due to the upwinding. Fezoui, Lanteri, Lohrengel, and Piperno [7] proposed a DG method based on central flux. This method preserves energy, but the convergence rate of the scheme is suboptimal. Recently, a new DG method has been developed for the wave equation in second order form; see Grote, Schneebeli, and Schötzau [9]. The method is also energy conserving

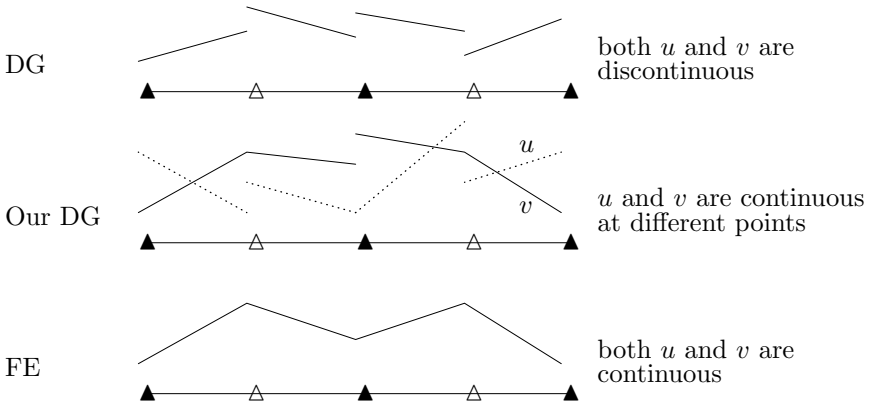


FIG. 1.1. *Comparison among standard DG, our new DG, and FE methods.*

in the sense of a newly defined energy. A space-time DG method has also been developed in Monk and Richter [14].

In this paper, we will develop and analyze a new class of DG methods which can be seen as a compromize between FE and DG methods. Our new DG method combines the advantages of FE and DG methods in the sense that it is both energy conserving and explicit. The idea is to use discontinuous functions with extra continuity. In the velocity-potential formulation of the scalar wave equation (A), we will add extra continuity to the velocity where the potential is discontinuous and add extra continuity to the potential where the velocity is discontinuous. For Maxwell's equations (E), a similar idea can be applied to the electric and magnetic fields. As a result, the flux integrals are evaluated exactly, which is the basis of energy conservation. However, the addition of the extra continuity cannot be done arbitrarily due to stability concerns. It has to be done in such a way that some inf-sup conditions are satisfied. In Figure 1.1, we illustrate this idea in one space dimension. For standard DG, both unknown functions u and v, which are velocity and potential for scalar wave equation and are electric and magnetic fields for Maxwell's equations, are discontinuous at cell boundaries. For FE methods, both u and v are continuous. For our new DG, the two functions are continuous at different points.

Yee's scheme [16] has been a very popular numerical method for computational electromagnetics. It is a second order central FD method on structured grids. The success of the scheme is due to the use of a staggered grid. Our new DG method is a FE method on staggered grids and can be seen as a higher generalization of Yee's scheme on unstructured grids. In particular, in one space dimension, our new DG method with piecewise constant approximation is the same as Yee's scheme. In two space dimensions, our new DG method in the lowest order is some averaged version of Yee's scheme.

The rest of the paper is organized as follows. In section 2, we will introduce the new FE spaces and prove the corresponding unisolvence and interpolation error estimates. The new DG is then derived in section 3. In section 4, under the assumption of some inf-sup conditions, the stability and convergence of the method are proved. The inf-sup conditions are then verified in section 5. Furthermore, some numerical experiments are presented in section 6. The paper ends with a conclusion.

Remark. We consider only two space dimensions in this paper. For three space dimensions, a careful choice of the two FE spaces U_h and V_h that verify (3.1) and

(3.2) as well as the two inf-sup conditions (4.1)–(4.2) are required. This work will be developed in a forthcoming paper.

2. FE spaces. Assume the domain Ω is triangulated by a family of triangles $\mathcal{T}$ so that $\Omega = \cup\{\tau \mid \tau \in \mathcal{T}\}$. Let $\tau \in \mathcal{T}$. We define h_τ as the diameter of τ and ρ_τ as the supremum of the diameters of the circles inscribed in τ. The mesh size h is defined as $h = \max_{\tau\in\mathcal{T}} h_\tau$. We will assume the set of triangles $\mathcal{T}$ forms a regular family of triangulation of Ω so that there exist a uniform constant K independent of the mesh size such that [3]

$$h_\tau \leq K\rho_\tau \quad \forall \tau \in \mathcal{T}.$$

In addition, we will assume the triangulation satisfies the inverse assumption [3].

Let $\mathcal{E}$ be the set of all edges and let $\mathcal{E}^0 \subset \mathcal{E}$ be the set of all interior edges of the triangles in $\mathcal{T}$. The length of $\sigma \in \mathcal{E}$ will be denoted by h_σ. We also denote by $\mathcal{N}$ the set of all interior nodes of the triangles in $\mathcal{T}$. Here, by interior edge and interior node, we mean any edge and node that does not lie on the boundary $\partial\Omega$. Let $\nu \in \mathcal{N}$. We define

$$\mathcal{S}(\nu) = \cup\{\tau \in \mathcal{T} \mid \nu \in \tau\}. \tag{2.1}$$

That is, $\mathcal{S}(\nu)$ is the union of all triangles having vertex ν. We will assume the triangulation of Ω satisfies the following condition.

Assumption on triangulation: There exists a subset $\mathcal{N}_1 \subset \mathcal{N}$ such that

(A1) $\Omega = \cup\{\mathcal{S}(\nu) \mid \nu \in \mathcal{N}_1\}$.

(A2) $\mathcal{S}(\nu_i) \cap \mathcal{S}(\nu_j) \in \mathcal{E}^0$ for all distinct $\nu_i, \nu_j \in \mathcal{N}_1$.

Let $\nu \in \mathcal{N}_1$. We define

$$\mathcal{E}_u(\nu) = \{\sigma \in \mathcal{E} \mid \nu \in \sigma\}. \tag{2.2}$$

That is, $\mathcal{E}_u(\nu)$ is the set of all edges that have ν as one of their endpoints. We further define

$$\mathcal{E}_u = \cup\{\mathcal{E}_u(\nu) \mid \nu \in \mathcal{N}_1\} \quad \text{and} \quad \mathcal{E}_v = \mathcal{E}\backslash\mathcal{E}_u. \tag{2.3}$$

Notice that $\mathcal{E}_u$ contains only interior edges since one of the endpoints of edges in $\mathcal{E}_u$ has a vertex from $\mathcal{N}_1$. On the other hand, $\mathcal{E}_v$ has both interior and boundary edges. So, we also define $\mathcal{E}_v^0 = \mathcal{E}_v \cap \mathcal{E}^0$ which contains elements from $\mathcal{E}_v$ that are interior edges. Notice that we have $\mathcal{E}_v\backslash\mathcal{E}_v^0 = \mathcal{E} \cap \partial\Omega$. Furthermore, for $\sigma \in \mathcal{E}_v^0$, we will let $\mathcal{R}(\sigma)$ be the union of the two triangles sharing the same edge σ. For $\sigma \in \mathcal{E}_v\backslash\mathcal{E}_v^0$, we will let $\mathcal{R}(\sigma)$ be the only triangle having the edge σ.

In practice, triangulations that satisfy assumptions (A1)–(A2) are not difficult to construct. In Figure 2.1, we illustrate how this kind of triangulation is generated. First, the domain Ω is triangulated by a family of triangles, called $\tilde{\mathcal{T}}$. Each triangle in this family is then subdivided into three subtriangles by connecting a point inside the triangle with its three vertices. Then we define the union of all these subtriangles to be our triangulation $\mathcal{T}$. Each triangle in $\tilde{\mathcal{T}}$ corresponds to an $\mathcal{S}(\nu)$ for some ν inside the triangle. In Figure 2.1, we show two of the triangles, enclosed by solid lines, in this family $\tilde{\mathcal{T}}$. This corresponds to 6 triangles in the triangulation $\mathcal{T}$. The dotted lines represent edges in the set $\mathcal{E}_u$ while solid lines represent edges in the set $\mathcal{E}_v$.

LEMMA 2.1. *Each $\tau \in \mathcal{T}$ has exactly two edges that belong to $\mathcal{E}_u$.*

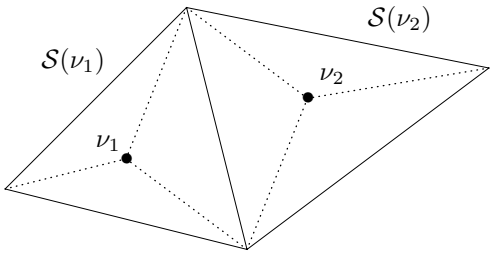


FIG. 2.1. *Triangulation.*

Proof. First of all, τ has at least one interior vertex. We will show that there is exactly one vertex of τ that belongs to $\mathcal{N}_1$. If none of the three vertices of τ belong to $\mathcal{N}_1$, then $\tau^0 \cap \mathcal{S}(\nu)$ is an empty set for all $\nu \in \mathcal{N}_1$, where τ^0 is the interior of τ. Then, $\cup\{\mathcal{S}(\nu) \mid \nu \in \mathcal{N}_1\} \cap \tau^0$ is an empty set. So, $\cup\{\mathcal{S}(\nu) \mid \nu \in \mathcal{N}_1\} \neq \Omega$, which violates assumption (A1). If τ has two vertices, ν_i and ν_j, that belong to $\mathcal{N}_1$, then $\mathcal{S}(\nu_i) \cap \mathcal{S}(\nu_j)$ contains τ. So, it violates assumption (A2). The case that τ has all vertices belonging to $\mathcal{N}_1$ can be discussed in the same way. In conclusion, τ has exactly one vertex which belongs to $\mathcal{N}_1$. So, by the definition of $\mathcal{E}_u$, the two edges having the vertex in $\mathcal{N}_1$ belong to $\mathcal{E}_u$. ∎

Given $\tau \in \mathcal{T}$, we will denote by $\nu(\tau)_1$, $\nu(\tau)_2$, and $\nu(\tau)_3$ the three vertices of τ. Moreover, $\nu(\tau)_1$ is the vertex that is one of the endpoints of the two edges of τ that belong to $\mathcal{E}_u$. Then $\nu(\tau)_2$ and $\nu(\tau)_3$ are named in a counterclockwise direction. In addition, $\lambda_{\tau,1}(x)$, $\lambda_{\tau,2}(x)$, and $\lambda_{\tau,3}(x)$ are the barycentric coordinates on τ with respect to the three vertices $\nu(\tau)_1$, $\nu(\tau)_2$, and $\nu(\tau)_3$.

Now, we will discuss the FE spaces. Let $k \geq 0$ be a nonnegative integer. Let $\tau \in \mathcal{T}$. We define $P^k(\tau)$ as the space of polynomials of degree less than or equal to k on τ. We also define

$$R^k(\tau) = P^k(\tau) \oplus \tilde{P}^{k+1}(\tau), \tag{2.4}$$

where $\tilde{P}^{k+1}(\tau)$ is the space of homogeneous polynomials of degree $k+1$ on τ in the two variables $\lambda_{\tau,2}$ and $\lambda_{\tau,3}$ such that the sum of the coefficients of $\lambda_{\tau,2}^{k+1}$ and $\lambda_{\tau,3}^{k+1}$ is equal to zero. That is, any function in $\tilde{P}^{k+1}(\tau)$ can be written as $\sum_{i+j=k+1, i\geq 0, j\geq 0} a_{i,j}\lambda_{\tau,2}^i\lambda_{\tau,3}^j$ such that $a_{k+1,0} + a_{0,k+1} = 0$. Now, we define

$$U_h = \{\phi \mid \phi|_\tau \in R^k(\tau); \phi \text{ is continuous at the } k+1 \text{ Gaussian points of } \sigma \ \forall \sigma \in \mathcal{E}_u\}.$$

For any edge σ, we use $P^k(\sigma)$ to represent the space of one dimensional polynomials of degree less than or equal to k on σ. We define the following degrees of freedom:

(UD1) For each edge $\sigma \in \mathcal{E}_u$, we have

$$\int_\sigma \phi p_k \, d\sigma$$

for all $p_k \in P^k(\sigma)$.

(UD2) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau \phi p_{k-1} \, dx$$

for all $p_{k-1} \in P^{k-1}(\tau)$ (for $k \geq 1$).

Notice that (UD1) is equivalent to $\phi(\alpha_i)$ where α_i, for $i = 1, 2, \dots, k+1$, are the $k+1$ Gaussian points of σ. For a smooth function ϕ, we will define $\mathcal{I}_u\phi \in U_h$ by the following degrees of freedom:

(U1) For each edge $\sigma \in \mathcal{E}_u$, we have

$$\int_\sigma (\mathcal{I}_u\phi - \phi)p_k = 0$$

for all $p_k \in P^k(\sigma)$.

(U2) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau (\mathcal{I}_u\phi - \phi)p_{k-1}\, dx = 0$$

for all $p_{k-1} \in P^{k-1}(\tau)$ (for $k \geq 1$).

THEOREM 2.2. *Let $\mathcal{I}_u : \Pi_{\nu\in\mathcal{N}_1} W^{k+1,p}(\mathcal{S}(\nu)) \to U_h$. Then $\mathcal{I}_u$ is uniquely determined by* (U1)–(U2)*. Moreover,*

$$|\phi - \mathcal{I}_u\phi|_{W^{m,p}(\mathcal{S}(\nu))} \leq Kh^{k+1-m}|\phi|_{W^{k+1,p}(\mathcal{S}(\nu))}. \tag{2.5}$$

Proof. Notice that $\dim(P^k) = \frac{1}{2}(k+1)(k+2)$. Then (UD1) gives $(k+1)|\mathcal{E}_u|$ conditions while (UD2) gives $\frac{1}{2}k(k+1)|\mathcal{T}|$ conditions where $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$. Notice that $|\mathcal{S}(\nu)| = |\mathcal{E}_u(\nu)|$ for all $\nu \in \mathcal{N}_1$. So, by the assumption (A1)–(A2) and the definition of $\mathcal{E}_u$, we have $|\mathcal{T}| = \sum_{\nu\in\mathcal{N}_1} |\mathcal{S}(\nu)| = \sum_{\nu\in\mathcal{N}_1} |\mathcal{E}_u(\nu)| = |\mathcal{E}_u|$. So, the total number of degrees of freedom defined by (UD1)–(UD2) is $\frac{1}{2}(k+1)(k+2)|\mathcal{T}|$. Next, we will find the $\dim(U_h)$. Notice that $\dim(\tilde{P}^{k+1}(\tau)) = k+1$. So, we have $\dim(U_h) = \frac{1}{2}(k+1)(k+2)|\mathcal{T}| + (k+1)|\mathcal{T}| - (k+1)|\mathcal{E}_u|$, where the subtraction of the third term is due to the continuity condition imposed on the $k+1$ Gaussian points of each edge in $\mathcal{E}_u$. Since $|\mathcal{T}| = |\mathcal{E}_u|$, we have $\dim(U_h) = \frac{1}{2}(k+1)(k+2)|\mathcal{T}|$, which is equal to the number of degrees of freedom defined by (UD1)–(UD2).

Next, we will show $\mathcal{I}_u\phi = 0$ if $\phi = 0$. Let $\tau \in \mathcal{T}$. Then the degree of freedom (UD1) implies that $\mathcal{I}_u\phi$ is zero at the $k+1$ Gaussian points of the two edges of τ that belong to $\mathcal{E}_u$. More precisely, we denote by α_j $(j = 1, 2, \dots, k+1)$ the $k+1$ Gaussian points of the edge of τ having endpoints $\nu(\tau)_1$ and $\nu(\tau)_2$. Then we define real numbers w_j $(j = 1, 2, \dots, k+1)$ such that $0 < w_1 < w_2 < \cdots < w_{k+1} < 1$ and $\alpha_j = (1-w_j)\nu(\tau)_1 + w_j\nu(\tau)_2$. Moreover, we denote by β_j $(j = 1, 2, \dots, k+1)$ the $k+1$ Gaussian points of the edge of τ having endpoints $\nu(\tau)_1$ and $\nu(\tau)_3$. Then the real numbers w_j also satisfy $\beta_j = (1-w_j)\nu(\tau)_1 + w_j\nu(\tau)_3$. So, we have

$$\mathcal{I}_u\phi = c\Pi_{j=1}^{k+1}(\lambda_{\tau,2} - w_j) + c\Pi_{j=1}^{k+1}(\lambda_{\tau,3} - w_j) - c(-1)^{k+1}\Pi_{j=1}^{k+1}w_j + \lambda_{\tau,2}\lambda_{\tau,3}q_{k-1}$$

for some $q_{k-1} \in P^{k-1}(\tau)$. By the definition of $R^k(\tau)$, the sum of the coefficients of $\lambda_{\tau,2}^{k+1}$ and $\lambda_{\tau,3}^{k+1}$ is zero. So, we have $c = 0$. Using (UD2), we have $\int_\tau \lambda_{\tau,2}\lambda_{\tau,3}q_{k-1}^2\, dx = 0$. Since $\lambda_{\tau,2}\lambda_{\tau,3} > 0$ in the interior of τ, we have $q_{k-1} = 0$. Hence, $\mathcal{I}_u\phi = 0$.

Now, we will prove (2.5). Let $\tau \in \mathcal{T}$ and let $p_k \in P^k(\tau)$. It suffices to show that $\mathcal{I}_u$ preserves polynomials. By (U1), $\mathcal{I}_u p_k - p_k$ is zero at the $k+1$ Gaussian points of the two edges that belong to $\mathcal{E}_u$. So,

$$\begin{aligned}\mathcal{I}_u p_k - p_k = {} & b\Pi_{j=1}^{k+1}(\lambda_{\tau,2} - w_j) + b\Pi_{j=1}^{k+1}(\lambda_{\tau,3} - w_j) \\ & -b(-1)^{k+1}\Pi_{j=1}^{k+1}w_j + \lambda_{\tau,2}\lambda_{\tau,3}r_{k-1}\end{aligned}$$

for some constant d and $r_{k-1} \in P^{k-1}(\tau)$. Since $\mathcal{I}_u p_k \in R^k(\tau)$, we have $b = 0$. Using (U2), we have $r_{k-1} = 0$. Hence, $\mathcal{I}_u p_k = p_k$. ∎

We define

$$V_h = \{\psi \mid \psi|_\tau \in P^k(\tau)^2; L\psi \text{ is continuous along } \sigma \forall \sigma \in \mathcal{E}_v; L\psi = 0 \text{ on } \partial\Omega\}.$$

We also define the following degrees of freedom:

(VD1) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau (L^\perp \psi) p_k \, dx \quad \forall p_k \in P^k(\tau).$$

(VD2) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau (L\psi) \lambda_{\tau,1} p_{k-1} \, dx \quad \forall p_{k-1} \in P^{k-1}(\tau).$$

(VD3) For each edge $\sigma \in \mathcal{E}_v^0$, we have

$$\int_\sigma (L\psi) q_k \, d\sigma \quad \forall q_k \in P^k(\sigma).$$

Furthermore, for a smooth vector field ψ, we define $\mathcal{I}_v \psi$ as the corresponding interpolation operator by the following degrees of freedom:

(V1) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau L^\perp (\mathcal{I}_v \psi - \psi) p_k \, dx = 0 \quad \forall p_k \in P^k(\tau).$$

(V2) For each triangle $\tau \in \mathcal{T}$, we have

$$\int_\tau L(\mathcal{I}_v \psi - \psi) \lambda_{\tau,1} p_{k-1} \, dx = 0 \quad \forall p_{k-1} \in P^{k-1}(\tau).$$

(V3) For each edge $\sigma \in \mathcal{E}_v^0$, we have

$$\int_\sigma L(\mathcal{I}_v \psi - \psi) q_k \, d\sigma = 0 \quad \forall q_k \in P^k(\sigma).$$

THEOREM 2.3. *Let $\mathcal{I}_v : \Pi_{\sigma \in \mathcal{E}_v^0} W^{k+1,p}(\mathcal{R}(\sigma))^2 \to V_h$. Then $\mathcal{I}_v$ is uniquely determined by* (V1), (V2), *and* (V3). *Moreover,*

$$|\psi - \mathcal{I}_v \psi|_{W^{m,p}(\mathcal{R}(\sigma))^2} \le K h^{k+1-m} |\psi|_{W^{k+1,p}(\mathcal{R}(\sigma))^2}. \tag{2.6}$$

Proof. First of all, the number of degrees of freedom defined by (VD1), (VD2), and (VD3) are $\frac{1}{2}(k+1)(k+2)|\mathcal{T}|$, $\frac{1}{2}k(k+1)|\mathcal{T}|$, and $(k+1)|\mathcal{E}_v^0|$, respectively. Also, the dimension of V_h is given by

$$\dim(V_h) = 2\frac{1}{2}(k+1)(k+2)|\mathcal{T}| - (k+1)|\mathcal{E}_v^0| - (k+1)|\mathcal{E}_v \backslash \mathcal{E}_v^0|,$$

where the second term on the right-hand side is due to the continuity condition that $L\psi$ is continuous on σ for all $\sigma \in \mathcal{E}_v^0$, and the third term on the right-hand side is due to the boundary condition $L\psi = 0$ on $\partial\Omega$. Notice that $|\mathcal{T}| = 2|\mathcal{E}_v^0| + |\mathcal{E}_v \backslash \mathcal{E}_v^0|$.

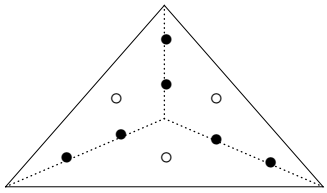


FIG. 2.2. $\mathcal{S}(\nu)$ *for* $k = 1$.

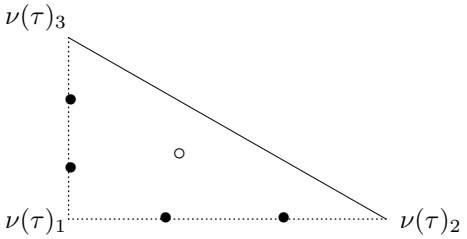


FIG. 2.3. *A single triangle for* $k = 1$.

Now, a direct calculation shows that $\dim(V_h)$ is equal to the total number of degrees of freedom defined by (VD1), (VD2), and (VD3).

Now, we will show $\mathcal{I}_v\psi = 0$ if $\psi = 0$. First, (VD1) implies that $L^\perp(\mathcal{I}_v\psi) = 0$ on each $\tau \in \mathcal{T}$. Using (VD3), we have $L(\mathcal{I}_v\psi) = 0$ on each $\sigma \in \mathcal{E}_v^0$. So, on each $\tau \in \mathcal{T}$, we have $L(\mathcal{I}_v\psi) = \lambda_{\tau,1}q_{k-1}$ for some $q_{k-1} \in P^{k-1}(\tau)$. Indeed, we can write

$$L(\mathcal{I}_v\psi) = a + \sum_{j=1}^{k}(b_j\lambda_{\tau,1}^j + c_j\lambda_{\tau,2}^j).$$

Since $L(\mathcal{I}_v\psi)|_\sigma = 0$, we have $a = 0$, $c_j = 0$, and $L(\mathcal{I}_v\psi) = \sum_{j=1}^{k} b_j\lambda_{\tau,1}^j$. Applying (VD2), we have $L(\mathcal{I}_v\psi) = 0$ on each $\tau \in \mathcal{T}$. Since $|\mathcal{I}_v\psi|^2 = |L^\perp(\mathcal{I}_v\psi)|^2 + |L(\mathcal{I}_v\psi)|^2$, we have $\mathcal{I}_v\psi = 0$.

The estimate (2.6) follows from the fact that the operator $\mathcal{I}_v$ preserves polynomials of degree k. ∎

Let us consider an example for $k = 1$. For U_h, the degrees of freedom are the two Gaussian points on each edge belonging to $\mathcal{E}_u$ and the cell center. In Figure 2.2, we illustrate three triangles in the triangulation which corresponds to an $\mathcal{S}(\nu)$ for some $\nu \in \mathcal{N}_1$. The dotted lines denote edges from the set $\mathcal{E}_u$ while solid lines denote edges from the set $\mathcal{E}_v$. The solid dots denote the continuity points defined by (UD1), which are the two Gaussian points of the edges in $\mathcal{E}_u$. The circle in each triangle represents the degree of freedom defined by (UD2). In Figure 2.3, we show the degrees of freedom on a single triangle. Any function $\phi \in R^k(\tau)$ can be expressed as

$$\phi = a + b\lambda_{\tau,2} + c\lambda_{\tau,3} + d(\lambda_{\tau,2}^2 - \lambda_{\tau,3}^2) + e\lambda_{\tau,2}\lambda_{\tau,3}.$$

For V_h, $L\psi$ is defined as a linear function which is continuous on the edge σ while $L^\perp$ is defined as a linear function on each triangle with no continuity requirement. In Figure 2.4, we illustrate an $\mathcal{R}(\sigma)$ for some $\sigma \in \mathcal{E}_v^0$, where σ is represented by the solid line. We represent the degrees of freedom of $L\psi$ by solid dots and the degrees of freedom of $L^\perp\psi$ by circles.

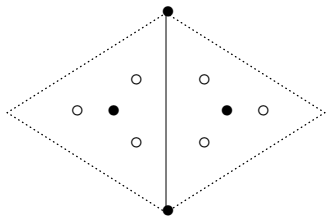


FIG. 2.4. $\mathcal{R}(\sigma)$ *for* $k = 1$.

3. The new scheme. In this section, we will derive the new discontinuous Galerkin method for the wave propagation problem (1.1)–(1.2). Multiplying both sides of (1.1) by ϕ, integrating the resulting equation on $\mathcal{S}(\nu)$, and using (1.3) yields

$$a_1 \int_{\mathcal{S}(\nu)} \frac{\partial u}{\partial t}\phi \, dx + \int_{\mathcal{S}(\nu)} (B^*\phi)v \, dx - \int_{\partial\mathcal{S}(\nu)} (Lv)\phi \, d\sigma = \int_{\mathcal{S}(\nu)} F_1\phi \, dx.$$

Summing over all $\nu \in \mathcal{N}_1$,

$$a_1 \int_{\Omega} \frac{\partial u}{\partial t}\phi \, dx + \int_{\Omega} (B^*\phi)v \, dx - \sum_{\nu\in\mathcal{N}_1} \int_{\partial\mathcal{S}(\nu)} (Lv)\phi \, d\sigma = \int_{\Omega} F_1\phi \, dx.$$

If Lv is continuous along each $\sigma \in \mathcal{E}_v^0$ and Lv is zero along each edge in $\mathcal{E}_v \backslash \mathcal{E}_v^0$, then

$$\sum_{\nu\in\mathcal{N}_1} \int_{\partial\mathcal{S}(\nu)} (Lv)\phi \, d\sigma = \sum_{\sigma\in\mathcal{E}_v^0} \int_{\sigma} (Lv)[\phi] \, d\sigma, \tag{3.1}$$

where $[\phi] = \phi^+ - \phi^-$ is the jump of ϕ along σ. Similarly, multiplying both sides of (1.2) by ψ, integrating the resulting equation on $\mathcal{R}(\sigma)$, and using (1.3) yields

$$a_2 \int_{\mathcal{R}(\sigma)} \frac{\partial v}{\partial t}\psi \, dx - \int_{\mathcal{R}(\sigma)} (B\psi)u \, dx - \int_{\partial\mathcal{R}(\sigma)} (L\psi)u \, d\sigma = \int_{\mathcal{R}(\sigma)} F_2\psi \, dx.$$

Summing for all $\sigma \in \mathcal{E}_v$,

$$a_2 \int_{\Omega} \frac{\partial v}{\partial t}\psi \, dx - \int_{\Omega} (B\psi)u \, dx - \sum_{\sigma\in\mathcal{E}_v} \int_{\partial\mathcal{R}(\sigma)} (L\psi)u \, d\sigma = \int_{\Omega} F_2\psi \, dx.$$

Now if $L\psi$ is a polynomial of degree k and u is a $(k+1)$th degree polynomial which is continuous at the $k+1$ Gaussian points of $\sigma \in \mathcal{E}_u$, then

$$\sum_{\sigma\in\mathcal{E}_v} \int_{\partial\mathcal{R}(\sigma)} (L\psi)u \, d\sigma = \sum_{\sigma\in\mathcal{E}_u} \int_{\sigma} [L\psi]u \, d\sigma, \tag{3.2}$$

where $[L\psi]$ denotes the jump of $L\psi$ along σ. Then, the new discontinuous Galerkin method is defined as follows.

The new discontinuous Galerkin method: Find $u_h \in U_h$ and $v_h \in V_h$ such that

$$a_1 \int_{\Omega} \frac{\partial u_h}{\partial t}\phi \, dx + B_h(v_h, \phi) = \int_{\Omega} F_1\phi \, dx, \tag{3.3}$$

$$a_2 \int_{\Omega} \frac{\partial v_h}{\partial t}\psi \, dx - B_h^*(u_h, \psi) = \int_{\Omega} F_2\psi \, dx \tag{3.4}$$

for all $\phi \in U_h$ and $\psi \in V_h$, where

$$B_h(v_h, \phi) = \int_\Omega (B^*\phi)v_h \, dx - \sum_{\sigma \in \mathcal{E}_v^0} \int_\sigma (Lv_h)[\phi] \, d\sigma, \tag{3.5}$$

$$B_h^*(u_h, \psi) = \int_\Omega (B\psi)u_h \, dx + \sum_{\sigma \in \mathcal{E}_u} \int_\sigma [L\psi]u_h \, d\sigma. \tag{3.6}$$

The initial conditions $u_h(0)$ and $v_h(0)$ will be defined as $u_h(0) = \mathcal{I}_u u_0$ and $v_h(0) = \mathcal{I}_v v_0$. We remark here that we define the spaces U_h and V_h so that (3.1) and (3.2) are valid. Furthermore, we define the discrete derivative operators B_h and B_h^* by

$$\langle B_h\psi, \phi\rangle = B_h(\psi, \phi) \quad \forall \phi \in U_h,$$
$$\langle B_h^*\phi, \psi\rangle = B_h^*(\phi, \psi) \quad \forall \psi \in V_h.$$

The two operators B_h and B_h^* are the discrete analogue of the two derivative operators B and B^*.

LEMMA 3.1. *For all $\phi \in U_h$ and $\psi \in V_h$, we have*

$$B_h(\psi, \phi) = B_h^*(\phi, \psi). \tag{3.7}$$

Proof. Let $\phi \in U_h$ and $\psi \in V_h$. Then, by the definition of B_h and (3.1),

$$\begin{aligned} B_h(\psi, \phi) &= \int_\Omega (B^*\phi)v_h \, dx - \sum_{\sigma \in \mathcal{E}_v^0} \int_\sigma (Lv_h)[\phi] \, d\sigma \\ &= \int_\Omega (B^*\phi)\psi \, dx - \sum_{\nu \in \mathcal{N}_1} \int_{\partial \mathcal{S}(\nu)} (L\psi)\phi \, d\sigma \\ &= \sum_{\nu \in \mathcal{N}_1} \left\{ \int_{\mathcal{S}(\nu)} (B^*\phi)\psi \, dx - \int_{\partial \mathcal{S}(\nu)} (L\psi)\phi \, d\sigma \right\}. \end{aligned}$$

Using integration by parts on each triangle,

$$B_h(\psi, \phi) = \sum_{\nu \in \mathcal{N}_1} \left\{ \int_{\mathcal{S}(\nu)} \phi(B\psi) \, dx + \sum_{\sigma \in \mathcal{E}_u(\nu)} \int_\sigma [L\psi]u \, d\sigma \right\} = B_h^*(\phi, \psi).$$

This completes the proof. ∎

We define the discrete L^2 norms and H^1 norms in the following ways. For all $\phi \in U_h$, we define

$$\|\phi\|_W^2 = \int_\Omega \phi^2 \, dx + \sum_{\sigma \in \mathcal{E}_u} (h_\sigma)^2 \sum_{j=1}^{k+1} \phi(\alpha_j)^2, \tag{3.8}$$

$$\|\phi\|_Z^2 = \int_\Omega (B^*\phi)^2 \, dx + \sum_{\sigma \in \mathcal{E}_v} h_\sigma^{-1} \int_\sigma [\phi]^2 \, d\sigma. \tag{3.9}$$

For all $\psi \in V_h$, we define

$$\|\psi\|_{W'}^2 = \int_\Omega \psi^2 \, dx + \sum_{\sigma \in \mathcal{E}_v} h_\sigma \int_\sigma (L\psi)^2 \, d\sigma, \tag{3.10}$$

$$\|\psi\|_{Z'}^2 = \int_\Omega (B\psi)^2 \, dx + \sum_{\sigma \in \mathcal{E}_u} (h_\sigma)^{-1} \int_\sigma [L\psi]^2 \, d\sigma. \tag{3.11}$$

With these definitions, we have the following continuity conditions for all $\phi \in U_h$ and $\psi \in V_h$:

$$|B_h(\psi,\phi)| \le K\|\psi\|_{W'}\|\phi\|_Z, \tag{3.12}$$

$$|B_h(\psi,\phi)| \le K\|\psi\|_{Z'}\|\phi\|_W. \tag{3.13}$$

Moreover, we have, in U_h, the norm $\|\phi\|_W$ is equivalent to the standard L^2 norm $\|\phi\|$, while in V_h, the norm $\|\psi\|_{W'}$ is equivalent to the standard L^2 norm $\|\psi\|$. That is, there are two uniform constants K_1 and K_2 such that

$$K_1\|\phi\| \le \|\phi\|_W \le K_2\|\phi\| \quad \forall \phi \in U_h, \tag{3.14}$$

$$K_1\|\psi\| \le \|\psi\|_{W'} \le K_2\|\psi\| \quad \forall \psi \in V_h. \tag{3.15}$$

Now, we will prove the following interpolation error estimates. The first one is the interpolation error in the discrete L^2 norms.

LEMMA 3.2. *Assume $(u,v) \in W^{k+1,\infty}(\Omega)^3$. Then for any integer m with $1 \le m \le k+1$,*

$$\|u - \mathcal{I}_u u\|_W \le Kh^m |u|_{W^{m,\infty}(\Omega)}, \quad \|v - \mathcal{I}_v v\|_{W'} \le Kh^m |v|_{W^{m,\infty}(\Omega)^2}. \tag{3.16}$$

Proof. By the definition of W-norm and W'-norm,

$$\|u - \mathcal{I}_u u\|_W \le K\|u - \mathcal{I}_u u\|_{L^\infty(\Omega)}, \quad \|v - \mathcal{I}_v v\|_{W'} \le K\|v - \mathcal{I}_v v\|_{L^\infty(\Omega)^2}.$$

The proof is complete by using (2.5) and (2.6). ∎

The second one is the interpolation error in the discrete H^1 norms.

LEMMA 3.3. *Assume $(u,v) \in H^{k+1}(\Omega)^3$. Then for any integer m with $1 \le m \le k$,*

$$\|u - \mathcal{I}_u u\|_Z \le Kh^m |u|_{H^{m+1}(\Omega)}, \quad \|v - \mathcal{I}_v v\|_{Z'} \le Kh^m |v|_{H^{m+1}(\Omega)^2}. \tag{3.17}$$

Proof. Let $\mathcal{I}_u u \in U_h$ be the interpolant of u. By the definition of Z-norm,

$$\|\mathcal{I}_u u - u\|_Z^2 = \int_\Omega (B^*(\mathcal{I}_u u - u))^2 \, dx + \sum_{\sigma \in \mathcal{E}_v^0} (h_\sigma)^{-1} \int_\sigma [\mathcal{I}_u u - u]^2 \, d\sigma.$$

The first term will be estimated by using the inverse inequality and the interpolation estimate (2.5)

$$\int_\Omega (B^*(\mathcal{I}_u u - u))^2 \, dx \le Kh^{2k} |u|^2_{H^{k+1}(\Omega)}.$$

For the second term, we will use the trace inequality

$$\int_\sigma (\mathcal{I}_u u - u)^2 \, d\sigma \le K(\|\mathcal{I}_u u - u\|_{L^2(\mathcal{R}(\tau))} \|\nabla(\mathcal{I}_u u - u)\|_{L^2(\mathcal{R}(\tau))} + h^{-1}\|\mathcal{I}_u u - u\|^2_{L^2(\mathcal{R}(\tau))}).$$

So,

$$\int_\sigma [\mathcal{I}_u u - u]^2 \, d\sigma \le Kh^{2k+1} |u|^2_{H^{k+1}(\Omega)},$$

and this implies

$$\|\mathcal{I}_u u - u\|_Z \le Kh^k |u|_{H^{k+1}(\Omega)}.$$

The estimate for $\|v - \mathcal{I}_v v\|_{Z'}$ can be proved by a similar argument. ∎

4. Stability and convergence analysis. In this section, we will prove the stability and convergence of the scheme (3.3)–(3.4). We write $\langle u, v\rangle = \int_\Omega uv\, dx$ and $\|u\| = \langle u, u\rangle^{\frac{1}{2}}$. In order to obtain an optimal error estimate, we will assume the following.

The inf-sup conditions:

$$\inf_{\psi\in V_h}\sup_{\phi\in U_h}\frac{B_h(\psi,\phi)}{\|\psi\|_{Z'}\|\phi\|_W}\geq K, \tag{4.1}$$

$$\inf_{\phi\in U_h}\sup_{\psi\in V_h}\frac{B_h^*(\phi,\psi)}{\|\phi\|_Z\|\psi\|_{W'}}\geq K. \tag{4.2}$$

For a general introduction to this topic; see Brezzi and Fortin [2].

Consequently, we have

$$\|B_h\psi\|_W \geq K\|\psi\|_{Z'}, \tag{4.3}$$

$$\|B_h^*\phi\|_{W'} \geq K\|\phi\|_Z. \tag{4.4}$$

By the continuity conditions (3.12) and (3.13), we have

$$K_1\|\psi\|_{Z'} \leq \|B_h\psi\|_W \leq K_2\|\psi\|_{Z'}, \tag{4.5}$$

$$K_1\|\phi\|_Z \leq \|B_h^*\phi\|_{W'} \leq K_2\|\phi\|_Z. \tag{4.6}$$

So, the discrete H^1 norm is equivalent to the discrete L^2 norm of the discrete derivative operator. Notice that, the above two inf-sup conditions (4.3)–(4.4) imply the existence of projection operators $\mathcal{P}_v$ and $\mathcal{P}_u$ such that

$$B_h(\mathcal{P}_v v - v, \phi) = 0 \quad \forall\phi\in U_h, \tag{4.7}$$

$$B_h^*(\mathcal{P}_u u - u, \psi) = 0 \quad \forall\psi\in V_h. \tag{4.8}$$

Regarding the initial condition $u_h(0)$ and $v_h(0)$, we can obtain them by solving the following:

$$B_h(\mathcal{P}_v v_0 - v_0, \phi) = 0 \quad \forall\phi\in U_h, \tag{4.9}$$

$$B_h^*(\mathcal{P}_u u_0 - u_0, \psi) = 0 \quad \forall\psi\in V_h; \tag{4.10}$$

then set $u_h(0) = \mathcal{P}_u u_0$ and $v_h(0) = \mathcal{P}_v v_0$. However, in order to retain the accuracy of the approximation, the initial conditions can also be defined as $u_h(0) = \mathcal{I}_u u_0$ and $v_h(0) = \mathcal{I}_v v_0$, where $\mathcal{I}_u$ and $\mathcal{I}_v$ are some interpolation operators with the same order of accuracy and stability estimates $\|u_h(0)\| \leq K\|u_0\|$ and $\|v_h(0)\| \leq K\|v_0\|$.

One important property of the numerical approximation (3.3)–(3.4) is that energy is conserved, as is the case for the continuous problem (1.1)–(1.2). In particular, the method (3.3)–(3.4) is stable in the discrete L^2 norm. Moreover, the convergence in the L^2 norm is optimal. We state these results in the following theorem.

THEOREM 4.1. *Let $u \in U$ and $v \in V$ be the solution to* (1.1)–(1.2) *and let $u_h \in U_h$ and $v_h \in V_h$ be the solution to the numerical scheme* (3.3)–(3.4). *Then, energy is conserved, namely*

$$\frac{d}{dt}(\|u_h\|^2 + \|v_h\|^2) = 0. \tag{4.11}$$

Moreover, for $0 \le t \le T$, *we have*

$$\begin{aligned}
&\|(u-u_h)(t)\| + \|(v-v_h)(t)\| \\
&\quad \le K\Big\{ \inf_{\phi\in U_h} \|u-\phi\|_W + \inf_{\psi\in V_h} \|v-\psi\|_{W'} \\
&\qquad + \int_0^t \Big(\inf_{\phi\in U_h} \|u_t-\phi\|_W + \inf_{\psi\in V_h} \|v_t-\psi\|_{W'} \Big)\, ds \Big\}.
\end{aligned} \tag{4.12}$$

Proof. Taking $\phi = u_h$ and $\psi = v_h$ in (3.3)–(3.4) yields

$$a_1\left\langle \frac{du_h}{dt}, u_h \right\rangle + B_h(v_h, u_h) = 0,$$
$$a_2\left\langle \frac{dv_h}{dt}, v_h \right\rangle - B_h^*(u_h, v_h) = 0.$$

Adding the two equations and using (3.7), we obtain (4.11).

In the following, we will prove (4.12). Let $\mathcal{I}_u u$ and $\mathcal{I}_v v$ be arbitrary elements in U_h and V_h, respectively. First, we have

$$a_1\left\langle \frac{d(u-u_h)}{dt}, \phi \right\rangle + B_h(v-v_h, \phi) = 0 \quad \forall \phi \in U_h, \tag{4.13}$$

$$a_2\left\langle \frac{d(v-v_h)}{dt}, \psi \right\rangle - B_h^*(u-u_h, \psi) = 0 \quad \forall \psi \in V_h. \tag{4.14}$$

By the definitions of the projection operators $\mathcal{P}_u$ and $\mathcal{P}_v$, we obtain

$$a_1\left\langle \frac{d(u-u_h)}{dt}, \phi \right\rangle + B_h(\mathcal{P}_v v - v_h, \phi) = 0 \quad \forall \phi \in U_h, \tag{4.15}$$

$$a_2\left\langle \frac{d(v-v_h)}{dt}, \psi \right\rangle - B_h^*(\mathcal{P}_u u - u_h, \psi) = 0 \quad \forall \psi \in V_h. \tag{4.16}$$

Let $Q_u^2 : U_h \to \ker(B_h^*)^\perp$ and $Q_v^2 : V_h \to \ker(B_h)^\perp$ be the projection operators. Taking $\phi = Q_u^2(\mathcal{P}_u u - u_h)$ and $\psi = Q_v^2(\mathcal{P}_v v - v_h)$, we have, by adding the two equations,

$$a_1\left\langle \frac{d(u-u_h)}{dt}, Q_u^2(\mathcal{P}_u u - u_h) \right\rangle + a_2\left\langle \frac{d(v-v_h)}{dt}, Q_v^2(\mathcal{P}_v v - v_h) \right\rangle = 0, \tag{4.17}$$

which implies

$$\begin{aligned}
&\frac{d}{dt}(a_1\|Q_u^2(\mathcal{P}_u u - u_h)\|^2 + a_2\|Q_v^2(\mathcal{P}_v v - v_h)\|^2) \\
&\quad = a_1\left\langle \frac{d(\mathcal{P}_u u - u)}{dt}, Q_u^2(\mathcal{P}_u u - u_h) \right\rangle + a_2\left\langle \frac{d(\mathcal{P}_v v - v)}{dt}, Q_v^2(\mathcal{P}_v v - v_h) \right\rangle.
\end{aligned}$$

This can be rewritten as

$$\begin{aligned}
&\frac{d}{dt}(a_1\|Q_u^2(\mathcal{P}_u u - u_h)\|^2 + a_2\|Q_v^2(\mathcal{P}_v v - v_h)\|^2) \\
&\quad = a_1\left\langle \frac{d(\mathcal{P}_u u - \mathcal{I}_u u)}{dt}, Q_u^2(\mathcal{P}_u u - u_h) \right\rangle + a_2\left\langle \frac{d(\mathcal{P}_v v - \mathcal{I}_v v)}{dt}, Q_v^2(\mathcal{P}_v v - v_h) \right\rangle \\
&\qquad + a_1\left\langle \frac{d(\mathcal{I}_u u - u)}{dt}, Q_u^2(\mathcal{P}_u u - u_h) \right\rangle + a_2\left\langle \frac{d(\mathcal{I}_v v - v)}{dt}, Q_v^2(\mathcal{P}_v v - v_h) \right\rangle.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\|a_1 Q_u^2(\mathcal{P}_u u - u_h)\| + \|a_2 Q_v^2(\mathcal{P}_v v - v_h)\| \\
&\quad \leq K \int_0^t \left\{ \|Q_u^2(\mathcal{P}_u u_t - \mathcal{I}_u u_t)\| + \|Q_v^2(\mathcal{P}_v v_t - \mathcal{P}_v v_t)\| \right\} ds \\
&\qquad + K \int_0^t \left\{ \|\mathcal{I}_u u_t - u_t\| + \|\mathcal{I}_v v_t - v_t\| \right\} ds.
\end{aligned}$$

Using the triangle inequality, we finally have

$$\begin{aligned}
&\|a_1 Q_u^2(\mathcal{I}_u u - u_h)\| + \|a_2 Q_v^2(\mathcal{I}_v v - v_h)\| \\
&\quad \leq K(\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\| + \|Q_v^2(\mathcal{P}_v v - \mathcal{I}_v v)\|) \\
&\qquad + K \int_0^t \left\{ \|Q_u^2(\mathcal{P}_u u_t - \mathcal{I}_u u_t)\| + \|Q_v^2(\mathcal{P}_v v_t - \mathcal{P}_v v_t)\| \right\} ds \\
&\qquad + K \int_0^t \left\{ \|\mathcal{I}_u u_t - u_t\| + \|\mathcal{I}_v v_t - v_t\| \right\} ds.
\end{aligned}$$

It suffices to estimate the norms $\|Q_v^2(\mathcal{P}_v v - \mathcal{I}_v v)\|$ and $\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\|$. In particular, we will prove $\|Q_v^2(\mathcal{P}_v v - \mathcal{I}_v v)\| \leq K\|v - \mathcal{I}_v v\|_{W'}$ and $\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\| \leq K\|u - \mathcal{I}_u u\|_W$.

We consider the following variational problem: Given $u_2 \in \ker(B_h^*)^\perp$, find $\tilde{\psi} \in V_h$ such that

$$B_h(\tilde{\psi}, \phi) = \langle u_2, \phi \rangle \quad \forall \phi \in U_h. \tag{4.18}$$

The existence of $\tilde{\psi}$ is ensured by the fact that $B_h : V_h \to \ker(B_h^*)^\perp$ is surjective. Taking the supremum in ϕ and using (4.3), we derive the following estimate:

$$\|\tilde{\psi}\|_{Z'} \leq K\|u_2\|. \tag{4.19}$$

Now, we have

$$\begin{aligned}
\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\|^2 &= B_h(\tilde{\psi}, Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)) \\
&= B_h^*(Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u), \tilde{\psi}) \\
&= B_h^*(\mathcal{P}_u u - \mathcal{I}_u u, \tilde{\psi}) \\
&= B_h^*(u - \mathcal{I}_u u, \tilde{\psi}) \\
&\leq \|\psi\|_{Z'} \|u - \mathcal{I}_u u\|_W \\
&\leq K\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\| \|u - \mathcal{I}_u u\|_W.
\end{aligned}$$

Hence, we have

$$\|Q_u^2(\mathcal{P}_u u - \mathcal{I}_u u)\| \leq \|u - \mathcal{I}_u u\|_W. \tag{4.20}$$

Replacing u by u_t, we have

$$\|Q_u^2(\mathcal{P}_u u_t - \mathcal{I}_u u_t)\| \leq \|u_t - \mathcal{I}_u u_t\|_W. \tag{4.21}$$

Similarly, we consider the problem: Given $v_2 \in \ker(B_h)^\perp$, find $\tilde{\phi} \in U_h$ such that

$$B_h^*(\tilde{\phi}, \psi) = \langle v_2, \psi \rangle \quad \forall \psi \in V_h \tag{4.22}$$

with the estimate

$$\|\tilde{\phi}\|_Z \le K\|v_2\|. \tag{4.23}$$

Hence, we have

$$\|Q_v^2(\mathcal{P}_v v - \mathcal{I}_v v)\| \le \|v - \mathcal{I}_v v\|_{W'}. \tag{4.24}$$

Replacing v by v_t, we have

$$\|Q_v^2(\mathcal{P}_v v_t - \mathcal{I}_v v_t)\| \le \|v_t - \mathcal{I}_v v_t\|_{W'}. \tag{4.25}$$

This finishes the proof of estimates of the components of the errors in $\ker(B_h^*)^\perp$ and $\ker(B_h)^\perp$. In the following, we will estimate the components of errors in $\ker(B_h^*)$ and $\ker(B_h)$.

Define $Q_u^1 : U_h \to \ker(B_h^*)$ and $Q_v^1 : V_h \to \ker(B_h)$ as the orthogonal projection operators. Taking $\phi = Q_u^1(\mathcal{I}_u u - u_h)$ and $\psi = Q_v^1(\mathcal{I}_v v - v_h)$, we have

$$a_1\left\langle \frac{d(u-u_h)}{dt}, Q_u^1(\mathcal{I}_u u - u_h)\right\rangle + a_2\left\langle \frac{d(v-v_h)}{dt}, Q_v^1(\mathcal{I}_v v - v_h)\right\rangle = 0.$$

Hence, we obtain

$$\|a_1 Q_u^1(\mathcal{I}_u u - u_h)\| + \|a_2 Q_v^1(\mathcal{I}_v v - v_h)\| \le \int_0^t \Big\{\|\mathcal{I}_u u_t - u_t\| + \|\mathcal{I}_v v_t - v_t\|\Big\}\, ds.$$

Combining all results,

$$\begin{aligned}\|\mathcal{I}_u u - u_h\|_u + \|\mathcal{I}_v v - v_h\|_v \le K\Big\{&\|Q_u^1(\mathcal{I}_u u - u_h)\|_u + \|Q_u^2(\mathcal{I}_u u - u_h)\|_u \\ &+ \|Q_v^1(\mathcal{I}_v v - v_h)\|_v + \|Q_v^2(\mathcal{I}_v v - v_h)\|_v\Big\}.\end{aligned}$$

The proof is complete by noticing that

$$\begin{aligned}\|u - u_h\| &\le \|u - \mathcal{I}_u u\| + \|\mathcal{I}_u u - u_h\|,\\ \|v - v_h\| &\le \|v - \mathcal{I}_v v\| + \|\mathcal{I}_v v - v_h\|. \qquad \square\end{aligned}$$

Now, we will state the convergence theorems. The following is the convergence in L^2 norm. We see that the numerical scheme is $O(h^{k+1})$ when the FE spaces U_h and V_h contain polynomials of degree k.

COROLLARY 4.2. *Let $(u, v) \in W^{1,1}(0, T; W^{k+1,\infty}(\Omega))^3$ be the exact solution to the wave propagation problem* (1.1)–(1.2) *and let (u_h, v_h) be the solution to* (3.3)–(3.4). *Then*

$$\|u - u_h\| + \|v - v_h\| \le Kh^{k+1}(\|u\|_{W^{1,1}(0,T;W^{k+1,\infty}(\Omega))} + \|v\|_{W^{1,1}(0,T;W^{k+1,\infty}(\Omega))^2}). \tag{4.26}$$

Theorem 4.1 and Corollary 4.2 state that the numerical scheme (3.3)–(3.4) is stable and convergent, with optimal rate, with respect to the discrete L^2-norms. The L^2 stability can be satisfied by a very large class of spaces U_h and V_h. However, with L^2 stability only, it is not sufficient to deduce the weak convergence to the true solution;

see Joly [11]. As a result, the numerical solution may behave badly and the optimal rate of convergence is not achieved. This fact can be seen by some numerical examples in the following sections. The extra conditions needed are the inf-sup conditions (4.1)–(4.2), as many mixed FE methods require some compatibility conditions between the two spaces U_h and V_h. These conditions yield a stability in the discrete H^1 norm and we state this result in the following theorem.

THEOREM 4.3. *Let $u_h \in U_h$ and $v_h \in V_h$ be the solution to the numerical scheme* (3.3)–(3.4). *Then*

$$\|u_h\|_Z + \|v_h\|_{Z'} \leq K\left\{\left\|\frac{du_0}{dt}\right\| + \left\|\frac{dv_0}{dt}\right\| + \int_0^t \left(\left\|\frac{dF_1}{dt}\right\| + \left\|\frac{dF_2}{dt}\right\|\right) ds\right\}. \tag{4.27}$$

Proof. Taking t-derivative in (3.3)–(3.4), we have

$$\begin{aligned} a_1\left\langle \frac{d^2u_h}{dt^2}, \phi\right\rangle + B_h\left(\frac{dv_h}{dt}, \phi\right) &= \left\langle \frac{dF_1}{dt}, \phi\right\rangle \quad \forall \phi \in U_h, \\ a_2\left\langle \frac{d^2v_h}{dt^2}, \psi\right\rangle - B_h^*\left(\frac{du_h}{dt}, \psi\right) &= \left\langle \frac{dF_2}{dt}, \psi\right\rangle \quad \forall \psi \in V_h. \end{aligned}$$

Taking $\phi = \frac{du_h}{dt}$ and $\psi = \frac{dv_h}{dt}$ and adding the two equations, we obtain

$$\frac{1}{2}\frac{d}{dt}\left(a_1\left\|\frac{du_h}{dt}\right\|^2 + a_2\left\|\frac{dv_h}{dt}\right\|^2\right) = \left\langle \frac{dF_1}{dt}, \frac{du_h}{dt}\right\rangle + \left\langle \frac{dF_2}{dt}, \frac{dv_h}{dt}\right\rangle,$$

and consequently

$$\left\|a_1\frac{du_h}{dt}\right\| + \left\|a_2\frac{dv_h}{dt}\right\| \leq K\left\{\left\|\frac{du_0}{dt}\right\| + \left\|\frac{dv_0}{dt}\right\| + \int_0^t \left(\left\|\frac{dF_1}{dt}\right\| + \left\|\frac{dF_2}{dt}\right\|\right) ds\right\}.$$

By using (4.3),

$$\|v_h\|_{Z'} \leq K\|B_h v_h\|_W = K \sup_{\phi \in U_h} \frac{|B_h(v_h, \phi)|}{\|\phi\|_W} \leq K \sup_{\phi \in U_h} \frac{|\langle \frac{du_h}{dt}, \phi\rangle|}{\|\phi\|_W} = K\left\|\frac{du_h}{dt}\right\|.$$

Similarly, we have

$$\|u_h\|_Z \leq K\left\|\frac{dv_h}{dt}\right\|.$$

This completes the proof. ∎

Before we state the convergence theorem, we will state a L^2 convergence result which is very similar to Theorem 4.1—that all functions are replaced by their time derivative. It can be proved in exactly the same way as the proof of Theorem 4.1.

LEMMA 4.4. *Let $u \in U$ and $v \in V$ be the solution to* (1.1)–(1.2) *and let $u_h \in U_h$ and $v_h \in V_h$ be the solution to the numerical scheme* (3.3)–(3.4). *Then for $0 \leq t \leq T$, we have*

$$\begin{aligned} \left\|\frac{d}{dt}(u - u_h)(t)\right\| + \left\|\frac{d}{dt}(v - v_h)(t)\right\| \leq K\Big\{ &\inf_{\phi \in U_h} \|u_t - \phi\|_W + \inf_{\psi \in V_h} \|v_t - \psi\|_{W'} \\ &+ \int_0^t \Big(\inf_{\phi \in U_h} \|u_{tt} - \phi\|_W + \inf_{\psi \in V_h} \|v_{tt} - \psi\|_{W'}\Big) ds\Big\}. \end{aligned} \tag{4.28}$$

The following theorem states the convergence of the method (3.3)–(3.4) in the discrete H^1 norm. It can be seen that the H^1 error is optimal with respect to the norms and the FE spaces.

THEOREM 4.5. *Let $u \in U$ and $v \in V$ be the solution to* (1.1)–(1.2) *and let $u_h \in U_h$ and $v_h \in V_h$ be the solution to the numerical scheme* (3.3)–(3.4). *Then for $0 \le t \le T$, we have*

$$\|v - v_h\|_{Z'} \le K\left(\left\|\frac{d}{dt}(u - u_h)\right\| + \inf_{\psi \in V_h} \|\psi - v\|_{Z'}\right), \tag{4.29}$$

$$\|u - u_h\|_{Z} \le K\left(\left\|\frac{d}{dt}(v - v_h)\right\| + \inf_{\phi \in U_h} \|\phi - u\|_{Z}\right). \tag{4.30}$$

Proof. By the inf-sup condition (4.3), we obtain

$$\|\mathcal{P}_v v - v_h\|_{Z'} \le K\|B_h(\mathcal{P}_v v - v_h)\|_W = K \sup_{\phi \in U_h} \frac{B_h(\mathcal{P}_v v - v_h, \phi)}{\|\phi\|_W}.$$

Recalling (4.15), we have

$$\left\langle \frac{d(u - u_h)}{dt}, \phi \right\rangle + B_h(\mathcal{P}_v v - v_h, \phi) = 0 \quad \forall \phi \in U_h,$$

and consequently,

$$\|\mathcal{P}_v v - v_h\|_{Z'} \le K\left\|\frac{d}{dt}(u - u_h)\right\|.$$

Let $\mathcal{I}_v v \in V_h$ be an arbitrary element of the FE space V_h, using the triangle inequality

$$\|v - v_h\|_{Z'} \le \|v - \mathcal{I}_v v\|_{Z'} + \|\mathcal{I}_v v - \mathcal{P}_v v\|_{Z'} + \|\mathcal{P}_v v - v_h\|_{Z'}.$$

Following the proof of Theorem 4.1, we have

$$\|\mathcal{I}_v v - \mathcal{P}_v v\|_{Z'} \le K\|\mathcal{I}_v v - v\|_{Z'}.$$

Hence, we obtain

$$\|v - v_h\|_{Z'} \le K\left(\left\|\frac{d}{dt}(u - u_h)\right\| + \|\mathcal{I}_v v - v\|_{Z'}\right).$$

Since $\mathcal{I}_v v$ is arbitrary,

$$\|v - v_h\|_{Z'} \le K\left(\left\|\frac{d}{dt}(u - u_h)\right\| + \inf_{\psi \in V_h} \|\psi - v\|_{Z'}\right).$$

So, (4.29) is proved. The estimte (4.30) can be proved in a similar fashion. ∎

Now, we state and prove the convergence in the discrete H^1 norm. We see that the numerical scheme is $O(h^k)$ in the discrete H^1 norm when the FE spaces U_h and V_h contain polynomials of degree k.

COROLLARY 4.6. *Assume $k \ge 0$ is the largest integer such that U_h and V_h contain polynomials of degree k. Let $(u, v) \in W^{1,p}(0, T; H^{k+1}(\Omega))^3 \cap W^{2,p}(0, T; W^{k,\infty}(\Omega))^3$,*

for $p > 1$, be the exact solution to the wave propagation problem (1.1)–(1.2)*, and let (u_h, v_h) be the solution to the numerical scheme* (3.3)–(3.4)*. Then*

$$\|u - u_h\|_Z + \|v - v_h\|_{Z'} \leq Kh^k(\|(u,v)\|_{W^{1,p}(0,T;H^{k+1}(\Omega))^3} + \|(u,v)\|_{W^{2,p}(0,T;W^{k,\infty}(\Omega))^3}). \tag{4.31}$$

Next, we prove a superconvergence result for some component of the derivative of u_h. We state this result as the following theorem.

THEOREM 4.7. *Let u be the exact solution to the wave propagation problem* (1.1)–(1.2) *and let u_h be the solution to the numerical method* (3.3)–(3.4)*. Then*

$$\begin{aligned}&\|L^\perp B^*(u - u_h)\|_{L^2(\Omega)} \\ &\quad \leq Kh^{k+1}\big\{\|(u,v)\|_{W^{2,1}(0,T;W^{k+1,\infty}(\Omega))^3} + \|u\|_{W^{1,1}(0,T;H^{k+2}(\Omega))}\big\}.\end{aligned} \tag{4.32}$$

Proof. To prove this, we observe that for all $\psi \in V_h$

$$a_2 \int_\Omega \frac{\partial}{\partial t}(v - v_h) \cdot \psi \, dx - B_h^*(u - u_h, \psi) = 0.$$

Let $\pi_h u$ be a function such that $\pi_h u|_\tau$ is a $(k+1)$th degree polynomial interpolant of $u|_\tau$. Then we have

$$B_h^*(u_h - \pi_h u, \psi) = B_h^*(u - \pi_h u, \psi) - a_2 \int_\Omega \frac{\partial}{\partial t}(v - v_h) \cdot \psi \, dx.$$

Recalling the definition of B_h^*, we have

$$B_h^*(u_h - \pi_h u, \psi) = \int_\Omega B^*(u_h - \pi_h u)\psi \, dx + \sum_{\sigma \in \mathcal{E}_u} \int_\sigma [u_h - \pi_h u] L\psi \, d\sigma.$$

Now, we choose $\psi \in V_h$ such that $L\psi = 0$ and $L^\perp \psi = L^\perp B^*(u_h - \pi_h u)$. This is equivalent to sets (VD2) and (VD3) to being zero. Therefore,

$$\begin{aligned}B_h^*(u_h - \pi_h u, \psi) &= \int_\Omega (L^\perp B^*(u_h - \pi_h u))^2 \, dx, \\ \|\psi\|_{L^2(\Omega)} &= \|L^\perp B^*(u_h - \pi_h u)\|_{L^2(\Omega)}.\end{aligned}$$

Consequently, we have

$$\|L^\perp B^*(u_h - \pi_h u)\|_{L^2(\Omega)}^2 = B_h^*(u - \pi_h u, \psi) - a_2 \int_\Omega \frac{\partial}{\partial t}(v - v_h) \cdot \psi \, dx. \tag{4.33}$$

Now we will estimate the right-hand side of (4.33). By Lemma 4.4 and interpolation error estimates (3.16), the second term on the right-hand side of (4.33) can be estimated by

$$\left\| \frac{\partial}{\partial t}(v - v_h) \right\|_{L^2(\Omega)} \leq Kh^{k+1}(\|u\|_{W^{2,1}(0,T;W^{k+1,\infty}(\Omega))} + \|v\|_{W^{2,1}(0,T;W^{k+1,\infty}(\Omega))^2}).$$

Using inverse type inequalities, we have

$$\|B^*(u - \pi_h u)\|_{L^2(\Omega)} \leq Kh^{k+1}|u|_{H^{k+2}(\Omega)}, \quad \|\psi \cdot l\|_{L^2(\sigma)} \leq Kh^{-\frac{1}{2}}\|\psi\|_{L^2(\tau')},$$

and by the trace inequality, we obtain

$$\|u - \pi_h u\|_{L^2(\sigma)} \le K h^{k+\frac{3}{2}} |u|_{H^{k+2}(\Omega)}.$$

Then the first term on the right-hand side of (4.33) can be estimated by

$$\begin{aligned} B_h^*(u - \pi_h u, \psi) &\le \|B^*(u - \pi_h u)\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)} \\ &\quad + \sum_{\sigma \in \mathcal{E}_u} \|u - \pi_h u\|_{L^2(\sigma)} \|L\psi\|_{L^2(\sigma)} \\ &\le K h^{k+1} |u|_{H^{k+2}(\Omega)} \|\psi\|_{L^2(\Omega)}. \end{aligned}$$

Combining the result, we prove the theorem. ∎

Now, we will discuss the condition (1.4). In the following theorem, we show that the numerical solution v_h satisfies (1.4) in a weak sense. Let S_h be the space of standard H^1-conforming FE space of degree $k+1$, namely $p \in S_h$ if $p|_\tau \in P^{k+1}(\tau)$ and p is continuous across each $\sigma \in \mathcal{E}$.

THEOREM 4.8. *Let v be the exact solution of the wave propagation problem* (1.1)–(1.2) *and let v_h be the numerical solution to the numerical scheme* (3.3)–(3.4). *Then*

$$\int_\Omega \frac{\partial (v - v_h)}{\partial t} \psi \, dx = 0 \tag{4.34}$$

if and only if $\psi = B^\perp p$ for $p \in S_h$.

Proof. If $\psi = \nabla p$ for $p \in S_h$, then $\psi \in V_h$. Using (4.16), we have proved (4.34). Assume (4.34) holds. Then, using (4.16), we have $B_h^*(\mathcal{P}_u u - u_h, \psi) = 0$. By (4.3), we have $\|\psi\|_{Z'} = 0$. By the definition of Z'-norm, we have $B\psi = 0$ and $[L\psi]|_\sigma = 0$ for all $\sigma \in \mathcal{E}_u$. Using $B\psi = 0$, we have $\psi = B^\perp p$ for some p. Since $\psi|_\tau \in P^k(\tau)^2$, we have $p|_\tau \in P^{k+1}(\tau)$. Notice that $L\psi$ is continuous on each edge in $\mathcal{E}_v^0$. Using this and $[L\psi]|_\sigma = 0$ for all $\sigma \in \mathcal{E}_u$, we have that p is continuous across each edge in $\mathcal{E}_u \cup \mathcal{E}_v^0$. So, the proof is complete. ∎

5. Verification of inf-sup conditions. Now, we are in a position to prove that the choice of U_h and V_h above satisfies the inf-sup condition (4.1)–(4.2).

THEOREM 5.1. *There is a uniform constant $K > 0$ such that*

$$\inf_{\psi \in V_h} \sup_{u \in U_h} \frac{B_h^*(u, \psi)}{\|u\|_W \|\psi\|_{Z'}} \ge K. \tag{5.1}$$

Proof. Let $\psi \in V_h$. It suffices to find $u \in U_h$ such that

$$B_h^*(u, \psi) \ge K \|\psi\|_{Z'}^2 \quad \text{and} \quad \|u\| \le K \|\psi\|_{Z'}. \tag{5.2}$$

Recalling the definition of B_h^*, we have

$$B_h^*(u, \psi) = \int_\Omega u (B\psi) \, dx + \sum_{\sigma \in \mathcal{E}_u} \int_\sigma [L\psi] u \, d\sigma.$$

First, we will define $u_1 \in U_h$ such that

$$\int_\Omega u_1 (B\psi) \, dx \ge K \int_\Omega (B\psi)^2 \, dx \quad \text{and} \quad \int_\Omega (u_1)^2 \, dx \le \int_\Omega (B\psi)^2 \, dx. \tag{5.3}$$

Let $\tau \in \mathcal{T}$. We define the function u_1 such that $u_1|_\tau = \lambda_{\tau,2}\lambda_{\tau,3}B\psi$. Notice that $B\psi|_\tau \in P^{k-1}(\tau)$ and $u_1|_\tau$ is zero on the two edges of τ that belong to $\mathcal{E}_u$, so $u_1|_\tau \in R^k(\tau)$. Since $\lambda_{\tau,2}\lambda_{\tau,3} \leq 1$, the second equation in (5.3) holds. Notice that the quantity $\int_\tau q_{k-1}^2 \lambda_{\tau,2}\lambda_{\tau,3}\, dx$ defines a norm for q_{k-1} in the space $P^{k-1}(\tau)$. Since norms in finite dimensional spaces are equivalent, we have

$$\int_\tau \lambda_{\tau,2}\lambda_{\tau,3}(B\psi)^2\, dx \geq K \int_\tau (B\psi)^2\, dx.$$

Summing up this equation for all $\tau \in \mathcal{T}$ proves the first equation in (5.3).

Next, we will define $u_2 \in U_h$ such that

$$\sum_{\sigma\in\mathcal{E}_u} \int_\sigma [L\psi]u_2\, d\sigma = \sum_{\sigma\in\mathcal{E}_u} (h_\sigma)^{-1} \int_\sigma [L\psi]^2\, d\sigma, \tag{5.4}$$

$$\int_\Omega (u_2)^2\, dx \leq K \sum_{\sigma\in\mathcal{E}_u} (h_\sigma)^{-1} \int_\sigma [L\psi]^2\, d\sigma. \tag{5.5}$$

To do so, we define u_2 so that

1. $u_2 = (h_\sigma)^{-1}[L\psi]$ at the $k+1$ Gaussian points of σ for all $\sigma \in \mathcal{E}_u$, and
2. $\int_\tau u_2 q_{k-1}\, dx = 0$ for all $q_{k-1} \in P^{k-1}(\tau)$ and $\tau \in \mathcal{T}$.

Then, clearly, (5.4) is satisfied. We will define a subspace U_h^0 of U_h by

$$U_h^0 = \left\{ \phi \in U_h \mid \int_\tau \phi q_{k-1}\, dx = 0, \forall q_{k-1} \in P^{k-1}(\tau), \forall \tau \in \mathcal{T} \right\}.$$

Then the following quantity

$$\sum_{\sigma\in\mathcal{E}_u} h_\sigma^2 \sum_{j=1}^{k+1} \phi(\alpha_j)^2$$

defines a norm for U_h^0. Since norms in finite dimensional spaces are equivalent, we have

$$\int_\tau u_2^2\, dx \leq K \sum_{\sigma\in\mathcal{E}_u} h_\sigma^2 \sum_{j=1}^{k+1} u_2(\alpha_j)^2.$$

By the definition of u_2,

$$\int_\tau u_2^2\, dx \leq K \sum_{\sigma\in\mathcal{E}_u} \sum_{j=1}^{k+1} [L\psi(\alpha_j)]^2.$$

Since $[L\psi]$ is a polynomial of degree k, we have

$$\sum_{j=1}^{k+1} [L\psi(\alpha_j)]^2 \leq K(h_\sigma)^{-1} \int_\sigma [L\psi]^2\, dx,$$

which follows from norm equivalence in finite dimensional spaces. So, (5.5) is proved.

To prove (5.2), we take $u = u_1 + u_2$. Notice that by the definitions of u_1 and u_2, we have

$$\int_\tau u_2(B\psi)\, dx = 0 \quad \text{and} \quad \int_\sigma [L\psi]u_1\, d\sigma = 0.$$

Using this together with (5.3), (5.4), and (5.5), we have proved (5.2). ∎

THEOREM 5.2. *There is a uniform constant $K > 0$ such that*

$$\inf_{\phi \in U_h} \sup_{v \in V_h} \frac{B_h(v,\phi)}{\|v\|_{W'} \|\phi\|_Z} \geq K. \tag{5.6}$$

Proof. Let $\phi \in U_h$. As in the proof of the previous theorem, we find $v \in V_h$ such that

$$B_h(v,\phi) \geq K\|\phi\|_Z^2 \quad \text{and} \quad \|v\| \leq K\|\phi\|_Z. \tag{5.7}$$

Recalling the definition of B_h, we have

$$B_h(v,\phi) = \int_\Omega (B^*\phi)v \, dx - \sum_{\sigma \in \mathcal{E}_v^0} \int_\sigma (Lv)[\phi] \, d\sigma.$$

We define $v_1 \in V_h$ such that

$$\int_\Omega (B^*\phi)v_1 \, dx \geq K_1 \int_\Omega |B^*\phi|^2 \, dx \quad \text{and} \quad \int_\Omega |v_1|^2 \, dx \leq K \int_\Omega |B^*\phi|^2 \, dx. \tag{5.8}$$

We define the set

$$V_h^1(\tau) = \{v|_\tau \mid v \in V_h; Lv|_\sigma = 0 \ \forall \sigma \in \mathcal{E}_v\}$$

and the linear functional

$$f_\tau(\eta) = \int_\tau (B^*\phi)\eta \, dx$$

for $\eta \in V_h^1(\tau)$. With the standard L^2 norm, $V_h^1(\tau)$ is a Hilbert space. By the Riesz representation theorem, there exist $v_{1,\tau} \in V_h^1(\tau)$ such that

$$f_\tau(\eta) = \int_\tau v_{1,\tau}\eta \, dx \quad \text{and} \quad \|v_{1,\tau}\|_{L^2(\tau)} = \|f_\tau\|_{L^2(\tau)^*},$$

where $*$ denotes a norm in the dual space. That is,

$$\|f_\tau\|_{L^2(\tau)^*} = \sup_{\eta \in V_h^1(\tau)} \frac{f_\tau(\eta)}{\|\eta\|_{L^2(\tau)}}.$$

We define v_1 such that $v_1|_\tau = v_{1,\tau}$. Then

$$\int_\Omega |v_1|^2 \, dx = \sum_{\tau \in \mathcal{T}} \int_\tau |v_{1,\tau}|^2 \, dx \leq \sum_{\tau \in \mathcal{T}} \int_\tau |B^*\phi|^2 \, dx = \int_\Omega |B^*\phi|^2 \, dx,$$

which proves the second inequality in (5.8). To prove the first inequality in (5.8), we will first show $\|f_\tau\|_{L^2(\tau)^*}$ defines a norm for B^*U_h on τ. So, it suffices to show $B^*\phi = 0$ if $\|f_\tau\|_{L^2(\tau)^*} = 0$. Assume $\|f_\tau\|_{L^2(\tau)^*} = 0$. Then we have $\int_\tau (B^*\phi)\eta \, dx = 0$ for all $\eta \in V_h^1(\tau)$. Notice that $B^*\phi \in P^k(\tau)^2$. Taking $L\eta = 0$ and $L^\perp\eta = L^\perp(B^*\phi)$ yields $\int_\tau (L^\perp(B^*\phi))^2 \, dx = 0$. So, we have $L^\perp(B^*\phi) = 0$ on τ. By the definitions of $L^\perp$ and B^*, we have $(\partial_{\lambda_{\tau,2}} - \partial_{\lambda_{\tau,3}})\phi = 0$. So, we have

$$\phi = \sum_{j=0}^{k+1} c_j(\lambda_{\tau,2} + \lambda_{\tau,3})^j.$$

Since $\phi \in U_h$, we have $c_{k+1} = 0$ and therefore $L(B^*\phi) \in P^{k-1}(\tau)$. Taking $L^\perp \eta = 0$ and $L\eta = \lambda_{\tau,1} L(B^*\phi)$, we have $\int_\tau \lambda_{\tau,1}(L(B^*\phi))^2\, dx = 0$, which implies $L(B^*\phi) = 0$. Hence $B^*\phi = 0$. Since norms in finite dimensional spaces are equivalent, we have $\|f_\tau\|_{L^2(\tau)^*} \geq K\|B^*\phi\|_{L^2(\tau)}$. Consequently, we obtain

$$\int_\Omega (B^*\phi)v_1\, dx = \sum_{\tau\in\mathcal{T}} \int_\tau (B^*\phi)v_{1,\tau}\, dx = \sum_{\tau\in\mathcal{T}} f_\tau(v_{1,\tau}) = \sum_{\tau\in\mathcal{T}} \int_\tau v_{1,\tau}^2\, dx,$$

which proves the first inequality in (5.8).

We will find $v_2 \in V_h$ and a function v^+ such that

$$-\sum_{\sigma\in\mathcal{E}_v} \int_\sigma (L(v_2 + v^+))[\phi]\, d\sigma \geq K_2 \sum_{\sigma\in\mathcal{E}_v^0} (h_\sigma)^{-1} \int_\sigma [\phi]^2\, d\sigma, \tag{5.9}$$

$$\int_\Omega |v_2 + v^+|^2\, dx \leq K_3 \sum_{\sigma\in\mathcal{E}_v^0} (h_\sigma)^{-1} \int_\sigma [\phi]^2\, d\sigma. \tag{5.10}$$

Let $\sigma \in \mathcal{E}_v^0$ and let $\tau, \tilde{\tau}$ be the two triangles sharing the same edge σ. We define $V_h^2(\tau \cup \tilde{\tau})$ as follows:

$$V_h^2(\tau\cup\tilde{\tau}) = \{v|_{\tau\cup\tilde{\tau}} + \nabla(\lambda_{\tau,2}\lambda_{\tau,3}q_k) \mid v \in V_h, (VD1) = (VD2) = 0; q_k \in P^k(\tau\cup\tilde{\tau})\}.$$

Here, by (VD1) = (VD2) = 0, we mean both the degrees of freedom defined by (VD1) and (VD2) are equal to zero. Also, the polynomial q_k is fixed and will be chosen in the following. With the standard $L^2(\tau\cup\tilde{\tau})^2$ norm, $V_h^2(\tau\cup\tilde{\tau})$ is a Hilbert space. We also define the following linear functional:

$$g_\tau(\eta) = \int_\sigma (L\eta)[\phi]\, d\sigma$$

for all $\eta \in V_h^2(\tau\cup\tilde{\tau})$. By the Riesz representation theorem, there is an element $v_{2,\tau} + v_\tau^+ \in V_h^2(\tau\cup\tilde{\tau})$ such that

$$g_\tau(\eta) = \int_{\tau\cup\tilde{\tau}} (v_{2,\tau} + v_\tau^+)\eta\, dx \quad \text{and} \quad \|v_{2,\tau} + v_\tau^+\|_{L^2(\tau\cup\tilde{\tau})^2} = \|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*},$$

where the norm $\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*}$ is defined as

$$\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*} = \sup_{\eta\in V_h^2(\tau\cup\tilde{\tau})} \frac{g_\tau(\eta)}{\|\eta\|_{L^2(\tau\cup\tilde{\tau})^2}}.$$

We then define v_2 by $v_2|_\tau = v_{2,\tau}$ and v^+ by $v^+|_\tau = v_\tau^+$. Since $\int_\sigma (L\eta)^2\, d\sigma \leq K(h_\sigma^{-1}) \int_{\tau\cup\tilde{\tau}} |\eta|^2\, dx$, we have

$$\|g_\tau\|^2_{(L^2(\tau\cup\tilde{\tau})^2)^*} \leq \frac{1}{\|\eta\|^2_{L^2(\tau\cup\tilde{\tau})^2}} \int_\sigma (L\eta)^2\, d\sigma \int_\sigma [\phi]^2\, d\sigma \leq K h_\sigma^{-1} \int_\sigma [\phi]^2\, d\sigma.$$

Summing up all $\tau \in \mathcal{T}$ proves (5.10). To prove (5.9), we will first show that $[\phi]|_\sigma = 0$ if $\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*} = 0$. Now, we assume $\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*} = 0$. Then we have

$\int_\sigma (L\eta)[\phi]\ d\sigma = 0$ for all $\eta \in V_h^2(\tau \cup \tilde{\tau})$. We take $\eta|_\tau \in P^k(\tau)^2$ and $\eta|_{\tilde{\tau}} \in P^k(\tilde{\tau})^2$ such that $L\eta = [\phi]$ at the $k+1$ Gaussian points of σ. Then, $\eta \in V_h^2(\tau \cup \tilde{\tau})$ and

$$h_\sigma \sum_{j=1}^{k+1} w_j [\phi(\alpha_j)]^2\ dx = \int_\sigma (L\eta)[\phi]\ dx = 0,$$

where the first equality follows from the Gaussian quadrature rule; here w_j denotes the weight and α_j denotes the quadrature point. Since the weights $w_j > 0$, we have $[\phi] = 0$ at the $k+1$ Gaussian points of σ. Notice that $[\phi]|_\sigma$ is a polynomial of degree $k+1$. So, we have that $[\phi]|_\sigma$ is a scalar multiple of the $(k+1)$th degree Legendre polynomial, namely $[\phi]|_\sigma = b\mathbb{P}_{k+1}$ for some constant b, where $\mathbb{P}_{k+1}$ is the Legendre polynomial of degree $k+1$. We take $\eta = \nabla(\lambda_{\tau,2}\lambda_{\tau,3}q_k)$ with q_k to be determined below. Notice that $L\eta$ is the tangential derivative of $\lambda_{\tau,2}\lambda_{\tau,3}q_k$ along σ. By a change of variable, we have

$$\begin{aligned}\int_\sigma (L\eta)[\phi]\ d\sigma &= d\int_{-1}^1 \frac{d}{dz}((1-z^2)q_k(z))[\phi(z)]\ dz \\ &= d\int_{-1}^1 \frac{d}{dz}((1-z^2)q_k(z))b\mathbb{P}_{k+1}(z)\ dz\end{aligned}$$

for some constant $d > 0$. Notice that $\mathbb{P}_{k+1}$ satisfies the Legendre differential equation

$$\frac{d}{dz}\left((1-z^2)\frac{d\mathbb{P}_{k+1}}{dz}\right) = -(k+1)(k+2)\mathbb{P}_{k+1}.$$

So, we take $q_k|_\sigma = \frac{d\mathbb{P}_{k+1}}{dz}$ and extend the definition of q_k over all of $\tau \cup \tilde{\tau}$. Then we obtain

$$\int_\sigma (L\eta)[\phi]\ dx = -bd(k+1)(k+2)\int_{-1}^1 \mathbb{P}_{k+1}^2(z)\ dz.$$

This implies that $b = 0$. Hence $[\phi] = 0$. So, $\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*}$ defines a norm on $L^2(\sigma)$. Since norms in finite dimensional spaces are equivalent, we obtain

$$\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*} \geq K(\tau \cup \tilde{\tau})\|[\phi]\|_{L^2(\sigma)}.$$

A standard scale change argument yields

$$\|g_\tau\|_{(L^2(\tau\cup\tilde{\tau})^2)^*} \geq K(h_\sigma)^{-\frac{1}{2}}\|[\phi]\|_{L^2(\sigma)},$$

which proves (5.9).

Combining (5.8), (5.9), and (5.10), we have

$$\begin{aligned}B_h(\delta v_1 + v_2 + v^+, \phi) \geq \delta K_1 \int_\Omega |B^*\phi|^2\ dx + K_2 \sum_{\sigma\in\mathcal{E}_v} (h_\sigma)^{-1}\int_\sigma [\phi]^2\ d\sigma \\ + \int_\Omega (B^*\phi)(v_2 + v^+)\ dx.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\int_\Omega (B^*\phi)(v_2 + v^+)\ dx \geq -\frac{K_2}{2K_3}\int_\Omega |v_2 + v^+|^2\ dx - \frac{K_3}{2K_2}\int_\Omega |B^*\phi|^2\ dx.$$

So, we have

$$B_h(\delta v_1 + v_2 + v^+, \phi) \geq \left(\delta K_1 - \frac{K_3}{2K_2}\right) \int_\Omega |B^*\phi|^2 \, dx + \frac{K_2}{2} \sum_{\sigma \in \mathcal{E}_v} (h_\sigma)^{-1} \int_\sigma [\phi]^2 \, d\sigma.$$

Now, we choose δ such that $\delta K_1 - \frac{K_3}{2K_2} = 1$. Then we have

$$B_h(\delta v_1 + v_2 + v^+, \phi) \geq \min\left(1, \frac{K_2}{2}\right) \|\phi\|_Z^2.$$

Since $B_h(v^+, \phi) = 0$, we have

$$B_h(\delta v_1 + v_2, \phi) \geq \min\left(1, \frac{K_2}{2}\right) \|\phi\|_Z^2.$$

We take $v = \delta v_1 + v_2 \in V_h$ so that the first inequality in (5.7) is proved. To prove the second inequality in (5.7), we first notice that $\|v_{2,\tau}\|_{L^2(\sigma)}$ defines a norm for $V_h^2(\tau \cup \tilde{\tau})$ since $\|v_{2,\tau}\|_{L^2(\sigma)} = 0$ implies (VD3) $= 0$ which in turn implies $v_{2,\tau} = 0$ by unisolvence of the FE space V_h. By norm equivalence in finite dimensional spaces, we have $\|v_{2,\tau}\|_{L^2(\tau)} \leq Kh^{\frac{1}{2}} \|v_{2,\tau}\|_{L^2(\sigma)} \leq \|v_{2,\tau}\|_{L^2(\tau)}$. Then we obtain

$$\|\delta v_1 + v_2\|_{L^2(\tau)} \leq \delta \|v_1\|_{L^2(\tau)} + \|v_2\|_{L^2(\tau)} \leq K\|v_1\|_{L^2(\tau)} + h^{\frac{1}{2}} \|v_2\|_{L^2(\sigma)}.$$

Furthermore, we have the following orthogonality condition:

$$\begin{aligned} \int_\sigma v_{2,\tau} v_\tau^+ \, d\sigma &= \int_{-1}^1 v_{2,\tau} \frac{d}{dz}\left((1-z^2)\frac{d\mathbb{P}_{k+1}}{dz}\right) dz \\ &= -(k+1)(k+2) \int_{-1}^1 v_{2,\tau} \mathbb{P}_{k+1} \, dz = 0 \end{aligned}$$

since the function $\mathbb{P}_{k+1}$ is equal to zero at the $k+1$ Gaussian points of σ. By the orthogonality condition,

$$\|v_2\|_{L^2(\sigma)}^2 \leq \|v_2\|_{L^2(\sigma)}^2 + \|v^+\|_{L^2(\sigma)}^2 = \|v_2 + v^+\|_{L^2(\sigma)}^2 \leq Kh^{-1} \|v_2 + v^+\|_{L^2(\tau \cup \tilde{\tau})}^2,$$

where the last inequality follows from trace inequality. So, we have

$$\|\delta v_1 + v_2\|_{L^2(\tau \cup \tilde{\tau})^2} \leq K\|v_1\|_{L^2(\tau \cup \tilde{\tau})^2} + K\|v_2 + v^+\|_{L^2(\tau \cup \tilde{\tau})^2}.$$

Summing up all $\tau \in \mathcal{T}$ and using estimates (5.8) and (5.10) completes the proof. ∎

6. Numerical examples. In this section we present a series of numerical experiments which give quantitative results and confirm the rate of convergence of the method (3.3)–(3.4). We will, in particular, consider the TE mode of Maxwell's equations (E) and set $\Omega = [0, 2\pi]^2$, $a_1 = a_2 = 1$, and $F_1 = F_2 = 0$. In addition, the function u is the magnetic field H while the vector v is the electric field E. The exact solution to Maxwell's equations is

$$\begin{aligned} H(x,t) &= \cos(t)\cos(x_1) + \cos(t)\cos(x_2), \\ E_1(x,t) &= -\sin(t)\sin(x_2), \\ E_2(x,t) &= \sin(t)\sin(x_1). \end{aligned}$$

TABLE 6.1
L^2 norm errors at $T = \pi/4$ for $k = 0$. Rate of convergence is 1.0298.

N	NT	L^2 error
10	100	1.311
20	200	0.4799
40	400	0.2782
80	800	0.1301
160	1600	0.06653
320	3200	0.03378

TABLE 6.2
Errors in various norms at $T = \pi/4$ for $k = 1$.

N	NT	L^2 error	$\|H - H_h\|_Z$	$\|E - E_h\|_{Z'}$
10	100	0.1809	1.526	0.7213
20	200	0.04528	0.6619	0.3472
40	400	0.01111	0.3498	0.1623
80	800	0.002797	0.1597	0.1019
160	1600	0.0007022	0.06220	0.04968

The domain Ω is triangulated in the following manner. First, we divide Ω into $N \times N$ uniform squares. Then, we subdivide each square by connecting the lower left corner and the upper right corner to obtain two triangles. We further subdivide each triangle into three triangles by connecting the center of the triangle to its three vertices. For the resulting ODE system in time, we use the standard leap-frog scheme. Below we use NT to represent the number of time steps.

We first consider an example for the first order method, that is, $k = 0$. We then test the rate of convergence by comparing the solution to the scheme (3.3)–(3.4) and the exact solution at $T = \frac{\pi}{4}$. Table 6.1 shows the discrete L^2 errors for various mesh sizes, from $N = 10$ to $N = 320$. Here, we choose the time step small enough so that a suitable CFL condition for the leap-frog scheme is satisfied. We will use the results from Table 6.1 and the least squares method to estimate the rates of convergence of the scheme in the discrete L^2 norm. More precisely, we assume the error is proportional to h^β for some $\beta \in \mathbb{R}$, and then perform a least square data fitting using the data from Table 6.1. Doing this, the numerical rate of convergence is 1.0298. This confirms that the scheme is first order convergence in the discrete L^2 norm.

Next, we consider an example with $k = 1$, that is, the FE spaces U_h and V_h, which contain all linear polynomials and a subset of quadratic polynomials. We will test the rates of convergence in various norms at $T = \pi/4$. Table 6.2 shows the results of error in various norms with various mesh sizes. In the third column of Table 6.2, we give the sum of the error for both H and E in the discrete L^2 norm. In the fourth column, we have the error for the magnetic field in the Z-norm. In the fifth column, we have the error for the electric field in the Z'-norm. Table 6.3 shows the estimated rates of convergence. From the table, we see that the estimated rate of convergence in the discrete L^2 norm is approximately 2. Moreover, the estimated rates of convergence in H^1 semi-norms are approximately equal to 1. Our theoretical statements are thus confirmed by this experiment.

In Table 6.4, we also show the error for the divergence of E as well as the normal jump of E. We have not proved convergence results in these two norms, but they are implied by the estimates that we proved in previous sections. The error in the

TABLE 6.3
Estimated rate of convergence at $T = \pi/4$ for $k = 1$.

Norm	Estimated rate
L^2 norm	2.004
$\|H - H_h\|_Z$	1.129
$\|E - E_h\|_{Z'}$	0.9489

TABLE 6.4
Normal jump and divergence errors at $T = \pi/4$ for $k = 1$.

N	NT	$\sum_{\sigma\in\mathcal{E}}(h_\sigma^{-\frac{1}{2}})\|L^\perp(E - E_h)\|_{L^2(\sigma)}$	$\|\mathrm{div}(E - E_h)\|_{L^2(\Omega)}$
10	100	1.212	0.2519
20	200	0.5692	0.1180
40	400	0.2610	0.05039
80	800	0.1387	0.03494
160	1600	0.06686	0.01632

divergence of E is measured by $\|\mathrm{div}(E - E_h)\|_{L^2(\Omega)}$ while the error in the normal jump of E is measured by $\sum_{\sigma\in\mathcal{E}}(h_\sigma^{-\frac{1}{2}})\|L^\perp(E - E_h)\|_{L^2(\sigma)}$. The estimated rates in both norms are 0.9652 and 1.040, respectively. So, the rates of convergence are indeed first order for these two norms.

In what follows, we will consider the one dimensional scalar wave equation on $\Omega = [0, 2\pi]$,

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial x}, \quad \frac{\partial v}{\partial t} = \frac{\partial u}{\partial x},$$

with periodic boundary condition. The purpose is to compare our new optimal DG with the central DG method. The central DG method is typically based on piecewise polynomial approximation without continuity requirement across cell interfaces. Flux integrals along cell boundaries are evaluated by using the average of two values of the numerical solutions from the two neighboring cells, or the so called central numerical flux; see, for example, [7]. We choose $u(x, t) = e^{\sin(x-t)}$ and $v(x, t) = -e^{\sin(x-t)}$ to be the exact solution. We will compare the numerical solutions by the two methods at $T = 20$ using 20 spatial cells and the leap-frog scheme for the time discretization. Figure 6.1 shows the numerical solutions. First, we see that both methods preserve energy. Second, we see that there are spurious modes in the numerical solutions obtained by the central DG. It can be shown that the central DG does not satisfy the inf-sup conditions that we introduce in this paper. With our new optimal DG, which verifies the inf-sup conditions, we see that there is no spurious mode appearing in the numerical solution.

Now we will compare our new DG with an upwind DG method. The upwind DG method is typically based on piecewise polynomial approximation without continuity requirement. Flux integrals along cell boundaries are evaluated by taking the upwind value of the numerical solution from the two neighboring cells, or the so called upwind numerical flux; see, for example, [10]. We will compare the numerical solutions by both methods using the same setting except that $T = 100$ and the 4th order Runge–Kutta method is used for time stepping for the upwind DG. Figure 6.2 shows the numerical results. We see that both methods contain no spurious mode. For the

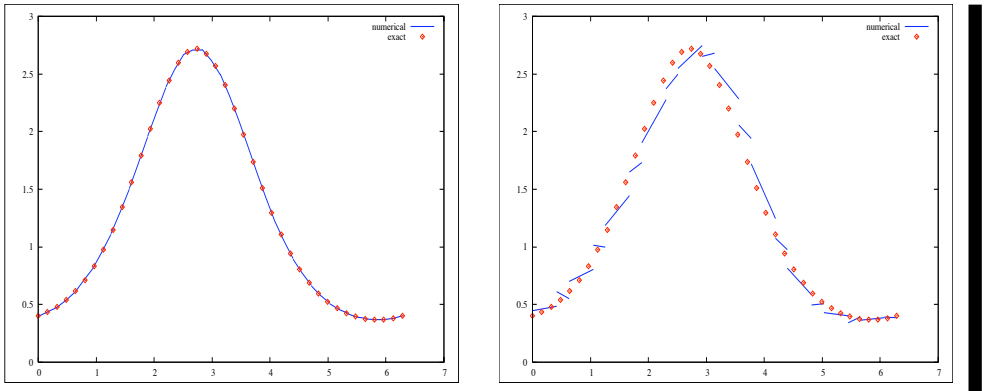


FIG. 6.1. *Comparison of the optimal DGM and central DGM.*

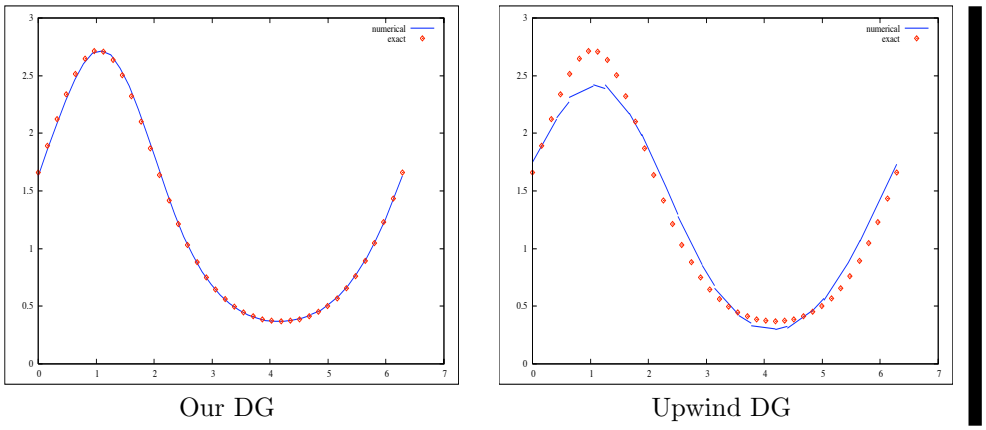


FIG. 6.2. *Comparison of the optimal DGM and upwind DGM.*

upwind DG, it is well known that it is dissipative as seen from the numerical result. On the contrary, our optimal DG preserves energy well.

Before ending this section, we will compare the upwind DG, central DG, and our new optimal DG with piecewise linear approximation. Due to the nature of the three schemes, they are all explicit and suitable for unstructured grids. Because of upwinding, the upwind DG is not energy preserving. In terms of the total number of degrees of freedom (DOF), both the upwind DG and the central DG need $4N$ unknowns. This is because there are 4 unknowns on each cell. On the contrary, owing to the extra continuity conditions, our new DG needs only $3N$ unknowns, which is more efficient in terms of memory storage. In addition, the central DG is only first order accurate, which is suboptimal since we are considering piecewise linear approximation. Both the upwind DG and our new DG have optimal order of convergence, namely, second order in the L^2 norm. We summarize all these properties in Table 6.5.

7. Conclusion. In this paper, we have developed and analyzed a new class of discontinuous Galerkin methods. This new DG can be seen as a compromise between

TABLE 6.5
Comparison among upwind, central, and our new DG with piecewise linear polynomials.

	Upwind	Central	Our
Explicit scheme	Y	Y	Y
Unstructured grid	Y	Y	Y
Energy conservation	N	Y	Y
DOF	$4N$	$4N$	$3N$
Order	$O(h^2)$	$O(h)$	$O(h^2)$

the standard DG and the FE in the sense that our new DG is explicit as standard DG and is energy conserving as FE. Energy conservation is an important property for a large class of applications that involves the numerical solutions of wave equations while explicitness provides a more efficient scheme where no matrix inversion is needed at each time step. We have shown that the new DG is stable in both the discrete L^2 norm and discrete H^1 norm. Moreover, the convergence rate is optimal with respect to the order of the polynomial space. To the best of our knowledge, our new DG is the first method that satisfies all of the following properties: explicit, energy conserving, suitable for unstructured grids, and optimal rate of convergence.

REFERENCES

[1] E. BÉCACHE, P. JOLY, AND C. TSOGKA, *An analysis of new mixed finite elements for the approximation of wave propagation problems*, SIAM J. Numer. Anal., 37 (2000), pp. 1053–1084.

[2] F. BREZZI AND M. FORTIN, *Mixed and hybrid finite element methods*, Springer-Verlag, New York, 1991.

[3] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland Publishing, Amsterdam, 1978.

[4] B. COCKBURN, F. LI, AND C.-W. SHU, *Local divergence-free discontinuous Galerkin methods for the Maxwell equations*, J. Comput. Phys., 194 (2004), pp. 588–610.

[5] G. COHEN, P. JOLY, N. TORJMAN, AND J. ROBERTS, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal., 38 (2001), pp. 2047–2078.

[6] G. COHEN AND P. MONK, *Gauss point mass lumping schemes for Maxwell's equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 63–88.

[7] L. FEZOUI, S. LANTERI, S. LOHRENGEL, AND S. PIPERNO, *Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes*, M2AN, Math. Model Numer. Anal., 39 (2005), pp. 1149–1176.

[8] T. GEVECI, *On the application of mixed finite element methods to the wave equations*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 243–250.

[9] M. GROTE, A. SCHNEEBELI, AND D. SCHÖTZAU, *Discontinuous Galerkin finite element method for the wave equation*, submitted.

[10] J. S. HESTHAVEN AND T. WARBURTON, *High-order nodal methods on unstructured grids, I. Time-domain solution of Maxwell's equations*, J. Comput. Phys., 181 (2002), pp. 186–221.

[11] P. JOLY, *Variational methods for time-dependent wave propagation problems*, Topics in Computational Wave Propagation, Lect. Notes Comput. Sci. Eng. 31, Springer-Verlag, Berlin, pp. 201–264.

[12] P. MONK, *A mixed method for approximating Maxwell's equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1610–1634.

[13] P. MONK, *An analysis of Nédélec's method for the spatial discretization of Maxwell's equations*, J. Comput. Appl. Math., 47 (1993), pp. 101–121.

[14] P. MONK AND G. R. RICHTER, *A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media*, J. Sci. Comput., 22 (2005), pp. 443–477.

[15] J. NÉDÉLEC, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[16] K. S. YEE, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas Propagat., 14 (1966), pp. 302–307.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2159–2197

ANALYSIS OF A STABILIZED FINITE ELEMENT APPROXIMATION OF THE TRANSIENT CONVECTION-DIFFUSION EQUATION USING AN ALE FRAMEWORK*

SANTIAGO BADIA† AND RAMON CODINA†

Abstract. In this paper we analyze a stabilized finite element method to approximate the convection-diffusion equation on moving domains using an arbitrary Lagrangian Eulerian (ALE) framework. As basic numerical strategy, we discretize the equation in time using first and second order backward differencing (BDF) schemes, whereas space is discretized using a stabilized finite element method (the *orthogonal subgrid scale* formulation) to deal with convection dominated flows. The semidiscrete problem (continuous in space) is first analyzed. In this situation it is easy to identify the error introduced by the ALE approach. After that, the fully discrete method is considered. We obtain optimal error estimates in both space and time in a mesh dependent norm. The analysis reveals that the ALE approach introduces an upper bound for the time step size for the results to hold. The results obtained for the fully discretized second order scheme (in time) are associated to a *weaker* norm than the one used for the first order method. Nevertheless, optimal convergence results have been proved. For fixed domains, we recover stability and convergence results with the strong norm for the second order scheme, stressing the aspects that make the analysis of this method much more involved.

Key words. stabilized finite elements, second order backward differencing, arbitrary Lagrangian Eulerian

AMS subject classifications. 65M12, 65M60

DOI. 10.1137/050643532

1. Introduction.

In this paper we propose and analyze two time integration schemes, of first and of second order, for the numerical approximation of the transient convection-diffusion equation in moving domains. This equation is written in an arbitrary Lagrangian Eulerian (ALE) framework, in which the temporal derivatives are expressed with respect to the reference of a moving domain Ω_t obtained from a mapping of the domain at the initial time. The space discretization is carried out using a stabilized finite element method (FEM) that allows us to deal with convection dominated flows.

The ALE framework, initially used with a finite element approximation in [14], has become widely popular when simulating fluid-structure interaction problems. Even though one can find a lot of numerical experimentation using the ALE approach, some aspects have remained in the dark for a long time. For instance, the meaning and effect of the *geometric conservation law* (GCL) and how the accuracy of a numerical method in fixed domains is spoiled when introducing moving domains with an ALE formulation were not clear. Farhat, Geuzaine, and Grandmont have shown in [15] that the GCL makes the numerical scheme preserve a maximum principle. In [18], the authors have shown that this condition is not necessary to obtain second order ALE

*Received by the editors October 26, 2005; accepted for publication (in revised form) May 26, 2006; published electronically November 3, 2006.

http://www.siam.org/journals/sinum/44-5/64353.html

†Universitat Politècnica de Catalunya, International Center for Numerical Methods in Engineering (CIMNE), Jordi Girona 1-3, Edifici C1, 08034 Barcelona, Spain (sbadia@cimne.upc.edu, ramon.codina@upc.edu). The first author's research was supported by the Departament d'Universitats, Recerca i Societat de la Informació of the Generalitat de Catalunya (Catalan Government) and the European Social Fund through a doctoral grant.

schemes in a finite volume framework. More recently, in a finite element setting where the transient convection-diffusion equation is taken as the *model* equation, works such as [16] and [25] have also clarified the effect of the GCL on the stability properties, identified the different behavior of conservative and nonconservative formulations, and proved some convergence results. Further analyses, for second order schemes, have been developed in later works, such as [17] and [3]. Herein we use the mathematical setting used, e.g., in [25] for the description of this method.

The ALE framework itself does not introduce any error at the continuous level. However, when the problem is discretized in time, some errors due to the ALE description arise. At this step, for fixed domains, the only source error is the time derivative of the unknown. In addition, for moving domains, the error from the evaluation of the mesh velocity also has to be accounted for. This velocity is calculated as the time derivative of the space position of a particle. Thus, an error is induced when this time derivative is calculated numerically.

On the other hand, in practical applications the mesh velocity belongs to the finite element space and does not introduce any interpolation error. Thus, we consider that the ALE formulation is better understood by analyzing the problem semidiscretized in time. However, most numerical analyses (see [16], [17], and [3]) first study the semidiscrete problem in space and then the fully discretized problem.

The convection-diffusion equation (as the Navier–Stokes equations) when discretized in space with the standard Galerkin formulation shows numerical oscillations if the convective term is dominant. With the aim of developing an FEM free of spurious oscillations many methods have been proposed during the last twenty years, such as streamline upwind/Petrov–Galerkin (SUPG) (see [6]), Galerkin/least-squares (see [24]), or the subgrid scale stabilization (see [22]). A comparison of different stabilization methods can be found in [7]. The *orthogonal subgrid scale* (OSS) method used in this paper belongs to this last family and was introduced by Codina in [8]. The method is designed by taking as starting point the subgrid scale variational setting proposed by Hughes et al. in [23] and modeling the subgrid problem in a certain way, in particular by taking the subgrid scales orthogonal to the finite element space. The common aspect of all of these methods is found in the convergence analysis of the discrete problem in space. For the Galerkin approximation, the error estimate bound depends on the physical properties (the Péclet number for the convection-diffusion equation) and increases as the convective term is more dominant. In fact, the stability bound blows up as diffusion goes to zero, reflecting the fact that the continuous problem is a singularly perturbed one. But when using stabilized methods this negative feature does not appear anymore. This is because the new terms introduced by the stabilization control the convective term norm. In the present analysis we have been able to obtain appropriate error estimates by controlling only a part of the convective term, which is an innovative result.

As far as we know, most of the existing stabilization techniques are extended to transient problems using the framework of the discontinuous Galerkin space-time formulation, increasing notably the computer cost for schemes in time of order two or higher. This situation has been improved by Guermond in [20], where he analyzes the introduction of a certain numerical subgrid viscosity. Optimal convergence results are obtained for an evolutionary equation. The key point is the uncoupling of the stabilization terms with the temporal derivative of the unknown. Another stabilization method with this feature is presented in [4].

Codina and Blasco analyze in [12] the transient convection-diffusion-reaction equation discretized in space using the OSS method and in time with the backward

Euler time integration. Further, they consider the *tracking* of the subscales in time. Optimal convergence and stability results are obtained.

The present paper can be viewed as an extension of [12]. We generalize the situation to moving domains (using an ALE approach). In addition, first and second order backward differencing (BDF) time integration schemes are considered, which will be denoted by BDF1 and BDF2, respectively. The blend of a stabilized FEM with the use of an ALE framework is one of the innovative aspects of this paper.

In order to analyze the stabilized method for transient problems, the following strategy is adopted in [12]: First the semidiscrete problem is studied (where no stabilization terms appear) and later the fully discrete method is analyzed. As shown in [12], this provides a natural way to deal with the subscales whose approximation enhances the stability and accuracy of the formulation. The main drawback of this strategy is that space regularity for the convergence analysis needs to be assumed for the semidiscrete solution, not for the continuous one.

The first time integration scheme considered uses the classical backward Euler formula for the approximation of both the time derivative of the unknown and the calculation of the mesh velocity. We label this method as follows:

- BDF1-BDF1$_{\delta t}$ for the problem semidiscretized in time,
- BDF1-BDF1$_{\delta t,h}$ for the fully discretized problem using the classical Galerkin approximation in space,
- BDF1-BDF1-OSS$_{\delta t,h}$ for the fully discretized problem using the OSS method in space,
- BDF1-OSS$_{\delta t,h}$ for the fully discretized problem using the OSS method in space on fixed domains (not in an ALE framework).

In the second method the time integration makes use of the second order BDF formula. Again, we use the following notation:

- BDF2-BDF2$_{\delta t}$ for the problem semidiscretized in time,
- BDF2-BDF2$_{\delta t,h}$ for the fully discretized problem using the classical Galerkin approximation in space,
- BDF2-BDF2-OSS$_{\delta t,h}$ for the fully discretized problem using the OSS method in space,
- BDF2-OSS$_{\delta t,h}$ for the fully discretized problem using the OSS method in space on fixed domains (not in an ALE framework).

Let us underline what is new in each case. The BDF1-BDF1$_{\delta t,h}$ method has been analyzed in [16]. As explained above, we change the order of the discretization: First we analyze BDF1-BDF1$_{\delta t}$ and then BDF1-BDF1-OSS$_{\delta t,h}$, introducing the appropriate stabilization terms. For fixed domains, BDF1-OSS$_{\delta t,h}$ has been analyzed in [12]. However, the analysis herein is slightly different. The analysis of convergence and stability of the semidiscrete method BDF2-BDF2$_{\delta t}$ is new, as it is for the method's fully discrete stabilized version BDF2-BDF2-OSS$_{\delta t,h}$. We specially note the fact that convergence results independent of the physical properties can be obtained without the full norm of the convective term. Even for fixed domains, the stability and convergence results for BDF2-OSS$_{\delta t,h}$ are new. In all cases the long-term behavior has been considered.

Numerical experimentation with the ALE methods (for diffusion dominated problems using the Galerkin method) BDF1-BDF1$_{\delta t,h}$ and BDF2-BDF2$_{\delta t,h}$ can be found in [17], [3], and [25], showing the expected behavior. The application of BDF1-OSS$_{\delta t,h}$ and BDF2-OSS$_{\delta t,h}$ can be found in [9] and [11] for the solution of fluid problems. Finally, the blend of these methods, BDF1-BDF1-OSS$_{\delta t,h}$ and BDF2-BDF2-OSS$_{\delta t,h}$, has been used for simulating engineering problems in [1], with excellent results.

TABLE 1.1
List of main results.

Method	Main result	Label
BDF1-BDF1$_{\delta t}$	Coercivity	Theorem 3.1
	Conditional stability	Corollary 3.3
	Convergence	Theorem 3.4
BDF2-BDF2$_{\delta t}$	Coercivity	Theorem 3.6
	Conditional stability	Corollary 3.7
	Convergence	Theorem 3.8
BDF1-BDF1-OSS$_{\delta t,h}$	*Weak* coercivity	Theorem 4.2
	Strong inf-sup	Corollary 4.7
	Strong conditional stability	Corollary 4.8
	Strong convergence	Theorem 4.11
BDF2-BDF2-OSS$_{\delta t,h}$	*Weak* coercivity	Theorem 4.12
	Weak conditional stability	Corollary 4.13
	Weak convergence	Theorem 4.17
BDF2-OSS$_{\delta t,h}$	Results of BDF2-BDF2-OSS$_{\delta t,h}$ +	
	Strong Λ-coercivity	Theorem 4.20
	Strong stability	Corollary 4.22
	Strong convergence	Theorem 4.25

The paper is organized as follows. In section 2 we state the governing equations for moving domains in an ALE framework. Some important ingredients needed to define the ALE approach are introduced. The semidiscrete problem is formulated for both BDF1 and BDF2. The section ends with the presentation of the OSS stabilization method and the fully discrete problem. Section 3 is devoted to the semidiscrete problem. First and second order methods are considered, for which stability and optimal convergence estimates are obtained. Section 4 presents an analogous analysis to that of section 3 but for the fully discrete problem. Finally, some conclusions are drawn in section 5.

In Table 1.1 we have summarized the main results proved in this paper in order to provide the reader with a *road map* for the subsequent discussion. The concepts used in this table (*weak*, *strong*, and Λ-coercivity) will be introduced later.

2. Problem statement.

2.1. The continuous problem. In order to study the ALE framework together with a stabilized FEM, we take as a model test problem the transient convection-diffusion equation. The problem written in an Eulerian framework consists in finding a function u such that

$$\frac{\partial u}{\partial t} - \nu\Delta u + \boldsymbol{a}\cdot\nabla u = f \quad \text{in } \Omega_t \times (0,T), \tag{2.1a}$$

$$u = 0 \quad \text{on } \partial\Omega_t \times (0,T), \tag{2.1b}$$

$$u(\boldsymbol{x}_0, 0) = u_0 \quad \text{in } \Omega_0 \times \{0\}, \tag{2.1c}$$

where $\Omega_t \subset \mathbb{R}^d$ (d=2,3) is a bounded and polyhedral domain (moving in time), $[0,T]$ is the time interval of analysis, $\boldsymbol{a}$ is a divergence-free velocity field, and $\nu > 0$ is the diffusion coefficient. Homogeneous boundary conditions are assumed to clarify the analysis. We also assume the following regularity of the data:

$$f \in L^2(0,T;H^{-1}(\Omega_t)), \quad u_0 \in L^2(\Omega_0), \quad \boldsymbol{a} \in \boldsymbol{L}^\infty(\Omega_t),$$

assuring the existence of a unique solution $u(t) \in L^2(0,T;H^1(\Omega_t)) \cap \mathcal{C}^0(0,T;L^2(\Omega_t))$.

We introduce some key ingredients of an ALE framework. Let $\mathcal{A}_t$ be a family of mappings, which for all $t \in [0,T]$ map a point $\boldsymbol{x}_0 \in \Omega_0$ into a point $\boldsymbol{x} \in \Omega_t$:

$$\mathcal{A}_t : \Omega_0 \longrightarrow \Omega_t, \quad \boldsymbol{x}(\boldsymbol{x}_0, t) = \mathcal{A}_t(\boldsymbol{x}_0).$$

We assume that $\mathcal{A}_t$ is invertible with inverse $\mathcal{A}_t^{-1}$. For $t_1, t_2 \in [0,T]$ we define

$$\mathcal{A}_{t_1,t_2} : \Omega_{t_1} \longrightarrow \Omega_{t_2}, \quad \mathcal{A}_{t_1,t_2} = \mathcal{A}_{t_2} \circ \mathcal{A}_{t_1}^{-1}.$$

We note that the family of mappings is arbitrary. Several techniques have been suggested in order to construct this ALE mapping. If $\mathcal{A}_t$ is the mapping arising from the motion of the particles, the resulting formulation would be of pure Lagrangian type.

Let us consider a function $f : \Omega_t \times [0,T] \longrightarrow \mathbb{R}$. We indicate with $\hat{f} = f \circ \mathcal{A}_t$ the corresponding function in the ALE frame:

$$\hat{f} : \Omega_0 \times [0,T] \longrightarrow \mathbb{R}, \quad \hat{f}(\boldsymbol{x}_0, t) = f(\mathcal{A}_t(\boldsymbol{x}_0), t).$$

Furthermore, the time derivatives in the ALE frame are defined as follows:

$$\left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{x}_0} : \Omega_t \times [0,T] \longrightarrow \mathbb{R}, \quad \left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{x}_0}(\boldsymbol{x}, t) = \frac{\partial \hat{f}}{\partial t}(\boldsymbol{x}_0, t).$$

The domain velocity $\boldsymbol{w}$ is calculated using the expression

$$\boldsymbol{w}(\boldsymbol{x}, t) = \left.\frac{\partial \boldsymbol{x}}{\partial t}\right|_{\boldsymbol{x}_0} = \frac{\partial \mathcal{A}_t(\boldsymbol{x}_0)}{\partial t},$$

and the Jacobian of the ALE mapping is given by

$$J_t = \det(\boldsymbol{J}_t), \quad \boldsymbol{J}_t = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}_0}.$$

We recall the *Reynolds transport formula*. Let $\psi(\boldsymbol{x}, t)$ be a function defined in Ω_t. Then, for any subdomain $V_t \subseteq \Omega_t$ such that $V_t = \mathcal{A}_t(V_0)$ with $V_0 \subseteq \Omega_0$, it holds that

$$\frac{d}{dt}\int_{V_t} \psi(\boldsymbol{x}, t)\, \mathrm{d}V = \int_{V_t} \left(\left.\frac{\partial \psi}{\partial t}\right|_{\boldsymbol{x}_0} + \psi \nabla \cdot \boldsymbol{w} \right) \mathrm{d}V.$$

In particular, if $v : \Omega_t \longrightarrow \mathbb{R}$, that is, if v does not depend explicitly on time, we have that

$$\frac{d}{dt}\int_{\Omega_t} v\, \mathrm{d}\Omega = \int_{\Omega_t} v \nabla \cdot \boldsymbol{w}\, \mathrm{d}\Omega. \tag{2.2}$$

With all this notation introduced, we are ready to write (2.1) in the ALE framework. It now reads

$$\left.\frac{\partial u}{\partial t}\right|_{\boldsymbol{x}_0} - \nu \Delta u + (\boldsymbol{a} - \boldsymbol{w}) \cdot \nabla u = \boldsymbol{f} \quad \text{in } \Omega_t \times (0,T), \tag{2.3a}$$

$$u = 0 \quad \text{on } \partial\Omega_t \times (0,T), \tag{2.3b}$$

$$\boldsymbol{u}(\boldsymbol{x}_0, 0) = \boldsymbol{u}_0 \quad \text{in } \Omega_0 \times \{0\}. \tag{2.3c}$$

The functional space

$$\mathcal{V}(\Omega_t) := \left\{ v : \Omega_t \to \mathbb{R}, v = \hat{v} \circ \mathcal{A}_t^{-1}, \hat{v} \in H_0^1(\Omega_0) \right\}, \quad t \in (0,T),$$

allows us to write (2.3) in its variational form. The variational problem reads as follows: find $u(t) \in \mathcal{V}(\Omega_t)$ for all $t \in (0,T)$ such that

$$\left(\frac{\partial u(t)}{\partial t}, v \right)_{\Omega_t} + \nu \left(\nabla u(t), \nabla v \right)_{\Omega_t} + \left((\boldsymbol{a} - \boldsymbol{w}(t)) \cdot \nabla u(t), v \right)_{\Omega_t} = \langle f(t), v \rangle_{\Omega_t}, \tag{2.4}$$

for all $v \in \mathcal{V}(\Omega_t)$, where $(\cdot,\cdot)_{\Omega_t}$ stands for the $L^2(\Omega_t)$ inner product and $\langle \cdot, \cdot \rangle_{\Omega_t}$ for the duality pairing in $H^{-1}(\Omega_t) \times H_0^1(\Omega_t)$.

Let us rescale the time variable as $t \leftarrow t/T$ so that the new time interval is $[0,1]$ and the coefficient $1/T$ has to be inserted in front of the time derivatives. The reason for this change is to display which terms in the stability and convergence results disappear as $T \to \infty$, that is, the long-term behavior. After rescaling, problem (2.4) is transformed into

$$\frac{1}{T} \left(\frac{\partial u(t)}{\partial t}, v \right)_{\Omega_t} + \nu \left(\nabla u(t), \nabla v \right)_{\Omega_t} + \left((\boldsymbol{a} - \boldsymbol{w}(t)) \cdot \nabla u(t), v \right)_{\Omega_t} = \langle f(t), v \rangle_{\Omega_t}, \tag{2.5}$$

and now the domain velocity is

$$\boldsymbol{w}(\boldsymbol{x}, t) = \frac{1}{T} \left. \frac{\partial \boldsymbol{x}}{\partial t} \right|_{\boldsymbol{x}_0}. \tag{2.6}$$

We take into account this rescaling in property (2.2), which now reads

$$\frac{1}{T} \frac{d}{dt} \int_{\Omega_t} v \, \mathrm{d}\Omega = \int_{\Omega_t} v \nabla \cdot \boldsymbol{w} \, \mathrm{d}\Omega. \tag{2.7}$$

2.2. The semidiscrete problem in time. Let us introduce some notation that we will use throughout the work. Consider a uniform partition of $[0,1]$ into N time intervals of length δt. Let us denote by f^n the approximation of a time dependent function f at time level $t^n = n\delta t$. We will also denote

$$\begin{aligned} \delta f^{n+1} &\equiv \delta^{(1)} f^{n+1} = f^{n+1} - f^n, \\ \delta^{(i+1)} f^{n+1} &= \delta^{(i)} f^{n+1} - \delta^{(i)} f^n, \ i = 1, 2, 3, \ldots . \end{aligned}$$

The discrete operators $\delta^{(i+1)}$ are centered. We will also use the backward difference operators

$$\begin{aligned} D_1 f^{n+1} &= \frac{\delta f^{n+1}}{\delta t} = \frac{f^{n+1} - f^n}{\delta t}, \\ D_2 f^{n+1} &= \frac{3}{2\delta t} \left(f^{n+1} - \frac{4}{3} f^n + \frac{1}{3} f^{n-1} \right). \end{aligned}$$

Let us discretize problem (2.5) in time, once t has been normalized. We assume the force term is continuous in time and denote the time level by a superscript. We start using the BDF1 time integration scheme. It leads to the following problem: for $n = 0, 1, \ldots, N-1$, given u^n, find $u^{n+1} \in \mathcal{V}(\Omega_{t^{n+1}})$ such that

$$\begin{aligned} &\frac{1}{T} \left(u^{n+1} - u^n, v^{n+1} \right)_{\Omega_{t^{n+1}}} + \delta t \nu \left(\nabla u^{n+1}, \nabla v^{n+1} \right)_{\Omega_{t^{n+1}}} \\ &\quad + \delta t \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla u^{n+1}, v^{n+1} \right)_{\Omega_{t^{n+1}}} = \delta t \langle f^{n+1}, v^{n+1} \rangle_{\Omega_{t^{n+1}}}, \end{aligned} \tag{2.8}$$

with $u^0 = u_0$ in $L^2(\Omega_0)$.

Furthermore, we discretize in time the ALE mapping using a linear interpolation. The discretized ALE mapping $\mathcal{A}_t^{n+1}$ is defined for a given time slab $[t^n, t^{n+1}]$ as

$$\mathcal{A}_t^{n+1}(\boldsymbol{x}_0, t) = \frac{t - t^n}{\delta t}\mathcal{A}_{t^{n+1}}(\boldsymbol{x}_0) + \frac{t^{n+1} - t}{\delta t}\mathcal{A}_{t^n}(\boldsymbol{x}_0).$$

Thus, the mesh velocity is constant on each time step and is given by

$$\hat{\boldsymbol{w}}^{n+1}(\boldsymbol{x}_0) = \frac{\mathcal{A}_{t^{n+1}}(\boldsymbol{x}_0) - \mathcal{A}_{t^n}(\boldsymbol{x}_0)}{T\delta t}$$

and $\boldsymbol{w}^{n+1}(\boldsymbol{x}, t) = \hat{\boldsymbol{w}}^{n+1}((\mathcal{A}_t^{n+1})^{-1}(\boldsymbol{x}))$ for $t \in (t^n, t^{n+1}]$. Equation (2.8) with this mesh velocity defines the BDF1-BDF1$_{\delta t}$ method. Note that the superscript $n+1$ in $\boldsymbol{w}$ denotes that it varies with time within the time interval $(t^n, t^{n+1}]$ where it is defined. However, in section 3 we will simply denote $\boldsymbol{w}^{n+1} \equiv \boldsymbol{w}^{n+1}(\boldsymbol{x}, t^{n+1})$. Since $\mathcal{A}_{t^{n+1}}^{n+1} = \mathcal{A}_{t^{n+1}}$, we will write $\boldsymbol{w}^{n+1}(\boldsymbol{x}, t^{n+1}) = \hat{\boldsymbol{w}}^{n+1}(\mathcal{A}_{t^{n+1}}^{-1}(\boldsymbol{x}))$ or, for $\boldsymbol{x}$ arbitrary, $\boldsymbol{w}^{n+1} = \hat{\boldsymbol{w}}^{n+1} \circ \mathcal{A}_{t^{n+1}}^{-1}$.

For the numerical analysis we rewrite the transient problem using a different setting. The sequence of problems (2.8) can be written in a unified manner as follows: find a sequence $U = \{u^0, u^1, u^2, \ldots, u^N\}$ such that

$$B(U, V) = L(V) \tag{2.9}$$

for all sequences V, where

$$\begin{aligned} B(U,V) := \frac{1}{2T}\left(u^0, v^0\right)_{\Omega_0} + \sum_{n=0}^{N-1}\Bigg[\frac{1}{T}\left(\delta u^{n+1}, v^{n+1}\right)_{\Omega_{t^{n+1}}} + \delta t\nu\left(\nabla u^{n+1}, \nabla v^{n+1}\right)_{\Omega_{t^{n+1}}} \\ + \delta t\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}, v^{n+1}\right)_{\Omega_{t^{n+1}}}\Bigg], \end{aligned} \tag{2.10}$$

$$L(V) := \frac{1}{2T}\left(u^0, v^0\right)_{\Omega_0} + \sum_{n=0}^{N-1}\delta t\langle f^{n+1}, v^{n+1}\rangle_{\Omega_{t^{n+1}}}. \tag{2.11}$$

Observe that the initial condition has been embedded in the variational problem.

In order to reach second order accuracy in time, the BDF2 integration scheme is used. It leads to the following time discretization of (2.5):

$$\begin{aligned} &\frac{1}{2T}\left(3u^{n+1} - 4u^n + u^{n-1}, v^{n+1}\right)_{\Omega_{t^{n+1}}} + \delta t\nu\left(\nabla u^{n+1}, \nabla v^{n+1}\right)_{\Omega_{t^{n+1}}} \\ &\quad + \delta t\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}, v^{n+1}\right)_{\Omega_{t^{n+1}}} = \delta t\langle f^{n+1}, v^{n+1}\rangle_{\Omega_{t^{n+1}}}. \end{aligned} \tag{2.12}$$

This problem has to be initialized. For instance, we can obtain u^1 with (2.8) and $u^0 = u_0$ in $L^2(\Omega_0)$ keeping the order of convergence of the method. In order to keep this accuracy, a quadratic interpolation is used to approximate the ALE mapping. For a given time slab $[t^n, t^{n+1}]$, this interpolation is given by

$$\begin{aligned} \mathcal{A}_t^{n+1}(\boldsymbol{x}_0, t) = &\frac{(t - t^n)(t - t^{n-1})}{2\delta t^2}\mathcal{A}_{t^{n+1}}(\boldsymbol{x}_0) \\ &- \frac{(t - t^{n+1})(t - t^{n-1})}{2\delta t^2}\mathcal{A}_{t^n}(\boldsymbol{x}_0) + \frac{(t - t^{n+1})(t - t^n)}{2\delta t^2}\mathcal{A}_{t^{n-1}}(\boldsymbol{x}_0). \end{aligned}$$

Thus, the mesh velocity on each time step is linear in time and is given by

$$\hat{\boldsymbol{w}}^{n+1}(\boldsymbol{x}_0,t) = \frac{2t-t^n-t^{n-1}}{2T\delta t^2}\mathcal{A}_{t^{n+1}}(\boldsymbol{x}_0) - \frac{2t-t^{n+1}-t^{n-1}}{2T\delta t^2}\mathcal{A}_{t^n}(\boldsymbol{x}_0) + \frac{2t-t^{n+1}-t^n}{2T\delta t^2}\mathcal{A}_{t^{n-1}}(\boldsymbol{x}_0)$$

and $\boldsymbol{w}^{n+1}(\boldsymbol{x},t) = \hat{\boldsymbol{w}}^{n+1}((\mathcal{A}_t^{n+1})^{-1}(\boldsymbol{x}),t)$ for $t \in (t^n, t^{n+1}]$. It is easily checked that at t^{n+1} we recover the BDF2 formula for the mesh velocity.

Again, we can rewrite the transient problem as an abstract "variational" problem (2.9), now with the bilinear form

$$\begin{aligned} B(U,V) = {} & \frac{1}{T}\left(u^1-u^0,v^1\right)_{\Omega_{t^1}} + \delta t\nu\left(\nabla u^1,\nabla v^1\right)_{\Omega_{t^1}} + \delta t\left((\boldsymbol{a}-\boldsymbol{w}^1)\cdot\nabla u^1,v^1\right)_{\Omega_{t^1}} \\ & + \frac{1}{2T}\left(u^0,v^0\right)_{\Omega_0} + \sum_{n=1}^{N-1}\left[\frac{1}{2T}\left(3u^{n+1}-4u^n+u^{n-1},v^{n+1}\right)_{\Omega_{t^{n+1}}}\right. \\ & \left. + \delta t\nu\left(\nabla u^{n+1},\nabla v^{n+1}\right)_{\Omega_{t^{n+1}}} + \delta t\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1},v^{n+1}\right)_{\Omega_{t^{n+1}}}\right] \end{aligned} \tag{2.13}$$

and the linear form

$$L(V) := \frac{1}{2T}\left(u^0,v^0\right)_{\Omega_0} + \sum_{n=0}^{N-1}\delta t\langle f^{n+1},v^{n+1}\rangle_{\Omega_{t^{n+1}}}. \tag{2.14}$$

We end this subsection by giving the norm for which stability and convergence results are obtained in section 3 for the previous semidiscrete problems:

$$|||V|||^2 = \frac{1}{T}\sup_{n\in[0,N]}\|v^n\|^2_{L^2(\Omega_{t^n})} + \sum_{n=0}^{N-1}\delta t\nu\left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}. \tag{2.15}$$

Given a normed space X, for $1 \le q < \infty$ we define the space $\ell^q(X)$ as that of sequences $V = \{v^n\}_{n=0}^N$ such that $\sum_{n=0}^N \delta t\|v^n\|_X^q < \infty$, and $\ell^\infty(X)$ the space of sequences such that $\sup_{n=0,\dots,N}\|v^n\|_X < \infty$. With this notation, the norm defined in (2.15) can be considered that of $\ell^\infty(L^2(\Omega_t)) \cap \ell^2(H_0^1(\Omega_t))$. Here, the subscript t has to be understood as t^n for the nth component of the sequence.

2.3. The fully discrete problem. At this point we treat the space discretization of systems (2.8) and (2.12). The BDF1-BDF1-OSS$_{\delta t,h}$ reads as follows: for $n = 0, 1, \dots, N-1$, given u_h^n, find $u_h^{n+1} \in \mathcal{V}_h(\Omega_t)$ such that

$$\begin{aligned} & \frac{1}{T}\left(u_h^{n+1}-u_h^n,v_h^{n+1}\right)_{\Omega_{t^{n+1}}} + \delta t\nu\left(\nabla u_h^{n+1},\nabla v_h^{n+1}\right)_{\Omega_{t^{n+1}}} \\ & \quad + \delta t\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u_h^{n+1},v_h^{n+1}\right)_{\Omega_{t^{n+1}}} \\ & \quad + \delta t\left(\Pi_h^\perp\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u_h^{n+1}\right),\tau^{n+1}(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla v_h^{n+1}\right)_{\Omega_{t^{n+1}}} \\ & \quad = \delta t\langle f^{n+1},v_h^{n+1}\rangle_{\Omega_{t^{n+1}}}, \end{aligned} \tag{2.16}$$

where $\mathcal{V}_h(\Omega_t)$ is a finite element approximation space of $\mathcal{V}(\Omega_t)$, τ^{n+1} is a mesh dependent parameter, which we will call the *stabilization parameter*, whose expression is detailed later, and $\Pi_h^\perp(\cdot) =: Id(\cdot) - \Pi_h(\cdot)$, with Id the identity in $L^2(\Omega_t)$ and $\Pi_h(\cdot)$ the L^2-projection onto this finite element space (and therefore $\Pi_h^\perp(\cdot)$ is the

projection orthogonal to the finite element space). The description and motivation of this formulation, which we call OSS stabilization, can be found in [10].

Let Θ_h^t be a finite element partition of the domain Ω_t in a family of elements $\{K_e\}_{e=1}^{n_{el}}$, n_{el} being the number of elements. We denote the diameter of the sphere that circumscribes element K by h_K and the diameter of the sphere inscribed in K by ϱ_K. We also call $h = \max_{K\in\Theta_h^t}(h_K)$ and $\varrho = \min_{K\in\Theta_h^t}(\varrho_K)$. We assume that all the element domains $K \in \Theta_h^t$ are the image of a reference element $\tilde{K}$ through polynomial mappings F_K, affine for simplicial elements, bilinear for quadrilaterals, and trilinear for hexahedra. On $\tilde{K}$ we define the polynomial spaces $R_p(\tilde{K})$, where R_p is, for simplicial elements, the set of polynomials in $x_1, \ldots, x_d$ of degree less than or equal to p, called P_p. For quadrilaterals and hexahedra, R_p consists of polynomials in $x_1, \ldots, x_d$ of degree less than or equal to p in each variable, a set called Q_p. The finite element spaces introduced before and that we will use in the following are

$$\begin{aligned}
\mathcal{V}_h^f(\Omega_0) &= \{\widehat{v}_h \in \mathcal{C}^0(\Omega_0) \mid \widehat{v}_h|_K = \tilde{v} \circ F_K^{-1},\ \tilde{v} \in R_p(\tilde{K}),\ K \in \Theta_h^t\},\\
\mathcal{V}_h(\Omega_0) &= \{v_h \in \mathcal{V}_h(\Omega_0) \mid v_h|_{\partial\Omega_0} = 0\},\\
\mathcal{V}_h^f(\Omega_t) &= \{v_h \in \mathcal{C}^0(\Omega_t) \mid v_h = \widehat{v}_h \circ \mathcal{A}_t^{-1},\ \widehat{v}_h \in \mathcal{V}_h(\Omega_0)\},\\
\mathcal{V}_h(\Omega_t) &= \{v_h \in \mathcal{C}^0(\Omega_t) \mid v_h = \widehat{v}_h \circ \mathcal{A}_t^{-1},\ \widehat{v}_h \in \mathcal{V}_{h,0}(\Omega_0)\}.
\end{aligned}$$

Moreover, Θ_h^t is assumed to be quasi-uniform; that is to say, there exists a constant $\varrho_2 > 0$, independent of h, such that $\frac{\varrho}{h} \geq \varrho_2 > 0$ as h tends to zero. This will simplify the analysis and, in particular, will allow us to use stabilization parameters constant in space.

Let us note that in practical applications $\mathcal{A}_{t^{n+1}}$ maps Θ_h^0 onto Θ_h^{n+1}. Therefore, it is easily checked that $\boldsymbol{w}^{n+1} \in (\mathcal{V}_h(\Omega_{t^{n+1}}))^d$. In the following we will not distinguish between $\boldsymbol{w}^{n+1}$ and $\boldsymbol{w}_h^{n+1}$.

Also in this case we can write the problem using a "variational" formalism. The fully discrete sequence of problems given by (2.16) can be written as follows: find a sequence $U_h = \{u_h^0, u_h^1, \ldots, u_h^N\}$ such that

$$B_h(U_h, V_h) = L(V_h) \tag{2.17}$$

for all sequences V_h, with the bilinear form B_h given by

$$\begin{aligned}
B_h(U_h, V_h) &= \frac{1}{2T}\left(u_h^0, v_h^0\right)_{\Omega_0}\\
&\quad + \sum_{n=0}^{N-1}\left[\frac{1}{T}\left(u_h^{n+1} - u_h^n, v_h^{n+1}\right)_{\Omega_{t^{n+1}}} + b_h\left(\boldsymbol{w}^{n+1}; u_h^{n+1}, v_h^{n+1}\right)_{\Omega_{t^{n+1}}}\right],
\end{aligned} \tag{2.18}$$

where b_h is defined as

$$\begin{aligned}
b_h\left(\boldsymbol{w}^{n+1}; u_h^{n+1}, v_h^{n+1}\right)_{\Omega_{t^{n+1}}} &= \delta t\nu\left(\nabla u_h^{n+1}, \nabla v_h^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&\quad + \delta t\left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla u_h^{n+1}, v_h^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&\quad + \delta t\left(\Pi_h^{\perp}\left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla u_h^{n+1}\right), \tau^{n+1}(\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v_h^{n+1}\right)_{\Omega_{t^{n+1}}}.
\end{aligned} \tag{2.19}$$

The OSS method modifies the discretized equation of the classical Galerkin method by introducing the last term, which enhances the stability of the original method. The value of the stabilization parameter τ^{n+1} has been justified in [10]. In an ALE

framework it depends on the difference between the advection velocity $\boldsymbol{a}$ and the mesh velocity $\boldsymbol{w}$. The expression we use is

$$\tau^{n+1} = \left(c_1 \frac{\nu}{h^2} + c_2 \frac{\|\boldsymbol{a} - \boldsymbol{w}\|_{L^\infty(\Omega_{t^{n+1}})}}{h} \right)^{-1}, \tag{2.20}$$

which is constant in space. Here, c_1 and c_2 are algorithmic constants that depend on the order of the finite element interpolation. As will be shown later (see (4.7)), they are related to the constant C_{inv} in the inverse estimate introduced in (4.1).

As in [12], we will make further assumptions. We assume that for each n the parameter τ^n satisfies

$$\tau^n \leq CT\delta t, \tag{2.21}$$

which in particular implies that we cannot let $\delta t \to 0$ without refining the finite element mesh. This condition is not only theoretical, but probably has practical consequences. It is shown in [2] in a particular numerical example that instabilities occur in the case of the transient Stokes problem if a condition similar to (2.21) is violated. Moreover, from the theoretical point of view there is a way to circumvent this, which consists in considering the subscales time dependent. This is the approach followed in [13], where stability of a stabilized FEM for the linearized Navier–Stokes equations is proved with and without condition (2.21).

For the space discretization of the second order method (2.12), the bilinear form is given by

(2.22)

$$\begin{aligned} &B_h\left(U_h, V_h\right) \\ &\quad = \sum_{n=1}^{N-1} \left[\frac{1}{2T} \left(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, v_h^{n+1}\right)_{\Omega_{t^{n+1}}} + b_h\left(\boldsymbol{w}^{n+1}; u_h^{n+1}, v_h^{n+1}\right)_{\Omega_{t^{n+1}}} \right] \\ &\qquad + \frac{1}{T}\left(u_h^1 - u_h^0, v_h^1\right)_{\Omega_{t^1}} + b_h\left(\boldsymbol{w}^1; u_h^1, v_h^1\right)_{\Omega_{t^1}} + \frac{1}{T}\left(u_h^0, v_h^0\right)_{\Omega_0}. \end{aligned}$$

We end this section with two norms that are useful in the following numerical analysis. The first is a norm that we will call *weak*, which is given by

$$\begin{aligned} |||V|||_w^2 = \frac{1}{T} \sup_{n\in[0,N]} \|v^n\|^2_{L^2(\Omega_{t^n})} &+ \sum_{n=0}^{N-1} \delta t\nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \\ &+ \sum_{n=0}^{N-1} \delta t\tau^{n+1} \left\|\Pi_h^\perp\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}. \end{aligned}$$

Observe that only the orthogonal projection of the convective term appears. The full convective term appears in the norm that we will call *strong*, given by

$$\begin{aligned} |||V|||_s^2 &= \frac{1}{T} \sup_{n\in[0,N]} \|v^n\|^2_{L^2(\Omega_{t^n})} + \sum_{n=0}^{N-1} \delta t\nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \\ &\quad + \sum_{n=0}^{N-1} \delta t\tau^{n+1} \left\|(\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \\ &= |||V|||_w^2 + \sum_{n=0}^{N-1} \delta t\tau^{n+1} \left\|\Pi_h\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}. \end{aligned}$$

3. Analysis of the semidiscrete problem. In this section we analyze problems BDF1-BDF1$_{\delta t}$ and BDF2-BDF2$_{\delta t}$. In both cases, stability and error estimates will be given. We denote by C a positive constant, possibly with different values at different appearances.

3.1. Analysis of BDF1-BDF1$_{\delta t}$. Let us define by $U_{\text{ex}} = \{u_0, u(t^1), u(t^2), \ldots, u(t^N)\}$ the sequence of solutions of the continuous problem (2.4) and by $U = \{u^0, u^1, u^2, \ldots, u^N\}$ the sequence of solutions of the semidiscrete problem (in time) (2.9)–(2.11). We start by obtaining a stability result for this method. With this aim, first we prove that the bilinear form (2.10) that governs the semidiscrete problem is coercive.

THEOREM 3.1 (coercivity). *There exists δt^1_{cr} such that for $0 < \delta t < \delta t^1_{\text{cr}}$ the bilinear form $B(\cdot,\cdot)$ defined in* (2.10) *is coercive; that is, for every sequence $V = \{v^n\}_{n=0}^N$, with $v^n \in \mathcal{V}(\Omega_{t^n})$,*

$$B(V,V) \geq \beta_1 |||V|||^2$$

for a certain constant $\beta_1 > 0$.

Proof. We know, from the definition of the bilinear form, that

$$\begin{aligned} B(V,V) = \sum_{n=0}^{N-1} \Bigg[& \frac{1}{T} \left(v^{n+1} - v^n, v^{n+1}\right)_{\Omega_{t^{n+1}}} + \delta t \nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \\ & + \delta t \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1}, v^{n+1}\right)_{\Omega_{t^{n+1}}} \Bigg] \\ & + \frac{1}{2T} \left\|v^0\right\|^2_{L^2(\Omega_0)}. \end{aligned}$$

We can rewrite the term coming from the time derivative as follows:

$$\begin{aligned} & \frac{1}{T} \left(v^{n+1} - v^n, v^{n+1}\right)_{\Omega_{t^{n+1}}} \\ & \quad = \frac{1}{2T} \left[\left\|v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} - \left\|v^n\right\|^2_{L^2(\Omega_{t^{n+1}})} + \left\|v^{n+1} - v^n\right\|^2_{L^2(\Omega_{t^{n+1}})} \right]. \end{aligned}$$

Integrating (2.7) from t^n to t^{n+1} for the function $(v^n)^2$, we get

$$\frac{1}{T} \left\|v^n\right\|^2_{L^2(\Omega_{t^{n+1}})} = \frac{1}{T} \left\|v^n\right\|^2_{L^2(\Omega_{t^n})} + \int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1})(v^n)^2 \, \mathrm{d}\Omega \, \mathrm{d}s,$$

where we have profited from the fact that the discrete mesh velocity is constant at every time step. On the other hand, due to the fact that the convective velocity $\boldsymbol{a}$ is divergence-free, we get

$$\begin{aligned} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1}, v^{n+1}\right)_{\Omega_{t^{n+1}}} &= -\frac{1}{2} \int_{\Omega_{t^{n+1}}} \boldsymbol{w}^{n+1} \cdot \nabla (v^{n+1})^2 \, \mathrm{d}\Omega \\ &= \frac{1}{2} \int_{\Omega_{t^{n+1}}} (\nabla \cdot \boldsymbol{w}^{n+1})(v^{n+1})^2 \, \mathrm{d}\Omega. \end{aligned}$$

We bound the terms associated to the mesh velocity as follows:

$$\int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla\cdot\boldsymbol{w}^{n+1})(v^n)^2 \,\mathrm{d}\Omega\,\mathrm{d}s \leq \delta t \sup_{s\in(t^n,t^{n+1})} \left\| J_{\mathcal{A}_{t^{n+1},s}} \nabla\cdot\boldsymbol{w}^{n+1} \right\|_{L^\infty(\Omega_{t^{n+1}})} \|v^n\|^2_{L^2(\Omega_{t^{n+1}})},$$

$$-\delta t \int_{\Omega_{t^{n+1}}} \boldsymbol{w}^{n+1}\cdot\nabla(v^{n+1})^2 \,\mathrm{d}\Omega = \delta t \int_{\Omega_{t^{n+1}}} (\nabla\cdot\boldsymbol{w}^{n+1})(v^{n+1})^2 \,\mathrm{d}\Omega \leq \delta t \left\|\nabla\cdot\boldsymbol{w}^{n+1}\right\|_{L^\infty(\Omega_{t^{n+1}})} \left\|v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}.$$

Let us define the parameters

$$\gamma_1^{n+1} = T \sup_{s\in(t^n,t^{n+1})} \left\| J_{\mathcal{A}_{t^{n+1},s}} \nabla\cdot\boldsymbol{w}^{n+1} \right\|_{L^\infty(\Omega_{t^{n+1}})} \tag{3.1}$$

for $n=-1,\dots,N-2$ and $\gamma_1^N=0$, together with

$$\gamma_2^{n+1} = T \left\|\nabla\cdot\boldsymbol{w}^n\right\|_{L^\infty(\Omega_{t^{n+1}})} \tag{3.2}$$

for $n=0,\dots,N-1$ and $\gamma_2^0=0$.

With the inequalities just proved we can easily obtain that

$$B(V,V) + \frac{1}{2T}\sum_{n=-1}^{N-1} \delta t(\gamma_1^{n+1}+\gamma_2^{n+1}) \left\|v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \geq \sup_{n\in[-1,N-1]} \frac{1}{2T}\left\|v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} + \sum_{n=0}^{N-1} 2\delta t\nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}.$$

If the maximum of $\|v^n\|_{L^2(\Omega_{t^n})}$ is achieved at $n=N_m$, the sequence

$$\{v^0, v^1, \dots, v^{N_m}, 0, \dots, 0\}$$

has to be added to the test sequence. Sometimes in the paper we obtain the maximum using this technique. Invoking the Gronwall lemma (see [21]), we can absorb the second term of the left-hand side with the first term of the right-hand side for a δt small enough. More precisely, the time step must be such that

$$\delta t < \frac{1}{\sup_{n\in[0,N]}(\gamma_1^n+\gamma_2^n)} =: \delta t^1_{\mathrm{cr}}.$$

We note that this is the time step size of the *normalized* problem in time. The original δt^1_{cr} does not depend on T any longer. ∎

This result, together with the continuity of $L(\cdot)$ proved in the next lemma, will lead us to a classical stability bound.

LEMMA 3.2 (continuity). *The following inequality holds:*

$$L(V) \leq \sum_{n=0}^{N-1} \frac{\delta t}{2\beta\nu}\left\|f^{n+1}\right\|^2_{H^{-1}(\Omega_{t^{n+1}})} + \sum_{n=0}^{N-1}\frac{\delta t\beta\nu}{2}\left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} + \frac{1}{4T}\left\|u^0\right\|^2_{L^2(\Omega_0)} + \frac{1}{4T}\left\|v^0\right\|^2_{L^2(\Omega_0)}$$

for all $\beta>0$.

Proof. The right-hand side has the following expression:

$$L(V) = \frac{1}{2T}\left(u^0, v^0\right)_{\Omega_0} + \sum_{n=0}^{N-1} \delta t \langle f^{n+1}, v^{n+1}\rangle_{\Omega_{t^{n+1}}}.$$

The Cauchy–Schwarz inequality leads to

$$L(V) \leq \left(\sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\|f^{n+1}\right\|^2_{H^{-1}(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}} \left(\sum_{n=0}^{N-1} \delta t \nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}} + \frac{1}{2T}\left\|u^0\right\|_{L^2(\Omega_0)} \left\|v^0\right\|_{L^2(\Omega_0)}.$$

The proof is finished by invoking Young's inequality. ∎

From Theorem 3.1 and Lemma 3.2 the following stability result is straightforward.

COROLLARY 3.3 (stability). *There exists $\delta t^1_{\rm cr}$ such that, for $0 < \delta t < \delta t^1_{\rm cr}$, the sequence U, solution of problem* (2.9)–(2.11)*, is bounded as follows:*

$$|||U|||^2 \leq C\left\{\frac{1}{T}\left\|u^0\right\|^2_{L^2(\Omega_0)} + \sum_{n=0}^{N-1} \frac{\delta t}{\nu}\left\|f^{n+1}\right\|^2_{H^{-1}(\Omega_{t^{n+1}})}\right\}.$$

Remark 3.1. The BDF1 method is unconditionally stable for fixed domains. However, for moving domains this property is not maintained anymore. In this case only conditional stability can be proved, with the critical time step value obtained above.

The next task is to obtain an optimal convergence result. In the following theorem, relying on the stability properties proved in Corollary 3.3, *optimal* error estimates are obtained. We denote by $e^{n+1} := u(t^{n+1}) - u^{n+1}$ the error introduced by the time integration at time t^{n+1}, and by $E := U_{\rm ex} - U$ the sequence of these errors.

THEOREM 3.4 (convergence). *There exists $\delta t^1_{\rm cr}$ such that, for $0 < \delta t < \delta t^1_{\rm cr}$, the sequence of errors $E = U_{\rm ex} - U$ satisfies the following error estimate:*

$$|||E|||^2 \leq C\frac{\delta t^2}{T}\sum_{n=0}^{N-1}\delta t\left(\left\|\left.\frac{\partial^2 u}{\partial t^2}\right|_{\boldsymbol{x}_0}\right\|^2_{L^2(\Omega_{t^{n+1}})} + \sup_{s\in(t^n,t^{n+1})}\left\|\left|\frac{\partial^2 \mathcal{A}_s}{\partial t^2}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}\right). \tag{3.3}$$

Proof. We start by taking the exact solution sequence $U_{\rm ex}$ in the bilinear form. We get

$$B(U_{\rm ex}, V) = L(V) + \sum_{n=0}^{N-1}\frac{1}{T}\left(u(t^{n+1}) - u(t^n) - \delta t\left.\frac{\partial u}{\partial t}\right|_{t^{n+1}}, v^{n+1}\right)_{\Omega_{t^{n+1}}} - \sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1} - \boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}), v^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

We subtract the equation for the semidiscrete sequence of solutions to the previous

equations and arrive at

$$B(U-U_{\rm ex},V) = -\sum_{n=0}^{N-1}\frac{1}{T}\left(u(t^{n+1})-u(t^n)-\delta t\,\frac{\partial u}{\partial t}\Big|_{t^{n+1}},v^{n+1}\right)_{\Omega_{t^{n+1}}} + \sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}),v^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

We test the previous equation with $V = U - U_{\rm ex} = E$, obtaining

$$B(E,E) = -\sum_{n=0}^{N-1}\frac{1}{T}\left(u(t^{n+1})-u(t^n)-\delta t\,\frac{\partial u}{\partial t}\Big|_{t^{n+1}},e^{n+1}\right)_{\Omega_{t^{n+1}}} + \sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}),e^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

Exploiting the fact that the bilinear form is coercive, the remaining ingredient is an appropriate bound for the error terms associated to the time discretization. Let us start with the terms related to the time derivative. We use the following Taylor formula for u:

(3.4)
$$\frac{u(\boldsymbol{x}_0,t^{n+1})-u(\boldsymbol{x}_0,t^n)}{T\delta t}-\frac{1}{T}\,\frac{\partial u}{\partial t}\Big|_{\boldsymbol{x}_0}(t^{n+1}) = -\frac{1}{T\delta t}\int_{t^n}^{t^{n+1}}(s-t^n)\,\frac{\partial^2 u}{\partial t^2}\Big|_{\boldsymbol{x}_0}(s)\,\mathrm{d}s.$$

For the mesh velocity, we use

$$\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}) = -\frac{1}{T\delta t}\left(\int_{t^n}^{t^{n+1}}(s-t^n)\frac{\partial^2\mathcal{A}_s}{\partial t^2}\,\mathrm{d}s\right)\circ\mathcal{A}^{-1}_{t^{n+1}}. \tag{3.5}$$

As explained in section 2, it is understood with this notation that this equality holds for arbitrary $\boldsymbol{x}\in\Omega_t$.

With (3.4) we get a bound for the term associated to the time derivative of u as follows:

$$\begin{aligned}
&\int_{\Omega_{t^{n+1}}} e^{n+1}\cdot\left(\int_{t^n}^{t^{n+1}}(s-t^n)\,\frac{\partial^2 u}{\partial t^2}\Big|_{\boldsymbol{x}_0}(\boldsymbol{x}_0,s)\,\mathrm{d}s\right)\circ\mathcal{A}^{-1}_{t^{n+1}}\,\mathrm{d}\Omega\\
&\quad\le\int_{t^n}^{t^{n+1}}\int_{\Omega_0}J_{\mathcal{A}_{t^{n+1}}}(s-t^n)\widehat{e^{n+1}}\,\widehat{\frac{\partial^2 u}{\partial t^2}}\,\mathrm{d}\Omega\,\mathrm{d}s\\
&\quad\le\left(\int_{t^n}^{t^{n+1}}(s-t^n)^2\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac12}\\
&\qquad\times\left(\int_{t^n}^{t^{n+1}}\int_{\Omega_0}J_{\mathcal{A}_{t^{n+1}}}\left(\widehat{\frac{\partial^2 u}{\partial t^2}}\right)^2\mathrm{d}\Omega\,\mathrm{d}s\right)^{\frac12}\\
&\quad\le\frac{\beta_1\delta t}{2}\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}+C\delta t^3\left\|\frac{\partial^2 u}{\partial t^2}\Big|_{\boldsymbol{x}_0}\right\|^2_{L^2(\Omega_{t^{n+1}})},
\end{aligned}$$

where β_1 is the coercivity constant introduced in Theorem 3.1. Similarly, using (3.5) for the term related to the time derivative of the mapping, we get

$$
\begin{aligned}
&-\int_{\Omega_{t^{n+1}}} e^{n+1}\left(\int_{t^n}^{t^{n+1}} (s-t^n)\frac{\partial^2 \mathcal{A}_s}{\partial t^2}\,\mathrm{d}s\right)\circ\mathcal{A}_{t^{n+1}}^{-1}\cdot\nabla u^{n+1}\,\mathrm{d}\Omega\\
&\quad\le \int_{t^n}^{t^{n+1}}\int_{\Omega_0} J_{\mathcal{A}_{t^{n+1}}}(s-t^n)\widehat{e^{n+1}}\frac{\partial^2\mathcal{A}_s}{\partial t^2}\cdot\nabla\widehat{u^{n+1}}\,\mathrm{d}\Omega\,\mathrm{d}s\\
&\quad\le \left(\int_{t^n}^{t^{n+1}}(s-t^n)^2\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\,\mathrm{d}s\right)^{\frac{1}{2}}\\
&\qquad\times\left(\int_{t^n}^{t^{n+1}}\left\|\left|\frac{\partial^2\mathcal{A}_s}{\partial t^2}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}\,\mathrm{d}s\right)^{\frac{1}{2}}\\
&\quad\le \frac{\beta_1\delta t}{2}\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}+C\delta t^3\sup_{s\in(t^n,t^{n+1})}\left\|\left|\frac{\partial^2\mathcal{A}_s}{\partial t^2}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}.
\end{aligned}
$$

With these results we can write

$$
\begin{aligned}
B(E,E)\le\frac{1}{T}\sum_{n=0}^{N-1}\Bigg(&\delta t\beta_1\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}+C\delta t^3\left\|\left.\frac{\partial^2 u}{\partial t^2}\right|_{\boldsymbol{x}_0}\right\|^2_{L^2(\Omega_{t^{n+1}})}\\
&+C\delta t^3\sup_{s\in(t^n,t^{n+1})}\left\|\left|\frac{\partial^2\mathcal{A}_s}{\partial t^2}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}\Bigg).
\end{aligned}
\tag{3.6}
$$

At this point we invoke the coercivity property of the bilinear form proved in Theorem 3.1. Thus, the first term of the right-hand side in (3.6) can be absorbed using the Gronwall lemma. We note that in this case we can apply the Gronwall lemma without any extra condition over the time step size (see [21]). ∎

Clearly, the second term in the right-hand side of (3.3) is bounded if the second time derivatives of the ALE mapping are uniformly bounded in $[0,T]$. In this case, its norm in the space $L^\infty(0,T;L^\infty(\Omega_0))$ can be taken out of the sum, and the stability estimate of Corollary 3.3 allows us to bound the remaining term. However, we have kept expression (3.3) to display the structure of the error bound.

We conclude this subsection with the following improved stability estimate.

COROLLARY 3.5 (stability in $\ell^\infty(H^2(\Omega_t))$). *Under the conditions of Theorem* 3.4, *suppose additionally that the right-hand side of* (3.3) *is bounded, that* $u\in L^\infty(0,T;H^2(\Omega_t))$, *and that the domain* Ω_t *is such that* $\Delta u\in L^2(\Omega_t)$ *implies* $u\in H^2(\Omega_t)$. *Then,* $U\in\ell^\infty(H^2(\Omega_t))$.

Proof. At each time step we can write the error equation

$$
\begin{aligned}
\nu\Delta(u^{n+1}-u(t^{n+1}))=&(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla(u^{n+1}-u(t^{n+1}))\\
&+(\boldsymbol{w}(t^{n+1})-\boldsymbol{w}^{n+1})\cdot\nabla u(t^{n+1})+\frac{1}{\delta t}(u^{n+1}-u^n)-\left.\frac{\partial u}{\partial t}\right|_{t^{n+1}}.
\end{aligned}
$$

By virtue of Theorem 3.4, all the terms in the right-hand side are bounded in $L^2(\Omega_{t^{n+1}})$

for $n = 0, \dots, N-1$. Since

$$\|\Delta u^{n+1}\|_{L^2(\Omega_{t^{n+1}})} \leq \|\Delta u^{n+1} - \Delta u(t^{n+1})\|_{L^2(\Omega_{t^{n+1}})} + \|\Delta u(t^{n+1})\|_{L^2(\Omega_{t^{n+1}})},$$

it follows that $\{\Delta u^{n+1}\}_{n=0}^{N-1} \in \ell^\infty(L^2(\Omega_t))$. The assumption on the domain Ω_t implies that $\{u^{n+1}\}_{n=0}^{N-1} \in \ell^\infty(H^2(\Omega_t))$. ∎

This justifies our strategy of first analyzing the problem semidiscretized in time and then the fully discrete problem. When we will require $U \in \ell^2(H^{p+1}(\Omega_t))$ to obtain optimal order of convergence in space, we know that at least for $p = 1$ this holds under the same condition on the domain Ω_t as for the sequence of solutions of the continuous problem, U_{ex}. It is well known that this condition on Ω_t holds, for example, if it is convex and polyhedral (see, for example, [19]).

3.2. Analysis of BDF2-BDF2$_{\delta t}$. For the second order method we follow the same procedure used above. In this case the problem that we analyze can be written using (2.9) together with the bilinear form (2.13) and the right-hand side linear form (2.14), and we denote by $U = \{u^0, u^1, u^2, \dots, u^N\}$ the sequence of solutions of this problem.

We start by again proving that the corresponding bilinear form is coercive.

THEOREM 3.6 (coercivity). *There exists δt_{cr}^2 such that for $0 < \delta t < \delta t_{\text{cr}}^2$ the bilinear form $B(\cdot,\cdot)$ defined in (2.12) is coercive; that is, for every sequence $V = \{v^n\}_{n=0}^N$, with $v^n \in \mathcal{V}(\Omega_{t^n})$,*

$$B(V,V) \geq \beta_2 |||V|||^2$$

for a certain constant $\beta_2 > 0$.

Proof. We know, from the definition of the bilinear form, that

$$\begin{aligned} B(V,V) = &\sum_{n=0}^{N-1} \left[\delta t \left\| \nabla v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} + \delta t \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1}, v^{n+1} \right)_{\Omega_{t^{n+1}}} \right] \\ &+ \sum_{n=1}^{N-1} \frac{1}{2T} \left(3v^{n+1} - 4v^n + v^{n-1}, v^{n+1} \right)_{\Omega_{t^{n+1}}} + \frac{1}{T} \left(v^1 - v^0, v^1 \right)_{\Omega_{t^1}} \\ &+ \frac{1}{2T} \left\| v^0 \right\|^2_{L^2(\Omega_0)}. \end{aligned} \tag{3.7}$$

Integrating (2.7) from t^n to t^{n+1} for the functions v^n and $2v^n - v^{n-1}$, we can express the term corresponding to the discrete time derivative as follows:

$$\begin{aligned} &\frac{1}{2T}(3v^{n+1} - 4v^n + v^{n-1}, 4v^{n+1})_{\Omega_{t^{n+1}}} \\ &\quad = \frac{1}{T} \left(\left\| v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} - \|v^n\|^2_{L^2(\Omega_{t^n})} + \left\| 2v^{n+1} - v^n \right\|^2_{L^2(\Omega_{t^{n+1}})} \right. \\ &\qquad \left. - \left\| 2v^n - v^{n-1} \right\|^2_{L^2(\Omega_{t^n})} + \left\| \delta^2 v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} \right) \\ &\qquad + \int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1}(s))(v^n)^2 \, \mathrm{d}\Omega \, \mathrm{d}s \\ &\qquad + \int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1}(s))(2v^n - v^{n-1})^2 \, \mathrm{d}\Omega \, \mathrm{d}s. \end{aligned} \tag{3.8}$$

The mesh velocity terms are bounded as follows:

(3.9)
$$\int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1}(s))(v^n)^2 \,\mathrm{d}\Omega \,\mathrm{d}s + \int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1}(s))(2v^n - v^{n-1})^2 \,\mathrm{d}\Omega \,\mathrm{d}s$$
$$\leq \delta t \sup_{s\in(t^n,t^{n+1})} \left\| J_{\mathcal{A}_{t^{n+1},s}} \nabla \cdot \boldsymbol{w}^{n+1}(s) \right\|_{L^\infty(\Omega_{t^{n+1}})} \times \left(\|v^n\|^2_{L^2(\Omega_{t^{n+1}})} + \left\|2v^n - v^{n-1}\right\|^2_{L^2(\Omega_{t^{n+1}})} \right).$$

On the other hand, we can exploit the fact that the convective velocity $\boldsymbol{a}$ is divergence-free, obtaining for the convective term that

$$\begin{aligned} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1}, 4v^{n+1}\right)_{\Omega_{t^{n+1}}} &= -2\delta t \int_{\Omega_{t^{n+1}}} \boldsymbol{w}^{n+1} \cdot \nabla (v^{n+1})^2 \,\mathrm{d}\Omega \\ &= 2\delta t \int_{\Omega_{t^{n+1}}} (\nabla \cdot \boldsymbol{w}^{n+1})(v^{n+1})^2 \,\mathrm{d}\Omega \\ &\leq 2 \left\|\nabla \cdot \boldsymbol{w}^{n+1}\right\|_{L^\infty(\Omega_{t^{n+1}})} \left\|u_h^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}. \end{aligned} \tag{3.10}$$

We use inequalities (3.9) and (3.10) in (3.7) and invoke again the Gronwall lemma. This leads to the desired bound for a time step size:

$$\delta t < \frac{1}{\sup\limits_{n\in[0,N]} (\gamma_1^n + 2\gamma_2^n)} =: \delta t^2_{\mathrm{cr}},$$

slightly different from the one obtained for the first order method. ∎

The previous theorem and Lemma 3.2 allow us to obtain the same stability result as for the previous case, stated in the next corollary.

COROLLARY 3.7 (stability). *There exists δt^2_{cr} such that for $0 < \delta t < \delta t^2_{\mathrm{cr}}$ the sequence U solution of problem* (2.9), (2.13), (2.14) *is bounded as follows:*

$$|||U|||^2 \leq C \sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\|f^{n+1}\right\|^2_{H^{-1}(\Omega_{t^{n+1}})}.$$

Furthermore, we can obtain optimal error estimates under some regularity assumptions. For the sake of clearness we assume that the initialization is calculated exactly. It can be easily checked from Theorem 3.4 that the error introduced by the initialization is optimal.

THEOREM 3.8 (convergence). *There exist δt^2_{cr} such that for $0 < \delta t < \delta t^2_{\mathrm{cr}}$ the sequence of errors $E = U_{ex} - U$ satisfies the following error estimate:*

$$|||E|||^2 \leq C \frac{\delta t^4}{T} \sum_{n=0}^{N-1} \delta t \left(\left\| \left. \frac{\partial^3 u}{\partial t^3} \right|_{\boldsymbol{x}_0} \right\|^2_{L^2(\Omega_{t^{n+1}})} + \sup_{s\in(t^n,t^{n+1})} \left\| \left| \frac{\partial^3 \mathcal{A}_s}{\partial t^3} \right|^2 \right\|_{L^\infty(\Omega_0)} \left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})} \right).$$

Proof. We start by taking the exact solution sequence U_{ex} in the bilinear form. We get

$$B(U_{ex},V)=L(V)+\sum_{n=0}^{N-1}\frac{1}{2T}\left(3u(t^{n+1})-4u(t^n)+u(t^{n-1})-\delta t\,\frac{\partial u}{\partial t}\bigg|_{t^{n+1}},v^{n+1}\right)_{\Omega_{t^{n+1}}}$$
$$-\sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}),v^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

Now we subtract the equation for the semidiscrete sequence of solutions to the previous equations and arrive at

$$B(U-U_{ex},V)=-\sum_{n=0}^{N-1}\frac{1}{T}\left(3u(t^{n+1})-4u(t^n)+u(t^{n-1})-\delta t\,\frac{\partial u}{\partial t}\bigg|_{t^{n+1}},v^{n+1}\right)_{\Omega_{t^{n+1}}}$$
$$+\sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}),v^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

We test the previous equation with $V=U-U_{ex}=E$, obtaining

$$B(E,E)=-\sum_{n=0}^{N-1}\frac{1}{T}\left(3u(t^{n+1})-4u(t^n)+u(t^{n-1})-\delta t\,\frac{\partial u}{\partial t}\bigg|_{t^{n+1}},e^{n+1}\right)_{\Omega_{t^{n+1}}}$$
$$+\sum_{n=0}^{N-1}\delta t\left((\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1}))\cdot\nabla u(t^{n+1}),e^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

The truncation error introduced by the time integration scheme BDF2 is evaluated using the following Taylor formula:

(3.11)
$$\frac{3u(\boldsymbol{x}_0,t^{n+1})-4u(\boldsymbol{x}_0,t^n)+u(\boldsymbol{x}_0,t^{n-1})}{T\delta t}-\frac{1}{T}\,\frac{\partial u}{\partial t}\bigg|_{\boldsymbol{x}_0}(t^{n+1})$$
$$=-\frac{1}{T\delta t}\int_{t^{n-1}}^{t^{n+1}}(s-t^n)^2\,\frac{\partial^3 u}{\partial t^3}\bigg|_{\boldsymbol{x}_0}(s)\,\mathrm{d}s-\frac{1}{T\delta t}\int_{t^n}^{t^{n+1}}(s-t^n)^2\,\frac{\partial^3 u}{\partial t^3}\bigg|_{\boldsymbol{x}_0}(s)\,\mathrm{d}s.$$

The evaluation of the mesh velocity (2.6) requires a time derivative. Its numerical approximation using the second order BDF2 scheme can be written again as a truncation error:

$$\boldsymbol{w}^{n+1}-\boldsymbol{w}(t^{n+1})$$
$$=-\frac{1}{T\delta t}\left(\int_{t^n-1}^{t^{n+1}}(s-t^n)^2\frac{\partial^3\mathcal{A}_s}{\partial t^3}\,\mathrm{d}s+\int_{t^n}^{t^{n+1}}(s-t^n)^2\frac{\partial^3\mathcal{A}_s}{\partial t^3}\,\mathrm{d}s\right)\circ\mathcal{A}^{-1}_{t^{n+1}},\tag{3.12}$$

which holds for all $\boldsymbol{x}\in\Omega_t$. Recall that $\boldsymbol{w}^{n+1}$ stands for the mesh velocity evaluated at t^{n+1}.

The error related to the time derivative of u can be bounded using the following inequality:

$$\begin{aligned}\int_{\Omega_{t^{n+1}}} e^{n+1}\cdot\Bigg(&\int_{t^{n-1}}^{t^{n+1}}(s-t^n)^2\left.\frac{\partial^3 u}{\partial t^3}\right|_{\boldsymbol{x}_0}(s)\,\mathrm{d}s\\ &-\frac{1}{T\delta t}\int_{t^n}^{t^{n+1}}(s-t^n)^2\left.\frac{\partial^3 u}{\partial t^3}\right|_{\boldsymbol{x}_0}(s)\,\mathrm{d}s\Bigg)\circ\mathcal{A}^{-1}_{t^{n+1}}\,\mathrm{d}\Omega\\ &\le\frac{\beta_2\delta t}{2}\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}+C\delta t^5\left\|\left.\frac{\partial^2 u}{\partial t^2}\right|_{\boldsymbol{x}_0}\right\|^2_{L^2(\Omega_{t^{n+1}})},\end{aligned}\tag{3.13}$$

where β_2 is the coercivity constant introduced in Theorem 3.6.

We obtain the following inequality in order to bound the error introduced by the evaluation of the mesh velocity,

$$\begin{aligned}-\int_{\Omega_{t^{n+1}}} e^{n+1}\Bigg(&\int_{t^n-1}^{t^{n+1}}(s-t^n)^2\frac{\partial^3\mathcal{A}_s}{\partial t^3}\,\mathrm{d}s\\ &+\int_{t^n}^{t^{n+1}}(s-t^n)^2\frac{\partial^3\mathcal{A}_s}{\partial t^3}\,\mathrm{d}s\Bigg)\circ\mathcal{A}^{-1}_{t^{n+1}}\cdot\nabla u^{n+1}\,\mathrm{d}\Omega\\ &\le\frac{\beta_2\delta t}{2}\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}+\delta t^5\sup_{s\in(t^{n-1},t^{n+1})}\left\|\left|\frac{\partial^3\mathcal{A}_s}{\partial t^3}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}.\end{aligned}\tag{3.14}$$

Using the error expressions (3.11) and (3.12) and bounds (3.13) and (3.14), we get

$$\begin{aligned}B(E,E)\le\frac{1}{T}\sum_{n=0}^{N-1}\delta t\beta_2\left\|e^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}&+C\frac{\delta t^4}{T}\sum_{n=0}^{N-1}\delta t\left\|\left.\frac{\partial^3 u}{\partial t^3}\right|_{\boldsymbol{x}_0}\right\|^2_{L^2(\Omega_{t^{n+1}})}\\ &+C\frac{\delta t^4}{T}\sum_{n=0}^{N-1}\delta t\sup_{s\in(t^n,t^{n+1})}\left\|\left|\frac{\partial^3\mathcal{A}_s}{\partial t^3}\right|^2\right\|_{L^\infty(\Omega_0)}\left\|u^{n+1}\right\|^2_{H^1(\Omega_{t^{n+1}})}.\end{aligned}$$

Again, we can apply the Gronwall lemma without any extra condition over the time step size. ∎

4. The fully discrete problem. In this section we analyze the fully discrete problems BDF1-BDF1-OSS$_{\delta t,h}$ and BDF2-BDF2-OSS$_{\delta t,h}$. In both cases, stability and error estimates are obtained.

Observe from (2.20) that τ^n has been taken constant in space. Further, we assume Θ_h^t quasi-uniform. In this case, the following inverse estimate holds (see [5]):

$$\|\nabla v_h\|_{L^2(\Omega_t)}\le\frac{C_{\mathrm{inv}}}{h}\|v_h\|_{L^2(\Omega_t)}.\tag{4.1}$$

In order to obtain optimal convergence results, we assume that $u^{n+1}\in H^{p+1}(\Omega_t)$ for $n=0,\dots,N-1$, where p is the degree of the polynomial defining the finite element

space $\mathcal{V}_h$. We also assume that for any function $v \in H^{p+1}(\Omega_t)$ there exists a finite element interpolation $\pi_h(v)$ such that

$$\|v - \pi_h(v)\|_{H^m(\Omega_t)} \leq C_h h^{p+1-m} \|v\|_{H^{p+1}(\Omega_t)}.$$

We need to prove that the L^2-projection onto the finite element space is an optimal interpolation in the $L^2(\Omega_t)$-norm and the seminorm $\|\nabla(\cdot)\|_{L^2(\Omega_t)}$. We show this in the following lemma.

LEMMA 4.1. *Given a function $v \in H^{p+1}(\Omega_t)$ with $p \geq 1$, its L^2-projection onto the finite element space $\Pi_h(v)$ satisfies*

$$\|v - \Pi_h(v)\|_{L^2(\Omega_t)} \leq C_h h^{p+1} \|v\|_{H^{p+1}(\Omega_t)} \tag{4.2}$$

and also

$$h^2 \|\Delta v - \Pi_h(\Delta v)\|_{L^2(\Omega_t)} \leq C h^{p+1} \|v\|_{H^{p+1}(\Omega_t)}. \tag{4.3}$$

If the inverse estimate (4.1) *holds true,*

$$\|\nabla(v - \Pi_h(v))\|_{L^2(\Omega_t)} \leq C h^p \|v\|_{H^{p+1}(\Omega_t)} \tag{4.4}$$

is satisfied.

The proof of this lemma is straightforward and relies on classical interpolation inequalities.

As in the previous section, C is a positive constant, possibly with different values at different appearances.

4.1. Analysis of BDF1-BDF1-OSS$_{\delta t,h}$. In this subsection we analyze the fully discrete problem (2.17) with the bilinear form $B_h(\cdot,\cdot)$ defined in (2.18) and right-hand side (2.14). We denote by $U = \{u^0, u^1, u^2, \ldots, u^N\}$ the sequence of solutions of the semidiscrete problem (in time) (2.9)–(2.11) and by $U_h = \{u_h^0, u_h^1, u_h^2, \ldots, u_h^N\}$ its fully discrete counterpart, solution of (2.17), (2.18), (2.14).

We start by proving the coercivity of the bilinear form for the *weak* norm $|||\cdot|||_w$. This result will be used in the convergence analysis.

THEOREM 4.2 (coercivity). *There exists δt_{cr}^1 such that for $0 < \delta t < \delta t_{\mathrm{cr}}^1$ the bilinear form $B_h(\cdot,\cdot)$ defined in* (2.18) *is coercive. That is, for every sequence $V = \{v^n\}_{n=0}^N$, with $v^n \in \mathcal{V}(\Omega_{t^n})$,*

$$B_h(V,V) \geq \beta_1 |||V|||_w^2$$

for a certain constant $\beta_1 > 0$ independent of h.

Proof. The bilinear form analyzed in this theorem is equal to the one for which coercivity is proved in Theorem 3.1 plus the stabilization term. We can easily get

$$\begin{aligned} B_h(V,V) = {} & \frac{1}{2T} \left\|v^N\right\|_{L^2(\Omega^N)}^2 + \frac{1}{2T} \sum_{n=0}^{N-1} \left\|\delta v^{n+1}\right\|_{L^2(\Omega_{t^{n+1}})}^2 \\ & + \sum_{n=0}^{N-1} \left[\delta t \nu \left\|\nabla v^{n+1}\right\|_{L^2(\Omega_{t^{n+1}})}^2 + \delta t \tau^{n+1} \left\|\Pi_h^{\perp}\left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1}\right)\right\|_{L^2(\Omega_{t^{n+1}})}^2\right] \\ & + \frac{1}{2} \sum_{n=0}^{N-1} \delta t \left(\nabla \cdot \boldsymbol{w}^{n+1}, (v^{n+1})^2\right)_{\Omega_{t^{n+1}}} + \frac{1}{2} \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \int_{\Omega_t} (\nabla \cdot \boldsymbol{w}^{n+1})(v^{n+1})^2 \, \mathrm{d}\Omega. \end{aligned} \tag{4.5}$$

Due to the fact that the stabilization term does not affect the treatment of the mesh velocity terms in Theorem 3.1, we refer to this theorem for the remainder of the proof. ∎

Let us define the Λ-coercivity property associated to a bilinear form that will be used in the following analysis.

DEFINITION 4.3 (Λ-coercivity). *Let $\mathcal{V}$ be a functional space and $\zeta : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$ a bilinear form. We say that ζ is Λ-coercive with respect to the norm $|||\cdot|||$ and the linear operator $\Lambda : \mathcal{V} \longrightarrow \mathcal{V}$ if there exists a constant $\beta > 0$ such that*

$$\zeta(v, \Lambda(v)) \geq \beta |||v|||^2 \quad \forall v \in \mathcal{V}. \tag{4.6}$$

The bilinear form $\zeta(\cdot,\cdot)$ also satisfies an *inf-sup* condition under the conditions of the following lemma.

LEMMA 4.4. *If Λ is continuous with respect to the norm $|||\cdot|||$ and $\zeta(\cdot,\cdot)$ is Λ-coercive, then there exists $\gamma > 0$ such that*

$$\inf_{u\in\mathcal{V}} \sup_{v\in\mathcal{V}} \frac{\zeta(u,v)}{|||u||| \, |||v|||} \geq \gamma.$$

The proof of the previous lemma is straightforward from Definition 4.3 and the continuity of the operator $\Lambda(\cdot)$.

We now show that the bilinear form $B_h(\cdot,\cdot)$ of our problem is Λ-coercive for the *strong* norm $|||\cdot|||_s$.

THEOREM 4.5 (Λ-coercivity). *Let $V = \{v^n\}_{n=0}^N$ be a sequence of functions such that $v^n \in \mathcal{V}(\Omega_{t^n})$ and consider the operator*

$$\Lambda(V) = V + \left\{0, \frac{1}{2}\left\{\tau^{n+1}\Pi_h\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\}_0^{N-1}\right\}.$$

Then, there exists δt^1_{cr} such that, for $0 < \delta t < \delta t^1_{\mathrm{cr}}$, the bilinear form $B_h(\cdot,\cdot)$ is Λ-coercive:

$$B_h\left(V, \Lambda(V)\right) \geq \beta_1 |||V|||_s^2$$

for a certain constant $\beta_1 > 0$ independent of h.

Proof. Testing (2.18) with the sequence of functions that belong to the finite element space

$$\Pi_0(\tau, V) := \{0, \{\tau^{n+1}\Pi_0(v^{n+1})\}\}_{n=0}^{N-1} := \{0, \{\tau^{n+1}\Pi_h\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\}_{n=0}^{N-1}\},$$

we have

$$\begin{aligned} B_h(V, \Pi_0(\tau, V)) \geq & \sum_{n=0}^{N-1} \phi^{n+1}\delta t\tau^{n+1} \left\|\Pi_h\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})} \\ & - \sum_{n=0}^{N-1}\left[\frac{1}{T}\left\|\delta v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} + \delta t\nu \left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\right. \\ & \left. + \delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left((\boldsymbol{a} - \boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right], \end{aligned} \tag{4.7}$$

where

$$\phi^{n+1} := 1 - \frac{1}{4}\frac{\tau^{n+1}}{T\delta t} - \frac{1}{4}\tau^{n+1}\frac{\nu C^2_{\mathrm{inv}}}{h^2} - \frac{1}{4}(\tau^{n+1})^2\frac{\left\|\boldsymbol{a} - \boldsymbol{w}^{n+1}\right\|^2_{L^\infty(\Omega_{t^{n+1}})} C^2_{\mathrm{inv}}}{h^2}. \tag{4.8}$$

To obtain (4.7) we have made use of Young's inequality and the inverse estimate (4.1). Assuming now that the constants c_1 and c_2 in (2.20) are such that $c_1 \leq C_{\text{inv}}^2$ and $c_2 \leq C_{\text{inv}}$ and the constant C in (2.21) is $C \leq 1$, it follows that $\phi^{n+1} \geq 1/4$.

The combination of (4.7) and (4.5) leads to

$$
\begin{aligned}
B_h(V, 2V + \Pi_0(\tau, V)) &\geq \frac{1}{T} \left\| v^N \right\|^2_{L^2(\Omega^N)} + \sum_{n=0}^{N-1} \delta t \nu \left\| \nabla v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + C \sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\| (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + \sum_{n=0}^{N-1} \delta t \left(\nabla \cdot \boldsymbol{w}^{n+1}, (v^{n+1})^2 \right)_{\Omega_{t^{n+1}}} + \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \int_{\Omega_s} (\nabla \cdot \boldsymbol{w}^{n+1})(v^{n+1})^2 \, \mathrm{d}\Omega \, \mathrm{d}s \\
&\geq \frac{1}{T} \left\| v^N \right\|^2_{L^2(\Omega^N)} + \sum_{n=0}^{N-1} \delta t \nu \left\| \nabla v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + C \sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\| \Pi_h \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1} \right) \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad - \sum_{n=0}^{N-1} \delta t \gamma_{n+1} \left\| v^{n+1} \right\|^2_{L^2(\Omega_{t^{n+1}})},
\end{aligned}
$$

with $\gamma_{n+1} := \gamma_1^{n+1} + \gamma_2^{n+1}$ and γ_1^{n+1}, γ_2^{n+1} defined in (3.1) and (3.2). Using the Gronwall lemma, we finally get the coercivity stated in the theorem. We point out that the critical time step δt_{cr}^1 in this case is identical to the one obtained for the semidiscrete problem. ∎

In order to satisfy the continuity of $\Lambda(\cdot)$ needed to obtain the *inf-sup* condition in Lemma 4.4, we have to restrict the situation to the discrete finite element space $\mathcal{V}_h$.

LEMMA 4.6 (continuity). *Let $V_h = \{v_h^n\}_{n=0}^N$ be a finite element sequence such that $v_h^n \in \mathcal{V}_h(\Omega_{t^n})$, and consider the operator Λ introduced in Theorem 4.5. Then, $\Lambda(\cdot)$ is continuous with respect to the norm $|||\cdot|||_s$ for every finite element sequence V_h:*

$$
|||\Lambda(V_h)|||_s \leq \rho |||V_h|||_s \tag{4.9}
$$

for a certain constant $\rho > 0$ independent of h.

Proof. Defining $\Pi_0(\tau, V_h)$ as in the proof of the previous theorem, we have from the definition of the norm that

$$
\begin{aligned}
|||\Pi_0(\tau, V_h)|||_s^2 &= \frac{1}{T} \sup_{n \in [0, N-1]} \left\| \tau^{n+1} \Pi_0(v_h^{n+1}) \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + \sum_{n=0}^{N-1} \delta t \nu \left\| \tau^{n+1} \nabla \Pi_0(v_h^{n+1}) \right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + \sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\| \tau^{n+1} (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla \Pi_0(v_h^{n+1}) \right\|^2_{L^2(\Omega_{t^{n+1}})}.
\end{aligned} \tag{4.10}
$$

Invoking the expression for τ^{n+1} and the inverse estimate (4.1), we can easily bound every term by $|||V|||_s^2$. ∎

Remark 4.1. The fact that we need to use the inverse estimate (4.1) in order to bound *the first term in* (4.10) restricts the continuity of $\Lambda(\cdot)$ to finite element sequences (for the rest of the terms the inverse estimate is applied to derivatives of $\Pi_h((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v^{n+1})$, a finite element function even if v^{n+1} is not in the finite element space). However, this restriction does not complicate the convergence analysis, where only the Λ-coercivity is invoked.

From Lemmas 4.4 and 4.6 we obtain the *discrete inf-sup* condition.

COROLLARY 4.7 (discrete inf-sup condition). *Let* $U_h = \{u_h^n\}_{n=0}^N$ *and* $V_h = \{v_h^n\}_{n=0}^N$ *be sequences of finite element functions such that* $u^n, v^n \in \mathcal{V}(\Omega_{t^n})$. *There exists* δt_{cr}^1 *such that, for* $0 < \delta t < \delta t_{\text{cr}}^1$, *the bilinear form* $B_h(\cdot,\cdot)$ *satisfies the following condition:*

$$\inf_{U_h \in \mathcal{V}_h} \sup_{V_h \in \mathcal{V}_h} \frac{B_h\left(U_h, V_h\right)}{|||U_h|||_s \, |||V_h|||_s} \geq \widetilde{\beta}_1$$

for a certain constant $\widetilde{\beta}_1 > 0$ *independent of* h.

At this point, the only other ingredient needed for a stability result is the continuity of the *force term*, provided by Lemma 3.2. The stability result is stated in the next corollary.

COROLLARY 4.8 (stability). *There exists* δt_{cr}^1 *such that, for* $0 < \delta t < \delta t_{\text{cr}}^1$, *the sequence* U_h, *solution of problem* (2.17), (2.18), (2.11), *is bounded as follows:*

$$|||U_h|||_s^2 \leq C \sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|_{H^{-1}(\Omega_{t^{n+1}})}^2 .$$

For the convergence analysis, let us define the difference between the solution of (2.8) and (2.16) as $e_d^{n+1} := u_h^{n+1} - u^{n+1}$, and the sequence of these errors by E_d. From Theorem 4.5, which proves the Λ-coercivity of the bilinear form B_h for Λ defined in this theorem, we know that

$$B_h\left(E_d, \Lambda(E_d)\right) \geq \beta_1 |||E_d|||_s^2. \tag{4.11}$$

We subtract the discrete bilinear form (2.18) from its semidiscrete counterpart (2.10) tested with finite element sequences in order to get

$$\begin{aligned} B_h\left(E_d, V_h\right) &= \epsilon_c(V_h) \\ &:= -\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left(\Pi_h^{\perp} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla u^{n+1} \right), (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla v_h^{n+1} \right)_{\Omega_{t^{n+1}}}, \end{aligned}$$

where $\epsilon_c(V_h)$ accounts for the consistency error. After some manipulations, we can write

$$\begin{aligned} B_h\left(E_d, \Lambda(E_d)\right) &= B_h\left(E_d, E_d\right) + \frac{1}{2} B_h\left(E_d, \Pi_0(\tau, E_d)\right) \\ &= B_h\left(E_d, \Pi_h(U) - U\right) + \epsilon_c(U_h - \Pi_h(U)) + \frac{1}{2}\epsilon_c(\Pi_0(\tau, E_d)), \end{aligned}$$

where $\Pi_h(U) := \{\Pi_h(u^n)\}_{n=0}^N$.

We distinguish between *interpolation* error, the first term of the right-hand side, and the *consistency* error associated to the second and third terms. In the following

two lemmas we bound these error terms. We start with the interpolation error, obtaining the result stated in the following lemma.

LEMMA 4.9 (interpolation error). *The error sequence $E_d = U_h - U$ satisfies the following inequality:*

(4.12)
$$B_h\left(E_d, \Pi_h(U) - U\right) \leq C|||E_d|||_w \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.$$

Proof. Let us expand the expression of the interpolation error, making use of the definition of the bilinear form associated to the problem we are analyzing:

$$\begin{aligned}
&B_h\left(E_d, \Pi_h(U) - U\right) \\
&= \sum_{n=0}^{N-1} \Bigg[\frac{1}{T}\left(e_d^{n+1} - e_d^n, \Pi_h\left(u^{n+1}\right) - u^{n+1}\right)_{\Omega_{t^{n+1}}} \\
&\quad + \delta t \nu \left(\nabla e_d^{n+1}, \nabla(\Pi_h\left(u^{n+1}\right) - u^{n+1})\right)_{\Omega_{t^{n+1}}} \\
&\quad + \delta t \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1}, \Pi_h\left(u^{n+1}\right) - u^{n+1}\right)_{\Omega_{t^{n+1}}} \\
&\quad + \delta t \tau^{n+1} \left(\Pi_h^{\perp}\left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1}\right), (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla \left(\Pi_h\left(u^{n+1}\right) - u^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \Bigg].
\end{aligned}$$

We must control each term separately. Let us start with the discrete time derivative term. Using assumption (2.21) we have that

$$\begin{aligned}
&\sum_{n=0}^{N-1} \frac{1}{T}\left(e_d^{n+1} - e_d^n, \Pi_h\left(u^{n+1}\right) - u^{n+1}\right)_{\Omega_{t^{n+1}}} \\
&\qquad \leq C \left(\sum_{n=0}^{N-1} \frac{1}{T} \left\|e_d^{n+1} - e_d^n\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}} \\
&\qquad\quad \times \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.
\end{aligned}$$

For the viscosity term, using the definition of τ^{n+1} and the inverse estimate (4.5), we have that

$$\begin{aligned}
&\sum_{n=0}^{N-1} \delta t \nu \left(\nabla e_d^{n+1}, \nabla(\Pi_h\left(u^{n+1}\right) - u^{n+1})\right)_{\Omega_{t^{n+1}}} \\
&\qquad \leq C \left(\sum_{n=0}^{N-1} \delta t \nu \left\|\nabla e_d^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}} \\
&\qquad\quad \times \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.
\end{aligned}$$

Similar arguments allow us to obtain a bound for the convective term,

$$\sum_{n=0}^{N-1} \delta t \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1}, \Pi_h \left(u^{n+1} \right) - u^{n+1} \right)_{\Omega_{t^{n+1}}}$$

$$\leq C \left(\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\| \Pi_h^{\perp} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1} \right) \right\|^2_{L^2(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}$$

$$\times \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}, \tag{4.13}$$

and for the stabilization term we obtain

$$\sum_{n=0}^{N-1} \delta t \tau^{n+1}$$

$$\times \left(\Pi_h^{\perp} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1} \right), (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla \left(\Pi_h \left(u^{n+1} \right) - u^{n+1} \right) \right)_{\Omega_{t^{n+1}}}$$

$$\leq C \left(\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\| \Pi_h^{\perp} \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla e_d^{n+1} \right) \right\|^2_{L^2(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}$$

$$\times \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}.$$

All the terms have been bounded by the right-hand side of (4.12), and therefore the proof is finished. ∎

Remark 4.2. Invoking the interpolation error (4.2) in (4.13) has allowed us to obtain an optimal bound for the interpolation error without the control of the full convective term in the norm $|||\cdot|||_w$. This fact will be used for the analysis of the second order method.

The following lemma is devoted to the control of the consistency error. Since we are interested in smooth solutions, say $u \in L^2(0,T;H^{p+1}(\Omega_t))$ (with the obvious modifications for u less regular), we assume that f is also smooth, in particular $f \in L^2(0,T;H^{p-1}(\Omega_t))$. Thus, for $p \geq 1$, $\langle f, v_h \rangle_{\Omega_t} = (\Pi_h(f), v_h)_{\Omega_t}$. Therefore, the finite element solution is not altered if we assume $\Pi_h^{\perp}(f) = 0$.

LEMMA 4.10 (consistency error). *The following inequality holds:*

$$\epsilon_c \left(U_h - \Pi_h(U) + \frac{1}{2} \Pi_0(\tau, E_d) \right)$$

$$\leq C \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}$$

$$\times \left(|||E_d|||_s^2 + h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}. \tag{4.14}$$

Proof. From the expression of the consistency error we arrive at

$$
\begin{aligned}
&-\epsilon_c(U_h-\Pi_h(U))\\
&=\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}\right),(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla(u_h^{n+1}-\Pi_h\left(u^{n+1}\right))\right)_{\Omega_{t^{n+1}}}\\
&=\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}\right),(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla e_d^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&+\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}\right),(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla(u^{n+1}-\Pi_h\left(u^{n+1}\right))\right)_{\Omega_{t^{n+1}}}.
\end{aligned}
\tag{4.15}
$$

On the other hand, from the equation for the semidiscrete unknown (2.8), we can easily check that

$$
\begin{aligned}
&\left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}\right),v^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&\qquad=\left(\Pi_h^{\perp}\left(\nu\Delta u^{n+1}-\frac{1}{T\delta t}(u^{n+1}-u^n)\right),v^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&\qquad=:\left(\Pi_h^{\perp}\left(\lambda(u^{n+1})\right),v^{n+1}\right)_{\Omega_{t^{n+1}}},
\end{aligned}
\tag{4.16}
$$

where $\lambda(\cdot):=\nu\Delta(\cdot)-\frac{\delta(\cdot)}{T\delta t}$. Note that we have not included $\Pi_h^{\perp}(f)$ in the previous equation.

Now, using (4.16) in (4.15) we can split the error into two different terms bounded as follows:

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left(\Pi_h^{\perp}\left(\lambda(u^{n+1})\right),(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla e_d^{n+1}\right)_{\Omega_{t^{n+1}}}\\
&\qquad\le\ C\left(\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\lambda(u^{n+1})\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}\\
&\qquad\quad\times\left(\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla e_d^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}},
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left(\Pi_h^{\perp}\left(\lambda(u^{n+1})\right),(\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla\left(u^{n+1}-\Pi_h\left(u^{n+1}\right)\right)\right)_{\Omega_{t^{n+1}}}\\
&\qquad\le\ C\left(\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\lambda(u^{n+1})\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}\\
&\qquad\quad\times\left(h^{2(p+1)}\sum_{n=0}^{N-1}\delta t(\tau^{n+1})^{-1}\left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.
\end{aligned}
$$

On the other hand, the term related to the perturbation of the test function $\Pi_0(\tau,E_d)$

appearing in (4.14) can be bounded using similar arguments, leading to

$$
\begin{aligned}
&\epsilon_c(\Pi_0(\tau, E_d)) \\
&= \sum_{n=0}^{N-1} \delta t \tau^{n+1} \left(\Pi_h^{\perp}\left(\lambda(u^{n+1})\right), (\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla\left(\tau^{n+1}\Pi_h\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla e_d^{n+1}\right)\right)\right)_{\Omega_{t^{n+1}}} \\
&\le C\left(\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\lambda(u^{n+1})\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}} \\
&\quad\times\left(\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla e_d^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.
\end{aligned}
$$

It remains only to prove that

$$
\sum_{n=0}^{N-1}\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\lambda(u^{n+1})\right)\right\|^2_{L^2(\Omega_{t^{n+1}})} \le Ch^{2(p+1)}\sum_{n=0}^{N-1}\delta t(\tau^{n+1})^{-1}\left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}.
$$

This inequality can be easily obtained from the expression of τ^{n+1}, assumption (2.21), and the interpolation error estimate (4.3). ∎

We end this section with the following main convergence result, which is a direct consequence of inequality (4.11) and Lemmas 4.9 and 4.10.

THEOREM 4.11 (convergence). *There exist δt^1_{cr} such that, for $0<\delta t<\delta t^1_{\mathrm{cr}}$, the sequence of errors $E_d = U_h - U$ satisfies the following error estimate:*

$$
|||E_d|||_s^2 \le Ch^{2(p+1)}\sum_{n=0}^{N-1}\delta t(\tau^{n+1})^{-1}\left\|u^{n+1}\right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}.
$$

4.2. Analysis of BDF2-BDF2-OSS$_{\delta t,h}$. In this subsection we analyze the fully discrete problem (2.17) with the bilinear form $B_h(\cdot,\cdot)$ defined in (2.22) and right-hand side (2.14). We denote by $U=\{u^0,u^1,u^2,\dots,u^N\}$ the sequence of solutions of the second order semidiscrete problem (in time) (2.9), (2.13), (2.14) and by $U_h = \{u_h^0,u_h^1,u_h^2,\dots,u_h^N\}$ its fully discrete counterpart, solution of (2.17), (2.22), (2.14).

We have obtained the results of this section using the *weak* norm $|||\cdot|||_w$. Let us start with a theorem proving coercivity under the *weaker* norm.

THEOREM 4.12 (coercivity). *There exists δt^2_{cr} such that, for $0<\delta t<\delta t^2_{\mathrm{cr}}$, the bilinear form $B_h(\cdot,\cdot)$ defined in (2.22) is coercive. That is, for every sequence $V=\{v^n\}_{n=0}^N$ with $v^n\in V(\Omega_{t^n})$*

$$
B_h(V,V) \ge \beta_2|||V|||_w^2
$$

for a certain constant $\beta_2>0$ independent of h.

Proof. It can be easily shown that

$$
\begin{aligned}
B_h(V,4V) &\ge \frac{1}{T}\left(\left\|v^N\right\|^2_{L^2(\Omega^N)} + \sum_{n=0}^{N-1}\left\|\delta^2 v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})}\right) \\
&\quad + \sum_{n=0}^{N-1} 4\delta t\nu\left\|\nabla v^{n+1}\right\|^2_{L^2(\Omega_{t^{n+1}})} + \sum_{n=0}^{N-1}4\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})} \\
&\quad + \int_{t^n}^{t^{n+1}}\int_{\Omega_s}(\nabla\cdot\boldsymbol{w}(s))(v^n)^2\,\mathrm{d}\Omega\,\mathrm{d}s + \int_{t^n}^{t^{n+1}}\int_{\Omega_s}(\nabla\cdot\boldsymbol{w}(s))(2v^n-v^{n-1})^2\,\mathrm{d}\Omega\,\mathrm{d}s.
\end{aligned}
$$

Manipulating the mesh velocity as for the BDF2-BDF2$_{\delta t}$ formulation (see Theorem 3.6) and applying the Gronwall lemma we obtain the desired result. ∎

Stability is now straightforward from Theorem 4.12 and Lemma 3.2.

COROLLARY 4.13 (stability). *There exists $\delta t_{\rm cr}^2$ such that, for $0 < \delta t < \delta t_{\rm cr}^2$, the sequence U_h, solution of problem* (2.17), (2.22), (2.14), *is bounded as follows:*

$$|||U_h|||_w^2 \leq C \sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|_{H^{-1}(\Omega_{t^{n+1}})}^2 .$$

This stability result can be considered *weak*. However, we will see that this result is enough to obtain error estimates that do not blow up for large Péclet numbers, the original motivation of stabilization methods for convection-diffusion problems.

Let us now obtain error estimates for the BDF2-BDF2-OSS$_{\delta t,h}$ formulation. We start with an auxiliary lemma that will be useful in what follows.

LEMMA 4.14. *Let $X = \{x^n\}_{n=0}^N$ and $V = \{v^n\}_{n=0}^N$ be two sequences of functions such that $x^n, v^n \in H^{p+1}(\Omega_{t^n})$. Then, the bilinear form* (2.22) *satisfies the following bound:*

$$B_h\left(X, \Pi_h^\perp(V)\right) \leq C \left(|||X|||_w^2 + \sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1} \|\Pi_h^\perp(x^{n+1})\|_{L^2(\Omega_{t^{n+1}})}^2 \right)^{\frac{1}{2}} \times \left(h^{2(p+1)} \sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| v^{n+1} \right\|_{H^{p+1}(\Omega_{t^{n+1}})}^2 \right)^{\frac{1}{2}} .$$

Proof. From (2.22) we have that

$$\begin{aligned} B_h\left(X, \Pi_h^\perp(V)\right) = & \sum_{n=0}^{N-1} b_h\left(\boldsymbol{w}^{n+1}; x^{n+1}, \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \\ & + \sum_{n=1}^{N-1} \frac{1}{2T}\left(3x^{n+1} - 4x^n + x^{n-1}, \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \\ & + \frac{1}{T}\left(x^1 - x^0, \Pi_h^\perp\left(v^1\right)\right)_{\Omega_{t^1}} + \frac{1}{T}\left(x^0, \Pi_h^\perp\left(v^0\right)\right)_{\Omega_0} , \end{aligned}$$

where

$$\begin{aligned} & \sum_{n=0}^{N-1} b_h\left(\boldsymbol{w}^{n+1}; x^{n+1}, \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \\ & = \sum_{n=0}^{N-1} \delta t \Big[\nu \left(\nabla x^{n+1}, \nabla \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} + \left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla x^{n+1}, \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \\ & \qquad + \tau^{n+1}\left(\Pi_h^\perp\left((\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla x^{n+1}\right), (\boldsymbol{a} - \boldsymbol{w}^{n+1}) \cdot \nabla \Pi_h^\perp\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} \Big] . \end{aligned}$$

Now we have to bound every term of the right-hand side in order to complete the proof. We start with the first term:

$$\begin{aligned} & \sum_{n=0}^{N-1} \delta t \nu \left(\nabla x^{n+1}, \nabla\left(\Pi_h^\perp\left(v^{n+1}\right)\right)\right)_{\Omega_{t^{n+1}}} \\ & \qquad \leq \left(\sum_{n=0}^{N-1} \delta t \nu \left\| \nabla x^{n+1} \right\|_{L^2(\Omega_{t^{n+1}})}^2 \right)^{\frac{1}{2}} \left(\sum_{n=0}^{N-1} \delta t \nu \left\| \nabla \Pi_h^\perp\left(v^{n+1}\right) \right\|_{L^2(\Omega_{t^{n+1}})}^2 \right)^{\frac{1}{2}} . \end{aligned}$$

The second term in the right-hand side can be bounded as

$$\sum_{n=0}^{N-1} \delta t \left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla x^{n+1}, \Pi_h^{\perp}\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}}$$
$$\leq \left(\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\|\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla x^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}$$
$$\times \left(\sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\|\Pi_h^{\perp}\left(v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}},$$

and the third term as

$$\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla x^{n+1}\right), (\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla\Pi_h^{\perp}\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}}$$
$$\leq \left(\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\|\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla x^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}$$
$$\times \left(\sum_{n=0}^{N-1} \delta t \tau^{n+1} \left\|\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla\Pi_h^{\perp}\left(v^{n+1}\right)\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.$$

The term related to the time derivative is bounded after recalling assumption (2.21) for the stabilization parameter τ^{n+1}:

$$\sum_{n=1}^{N-1} \frac{1}{2T}\left(3x^{n+1}-4x^n+x^{n-1}, \Pi_h^{\perp}\left(v^{n+1}\right)\right)_{\Omega_{t^{n+1}}} + \frac{1}{T}\left(x^1-x^0, \Pi_h^{\perp}\left(v^1\right)\right)_{\Omega_{t^1}}$$
$$+ \frac{1}{T}\left(x^0, \Pi_h^{\perp}\left(v^0\right)\right)_{\Omega_0} \leq C\left(\sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1}\left\|\Pi_h^{\perp}\left(x^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}$$
$$\times \left(\sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1}\left\|\Pi_h^{\perp}\left(v^{n+1}\right)\right\|^2_{L^2(\Omega_{t^{n+1}})}\right)^{\frac{1}{2}}.$$

We now have to use (4.4) of Lemma 4.1 and the expression (2.20) of the stabilization parameter τ^{n+1} to conclude the proof. ∎

To obtain the error estimate, we also need to invoke the coercivity of $B_h(\cdot,\cdot)$, which leads to

$$B_h\left(E_d, E_d\right) \geq \beta_2 |||E_d|||_w^2.$$

Subtracting the equation for the semidiscrete velocity and the discrete velocity, we get

$$B_h\left(E_d, V_h\right) =: \epsilon_c(V_h)$$
$$= -\sum_{n=0}^{N-1} \delta t \tau^{n+1}\left(\Pi_h^{\perp}\left((\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla u^{n+1}\right), (\boldsymbol{a}-\boldsymbol{w}^{n+1})\cdot\nabla v_h^{n+1}\right)_{\Omega_{t^{n+1}}}.$$

Using the previous equation, we can obtain

$$B_h\left(E_d, E_d\right) = B_h\left(E_d, \Pi_h(U) - U\right) + \epsilon_c(U_h - \Pi_h(U)).$$

The first term is due to the interpolation error, whereas the second is the consistency error. In the following lemma we obtain a bound for the interpolation error.

LEMMA 4.15 (interpolation error). *The following inequality holds:*

$$B_h\left(E_d, \Pi_h^{\perp}\left(U\right)\right) \leq C \left(|||E_d|||_w^2 + h^{2(p+1)} \sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}} \times \left(h^{2(p+1)} \sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}.$$

Proof. Invoking Lemma 4.14 and using the fact that $\Pi_h\left(U\right) - U = -\Pi_h^{\perp}\left(U\right)$ and $\Pi_h^{\perp}\left(E_d\right) = -\Pi_h^{\perp}\left(U\right)$, we immediately get the result. ∎

In order to bound the consistency error we again follow the technique developed in Lemma 4.10. The only difference between these two cases is the term associated to the time derivative, which does not essentially affect the proof.

LEMMA 4.16 (consistency error). *The following inequality holds:*

$$\epsilon_c(U_h - \Pi_h(U)) \leq C \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}} \times \left(|||E_d|||_w^2 + h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})} \right)^{\frac{1}{2}}.$$

Again, we end with the following desired convergence result, which is straight from Lemma 4.16 for the bound of the consistency error, Lemma 4.15 for the bound of the interpolation error, and Theorem 4.12, which gives coercivity of the bilinear form.

THEOREM 4.17 (convergence). *There exists δt_{cr}^2 such that, for $0 < \delta t < \delta t_{\text{cr}}^2$, the sequence of errors $E_d = U_h - U$ satisfies the following error estimate:*

$$|||E_d|||_w^2 \leq C h^{2(p+1)} \sum_{n=-1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| u^{n+1} \right\|^2_{H^{p+1}(\Omega_{t^{n+1}})}.$$

This error estimate is optimal.

From this analysis, we can easily obtain stability and convergence results when the domain is fixed, that is, when the mesh velocity vanishes.

4.3. Analysis of BDF2-OSS$_{\delta t,h}$. The previous results are new even for fixed domains. The OSS stabilization method was analyzed in [12] using the backward Euler time integration. It can be easily seen that for fixed domains, i.e., when $\boldsymbol{w}^{n+1} = \mathbf{0}$, there is no critical time step size, the method becoming unconditionally stable. In this case, the problem to be solved reads as follows: find a sequence of finite element functions U_h such that

$$B_h(U_h, V_h) = L(V_h) \tag{4.17}$$

with the bilinear form

$$B_h\left(U_h, V_h\right) = \sum_{n=1}^{N-1}\left[\frac{1}{2T}\left(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, v_h^{n+1}\right) + b_h\left(u_h^{n+1}, v_h^{n+1}\right)\right]$$
$$+ \frac{1}{T}\left(u_h^1 - u_h^0, v_h^1\right) + b_h\left(u_h^1, v_h^1\right) + \frac{1}{T}\left(u_h^0, v_h^0\right), \tag{4.18}$$

where now $b_h(u_h^{n+1}, v_h^{n+1})$ denotes $b_h(\mathbf{0}; u_h^{n+1}, v_h^{n+1})$, with $b_h(\boldsymbol{w}^{n+1}; u_h^{n+1}, v_h^{n+1})$ defined in (2.19). The right-hand side linear form is given again by (2.14).

In this case two different sets of results are obtained. The first one with the *weak* norm $|||\cdot|||_w$, and the second one with the *strong* norm $|||\cdot|||_s$. The main difference is that in the second norm, $B_h(\cdot,\cdot)$ loses coercivity. This complicates the analysis.

We state the results with the norm $|||\cdot|||_w$ in the following corollaries. Their proofs are straightforward from the previous analysis.

COROLLARY 4.18 (stability). *The sequence U_h solution of problem* (4.17) *is bounded as follows:*

$$|||U_h|||_w^2 \leq C\sum_{n=0}^{N-1}\frac{\delta t}{\nu}\left\|f^{n+1}\right\|_{H^{-1}(\Omega)}^2$$

for all $\delta t > 0$.

Again, we denote by $U = \{u^0, u^1, u^2, \ldots, u^N\}$ the sequence of solutions of the second order semidiscrete problem (in time) (2.9), (2.13), (2.14), now with $\Omega_t \equiv \Omega$.

COROLLARY 4.19 (convergence). *The error sequence $E_d = U_h - U$ satisfies the following error estimate:*

$$|||E_d|||_w^2 \leq Ch^{2(p+1)}\sum_{n=-1}^{N-1}\delta t(\tau^{n+1})^{-1}\|u^{n+1}\|_{H^{p+1}(\Omega)}^2$$

for all $\delta t > 0$.

The remainder of this section is devoted to improving these stability and convergence estimates. The improvement consists in obtaining estimates in the *stronger* norm $|||\cdot|||_s$. This is possible for *fixed* domains, but we have not been able to obtain estimates similar to those presented next for *moving* domains. Nevertheless, some additional assumptions will be required. We will also note the aspects that make the analysis of the BDF2-OSS$_{\delta t,h}$ method much more involved than that of the BDF1-OSS$_{\delta t,h}$ formulation.

Let us introduce some new notation. We modify the bilinear form as

$$B_h^*\left(U_h, V_h\right) = \sum_{n=0}^{N-1} b_h\left(u_h^{n+1}, v_h^{n+1}\right) + \sum_{n=1}^{N-1}\frac{1}{2T}(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, v_h^{n+1})$$
$$+ \frac{1}{T}(u_h^1 - u_h^0, v_h^1) + \frac{1}{T}(u_h^0, v_h^0) + \frac{1}{T\delta t}(u_h^{-1}, v_h^{-1}), \tag{4.19}$$

and the right-hand-side linear form as

$$L^*(V_h) = \sum_{n=0}^{N-1}\delta t\langle f^{n+1}, v_h^{n+1}\rangle + \frac{1}{T}(u_0, v_h^0) + \frac{1}{T\delta t}(u_{1,h} - \Pi_h(u_0), v_h^{-1}), \tag{4.20}$$

where u_0 is obviously the initial condition and $u_{1,h}$ is the solution at the first time step obtained with the scheme used to initialize the BDF2 scheme. For example, the BDF1 scheme can be used, and this is precisely what is assumed in the expression of $B_h^*(\cdot,\cdot)$. Note that now the sequences of finite element functions start at $n=-1$.

It is easily checked that the solution of (2.9) with the bilinear form (2.12) is equivalent to

$$B_h^*(U_h, V_h) = L^*(V_h).$$

Observe that this problem yields $u_h^{-1} = u_{1,h} - \Pi_h(u_0)$, $u_h^0 = \Pi_h(u_0)$, and $u_h^1 = u_{1,h}$. The rest of the terms of the sequence of unknowns $U = \{u_h^{-1}, u_h^0, u_h^2, \dots, u_h^N\}$ are the same as those in the solution of problem (4.17).

Let us introduce some additional ingredients. Given a sequence

$$V = \{v^{-1}, v^0, v^1, v^2, \dots, v^N\},$$

we define

$$d^{1,*}(V) = \{0, 0, 0, \delta v^2, \delta v^3, \dots, \delta v^{N-1}, \delta v^N\},$$
$$d^{2,*}(V) = \{0, 0, -\delta v^2, -\delta^2 v^2, -\delta^2 v^3, \dots, -\delta^2 v^N, \delta v^N\}.$$

These operators on sequences have the following property: for all sequences $X = \{x^n\}_{n=-1}^N$ it holds that

$$\begin{aligned} B_h^*\left(X, d^{2,*}(V)\right) &= \sum_{n=1}^{N-1} b_h\left(\delta x^{n+1}, \delta v^{n+1}\right) + \sum_{n=2}^{N-1} \frac{1}{2T}(3\delta x^{n+1} - 4\delta x^n + \delta x^{n-1}, \delta v^{n+1}) \\ &\quad + \frac{3}{2T}(\delta x^2 - \delta x^1, \delta v^2) \\ &= B_h^*\left(d^{1,*}(X), d^{1,*}(V)\right) + \frac{1}{2T}(\delta x^1, \delta v^3 - 3\delta v^2). \end{aligned} \tag{4.21}$$

Remark 4.3. The previous property is not satisfied for moving domains due to the fact that the convective velocity changes at every time step. It introduces an extra term $b_h\left(\delta \boldsymbol{w}^{n+1}; u^n, \delta v^{n+1}\right)_{\Omega_{t^{n+1}}}$ that cannot be bounded as required in the following analysis.

In the next theorem we obtain Λ-coercivity for the norm $|||\cdot|||_s$.

THEOREM 4.20 (Λ-coercivity). *Let $V = \{v^n\}_{n=-1}^N$ be a sequence of functions such that $v^n \in \mathcal{V}(\Omega)$, $n = 0, 1, \dots, N$, and $v^{-1} = v^1 - v^0$, and consider the operator*

$$\Lambda(V) = V + \left\{0, 0, \frac{1}{4}\{\tau^{n+1}\Pi_h\left(\boldsymbol{a}\cdot\nabla v^{n+1}\right)\}_0^{N-1}\right\} + \delta t^{-1} d^{2,*}(V).$$

Then, the bilinear form $B_h^*(\cdot,\cdot)$ is Λ-coercive. In particular, the following inequality holds:

$$B_h^*(V, \Lambda(V)) \geq \beta_2\left(|||V|||_s^2 + \delta t^{-1}|||d^{1,*}(V)|||_w^2 + \frac{1}{T\delta t}\|v^{-1}\|_{L^2(\Omega)}^2\right)$$

for a certain constant $\beta_2 > 0$ independent of h.

Proof. It can be easily shown that

$$
\begin{aligned}
&B_h^*(V,4V)\\
&\quad=\sum_{n=0}^{N-1}4b_h\left(v^{n+1},v^{n+1}\right)+\sum_{n=2}^{N-1}\frac{4}{2T}\left(3v^{n+1}-4v^n+v^{n-1},v^{n+1}\right)\\
&\qquad+\frac{4}{T}\left(v^1-v^0,v^1\right)+\frac{4}{T}\left(v^0,v^0\right)+\frac{4}{T\delta t}\left(v^{-1},v^{-1}\right)\\
&\quad\geq\sum_{n=0}^{N-1}4\left[\delta t\nu\left\|\nabla v^{n+1}\right\|_{L^2(\Omega)}^2+\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\boldsymbol{a}\cdot\nabla v^{n+1}\right)\right\|_{L^2(\Omega)}^2\right]\\
&\qquad+\frac{1}{T}\left[\left\|v^{N+1}\right\|_{L^2(\Omega)}^2+\sum_{n=1}^{N-1}\left\|\delta^2v^{n+1}\right\|_{L^2(\Omega)}^2+2\left\|v^0\right\|_{L^2(\Omega)}^2+\frac{4}{\delta t}\|v^{-1}\|_{L^2(\Omega)}^2\right].
\end{aligned}
$$

In order to obtain stability for the component of the convective term in the finite element space, we use as test function the sequence $\{0,0,\{\tau^{n+1}\Pi_h(\boldsymbol{a}\cdot\nabla v^{n+1})\}_0^{N-1}\}=:\Pi_0(\tau,V)$ which starts with 0 in the components -1 and 0. Exactly as in the proof of Theorem 4.5, we now obtain

$$
\begin{aligned}
B_h^*\left(V,\Pi_0(\tau,V)\right)&\geq\sum_{n=0}^{N-1}\phi^{n+1}\delta t\tau^{n+1}\left\|\Pi_h\left(\boldsymbol{a}\cdot\nabla v^{n+1}\right)\right\|_{L^2(\Omega)}^2\\
&\quad-\sum_{n=0}^{N-1}\left[\delta t\nu\left\|\nabla v^{n+1}\right\|_{L^2(\Omega)}^2+\delta t\tau^{n+1}\left\|\Pi_h^{\perp}\left(\boldsymbol{a}\cdot\nabla v^{n+1}\right)\right\|_{L^2(\Omega)}^2\right]\\
&\quad-\sum_{n=1}^{N-1}\frac{1}{4T}\left\|3v^{n+1}-4v^n+v^{n-1}\right\|_{L^2(\Omega)}^2-\frac{1}{T}\left\|v^1-v^0\right\|_{L^2(\Omega)}^2,
\end{aligned}
\tag{4.22}
$$

with the expression of ϕ^{n+1} given in (4.8). We do not have control over the term related to the time derivative needing a further step. We now use as test function $d^{2,*}(V)$. From the first step in (4.21) it follows that

$$
\begin{aligned}
\delta t^{-1}B_h^*\left(V,4d^{2,*}(V)\right)&\geq\delta t^{-1}|||d^{1,*}(V)|||_w^2-\frac{3}{T\delta t}\left\|\delta v^1\right\|_{L^2(\Omega)}^2\\
&=\delta t^{-1}|||d^{1,*}(V)|||_w^2-\frac{3}{T\delta t}\left\|v^{-1}\right\|_{L^2(\Omega)}^2.
\end{aligned}
\tag{4.23}
$$

Combining the previous inequalities and invoking the Gronwall lemma (without any assumption over the time step size) we can conclude the proof of the theorem. ∎

Remark 4.4. In (4.22) we do not have control over the term associated to the time derivative. It makes the analysis for the second order method more intricate than for the first order method, for which the time derivative term is easily controlled (see (4.7)). The control of this term has motivated the introduction of $d^{2,*}(V)$ in the test sequence used.

In order to obtain stability it remains to prove some kind of continuity with respect to the operator Λ. This is what the next theorem states.

THEOREM 4.21 (Λ-continuity). *The following inequality holds:*

$$L^*(\Lambda(V)) \leq \left(\sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N-1} \frac{\delta t^2}{\nu} \left\| D_1 f^{n+1} \right\|^2_{H^{-1}(\Omega)} \right.$$
$$\left. + \frac{1}{T} \left\| u_0 \right\|^2_{L^2(\Omega)} + \frac{\delta t}{T} \left\| \frac{u_{1,h} - \Pi_h\left(u_0\right)}{\delta t} \right\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}$$
$$\times \left(|||V|||_s^2 + \delta t^{-1} |||d^{1,*}(V)|||_w^2 + \frac{1}{T} \left\| v^0 \right\|^2_{L^2(\Omega)} + \frac{1}{T\delta t} \| v^{-1} \|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}.$$

Proof. The following inequalities can be easily obtained:

$$L^*(V) \leq \left(\sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|^2_{H^{-1}(\Omega)} + \frac{1}{T} \left\| u_0 \right\|^2_{L^2(\Omega)} + \frac{\delta t}{T} \left\| \frac{u_{1,h} - \Pi_h\left(u_0\right)}{\delta t} \right\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}$$
$$\times \left(\sum_{n=0}^{N-1} \delta t \nu \left\| \nabla v^{n+1} \right\|^2_{L^2(\Omega)} + \frac{1}{T} \left\| v^0 \right\|^2_{L^2(\Omega)} + \frac{1}{T\delta t} \left\| v^{-1} \right\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}},$$

$$L^*(\delta t^{-1} d^{2,*}(V)) \leq \left(\sum_{n=1}^{N-1} \frac{\delta t^2}{\nu} \left\| D_1 f^{n+1} \right\|^2_{H^{-1}(\Omega)} \right)^{\frac{1}{2}} \left(\sum_{n=1}^{N-1} \delta t^2 \nu \left\| D_1 v^{n+1} \right\|^2_{H^1(\Omega)} \right)^{\frac{1}{2}},$$

$$L^*(\Pi_0(\tau, V)) \leq \left(\sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|^2_{H^{-1}(\Omega)} \right)^{\frac{1}{2}}$$
$$\times \left(\sum_{n=0}^{N-1} \delta t \nu \left\| \tau^{n+1} \nabla (\Pi_h \left(\boldsymbol{a} \cdot \nabla v^{n+1} \right)) \right\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}},$$

and

$$\nu \left\| \tau^{n+1} \nabla \left(\Pi_h \left(\boldsymbol{a} \cdot \nabla v^{n+1} \right) \right) \right\|^2_{L^2(\Omega)} \leq \frac{C_{\text{inv}}^2 \nu}{h^2} (\tau^{n+1})^2 \left\| \Pi_h \left(\boldsymbol{a} \cdot \nabla v^{n+1} \right) \right\|^2_{L^2(\Omega)}$$
$$\leq C \tau^{n+1} \left\| \boldsymbol{a} \cdot \nabla v^{n+1} \right\|^2_{L^2(\Omega)}.$$

From all these inequalities the theorem follows easily. ∎

The two previous theorems lead to the following stability result.

COROLLARY 4.22 (stability II). *The sequence U_h, solution of problem* (4.17), *is bounded as follows:*

$$|||U_h|||_s^2 + \delta t^{-1} |||d^{1,*} U_h|||_w^2$$
$$\leq C \left(\sum_{n=0}^{N-1} \frac{\delta t}{\nu} \left\| f^{n+1} \right\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N-1} \frac{\delta t^2}{\nu} \left\| D_1 f^{n+1} \right\|^2_{H^{-1}(\Omega)} \right.$$
$$\left. + \frac{1}{T} \left\| u^0 \right\|^2_{L^2(\Omega)} + \frac{\delta t}{T} \left\| \frac{u_{1,h} - \Pi_h\left(u_0\right)}{\delta t} \right\|^2_{L^2(\Omega)} \right)$$

for all $\delta t > 0$.

Obviously, this stability bound makes sense if the initialization is such that the last term on the right-hand side is bounded. Using, for example, the backward Euler scheme, it is easy to show that this last term is bounded if $h^{p+1} \leq CT\delta t$, and this condition is automatically satisfied thanks to assumption (2.21).

The final result we obtain is an error estimate in the *strong* norm $|||\cdot|||_s$. At this point we introduce the sequence $U = \{u_h^{-1}, u^0, u^1, u^2, \dots, u^N\}$, which consists of the sequence of solutions of the semidiscrete problem (2.9)–(2.11) supplemented with u_h^{-1} at $n = -1$. It can be easily checked that this sequence satisfies

$$B_h^*(U, V) = L^*(V) - \epsilon_c(V).$$

Thus, $E_d := U_h - U = \{0, u_h^0 - u^0, u_h^1 - u^1, \dots, u_h^N - u^N\}$ satisfies

$$B_h^*(E_d, V_h) = \epsilon_c(V_h).$$

We point out that for fixed domains the critical time step size does not appear anymore due to the fact that $\boldsymbol{w} = \mathbf{0}$. The method is unconditionally stable, as expected.

We stress the fact that $e_d^{-1} \neq e_d^1 - e_d^0$, and therefore E_d does not verify the statement of Theorem 4.20. The only place where the fact that $v^{-1} = v^1 - v^0$ is used is in (4.23). When the test sequence does not satisfy the assumption $v^{-1} = v^1 - v^0$ of Theorem 4.20, we have to modify the Λ-coercivity proved in this theorem as follows:

$$\frac{4}{T\delta t}\left\|\delta e_d^1\right\|_{L^2(\Omega)}^2 + B_h^*(E_d, \Lambda(E_d)) \geq \beta_2\left(|||E_d|||_s^2 + \delta t^{-1}|||d^{1,*}(E_d)|||_w^2\right). \tag{4.24}$$

With the expression of $\Lambda(\cdot)$ given in Theorem 4.20, we arrive at

$$\begin{aligned} B_h^*(E_d, \Lambda(E_d)) &= B_h^*(E_d, E_d) + \frac{1}{4}\epsilon_c(\Pi_0(\tau, E_d)) + \delta t^{-1} B_h^*\left(E_d, d^{2,*}E_d\right) \\ &= B_h^*(E_d, \Pi_h(U) - U) + \epsilon_c(U_h - \Pi_h(U)) + \frac{1}{4}\epsilon_c(\Pi_0(\tau, E_d)) \\ &\quad + \delta t^{-1} B_h^*\left(E_d, d^{2,*}(\Pi_h(U) - U)\right) + \delta t^{-1}\epsilon_c(d^{2,*}(U_h - \Pi_h(U))). \end{aligned} \tag{4.25}$$

Again, we group the different terms as interpolation and consistency errors and bound them separately in the next lemmas.

LEMMA 4.23 (interpolation error). *The following inequality holds:*

$$\begin{aligned} &B_h^*\left(E_d, \Pi_h^\perp(U)\right) + \delta t^{-1} B_h^*\left(E_d, d^{2,*}(\Pi_h^\perp(U)\right) \\ &\quad \leq \Bigg(|||E_d|||_w^2 + \delta t^{-1}|||d^{1,*}(E_d)|||_w^2 \\ &\qquad + h^{2(p+1)} \sum_{n=0}^{N-1} \delta t(\tau^{n+1})^{-1}\left(\left\|u^{n+1}\right\|_{H^{p+1}(\Omega)}^2 + \left\|\sqrt{\delta t} D_1 u^{n+1}\right\|_{H^{p+1}(\Omega)}^2\right)\Bigg)^{\frac{1}{2}} \\ &\qquad \times \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t(\tau^{n+1})^{-1}\left(\left\|u^{n+1}\right\|_{H^{p+1}(\Omega)}^2 + \left\|\sqrt{\delta t} D_1 u^{n+1}\right\|_{H^{p+1}(\Omega)}^2\right)\right)^{\frac{1}{2}}. \end{aligned}$$

Proof. The bound for the first term of the left-hand side of the inequality is easily obtained from the proof of Lemma 4.15, since $e_d^{-1} = 0$. For the second term we use property (4.21) and again the fact that $e_d^{-1} = 0$, getting

$$\begin{aligned} &B_h^*\left(E_d, d^{2,*}(\Pi_h^\perp(U))\right) \\ &\quad = B_h(d^{1,*}(E_d), d^{1,*}(\Pi_h^\perp(U))) - \frac{1}{2T}\left(\delta e_d^1, \delta\left(\Pi_h\left(u^3\right) - 3\Pi_h\left(u^2\right)\right)\right). \end{aligned}$$

Note that when we write $B_h(d^{1,*}(E_d), d^{1,*}(\Pi_h^\perp(U)))$ we eliminate the element -1 of the sequences to apply the bilinear form $B_h(\cdot,\cdot)$.

Using Lemma 4.14 we get

$$\begin{aligned}
&\delta t^{-1} B_h(d^{1,*}(E_d), d^{1,*}(\Pi_h^\perp(U))) \\
&\quad \leq C \left(\delta t^{-1} |||d^{1,*} E_d|||_w^2 + h^{2(p+1)} \sum_{n=1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right)^{\frac{1}{2}} \\
&\qquad \times \left(h^{2(p+1)} \sum_{n=1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

Exploiting the fact that $\Pi_h^\perp\left(e_d^1\right) = \Pi_h\left(u^1\right) - u^1$, we can easily get that

$$\begin{aligned}
&\frac{1}{2T\delta t} \left(\delta e_d^1, \delta \left(\Pi_h\left(u^3\right) - 3\Pi_h\left(u^2\right) \right) \right) \\
&\qquad \leq C h^{2(p+1)} \sum_{n=0}^{2} \delta t (\tau^{n+1})^{-1} \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2.
\end{aligned}$$

The proof is concluded. ∎

LEMMA 4.24 (consistency error). *The following inequality holds:*

$$\begin{aligned}
&\epsilon_c \left(U_h - \Pi_h(U) + \frac{1}{4} \Pi_0(\tau, E_d) + \delta t^{-1} d^{2,*}(U_h - \Pi_h(U)) \right) \\
&\leq C \left(h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left(\left\| u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 + \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right) \right)^{\frac{1}{2}} \\
&\times \left(|||E_d|||_w^2 + h^{2(p+1)} \sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left(\left\| u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 + \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right) \right)^{\frac{1}{2}}.
\end{aligned}$$

Proof. Due to the fact that $e_d^{-1} = 0$, we can profit from the bounds obtained in Lemmas 4.10 and 4.16. The remaining term associated to $d^{2,*}(\cdot)$ can be bounded as follows:

$$\begin{aligned}
\epsilon_c(\delta t^{-1} d^{2,*}(U_h - \Pi_h(U))) &= \sum_{n=1}^{N-1} \tau^{n+1} \left(\Pi_h^\perp \left(\boldsymbol{a} \cdot \nabla \delta u^{n+1} \right), \boldsymbol{a} \cdot \nabla \Pi_h^\perp \left(\delta u^{n+1} \right) \right) \\
&= \sum_{n=1}^{N-1} \tau^{n+1} \left(\Pi_h^\perp \left(\lambda(\delta u^{n+1}) \right), \boldsymbol{a} \cdot \nabla \Pi_h^\perp \left(\delta u^{n+1} \right) \right) \\
&\leq C \left(\sum_{n=1}^{N-1} \delta t \tau^{n+1} \left\| \Pi_h^\perp \left(\sqrt{\delta t} \lambda(D_1 u^{n+1}) \right) \right\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}} \\
&\quad \times \left(h^{2(p+1)} \sum_{n=1}^{N-1} \delta t (\tau^{n+1})^{-1} \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right)^{\frac{1}{2}},
\end{aligned}$$

with $\lambda(\cdot)$ introduced in Lemma 4.10. The term related to $\lambda(D_1 u^{n+1})$ can be easily bounded from the expression of τ^{n+1}, assumption (2.19), and the interpolation error estimate (4.3), as pointed out in Lemma 4.10. ∎

We end with the convergence result of the method in the norm $|||\cdot|||_s$.

THEOREM 4.25 (convergence II). *The sequence of errors $E_d = U_h - U$ satisfies the following error estimate:*

$$|||E_d|||_s^2 \leq Ch^{2(p+1)} \left[\sum_{n=0}^{N-1} \delta t (\tau^{n+1})^{-1} \left(\left\| u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 + \left\| \sqrt{\delta t} D_1 u^{n+1} \right\|_{H^{p+1}(\Omega)}^2 \right) \right.$$

$$\left. + (\tau^1)^{-1} \left\| u^1 \right\|_{H^{p+1}(\Omega)} + (\tau^1)^{-1} \left\| u^0 \right\|_{H^{p+1}(\Omega)} \right] \tag{4.26}$$

for all $\delta t > 0$.

Proof. Using Lemmas 4.23 and 4.24 in expressions (4.24) and (4.25), we can easily get the desired bound for $|||E_d|||_s^2$ in terms of $\frac{4}{T\delta t} \left\| \delta e_d^1 \right\|_{L^2(\Omega)}^2$. Using as initialization the backward Euler scheme and the convergence result of Theorem 4.11 for the semidiscrete problem, it follows that

$$\begin{aligned} \frac{1}{T\delta t} \left\| \delta e_d^1 \right\|_{L^2(\Omega)}^2 &\leq \frac{C}{T\delta t} \left(\left\| e_d^1 \right\|_{L^2(\Omega)}^2 + \left\| u^0 - \Pi_h(u^0) \right\|_{L^2(\Omega)}^2 \right) \\ &\leq C(\tau^1)^{-1} h^{2(p+1)} \left(\left\| u^1 \right\|_{H^{p+1}(\Omega)} + \left\| u^0 \right\|_{H^{p+1}(\Omega)} \right), \end{aligned}$$

from which we obtain the desired result. ∎

Remark 4.5. From (4.26) it is seen that we need $\{\sqrt{\delta t} D_1 u^{n+1}\}$ bounded in the norm of $\ell^2(H^{p+1}(\Omega))$. This can be understood as additional regularity on the data or as an additional assumption on the asymptotic behavior of the time step size in terms of h. From the semidiscrete equation, it is immediate to bound $\|D_1 u^{n+1}\|_{H^q(\Omega)}$ in terms of the $H^q(\Omega)$-norm of the rest of the terms of the equation. In particular, the viscous term implies that the $H^q(\Omega)$-norm of $D_1 u^{n+1}$ can be bounded in terms of the $H^{q+2}(\Omega)$-norm of u^{n+1}. If only the $H^{p+1}(\Omega)$-norm of u^{n+1} is bounded, we have to take $q = p-1$, and thus $h^{2(p+1)} \|\sqrt{\delta t} D_1 u^{n+1}\|_{H^{p+1}(\Omega)}^2$ has to be replaced by $h^{2(p-1)} \|\sqrt{\delta t} D_1 u^{n+1}\|_{H^{p-1}(\Omega)}^2$, and therefore we need $\delta t \leq Ch^4$ in order to maintain the optimal order of accuracy.

5. Conclusions. In this paper we have analyzed a stabilized FEM to approximate the convection-diffusion equation on moving domains. The OSS formulation has been used as a stabilization technique, and an ALE framework has been used in order to deal with moving domains.

In the first part of the paper we have analyzed the semidiscrete problem (in time). Two methods have been considered: a first order accurate method, where the time derivatives are computed using the BDF1 scheme, and a second order accurate method, where the BDF2 scheme has been used. In this analysis it is easy to identify the error introduced by the ALE formulation. The mesh velocity is computed as the time derivative of the mesh displacement. The numerical approximation of this time derivative is the only source of error introduced by the ALE formulation. As a conclusion, *in order to keep the accuracy of a kth order (in time) method on fixed domains, we must compute the mesh velocity using a time integration scheme of at least order k of accuracy.* The only negative aspect is that *unconditional stable methods for fixed domains become conditionally stable.*

In the second part of the paper we have analyzed a stabilized transient convection-diffusion equation in an ALE framework. We have introduced the concept of Λ-coercivity that has been used for obtaining stability results and error estimates. It has

been shown that the OSS method can be easily extended to transient problems. For the BDF1 time integration scheme we have stability of the convective term norm, as is usual when using stabilization techniques. The analysis of BDF2 is more complicated. We have control over only the *orthogonal projection* of the convective term. However, optimal convergence results with constants that do not depend on the Péclet number can be proved. Finally, for fixed domains, we have been able to recover *stronger* stability and convergence involving the full norm of the convective term, but the analysis is much more involved and requires more regularity assumptions.

REFERENCES

[1] S. Badia and R. Codina, *On some fluid-structure iterative algorithms using pressure segregation methods. Application to aeroelasticity*, Internat. J. Numer. Methods Engrg., submitted.

[2] P. Bochev, M. Gunzburger, and R. Lehoucq, *On stabilized finite element methods for the Stokes problem in the small time-step limit*, Internat. J. Numer. Methods Fluids, to appear.

[3] D. Boffi and L. Gastaldi, *Stability and geometric conservation laws for ALE formulation*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 4717–4739.

[4] M. Braack and E. Burman, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method*, SIAM J. Numer. Anal., 43 (2006), pp. 2544-2566.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[6] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[7] R. Codina, *Comparison of some finite element methods for solving the diffusion-convection-reaction equation*, Comput. Methods Appl. Mech. Engrg., 156 (1998), pp. 185–210.

[8] R. Codina, *Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1579–1599.

[9] R. Codina, *Pressure stability in fractional step finite element methods for incompressible flows*, J. Comput. Phys., 170 (2001), pp. 112–140.

[10] R. Codina, *A stabilized finite element method for generalized stationary incompressible flows*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2681–2706.

[11] R. Codina and S. Badia, *On some pressure segregation methods of fractional-step type for the finite element approximation of incompressible flow problems*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2900–2918.

[12] R. Codina and J. Blasco, *Analysis of a stabilized finite element approximation of the transient convection-diffusion-reaction equation using orthogonal subscales*, Comput. Vis. Sci., 4 (2002), pp. 167–174.

[13] R. Codina, J. Principe, O. Guasch, and S. Badia, *Time dependent subscales in the stabilized finite element approximation of incompressible flow problems*, Comput. Methods Appl. Mech. Engrg., submitted.

[14] J. Donea, P. Fasoli-Stella, and S. Giuliani, *Lagrangian and Eulerian finite element techniques for transient fluid structure interaction problems*, in Transactions of the Fourth SMIRT Conference, 1977, p. B1/2.

[15] C. Farhat, P. Geuzaine, and C. Grandmont, *The discrete geometric conservation law and the non-linear stability of ALE schemes for the solution of flow problems on moving grids*, J. Comput. Phys., 174 (2001), pp. 669–692.

[16] L. Formaggia and F. Nobile, *A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements*, East-West J. Numer. Math., 7 (1999), pp. 105–132.

[17] L. Formaggia and F. Nobile, *Stability analysis of second-order time accurate schemes for ALE-FEM*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 4097–4116.

[18] P. Geuzaine, C. Grandmont, and C. Farhat, *Design and analysis of ALE schemes with provable second-order time accuracy for inviscid and viscous flow simulations*, J. Comput. Phys., 191 (2003), pp. 206–227.

[19] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithm*, Springer-Verlag, Berlin, 1986.

[20] J. L. Guermond, *Subgrid stabilization of Galerkin approximations of linear contraction semigroups of class $\mathcal{C}^0$*, Comput. Vis. Sci., 2 (1999), pp. 131–138.

[21] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem. Part* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[22] T. J. R. Hughes, *Multiscale phenomena: Green's function, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized formulations*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[23] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J. B. Quincy, *The variational multiscale method—A paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[24] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics:* VIII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[25] F. Nobile, *Numerical Approximation of Fluid-Structure Interaction Problems with Application to Haemodynamics*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2001.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2198–2226

DISCONTINUOUS GALERKIN APPROXIMATION OF THE MAXWELL EIGENPROBLEM*

ANNALISA BUFFA† AND ILARIA PERUGIA‡

Abstract. A theoretical framework for the analysis of discontinuous Galerkin approximations of the Maxwell eigenproblem with discontinuous coefficients is presented. Necessary and sufficient conditions for a spurious-free approximation are established, and it is shown that, at least on conformal meshes, basically all the discontinuous Galerkin methods in the literature actually fit into this framework. Relations with the classical theory for conforming approximations are also discussed.

Key words. discontinuous Galerkin methods, Maxwell's equations, eigenvalue problems

AMS subject classifications. 65N30, 65N25

DOI. 10.1137/050636887

1. Introduction.

One of the most relevant problems in computational electromagnetics is the one of computing eigenfrequencies of the Maxwell equations in a cavity: find $\mathbf{u} \neq 0$ and ω such that

$$\nabla \times (\mu^{-1} \nabla \times \mathbf{u}) - \omega^2 \varepsilon \mathbf{u} = 0, \tag{1.1}$$

with suitable boundary conditions, where μ and ε are the magnetic permeability and the electric permittivity, respectively.

Finite element techniques are widely used to approximate problem (1.1), and, in recent years, a complete mathematical theory has been developed for *conforming* approximations, identifying the properties that the underlying finite element spaces need to fulfill in order to guarantee spurious-free approximations. We refer the reader to the pioneering work [12] and to [33] or [38] and the references therein (we point, in particular, to the fundamental papers [10], [18], [24], and [16]).

On the other hand, the use of discontinuous Galerkin (DG) methods in electromagnetism is attractive thanks to their flexibility in the mesh design and in the choice of shape functions. A unified presentation and analysis of all the DG methods available in the literature, in the elliptic context, are contained in [5], whereas the extension of these methods to the time-domain and frequency-domain Maxwell equations is the object of ongoing research (see, among others, [43], [34], [30], and [31]).

The main difficulties encountered in the analysis of DG approximations of the Maxwell equations are related to the lack of ellipticity and underlying compactness property of the Maxwell operator, which is "amplified" by the use of *nonconforming* approximation spaces.

The first studies on DG approximations of the Maxwell eigenproblem are contained in the recent papers [32] and [45]. There, the main goal was to investigate the role of the penalty parameter appearing in the local discontinuous Galerkin method in avoiding the pollution of the lowest part of the spectrum by eigenvalues related to the

*Received by the editors July 26, 2005; accepted for publication (in revised form) March 31, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sinum/44-5/63688.html

†Istituto di Matematica Applicata e Tecnologie Informatiche - CNR, Via Ferrata 1, 27100 Pavia, Italy (annalisa@imati.cnr.it).

‡Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (ilaria.perugia@unipv.it).

nonconformity of the approximation spaces for a fixed mesh size. That analysis approach and thorough numerical tests have highlighted the links between the spectral properties of DG and curl-conforming methods. In this paper, we aim at developing an *asymptotic* analysis (i.e., for mesh sizes which tend to zero) of DG approximations of the eigenproblem (1.1) in the spirit of [10], [18], [24], and [16].

The spectral theory for DG methods developed in [3] for elliptic problems (with associated compact inverse operators) needs to be extended to treat problems with noncompact inverse operators of the type (1.1). In this case, the lack of compactness results in the presence of an essential spectrum $\sigma^{ess} = \{0\}$, the eigenspace associated with the eigenvalue 0 being infinite dimensional. More precisely, we provide a general framework with a set of sufficient (and necessary) conditions for a DG method to provide a *spurious-free* approximation of problem (1.1), i.e., an approximation with the following properties (see [18]):

(i) *isolation of discrete kernel*; i.e., all discrete eigenvalues approaching the essential spectrum $\sigma^{ess} = \{0\}$ are separated from the other ones (see section 4.1 for a precise definition);
(ii) *nonpollution of the spectrum*; i.e., there are no discrete spurious eigenvalues;
(iii) *completeness of the spectrum*; i.e., all continuous eigenvalues smaller than an arbitrarily large fixed number are approximated for sufficiently fine meshes;
(iv) *nonpollution and completeness of the eigenspaces*; i.e., there are no spurious eigenfunctions, and the eigenspace approximations associated with eigenvalues which are not approaching $\sigma^{ess} = \{0\}$ have the right dimension.

The analysis presented in this paper is carried out along the lines of [18] and [16], and it is based on the theory developed in [25] and [26]. It is worth noting that our general framework applies to both hermitian and non-hermitian DG methods. The two key assumptions which ensure spurious-free DG approximations are (i) a discrete Friedrichs inequality (see Assumption 5) and (ii) a gap property (see Assumption 6). They are the DG analogue of the discrete Friedrichs inequality for discrete, weakly divergence-free curl-conforming vector fields and of the discrete compactness property (see, e.g., [37]), respectively, which have been proved to be necessary and sufficient conditions to have conforming spurious-free approximations to the Maxwell eigenproblem (1.1) (see [10] and [18]). Like for conforming approximations, we show the necessity of these assumptions restricting ourselves, for simplicity, to hermitian DG methods only.

We point out that our theory is able to treat general piecewise smooth material coefficients. In this respect, the appendix is devoted to the analysis of the approximation properties of the DG solutions under minimal regularity assumptions on the solutions of the corresponding continuous problem. This analysis is technical, extends the results of [42], and is, to our knowledge, new.

As a direct consequence of the spectral theory developed in this paper, we obtain well-posedness and quasi-optimal error estimates for DG discretizations, for sufficiently fine meshes, of the Maxwell source problem

$$\nabla \times (\mu^{-1} \nabla \times \mathbf{u}) - \omega^2 \varepsilon \mathbf{u} = \mathbf{f}, \tag{1.2}$$

with suitable boundary conditions, where ω is a fixed frequency, away from the eigen-

frequencies of the continuous problem. Indeed, the fact that a spurious-free finite element method is also a stable and convergent method for (1.2) is based on a general reasoning which is, to our knowledge, new.

Finally, applying our theory, we analyze the spectral approximation properties of several DG methods, such as the methods of the interior penalty family (interior penalty (IP), nonsymmetric interior penalty (NIP), and incomplete interior penalty (IIP); see [4], [44], and [23], respectively) and the local discontinuous Galerkin method (LDG; see [21]). Our theoretical results can be summarized as follows:

1. on conformal tetrahedral/triangular meshes, these methods are spurious-free when the approximation spaces are made of elementwise polynomials of degree ℓ in each variable as well as of elementwise Nédélec elements of the first family [39];
2. on conformal hexahedral/quadrilateral meshes, these methods are spurious-free when the approximation spaces are made of elementwise Nédélec elements of the first family, whereas they produce spurious modes when the approximation spaces are made of elementwise polynomials of degree ℓ in each variable;
3. the convergence rates of the eigenfunction approximations are optimal, i.e., for smooth solutions, $\mathcal{O}(h^{\ell})$ for elements of degree ℓ, whereas the convergence rates of the eigenvalue approximations are optimal, i.e., for smooth solutions, $\mathcal{O}(h^{2\ell})$ for hermitian DG methods, and suboptimal $\mathcal{O}(h^{\ell})$ for non-hermitian DG methods.

We point out that all the results obtained here for the DG spectral approximations of the *curl-curl* operator carry over to the DG spectral approximations of the *grad-div* operator encountered, for instance, in fluid-structure problems (see, e.g., [9] and [8]).

Some questions still remain open and are the object of ongoing research: (i) Can one use a mesh with hanging nodes? (ii) Can one use approximation spaces made of elementwise divergence-free polynomial spaces (see [6] and [20])? A partial answer to the first question has been provided by numerical tests performed while this paper was undergoing the review process (see [17]).

The paper is organized as follows: in sections 2 and 3 we set the notation and the definitions for the continuous and the discrete problems, respectively. Section 4 is the core of the paper and contains the analysis of the DG spectral approximation, under a minimal set of assumptions, which are indeed proved to also be necessary for spurious-free approximations in section 5. In section 6 we analyze the consequences of our theory on the Maxwell source problem (1.2), and finally in section 7 we apply our framework to the most used DG methods applied to the Maxwell equations. Here, the link between our assumptions and their conforming analogue is made clear for the interested reader. Finally, in section 8, we summarize our results.

2. Continuous problem. For a bounded domain D in $\mathbb{R}^d$, $d = 2, 3$, we denote by $H^s(D)$ the standard Sobolev space of order $s \geq 0$ of real or complex functions and by $\|\cdot\|_{s,D}$ the usual Sobolev norm. For $s = 0$, we write $L^2(D)$ in lieu of $H^0(D)$. We also use $\|\cdot\|_{s,D}$ to denote the norm for the space $H^s(D)^d$.

We denote by Ω the problem domain, which we assume to be a bounded Lipschitz polygonal or polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$, and by $\mathbf{n}$ the normal unit vector to its boundary $\partial\Omega$, pointing outside Ω. Whenever $\partial\Omega$ is not connected, we denote by Γ_i, $i = 1, \ldots, n_\Gamma$, its connected components.

If $d = 3$, we assume Ω to be occupied by inhomogeneous, anisotropic materials, i.e., for which the electric permittivity $\varepsilon = \varepsilon(\mathbf{x})$ and magnetic permeability $\mu = \mu(\mathbf{x})$

are second order, real, symmetric, tensor-valued functions, satisfying

$$0 < \varepsilon_\star(\mathbf{x}) \le \sum_{i,j=1}^{3} \varepsilon_{i,j}\xi_i\xi_j \le \varepsilon^\star(\mathbf{x}) \qquad \text{a.e. in } \Omega \quad \forall \xi \in \mathbb{R}^3,\ \|\xi\| = 1, \tag{2.1}$$

$$0 < \mu_\star(\mathbf{x}) \le \sum_{i,j=1}^{3} \mu_{i,j}\xi_i\xi_j \le \mu^\star(\mathbf{x}) \qquad \text{a.e. in } \Omega \quad \forall \xi \in \mathbb{R}^3,\ \|\xi\| = 1, \tag{2.2}$$

where $\varepsilon^\star,\ \varepsilon_\star,\ \mu^\star,\ \mu_\star \in L^\infty(\Omega)$. If $d = 2$, $\varepsilon = \varepsilon(\mathbf{x})$ is again a second order tensor, whereas $\mu = \mu(\mathbf{x})$ is a scalar; therefore, the conditions on ε are analogous to (2.1), whereas (2.2) becomes $0 < \mu_\star(\mathbf{x}) = \mu(\mathbf{x}) = \mu^\star(\mathbf{x})$. Finally, we assume that there exists a partition of Ω into Lipschitz subdomains such that in each of them ε, μ, and μ^{-1} are smooth.

We define, as usual, the following spaces of complex functions:

$$\begin{aligned}
H(\mathrm{curl};\Omega) &= \{\mathbf{v} \in L^2(\Omega)^d :\ \nabla \times \mathbf{v} \in L^2(\Omega)^{2d-3}\},\\
H_0(\mathrm{curl};\Omega) &= \{\mathbf{v} \in H(\mathrm{curl};\Omega) :\ \mathbf{n} \times \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\},\\
H_0(\mathrm{curl}^0;\Omega) &= \{\mathbf{v} \in H_0(\mathrm{curl};\Omega) :\ \nabla \times \mathbf{v} = \mathbf{0}\},\\
H(\mathrm{div}^0_\varepsilon;\Omega) &= \left\{\mathbf{v} \in L^2(\Omega)^d :\ \nabla \cdot (\varepsilon\mathbf{v}) = 0,\ \int_{\Gamma_i} (\varepsilon\mathbf{v}) \cdot \mathbf{n}\, ds = 0,\ i = 1,\dots,n_\Gamma\right\};
\end{aligned}$$

if $\partial\Omega$ is connected, then $H(\mathrm{div}^0_\varepsilon;\Omega) = \{\mathbf{v} \in L^2(\Omega)^d :\ \nabla \cdot (\varepsilon\mathbf{v}) = 0\}$. Moreover, we set

$$\mathbf{V} = H_0(\mathrm{curl};\Omega), \qquad \mathbf{V}^0 = H_0(\mathrm{curl}^0;\Omega), \qquad \mathbf{W} = \mathbf{V} \cap H(\mathrm{div}^0_\varepsilon;\Omega).$$

Finally, we denote by $(\cdot,\cdot)$ the standard inner product in $L^2(\Omega)^d$ given by $(\mathbf{u},\mathbf{v}) = \int_\Omega \mathbf{u} \cdot \overline{\mathbf{v}}\, d\mathbf{x}$ and write $L^2_\varepsilon(\Omega)^d$ for the space $L^2(\Omega)^d$ endowed with the ε-weighted inner product $(\mathbf{u},\mathbf{v})_\varepsilon = \int_\Omega \varepsilon\mathbf{u} \cdot \overline{\mathbf{v}}\, d\mathbf{x}$. The L^2-norm and the L^2_ε-norm are clearly equivalent, due to the assumptions on ε.

We endow $\mathbf{V}$ with the seminorm $|\mathbf{v}|_{\mathbf{V}} = \|\mu^{-1/2}\nabla \times \mathbf{v}\|_{0,\Omega}$, the inner product $(\mathbf{u},\mathbf{v})_{\mathbf{V}} = (\mu^{-1}\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) + (\mathbf{u},\mathbf{v})_\varepsilon$, and the norm $\|\mathbf{v}\|^2_{\mathbf{V}} = \|\mu^{-1/2}\nabla \times \mathbf{v}\|^2_{0,\Omega} + \|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega}$.

The following decompositions are L^2_ε-orthogonal (see [28]):

$$L^2(\Omega)^d = H(\mathrm{div}^0_\varepsilon;\Omega) \oplus \mathbf{V}^0, \qquad \mathbf{V} = \mathbf{W} \oplus \mathbf{V}^0. \tag{2.3}$$

Define the (hermitian) bilinear forms $a : \mathbf{V} \times \mathbf{V} \to \mathbb{C}$ and $b : \mathbf{V} \times \mathbf{V} \to \mathbb{C}$ as

$$\begin{aligned}
a(\mathbf{u},\mathbf{v}) &= (\mu^{-1}\nabla \times \mathbf{u}, \nabla \times \mathbf{v}),\\
b(\mathbf{u},\mathbf{v}) &= a(\mathbf{u},\mathbf{v}) + (\mathbf{u},\mathbf{v})_\varepsilon = (\mathbf{u},\mathbf{v})_{\mathbf{V}}.
\end{aligned}$$

The variational formulation of the eigenproblem we are interested in is the following: find $(\mathbf{0} \neq \mathbf{u}, \omega) \in \mathbf{W} \times \mathbb{C}$ such that

$$a(\mathbf{u},\mathbf{v}) = \omega^2(\mathbf{u},\mathbf{v})_\varepsilon \qquad \forall \mathbf{v} \in \mathbf{W}.$$

A standard way to discretize this problem consists in neglecting the constraint $\mathbf{u} \in \mathbf{W}$ and adding a zero frequency eigenspace corresponding to the *infinite-dimensional* space $\mathbf{V}^0$, leading to the following variational problem.

PROBLEM 1. *Find* $(\mathbf{0} \neq \mathbf{u}, \omega) \in \mathbf{V} \times \mathbb{C}$*:*

$$a(\mathbf{u}, \mathbf{v}) = \omega^2 (\mathbf{u}, \mathbf{v})_\varepsilon \qquad \forall \mathbf{v} \in \mathbf{V}.$$

Clearly, $\omega^2 = 0$ is an eigenvalue of Problem 1 with associated eigenspace $\mathbf{V}^0$. Moreover, the eigenvalue $\omega^2 = 0$ is isolated, all the other eigenvalues are real, positive, and isolated and form a sequence accumulating only at $+\infty$, and their associated eigenspaces are finite dimensional. Finally, eigenspaces associated with different eigenvalues are L^2_ε-orthogonal and $\mathbf{V}$-orthogonal (see, e.g., [38, Section 4.7]).

For the purpose of the analysis, following [18], we introduce the following auxiliary eigenproblem with a positive definite operator.

PROBLEM 2. *Find* $(\mathbf{0} \neq \mathbf{u}, \widetilde{\omega}) \in \mathbf{V} \times \mathbb{C}$*:*

$$b(\mathbf{u}, \mathbf{v}) = \widetilde{\omega}^2 (\mathbf{u}, \mathbf{v})_\varepsilon \qquad \forall \mathbf{v} \in \mathbf{V}.$$

The eigenvalues of Problem 1 and those of Problem 2 are such that $\widetilde{\omega}^2 = \omega^2 + 1$; thus, $\widetilde{\omega}^2 = 1$ is an eigenvalue of Problem 2 with infinite multiplicity and associated eigenspace $\mathbf{V}^0$.

Define the solution operator $A : L^2(\Omega)^d \to \mathbf{V}$ as follows: given $\mathbf{f} \in L^2(\Omega)^d$, $A\mathbf{f}$ is the (unique) element of $\mathbf{V}$ which satisfies

$$b(A\mathbf{f}, \mathbf{v}) = (\mathbf{f}, \mathbf{v})_\varepsilon \qquad \forall \mathbf{v} \in \mathbf{V}.$$

We have that $A \in \mathcal{L}(L^2(\Omega)^d, \mathbf{V})$. Notice that $(\mathbf{u}, \omega)$ is an eigenpair of Problem 1 if and only if $(\mathbf{u}, \lambda = \frac{1}{\omega^2+1})$ is an eigenpair of A.

Denote by $\sigma(A)$ and $\rho(A)$ the spectrum and the resolvent set (in the complex plane), respectively, of the solution operator A. Finally, for any $z \in \rho(A)$, we define the resolvent operator $R_z(A) = (z - A)^{-1}$ from $\mathbf{V}$ to $\mathbf{V}$.

3. Discontinuous Galerkin approximation: Assumptions. Let $\mathcal{T}_h$ be a shape-regular, not necessarily conformal, triangular ($d = 2$) or tetrahedral ($d = 3$) mesh aligned with the possible discontinuities of ε and μ. We suppose that there exists a $\overline{\mu} > 0$, independent of the mesh size, such that

$$\max_{\mathbf{x} \in K} \frac{\mu^\star(\mathbf{x})}{\mu_\star(\mathbf{x})} \leq \overline{\mu} \qquad \forall\, K \in \mathcal{T}_h. \tag{3.1}$$

We consider a complex vector-valued DG finite element space $\mathbf{V}_h$ (i.e., a discontinuous piecewise polynomial space on $\mathcal{T}_h$) and define the sum space $\mathbf{V}(h) = \mathbf{V} + \mathbf{V}_h$.

Given a seminorm $|\cdot|_{\mathbf{V}(h)}$ on $\mathbf{V}(h)$, we endow both $\mathbf{V}_h$ and $\mathbf{V}(h)$ with the norm

$$\|\mathbf{v}\|^2_{\mathbf{V}(h)} = |\mathbf{v}|^2_{\mathbf{V}(h)} + \|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega} \qquad \forall \mathbf{v} \in \mathbf{V}(h),$$

which we assume to be Hilbertian; we denote by $(\cdot,\cdot)_{\mathbf{V}(h)}$ the associated inner product.

Let $a_h : \mathbf{V}_h \times \mathbf{V}_h \to \mathbb{C}$ be the DG bilinear form obtained by discretizing $a : \mathbf{V} \times \mathbf{V} \to \mathbb{C}$ by a DG method, and define

$$b_h(\mathbf{u}, \mathbf{v}) = a_h(\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{v})_\varepsilon \qquad \forall \mathbf{u}, \mathbf{v} \in \mathbf{V}_h.$$

In this section we formulate general assumptions on the space $\mathbf{V}_h$ and on the bilinear form $a_h(\cdot,\cdot)$ under which our theory is developed.

Assumption 1 (norm compatibility). If $\mathbf{v} \in \mathbf{V}(h)$ and $|\mathbf{v}|_{\mathbf{V}(h)} = 0$, then $\mathbf{v} \in \mathbf{V}^0$; moreover, if $\mathbf{v} \in \mathbf{V}$, then $|\mathbf{v}|_{\mathbf{V}(h)} = |\mathbf{v}|_{\mathbf{V}}$.

Notice that Assumption 1 implies that $|\mathbf{v}|_{\mathbf{V}(h)} = 0$ if and only if $\mathbf{v} \in \mathbf{V}^0$. The space $\mathbf{V}(h)$ is a Hilbert space and the $\mathbf{V}(h)$-norm coincides with the $\mathbf{V}$-norm on $\mathbf{V}$.

For the DG space $\mathbf{V}_h$, we make the following approximation assumption.

Assumption 2 (approximation property of $\mathbf{V}_h$). There holds

$$\lim_{h\to 0} \inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\mathbf{v} - \mathbf{v}_h\|_{\mathbf{V}(h)} = 0 \qquad \forall \mathbf{v} \in \mathbf{W}.$$

We assume the following properties to be satisfied.

Assumption 3 (coercivity in seminorm and continuity). There exist positive constants α, γ independent of the mesh size such that

$$\begin{aligned} \text{Re}\,[a_h(\mathbf{v},\mathbf{v})] &\geq \alpha\,|\mathbf{v}|^2_{\mathbf{V}(h)} && \forall \mathbf{v} \in \mathbf{V}_h, \\ |a_h(\mathbf{u},\mathbf{v})| &\leq \gamma \|\mathbf{u}\|_{\mathbf{V}(h)} \|\mathbf{v}\|_{\mathbf{V}(h)} && \forall \mathbf{u},\mathbf{v} \in \mathbf{V}_h. \end{aligned}$$

Define the kernel of $a_h(\cdot,\cdot)$ and its $\mathbf{V}(h)$-orthogonal complement as follows:

$$\begin{aligned} K_h &= \{\mathbf{v} \in \mathbf{V}_h : a_h(\mathbf{v},\mathbf{w}) = 0 \quad \forall \mathbf{w} \in \mathbf{V}_h\}, \\ K_h^\perp &= \{\mathbf{v} \in \mathbf{V}_h : (\mathbf{v},\mathbf{w})_{\mathbf{V}(h)} = 0 \quad \forall \mathbf{w} \in K_h\}. \end{aligned}$$

If $a_h(\cdot,\cdot)$ is non-hermitian, we also assume that left and right kernels coincide, i.e.,

$$a_h(\mathbf{v},\mathbf{w}) = 0 \qquad \forall \mathbf{v} \in \mathbf{V}_h, \mathbf{w} \in K_h. \tag{3.2}$$

Remark 3.1. From Assumption 3 it follows that

$$\text{Re}\,[b_h(\mathbf{v},\mathbf{v})] \geq \min\{\alpha, 1\} \|\mathbf{v}\|^2_{\mathbf{V}(h)} \qquad \forall \mathbf{v} \in \mathbf{V}_h \tag{3.3}$$

and that

$$|\mathbf{v}|_{\mathbf{V}(h)} = 0 \qquad \forall \mathbf{v} \in K_h. \tag{3.4}$$

The coercivity property (3.3) guarantees that, for any given $\mathbf{f} \in L^2(\Omega)^d$, there exists a unique $\mathbf{u}_h \in \mathbf{V}_h$ such that $b_h(\mathbf{u}_h,\mathbf{v}) = (\mathbf{f},\mathbf{v})_\varepsilon$ for all $\mathbf{v} \in \mathbf{V}_h$, and $\|\mathbf{u}_h\|_{\mathbf{V}(h)} \leq C\|\mathbf{f}\|_{0,\Omega}$, with $C > 0$ independent of the mesh size and of the right-hand side $\mathbf{f}$. The identity (3.4), together with Assumption 1, implies that $K_h \subset \mathbf{V}^0$; consequently,

$$K_h^\perp = \{\mathbf{v} \in \mathbf{V}_h : (\mathbf{v},\mathbf{w})_{\mathbf{V}(h)} = (\mathbf{v},\mathbf{w})_\varepsilon = 0 \quad \forall \mathbf{w} \in K_h\}.$$

For the following assumption on the DG method, we introduce the broken spaces

$$\begin{aligned} H^s(\mathcal{T}_h)^d &= \{\mathbf{v} \in L^2(\Omega)^d : \mathbf{v}|_K \in H^s(K)^d \quad \forall K \in \mathcal{T}_h\} \quad \text{for } s \geq 0, \\ H^r(\text{curl};\mathcal{T}_h) &= \{\mathbf{v} \in L^2(\Omega)^d : \varepsilon\mathbf{v}|_K \in H^r(K)^d, \\ &\qquad \mu^{-1}\nabla\times\mathbf{v}|_K \in H^r(K)^{2d-3} \quad \forall K \in \mathcal{T}_h\} \quad \text{for } r > 0 \end{aligned}$$

and the norms

$$\begin{aligned} \|\mathbf{v}\|^2_{H^s(\mathcal{T}_h)^d} &= \sum_{K\in\mathcal{T}_h} \|\mathbf{v}\|^2_{s,K}, \\ \|\mathbf{v}\|^2_{H^r(\text{curl};\mathcal{T}_h)} &= \sum_{K\in\mathcal{T}_h} \left(\|\varepsilon^{1/2}\mathbf{v}\|^2_{r,K} + \|\mu^{-1/2}\nabla\times\mathbf{v}\|^2_{r,K} \right). \end{aligned}$$

Assumption 4 (convergence). Let $\mathbf{f}$ be in $H(\mathrm{div}^0_\varepsilon;\Omega)$; denote by $\mathbf{u}_s \in \mathbf{V}$ the solution to the coercive source problem $b(\mathbf{u}_s,\mathbf{v}) = (\mathbf{f},\mathbf{v})_\varepsilon$ for all $\mathbf{v} \in \mathbf{V}$ and by $\mathbf{u}_h \in \mathbf{V}_h$ its Galerkin projection which satisfies $b_h(\mathbf{u}_h,\mathbf{v}) = (\mathbf{f},\mathbf{v})_\varepsilon$ for all $\mathbf{v} \in \mathbf{V}_h$. Whenever $\mathbf{u}_s \in H^r(\mathrm{curl};\mathcal{T}_h)$, with $r > 0$, and $\mathbf{f} \in H^s(\mathcal{T}_h)^d$, with $s \geq 0$, then

$$\exists\, t > 0 : \qquad \|\mathbf{u}_s - \mathbf{u}_h\|_{\mathbf{V}(h)} \leq C h^t \left(\|\mathbf{u}_s\|_{H^r(\mathrm{curl};\mathcal{T}_h)} + \|\mathbf{f}\|_{H^s(\mathcal{T}_h)^d} \right), \tag{3.5}$$

where $C > 0$ is independent of the mesh size. The bound (3.5), together with the regularity results in [22], implies that

$$\exists\, \sigma > 0 : \qquad \|\mathbf{u}_s - \mathbf{u}_h\|_{\mathbf{V}(h)} \leq C h^\sigma \|\mathbf{f}\|_{0,\Omega} \qquad \forall \mathbf{f} \in H(\mathrm{div}^0_\varepsilon;\Omega),$$

where $C > 0$ is independent of the mesh size.

For the most common DG methods, the proof that Assumption 4 holds true makes use of results proved in the appendix (see Proposition 7.3).

We define the DG solution operator $A_h : L^2(\Omega)^d \to \mathbf{V}_h$ as follows: given $\mathbf{f} \in L^2(\Omega)^d$, $A_h\mathbf{f}$ is the (unique) element of $\mathbf{V}_h$ which satisfies

$$b_h(A_h\mathbf{f},\mathbf{v}) = (\mathbf{f},\mathbf{v})_\varepsilon \qquad \forall \mathbf{v} \in \mathbf{V}_h.$$

The operator A_h is well defined and $A_h \in \mathcal{L}(L^2(\Omega)^d, \mathbf{V}_h)$ (see Remark 3.1).

As in the continuous case, we denote by $\sigma(A_h)$ and $\rho(A_h)$ the spectrum and the resolvent set, respectively, of the DG solution operator A_h. Finally, for any $z \in \mathbb{C}$, we formally define the resolvent operator $R_z(A_h) = (z - A_h)^{-1}$ from $\mathbf{V}_h$ to $\mathbf{V}_h$.

The previous assumptions imply the following properties of the discrete eigenvalues and eigenfunctions.

PROPOSITION 3.2. *If $\lambda_h \in \sigma(A_h)$, then $0 < \mathrm{Re}\,[\lambda_h] \leq 1$. Moreover, $1 \in \sigma(A_h)$ and its associated eigenspace is K_h.*

Proof. Let $\mathbf{v} \neq \mathbf{0}$ be an eigenfunction associated with $\lambda_h \in \sigma(A_h)$. We have

$$\begin{aligned} \overline{\lambda}_h \|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega} &= (\mathbf{v},\lambda_h\mathbf{v})_\varepsilon = b_h(A_h\mathbf{v},\lambda_h\mathbf{v}) = b_h(\lambda_h\mathbf{v},\lambda_h\mathbf{v}) \\ &= a_h(\lambda_h\mathbf{v},\lambda_h\mathbf{v}) + |\lambda_h|^2 \|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega}, \end{aligned}$$

and thus $\mathrm{Re}\,[\lambda_h]\|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega} = \mathrm{Re}\,[a_h(\lambda_h\mathbf{v},\lambda_h\mathbf{v})] + |\lambda_h|^2\|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega}$. Since, owing to Assumption 3, $\mathrm{Re}\,[a_h(\lambda_h\mathbf{v},\lambda_h\mathbf{v})] \geq 0$, we readily have $\mathrm{Re}\,[\lambda_h] > 0$ and $\mathrm{Re}\,[\lambda_h] \geq |\lambda_h|^2$, from which $\mathrm{Re}\,[\lambda_h] \leq 1$. The second part of the statement is obvious. ∎

Clearly, whenever $a_h(\cdot,\cdot)$ is hermitian, then all the discrete eigenvalues are real.

PROPOSITION 3.3. (i) *Let $\mathbf{v} \neq 0$ be an eigenfunction of A_h associated with an eigenvalue $\lambda_h \neq 1$. Then $(\mathbf{v},\mathbf{w})_\varepsilon = (\mathbf{v},\mathbf{w})_{\mathbf{V}(h)} = b_h(\mathbf{v},\mathbf{w}) = 0$ for all $\mathbf{w} \in K_h$.*

(ii) *If $a_h(\cdot,\cdot)$ is hermitian, for all eigenfunctions $\mathbf{v}_1$, $\mathbf{v}_2$ associated with different eigenvalues, it holds that $(\mathbf{v}_1,\mathbf{v}_2)_\varepsilon = b_h(\mathbf{v}_1,\mathbf{v}_2) = 0$.*

Proof. For the proof of (i), let $\mathbf{w} \in K_h$. Since $a_h(\mathbf{v},\mathbf{w}) = 0$, we can write

$$(\mathbf{v},\mathbf{w})_\varepsilon = \lambda_h b_h(\mathbf{v},\mathbf{w}) = \lambda_h(\mathbf{v},\mathbf{w})_\varepsilon;$$

then, $\lambda_h \neq 1$ implies $(\mathbf{v},\mathbf{w})_\varepsilon = 0$ and $b_h(\mathbf{v},\mathbf{w}) = 0$. Moreover, since $|\mathbf{w}|_{\mathbf{V}(h)} = 0$, we also have $(\mathbf{v},\mathbf{w})_{\mathbf{V}(h)} = (\mathbf{v},\mathbf{w})_\varepsilon$. The proof of (ii) is trivial. ∎

4. Spurious-free discontinuous Galerkin approximations. In order to guarantee a *spurious-free* DG approximation to Problem 1 (see the introduction), in addition to Assumptions 1–4, we need to make sure that the two additional properties are verified.

PROPERTY 1 (isolation of discrete kernel). *There exists* $0 < \beta < 1$ *independent of the mesh size such that if* $1 \neq \lambda_h \in \sigma(A_h)$*, then*

$$\text{Re } [\lambda_h] \leq \beta.$$

For a linear, continuous operator $L : V_1 \to V_2$, with V_1 and V_2 Hilbert spaces, we define

$$\|L\|_{\mathcal{L}(V_1,V_2)} = \sup_{\substack{v \in V_1 \\ \|v\|_{V_1}=1}} \|Lv\|_{V_2}.$$

PROPERTY 2 (convergence in mesh-dependent norm).

$$\lim_{h \to 0} \|A - A_h\|_{\mathcal{L}(\mathbf{V}_h, \mathbf{V}(h))} = 0.$$

We remark that Property 2 is the DG analogue of [15, P1, p. 100] and that the norm $\|\cdot\|_{\mathcal{L}(\mathbf{V}_h,\mathbf{V}(h))}$ coincides with the mesh-dependent norm $\|\cdot\|_h$ of [25].

In the following two sections we formulate key assumptions on the DG spaces and bilinear forms which guarantee the validity of Properties 1 and 2, respectively.

4.1. Isolation of discrete kernel. We prove that the following assumption implies Property 1.

Assumption 5 (discrete Friedrichs inequality). There exists $C > 0$ independent of the mesh size such that

$$\|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega}^2 \leq C \,\text{Re } [a_h(\mathbf{v}, \mathbf{v})] \qquad \forall \mathbf{v} \in K_h^{\perp}.$$

PROPOSITION 4.1. *Assumption* 5 *implies Property* 1.

Proof. If $\mathbf{v}$ is an eigenfunction of A_h associated with an eigenvalue $\lambda_h \neq 1$, then $\mathbf{v}$ belongs to $K_h^{\perp}$ (see Proposition 3.3). From Assumption 5 we have

$$\begin{aligned}\frac{|\lambda_h|^2}{C} \|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega}^2 &\leq \text{Re } [a_h(\lambda_h\mathbf{v}, \lambda_h\mathbf{v})] = \text{Re } [a_h(A_h\mathbf{v}, \lambda_h\mathbf{v})] \\ &= \text{Re } [b_h(A_h\mathbf{v}, \lambda_h\mathbf{v}) - (A_h\mathbf{v}, \lambda_h\mathbf{v})_\varepsilon] \\ &= \text{Re } [(\mathbf{v}, \lambda_h\mathbf{v})_\varepsilon - (\lambda_h\mathbf{v}, \lambda_h\mathbf{v})_\varepsilon] = (\text{Re } [\lambda_h] - |\lambda_h|^2)\|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega}^2.\end{aligned}$$

Property 1 readily follows with $\beta = C/(1 + C)$. ∎

Remark 4.2. The $\mathbf{V}(h)$-ellipticity of $a_h(\cdot, \cdot)$ on $K_h^{\perp}$ follows from Assumptions 3 and 5. In fact, if $\mathbf{v} \in K_h^{\perp}$, the definition of $\|\cdot\|_{\mathbf{V}(h)}$, Assumptions 3 and 5 give

$$\|\mathbf{v}\|_{\mathbf{V}(h)}^2 = |\mathbf{v}|_{\mathbf{V}(h)}^2 + \|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega}^2 \leq \left(\frac{1}{\alpha} + C\right) \text{Re } [a_h(\mathbf{v}, \mathbf{v})].$$

4.2. Convergence of solution operators in mesh-dependent norm. First, we note that Property 2 can be rephrased as follows: for all h small enough,

$$\|(A - A_h)\mathbf{f}_h\|_{\mathbf{V}(h)} \leq \xi_h \|\mathbf{f}_h\|_{\mathbf{V}(h)} \qquad \forall \mathbf{f}_h \in \mathbf{V}_h, \tag{4.1}$$

with $\xi_h \to 0$ as $h \to 0$.

The aim of this section is to prove that the following key assumption implies Property 2.

Assumption 6 (gap property). For all h small enough, for any $\mathbf{w}_h \in K_h^\perp$ there exists $\mathbf{w} = \mathbf{w}(h) \in H(\mathrm{div}_\varepsilon^0;\Omega)$ such that

$$\|\mathbf{w} - \mathbf{w}_h\|_{0,\Omega} \le \eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)},$$

with $\eta_h \to 0$ as $h \to 0$.

Assumption 6 is related to the approximation properties of $K_h^\perp$ and K_h in $\mathbf{W}$ and in $\mathbf{V}^0$, respectively (see section 5).

In order to prove Property 2, we state the following lemma.

LEMMA 4.3. *For all* $\mathbf{f}_h^0 \in K_h$, *we have* $(A - A_h)\mathbf{f}_h^0 = 0$.

Proof. The condition $\mathbf{f}_h^0 \in K_h$ implies that $\mathbf{f}_h^0 \in \mathbf{V}_h \cap \mathbf{V}^0$ (see Remark 3.1). Therefore, $b(A\mathbf{f}_h^0, \mathbf{v}) = (\mathbf{f}_h^0, \mathbf{v})_\varepsilon$ for all $\mathbf{v} \in \mathbf{V}$ implies that $A\mathbf{f}_h^0$ is solution to $b(\mathbf{u}, \mathbf{v}) = (\mathbf{f}_h^0, \mathbf{v})_\varepsilon$ for all $\mathbf{v} \in \mathbf{V}$. Since $\mathbf{f}_h^0 \in \mathbf{V}^0$, then $\mathbf{u} = \mathbf{f}_h^0$ is a solution; uniqueness implies that $A\mathbf{f}_h^0 = \mathbf{f}_h^0$. Therefore, we need only prove that $A_h\mathbf{f}_h^0 = \mathbf{f}_h^0$. But $a_h(\mathbf{f}_h^0, \mathbf{v}) = 0$ for all $\mathbf{v} \in \mathbf{V}_h$ implies that

$$b_h(A_h\mathbf{f}_h^0, \mathbf{v}) = (\mathbf{f}_h^0, \mathbf{v})_\varepsilon = b_h(\mathbf{f}_h^0, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{V}_h,$$

from which $A_h\mathbf{f}_h^0 = \mathbf{f}_h^0$, owing to the well-posedness in Remark 3.1, and the proof is complete. ∎

PROPOSITION 4.4. *Property* 2 *holds true.*

Proof. Decompose $\mathbf{f}_h \in \mathbf{V}_h$ as $\mathbf{f}_h = \mathbf{f}_h^0 + \mathbf{f}_h^\perp$, with $\mathbf{f}_h^0 \in K_h$ and $\mathbf{f}_h^\perp \in K_h^\perp$ and $\|\mathbf{f}_h\|_{\mathbf{V}(h)}^2 = \|\mathbf{f}_h^0\|_{\mathbf{V}(h)}^2 + \|\mathbf{f}_h^\perp\|_{\mathbf{V}(h)}^2$. Owing to Lemma 4.3, it is enough to prove that, for all h small enough,

$$\|(A - A_h)\mathbf{f}_h^\perp\|_{\mathbf{V}(h)} \le \xi_h \|\mathbf{f}_h^\perp\|_{\mathbf{V}(h)} \qquad \forall \mathbf{f}_h^\perp \in K_h^\perp, \tag{4.2}$$

with $\xi_h \to 0$ as $h \to 0$. For h small enough, we can write

$$\|(A - A_h)\mathbf{f}_h^\perp\|_{\mathbf{V}(h)} \le \|(A - A_h)(\mathbf{f} - \mathbf{f}_h^\perp)\|_{\mathbf{V}(h)} + \|(A - A_h)\mathbf{f}\|_{\mathbf{V}(h)}, \tag{4.3}$$

with $\mathbf{f} \in H(\mathrm{div}_\varepsilon^0;\Omega)$ as in Assumption 6.

For the first term on the right-hand side in (4.3), we have

$$\begin{aligned}\|(A - A_h)(\mathbf{f} - \mathbf{f}_h^\perp)\|_{\mathbf{V}(h)} &\le \left(\|A\|_{\mathcal{L}(L^2(\Omega)^d,\mathbf{V})} + \|A_h\|_{\mathcal{L}(L^2(\Omega)^d,\mathbf{V}_h)}\right) \|\mathbf{f} - \mathbf{f}_h^\perp\|_{0,\Omega} \\ &\le C\,\eta_h\,\|\mathbf{f}_h^\perp\|_{\mathbf{V}(h)},\end{aligned}$$

owing to the continuity of A_h (see Remark 3.1) and Assumption 6.

For the second term on the right-hand side in (4.3), since $\mathbf{f} \in H(\mathrm{div}_\varepsilon^0;\Omega)$, from Assumption 4 we have that there exists a $\sigma > 0$ such that

$$\begin{aligned}\|(A - A_h)\mathbf{f}\|_{\mathbf{V}(h)} &\le Ch^\sigma \|\mathbf{f}\|_{0,\Omega} \le Ch^\sigma(\|\mathbf{f} - \mathbf{f}_h^\perp\|_{0,\Omega} + \|\mathbf{f}_h^\perp\|_{0,\Omega}) \\ &\le Ch^\sigma(\eta_h + 1)\|\mathbf{f}_h^\perp\|_{\mathbf{V}(h)},\end{aligned}$$

where we have again used Assumption 6 and the definition of the $\mathbf{V}(h)$-norm.

Therefore, (4.2) holds true with $\xi_h = h^\sigma(\eta_h + 1)$. ∎

4.3. Nonpollution of the spectrum. This section is devoted to the proof of the following theorem.

THEOREM 4.5 (nonpollution of the spectrum). *Let* $G \subset \mathbb{C}$ *be an open set containing* $\sigma(A)$. *Then, for* h *small enough,* $\sigma(A_h) \subset G$.

We proceed by establishing few intermediate results.

LEMMA 4.6. *Fix $0 \neq z \in \rho(A)$. There exists a positive constant C depending only upon Ω and $|z|$ such that, for all $\mathbf{f} \in \mathbf{V}(h)$,*

$$\|(z - A)\mathbf{f}\|_{\mathbf{V}(h)} \geq C\|\mathbf{f}\|_{\mathbf{V}(h)}.$$

Proof. The proof is similar to the one of [3, Lemma 4.2]. Let $\mathbf{f} \in \mathbf{V}(h)$ and $\mathbf{g} := (z - A)\mathbf{f}$. By construction, $\mathbf{g} \in \mathbf{V}(h)$ and $z\mathbf{f} - \mathbf{g} \in \mathbf{V}$. Moreover, $(z\mathbf{f} - \mathbf{g})$ solves

$$B^{-1}(z\mathbf{f} - \mathbf{g}) - \frac{1}{z}(z\varepsilon\mathbf{f} - \varepsilon\mathbf{g}) = \frac{1}{z}\varepsilon\mathbf{g},$$

where B^{-1} is the operator $\nabla \times (\mu^{-1}\nabla \times (\cdot)) + \varepsilon(\cdot)$. Since $z \in \rho(A)$, and $z\mathbf{f} - \mathbf{g}$ verifies homogeneous Dirichlet boundary condition, well-posedness implies that

$$\|z\mathbf{f} - \mathbf{g}\|_{\mathbf{V}} \leq \frac{C}{|z|}\|\varepsilon^{1/2}\mathbf{g}\|_{0,\Omega} \leq \frac{C}{|z|}\|\mathbf{g}\|_{\mathbf{V}(h)}.$$

Owing to Assumption 1, it holds that $\|z\mathbf{f} - \mathbf{g}\|_{\mathbf{V}} = \|z\mathbf{f} - \mathbf{g}\|_{\mathbf{V}(h)}$. Therefore

$$\|\mathbf{f}\|_{\mathbf{V}(h)} \leq \frac{1}{|z|}\left(\|z\mathbf{f} - \mathbf{g}\|_{\mathbf{V}} + \|\mathbf{g}\|_{\mathbf{V}(h)}\right) \leq C(|z|)\|\mathbf{g}\|_{\mathbf{V}(h)}. \qquad \square$$

THEOREM 4.7. *Fix $0 \neq z \in \rho(A)$. For h small enough, there exists a positive constant C depending only upon Ω and $|z|$ such that, for all $\mathbf{f} \in \mathbf{V}_h$,*

$$\|(z - A_h)\mathbf{f}\|_{\mathbf{V}(h)} \geq C\|\mathbf{f}\|_{\mathbf{V}(h)}.$$

Proof. By triangle inequality, we have

$$\|(z - A_h)\mathbf{f}\|_{\mathbf{V}(h)} \geq \|(z - A)\mathbf{f}\|_{\mathbf{V}(h)} - \|(A - A_h)\mathbf{f}\|_{\mathbf{V}(h)}.$$

Lemma 4.6 and the continuity of the operator $A - A_h$ yield

$$\|(z - A_h)\mathbf{f}\|_{\mathbf{V}(h)} \geq (C - \|A - A_h\|_{\mathcal{L}(\mathbf{V}_h, \mathbf{V}(h))})\|\mathbf{f}\|_{\mathbf{V}(h)},$$

and Property 2 allows us to conclude. $\square$

Theorem 4.7 implies that, for any $0 \neq z \in \rho(A)$ and h small enough, $(z - A_h)$ is an invertible operator and the following result holds true.

COROLLARY 4.8. *Let $F \subset \rho(A)$ be closed. Then, there exists a positive constant C independent of the mesh size such that, for h small enough, we have*

$$\|R_z(A_h)\|_{\mathcal{L}(\mathbf{V}_h, \mathbf{V}_h)} \leq C$$

for all $z \in F$, with $C > 0$ independent of the mesh size.

Proof. We observe that if $\mathbf{f} \in \mathbf{V}_h$, then $(z - A_h)^{-1}\mathbf{f} \in \mathbf{V}_h$. In fact, $\mathbf{g} := (z - A_h)^{-1}\mathbf{f}$ $\Rightarrow (z - A_h)\mathbf{g} = \mathbf{f} \Rightarrow z\mathbf{g} = \mathbf{f} + A_h\mathbf{g} \in \mathbf{V}_h$. Theorem 4.7 then says that, for all $z \in F$ and h sufficiently small, the continuous operator $(z - A_h) : \mathbf{V}_h \to \mathbf{V}_h$ is invertible with continuous inverse and continuity constant independent of the mesh size. The statement readily follows. $\square$

Theorem 4.5 is a direct consequence of Corollary 4.8.

Remark 4.9. For fixed $z \in \rho(A)$ and $\mathbf{f} \in \mathbf{V}(h)$, we can write

$$\|(z - A)\mathbf{f}\|_{\mathbf{V}(h)} \leq |z|\|\mathbf{f}\|_{\mathbf{V}(h)} + \|A\mathbf{f}\|_{\mathbf{V}} \leq |z|\|\mathbf{f}\|_{\mathbf{V}(h)} + C\|\mathbf{f}\|_{0,\Omega} \leq C(|z|)\|\mathbf{f}\|_{\mathbf{V}(h)},$$

owing to the stability estimate of the continuous problem and the definition of the $\mathbf{V}(h)$-norm. This, together with the result of Lemma 4.6, implies that, for all fixed $0 \neq z \in \rho(A)$, $(z - A) : \mathbf{V}(h) \to \mathbf{V}(h)$ is a continuous invertible operator with continuous inverse. An immediate consequence of this fact is the analogue of Corollary 4.8: Let $F \subset \rho(A)$ be closed. Then, there exists a positive constant C independent of the mesh size such that, for all $z \in F$,

$$\|R_z(A)\|_{\mathcal{L}(\mathbf{V}(h),\mathbf{V}(h))} \leq C.$$

4.4. Nonpollution and completeness of the eigenspaces, and completeness of the spectrum. Let λ be an eigenvalue of A with algebraic multiplicity m, and let Γ be a circle in the complex plane centered at λ which lies in $\rho(A)$ and does not enclose any other point of $\sigma(A)$. According to [36, p. 178], we define the *spectral projections* E and, for h small enough, E_h from $\mathbf{V}_h$ into $\mathbf{V}(h)$ by

$$E = E_\lambda = \frac{1}{2\pi i}\int_\Gamma R_z(A)\,dz, \qquad E_h = E_{h,\lambda} = \frac{1}{2\pi i}\int_\Gamma R_z(A_h)\,dz, \tag{4.4}$$

respectively. Theorem 4.5 guarantees that, for h small enough, E_h is well defined.

We have the following uniform convergence result, analogous to [25, Lemma 2].

THEOREM 4.10. *We have*

$$\lim_{h\to 0} \|E - E_h\|_{\mathcal{L}(\mathbf{V}_h,\mathbf{V}(h))} = 0.$$

Proof. Since $(z - A)^{-1} - (z - A_h)^{-1} = (z - A)^{-1}(A - A_h)(z - A_h)^{-1}$, we have

$$R_z(A) - R_z(A_h) = R_z(A)(A - A_h)R_z(A_h).$$

Therefore, for $\mathbf{f} \in \mathbf{V}_h$,

$$\begin{aligned}&\|R_z(A)(A - A_h)R_z(A_h)\mathbf{f}\|_{\mathbf{V}(h)}\\ &\quad\leq \|R_z(A)\|_{\mathcal{L}(\mathbf{V}(h),\mathbf{V}(h))}\|A - A_h\|_{\mathcal{L}(\mathbf{V}_h,\mathbf{V}(h))}\|R_z(A_h)\|_{\mathcal{L}(\mathbf{V}_h,\mathbf{V}_h)}\|\mathbf{f}\|_{\mathbf{V}(h)}.\end{aligned}$$

Owing to Remark 4.9, Property 2, and Corollary 4.8, we get the result. ∎

If Y and Z are closed subspaces of $\mathbf{V}(h)$, we define

$$\delta_h(\mathbf{x}, Y) := \inf_{\mathbf{y}\in Y} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{V}(h)}, \qquad \delta_h(Y, Z) := \sup_{\substack{\mathbf{y}\in Y\\ \|\mathbf{y}\|_{\mathbf{V}(h)}=1}} \delta_h(\mathbf{y}, Z),$$

$$\widehat{\delta}_h(Y, Z) := \max\{\delta_h(Y, Z), \delta_h(Z, Y)\}.$$

The following result holds true (compare with [25, Theorem 2]).

THEOREM 4.11 (nonpollution of the eigenspaces). *We have*

$$\lim_{h\to 0} \delta_h(E_h(\mathbf{V}_h), E(\mathbf{V})) = 0.$$

Proof. We start by observing that $E(\mathbf{V}) = E(L^2(\Omega)^d)$. Indeed, $E(\mathbf{V})$ is the projection onto the eigenspace associated with the eigenvalue λ of the operator $A : \mathbf{V} \to \mathbf{V}$, and $E(L^2(\Omega)^d)$ is the projection onto the eigenspace associated with the eigenvalue λ of the operator $A : L^2(\Omega)^d \to L^2(\Omega)^d$ (see, e.g., [27, Theorem 5, p. 579]).

Since all eigenfunctions of $A : L^2(\Omega)^d \to L^2(\Omega)^d$ are in $\mathbf{V}$, the two eigenspaces coincide, i.e., $E(\mathbf{V}) = E(L^2(\Omega)^d)$. Therefore,

$$\begin{aligned}\sup_{\substack{\mathbf{y}_h \in E_h(\mathbf{V}_h)\\ \|\mathbf{y}_h\|_{\mathbf{V}(h)}=1}} \inf_{\mathbf{x}\in E(\mathbf{V})} \|\mathbf{y}_h - \mathbf{x}\|_{\mathbf{V}(h)} &= \sup_{\substack{\mathbf{y}_h \in E_h(\mathbf{V}_h)\\ \|\mathbf{y}_h\|_{\mathbf{V}(h)}=1}} \inf_{\mathbf{x}\in E(L^2(\Omega)^d)} \|\mathbf{y}_h - \mathbf{x}\|_{\mathbf{V}(h)}\\ &= \sup_{\substack{\mathbf{y}_h \in E_h(\mathbf{V}_h)\\ \|y_h\|_{\mathbf{V}(h)}=1}} \inf_{\mathbf{x}\in L^2(\Omega)^d} \|E_h\mathbf{y}_h - E\mathbf{x}\|_{\mathbf{V}(h)},\end{aligned}$$

where in the last step we have used that $E_h\mathbf{y}_h = \mathbf{y}_h$ for all $\mathbf{y}_h \in E_h(\mathbf{V}_h)$. Taking $\mathbf{x} = \mathbf{y}_h$, Theorem 4.10 allows us to conclude. ∎

For eigenspaces associated with eigenvalues $\lambda \neq 1$, we have the following result (compare with [25, Theorem 3]).

THEOREM 4.12 (completeness of the eigenspaces). *If $E = E_\lambda$ is associated with an eigenvalue $\lambda \neq 1$, then*

$$\lim_{h\to 0} \delta_h(E(\mathbf{V}), E_h(\mathbf{V}_h)) = 0.$$

Proof. Since $EE\mathbf{y} = E\mathbf{y}$ for all $\mathbf{y} \in \mathbf{V}$, we can write

$$\delta_h(E(\mathbf{V}), E_h(\mathbf{V}_h)) = \sup_{\substack{\mathbf{x}\in E(\mathbf{V})\\ \|\mathbf{x}\|_{\mathbf{V}(h)}=1}} \inf_{\mathbf{x}_h\in\mathbf{V}_h} \|E\mathbf{x} - E_h\mathbf{x}_h\|_{\mathbf{V}(h)}.$$

Fix $\mathbf{x} \in E(\mathbf{V})$. Then, $E\mathbf{x} = \mathbf{x}$ and $\mathbf{x} \in \mathbf{W}$. By Assumption 2, there exists $\widetilde{\mathbf{x}}_h \in \mathbf{V}_h$ such that

$$\lim_{h\to 0} \|\mathbf{x} - \widetilde{\mathbf{x}}_h\|_{\mathbf{V}(h)} = 0. \tag{4.5}$$

Therefore,

$$\begin{aligned}\inf_{\mathbf{x}_h\in\mathbf{V}_h} \|E\mathbf{x} - E_h\mathbf{x}_h\|_{\mathbf{V}(h)} &\le \|E\mathbf{x} - E_h\widetilde{\mathbf{x}}_h\|_{\mathbf{V}(h)}\\ &\le \|E(\mathbf{x} - \widetilde{\mathbf{x}}_h)\|_{\mathbf{V}(h)} + \|(E - E_h)\widetilde{\mathbf{x}}_h\|_{\mathbf{V}(h)}\\ &\le \|E\|_{\mathcal{L}(\mathbf{V}(h),\mathbf{V}(h))}\|\mathbf{x} - \widetilde{\mathbf{x}}_h\|_{\mathbf{V}(h)} + \|E - E_h\|_{\mathcal{L}(\mathbf{V}_h,\mathbf{V}(h))}\|\widetilde{\mathbf{x}}_h\|_{\mathbf{V}(h)}.\end{aligned}$$

The first term on the right-hand side tends to zero, as $h \to 0$, due to (4.5), whereas the second term tends to zero, as $h \to 0$, owing to Theorem 4.10. Since $E(\mathbf{V})$ is the eigenspace associated with $\lambda \neq 1$, it is finite dimensional; therefore, pointwise convergence implies uniform convergence in $E(\mathbf{V})$, and the result readily follows. ∎

Finally, we have the following result.

THEOREM 4.13 (completeness of the spectrum). *For all $\lambda \in \sigma(A)$,*

$$\lim_{h\to 0} \delta_h(\lambda, \sigma(A_h)) = 0.$$

Proof. For $\lambda = 1$, since $\lambda_h = 1 \in \sigma(A_h)$, the result is obvious. For $\lambda \neq 1$, Theorems 4.11 and 4.12 imply that, for $E = E_\lambda$,

$$\lim_{h\to 0} \widehat{\delta}_h(E(\mathbf{V}), E_h(\mathbf{V}_h)) = 0. \tag{4.6}$$

Now, let m and m_h be the (finite) dimensions of $E(\mathbf{V})$ and $E_h(\mathbf{V}_h)$, respectively. Then, (4.6) implies that, for h small enough, $m_h = m$ (see [36, p. 200]). In particular, denoting by D_Γ the domain of $\mathbb{C}$ bounded by Γ, if $D_\Gamma \cap (\sigma(A) \setminus \{1\}) \neq \emptyset$, then, for h small enough, $D_\Gamma \cap (\sigma(A_h) \setminus \{1\}) \neq \emptyset$. The fact that all the eigenvalues are isolated allows us to conclude. ∎

4.5. Approximation of eigenvalues and eigenfunctions. In this section we report the consequences of the results obtained in the previous section on the approximation of the eigenvalues and the eigenfunctions. The results in this section are stated without proof, since their proofs are standard and the paper [26] can be used as a reference; the proof of the eigenvalue estimates is also reported in [3].

Let $\lambda \neq 1$ be an eigenvalue of A, and let m be its (finite) multiplicity. We denote by E and E_h the associated continuous and discrete spectral projections, respectively. At the end of the previous section, we have proved that there exist exactly m eigenvalues $\{\lambda_{1,h}, \dots, \lambda_{m,h}\}$ of A_h (repeated with their multiplicities) which converge to λ, i.e.,

$$\lim_{h\to 0} \sup_{1\le i\le m} |\lambda - \lambda_{i,h}| = 0.$$

In the following theorem, we analyze the convergence rate of this limit (convergence of eigenvalues) and the one of the limits in Theorem 4.12 (convergence of eigenfunctions).

THEOREM 4.14. *Let $\lambda \neq 1$ be an eigenvalue of A, and let E and E_h be the associated continuous and discrete spectral projections, respectively. Then, for h small enough, it holds that*

$$\delta_h(E(\mathbf{V}), E_h(\mathbf{V}_h)) \le Ch^t,$$
$$\sup_{1\le i\le m} |\lambda - \lambda_{i,h}| \le Ch^t,$$

where t is the maximal exponent which can be used in the bound (3.5) *of Assumption* 4 *for all $\mathbf{f} \in E(\mathbf{V})$, and the constant C depends only on λ (and deteriorates for small values of λ). Moreover, for hermitian DG methods, we have*

$$\sup_{1\le i\le m} |\lambda - \lambda_{i,h}| \le Ch^{2t}.$$

5. Remarks on Assumptions 5 and 6. In this section we make some remarks on our key assumptions, Assumptions 5 and 6. More precisely, in sections 5.1 and 5.2, respectively, we show that

(i) Assumptions 5 and 6 are not only sufficient but also necessary for a spurious-free DG approximation of Problem 1; therefore, provided that Assumptions 1–4 are satisfied, Assumptions 5 and 6 are *necessary and sufficient* for a DG method to provide a spurious-free approximation of Problem 1;

(ii) Assumption 6 implies that $K_h^\perp$ and K_h are approximating in $\mathbf{W}$ and in $\mathbf{V}^0$, respectively (see (2.3)), provided that $\mathbf{V}_h$ is approximating in $\mathbf{V}$.

5.1. Necessity of Assumptions 5 and 6. For simplicity, we restrict ourselves to hermitian formulations and prove the necessity of Assumptions 5 and 6 for a spurious-free DG approximation of Problem 1.

PROPOSITION 5.1. *Any spurious-free hermitian DG method satisfies Assumption* 5.

Proof. We proceed as in the proof of [18, Lemma 6.5]. Let $\mathbf{v}$ be in $K_h^\perp$, and consider its spectral decomposition

$$\mathbf{v} = \sum_{1\neq\lambda_h\in\sigma(A_h)} \mathbf{v}_{\lambda_h},$$

with $\mathbf{v}_{\lambda_h}$ being an eigenfunction associated with $\lambda_h \in \sigma(A_h)$. Since $a_h(\cdot,\cdot)$ is hermitian, $a_h(\mathbf{v},\mathbf{v})$ is real; thus we can write

$$\begin{aligned} a_h(\mathbf{v},\mathbf{v}) &= \sum_{1\neq\lambda_h\in\sigma(A_h)} \sum_{1\neq\nu_h\in\sigma(A_h)} a_h(\mathbf{v}_{\lambda_h},\mathbf{v}_{\nu_h}) \\ &= \sum_{1\neq\lambda_h\in\sigma(A_h)} \sum_{1\neq\nu_h\in\sigma(A_h)} \left(\frac{1}{\lambda_h}-1\right)(\mathbf{v}_{\lambda_h},\mathbf{v}_{\nu_h})_\varepsilon \geq \frac{1-\beta}{\beta}\|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega}^2, \end{aligned}$$

due to Property 1; therefore Assumption 5 is satisfied with $C = \beta/(1-\beta)$. ∎

PROPOSITION 5.2. *Any spurious-free hermitian DG method satisfies Assumption* 6.

Proof. The proof is similar to the one of [18, Lemma 6.3]. Assumption 6 can be rewritten as follows: for all $\eta > 0$, there is $\overline{h} > 0$ such that, for all $h \in (0,\overline{h})$, for any $\mathbf{w}_h \in K_h^\perp$ with $\|\mathbf{w}_h\|_{\mathbf{V}(h)} = 1$, there exists $\mathbf{w} \in H(\mathrm{div}_\varepsilon^0;\Omega)$ such that

$$\|\varepsilon^{1/2}(\mathbf{w}-\mathbf{w}_h)\|_{0,\Omega} \leq \eta$$

(we have used the equivalence between the L^2-norm and the L^2_ε-norm).

Let $\{\lambda_j\}_{j=1}^\infty$ be the decreasing sequence of all the continuous eigenvalues $1 \neq \lambda_j \in \sigma(A)$, where each distinct λ_j appears only once in the sequence, independently of its multiplicity m_j. Denoting by n_h the dimension of $K_h^\perp$, let $\{\lambda_{i,h}\}_{i=1}^{n_h}$ be the nonincreasing sequence of all the discrete eigenvalues $1 \neq \lambda_{i,h} \in \sigma(A_h)$, repeated according to their multiplicity.

Fix $\eta > 0$. Since $\lambda_1 \leq \frac{1}{1+\alpha_c} < 1$, with α_c being the $\mathbf{V}$-ellipticity constant of $a(\cdot,\cdot)$ in $\mathbf{W}$, there exists $k > 0$ such that $\lambda_k < \frac{\eta^2}{8\gamma} - \frac{1}{4}\frac{\eta^2}{8\gamma}$, where γ is the continuity constant of the form $a_h(\cdot,\cdot)$ (see Assumption 3); moreover, we can choose mutually disjoint neighborhoods $N(\lambda_j)$ of λ_j, $1 \leq j \leq k$, such that $N(\lambda_j) \subset (\lambda_j - \frac{1}{4}\frac{\eta^2}{8\gamma}, \lambda_j + \frac{1}{4}\frac{\eta^2}{8\gamma})$. From Theorem 4.5 and Property 1, there is $h_1 > 0$ such that, for all $h < h_1$, $N(\lambda_j)$ contains exactly m_j discrete eigenvalues, $1 \leq j \leq k$; moreover, $N(1)$ can be chosen in such a way that $N(1) \cap \{\lambda_{i,h}\}_{i=1}^{n_h} = \emptyset$. Set $m = \sum_{j=1}^k m_j$; obviously, $m \leq n_h$.

Now, take $h < h_1$, fix $\mathbf{w}_h \in K_h^\perp$, with $\|\mathbf{w}_h\|_{\mathbf{V}(h)} = 1$, and consider its spectral decomposition

$$\mathbf{w}_h = \sum_{i=1}^{n_h} \mathbf{w}_{i,h} = \sum_{i=1}^{m} \mathbf{w}_{i,h} + \sum_{i=m+1}^{n_h} \mathbf{w}_{i,h} =: \mathbf{w}_h^1 + \mathbf{w}_h^2.$$

For the term $\mathbf{w}_h^2$, we use Proposition 3.3(ii). Denoting by $\lambda_{\ell,h}$ the eigenvalue corresponding to $\mathbf{w}_{\ell,h}$, we can write

$$\begin{aligned} \|\varepsilon^{1/2}\mathbf{w}_h^2\|_{0,\Omega}^2 &= \sum_{i=m+1}^{n_h} \|\varepsilon^{1/2}\mathbf{w}_{i,h}\|_{0,\Omega}^2 = \sum_{i=m+1}^{n_h} b_h(A_h\mathbf{w}_{i,h},\mathbf{w}_{i,h}) \\ &= \sum_{i=m+1}^{n_h} \lambda_{i,h} b_h(\mathbf{w}_{i,h},\mathbf{w}_{i,h}) \leq \lambda_{m,h} \sum_{i=m+1}^{n_h} b_h(\mathbf{w}_{i,h},\mathbf{w}_{i,h}) \\ &= \lambda_{m,h} b_h(\mathbf{w}_h^2,\mathbf{w}_h^2) \leq 2\gamma\lambda_{m,h}\|\mathbf{w}_h^2\|_{\mathbf{V}(h)}^2. \end{aligned}$$

Since $\|\mathbf{w}_h^2\|_{\mathbf{V}(h)} \leq \|\mathbf{w}_h\|_{\mathbf{V}(h)} = 1$ and $\lambda_{m,h} < \frac{\eta^2}{8\gamma}$, due to $\lambda_{m,h} \in N(\lambda_k)$, we obtain

$$\|\varepsilon^{1/2}\mathbf{w}_h^2\|_{0,\Omega} \leq \frac{\eta}{2}. \tag{5.1}$$

Let us turn now to the term $\mathbf{w}_h^1$, and consider its spectral decomposition

$$\mathbf{w}_h^1 = \sum_{i=1}^{m} \mathbf{w}_{i,h} = \sum_{j=1}^{k} \sum_{i=a_j}^{b_j} \mathbf{w}_{i,h} =: \sum_{j=1}^{k} \widetilde{\mathbf{w}}_{j,h},$$

where $a_j = \sum_{i=1}^{j-1} m_i$ and $b_j = \sum_{i=1}^{j} m_i$. Owing to Theorem 4.11, in correspondence with η, there is $h_2 > 0$ such that, for all $h < h_2$, for each $1 \le j \le k$, there exists a continuous eigenfunction $\mathbf{w}_j$ associated with λ_j such that $\|\mathbf{w}_j - \widetilde{\mathbf{w}}_{j,h}\|_{\mathbf{V}(h)} \le \frac{\eta}{2k}$. Set $\mathbf{w} = \sum_{j=1}^{k} \mathbf{w}_j$; clearly, $\mathbf{w} \in \mathbf{W}$. Then,

$$\|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h^1)\|_{0,\Omega} \le \sum_{j=1}^{k} \|\varepsilon^{1/2}(\mathbf{w}_j - \widetilde{\mathbf{w}}_{j,h})\|_{0,\Omega} \le \frac{\eta}{2}. \tag{5.2}$$

Therefore, for all $h < \overline{h} = \min\{h_1, h_2\}$, in correspondence to any $\mathbf{w}_h \in K_h^\perp$ with $\|\mathbf{w}_h\|_{\mathbf{V}(h)=1}$, we have found $\mathbf{w} \in \mathbf{W} \subset H(\mathrm{div}_\varepsilon^0; \Omega)$ such that

$$\|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega} \le \|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h^1)\|_{0,\Omega} + \|\varepsilon^{1/2}\mathbf{w}_h^2\|_{0,\Omega} \le \eta,$$

owing to (5.1) and (5.2), which concludes the proof. ∎

Remark 5.3. From the proof of Proposition 5.2 it is clear that a spurious-free hermitian DG method satisfies Assumption 6 with $\mathbf{w} \in \mathbf{W}$.

5.2. Gap properties. Let $P : L^2(\Omega)^d \to H(\mathrm{div}_\varepsilon^0; \Omega)$ and $Q = I - P$ be the projection operators associated with the first decomposition in (2.3). Notice that, for all $\mathbf{v} \in \mathbf{V}(h)$, $P\mathbf{v}$ and $Q\mathbf{v}$ belong to $\mathbf{V}(h)$, and $Q \in \mathcal{L}(L^2(\Omega)^d, \mathbf{V}(h))$. The restrictions of P and Q to $\mathbf{V}$ are onto $\mathbf{W}$ and $\mathbf{V}^0$, respectively, and coincide with the projection operators associated with the second decomposition in (2.3). We will make use of the following lemma.

LEMMA 5.4. *Assumption* 6 *implies that, for all* h *small enough,*

$$\|\mathbf{w}_h - P\mathbf{w}_h\|_{\mathbf{V}(h)} = \|Q\mathbf{w}_h\|_{\mathbf{V}(h)} \le \eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)} \qquad \forall \mathbf{w}_h \in K_h^\perp,$$

with $\eta_h \to 0$ *as* $h \to 0$.

Proof. Let us rewrite Assumption 6 as follows: for all h small enough, there exists an operator $\Pi_h : K_h^\perp \to H(\mathrm{div}_\varepsilon^0; \Omega)$ such that $\Pi_h \in \mathcal{L}(\mathbf{V}(h), L^2(\Omega)^d)$ and

$$\|\mathbf{w}_h - \Pi_h \mathbf{w}_h\|_{0,\Omega} \le \eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)}, \tag{5.3}$$

with $\eta_h \to 0$ as $h \to 0$.

Then, for all h small enough, due to $\Pi_h \mathbf{w}_h \in H(\mathrm{div}_\varepsilon^0; \Omega)$ and to (5.3), we have

$$\begin{aligned} \|\mathbf{w}_h - P\mathbf{w}_h\|_{\mathbf{V}(h)} &= \|Q\mathbf{w}_h\|_{\mathbf{V}(h)} = \|Q(\mathbf{w}_h - \Pi_h \mathbf{w}_h)\|_{\mathbf{V}(h)} \\ &\le \|Q\|_{\mathcal{L}(L^2(\Omega)^d, \mathbf{V}(h))} \|\mathbf{w}_h - \Pi_h \mathbf{w}_h\|_{0,\Omega} \le C\eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)}, \end{aligned}$$

with $\eta_h \to 0$, as $h \to 0$. ∎

We have the following result.

PROPOSITION 5.5. *Assumptions* 1, 2, 3, *and* 6 *imply that* $K_h^\perp$ *is approximating in* $\mathbf{W}$, *i.e.,*

$$\lim_{h \to 0} \inf_{\mathbf{w}_h \in K_h^\perp} \|\mathbf{w} - \mathbf{w}_h\|_{\mathbf{V}(h)} = 0 \qquad \forall \mathbf{w} \in \mathbf{W}.$$

Moreover, provided that, in addition, Assumption 2 *holds true for all* $\mathbf{v} \in \mathbf{V}$*, we also have that* $K_h \subset \mathbf{V}^0$ *is approximating in* $\mathbf{V}^0$*, i.e.,*

$$\lim_{h\to 0} \inf_{\mathbf{k}_h \in K_h} \|\mathbf{k} - \mathbf{k}_h\|_{0,\Omega} = 0 \qquad \forall \mathbf{k} \in \mathbf{V}^0.$$

Proof. We use similar arguments as in [16, Theorem 3.3]. Let $P_h : \mathbf{V}_h \to K_h^\perp$ and $Q_h = I - P_h$ be the projection operators associated with the $\mathbf{V}(h)$-orthogonal decomposition $\mathbf{V}_h = K_h \oplus K_h^\perp$, and let $I_h : \mathbf{V} \to \mathbf{V}_h$ be the $\mathbf{V}(h)$-orthogonal projection. We proceed in two steps.

(i) $K_h^\perp$ *is approximating in* $\mathbf{W}$. We start by observing that if $\mathbf{w} \in \mathbf{W}$, then $I_h\mathbf{w} \in K_h^\perp$; in fact, for all $\mathbf{k}_h \in K_h$, since $K_h \subset \mathbf{V}^0$, we have $(I_h\mathbf{w}, \mathbf{k}_h)_{\mathbf{V}(h)} = (\mathbf{w}, \mathbf{k}_h)_{\mathbf{V}(h)} = (\mathbf{w}, \mathbf{k}_h)_\varepsilon$, which is equal to zero, due to the L^2_ε-orthogonality between $\mathbf{V}^0$ and $\mathbf{W}$. Now, given $\mathbf{w} \in \mathbf{W}$, we let $\mathbf{w}_h \in K_h^\perp$ be defined by $\mathbf{w}_h = I_h\mathbf{w}$. Assumption 2 ensures that $\|\mathbf{w} - I_h\mathbf{w}\|_{\mathbf{V}(h)}$ converges to zero, as $h \to 0$, and the proof of (i) is complete.

(ii) K_h *is approximating in* $\mathbf{V}^0$. Given $\mathbf{k} \in \mathbf{V}^0$, we let $\mathbf{k}_h \in K_h$ be defined by $\mathbf{k}_h = Q_h I_h \mathbf{k}$. Since $\mathbf{k} - \mathbf{k}_h = Q(\mathbf{k} - I_h\mathbf{k}) + (Q - Q_h)I_h\mathbf{k}$, we have

$$\|\mathbf{k} - \mathbf{k}_h\|_{0,\Omega} \le \|Q(\mathbf{k} - I_h\mathbf{k})\|_{0,\Omega} + \|(Q - Q_h)I_h\mathbf{k}\|_{0,\Omega}.$$

For the first term, we have

$$\|Q(\mathbf{k} - I_h\mathbf{k})\|_{0,\Omega} \le \|Q\|_{\mathcal{L}(L^2(\Omega)^d, L^2(\Omega)^d)} \|\mathbf{k} - I_h\mathbf{k}\|_{0,\Omega},$$

which converges to zero, as $h \to 0$, since we have supposed Assumption 2 to be satisfied for all functions in $\mathbf{V}$. For the second term, we have

$$(Q - Q_h)I_h\mathbf{k} = (Q - Q_h)(P_h I_h\mathbf{k} + Q_h I_h\mathbf{k}) = QP_h I_h\mathbf{k},$$

since, due to $K_h \subset \mathbf{V}^0$, Q is identity on K_h. Therefore,

$$\|(Q - Q_h)I_h\mathbf{k}\|_{0,\Omega} = \|QP_h I_h\mathbf{k}\|_{0,\Omega} \le \eta_h \|P_h I_h\mathbf{k}\|_{0,\Omega} \le C\eta_h \|\mathbf{k}\|_{0,\Omega},$$

owing to Lemma 5.4 and the fact that P_h and I_h are $\mathbf{V}(h)$-orthogonal projections and $\mathbf{k} \in \mathbf{V}^0$. This completes the proof. ∎

6. The indefinite Maxwell source problem. Consider the following indefinite Maxwell source problem: given $\mathbf{f} \in L^2(\Omega)^d$ and $\omega \in \mathbb{R}$ such that ω^2 is not an eigenvalue of Problem 1, find $\mathbf{u} \in \mathbf{V}$ such that

$$\nabla \times (\mu^{-1}\nabla \times \mathbf{u}) - \omega^2 \varepsilon \mathbf{u} = \mathbf{f}. \tag{6.1}$$

The theory developed so far guarantees that, for the DG method for (6.1), i.e., find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$a_h(\mathbf{u}_h, \mathbf{v}) - \omega^2 (\mathbf{u}_h, \mathbf{v})_\varepsilon = (\mathbf{f}, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{V}_h, \tag{6.2}$$

the following result holds true.

THEOREM 6.1. *Provided that Assumptions* 1–6 *are satisfied, for* h *small enough, the DG method* (6.2) *is well-posed.*

Proof. Let $\mathbf{g}_h$ be the (unique) element of $\mathbf{V}_h$ such that

$$(\mathbf{g}_h, \mathbf{v})_\varepsilon = (\mathbf{f}, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{V}_h. \tag{6.3}$$

Then, since $a_h(\mathbf{u}_h,\mathbf{v})-\omega^2(\mathbf{u}_h,\mathbf{v})_\varepsilon = b_h(\mathbf{u}_h,\mathbf{v})-(1+\omega^2)(\mathbf{u}_h,\mathbf{v})_\varepsilon$, setting $z=1/(1+\omega^2)$, we can write (6.2) as

$$b_h(z\mathbf{u}_h,\mathbf{v}) = z(\mathbf{g}_h,\mathbf{v})_\varepsilon + (\mathbf{u}_h,\mathbf{v})_\varepsilon \qquad \forall \mathbf{v}\in\mathbf{V}_h$$

or, equivalently, using the definition of the solution operator A_h,

$$b_h(A_h\mathbf{u}_h + zA_h\mathbf{g}_h - z\mathbf{u}_h,\mathbf{v}) = 0 \qquad \forall \mathbf{v}\in\mathbf{V}_h.$$

From this, due to the coercivity of $b_h(\cdot,\cdot)$ (see Assumption 3 and Remark 3.1), it holds that

$$(z-A_h)\mathbf{u}_h = zA_h\mathbf{g}_h. \tag{6.4}$$

Since ω^2 is not an eigenvalue of Problem 1, then $0\neq z\in\rho(A)$; thus Theorem 4.7 applies, and we have that (6.4) admits the unique solution $\mathbf{u}_h = z(z-A_h)^{-1}A_h\mathbf{g}_h$ for h small enough. Moreover, due to Corollary 4.8 and $A_h\in\mathcal{L}(L^2(\Omega)^d,\mathbf{V}_h)$, there exists $C>0$ independent of the mesh size such that

$$\|\mathbf{u}_h\|_{\mathbf{V}(h)} \le C\|\mathbf{g}_h\|_{0,\Omega} \le C\|\mathbf{f}\|_{0,\Omega}, \tag{6.5}$$

where the second inequality follows from (6.3) and the equivalence between the L^2-norm and the L^2_ε-norm. ∎

We end this section by proving the following inf-sup condition.

PROPOSITION 6.2. *With the assumptions of Theorem* 6.1*, for h small enough, there exists a constant $\kappa>0$ independent of h such that*

$$\inf_{\mathbf{0}\neq\mathbf{u}_h\in\mathbf{V}_h}\sup_{\mathbf{0}\neq\mathbf{v}_h\in\mathbf{V}_h}\frac{\mathrm{Re}\,[a_h(\mathbf{u}_h,\mathbf{v}_h)-\omega^2(\mathbf{u}_h,\mathbf{v}_h)_\varepsilon]}{\|\mathbf{u}_h\|_{\mathbf{V}(h)}\|\mathbf{v}_h\|_{\mathbf{V}(h)}} \ge \kappa. \tag{6.6}$$

Proof. Theorem 6.1 implies that, for h small enough, there exists a constant $\kappa'>0$ independent of h such that

$$\inf_{\mathbf{0}\neq\mathbf{v}_h\in\mathbf{V}_h}\sup_{\mathbf{0}\neq\mathbf{u}_h\in\mathbf{V}_h}\frac{\mathrm{Re}\,[a_h(\mathbf{u}_h,\mathbf{v}_h)-\omega^2(\mathbf{u}_h,\mathbf{v}_h)_\varepsilon]}{\|\mathbf{u}_h\|_{\mathbf{V}(h)}\|\mathbf{v}_h\|_{\mathbf{V}(h)}} \ge \kappa'. \tag{6.7}$$

In fact, fix $\mathbf{v}_h\in\mathbf{V}_h$ and set $\mathbf{u}_h=\mathbf{u}_h^1+(1+\omega^2)\mathbf{u}_h^2$, with $\mathbf{u}_h^1=\mathbf{v}_h$ and $\mathbf{u}_h^2$ solution to (6.2) with $\mathbf{f}=\varepsilon\mathbf{v}_h$; the stability estimate (6.5) and the coercivity in Assumption 3 lead to (6.7).

If $a_h(\cdot,\cdot)$ is hermitian, (6.6) coincides with (6.7), and the proof is complete. Otherwise, we prove the well-posedness of the adjoint problem as follows: given $\mathbf{f}\in L^2(\Omega)^d$ and $\omega\in\mathbb{R}$ such that ω^2 is not an eigenvalue of Problem 1, find $\mathbf{v}_h\in\mathbf{V}_h$ such that

$$a_h(\mathbf{w}_h,\mathbf{v}_h)-\omega^2(\mathbf{w}_h,\mathbf{v}_h)_\varepsilon = (\mathbf{f},\mathbf{w}_h) \qquad \forall \mathbf{w}_h\in\mathbf{V}_h. \tag{6.8}$$

Existence and uniqueness of the solution of (6.8), for h small enough, immediately follow from Theorem 6.1, due to finite dimensionality. For the stability, due to (6.7), in correspondence to $\mathbf{v}_h$, we can find $\mathbf{0}\neq\mathbf{w}_h\in\mathbf{V}_h$ such that

$$\begin{aligned}\kappa'\|\mathbf{w}_h\|_{\mathbf{V}(h)}\|\mathbf{v}_h\|_{\mathbf{V}(h)} &\le \mathrm{Re}\,[a_h(\mathbf{w}_h,\mathbf{v}_h)-\omega^2(\mathbf{w}_h,\mathbf{v}_h)_\varepsilon] = \mathrm{Re}\,[(\mathbf{f},\mathbf{w}_h)]\\ &\le \|\mathbf{f}\|_{L^2(\Omega)^d}\|\mathbf{w}_h\|_{\mathbf{V}(h)},\end{aligned}$$

which immediately gives $\|\mathbf{v}_h\|_{\mathbf{V}(h)}\le C\|\mathbf{f}\|_{L^2(\Omega)^d}$, with $C>0$ independent of h.

Therefore, the inf-sup condition (6.6) follows from the well-posedness of the adjoint problem (6.8), the same way as the inf-sup condition (6.7) follows from the well-posedness of problem (6.2), and the proof is complete. ∎

Remark 6.3. It is well known that the inf-sup condition (6.6) is a key ingredient in the proof of error estimates; see Remark 7.11.

7. Application to some discontinuous Galerkin methods. In this section we apply the theory developed in the previous sections to some of the DG methods present in the literature, more precisely, to the methods of the interior penalty family (interior penalty (IP), nonsymmetric interior penalty (NIP), and incomplete interior penalty (IIP); see [4], [44], and [23], respectively) and to the local discontinuous Galerkin method (LDG; see [21]). We point out that everything stated below holds true also for the variants of the IP and LDG methods introduced in [7] and [14], respectively. We restrict ourselves to the case of *conformal meshes*, i.e., with no *hanging nodes*.

This section is organized as follows: in sections 7.1 and 7.2 the DG spaces and bilinear forms are defined and proved to fulfill the assumptions in section 3; in section 7.3 we prove that Assumption 5 is satisfied. In section 7.4, Assumption 6 is proved, and a few remarks aiming at specializing the results of our theory to the examples presented here are provided. Finally, in section 7.5 we investigate the relation of Assumption 6 with the discrete compactness property (see Property 3).

7.1. Meshes, trace operators, finite element spaces, and norms. Consider conformal, shape-regular partitions $\mathcal{T}_h$ of Ω into simplices $\{K\}$, where $h = \max_{K\in\mathcal{T}_h} h_K$, with $h_K = \operatorname{diam}(K)$ for all $K \in \mathcal{T}_h$. We denote by $\mathcal{F}_h^{\mathcal{I}}$ the set of all interior faces (edges if $d = 2$) of $\mathcal{T}_h$ and by $\mathcal{F}_h^{\mathcal{B}}$ the set of all boundary faces of $\mathcal{T}_h$, and we set $\mathcal{F}_h = \mathcal{F}_h^{\mathcal{I}} \cup \mathcal{F}_h^{\mathcal{B}}$. For a piecewise smooth vector-valued function $\mathbf{v}$, we introduce the following trace operators. Let $f \in \mathcal{F}_h^{\mathcal{I}}$ be an interior face shared by two neighboring elements K^+ and K^-; we write $\mathbf{n}^\pm$ to denote the outward normal unit vectors to the boundaries $\partial K^\pm$, respectively. Denoting by $\mathbf{v}^\pm$ the traces of $\mathbf{v}$ taken from within $K^\pm$, respectively, we define the tangential jumps and averages across f by

$$[\![\mathbf{v}]\!]_T := \mathbf{n}^+ \times \mathbf{v}^+ + \mathbf{n}^- \times \mathbf{v}^-, \qquad \{\!\{\mathbf{v}\}\!\} := (\mathbf{v}^+ + \mathbf{v}^-)/2,$$

respectively; if $d = 2$, defining the tangential vectors $\mathbf{t}^\pm = (-n_2^\pm, n_1^\pm)$, we understand $[\![\mathbf{v}]\!]_T$ as $\mathbf{v}^+ \cdot \mathbf{t}^+ + \mathbf{v}^- \cdot \mathbf{t}^-$.

On a boundary face $f \in \mathcal{F}_h^{\mathcal{B}}$, we set $[\![\mathbf{v}]\!]_T := \mathbf{n} \times \mathbf{v}$ and $\{\!\{\mathbf{v}\}\!\} := \mathbf{v}$.

For a given partition $\mathcal{T}_h$ of Ω and an approximation order $\ell \geq 1$, we define the complex vector-valued discontinuous finite element space

$$\mathbf{V}_h := \{\mathbf{v} \in L^2(\Omega)^d : \mathbf{v}|_K \in \mathcal{P}^\ell(K)^d \quad \forall K \in \mathcal{T}_h\}, \tag{7.1}$$

where $\mathcal{P}^\ell(K)$ is the space of complex polynomials of total degree at most ℓ on K. We also need to define the complex scalar-valued discontinuous finite element space

$$Q_h = \{q \in L^2(\Omega) : \mathbf{v}|_K \in \mathcal{P}^{\ell+1}(K) \quad \forall K \in \mathcal{T}_h\}.$$

We point out that all the results of this section hold true also with the choice of the local Nédélec elements of the first type [40], instead of the full polynomials of degree ℓ, in (7.1). For the case of parallelograms or parallelepipeds, see Remark 7.14.

We endow both $\mathbf{V}_h$ and $\mathbf{V}(h) = \mathbf{V} + \mathbf{V}_h$ with the seminorm and norm

$$|\mathbf{v}|^2_{\mathbf{V}(h)} = \|\mu^{-1/2}\nabla_h \times \mathbf{v}\|^2_{0,\Omega} + \|\mathtt{h}^{-1/2}[\![\mathbf{v}]\!]_T\|^2_{0,\mathcal{F}_h},$$
$$\|\mathbf{v}\|^2_{\mathbf{V}(h)} = |\mathbf{v}|^2_{\mathbf{V}(h)} + \|\varepsilon^{1/2}\mathbf{v}\|^2_{0,\Omega},$$

where we have denoted by ∇_h the elementwise application of the ∇ operator and used the notation $\|\varphi\|^2_{0,\mathcal{F}_h} := \sum_{f\in\mathcal{F}_h} \|\varphi\|^2_{0,f}$. In the following, we will also use the notation $\int_{\mathcal{F}_h} \varphi\, ds := \sum_{f\in\mathcal{F}_h} \int_f \varphi\, ds$.

The mesh function $\mathtt{h} \in L^\infty(\mathcal{F}_h)$ is defined by

$$\mathtt{h}(\mathbf{x}) := h_f\, \mathtt{m}(\mathbf{x}), \qquad \mathbf{x} \in f,\ f \in \mathcal{F}_h,$$

with h_f denoting the diameter of the face f and the function $\mathtt{m} \in L^\infty(\mathcal{F}_h)$ being defined as follows: if μ_K denotes the extension of $\mu_{|_K}$ up to ∂K, and $|\mu_K(\mathbf{x})|$ denotes the spectral norm of the tensor $\mu_K(\mathbf{x})$, then $\mathtt{m}(\mathbf{x}) = \min\{|\mu_{K^+}(\mathbf{x})|, |\mu_{K^-}(\mathbf{x})|\}$ if $\mathbf{x}$ is in the interior of $\partial K^+ \cap \partial K^-$ and $\mathtt{m}(\mathbf{x}) = |\mu_K(\mathbf{x})|$ if $\mathbf{x}$ is in the interior of $\partial K \cap \partial\Omega$.

The following result is then evident.

PROPOSITION 7.1. *Assumptions* 1 *and* 2 *are satisfied.*

7.2. Discontinuous Galerkin bilinear forms. We recall the expressions of the DG bilinear forms associated with the IP methods and with the LDG method applied to the Maxwell equations, pointing for further details to [43], [35] for the IP method and to [42], [31], [32] for the LDG method.

Define the IP, NIP, and IIP forms $a_h^{IP(k)} : \mathbf{V}_h \times \mathbf{V}_h \to \mathbb{C}$

$$\begin{aligned} a_h^{IP(k)}(\mathbf{u},\mathbf{v}) :=& (\mu^{-1}\nabla_h \times \mathbf{u}, \nabla_h \times \mathbf{v}) - \int_{\mathcal{F}_h} [\![\overline{\mathbf{v}}]\!]_T \cdot \{\!\{\mu^{-1}\nabla_h \times \mathbf{u}\}\!\}\, ds \\ & - k \int_{\mathcal{F}_h} [\![\mathbf{u}]\!]_T \cdot \{\!\{\mu^{-1}\nabla_h \times \overline{\mathbf{v}}\}\!\}\, ds + \int_{\mathcal{F}_h} \mathtt{a}\, [\![\mathbf{u}]\!]_T \cdot [\![\overline{\mathbf{v}}]\!]_T\, ds, \end{aligned}$$

where $k = 1$ for the IP method, $k = -1$ for the NIP method, and $k = 0$ for the IIP method, and the stabilization function $\mathtt{a} \in L^\infty(\mathcal{F}_h)$ is defined by

$$\mathtt{a} := a_{\mathrm{stab}}\, \mathtt{h}^{-1}, \tag{7.2}$$

with $a_{\mathrm{stab}} > 0$ independent of the mesh size and the material coefficients.

The LDG form is defined as follows:

$$a_h^{LDG}(\mathbf{u},\mathbf{v}) := (\mu^{-1}(\nabla_h \times \mathbf{u} - \mathcal{L}(\mathbf{u})), \nabla_h \times \mathbf{v} - \mathcal{L}(\mathbf{v})) + \int_{\mathcal{F}_h} \mathtt{a}\, [\![\mathbf{u}]\!]_T \cdot [\![\overline{\mathbf{v}}]\!]_T\, ds, \tag{7.3}$$

with $\mathtt{a}$ again as in (7.2), and $\mathcal{L}$ is the *lifting operator* from $\mathbf{V}(h)$ into $\mathbf{V}_h$ defined by

$$(\mathcal{L}(\mathbf{v}), \mathbf{w}) = \int_{\mathcal{F}_h^{\mathcal{I}}} \mathtt{b} [\![\mathbf{v}]\!]_T \cdot [\![\overline{\mathbf{w}}]\!]_T\, ds + \int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\overline{\mathbf{w}}\}\!\}\, ds \qquad \forall \mathbf{w} \in \mathbf{V}_h;$$

here, $\mathtt{b} \in L^\infty(\mathcal{F}_h)$ is a bounded function independent of the mesh size.

Remark 7.2. The LDG method is usually defined by introducing the auxiliary variable $\mathbf{s} := \mu^{-1}\nabla \times \mathbf{u}$ and rewriting the second order problem in mixed form as a first order system; then an element-by-element integration by parts is performed, and the traces along the elemental boundaries are replaced by the so-called *numerical fluxes*, obtaining an $(\mathbf{s},\mathbf{u})$-formulation of the method, which is equivalent to the $\mathbf{u}$-formulation $a_h^{LDG}(\mathbf{u},\mathbf{v}) = \omega(\mathbf{u},\mathbf{v})_\varepsilon$, with $a_h^{LDG}(\mathbf{u},\mathbf{v})$ as in (7.3), after elimination of the auxiliary variable $\mathbf{s}$ in terms of $\mathbf{u}$ (see [42] for details). Here, we concentrate on the $\mathbf{u}$-formulation because we are concerned only with the analysis of the method in the framework presented in this paper.

We prove that the DG bilinear forms in this section fulfill Assumptions 3 and 4.

PROPOSITION 7.3. *Provided that* a_{stab} *in* (7.2) *is large enough, in the case of the IP and IIP methods, for all the considered DG bilinear forms Assumptions* 3 *and* 4 *are*

satisfied. Moreover, the exponent t *in* (3.5) *can be chosen as* $t = \min\{\ell, r\}$. *Finally, the condition* (3.2) *is satisfied.*

Proof. The validity of Assumption 3 is standard and the one of (3.2) is straightforward. The proof of Assumption 4 is technical, and we postpone it to the appendix. Note that existent results (see [42], [35], or [34]) apply only when $r > 1/2$. ∎

7.3. Discrete Friedrichs inequality (Assumption 5). Denote by $H^1_\Gamma(\Omega)$ the subspace of $H^1(\Omega)$ whose functions have zero trace on Γ_1, the outer boundary of Ω, and constant traces on the other connected components Γ_i of $\partial\Omega$, $i = 2, \dots, n_\Gamma$; notice that if $\partial\Omega$ is connected, then $H^1_\Gamma(\Omega) = H^1_0(\Omega)$. We set $\mathbf{V}_h^c = \mathbf{V}_h \cap \mathbf{V}$ and $Q_h^c = Q_h \cap H^1_\Gamma(\Omega)$; notice that $\mathbf{V}_h^c$ coincides with the $H_0(\mathrm{curl};\Omega)$-conforming Nédélec elements of the second family of degree ℓ (see [41]), and Q_h^c coincides with the space of continuous nodal elements of degree $\ell+1$ with zero trace on Γ_1 and constant traces on Γ_i, $i = 2, \dots, n_\Gamma$. Notice that, due to Assumption 1, we have

$$K_h = \mathbf{V}_h^c \cap \mathbf{V}^0 = \nabla Q_h^c,$$

and we denote by $\mathbf{W}_h^c$ its L^2_ε-orthogonal complement in $\mathbf{V}_h^c$, i.e.,

$$\mathbf{W}_h^c = \{\mathbf{v} \in \mathbf{V}_h^c : (\mathbf{v}, \nabla q)_\varepsilon = 0 \quad \forall q \in Q_h^c\}.$$

By definition, the splitting

$$\mathbf{V}_h^c = \mathbf{W}_h^c \oplus \nabla Q_h^c \tag{7.4}$$

is orthogonal in both the L^2_ε-norm and the $\mathbf{V}$-norm. Moreover, the discrete Friedrichs inequality holds in $\mathbf{W}_h^c$ (see [38, Corollary 7.22]):

$$\|\varepsilon^{1/2}\mathbf{w}\|_{0,\Omega} \le C|\mathbf{w}|_{\mathbf{V}} \qquad \forall\, \mathbf{w} \in \mathbf{W}_h^c. \tag{7.5}$$

We first establish a decomposition of $\mathbf{V}_h$ which will be used in order to prove both Assumptions 5 and 6; the proof is based on the following result (see [34, Proposition 4.5 and the appendix]).

THEOREM 7.4. *There exists an operator* $\Pi_h^c : \mathbf{V}_h \to \mathbf{V}_h^c$ *such that*

$$\|\mathbf{v} - \Pi_h^c \mathbf{v}\|^2_{0,\Omega} \le C \int_{\mathcal{F}_h} \mathtt{h} |[\![\mathbf{v}]\!]_T|^2 \, ds, \tag{7.6}$$

$$\|\mathbf{v} - \Pi_h^c \mathbf{v}\|^2_{\mathbf{V}(h)} \le C \int_{\mathcal{F}_h} \mathtt{h}^{-1} |[\![\mathbf{v}]\!]_T|^2 \, ds \tag{7.7}$$

for all $\mathbf{v} \in \mathbf{V}_h$, *with a constant* $C > 0$ *independent of the mesh size.*

The following proposition is a consequence of Theorem 7.4.

PROPOSITION 7.5. *There exists a complement* $\mathbf{V}_h^\perp$ *of* $\mathbf{V}_h^c := \mathbf{V}_h \cap \mathbf{V}$ *in* $\mathbf{V}_h$ *such that the decomposition* $\mathbf{V}_h = \mathbf{V}_h^c \oplus \mathbf{V}_h^\perp$ *is stable in* $\mathbf{V}_h$, *i.e.,*

$$\mathbf{v}_h = \mathbf{v}_h^c + \mathbf{v}_h^\perp, \qquad \|\mathbf{v}_h^c\|_{\mathbf{V}(h)} + \|\mathbf{v}_h^\perp\|_{\mathbf{V}(h)} \le C\|\mathbf{v}_h\|_{\mathbf{V}(h)}. \tag{7.8}$$

Moreover, it holds that

$$\|\mathbf{v}_h^\perp\|_{0,\Omega} \le Ch|\mathbf{v}_h^\perp|_{\mathbf{V}(h)} \qquad \forall \mathbf{v}_h^\perp \in \mathbf{V}_h^\perp. \tag{7.9}$$

The constant $C > 0$ *is independent of the mesh size.*

Proof. The operator Π_h^c defined in Theorem 7.4 is continuous thanks to (7.7). Moreover, let $\mathbf{v}_h^c \in \mathbf{V}_h^c$, and using again (7.7), we have $\Pi_h^c \mathbf{v}_h^c = \mathbf{v}_h^c$, since $[\![\mathbf{v}_h^c]\!]_T = 0$. This proves that Π_h^c is a projection and that it is surjective. Thus, it defines a stable decomposition of $\mathbf{V}_h$ as $\mathbf{V}_h = \mathbf{V}_h^c \oplus \mathbf{V}_h^\perp$, with $\mathbf{V}_h^\perp = \ker\{\Pi_h^c\}$. In other words, any $\mathbf{v}_h \in \mathbf{V}_h$ is decomposed as $\mathbf{v}_h = \mathbf{v}_h^c + \mathbf{v}_h^\perp$, $\mathbf{v}_h^c = \Pi_h^c \mathbf{v}_h$ and $\mathbf{v}_h^\perp = (I - \Pi_h^c)\mathbf{v}_h$. The estimate (7.7) provides (7.8), and (7.6) provides (7.9), since

$$\|\mathbf{v}_h^\perp\|_{0,\Omega}^2 \le C \int_{\mathcal{F}_h} \mathtt{h} |[\![\mathbf{v}_h]\!]_T|^2 \, ds \le C h^2 \int_{\mathcal{F}_h} \mathtt{h}^{-1} |[\![\mathbf{v}_h^\perp]\!]_T|^2 \, ds \le C h^2 |\mathbf{v}_h^\perp|_{\mathbf{V}(h)}^2. \qquad \square$$

We proceed now by proving Assumption 5. For this, we need the following lemma.

LEMMA 7.6. *We have*

$$\|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega} \le C |\mathbf{v}|_{\mathbf{V}(h)} \qquad \forall \mathbf{v} \in K_h^\perp,$$

with a positive constant C independent of the mesh size.

Proof. Fix $\mathbf{v} \in K_h^\perp$ and decompose it, according to the decompositions (7.8) and (7.4), as $\mathbf{v} = \mathbf{v}^c + \mathbf{v}^\perp = \mathbf{w} + \nabla p + \mathbf{v}^\perp$, with $\mathbf{w} \in \mathbf{W}_h^c$, $p \in Q_h^c$, and $\mathbf{v}^\perp = \mathbf{v} - \Pi_h^c \mathbf{v}$, where Π_h^c is the operator defined in Theorem 7.4. Since $(\mathbf{w}, \nabla q)_\varepsilon = 0$ for all $q \in Q_h^c$, the condition $(\mathbf{v}, \nabla q)_\varepsilon = 0$ for all $q \in Q_h^c$ becomes

$$(\nabla p, \nabla q)_\varepsilon = -(\mathbf{v}^\perp, \nabla q)_\varepsilon \qquad \forall q \in Q_h^c.$$

By taking $q = p$, we obtain that $\|\varepsilon^{1/2}\nabla p\|_{0,\Omega} \le \|\varepsilon^{1/2}\mathbf{v}^\perp\|_{0,\Omega}$, and thus

$$\|\varepsilon^{1/2}\mathbf{v}\|_{0,\Omega} \le \|\varepsilon^{1/2}\mathbf{w}\|_{0,\Omega} + 2\,\|\varepsilon^{1/2}\mathbf{v}^\perp\|_{0,\Omega}. \tag{7.10}$$

For the first term on the right-hand side of (7.10), from the discrete Friedrichs inequality for the conforming Nédélec elements (7.5), the triangle inequality, and (7.7), we get

$$\begin{aligned} \|\varepsilon^{1/2}\mathbf{w}\|_{0,\Omega} &\le C\,\|\mu^{-1/2}\nabla \times \mathbf{w}\|_{0,\Omega} \\ &\le C\left(\|\mu^{-1/2}\nabla_h \times (\mathbf{w} + \mathbf{v}^\perp)\|_{0,\Omega} + \|\mu^{-1/2}\nabla_h \times \mathbf{v}^\perp\|_{0,\Omega}\right) \\ &\le C\left(\|\mu^{-1/2}\nabla_h \times \mathbf{v}\|_{0,\Omega} + \|\mathtt{h}^{-1/2}[\![\mathbf{v}^\perp]\!]_T\|_{0,\mathcal{F}_h}\right) \le C\,|\mathbf{v}|_{\mathbf{V}(h)}. \end{aligned} \tag{7.11}$$

Using again (7.7), we bound the second term on the right-hand side of (7.10) as

$$\|\varepsilon^{1/2}\mathbf{v}^\perp\|_{0,\Omega} \le C\,|\mathbf{v}|_{\mathbf{V}(h)}. \tag{7.12}$$

Inserting (7.11) and (7.12) into (7.10) proves the lemma. $\square$

The following proposition is an immediate consequence of the coercivity in Assumption 3 and Lemma 7.6.

PROPOSITION 7.7. *Assumption* 5 *holds true.*

7.4. Gap property (Assumption 6). The following proposition concludes the analysis of the IP methods and the LDG method.

PROPOSITION 7.8. *Assumption* 6 *holds true.*

Proof. Due to the discrete compactness property for the conforming Nédélec elements (we refer the reader to [18] and [19] for the case of varying coefficients; see also [38, Theorem 7.18]), it can be seen as in [16] that Assumption 6 for the conforming

Nédélec elements holds true (see also the proof of (i) in Proposition 7.13): for all h small enough, for any $\mathbf{w}_h \in \mathbf{W}_h^c$ there exists $\mathbf{w} = \mathbf{w}(h) \in H(\mathrm{div}_\varepsilon^0; \Omega)$ such that

$$\|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega} \leq \eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)}, \tag{7.13}$$

with $\eta_h \to 0$ as $h \to 0$ (we have used the equivalence between the L^2-norm and the L^2_ε-norm).

Now, fix $\mathbf{w}_h \in K_h^\perp$ and decompose it, according to (7.8) and (7.4), as $\mathbf{w}_h = \mathbf{w}_h^c + \mathbf{w}_h^\perp = \mathbf{w}_h^0 + \nabla p_h + \mathbf{w}_h^\perp$, with $\mathbf{w}_h^0 \in \mathbf{W}_h^c$, $p_h \in Q_h^c$, and $\mathbf{w}_h^\perp = \mathbf{w}_h - \Pi_h^c \mathbf{w}_h$, where Π_h^c is the operator defined in Theorem 7.4. For all h small enough, in correspondence to $\mathbf{w}_h^0$, let $\mathbf{w}$ be an element of $H(\mathrm{div}_\varepsilon^0; \Omega)$ which satisfies (7.13). The Cauchy–Schwarz inequality and the L^2_ε-orthogonality of ∇Q_h^c to both $K_h^\perp$ and $H(\mathrm{div}_\varepsilon^0; \Omega)$ give

$$\begin{aligned}
\|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 &= (\mathbf{w} - \mathbf{w}_h, \mathbf{w} - \mathbf{w}_h)_\varepsilon \\
&= (\mathbf{w} - \mathbf{w}_h, \mathbf{w} - \mathbf{w}_h^0)_\varepsilon - (\mathbf{w} - \mathbf{w}_h, \nabla p_h)_\varepsilon - (\mathbf{w} - \mathbf{w}_h, \mathbf{w}_h^\perp)_\varepsilon \\
&\leq \|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega} (\|\varepsilon^{1/2}(\mathbf{w} - \mathbf{w}_h^0)\|_{0,\Omega} + \|\varepsilon^{1/2}\mathbf{w}_h^\perp\|_{0,\Omega}).
\end{aligned}$$

The bounds (7.13) and (7.9), together with the $\mathbf{V}(h)$-stability of the decompositions (7.4) and (7.8), give Assumption 6. ∎

Remark 7.9. With our choice of $\mathbf{V}(h)$, if σ_λ is the regularity exponent of the eigenspace $E(\mathbf{V})$ associated with an eigenvalue $\lambda \neq 1$ of the operator A, i.e., $\mathbf{u} \in H^{\sigma_\lambda}(\mathrm{curl}; \mathcal{T}_h)$ for all $\mathbf{u} \in E(\mathbf{V})$, the exponent t in the eigenvalue and eigenfunction estimates of Theorem 4.14 is given by

$$t = \min\{\ell, \sigma_\lambda\}.$$

Remark 7.10. Numerical results reported in [3] for DG spectral approximations of the Laplace operator have shown that the suboptimal eigenvalue convergence rate of Theorem 4.14 in the case of non-hermitian DG methods (t instead of $2t$) is actually sharp, at least for even approximation polynomial degrees; for odd degrees, one order of convergence better than expected has been observed for smooth solutions. The same behavior has been reported in [29] in the context of error estimation of linear target functionals of the solutions to advection-diffusion-reaction problems.

Remark 7.11. Well-posedness of the DG discretization, for h small enough, of the indefinite source problem (6.1), with ω away from the eigenfrequencies of the continuous problem, has been established in our abstract framework in section 6, together with an inf-sup condition. The result provided in the appendix, together with consistency, guarantees the validity of a quasi-optimal error estimate.

7.5. Relations between Assumption 6 and the discrete compactness property. We conclude this section by establishing directly the relations between Assumption 6 and the so-called *discrete compactness property*.

The discrete compactness property plays a crucial role in the theory of *conforming* finite element methods for the Maxwell eigenproblem (1) (see, e.g., [37], [10], [24]). Here we rephrase this property in the context of nonconforming approximations.

PROPERTY 3 (discrete compactness property). *Let $\{h_n\}_{n=1}^\infty$ be a sequence of decreasing mesh sizes, with $h_n \to 0$ as $n \to \infty$, and let $\{\mathbf{w}_{h_n}\}_{n=1}^\infty$ be a sequence such that $\mathbf{w}_{h_n} \in K_{h_h}^\perp$ and $\|\mathbf{w}_{h_n}\|_{\mathbf{V}(h)} \leq 1$ for all h_n. Then, there exists a subsequence, still denoted $\{\mathbf{w}_{h_n}\}_{n=1}^\infty$, and an element $\mathbf{v} \in L^2(\Omega)^3$ such that*

$$\lim_{h_n \to 0} \|\mathbf{w}_{h_n} - \mathbf{v}\|_{0,\Omega} = 0.$$

Note that if $\mathbf{V}_h \subset \mathbf{V}$ and $\|\cdot\|_{\mathbf{V}(h)} = \|\cdot\|_{\mathbf{V}}$, Property 3 is the standard discrete compactness property for conforming spaces.

It is known that Property 3, the completeness of the approximation spaces (cf. Assumption 2), and the discrete Friedrichs inequality (cf. Assumption 5) are *necessary* and *sufficient* conditions for spurious-free *conforming* approximations (see [18, Theorem 6.8]).

Remark 7.12. If the completeness of the approximation spaces and the discrete Friedrichs inequality hold true, Propositions 2.18 and 2.21 of [18] apply and guarantee that, for conforming approximations, the limit in Property 3 actually belongs to $\mathbf{W}$.

In the following proposition we establish directly the relations between Assumption 6 and Property 3.

PROPOSITION 7.13. *Let Assumption* 1 *hold true. Then, Assumption* 6 *is equivalent to Property* 3 *with strong limit in* $\mathbf{W}$.

Proof. (i) *Property* 3 $\Rightarrow$ *Assumption* 6. We proceed by contradiction. Let Assumption 6 be false; then there exists $\eta > 0$ such that, for all $\bar{h} > 0$, there is $h \in (0, \bar{h})$ and $\mathbf{w}_h \in K_h^\perp$ with $\|\mathbf{w}_h\|_{\mathbf{V}(h)} \le 1$ such that

$$\|\mathbf{w}_h - \mathbf{w}\|_{0,\Omega} > \eta \qquad \forall \mathbf{w} \in H(\mathrm{div}_\varepsilon^0; \Omega). \tag{7.14}$$

Now, select a sequence $\{h_n\}_{n=1}^\infty$ with $h_n \to 0$ as $n \to \infty$. The previous assertion allows us to construct, in correspondence with $\{h_n\}_{n=1}^\infty$, a sequence $\{\mathbf{w}_{h_n}\}_{n=1}^\infty$ with $\mathbf{w}_{h_n} \in K_{h_n}^\perp$ and $\|\mathbf{w}_{h_n}\|_{\mathbf{V}(h_n)} \le 1$ for all h_n, which does not contain any subsequence converging to an element $\mathbf{w} \in H(\mathrm{div}_\varepsilon^0; \Omega)$, owing to (7.14). This contradicts Property 3 with strong limit in $\mathbf{W}$.

(ii) *Assumption* 6 $\Rightarrow$ *Property* 3. Let $\mathbf{w}_h$ be in $K_h^\perp$, and select $\mathbf{w} \in H(\mathrm{div}_\varepsilon^0; \Omega)$ as $\mathbf{w} = P\mathbf{w}_h$, with P being the operator defined at the beginning of section 5.2. Owing to Lemma 5.4, we know that

$$\|\mathbf{w}_h - \mathbf{w}\|_{\mathbf{V}(h)} \le \eta_h \|\mathbf{w}_h\|_{\mathbf{V}(h)}, \tag{7.15}$$

with $\eta_h \to 0$, as $h \to 0$.

Now, let $\{\mathbf{w}_{h_n}\}_{n=1}^\infty$ be a sequence in $K_{h_n}^\perp$, bounded in the $\mathbf{V}(h)$-norm. Decompose $\mathbf{w}_{h_n}$ as $\mathbf{w}_{h_n} = \mathbf{w}_{h_n}^c + \mathbf{w}_{h_n}^\perp$, according to (7.8). The sequence $\{\mathbf{w}_n\}_{n=1}^\infty := \{P\mathbf{w}_{h_n}^c\}_{n=1}^\infty \subset \mathbf{W}$ also is bounded in the $\mathbf{V}(h)$-norm, owing to (7.15). From the compactness of $\mathbf{W}$, endowed with the $\mathbf{V}(h)$-norm, in $L^2(\Omega)^d$, there exists a subsequence still denoted $\{\mathbf{w}_n\}_{n=1}^\infty$ and an element $\mathbf{v} \in \mathbf{W}$ such that

$$\lim_{n\to\infty} \|\mathbf{w}_n - \mathbf{v}\|_{0,\Omega} = 0. \tag{7.16}$$

If $\{\mathbf{w}_{h_n}\}_{n=1}^\infty$ is such that $\mathbf{w}_{h_n} = \mathbf{w}_{h_n}^c + \mathbf{w}_{h_n}^\perp$, according to (7.8), and $P\mathbf{w}_{h_n}^c = \mathbf{w}_n$ for all n, by the triangle inequality we have

$$\|\mathbf{w}_{h_n} - \mathbf{v}\|_{0,\Omega} \le \|\mathbf{w}_{h_n} - P\mathbf{w}_{h_n}\|_{0,\Omega} + \|\mathbf{w}_n - \mathbf{v}\|_{0,\Omega} + \|P\mathbf{w}_{h_n}^\perp\|_{0,\Omega}.$$

The first two terms on the right-hand side converge to zero, owing to (7.15) and (7.16), respectively; since the projector P is L^2-stable, also the third term converges to zero, due to (7.9) and the $\mathbf{V}(h)$-stability of the decomposition (7.8). Thus, Property 3 holds true with strong limit in $\mathbf{W}$. ∎

Remark 7.14. On parallelograms or parallelepipeds, all the results in this section apply to the choice of $\mathbf{V}_h$ in (7.1) with the local Nédélec elements of the first type of degree ℓ, instead of the full polynomials of degree ℓ, allowing us to conclude that

the obtained approximation of Problem 1 is spurious-free. This is not true for the full polynomials of degree ℓ in each variable, namely the local Nédélec elements of the second type of degree ℓ. In fact, let K_h^c be the discrete kernel of the corresponding conforming approximation, and consider the $\mathbf{V}$-orthogonal decomposition $\mathbf{V}_h^c = K_h^c \oplus K_h^{\perp,c}$. Since $\mathbf{V}_h^c \subseteq \mathbf{V}_h$ and $K_h = K_h^c$, then $K_h^{\perp,c} \subseteq K_h^\perp$. Recalling that the conforming Nédélec elements of the second type do not satisfy the discrete compactness property (see [11]), Proposition 7.13 says that they do not satisfy Assumption 6; the inclusion $K_h^{\perp,c} \subseteq K_h^\perp$ implies that also for their discontinuous counterpart Assumption 6 is not satisfied, and then the obtained method cannot be spurious-free.

8. Conclusions. We have presented a theoretical framework for the analysis of DG approximations of the Maxwell eigenproblem with possibly discontinuous coefficients. In particular, we have restricted our attention to DG methods satisfying the usual assumptions for a correct approximation to the coercive Maxwell source problem $\nabla \times (\mu^{-1}\nabla \times \mathbf{u}) + \varepsilon \mathbf{u} = \mathbf{f}$ in the domain Ω with suitable boundary conditions. For these methods, necessary and sufficient conditions for a spurious-free approximation are (i) a discrete Friedrichs inequality and (ii) a gap property between the orthogonal complement of the discrete kernel and the space of divergence-free functions. We have also proved that basically all the DG methods present in the literature actually fit into this framework, at least on meshes with no hanging nodes (the extension to meshes with hanging nodes is currently under investigation). It is worth pointing out that all these methods provide optimal convergence of the eigenfunctions, while the convergence of the eigenvalues is optimal for hermitian DG methods and suboptimal for non-hermitian DG methods. Another consequence of the theory developed in this paper is that all these methods provide a correct approximation to the indefinite Maxwell source problem $\nabla \times (\mu^{-1}\nabla \times \mathbf{u}) - \omega^2 \varepsilon \mathbf{u} = \mathbf{f}$ in Ω, with suitable boundary conditions, also in the case of discontinuous coefficients ε and μ, extending in this way the results obtained in [34] for smooth coefficients.

Appendix. The aim of this appendix is to prove a continuity estimate for all the DG bilinear forms $a_h(\cdot,\cdot)$ introduced in section 7.2. More precisely, we prove that, given $r > 0$ and σ such that $0 < \sigma < \min\{1/2, r\}$, there exists a mesh-dependent seminorm $|\cdot|_{+,\sigma}$ such that the norm

$$\|\boldsymbol{\xi}\|_{+,\sigma}^2 = \|\boldsymbol{\xi}\|_{\mathbf{V}(h)}^2 + |\boldsymbol{\xi}|_{+,\sigma}^2, \tag{8.1}$$

defined for functions $\boldsymbol{\xi} \in H^r(\mathrm{curl}; \mathcal{T}_h)$ with $\nabla_h \times (\mu^{-1}\nabla_h \times \boldsymbol{\xi}) \in L^2(\Omega)^d$, satisfies

$$|a_h(\boldsymbol{\xi}, \mathbf{v}_h)| \le C \|\boldsymbol{\xi}\|_{+,\sigma} \|\mathbf{v}_h\|_{\mathbf{V}(h)} \tag{8.2}$$

for all $\boldsymbol{\xi} \in H^r(\mathrm{curl}; \mathcal{T}_h)$ with $\nabla_h \times (\mu^{-1}\nabla_h \times \boldsymbol{\xi}) \in L^2(\Omega)^d$, and $\mathbf{v}_h \in \mathbf{V}_h$, with a constant $C > 0$ independent of the mesh size. Moreover, for any $s \ge 0$, there holds that

$$\begin{aligned} &\inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\boldsymbol{\xi} - \mathbf{v}_h\|_{+,\sigma} \\ &\qquad \le C h^{\min\{r,\ell,s+1\}} \left(\|\boldsymbol{\xi}\|_{H^r(\mathrm{curl};\mathcal{T}_h)} + \left(\sum_{K \in \mathcal{T}_h} M_K \|\nabla \times (\mu^{-1}\nabla \times \boldsymbol{\xi})\|_{s,K}^2 \right)^{1/2} \right), \end{aligned} \tag{8.3}$$

for all $\boldsymbol{\xi} \in H^r(\mathrm{curl}; \mathcal{T}_h)$ with $\nabla_h \times (\mu^{-1}\nabla_h \times \boldsymbol{\xi}) \in H^s(\mathcal{T}_h)^d$, where M_K is defined as $M_K = \max_{\mathbf{x} \in \overline{K}} |\mu_K(\mathbf{x})|$, and the constant C is independent of the mesh size.

Note that the continuity property (8.2) and the best approximation estimate (8.3) provide a proof for Proposition 7.3 (indeed, they also prove that DG methods provide quasi-optimal approximations for the coercive source problem introduced in Assumption 4) and jointly with consistency and the inf-sup condition (6.6) provide quasi-optimal error estimates for DG solutions to the indefinite problem (6.1).

PROPOSITION 8.1. *With the notation introduced here above,* (8.2) *and* (8.3) *hold true with the seminorm in* (8.1) *defined by*

$$|\boldsymbol{\xi}|^2_{+,\sigma} = \sum_{K\in\mathcal{T}_h} \left(h_K^{2\sigma} M_K \|\mu^{-1}\nabla\times\boldsymbol{\xi}\|^2_{\sigma,K} + h_K^2 M_K \|\nabla\times(\mu^{-1}\nabla\times\boldsymbol{\xi})\|^2_{0,K}\right),$$

where we have set $M_K = \max_{\mathbf{x}\in\overline{K}}|\mu_K(\mathbf{x})|$.

We remark that, whenever μ is an elementwise constant tensor, then

$$|\boldsymbol{\xi}|^2_{+,\sigma} = \sum_{K\in\mathcal{T}_h} \left(h_K^{2\sigma} \|\mu^{-1/2}\nabla\times\boldsymbol{\xi}\|^2_{\sigma,K} + h_K^2 M_K \|\nabla\times(\mu^{-1}\nabla\times\boldsymbol{\xi})\|^2_{0,K}\right).$$

In order to prove Proposition 8.1, we need the following technical lemma.

LEMMA 8.2. *With the notation introduced above, for any* $f\in\mathcal{F}_h$, *we have*

(8.4)
$$\int_f [\![\overline{\mathbf{v}}]\!]_T\cdot(\mu^{-1}\nabla_h\times\boldsymbol{\xi})^{\pm}\,ds \le C|\mathbf{v}|_{\mathbf{V}(h)}(h^{\sigma}_{K^\pm}M^{1/2}_{K^\pm}\|\mu^{-1}\nabla\times\boldsymbol{\xi}\|_{\sigma,K} + h_{K^\pm}M^{1/2}_{K^\pm}\|\nabla\times(\mu^{-1}\nabla\times\boldsymbol{\xi})\|_{0,K^\pm}),$$

where $K^\pm$ *are the two tetrahedra sharing the face* f, *and* $C>0$ *is independent of the mesh size.*

Proof. We assume, to fix the ideas, that $d=3$. We start by introducing some notation. Let K be a tetrahedron, f one of its faces, and $\mathbf{n}$ the normal at f pointing outside K. Let $\boldsymbol{\eta}\in H^{1/2}(f)^3$ be such that $\boldsymbol{\eta}\cdot\mathbf{n}=0$ and $\boldsymbol{\eta}\times\mathbf{n} = [\![\overline{\mathbf{v}}]\!]_T$, and set $\boldsymbol{\phi}=\mu^{-1}\nabla\times\boldsymbol{\xi}$ on K; we know that $\boldsymbol{\phi}\in H^\sigma(K)^3$ and $\nabla\times\boldsymbol{\phi}\in L^2(K)^3$.

We decompose $\boldsymbol{\eta}$ as $\boldsymbol{\eta}=\boldsymbol{\eta}_0+\boldsymbol{\eta}_M$, with $\boldsymbol{\eta}_M = \frac{1}{|f|}\int_f\boldsymbol{\eta}\,ds$, and $\boldsymbol{\eta}_0=\boldsymbol{\eta}-\boldsymbol{\eta}_M$. Note that $\boldsymbol{\eta}_0$ has zero mean value on f, and it holds that

$$\|\boldsymbol{\eta}\|^2_{0,f} = \|\boldsymbol{\eta}_0\|^2_{0,f} + \|\boldsymbol{\eta}_M\|^2_{0,f}. \tag{8.5}$$

We can write

$$\int_f(\boldsymbol{\eta}\times\mathbf{n})\cdot\boldsymbol{\phi}\,ds = \int_f(\boldsymbol{\eta}_0\times\mathbf{n})\cdot\boldsymbol{\phi}\,ds + \int_f(\boldsymbol{\eta}_M\times\mathbf{n})\cdot\boldsymbol{\phi}\,ds = \text{(I)}+\text{(II)}. \tag{8.6}$$

Estimate of (I). If we map vector fields onto the reference tetrahedron $\widehat{K}$ by means of the standard curl-conforming transformation (see [38, p. 77]), we have (with self-evident notation)

$$\int_f(\boldsymbol{\eta}_0\times\mathbf{n})\cdot\boldsymbol{\phi}\,ds = \pm\int_{\widehat{f}}(\widehat{\boldsymbol{\eta}}_0\times\widehat{\mathbf{n}})\cdot\widehat{\boldsymbol{\phi}}\,d\widehat{s}$$

(see [38, p. 80]). Notice that $\int_f\boldsymbol{\eta}_0\,ds=0$ implies $\int_{\widehat{f}}\widehat{\boldsymbol{\eta}}_0\,ds=0$, and

$$|\widehat{\boldsymbol{\eta}}_0|_{1/2,f}\le\|\widehat{\boldsymbol{\eta}}_0\|_{1/2,f}\le C|\widehat{\boldsymbol{\eta}}_0|_{1/2,f}. \tag{8.7}$$

Let $\mathrm{R} : H^{1/2-\sigma}(\widehat{f})^2 \to H^{1-\sigma}(\widehat{K})^3$ be a continuous lifting operator from $\widehat{f}$ to $\widehat{K}$ such that $\mathrm{R}\widehat{\boldsymbol{\eta}}_0$ has zero tangential trace on $\partial\widehat{K} \setminus \widehat{f}$ (note that this lifting is the standard one, component by component). By continuity of R, we have

$$\|\mathrm{R}\widehat{\boldsymbol{\eta}}_0\|_{0,\widehat{K}} + \|\widehat{\nabla} \times \mathrm{R}\widehat{\boldsymbol{\eta}}_0\|_{-\sigma,\widehat{K}} \le C(\sigma)|\widehat{\boldsymbol{\eta}}_0|_{1/2-\sigma,\widehat{f}} \le C(\sigma)|\widehat{\boldsymbol{\eta}}_0|_{1/2,\widehat{f}},$$

where $\widehat{\nabla} \times \cdot$ denotes the curl operator with respect to the reference coordinates. By integration by parts, since $\mathrm{R}\widehat{\boldsymbol{\eta}}_0$ has zero trace on $\partial\widehat{K} \setminus \widehat{f}$, we get

$$\int_{\widehat{K}} \widehat{\nabla} \times \mathrm{R}\widehat{\boldsymbol{\eta}}_0 \cdot \mathbf{e}_i = \int_{\partial\widehat{K}} (\mathbf{e}_i \times \widehat{\mathbf{n}}) \cdot \mathrm{R}\widehat{\boldsymbol{\eta}}_0 = \int_{\widehat{f}} (\mathbf{e}_i \times \widehat{\mathbf{n}}) \cdot \widehat{\boldsymbol{\eta}}_0 = 0\,, \qquad i = 1,2,3, \tag{8.8}$$

where $\mathbf{e}_i$, $i = 1,2,3$, are the canonical basis vectors in $\mathbb{R}^3$. This means each component of $\widehat{\nabla} \times \mathrm{R}\widehat{\boldsymbol{\eta}}_0$ has zero mean value. Thus, it holds that

$$\begin{aligned}
\int_{\widehat{f}} (\widehat{\boldsymbol{\eta}}_0 \times \widehat{\mathbf{n}}) \cdot \widehat{\boldsymbol{\phi}}\, d\widehat{s} &= \int_{\widehat{K}} (\widehat{\nabla} \times \widehat{\boldsymbol{\phi}} \cdot \mathrm{R}\widehat{\boldsymbol{\eta}}_0 - \widehat{\boldsymbol{\phi}} \cdot \widehat{\nabla} \times \mathrm{R}\widehat{\boldsymbol{\eta}}_0)\, d\widehat{\mathbf{x}} \\
&\le C\|\widehat{\nabla} \times \widehat{\boldsymbol{\phi}}\|_{0,\widehat{K}} \|\mathrm{R}\widehat{\boldsymbol{\eta}}_0\|_{0,\widehat{K}} + |\widehat{\boldsymbol{\phi}}|_{\sigma,K} \|\nabla \times \mathrm{R}\widehat{\boldsymbol{\eta}}_0\|_{-\sigma,\widehat{K}} \\
&\le C(\|\widehat{\nabla} \times \widehat{\boldsymbol{\phi}}\|_{0,\widehat{K}} + |\widehat{\boldsymbol{\phi}}|_{\sigma,\widehat{K}}) |\widehat{\boldsymbol{\eta}}_0|_{1/2,\widehat{f}},
\end{aligned}$$

where we have used the continuity estimate for R, (8.7), and (8.8). Scaling arguments can be applied (see [2, Lemma 5.5]), using the shape regularity of the meshes

$$\begin{aligned}
|\widehat{\boldsymbol{\eta}}_0|_{1/2,\widehat{f}} &\le C h_K^{1/2} |\boldsymbol{\eta}_0|_{1/2,f}, \\
\|\widehat{\nabla} \times \widehat{\boldsymbol{\phi}}\|_{0,\widehat{K}} &\le C h_K^{1/2} \|\nabla \times \boldsymbol{\phi}\|_{0,K}, \\
|\widehat{\boldsymbol{\phi}}|_{\sigma,\widehat{K}} &\le C h_K^{-1/2+\sigma} |\boldsymbol{\phi}|_{\sigma,K},
\end{aligned}$$

and we obtain

$$\int_f (\boldsymbol{\eta}_0 \times \mathbf{n}) \cdot \boldsymbol{\phi}\, ds \le C(h_K \|\nabla \times \boldsymbol{\phi}\|_{0,K} + h_K^{\sigma} |\boldsymbol{\phi}|_{\sigma,K}) |\boldsymbol{\eta}_0|_{1/2,f}.$$

Since, by inverse inequality and (8.5), it holds that

$$|\boldsymbol{\eta}_0|_{1/2,f} \le C\|h_f^{-1/2}\boldsymbol{\eta}_0\|_{0,f} \le C\|h_f^{-1/2}\boldsymbol{\eta}\|_{0,f},$$

and the definition of $\mathtt{h}$ implies that

$$\|h_f^{-1/2}\boldsymbol{\eta}\|_{0,f} = \|\mathtt{m}^{1/2} h_f^{-1/2} \mathtt{m}^{-1/2} \boldsymbol{\eta}\|_{0,f} \le M_K^{1/2} \|\mathtt{h}^{-1/2}\boldsymbol{\eta}\|_{0,f}, \tag{8.9}$$

we have

$$\text{(I)} \le C(h_K M_K^{1/2} \|\nabla \times \boldsymbol{\phi}\|_{0,K} + h_K^{\sigma} M_K^{1/2} |\boldsymbol{\phi}|_{\sigma,K}) \|\mathtt{h}^{-1/2}\boldsymbol{\eta}\|_{0,f}. \tag{8.10}$$

Estimate of (II). From the "lifting property" proved in [13], we know that, for any fixed $t < 2$, there exists a function $\varphi_f \in H^1(K)$ such that

$$\begin{aligned}
\varphi_f &= 1 \text{ on } f\,, \qquad \varphi_f = 0 \text{ on } \partial K \setminus f, \\
\|\varphi_f\|_{0,K} &\le C h_K^{3/2}, \qquad \|\nabla\varphi_f\|_{L^t(K)^3} \le C h_K^{3/t-1}.
\end{aligned}$$

By taking t' such that $1/t' + 1/t = 1$, it holds that

$$\begin{aligned}\int_f (\boldsymbol{\eta}_M \times \mathbf{n}) \cdot \boldsymbol{\phi}\, ds &= \int_f (\varphi_f \boldsymbol{\eta}_M \times \mathbf{n}) \cdot \boldsymbol{\phi}\, ds \\ &= \int_K \nabla \times \boldsymbol{\phi} \cdot (\varphi_f \boldsymbol{\eta}_M)\, d\mathbf{x} - \int_K \boldsymbol{\phi} \cdot \nabla \times (\varphi_f \boldsymbol{\eta}_M)\, d\mathbf{x} \\ &\le C(h_K^{3/2} \|\nabla \times \boldsymbol{\phi}\|_{0,K} |\boldsymbol{\eta}_M| + h_K^{3/t-1} \|\boldsymbol{\phi}\|_{L^{t'}(K)^3} |\boldsymbol{\eta}_M|) \\ &\le C(h_K \|\nabla \times \boldsymbol{\phi}\|_{0,K} + h_K^{3/t-3/2} \|\boldsymbol{\phi}\|_{L^{t'}(K)^3})(h_K^{-1/2} \|\boldsymbol{\eta}_M\|_{0,f}),\end{aligned}$$

where $|\boldsymbol{\eta}_M|$ denotes the modulus of the constant vector $\boldsymbol{\eta}_M$ and where we have used $|\boldsymbol{\eta}_M| \le C h_K^{-1} \|\boldsymbol{\eta}_M\|_{0,f}$. Now, let $t' = \frac{6}{3-2\sigma}$; by Sobolev embedding theorem (see, e.g., [1, p. 217]), we have that

$$\|\boldsymbol{\phi}\|_{L^{t'}(K)^3} \le C \|\boldsymbol{\phi}\|_{\sigma,K}.$$

The shape regularity of the meshes and (8.5) imply $h_K^{-1/2} \|\boldsymbol{\eta}_M\|_{0,f} \le C \|h_f^{-1/2} \boldsymbol{\eta}\|_{0,f}$; thus, by using (8.9) and simple algebra, we obtain

$$\text{(II)} \le C(h_K M_K^{1/2} \|\nabla \times \boldsymbol{\phi}\|_{0,K} + h_K^{\sigma} M_K^{1/2} \|\boldsymbol{\phi}\|_{\sigma,K}) \|\mathfrak{h}^{-1/2} \boldsymbol{\eta}\|_{0,f}. \tag{8.11}$$

Taking into account the definitions of $\boldsymbol{\phi}$, $\boldsymbol{\eta}$ and of $|\cdot|_{\mathbf{V}(h)}$, the expression (8.6) and the estimates (8.10) and (8.11) give (8.4), and the proof is complete. ∎

We are now in a position to prove Proposition 8.1.

Proof of Proposition 8.1. For all the DG methods of section 7.2, it is easy to see that, for all $\boldsymbol{\xi} \in H^r(\mathrm{curl}; \mathcal{T}_h)$ with $\nabla_h \times (\mu^{-1} \nabla_h \times \boldsymbol{\xi}) \in L^2(\Omega)^3$ and $\mathbf{v} \in \mathbf{V}_h$, it holds that

$$a_h(\boldsymbol{\xi}, \mathbf{v}_h) \le C \|\boldsymbol{\xi}\|_{\mathbf{V}(h)} \|\mathbf{v}_h\|_{\mathbf{V}(h)} + C' \int_{\mathcal{F}_h} [\![\overline{\mathbf{v}_h}]\!]_T \cdot \{\!\{\mu^{-1} \nabla_h \times \boldsymbol{\xi}\}\!\}\, ds. \tag{8.12}$$

Summing (8.4) of Lemma 8.2 over all $f \in \mathcal{F}_h$ gives

$$\int_{\mathcal{F}_h} [\![\overline{\mathbf{v}_h}]\!]_T \cdot \{\!\{\mu^{-1} \nabla_h \times \boldsymbol{\xi}\}\!\}\, ds \le C |\boldsymbol{\xi}|_{+,\sigma} \|\mathbf{v}_h\|_{\mathbf{V}(h)}.$$

Inserting this into (8.12) completes the proof of (8.2). The best approximation estimate (8.3) is a direct consequence of standard polynomial approximation properties. ∎

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Alonso and A. Valli, *A domain decomposition approach for heterogeneous time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 97–112.

[3] P. F. Antonietti, A. Buffa, and I. Perugia, *Discontinuous Galerkin approximation of the Laplace eigenproblem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3483–3503.

[4] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[5] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[6] G. A. Baker, W. N. Jureidini, and O. A. Karakashian, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.

[7] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, R. Decuypere and G. Dibelius, eds., Technologisch Instituut, Antwerpen, Belgium, 1997, pp. 99–108.

[8] K.-J. Bathe and X. Wang, *On mixed elements for acoustic fluid-structure interactions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 329–344.

[9] A. Bermúdez, R. Durán, M. A. Muschietti, R. Rodríguez, and J. Solomin, *Finite element vibration analysis of fluid–solid systems without spurious modes*, SIAM J. Numer. Anal., 32 (1995), pp. 1280–1295.

[10] D. Boffi, *Fortin operator and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[11] D. Boffi, M. Costabel, M. Dauge, and L. Demkowicz, *Discrete compactness for the hp version of rectangular edge finite elements*, SIAM J. Numer. Anal., 44 (2006), pp. 979–1004.

[12] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[13] F. Brezzi, K. Lipnikov, and M. Shashkov, *Convergence of mimetic finite difference methods for diffusion problems on polyhedral meshes*, SIAM J. Numer. Anal., 43 (2005), pp. 1872–1896.

[14] F. Brezzi, D. Marini, P. Pietra, and A. Russo, *Discontinuous finite elements for diffusion problems*, in Atti Convegno in onore di F. Brioschi (Milano 1997), Istituto Lombardo, Accademia di Scienze e Lettere, Milano, Italy, 1999, pp. 197–217.

[15] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite dimensional approximation of nonlinear problems. Part* I: *Branches of nonsingular solutions*, Numer. Math., 36 (1980/81), pp. 1–25.

[16] A. Buffa, *Remarks on the discretization of some noncoercive operator with applications to heterogeneous Maxwell equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1–18.

[17] A. Buffa, P. Houston, and I. Perugia, *Discontinuous Galerkin computation of the Maxwell eigenvalues on simplicial meshes*, J. Comput. Appl. Math., to appear.

[18] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.

[19] S. Caorsi, P. Fernandes, and M. Raffetto, *Spurious-free approximations of electromagnetic eigenproblems by means of Nédélec-type elements*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 331–354.

[20] B. Cockburn, F. Li, and C.-W. Shu, *Locally divergence-free discontinuous Galerkin methods for the Maxwell equations*, J. Comput. Phys., 194 (2004), pp. 588–610.

[21] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[22] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.

[23] C. Dawson, S. Sun, and M. F. Wheeler, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[24] L. Demkowicz and P. Monk, *Discrete compactness and the approximation of Maxwell's equations in* $\mathbb{R}^3$, Math. Comp., 70 (2001), pp. 507–523.

[25] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation Part* 1. *The problem of convergence*, RAIRO Anal. Numér., 12 (1978), pp. 97–112.

[26] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation Part* 2. *Error estimates for the Galerkin method convergence*, RAIRO Anal. Numér., 12 (1978), pp. 113–119.

[27] N. Dunford and J. T. Schwartz, *Spectral Theory: Self Adjoint Operators in Hilbert Space*, Interscience, New York, 1963.

[28] P. Fernandes and G. Gilardi, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.

[29] K. Harriman, P. Houston, B. Senior, and E. Süli, *hp-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form*, in Recent Advances in Scientific Computing and Partial Differential Equations, Contemp. Math. 330, C.-W. Shu, T. Tang, and S.-Y. Cheng, eds., AMS, Providence, RI, 2003, pp. 89–119.

[30] J. S. Hesthaven and T. Warburton, *Nodal high-order methods on unstructured grids. Part* I.

Time-domain solution of Maxwell's equations, J. Comput. Phys., 181 (2002), pp. 186–221.
[31] J. S. Hesthaven and T. Warburton, *High-order accurate methods for time-domain electromagnetics*, CMES Comput. Model. Eng. Sci., 5 (2004), pp. 395–408.
[32] J. S. Hesthaven and T. Warburton, *High order nodal discontinuous Galerkin methods for the Maxwell eigenvalue problem*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 493–524.
[33] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.
[34] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau, *Interior penalty method for the indefinite time-harmonic Maxwell equations*, Numer. Math., 100 (2005), pp. 485–518.
[35] P. Houston, I. Perugia, and D. Schötzau, *Mixed discontinuous Galerkin approximation of the Maxwell operator: Non-stabilized formulation*, J. Sci. Comput., 22 (2005), pp. 325–356.
[36] T. Kato, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.
[37] F. Kikuchi, *Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism*, Comput. Methods Appl. Mech. Engrg., 64 (1987), pp. 509–512.
[38] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, New York, 2003.
[39] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[40] J. C. Nédélec, *Éléments finis mixtes incompressibles pour l'équation de Stokes dans* $\mathbb{R}^3$, Numer. Math., 39 (1982), pp. 97–112.
[41] J. C. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.
[42] I. Perugia and D. Schötzau, *The hp-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell equations*, Math. Comp., 72 (2003), pp. 1179–1214.
[43] I. Perugia, D. Schötzau, and P. Monk, *Stabilized interior penalty methods for the time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4675–4697.
[44] B. Rivière, M. F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems, Part* I, Comput. Geosci., 3 (1999), pp. 337–360.
[45] T. Warburton and M. Embree, *The role of the penalty in the local discontinuous Galerkin method for Maxwell's eigenvalue problem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3205–3223.

SIAM J. NUMER. ANAL.
Vol. 44, No. 5, pp. 2227–2243

AN ALGEBRAIC PROCEDURE FOR THE SPECTRAL CORRECTIONS USING THE MISS-DISTANCE FUNCTIONS IN REGULAR AND SINGULAR STURM–LIOUVILLE PROBLEMS*

LIDIA ACETO†, PAOLO GHELARDONI†, AND GIOVANNI GHERI†

Abstract. A general method based on the evaluation of the zeros of a suitable polynomial is suggested in order to have an estimation of the spectral error in the numerical treatment of Sturm–Liouville problems. The method is strictly concerned with the miss-distance function arising in the shooting algorithm for eigenvalues. The error correcting procedure derived from the method is particularly helpful when difficulties arise in the numerical integration. Two kinds of Sturm–Liouville problems are considered: the standard regular problems on a closed interval and the problems where an eigenvalue is nonlinearly involved and embedded in an essential spectrum giving origin to an inner singularity. Numerical experiments clearly highlight the efficaciousness of the proposed method both in the regular and singular case.

Key words. Sturm–Liouville problem, algebraic spectral correction, miss-distance

AMS subject classifications. 65L10, 65L12, 65L15

DOI. 10.1137/050635092

1. Introduction.

The subject matter of this paper is the numerical computation of the eigenvalues in a Sturm–Liouville problem (SLP) by the shooting technique and a corresponding spectral error correcting procedure based on the evaluation of the zeros of a suitable polynomial. Two classes of problems are considered here. The first is concerned with the classical SLP in its regular form given by the differential equation

$$-(p(x)y'(x))' + q(x)y(x) = \lambda w(x)y(x) \tag{1.1}$$

on a finite interval $a < x < b$ with the boundary conditions (BCs)

$$\begin{aligned} a_1 y(a) - a_2 p(a) y'(a) &= 0, \\ b_1 y(b) - b_2 p(b) y'(b) &= 0. \end{aligned} \tag{1.2}$$

In the differential equation the real functions $q(x)$ and $w(x)$ are continuous on the interval $[a, b]$ with $w(x) > 0$, while the real function $p(x)$ is strictly positive and almost once differentiable. In the BCs the constants a_1 and a_2 are real and not both equal to zero; similarly for b_1 and b_2.

In section 4 these conditions will be sharpened and the problem (1.1)–(1.2) mainly considered in the following equivalent matrix form:

$$z' = H(x, \lambda)z, \quad a < x < b, \tag{1.3}$$

where

$$z(x) = \begin{bmatrix} y(x) \\ y'(x) \end{bmatrix}, \quad H(x, \lambda) = \begin{bmatrix} 0 & 1 \\ \frac{q - \lambda w}{p} & -\frac{p'}{p} \end{bmatrix}. \tag{1.4}$$

*Received by the editors July 4, 2005; accepted for publication (in revised form) May 3, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sinum/44-5/63509.html
†Dipartimento di Matematica Applicata, Università di Pisa, Via Buonarroti 1/c, I-56127 Pisa, Italy (l.aceto@dma.unipi.it, ghelardoni@dma.unipi.it, gheri@dma.unipi.it).

The BCs (1.2) will be written as

$$\alpha^T Jz(a) = 0, \quad \beta^T Jz(b) = 0, \tag{1.5}$$

with

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \alpha = \begin{bmatrix} a_2 p(a) \\ a_1 \end{bmatrix}, \quad \beta = \begin{bmatrix} b_2 p(b) \\ b_1 \end{bmatrix}. \tag{1.6}$$

The second kind of SLP which we consider here is given by the differential equation

$$-y''(x) = g(x,\lambda)y(x), \quad a < x < b, \tag{1.7}$$

with the BCs

$$\begin{aligned} a_1 y(a) - a_2 y'(a) &= 0, \\ b_1 y(b) - b_2 y'(b) &= 0, \end{aligned} \tag{1.8}$$

and a_1, a_2, b_1, b_2 satisfying the same conditions as in the regular case; however in sections 3, 4, and 5 we will require more strong conditions.

In this section we assume that

$$g(x,\lambda) = \lambda + \frac{q(x)}{u(x) - \lambda}$$

with $\lambda \in \mathbb{R}$, $q(x), u(x) \in C^1([a,b])$ and $u(x)$ strictly monotone increasing. Thus if $\lambda \in [u(a), u(b)]$, then there is a unique point $x_\lambda \in [a,b]$ such that $u(x_\lambda) = \lambda$.

In what follows we also require $q(x) > 0$ for all $x \in [a,b]$. Then it is well known that the problem (1.7)–(1.8) can be set in the so-called λ-linear block operator problem

$$\begin{bmatrix} -\frac{d^2}{dx^2} & \sqrt{q(x)} \\ \sqrt{q(x)} & u(x) \end{bmatrix} \tilde{y} = \lambda \tilde{y}, \tag{1.9}$$

where $\tilde{y} = \left[y, -\frac{y\sqrt{q}}{u-\lambda} \right]^T$.

Problems of this type, both (1.7) and (1.9), have been investigated recently (see, for instance, [1, 13, 14]) and play an important role in magnetohydrodynamics such as Hain–Lüst equations (see, for example, [15]) and also go by the name of λ-rational SLPs.

For a given $\lambda \in [u(a), u(b)]$ a solution of (1.7) is a function $y(x)$ satisfying (1.7) for $x \neq x_\lambda$ and such that

$$\lim_{x \to x_\lambda} g(x,\lambda)y(x) \quad \text{exists.} \tag{1.10}$$

If $y(x)$ satisfies the BCs (1.8), then we say that $y(x)$ is an eigenfunction associated with the eigenvalue λ embedded in the essential spectrum.

As for the regular case, beside the scalar form (1.7)–(1.8) we consider the matrix representation

$$z' = K(x,\lambda)z, \quad a < x < b\,, \tag{1.11}$$

with

$$z(x) = \begin{bmatrix} y(x) \\ y'(x) \end{bmatrix}, \quad K(x,\lambda) = \begin{bmatrix} 0 & 1 \\ -g(x,\lambda) & 0 \end{bmatrix} \tag{1.12}$$

and the boundary conditions

$$\tilde{\alpha}^T Jz(a) = 0, \quad \tilde{\beta}^T Jz(b) = 0, \tag{1.13}$$

where

$$\tilde{\alpha} = \begin{bmatrix} a_2 \\ a_1 \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} b_2 \\ b_1 \end{bmatrix}, \tag{1.14}$$

and J as in (1.6).

The key for the shooting is given by the Theorem 4.1 in [1] which we quote here for completeness in the following simplified form.

THEOREM 1.1. *For $\lambda \in [u(a), u(b)]$ there is a unique solution $y(x)$ of (1.7) fulfilling the conditions $y(x_\lambda) = 0$, $y'(x_\lambda) = 1$.*

Thus, with regard to the matrix formulation (1.11) of the problem, the condition for $z(x)$ when $x = x_\lambda$ is given by

$$z(x_\lambda) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{1.15}$$

From this we see that an eigenvalue embedded in the essential spectrum represents a pathological case because its eigenfunction must satisfy not only the BCs (1.13) but also the inner condition (1.15).

As far as the problem approximating the eigenvalues in a general SLP is concerned, several papers have been produced (see, for instance, [2, 3, 12, 18, 19]).

In [7, 8] a technique is proposed to obtain a spectral correction based on a particular discretization parameter. Nevertheless the present paper differs from the other ones because it is a generalization of a spectral correcting procedure obtained in [9] starting from a straight computation of the discretization error and an extension to the parallel shooting.

In the next section we look onto the parallel shooting both for the regular and singular case and we build up the corresponding miss-distance functions, namely the functions describing the extent to which some matching conditions fail to be satisfied at the shooting nodes or at the boundary points.

In section 3 we describe the discretization procedure based on the boundary value method (BVM) [4]. We state the reason for this choice observing that BVM is a high order procedure without barriers to the numerical stability. Furthermore the use of a BVM having symmetric scheme [4, p. 159] leads to an approximate solution analytic in a discretization parameter (see, for instance, Gragg's theorem in [10]) allowing particular facilities in theorem proving.

In that section we also highlight that the BVM experiences a decay of its performances in the presence of the inner singularity rendering the correcting procedure particularly useful.

In sections 4 and 5 the correcting procedures in the regular and singular cases are given.

Section 6 is devoted to some numerical experiments underlining the effectiveness of the methods proposed.

2. Building up the miss-distances. We begin with the regular case splitting the problem (1.3)–(1.5) into two initial value problems (IVPs)

$$z' = H(x, \lambda)z, \quad z(a) = \alpha, \tag{2.1}$$

$$z' = H(x, \lambda)z, \quad z(b) = \beta. \tag{2.2}$$

Let c be a point of $[a,b]$. In order to integrate (2.1) left to right we consider m_a+1 shooting nodes ξ_i, $i=0,1,\dots,m_a$, such that $a=\xi_0<\xi_1<\cdots<\xi_{m_a}=c$ and the IVPs

$$\begin{aligned} U_j'(x) &= H(x,\lambda)U_j(x), \quad \xi_{j-1}<x\le\xi_j,\\ U_j(\xi_{j-1}) &= I, \qquad\qquad\qquad j=1,2,\dots,m_a, \end{aligned} \tag{2.3}$$

I being the identity matrix of order two. Thus the solution of (2.1) is

$$z_L(x)\equiv z_j(x)=U_j(x)s_j, \quad j=1,2,\dots,m_a, \tag{2.4}$$

where the vectors $s_1,s_2,\dots,s_{m_a}$ are chosen to ensure the continuity of $z_L(x)$ across the interior nodes. Namely the conditions $z_j(\xi_j)=z_{j+1}(\xi_j)$, $j=1,2,\dots,m_a-1$ are equivalent to the conditions $U_j(\xi_j)s_j=s_{j+1}$, $j=1,2,\dots,m_a-1$. By recurrence, we have $s_{m_a}=U_{m_a-1}(\xi_{m_a-1}),\dots,U_1(\xi_1)s_1$, and from (2.4) $z_L(\xi_{m_a})=U_{m_a}(\xi_{m_a})s_{m_a}=U_{m_a}(\xi_{m_a}),\dots,U_1(\xi_1)s_1$. Because $z_L(a)=U_1(a)s_1=s_1=\alpha$, setting

$$U=U_{m_a}(\xi_{m_a}),\dots,U_1(\xi_1), \tag{2.5}$$

we can write $z_L(\xi_{m_a})=U\alpha$.

We integrate the IVP (2.2) right to left considering the m_b+1 shooting nodes θ_i, $i=0,1,\dots,m_b$, such that $c=\theta_{m_b}<\theta_{m_b-1}<\cdots<\theta_0=b$ and the IVPs

$$\begin{aligned} V_j'(x) &= H(x,\lambda)V_j(x), \quad \theta_j\le x<\theta_{j-1},\\ V_j(\theta_{j-1}) &= I, \qquad\qquad\qquad j=1,2,\dots,m_b. \end{aligned} \tag{2.6}$$

Then the solution of (2.2) is

$$z_R(x)\equiv z_j(x)=V_j(x)\sigma_j, \quad j=1,2,\dots,m_b,$$

where the vectors $\sigma_1,\sigma_2,\dots,\sigma_{m_b}$ play the same role in the right side of the vectors as $s_1,s_2,\dots,s_{m_a}$ in the left one. Repeating almost verbatim the foregoing considerations, we find that $z_R(\theta_{m_b})=V\beta$, where

$$V=V_{m_b}(\theta_{m_b}),\dots,V_1(\theta_1). \tag{2.7}$$

Consider now the Wronskian determinant

$$\det(z_R(\theta_{m_b}),z_L(\xi_{m_a}))=z_L^T(\xi_{m_a})Jz_R(\theta_{m_b})=\alpha^TU^TJV\beta. \tag{2.8}$$

The matching condition at $\xi_{m_a}=\theta_{m_b}=c$ is satisfied if the values of λ are eigenvalues of the SLP (1.3)–(1.5), namely solutions of the equation

$$\alpha^TU^TJV\beta=0.$$

In what follows we no longer mark the left solution as $z_L(x)$ and the right one as $z_R(x)$ because it will be clear from the context which is the $z(x)$ to consider.

In practice we integrate the IVPs (2.3) and (2.6) using a numerical method with a constant stepsize h and starting up with an initial guess μ of λ.

In the left side we consider $x_n=a+nh$, $n=0,1,\dots,n_a=\frac{c-a}{h}$, assuming, for the sake of simplicity, c and h such that $\frac{c-a}{h}$ is an integer number. Thus the shooting nodes can be defined as $\xi_j=x_{jl_a}$, $j=0,1,\dots,m_a$ for some integer l_a with $m_al_a=n_a$.

Let $U_{j,n}$ be the approximation to $U_j(x_n)$ and z_n the one to $z(x_n)$ obtained using a numerical method. We assume $U_{j,n}$ and z_n depend on μ and on some discretization parameter t, that is $U_{j,n} = U_{j,n}(\mu, t)$ and $z_n = z_n(\mu, t)$, with $(j-1)l_a \le n \le jl_a$, $j = 1, 2, \ldots, m_a$.

Furthermore we assume t to be a function on h, i.e., $t = t(h)$ with the property

$$t(h) \neq 0 \quad \text{if} \quad h \neq 0, \quad \lim_{h\to 0} t(h) = 0. \tag{2.9}$$

For a classical method of order p we have $t(h) = h^p$.

Analogously in the right side we consider $x_n = b - nh$, $n = 0, 1, \ldots, n_b = \frac{b-c}{h}$, with $\frac{b-c}{h}$ an integer number, and choosing the shooting nodes as $\theta_j = x_{jl_b}$, $j = 0, 1, \ldots, m_b$ for some integer l_b such that $m_b l_b = n_b$.

Using the same numerical method we integrate the IVPs in (2.6) with μ instead of λ obtaining $V_{j,n}(\mu, t) \simeq V_j(x_n)$ and $z_n(\mu, t) \simeq z(x_n)$, $(j-1)l_b \le n \le jl_b$, $j = 1, 2, \ldots, m_b$.

Denoting $U_{j,jl_a}(\mu, t)$ by $U_j(\mu, t)$, $j = 1, 2, \ldots, m_a$, and $V_{j,jl_b}(\mu, t)$ by $V_j(\mu, t)$, $j = 1, 2, \ldots, m_b$, (2.5) and (2.7) become, respectively,

$$U(\mu, t) = U_{m_a}(\mu, t), \ldots, U_1(\mu, t) \quad \text{and} \quad V(\mu, t) = V_{m_b}(\mu, t), \ldots, V_1(\mu, t).$$

With these notations the Wronskian determinant (2.8) takes the form

$$F(\mu, t) = \alpha^T U(\mu, t)^T J V(\mu, t)\beta, \tag{2.10}$$

and goes by the name of *miss-distance*.

Assuming that

$$\lim_{t\to 0} U_{j,n}(\mu, t) = U_j(x_n), \quad \lim_{t\to 0} V_{j,n}(\mu, t) = V_j(x_n), \quad \lim_{t\to 0} z_n(\mu, t) = z(x_n),$$

then, for a fixed t, μ is an approximation to λ if μ is a zero of the miss-distance $F(\mu, t)$. Thus we admit that μ is dependent on t as well, i.e., $\mu = \mu(t)$ and $\lim_{t\to 0} \mu(t) = \lambda$. By consequence we are allowed to state by definition

$$U(\lambda, 0) = U, \quad V(\lambda, 0) = V, \quad F(\lambda, 0) = 0.$$

We remark that if $c = b$ we have the forward parallel shooting, while if $c = a$ we have the backward parallel shooting. If $l_a = n_a$ or $l_b = n_b$ the corresponding shooting is simple.

We consider now the λ-rational SLP (1.11)–(1.14).

Owing to (1.10) and Theorem 1.1, the parallel shooting is carried out starting from the singular point. Let c_a, c_b be points of $[a, x_\lambda)$ and $(x_\lambda, b]$, respectively. We split the SLP into two couples of IVPs:

$$z'(x) = K(x, \lambda)z(x), \quad a < x \le c_a, \quad z(a) = \tilde{\alpha}, \tag{2.11}$$

$$z'(x) = K(x, \lambda)z(x), \quad c_a \le x < x_\lambda, \quad z(x_\lambda) = \gamma, \tag{2.12}$$

where, according to (1.15), we have indicated that $\gamma = [0, 1]^T$ and

$$z'(x) = K(x, \lambda)z(x), \quad x_\lambda < x \le c_b, \quad z(x_\lambda) = \gamma, \tag{2.13}$$

$$z'(x) = K(x, \lambda)z(x), \quad c_b \le x < b, \quad z(b) = \tilde{\beta}. \tag{2.14}$$

Let μ be an initial guess of λ somehow estimated and x_μ the solution of the equation $u(x)-\mu=0$. Following a procedure quite similar to that previously seen, we consider the leftward problem (2.11) with μ in place of λ setting $x_n = a+nh$, $n=0,1,\dots,n_a = \frac{c_a-a}{h}$, and assuming as shooting nodes $\xi_j = x_{jk_a}$, $j=0,1,\dots,r_a$, with $r_a k_a = n_a$. Then we integrate with a numerical method the IVPs

$$\begin{aligned} X_j'(x) &= K(x,\mu)X_j(x), \quad && \xi_{j-1} < x \le \xi_j, \\ X_j(\xi_{j-1}) &= I, && j=1,2,\dots,r_a, \end{aligned} \tag{2.15}$$

obtaining $X_{j,n}(\mu,t) \simeq X_j(x_n)$, $(j-1)k_a \le n \le jk_a$, $j=1,2,\dots,r_a$. Then denoting $X_{j,jk_a}(\mu,t)$ by $X_j(\mu,t)$ we are able to define $X(\mu,t) = X_{r_a}(\mu,t),\dots,X_1(\mu,t)$.

In the same way we consider the problem (2.12) with μ in place of λ and x_μ in place of x_λ setting $x_n = x_\mu - nh$, $n=0,1,\dots,n_{\mu_a} = \frac{x_\mu - c_a}{h}$ and assuming as shooting nodes $\theta_j = x_{jk_{\mu_a}}$, $j=0,1,\dots,r_\mu$ with $r_\mu k_{\mu_a} = n_{\mu_a}$. With the same numerical method we integrate the IVPs

$$\begin{aligned} Y_j'(x) &= K(x,\mu)Y_j(x), \quad && \theta_j \le x < \theta_{j-1}, \\ Y_j(\theta_{j-1}) &= I, && j=1,2,\dots,r_\mu. \end{aligned} \tag{2.16}$$

With obvious notations we define $Y(\mu,t) = Y_{r_\mu}(\mu,t),\dots,Y_1(\mu,t)$.

Thus the miss-distance function for the interval $[a,x_\mu]$ is

$$F_a(\mu,t) = \tilde{\alpha}^T X^T(\mu,t) J Y(\mu,t)\gamma. \tag{2.17}$$

Referring to the problem (2.13) with μ and x_μ instead of λ and x_λ, we consider $x_n = x_\mu + nh$, $n=0,1,\dots,n_{\mu_b} = \frac{c_b - x_\mu}{h}$ and the shooting nodes $\xi_j = x_{jk_{\mu_b}}$, $j=0,1,\dots,s_\mu$, with $s_\mu k_{\mu_b} = n_{\mu_b}$. Then, upon numerical integration of the IVPs

$$\begin{aligned} R_j'(x) &= K(x,\mu)R_j(x), \quad && \xi_{j-1} < x \le \xi_j, \\ R_j(\xi_{j-1}) &= I, && j=1,2,\dots,s_\mu, \end{aligned} \tag{2.18}$$

it is possible to define $R(\mu,t) = R_{s_\mu}(\mu,t),\dots,R_1(\mu,t)$.

Finally, denoting by $S_j(x)$, $j=1,2,\dots,s_b$, the fundamental solutions associated to the problem (2.14), where λ is replaced with μ, and following the same procedure as before, we integrate

$$\begin{aligned} S_j'(x) &= K(x,\mu)S_j(x), \quad && \theta_j \le x < \theta_{j-1}, \\ S_j(\theta_{j-1}) &= I, && j=1,2,\dots,s_b, \end{aligned} \tag{2.19}$$

where $\theta_0 = b$, so that we can define $S(\mu,t) = S_{s_b}(\mu,t),\dots,S_1(\mu,t)$.

By consequence the miss-distance in the interval $[x_\mu,b]$ is

$$F_b(\mu,t) = \gamma^T R^T(\mu,t) J S(\mu,t)\tilde{\beta}. \tag{2.20}$$

Thus μ is an approximation to λ if μ is a zero both for $F_a(\mu,t)$ and $F_b(\mu,t)$.

3. The discretization method and its behavior in the singular case. We use a BVM endowed with a symmetric scheme in order to integrate the IVPs (2.3), (2.6) in the regular case and the IVPs (2.15), (2.16), (2.18), (2.19) in the singular case. In any event, the general form of these problems in each shooting interval is

$$\begin{aligned} Z'(x) &= \Gamma(Z(x),\lambda), \\ Z(x_0) &= I, \end{aligned}$$

where $\Gamma(Z(x),\lambda) = H(x,\lambda)Z(x)$ or $\Gamma(Z(x),\lambda) = K(x,\lambda)Z(x)$ depending on whether we consider the regular or singular SLP.

Thus the general form of BVM [4] with (k_1,k_2)-boundary conditions is

$$\begin{aligned}
&\sum_{i=0}^{r^*}\alpha_{i\nu}Z_i = h\sum_{i=0}^{r^*}\beta_{i\nu}\Gamma_i, \quad \nu = 1,\dots,k_1-1,\\
&\sum_{i=0}^{k^*}\alpha_i Z_{\nu+i-k_1} = h\sum_{i=0}^{k^*}\beta_i\Gamma_{\nu+i-k_1}, \quad \nu = k_1,\dots,N-k_2,\\
&\sum_{i=0}^{s^*}\alpha_{i\nu}Z_{\nu+i-s^*} = h\sum_{i=0}^{s^*}\beta_{i\nu}\Gamma_{\nu+i-s^*}, \quad \nu = N-k_2+1,\dots,N,
\end{aligned} \tag{3.1}$$

where N is the number of subintervals in the shooting interval considered, $r^*, s^* \le k^* = k_1 + k_2$, $Z_i \simeq Z(x_i)$, $\Gamma_i = \Gamma(Z_i,\lambda)$, $i = 0,1,\dots,N$.

In the case of a λ-rational SLP, if the BVM (3.1) numbers among its nodes the point x_λ, we then need to compute $K(x,\lambda)$ in the singular point. Thus, in order to avoid numerical overflow, we introduce an artificial layer δ in width with

$$0 < \delta < h \le h^* < 1 \tag{3.2}$$

and we choose as the starting point for the shooting $x_0 = x_\lambda - \delta$ or $x_0 = x_\lambda + \delta$ according to whether we integrate right to left or left to right.

The layer option partly inhibits the classical order of convergence of a BVM when applied to the SLP (1.11)–(1.13).

Actually the general solution of the differential equation $-v'' = g(x,\lambda)v$ in $[a,x_0]$ or in $[x_0,b]$ with the initial conditions $v(x_0) = 0$ and $v'(x_0) = 1$ takes the form $v(x) = Ay(x) + Cw(x)$, where

$$w(x) = -1 + \frac{q(x_\lambda)}{u'(x_\lambda)}(x - x_\lambda)\log|x - x_\lambda| + O(x - x_\lambda)$$

for $x \to x_\lambda$ and A and C depend on λ [1, Theorem 4.2].

Thus the explanation for this behavior of the BVMs is that the "nice" solution of (1.7) satisfying the conditions of Theorem 1.1 lies, in fact, in $C^\infty([a,b])$ whereas the solution $v(x)$ has singularities in the higher order derivatives inherited from the function $w(x)$.

A general discussion on this topic is in [9, section 3]. More detailed considerations show that for a class of BVMs, when applied to the SLP (1.11)–(1.13) using the simple shooting, the best accuracy is $t(h) = h^2\log h$ for a method whose classical order is 2 and $t(h) = h^2$ for a method whose classical order is greater than 2 [8, Theorem 5.2]. In the parallel shooting the starting shooting subinterval bordering on a singular point is precisely in the aforementioned conditions.

4. The errors polynomial: the regular case. In what follows we consider the miss-distance $F(\mu,t)$ as in (2.10) differentiable with respect to t as we need and the same for $U(\mu,t)$ and $V(\mu,t)$ with respect to μ.

LEMMA 4.1. *Let $q_0 = 0$, $q_j = i_1 + i_2 + \cdots + i_j$ be for some nonnegative integers $i_1, i_2, \dots, i_j$. Denote by $W_j(\mu,t)$, $j = 1,2,\dots,m$, second order matrices depending on μ and t. Let $W_j^{(s)}(\mu,t) = (\partial^s/\partial\mu^s)\,W_j(\mu,t)$ and $W(\mu,t) = W_m(\mu,t),\dots,W_1(\mu,t)$.*

Define the formal product operators

$$p_j(m) = \prod_{r=0}^{m-2} \sum_{i_{r+1}=0}^{j-q_r} \binom{j-q_r}{i_{r+1}}, \quad j \geq q_r,$$

acting on the matrices

$$P_j(W,m) = W_m^{(j-q_{m-1})}(\mu,t) \prod_{r=0}^{m-2} W_{m-1-r}^{(i_{m-1-r})}(\mu,t), \quad j \geq q_{m-1}.$$

Then

$$\begin{aligned} F^{(k)}(\mu,t) &= \left(\partial^k/\partial\mu^k\right) F(\mu,t) \\ &= \alpha^T \left\{ \sum_{i=0}^{k} \binom{k}{i} \left[p_{k-i}(m_a) P_{k-i}(U,m_a)\right]^T J p_i(m_b) P_i(V,m_b) \right\} \beta. \end{aligned} \tag{4.1}$$

Proof. Using the Newton–Leibnitz binomial expansion we get

$$F^{(k)}(\mu,t) = \alpha^T \left\{ \sum_{i=0}^{k} \binom{k}{i} \left[U^{(k-i)}(\mu,t)\right]^T J V^{(i)}(\mu,t) \right\} \beta. \tag{4.2}$$

The term $U^{(k-i)}(\mu,t)$ in this equation can be expressed by again using the Newton–Leibnitz binomial expansion and taking advantage of the associative property for matrices. That is to say, with $s = k - i$,

$$U^{(s)}(\mu,t) = \left(\partial^s/\partial\mu^s\right)\left[A_2(\mu,t)U_1(\mu,t)\right] = \sum_{i_1=0}^{s} \binom{s}{i_1} A_2^{(s-i_1)}(\mu,t) U_1^{(i_1)}(\mu,t),$$

where $A_2(\mu,t) = U_m(\mu,t), \ldots, U_2(\mu,t)$, then considering

$$A_2^{(s-i_1)}(\mu,t) = \left(\partial^{s-i_1}/\partial\mu^{s-i_1}\right)\left[A_3(\mu,t)U_2(\mu,t)\right],$$

where $A_3(\mu,t) = U_m(\mu,t), \ldots, U_3(\mu,t)$, and so on as far as the term $U^{(s)}(\mu,t) = p_s(m_a)$ $P_s(U,m_a)$ is obtained. By the same way we get $V^{(i)}(\mu,t) = p_i(m_b)P_i(V,m_b)$. ∎

LEMMA 4.2. *Denote by $U_j(x,\mu)$, $j = 1,2,\ldots,m_a$ the solutions of the IVPs* (2.3) *with μ in place of λ, namely the IVPs*

$$\begin{aligned} &U_j'(x,\mu) = H(x,\mu)U_j(x,\mu), \quad \xi_{j-1} < x \leq \xi_j, \\ &U_j(\xi_{j-1},\mu) = I, \quad j = 1,2,\ldots,m_a. \end{aligned} \tag{4.3}$$

Then the terms $U_j^{(s)}(x,\mu)$, $s = k-i$, $j = 1,2,\ldots,m_a$, in (4.2) *are the numerical solutions of the IVPs*

$$\begin{aligned} &\left(U_j^{(s)}(x,\mu)\right)' = H(x,\mu)U_j^{(s)}(x,\mu) + sH^{(1)}(x,\mu)U_j^{(s-1)}(x,\mu), \xi_{j-1} < x \leq \xi_j, \\ &U_j^{(s)}(\xi_{j-1},\mu) = 0, \quad j = 1,2,\ldots,m_a, \end{aligned} \tag{4.4}$$

where $H^{(1)}(x,\mu) = (\partial/\partial\mu)\, H(x,\mu)$.

Proof. Consider $s \geq 1$, the case $s = 0$ being trivial. Equation (4.4) is obtained by taking the s-derivative with respect to μ of the left and right-hand sides of (4.3), owing to the linearity of $H(x, \mu)$ with respect to μ, and using the Schwartz theorem. Then a numerical integration of (4.4) with a method having t as the discretization parameter gives rise to $U_j^{(s)}(x, \mu)$. □

With obvious changes this lemma can be restated in order to compute the terms $V_j^{(s)}(\mu, t)$, $s = i$, $j = 1, 2, \ldots, m_b$.

The following theorem gives an explicit form of the polynomial whose zeros are reliable evaluations of the error $\lambda - \mu$.

THEOREM 4.3. *Let μ be an estimation of λ and $t(h)$ as in* (2.9). *Let h_i, $i = 1, 2, \ldots, r$, be distinct values of the stepsize h with*

$$0 < h_1, h_2, \ldots, h_r \leq h^*$$

and suppose that

$$t(h_i) \neq t(h_j), \quad 1 \leq i \neq j \leq r.$$

Define the vectors $\gamma^{(k)} = [\gamma_1^{(k)}, \gamma_2^{(k)}, \ldots, \gamma_r^{(k)}]^T$, $k = 0, 1, \ldots, m$, whose components are

$$\gamma_i^{(k)} = F^{(k)}(\mu, t(h_i)), \quad i = 1, 2, \ldots, r,$$

$F^{(k)}(\mu, t)$ being as in (4.1). *Let T be the square matrix whose elements are given by*

$$(T)_{ij} = [t(h_i)]^{j-1}, \quad 1 \leq i, j \leq r.$$

Then

$$\lambda = \mu + \epsilon, \tag{4.5}$$

ϵ being a solution of

$$\hat{\phi}(\epsilon) + O(\epsilon^{m+1}) = 0, \tag{4.6}$$

where

$$\hat{\phi}(\epsilon) = \sum_{k=0}^{m} \hat{\phi}_k \epsilon^k. \tag{4.7}$$

The coefficients of the polynomial (4.7) *are*

$$\hat{\phi}_k = \frac{1}{k!} e^{(1)^T} T^{-1} \left(\gamma^{(k)} + O(t_\rho^r)\right), \tag{4.8}$$

$e^{(1)}$ being the first column of the identity matrix of order r and t_ρ such that

$$t_\rho = \max \left\{ |t(h_1)|, |t(h_2)|, \ldots, |t(h_r)| \right\}. \tag{4.9}$$

Proof. With $\epsilon = \lambda - \mu$ and $F^{(k)}(\mu, 0) = (\partial^k / \partial \mu^k) F(\mu, 0)$ we have

$$F(\lambda, 0) = \sum_{k=0}^{m} \frac{1}{k!} F^{(k)}(\mu, 0) \epsilon^k + O(\epsilon^{m+1}). \tag{4.10}$$

Furthermore

$$F^{(k)}(\mu, t(h)) = \sum_{i=0}^{r-1} w_i^{(k)} \left[t(h)\right]^i + O\left(\left[t(h)\right]^r\right), \tag{4.11}$$

where

$$w_i^{(k)} = \frac{1}{i!}\left[(\partial^i/\partial t^i) F^{(k)}(\mu, t)\right]_{t=0}, \quad i = 0, 1, \ldots, r-1.$$

Then writing (4.11) for $t(h) = t(h_j)$, $j = 1, 2, \ldots, r$, and defining the vector $w^{(k)} = (w_0^{(k)}, w_1^{(k)}, \ldots, w_{r-1}^{(k)})^T$ we have

$$\gamma^{(k)} = T w^{(k)} + O(t_\rho^r). \tag{4.12}$$

Noticing that $w_0^{(k)} = F^{(k)}(\mu, 0)$ and because T is a nonsingular Vandermonde matrix, from (4.12) and (4.10) (4.8) is obtained. ∎

We remark that in (4.7) a root of $\hat{\phi}(\epsilon)$ going to zero when $\mu \to \lambda$ at least exists. As a matter of fact $\hat{\phi}_0 = F(\mu, 0)$ from (4.8) and $F(\mu, 0) \to F(\lambda, 0) = 0$ by definition when $\mu \to \lambda$.

In practice, for h^* sufficiently small we are allowed to leave out the term $O(\epsilon^{m+1})$ in (4.6) and $O(t_\rho^r)$ in (4.8).

Thus (4.5) can be written as

$$\lambda \simeq \lambda^{(c)} = \mu + \epsilon, \tag{4.13}$$

where now ϵ is a zero of the polynomial

$$\phi(\epsilon) = \sum_{k=0}^{m} \phi_k \epsilon^k \tag{4.14}$$

with

$$\phi_k = \frac{1}{k!} e^{(1)^T} T^{-1} \gamma^{(k)}.$$

A very simple and useful case is $m = 2$, so that it is possible to pick out as correcting term in (4.13) the root

$$\epsilon = \frac{-\phi_1 + \frac{\phi_1}{|\phi_1|}\sqrt{\phi_1^2 - 4\phi_0\phi_2}}{2\phi_2}$$

which is, for h^* small, real and goes to zero when $\mu \to \lambda$.

A further correction procedure that is easy to use is obtained with $m = 3$ because at least one real zero of (4.14) exists. Going into details, if we set

$$\begin{aligned}
\omega_0 &= \frac{1}{27\phi_3^3}(2\phi_2^3 - 9\phi_1\phi_2\phi_3 + 27\phi_0\phi_3^2), \\
\omega_1 &= \frac{1}{\phi_3^2}\left(\phi_1\phi_3 - \frac{1}{3}\phi_2^2\right), \\
\varrho &= \sqrt{\left(\frac{\omega_0}{2}\right)^2 + \left(\frac{\omega_1}{3}\right)^3} - \frac{\omega_0}{2},
\end{aligned}$$

then the roots of (4.14) are (see, for instance, [5, 6])

$$\epsilon_{1,2,3} = \varrho^{1/3} - \frac{\omega_1}{3}\varrho^{-1/3} - \frac{1}{3}\frac{\phi_2}{\phi_3} \tag{4.15}$$

with the same determination for the cubic root. If $\left(\frac{\omega_0}{2}\right)^2 + \left(\frac{\omega_1}{3}\right)^3 \geq 0$ (the reducible case), a suitable choice in (4.13) for ϵ is the one in (4.15) with real value for $\varrho^{1/3}$.

5. The errors polynomial: the singular case. We consider the IVPs (2.11), (2.12), (2.13), (2.14), and the IVPs giving the corresponding fundamental solutions in the intervals $[a, c_a]$, $[c_a, x_\mu]$, $[x_\mu, c_b]$, $[c_b, b]$. Let $F(\mu, t)$ be the left miss-distance (2.17) or the right one (2.20) as the case may be. At present the matrix $K(x, \mu)$ depends nonlinearly on μ so that the derivatives with respect to μ look very knotty to be computed. Moreover, the singular point x_μ is μ-depending so that the initial condition in (4.4) represents a doubtful issue. By consequence Lemmas 4.1 and 4.2 are all but useless and Theorem 4.3 can be restated using a different way to compute $\gamma^{(k)}$, $k = 0, 1, \ldots, m$. To this end the following procedure based on the numerical differentiation using the method of undetermined coefficients (see, for instance, [11, chap. 6, sec. 5]) provides a simple and helpful alternative to overcome the obstacle.

Consider $\nu + 1$ distinct values of μ obtained during the same iterative rootfinding process and, without loss of generality, take them ordered by $\mu_0 < \mu_1 < \cdots < \mu_\nu$. Define the functions

$$g_j(\mu) = F(\mu, t(h_j)), \quad j = 1, 2, \ldots, r,$$

where $t(h_1), t(h_2), \ldots, t(h_r)$ are as in Theorem 4.3. Denote by $p_j(\mu)$, $j = 1, 2, \ldots, r$, the νth degree interpolation polynomials for $g_j(\mu)$ with respect to the $\nu + 1$ points $\mu_0, \mu_1, \ldots, \mu_\nu$, namely $p_j(\mu_l) = g_j(\mu_l)$, $l = 0, 1, \ldots, \nu$. We seek a formula of the form

$$p_j^{(k)}(\mu_\eta) = \sum_{l=0}^{\nu} c_l g_j(\mu_l), \quad j = 1, 2, \ldots, r, \tag{5.1}$$

where $p_j^{(k)}(\mu) = (\partial^k/\partial\mu^k)p_j(\mu)$ and $0 \leq \eta \leq \nu$.

For this purpose we choose the c_l coefficients in order that (5.1) will be the most accurate approximations to $p_j^{(k)}(\mu_\eta)$ when

$$g_j(\mu) = (\mu - \mu_\eta)^l, \quad l = 0, 1, \ldots, \nu.$$

Because

$$\left[(\partial^k/\partial\mu^k)g_j(\mu)\right]_{\mu=\mu_\eta} = k!\delta_{lk},$$

we obtain the linear system

$$M(\mu_\eta)c = b^{(k)}, \tag{5.2}$$

where $c = (c_0, c_1, \ldots, c_\nu)^T$, $b^{(k)} = k!(\delta_{0k}, \delta_{1k}, \ldots, \delta_{\nu k})^T$ and

$$(M(\mu))_{lr} = (\mu_r - \mu)^l, \quad 0 \leq l, r \leq \nu. \tag{5.3}$$

It is clear that the system (5.2) has a unique solution since (5.3) is a nonsingular Vandermonde matrix. Besides, it is possible to prove [11, chap. 6, sec. 5, Theorem 1] that

$$F^{(k)}(\mu, t(h_j)) = p_j^{(k)}(\mu) + \chi_j^{(k)}(\mu), \quad j = 1, 2, \ldots, r, \tag{5.4}$$

having set

$$\chi_j^{(k)}(\mu) = \prod_{i=0}^{\nu-k} (\mu - \tau_i) \frac{F^{(\nu+1)}(\xi, t(h_j))}{(\nu+1-k)!}, \tag{5.5}$$

when the $\nu + 1 - k$ distinct points τ_i are independent of μ and lie in the intervals (μ_i, μ_{i+k}), $i = 0, 1, \ldots, \nu - k$, and $\xi = \xi(\mu)$ is some point in the interval containing μ and τ_i.

We are now able to restate Theorem 4.3 in the following form suitably modified for the present case.

THEOREM 5.1. *Let μ be an estimation of λ and $t(h_i)$, $i = 1, 2, \ldots, r$, as in Theorem* 4.3. *Let μ_l, $l = 0, 1, \ldots, \nu$, be values of μ obtained during the same rootfinding process and ordered by $\mu_0 < \mu_1 < \cdots < \mu_\nu$. Define the matrix G whose elements are*

$$(G)_{jl} = g_{jl} = F(\mu_l, t(h_j)), \quad 0 \le l \le \nu, \quad 1 \le j \le r,$$

and let $M(\mu)$ be as in (5.3). *Denote again by $\gamma^{(k)}$ the vector whose components are $\gamma_j^{(k)} = F^{(k)}(\mu, t(h_j))$, $j = 1, 2, \ldots, r$, and let T be the Vandermonde matrix as in Theorem* 4.3.

Then

$$\lambda = \mu + \epsilon, \tag{5.6}$$

ϵ being a solution of

$$\hat{\psi}(\epsilon) + O(\epsilon^{m+1}) = 0, \tag{5.7}$$

where

$$\hat{\psi}(\epsilon) = \sum_{k=0}^{m} \hat{\psi}_k \epsilon^k. \tag{5.8}$$

The coefficients in (5.8) *are given by*

$$\hat{\psi}_k = \frac{1}{k!} e^{(1)^T} T^{-1} \big(G M^{-1}(\mu) b^{(k)} + \chi^{(k)} + O(t_\rho^r) \big), \tag{5.9}$$

where, from (5.5),

$$\chi^{(k)} = \big(\chi_1^{(k)}(\mu), \chi_2^{(k)}(\mu), \ldots, \chi_r^{(k)}(\mu) \big)^T$$

and t_ρ being as in (4.9).

Proof. The proof runs as the proof of Theorem 4.3. We have to use only for components of $\gamma^{(k)}$ (5.4) instead of (4.2) and observe that from (5.1) and (5.2) we obtain

$$p_j^{(k)}(\mu) = \big(G M^{-1}(\mu) b^{(k)} \big)_j, \quad j = 1, 2, \ldots, r. \qquad \square$$

Leaving out the terms $O(\epsilon^{m+1})$ in (5.7), $\chi^{(k)}$, and $O(t_\rho^r)$ in (5.9), (5.6) can be written as

$$\lambda \simeq \lambda^{(c)} = \mu + \epsilon,$$

where now ϵ is a zero of the polynomial

$$\psi(\epsilon) = \sum_{k=0}^{m} \psi_k \epsilon^k \tag{5.10}$$

with

$$\psi_k = \frac{1}{k!} e^{(1)^T} T^{-1}\big(GM^{-1}(\mu) b^{(k)}\big).$$

As in the regular case a simple but powerful correcting procedure is obtained with $m = 2$. If $r = 3$ and μ_0, μ_2 are such that $|\mu_0 - \mu_2|$ is sufficiently small, define $\mu_1 = \frac{1}{2}(\mu_0 + \mu_2)$ and $\sigma = \mu_0 - \mu_1 = \mu_1 - \mu_2$. Therefore it is not difficult to verify that

$$\gamma_j^{(1)} = \frac{g_{j0} - g_{j2}}{2\sigma} + O(\sigma^2), \quad j = 1, 2, 3,$$
$$\gamma_j^{(2)} = \frac{g_{j0} - 2g_{j1} + g_{j2}}{\sigma^2} + O(\sigma^2), \quad j = 1, 2, 3,$$

where $\gamma_j^{(0)}$, $\gamma_j^{(1)}$, and $\gamma_j^{(2)}$ are computed for $\mu = \mu_1$. Then the correction is given by

$$\lambda \simeq \lambda^{(c)} = \mu_1 + \epsilon,$$

where

$$\epsilon = \frac{-\psi_1 + \frac{\psi_1}{|\psi_1|}\sqrt{\psi_1^2 - 4\psi_0\psi_2}}{2\psi_2}.$$

With $m = 3$, since ψ_k have been computed in (5.10), the correction can be performed as at the end of section 4.

We remark that the procedure described in this section to compute the vectors $\gamma^{(k)}$ can be used in the regular case too. Still it is worth mentioning that the use of Lemmas 4.1 and 4.2 for the regular case seems to be theoretically more correct. As a matter of fact the vectors $\gamma^{(k)}$ depend on the solutions of the differential equations describing the analytical behavior of the derivatives $F^{(k)}(\mu, t)$ of the miss-distance itself. Consequently, if these solutions are carried out by means of the same discretization method employed for the "principal" problems (2.1) and (2.2), no additional truncation error substantially different from the "principal" one is introduced.

Nevertheless, when k is increasing, the bigger computational cost could be disguising for this distinctive feature.

6. Numerical experiments. We apply the techniques exposed in sections 4 and 5 both to a regular and λ-rational SLP.

In the following examples the used BVM is the fourth order extended trapezoidal rule of second kind (ETR_2) coupled with its boundary conditions [4, p. 164]:

$$\frac{1}{24}(-17Z_0 + 9Z_1 + 9Z_2 - Z_3) = \frac{h}{4}(\Gamma_0 + 3\Gamma_1),$$
$$\frac{1}{12}(-Z_{\nu-2} - 9Z_{\nu-1} + 9Z_\nu + Z_{\nu+1}) = \frac{h}{2}(\Gamma_{\nu-1} + \Gamma_\nu), \quad \nu = 2, \ldots, N-1,$$
$$\frac{1}{24}(Z_{N-3} - 9Z_{N-2} - 9Z_{N-1} + 17Z_N) = \frac{h}{4}(3\Gamma_{N-1} + \Gamma_N).$$

Moreover, the approximations to the eigenvalues are obtained by choosing the secant method as rootfinding process.

Example 1 (see [16, p. 134]). Consider the regular SLP

$$-y''(x) + \exp(x)y(x) = \lambda y(x), \quad 0 \le x \le \pi,$$

$$y(0) = 0, \quad y(\pi) = 0.$$

We shall denote by

- λ_k the kth eigenvalue computed with the SLEIGN code (see [17, Appendix C]) and considered as "exact";
- $\mu_k \equiv \mu_k^{(j+1)}$ the approximation to λ_k obtained by using the rootfinding process with a constant stepsize h and the stopping criterion given by
$$| \mu_k^{(j+1)} - \mu_k^{(j)} | \le 10^{-4};$$
- $\lambda_k^{(c)}$ the corrected eigenvalue derived by fixing $h_i = h/i$, $i = 1, 2, \ldots, r$ (see Theorem 4.3).

In Table 6.1 we get the results choosing $h = \pi/1000$, $r = 2, 3$, the polynomials (4.14) having degree $m = 1, 2, 3$. Moreover, we fix 10 shooting intervals and the matching point $c = \pi/2$. It is worth noting that the correction does not depend on the selected value of c. Although the used stopping criterion leads to a rough enough approximation μ_k, the correction technique is able to give a value $\lambda_k^{(c)}$ which emphasizes a remarkable improvement of accuracy. However, as the index k increases, we observe a loss of precision as a consequence of the rounding off errors in the computation of the coefficients defining the polynomials (4.14).

Example 2 (see [9, p. 375]). We consider the λ-rational SLP

$$-y''(x) = \left(\lambda + \frac{\exp(-x)}{\exp(-x_*) - \exp(-x) - \lambda}\right) y(x), \quad -5 \le x \le 5,$$

$$\exp(5)y(-5) - (\exp(-x_*) - \exp(5))y'(-5) = 0,$$

$$\exp(-5)y(5) - (\exp(-x_*) - \exp(-5))y'(5) = 0.$$

TABLE 6.1
Example 1.

		$r = 2$		$r = 3$		
k	$\lvert \lambda_k - \mu_k \rvert$	$\lvert \lambda_k - \lambda_k^{(c)} \rvert$	$\frac{\lvert\lambda_k^{(c)} - \mu_k\rvert}{\lvert\lambda_k - \mu_k\rvert}$	$\lvert \lambda_k - \lambda_k^{(c)} \rvert$	$\frac{\lvert\lambda_k^{(c)} - \mu_k\rvert}{\lvert\lambda_k - \mu_k\rvert}$	m
5	$3.14\ 10^{-5}$	$4.44\ 10^{-8}$	1.0014	$2.76\ 10^{-8}$	1.0008	1
		$4.42\ 10^{-8}$	1.0014	$2.73\ 10^{-8}$	1.0008	2
		$4.42\ 10^{-8}$	1.0014	$2.73\ 10^{-8}$	1.0008	3
10	$7.42\ 10^{-5}$	$1.82\ 10^{-8}$	1.0002	$9.27\ 10^{-9}$	1.0001	1
		$1.67\ 10^{-8}$	1.0002	$7.86\ 10^{-9}$	1.0002	2
		$1.67\ 10^{-8}$	1.0002	$7.86\ 10^{-9}$	1.0002	3
30	$1.10\ 10^{-3}$	$2.58\ 10^{-6}$	0.9976	$1.56\ 10^{-7}$	0.9998	1
		$2.28\ 10^{-6}$	0.9979	$1.50\ 10^{-7}$	1.0001	2
		$2.28\ 10^{-6}$	0.9979	$1.47\ 10^{-7}$	1.0001	3
50	$2.37\ 10^{-2}$	$5.38\ 10^{-5}$	0.9977	$3.36\ 10^{-6}$	0.9998	1
		$8.82\ 10^{-5}$	1.0037	$1.39\ 10^{-4}$	1.0058	2
		$7.84\ 10^{-5}$	1.0033	$1.29\ 10^{-4}$	1.0054	3

TABLE 6.2
Example 2: Interval $[-5, x_\lambda]$.

	$r=2$		$r=3$		
h	$\mid \lambda - \lambda^{(c)} \mid$	$\frac{\mid\lambda^{(c)}-\mu\mid}{\mid\lambda-\mu\mid}$	$\mid \lambda - \lambda^{(c)} \mid$	$\frac{\mid\lambda^{(c)}-\mu\mid}{\mid\lambda-\mu\mid}$	m
0.1	$2.39\ 10^{-9}$	0.9974	$3.84\ 10^{-10}$	1.0004	1
	$2.39\ 10^{-9}$	0.9974	$3.84\ 10^{-10}$	1.0004	2
	$2.39\ 10^{-9}$	0.9974	$3.84\ 10^{-10}$	1.0004	3
0.05	$5.70\ 10^{-11}$	1.0000	$1.04\ 10^{-11}$	1.0000	1
	$5.67\ 10^{-11}$	1.0000	$1.01\ 10^{-11}$	1.0000	2
	$5.67\ 10^{-11}$	1.0000	$1.01\ 10^{-11}$	1.0000	3
0.01	$9.61\ 10^{-13}$	1.0000	$5.85\ 10^{-13}$	1.0000	1
	$6.78\ 10^{-13}$	1.0000	$3.01\ 10^{-13}$	1.0000	2
	$6.77\ 10^{-13}$	1.0000	$3.01\ 10^{-13}$	1.0000	3

This problem has an embedded eigenvalue $\lambda = 0$ corresponding to the solution $y(x) = \exp(-x_*) - \exp(-x)$ in the interval $[-5, 5]$.

In Table 6.2 we report the results obtained by approximating the eigenvalue λ on the interval $[-5, x_\lambda]$ with $x_\lambda \equiv x_* = 0$ and introducing the artificial layer δ (see (3.2)) equal to 10^{-15}. The involved differential problems are solved fixing 10 shooting intervals and matching point $c = -5$. In Table 6.3 the interval $[x_\lambda, 5]$ has been considered choosing the same parameters as before, except for the matching point here taken as $c = 5$. In both tables we have denoted with μ and $\lambda^{(c)}$ the approximation and correction to λ, respectively. In accordance with what we have pointed out on the order of convergence in section 3, we have used $t(h) = h^2$. The approximation μ is obtained by the same stopping criterion used in the regular case.

Remark 6.1. Despite that this correction technique may be easily used for polynomials (5.10) having any degree, from the shown results we see that the approximation of the eigenvalue does not noticeably improve as the degree m of (5.10) becomes greater than two. This behavior can appear fairly surprising. Nevertheless, it is worth taking into account that the correction is rather strong even if $m = 1$ or $m = 2$. Thus we are allowed to interpret these results as a kind of numerical saturation of the algorithm as regards the attainable accuracy. Consequently, in the present case, it seems insignificant to consider $m > 2$.

TABLE 6.3
Example 2: Interval $[x_\lambda, 5]$.

	$r=2$		$r=3$		
h	$\mid \lambda - \lambda^{(c)} \mid$	$\frac{\mid\lambda^{(c)}-\mu\mid}{\mid\lambda-\mu\mid}$	$\mid \lambda - \lambda^{(c)} \mid$	$\frac{\mid\lambda^{(c)}-\mu\mid}{\mid\lambda-\mu\mid}$	m
0.1	$2.07\ 10^{-6}$	1.0086	$5.09\ 10^{-8}$	1.0002	1
	$1.98\ 10^{-6}$	1.0082	$4.35\ 10^{-8}$	0.9998	2
	$1.98\ 10^{-6}$	1.0082	$4.35\ 10^{-8}$	0.9998	3
0.05	$7.98\ 10^{-8}$	1.0009	$2.99\ 10^{-8}$	1.0003	1
	$6.77\ 10^{-8}$	1.0007	$1.79\ 10^{-8}$	1.0002	2
	$6.77\ 10^{-8}$	1.0007	$1.79\ 10^{-8}$	1.0002	3
0.01	$4.81\ 10^{-9}$	1.0001	$4.39\ 10^{-9}$	1.0001	1
	$2.62\ 10^{-9}$	1.0000	$2.19\ 10^{-9}$	1.0000	2
	$2.62\ 10^{-9}$	1.0000	$2.19\ 10^{-9}$	1.0000	3

TABLE 6.4
Example 1*: Simple shooting,* $h = \pi/500$.

		$r = 2$		$r = 3$		
k	$\mid \lambda_k - \mu_k \mid$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	m
50	$4.03\ 10^{-1}$	$1.15\ 10^{-1}$	0.7127	$1.15\ 10^{-1}$	0.7127	1
		complex	–	complex	–	2
		$8.28\ 10^{-2}$	0.7943	$8.28\ 10^{-2}$	0.7944	3

TABLE 6.5
Example 1*: Multiple shooting,* $h = \pi/500$.

		$r = 2$		$r = 3$		
k	$\mid \lambda_k - \mu_k \mid$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	m
50	$1.44\ 10^{-1}$	$2.46\ 10^{-2}$	0.8289	$1.86\ 10^{-2}$	0.8708	1
		$3.72\ 10^{-3}$	0.9742	$5.09\ 10^{-3}$	1.0352	2
		$1.27\ 10^{-2}$	0.9115	$5.71\ 10^{-3}$	0.9604	3

TABLE 6.6
Example 1*: Simple shooting,* $h = \pi/600$.

		$r = 2$		$r = 3$		
k	$\mid \lambda_k - \mu_k \mid$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	m
50	$1.94\ 10^{-1}$	$3.16\ 10^{-2}$	0.8371	$3.16\ 10^{-2}$	0.8372	1
		$1.63\ 10^{-2}$	1.0842	$1.64\ 10^{-2}$	1.0844	2
		$1.24\ 10^{-2}$	0.9358	$1.24\ 10^{-2}$	0.9359	3

TABLE 6.7
Example 1*: Multiple shooting,* $h = \pi/600$.

		$r = 2$		$r = 3$		
k	$\mid \lambda_k - \mu_k \mid$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	$\mid \lambda_k - \lambda_k^{(c)} \mid$	$\frac{\mid \lambda_k^{(c)} - \mu_k \mid}{\mid \lambda_k - \mu_k \mid}$	m
50	$6.79\ 10^{-2}$	$7.08\ 10^{-3}$	0.8956	$3.98\ 10^{-3}$	0.9413	1
		$2.70\ 10^{-3}$	0.9601	$9.04\ 10^{-4}$	1.0133	2
		$3.60\ 10^{-3}$	0.9469	$1.47\ 10^{-4}$	0.9978	3

Finally, in order to stress that the multiple shooting works better than the simple one, we have quoted in Table 6.4 the results obtained with Example 1 when the simple shooting is used, and in Table 6.5 the same example in the case of multiple shooting. In both cases we have selected $h = \frac{\pi}{500}$ and λ_{50} the eigenvalue to be approximated. In the multiple shooting we have chosen 50 shooting intervals (and 10 steps in each interval) and $c = \pi$ as matching point. We see that in the simple shooting the second degree polynomial has complex zeroes. Thus, we are compelled to gain the advantage from the use of first or third degree polynomials where a real zero always exists. Tables 6.6 and 6.7 are the same as the previous ones but with $h = \frac{\pi}{600}$ and 60 shooting intervals in the multiple shooting.

Acknowledgment. The authors would like to thank the anonymous referee very much for his valuable comments and suggestions enabling an improvement of the technical correctness of the paper.

REFERENCES

[1] V. Adamyan, H. Langer, and M. Langer, *A spectral theory for a λ-rational Sturm-Liouville problem*, J. Differential Equations, 171 (2001), pp. 315–345.

[2] R. S. Anderssen and F. R. de Hoog, *On the correction of finite difference eigenvalue approximations for Sturm-Liouville problems with general boundary conditions*, BIT, 24 (1984), pp. 401–412.

[3] A. L. Andrew, *Asymptotic correction of computed eigenvalues of differential equations*, Ann. Numer. Math., 1 (1994), pp. 41–51.

[4] L. Brugnano and D. Trigiante, *Solving Differential Problems by Multistep Initial and Boundary Value Methods*, Gordon and Breach, Amsterdam, 1998.

[5] H. Cardani, *Artis Magnae, sive de Regulis Algebraicis, Liber unus*, Norimbergae, 1545.

[6] L. Euler, *De formis radicum aequationum cuiusque ordinis coniectatio*, in Comm. Acad. Sci. Petrop., 6 (1732–33), pp. 217–231, reprinted in Opera Omnia Ser. 1, 6 (1738), 1–19.

[7] P. Ghelardoni, G. Gheri, and M. Marletta, *Spectral corrections for Sturm-Liouville problems*, J. Comput. Appl. Math., 132 (2001), pp. 443–459.

[8] P. Ghelardoni, G. Gheri, and M. Marletta, *Numerical solution of a λ-rational Sturm-Liouville problem*, IMA J. Numer. Anal., 23 (2003), pp. 29–53.

[9] P. Ghelardoni, G. Gheri, and M. Marletta, *A polynomial approach to the spectral corrections for Sturm-Liouville problems*, J. Comput. Appl. Math., 185 (2006), pp. 360–376.

[10] W. B. Gragg, *On extrapolation algorithms for ordinary initial value problems*, J. Soc. Indust. Appl. Math., Ser. B, Numer. Anal., 2 (1965), pp. 384–403.

[11] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, Inc., New York, 1966.

[12] L. Gr. Ixaru, H. De Meyer, and G. Vanden Berghe, *SLCPM*12 - *A program for solving regular Sturm-Liouville problems*, Comput. Phys. Comm., 118 (1999), pp. 259–277.

[13] H. Langer, R. Mennicken, and M. Möller, *A second order differential operator depending nonlinearly on the eigenvalue parameter*, Oper. Theory Adv. Appl., Birkhäuser, Basel, 48 (1990), pp. 319–332.

[14] M. Langer, *Eigenvalues of a λ-rational Sturm-Liouville problem*, Math. Nachr., 210 (2000), pp. 163–176.

[15] A. Lifschitz, *Magnetohydrodynamics and Spectral Theory*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[16] J. W. Paine, F. R. de Hoog, and R. S. Anderssen, *On the correction of finite difference eigenvalue approximations for Sturm-Liouville problems*, Computing, 26 (1981), pp. 123–139.

[17] J. D. Pryce, *Numerical Solution of Sturm-Liouville Problems*, Clarendon Press, Oxford University Press, New York, 1993.

[18] G. Vanden Berghe and H. De Meyer, *A finite-element estimate with trigonometric hat functions for Sturm-Liouville eigenvalues*, J. Comput. Appl. Math., 53 (1994), pp. 389–396.

[19] A. Zettl, *Sturm-Liouville Problems*, in Spectral Theory and Computational Methods of Sturm-Liouville Problems, D. Hinton and P. H. Schefer, eds., Dekker, New York, 1997.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2245–2269

CFL CONDITION AND BOUNDARY CONDITIONS FOR DGM APPROXIMATION OF CONVECTION-DIFFUSION*

JEAN-BAPTISTE APOUNG KAMGA† AND BRUNO DESPRÉS‡

Abstract. We propose a general method for the design of discontinuous Galerkin methods (DGMs) for nonstationary linear equations. The method is based on a particular splitting of the bilinear forms that appear in the weak DGM. We prove that an appropriate time splitting gives a stable linear explicit scheme whatever the order of the polynomial approximation. Numerical results are presented.

Key words. discontinuous Galerkin method, advection diffusion, stability, CFL condition

AMS subject classifications. 65M12, 65M60

DOI. 10.1137/050633159

1. Introduction.

The convection-diffusion equation is widely used in real-life problems such as contaminant transport in porous media [1, 8, 26]. Due to the geological structure of the problem, the equation is convection-dominant in random distributed parts of the media. This makes its numerical resolution difficult. While difference schemes suffer from the complex geometry of the domain, ordinary finite element methods suffer from their lack of local conservativity [28], and finite volume methods suffer from their low order of accuracy (due to low order polynomial approximation). The discontinuous Galerkin method (DGM or DG), introduced in 1973 by Reed and Hill [32], in its development [24] found here a good field of application. In a computational aspect, the DGM can be used efficiently to handle the advection part in an operator splitting technique scheme [27]. But this strategy may break apart at boundary conditions of mixed type, where it is difficult to determine whether the boundary condition is more in the advection step or in the diffusion step. For real-life problems [8], the Dirichlet part of the boundary can also be split into inflow and outflow parts. This boundary condition can astutely be distributed in between the advection terms and diffusion terms [5, 6]. In a mathematical aspect, it is more convenient to have a unique bilinear form even if the splitting technique is used [20, 37, 33]. This leads to an ordinary differential equation, a different approach is [39]. Assuming for example that the DGM is used only in space to exploit the block diagonal mass matrix obtained, most time discretizations are explicit and therefore require a CFL condition.

In the one-dimensional case, using Von Neumann analysis, Chavent & Cockburn [11] proved that explicit linear Euler time integration of the DGM is unconditionally unstable if the ratio $\frac{\Delta t}{\Delta x}$ is held constant. To overcome this striking difficulty and still keep high order accuracy, Cockburn and Shu [19, 20, 21, 22] introduced the RKDG (Runge–Kutta discontinuous Galerkin method). It uses at each time step an

*Received by the editors June 6, 2005; accepted for publication (in revised form) April 25, 2006; published electronically November 24, 2006.

http://www.siam.org/journals/sinum/44-6/63315.html

†University Paris VI, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France (apoung@ann.jussieu.fr). This author's research was supported by BQR of the University of Paris VI and GDR MOMAS under the guidance of O. Pironneau.

‡CEA 31/33 rue de la Fédération, 75752 Paris Cedex 15, France, and University Paris VI, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France (despres@ann.jussieu.fr).

explicit Euler scheme, stabilized by a particular slope limiter, which makes the scheme nonlinear. Due to this nonlinearity, proof of convergence of the fully discrete explicit DGM is not possible except perhaps in very rare and special cases. We refer the reader to Cockburn [18] for a presentation of the convergence theory for the DGM. Despite this lack of theory, numerical experiments show the convergence. For example, in the one-dimensional case for advection, the convergence is observed if the CFL condition is of the form $\frac{1}{2k+1}$ for polynomials of order k [18]. To the best of our knowledge, the analysis of the fully discrete explicit DGM scheme remains an open problem.

In this work we propose a way to solve this problem. We propose an abstract functional formalism. Within this formalism, it is easy to design explicit (only local-in-the-cell) computations, which are linear and stable under CFL DGMs. Then we apply this method to our model problem, which is advection diffusion in two dimensions,

$$\partial_t c + \mathbf{u}.\nabla c - \nabla.(K\nabla c) = 0, \quad x \in \mathbf{R}^2,\ t > 0. \tag{1.1}$$

The diffusion coefficient is nonnegative $K \geq 0$, and the velocity is divergence-free $\nabla.\mathbf{u} = 0$. Boundary conditions are general and are specified in the core of the paper. Due to the stability (under the CFL condition) and linearity of our explicit DGM scheme, we are able to prove the convergence by a standard method. For example, we obtain the estimate of convergence in two dimensions for the advection case ($K = 0$),

$$\|c(n\Delta t) - c_h^n\|_{L^2} \leq C_1 \Delta t^2 + C_2 h^p + E.$$

E is an error term due the discretization of the initial condition and can be taken as small as desired. This estimate is true for the second order in time discretization. The order in space is p, which is the degree of the polynomial basis. Since the optimal order in space is $p + 1/2$, we think this loss of $1/2$ is an artifact of the analysis, which could be corrected with a more sophisticated technique [13, 14, 15]. To our knowledge, such an estimate is new and was not possible to get for previous fully discrete explicit DGM schemes.

At the theoretical level the key idea is to reformulate (1.1) as a weak problem

$$\left(\frac{\partial}{\partial t}U, V\right) + \mathcal{A}_0(U,V) + \mathcal{A}_1(U,V) - \mathcal{A}_2(U,V) = 0 \ \ \forall V \in \mathcal{V}, \tag{1.2}$$

where U is the solution, V is a test function, $(.,.)$ is the standard L^2 scalar product, and $\mathcal{A}_{0,1,2}$ are some bilinear forms defined later in this paper. The space is $\mathcal{V} \subset \sum_k L^2(\Omega_k)$, where (Ω_k) is a partition of the plane, i.e., is the mesh. Among other properties, the local bilinear forms $\mathcal{A}_0(U,V)$, $\mathcal{A}_1(U,V)$, and $\mathcal{A}_2(U,V)$ satisfy

$$\mathcal{A}_0(U,U) + \mathcal{A}_1(U,U) - \mathcal{A}_2(U,U) \geq 0. \tag{1.3}$$

The first order time discretization of (1.3) is as follows: Find $U_h^n, U_h^{n+1} \in \mathcal{V}_h$ such that for all test functions $V_h \in \mathcal{V}_h$,

$$\left(\frac{U_h^{n+1} - U_h^n}{\Delta t}, V_h\right) + \mathcal{A}_0(U_h^{n+1}, V_h) + \mathcal{A}_1(U_h^n, V_h) - \mathcal{A}_2(U_h^n, V_h) = 0. \tag{1.4}$$

When applied to (1.1), the bilinear form $\mathcal{A}_0$ is local-in-the-cell, and this is why the scheme is explicit. The main stability property that we prove is the inequality

$$||U_h^{n+1}||_{L^2(\mathbf{R}^2)} \leq ||U_h^n||_{L^2(\mathbf{R}^2)} \quad \forall n \ \in \mathbb{N}, \tag{1.5}$$

which is true under a CFL condition that is studied in detail. It guarantees stability whatever the order of the polynomial approximation. Since $\mathcal{A}_0$ is in practice a local-in-the-cell bilinear form, the scheme is explicit at the price of the resolution of a local-in-the-cell linear system. At the implementation level, it does not cost more than inverting the local mass matrix. We also study the second order discretization in time,

$$\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},V_h\right)+\frac{2}{3}\mathcal{A}_0(U_h^{n+1},V_h) + \frac{2}{3}\mathcal{A}_1(2U_h^n-U_h^{n-1},V_h)-\frac{2}{3}\mathcal{A}_2(2U_h^n-U_h^{n-1},V_h)=0. \tag{1.6}$$

The CFL condition is twice as stringent for (1.6) than for (1.4). It is possible to define all the parameters of the method in order to optimize the CFL condition. We will apply this method for our convection-diffusion problem.

The paper is organized as follows. In section 2 we consider a general setting. We present the properties which the bilinear forms should satisfy in this framework. Assuming these properties, we discuss some time schemes and derive the abstract CFL condition that guarantees their stability. In section 3 we address the convection-diffusion equation within the discontinuous Galerkin approximation and show how to cast the bilinear form to fit within the abstract formalism. We show how to introduce commonly used boundary conditions. In section 4 we analyze the abstract CFL condition in the case of a uniform grid and give values to all constants. We give the bilinear forms in particular cases of pure advection and pure diffusion. We conclude that the totally discrete schemes introduced for the convection-diffusion equation make up a continuous interpolation between the scheme for pure advection and the scheme for pure diffusion. In section 5 we analyze the convergence of the second order schemes in the case of the pure advection equation. Finally, in section 6 we present numerical results for advection and diffusion and compare them with other DGMs.

2. The abstract discontinuous Galerkin formalism. We first consider an abstract formalism in a more general setting and derive some time discretization, which will be stable under an abstract CFL condition.

2.1. Abstract formalism. Let us define the spaces

$$\mathcal{V}\subset\mathcal{H}. \tag{2.1}$$

$\mathcal{H}$ is endowed with a scalar product, namely $(.,.)$. In practice $\mathcal{H}=L^2(\Omega)$.

DEFINITION 2.1. *A sequence $(U^p)_p\in\mathcal{V}$ will be said to be L^2 stable if there exists a constant $C\in\mathbb{R}$ such that $(U^p,U^p)\le C$ for all $p\in\mathbb{N}$.*

Let $\mathcal{A}_i, i=0,1,2$, be three bilinear forms on $\mathcal{V}$ satisfying the following properties:

(2.2)
$$\begin{cases}\mathcal{A}_1 \text{ is symmetric nonnegative.}\\ \text{There exist a bilinear form } \mathcal{A}_3 \text{ also defined on } \mathcal{V} \text{ such that}\\ \mathcal{A}_0(U,U)\ge\frac{1}{2}\left(-\mathcal{A}_1(U,U)+\mathcal{A}_3(U,U)\right) \text{ and } \mathcal{A}_2(U,V)\le\dfrac{1}{2}\left(\mathcal{A}_1(U,U)+\mathcal{A}_3(V,V)\right).\end{cases}$$

A consequence of (2.2) is

$$\mathcal{A}_0(U,U)+\mathcal{A}_1(U,U)-\mathcal{A}_2(U,U)\ge 0\quad\forall U\in\mathcal{V}.$$

We now consider the problem (2.3):

$$\text{(2.3)}\qquad \begin{cases} \text{Given} \quad U_0 \in \mathcal{V}, \\ \text{find} \quad U \in C^1(0,T;\mathcal{V}) \quad \text{such that} \quad \forall V \in \mathcal{V}, \\ \left(\dfrac{\partial}{\partial t}U, V\right) + \mathcal{A}_0(U,V) + \mathcal{A}_1(U,V) - \mathcal{A}_2(U,V) = 0, \\ U = U_0 \quad \text{at} \quad t = 0. \end{cases}$$

In what follows we will assume that it has a unique solution.

LEMMA 2.2. *Assume that the bilinear forms $\mathcal{A}_i, i = 0,1,2$, satisfy* (2.2). *Then the solution to* (2.3) *is L^2 stable.*

Proof. Choosing $V = U$ and using the property of (2.2) one gets directly that $d_t\left[\frac{1}{2}(U,U)(t)\right] \leq 0$. Therefore the energy $t \mapsto (U,U)(t)$ decreases. ∎

2.2. Time and space discretizations and abstract CFL conditions. Let $\mathcal{V}_h \subset \mathcal{V}$ be a finite-dimensional vectorial subspace of $\mathcal{V}$. The unknown at time step n is $U_h^n \in \mathcal{V}_h$. The test function is denoted by $V_h^n \in \mathcal{V}_h$. Under assumptions (2.2) on bilinear forms $\mathcal{A}_i, i = 0,1,2$, we can now derive some fully discrete schemes, which are stable under abstract CFL conditions.

2.2.1. First order scheme. The first order scheme reads

$$\text{(2.4)}\qquad \left(\frac{U_h^{n+1} - U_h^n}{\Delta t}, V_h\right) + \mathcal{A}_0(U_h^{n+1}, V_h) + \mathcal{A}_1(U_h^n, V_h) - \mathcal{A}_2(U_h^n, V_h) = 0 \quad \forall V_h.$$

We have the following result.

THEOREM 2.3. *Assuming that the bilinear forms $\mathcal{A}_i, i = 0,1,2$, satisfy the properties* (2.2), *we assume that the time step satisfies the abstract CFL requirement*

$$\text{(2.5)}\qquad \Delta t \mathcal{A}_1(U_h, U_h) \leq (U_h, U_h) \quad \forall U_h \in \mathcal{V}_h.$$

Then scheme (2.4) *is L^2 stable and*

$$\text{(2.6)}\qquad \left(U_h^{n+1}, U_h^{n+1}\right) \leq \left(U_h^n, U_h^n\right).$$

$\Delta t > 0$ exists because the dimension of $\mathcal{V}_h$ is finite.

Proof. The proof explicitly uses the inequalities of (2.2). The scalar product of (2.4) with U_h^{n+1} gives

$$\begin{aligned} &\left(U_h^{n+1}, U_h^{n+1}\right) \\ &= \left(U_h^n, U_h^{n+1}\right) - \Delta t \mathcal{A}_0(U_h^{n+1}, U_h^{n+1}) - \Delta t \mathcal{A}_1(U_h^n, U_h^{n+1}) + \Delta t \mathcal{A}_2(U_h^n, U_h^{n+1}) \\ &\leq \left(U_h^n, U_h^{n+1}\right) - \Delta t \mathcal{A}_0(U_h^{n+1}, U_h^{n+1}) - \Delta t \mathcal{A}_1(U_h^n, U_h^{n+1}) \\ &\quad + \frac{\Delta t}{2}\left(\mathcal{A}_1(U_h^n, U_h^n) + \mathcal{A}_3(U_h^{n+1}, U_h^{n+1})\right) \\ &\leq \left(U_h^n, U_h^{n+1}\right) + \frac{\Delta t}{2}\left(\mathcal{A}_1(U_h^n, U_h^n) - 2\mathcal{A}_1(U_h^n, U_h^{n+1}) + \mathcal{A}_1(U_h^{n+1}, U_h^{n+1})\right). \end{aligned}$$

Using the symmetry of bilinear form $\mathcal{A}_1$ and the scalar product, we rewrite the previous inequality as

$$\begin{aligned} &\left(U_h^{n+1}, U_h^{n+1}\right) \leq \left(U_h^{n+1}, U_h^{n+1}\right) \\ &- \left(\left(U_h^{n+1} - U_h^n, U_h^{n+1} - U_h^n\right) - \Delta t \mathcal{A}_1\left(U_h^{n+1} - U_h^n, U_h^{n+1} - U_h^n\right)\right). \end{aligned}$$

Assuming the abstract CFL-like condition (2.5), the result is proved. ∎

2.2.2. Second order scheme. Extending to second order time discretization the abstract DGM already mentioned is not easy. After numerous attempts, we focused on the following approach, which is based on the theory of A-stable time integration for stiff equations; see [25]. First, we begin with the retrograde second order time integration,

$$\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},V_h\right)+\frac{2}{3}\left(\mathcal{A}_0+\mathcal{A}_1-\mathcal{A}_2\right)(U_h^{n+1},V_h)=0\quad\forall V_h. \tag{2.7}$$

Its stability can be proved, by taking $V_h=U_h^{n+1}$ in (2.7). The scheme is fully implicit in the sense that it requires the inversion of a global linear system to get the new value. Let us now define a semi-implicit second order time scheme. The idea is to get rid of the cell-to-cell coupling that appears in (2.7). For this we use the relation $U((n+1)\Delta t)=2U(n\Delta t)-U((n-1)\Delta t)+O(\Delta t^2)$, which is true provided that U is smooth. Then we eliminate some occurrences of U_h^{n+1} in (2.7) using transformation $U_h^{n+1}\leftarrow 2U_h^n-U_h^{n-1}$. It gives the scheme

$$\begin{aligned}\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},V_h\right)+\frac{2}{3}\mathcal{A}_0(U_h^{n+1},V_h)\\ +\frac{2}{3}\mathcal{A}_1(2U_h^n-U_h^{n-1},V_h)-\frac{2}{3}\mathcal{A}_2(2U_h^n-U_h^{n-1},V_h)=0\quad\forall V_h.\end{aligned} \tag{2.8}$$

We will see that in practice, $\mathcal{A}_0$ is of local-in-the-cell bilinear form. In this case, scheme (2.8) is only locally implicit, and we need only inverse local linear systems to get the new solution. Hence scheme (2.8) is in practice an explicit one.

THEOREM 2.4. *Assume the bilinear forms $\mathcal{A}_i, i=0,1,2$, satisfy the properties* (2.2)*, and assume the time step satisfies the abstract CFL requirement*

$$2\Delta t\mathcal{A}_1(U_h,U_h)\le(U_h,U_h)\quad\forall U_h\in\mathcal{V}_h. \tag{2.9}$$

Then scheme (2.8) *is L^2 stable and*

$$\begin{aligned}\left(U_h^{n+1},U_h^{n+1}\right)+\left(2U_h^{n+1}-U_h^n,2U_h^{n+1}-U_h^n\right)\\ \le(U_h^n,U_h^n)+\left(2U_h^n-U_h^{n-1},2U_h^n-U_h^{n-1}\right).\end{aligned} \tag{2.10}$$

Proof. Let us take $V_h=U_h^{n+1}$ in (2.8). We get

$$\begin{aligned}\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},U_h^{n+1}\right)+\frac{2}{3}\mathcal{A}_0(U_h^{n+1},U_h^{n+1})\\ +\frac{2}{3}\mathcal{A}_1(2U_h^n-U_h^{n-1},U_h^{n+1})-\frac{2}{3}\mathcal{A}_2(2U_h^n-U_h^{n-1},U_h^{n+1})=0.\end{aligned}$$

We can give a lower bound to $\mathcal{A}_0(U_h^{n+1},U_h^{n+1})$ and $-\mathcal{A}_2(2U_h^n-U_h^{n-1},U_h^{n+1})$ using (2.2). Therefore

$$\begin{aligned}\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},U_h^{n+1}\right)+\frac{1}{3}(\mathcal{A}_3-\mathcal{A}_1)(U_h^{n+1},U_h^{n+1})\\ +\frac{2}{3}\mathcal{A}_1(2U_h^n-U_h^{n-1},U_h^{n+1})-\frac{1}{3}\mathcal{A}_1(2U_h^n-U_h^{n-1},2U_h^n-U_h^{n-1})-\frac{1}{3}\mathcal{A}_3(U_h^{n+1},U_h^{n+1})\le 0,\end{aligned}$$

that is,

$$\frac{1}{3}\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},U_h^{n+1}\right) - \frac{1}{3}\mathcal{A}_1(U_h^{n+1}-2U_h^n+U_h^{n-1},U_h^{n+1}-2U_h^n+U_h^{n-1})\le 0.$$

Let us define the energy

$$E(n+1)=\left(U_h^{n+1},U_h^{n+1}\right)+\left(2U_h^{n+1}-U_h^n,2U_h^{n+1}-U_h^n\right).$$

One has the equality

$$E(n+1)-E(n)=6\left(\frac{3U_h^{n+1}-4U_h^n+U_h^{n-1}}{\Delta t},U_h^{n+1}\right) - \left(U_h^{n+1}-2U_h^n+U_h^{n-1},U_h^{n+1}-2U_h^n+U_h^{n-1}\right).$$

Plugging in the previous inequality, we obtain

$$E(n+1)\le E(n)-\left(U_h^{n+1}-2U_h^n+U_h^{n-1},U_h^{n+1}-2U_h^n+U_h^{n-1}\right) + 2\Delta t\mathcal{A}_1(U_h^{n+1}-2U_h^n+U_h^{n-1},U_h^{n+1}-2U_h^n+U_h^{n-1}).$$

Under the abstract CFL condition (2.9), the result is proved. ∎

2.2.3. Implicit scheme. The implicit scheme is

$$\left(\frac{U_h^{n+1}-U_h^n}{\Delta t},V_h\right)+\mathcal{A}_0(U_h^{n+1},V_h)+\mathcal{A}_1(U_h^{n+1},V_h)-\mathcal{A}_2(U_h^{n+1},V_h)=0. \tag{2.11}$$

LEMMA 2.5. *The implicit scheme* (2.11) *is L^2 stable unconditionally.*

Proof. The proof is left to the reader. ∎

2.3. Optimization of numerical parameters. It is well known that the DGM applied to convection-diffusion needs the definition of some arbitrary numerical parameters in order to completely define the bilinear forms at interfaces. We refer to [30, 12], where the dependence between the convergence of the DGM for stationary problems and the numerical parameters is analyzed. In what follows, we analyze the influence of the numerical parameters on the CFL condition (for nonstationary problems, of course). An open problem is to show that the parameter which is optimal with respect to the CFL condition is also optimal for convergence.

By inspection of the bilinear forms defined in the following section for convection-diffusion, it is enough to consider the abstract problem

$$\left(\frac{\partial}{\partial t}U,V\right)+\mathcal{A}_0(U,V)+\mathcal{A}_1^\alpha(U,V)-\mathcal{A}_2^\alpha(U,V)=0\ \ \forall V\in\mathcal{V}. \tag{2.12}$$

The bilinear forms $\mathcal{A}_0,\mathcal{A}_1^\alpha,\mathcal{A}_2^\alpha$ satisfy (2.2). The dependence to the arbitrary parameters is represented by α. The CFL condition takes the form

$$\left(\max_{U_h\in\mathcal{V}_h,\,U_h\neq 0}\frac{\mathcal{A}_1^\alpha(U_h,U_h)}{(U_h,U_h)}\right)\Delta t\le C, \tag{2.13}$$

where $C = 1$ for the first order scheme (2.5) and $C = \frac{1}{2}$ for the second order scheme (2.9). So the best α, denoted as $\alpha_{\rm opt}$, is the one that minimizes the constant in this inequality. We obtain the min-max problem for $\alpha_{\rm opt}$,

$$\left(\max_{U_h \in \mathcal{V}_h,\, U_h \neq 0} \frac{\mathcal{A}_1^{\alpha_{\rm opt}}(U_h, U_h)}{(U_h, U_h)} \right) \leq \left(\max_{U_h \in \mathcal{V}_h,\, U_h \neq 0} \frac{\mathcal{A}_1^{\alpha}(U_h, U_h)}{(U_h, U_h)} \right) \quad \forall \alpha.$$

We will apply this method in order to define optimized coefficients for DGM discretization for convection-diffusion in section 3.

3. Advection-diffusion with discontinuous coefficients and boundary conditions. In what follows, we describe the introduction of mixed-type boundary conditions in an advection-diffusion problem. We show that physically correct boundary conditions fit into the framework. So the stability of the scheme is guaranteed for all boundary conditions described below. Let us recall the model equation

$$\partial_t c + \mathbf{u}.\nabla c - \nabla.(K \nabla c) = 0, \quad x \in \Omega \subset \mathbf{R}^2,\ t > 0. \tag{3.1}$$

Ω is a bounded smooth open set of $\mathbf{R}^2$.

3.1. Abstract discontinuous Galerkin formalism of problem (3.1). We are now going to show how to cast the discontinuous Galerkin formulation of problem (3.1) so that the bilinear forms fit with properties (2.2).

3.1.1. Notation. We begin with some notation. Let (Ω_k) be a mesh of the plane. The cells Ω_k do not overlap. They cover the plane. The boundary of cell Ω_k is $\partial\Omega_k$. The intersection of the boundary of cell Ω_j and cell Ω_k is referred to as $\Sigma_{jk} = \Sigma_{kj}$. The outgoing normal from Ω_k is $\mathbf{n}_k$.

The velocity field $\mathbf{u}$ is not necessarily constant but is divergence-free. Therefore the degrees of freedom of $\mathbf{u}$ are naturally described in terms of its fluxes $(\mathbf{u}_{kj}, \mathbf{n}_k)$ across Σ_{jk}. The diffusion coefficient is assumed to be positive and lower bounded, but not necessarily constant. Let K_k denote the value of the diffusion coefficient in cell Ω_k. For simplicity, K_k is considered constant in the cell, but there is no real issue if it is not, except at the implementation level. We will describe the boundary conditions later on. If necessary we will assumed that the outgoing unit normal is split into two parts

$$\begin{cases} \text{if } (\mathbf{u}, \mathbf{n}_k) \geq 0, & \text{then } \mathbf{n}_k^+ = \mathbf{n}_k & \text{and } \mathbf{n}_k^- = 0, \\ \text{if } (\mathbf{u}, \mathbf{n}_k) < 0, & \text{then } \mathbf{n}_k^+ = 0 & \text{and } \mathbf{n}_k^- = \mathbf{n}_k. \end{cases} \tag{3.2}$$

Let us define the spaces

$$\mathcal{V} = \oplus_k H^2(\Omega_k) \subset \mathcal{H} = \oplus_k L^2(\Omega_k). \tag{3.3}$$

$\mathcal{H}$ is endowed with a scalar product $(U, V) = \sum_k \int_{\Omega_k} u_k(x) v_k(x) dx$.

3.1.2. Construction of the bilinear forms. Next we assume that c is smooth. Let us define $U = (u_k)$ with $u_k = c_{|\Omega_k}$. The test function is $V = (v_k)$. Let us define the local bilinear form

$$\mathcal{A}_0(U, V) = \sum_k \int_{\Omega_k} \left(-u_k(t,x) \mathbf{u}.\nabla v_k(x) + u_k \nabla.(K_k \nabla v_k) + 2 K_k \nabla u_k.\nabla v_k \right) dx. \tag{3.4}$$

We also need to define $\mathcal{A}_1$ and $\mathcal{A}_2$. So let us compute

$$(\partial_t U, V) + \mathcal{A}_0(U,V) = \sum_k \int_{\Omega_k} \left(-\mathbf{u}.\nabla u_k + \nabla.(K_k \nabla u_k)\right) v_k$$
$$+ \sum_k \int_{\Omega_k} \left(-u_k(t,x)\mathbf{u}.\nabla v_k(x) + u_k \nabla.(K_k \nabla v_k) + 2K_k \nabla u_k.\nabla v_k\right) dx$$
$$= \sum_k \int_{\partial\Omega_k} \left(-u_k v_k(\mathbf{u}_{kj},\mathbf{n}_k) + u_k K_k \frac{\partial}{\partial n_k} v_k + v_k K_k \frac{\partial}{\partial n_k} u_k.\right) d\sigma = \text{R.H.S.}$$

Next we need to transform the right-hand side (R.H.S.) in order to be able to define $\mathcal{A}_1$ and $\mathcal{A}_2$. For this task we define

$$\begin{cases} w_k^+ = K_k \frac{\partial}{\partial n_k} u_k - \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) u_k + \alpha_{jk} u_k, \\ w_k^- = -K_k \frac{\partial}{\partial n_k} u_k + \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) u_k + \alpha_{jk} u_k \end{cases}$$

and

$$\begin{cases} z_k^+ = K_k \frac{\partial}{\partial n_k} v_k - \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) v_k + \alpha_{jk} v_k, \\ z_k^- = -K_k \frac{\partial}{\partial n_k} v_k + \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) v_k + \alpha_{jk} v_k. \end{cases}$$

The value of the positive parameter $\alpha_{jk} = \alpha_{kj}$ will be specified later on. Then the R.H.S. is also

$$\text{R.H.S.} = \sum_k \int_{\partial\Omega_k} \left[u_k \left(K_k \frac{\partial}{\partial n_k} v_k - \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) v_k \right) + v_k \left(K_k \frac{\partial}{\partial n_k} u_k - \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) u_k \right) \right] d\sigma,$$
$$\text{R.H.S.} = \sum_k \int_{\partial\Omega_k} \frac{1}{2\alpha_{jk}} (w_k^+ z_k^+ - w_k^- z_k^-) d\sigma.$$

The nonnegative symmetric bilinear form is given by the $w^- z^-$ part of the integral. Therefore we define

$$\mathcal{A}_1(U,V) = \sum_k \int_{\partial\Omega_k} \frac{1}{2\alpha_{jk}} \left(-K_k \frac{\partial}{\partial n_k} u_k + \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) u_k + \alpha_{jk} u_k \right)$$
$$\times \left(-K_k \frac{\partial}{\partial n_k} v_k + \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) v_k + \alpha_{jk} v_k \right) d\sigma \tag{3.5}$$

so that we now have the relation

$$(\partial_t U, V) + \mathcal{A}_0(U,V) + \mathcal{A}_1(U,V) - \sum_k \int_{\partial\Omega_k} \frac{1}{2\alpha_{jk}} w_k^+ z_k^+ d\sigma = 0. \tag{3.6}$$

It is the place into which boundary conditions must be plugged. Let us start with some notation. The boundary between two cells Ω_k and Ω_j is still referred to as Σ_{jk}. The exterior boundary of cell Ω_k is Γ_k,

$$\Gamma_k = \partial\Omega_k \cap \partial\Omega, \quad \partial\Omega_k = (\cup_j \Sigma_{jk}) \cup \Gamma_k. \tag{3.7}$$

To transform the residual in (3.6) we use the continuity equation

$$w_k^+ = w_j^- \text{ on } \Sigma_{jk} \tag{3.8}$$
$$\iff K_k \frac{\partial}{\partial n_k} u_k - \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_k) u_k + \alpha_{jk} u_k = -K_k \frac{\partial}{\partial n_j} u_j + \frac{1}{2}(\mathbf{u}_{kj},\mathbf{n}_j) u_j + \alpha_{jk} u_j.$$

For mathematical convenience we consider that all boundary conditions may be rewritten as

$$w_k^+ = R_k^\alpha w_k^- \text{ on } \Gamma_k, \tag{3.9}$$

where $R_k^\alpha \in \mathbf{R}$ characterizes the boundary condition. This coefficient R_k^α is very similar to a reflexion coefficient in time-harmonic wave equations. It will be more obvious later on that physically correct boundary conditions are such that $|R_k^\alpha| \le 1$. α_{kk} stands for the value of the artificial parameter on Γ_k, and $(\mathbf{u}_{kj}, \mathbf{n}_k)$ stands for the value of the velocity flux on the boundary. We now define

$$\begin{aligned}\mathcal{A}_2(U,V) = \sum_{kj} \int_{\Sigma_{kj}} \frac{1}{2\alpha_{jk}} \left(-K_j \frac{\partial}{\partial n_j} u_j + \frac{1}{2}(\mathbf{u}_{kj}, \mathbf{n}_j) u_j + \alpha_{jk} u_j\right) \\ \times \left(K_k \frac{\partial}{\partial n_k} v_k - \frac{1}{2}(\mathbf{u}_{kj}, \mathbf{n}_k) v_k + \alpha_{jk} v_k\right) d\sigma \\ + \sum_k \int_{\Gamma_k} \frac{R_k^\alpha}{2\alpha_{kk}} \left(-K_k \frac{\partial}{\partial n_k} u_k + \frac{1}{2}(\mathbf{u}_{kk}, \mathbf{n}_k) u_k + \alpha_{kk} u_k\right) \\ \times \left(K_k \frac{\partial}{\partial n_k} v_k - \frac{1}{2}(\mathbf{u}_{kk}, \mathbf{n}_k) v_k + \alpha_{kk} v_k\right) d\sigma.\end{aligned} \tag{3.10}$$

The bilinear form $\mathcal{A}_3$ is

$$\begin{aligned}\mathcal{A}_3(U,V) = \sum_k \int_{\partial\Omega_k} \frac{1}{2\alpha_{jk}} \left(K_k \frac{\partial}{\partial n_k} u_k - \frac{1}{2}(\mathbf{u}_{kj}, \mathbf{n}_k) u_k + \alpha_{jk} u_k\right) \\ \times \left(K_k \frac{\partial}{\partial n_k} v_k - \frac{1}{2}(\mathbf{u}_{kj}, \mathbf{n}_k) v_k + \alpha_{jk} v_k\right) d\sigma.\end{aligned} \tag{3.11}$$

Now that we have defined all the bilinear forms, let us show that they satisfy the required properties.

LEMMA 3.1. *Consider the bilinear forms* (3.4), (3.5), (3.10), (3.11). *Assume that* $|R_k^\alpha| \le 1$. *Then properties* (2.2) *are satisfied.*

Proof. One has

$$\begin{aligned}\mathcal{A}_0(U,U) &= \sum_k \int_{\Omega_k} (-u_k(t,x)\mathbf{u}.\nabla u_k(x) + u_k \nabla.(K_k \nabla u_k) + 2K_k \nabla u_k.\nabla u_k)\, dx \\ &\ge \sum_k \int_{\partial\Omega_k} \left(-\frac{1}{2}(\mathbf{u}, \mathbf{n}_k) u_k^2 + u_k K \frac{\partial}{\partial n_k} u_k\right) d\sigma = \frac{1}{2}\left(-\mathcal{A}_1(U,U) + \mathcal{A}_3(U,U)\right),\end{aligned}$$

which proves the first part of (2.2). Then using the Cauchy–Schwarz inequality and property $|R_k^\alpha| \le 1$, one gets $\mathcal{A}_2(U,V) \le \frac{1}{2}(\mathcal{A}_1(U,U) + \mathcal{A}_3(V,V))$, which is the second part of (2.2). $\mathcal{A}_1$ is obviously symmetric nonnegative. ∎

3.1.3. Boundary conditions. One major particularity of this formalism is the way boundary conditions are introduced. They are all defined by giving different values to parameter R_k^α. Equation (3.9) shows how to introduce homogeneous boundary conditions. The expressions of R_k^α for commonly used boundary conditions are given in Table 3.1. For the Robin-type boundary condition, we need to restrict the admissible boundary conditions to $\frac{1}{2}(\mathbf{u}, \mathbf{n}) + \sigma \ge 0$ so that $|R_k^\alpha| \le 1$.

LEMMA 3.2. *All* R_k^α *given in Table* 3.1 *satisfy the inequality* $|R_k^\alpha| \le 1$.

Proof. The proof is obtained by straightforward computation. ∎

TABLE 3.1
Values of R_k^α for commonly used boundary conditions in the convection-diffusion equation.

Outgoing	$K_k = 0$, $(\mathbf{u}, \mathbf{n}) > 0$	$R_k^\alpha = \frac{-(\mathbf{u},\mathbf{n})+\alpha}{(\mathbf{u},\mathbf{n})+\alpha}$
Ingoing Dirichlet	$K_k = 0$, $(\mathbf{u}, \mathbf{n}) < 0$	$R_k^\alpha = 0$
Dirichlet	$K_k > 0$, $(\mathbf{u}, \mathbf{n}) = 0$	$R_k^\alpha = -1$
Neumann	$K_k > 0$, $(\mathbf{u}, \mathbf{n}) = 0$	$R_k^\alpha = 1$
Mixed or Robin	$K_k \frac{\partial}{\partial n} c + \sigma c = 0$	$R_k^\alpha = \frac{\alpha - \frac{1}{2}(\mathbf{u},\mathbf{n}) - \sigma}{\alpha + \frac{1}{2}(\mathbf{u},\mathbf{n}) + \sigma}$

3.2. Fully discrete DGM. Now we need to choose the space $\mathcal{V}_h$. The standard choice for DGMs is $\mathcal{V}_h = \mathcal{V}_p \subset \mathcal{V}$ with

$$\mathcal{V}_p = \oplus_k P_p(\Omega_k), \tag{3.12}$$

where $P_p(\Omega_k)$ is the space of all polynomial functions of degree $p \in \mathbf{N}$ or less on cell Ω_k. Applying the time discretization defined in section 2.2, we obtain the fully discrete DGM. By construction, this DGM is L^2 stable for all p and without the need of any limiter. Therefore this method is different from the standard RKDG approach. The bilinear form $\mathcal{A}_0$ is local-to-one-cell so that both the first (2.4) and the second (2.7) order schemes are semi-implicit. In fact, one needs only inverse local linear systems to get the new solutions. Let us now analyze the abstract CFL condition in the case of uniform meshes.

4. CFL analysis. In this section we show that the abstract CFL condition (2.5) is equivalent to standard CFL requirements for the convection-diffusion equation, which is a kind of interpolation between pure convection and pure diffusion.

LEMMA 4.1. *Consider (for simplicity) a sequence of triangular and conformal meshes. Assume the sequence of meshes is uniformly regular. Denote by h a characteristic length of the mesh. Consider the first order scheme* (2.4) *with bilinear forms* (3.4), (3.5), (3.10).

For all $p \in \mathbf{N}$, there exists two constants $C_p^1 > 0, C_p^2 > 0$ such that if

$$\frac{3}{2}\Delta t \max_k \left(\frac{\alpha_{kj}}{C_p^1 h} + \frac{|\mathbf{u}|^2}{4\alpha_{kj} C_p^1 h} + \frac{K_k^2}{\alpha_{kj} C_p^2 h^3} \right) \leq 1, \tag{4.1}$$

then the abstract CFL condition (2.5) *holds, and* (2.4) *is L^2 stable. Assuming that K is constant for simplicity, the optimal value of α corresponding to the least stringent CFL constraint is*

$$\alpha_{opt} = \sqrt{\frac{|\mathbf{u}|^2}{4} + \frac{K^2 C_p^1}{C_p^2 h^2}}. \tag{4.2}$$

Proof. First, the abstract CFL condition (2.5) is

$$\Delta t \max_k \left(\max_{\text{degree}(u_k) \leq p} \frac{1}{2\alpha_{kj}} \frac{\int_{\partial\Omega_k} (\alpha_{kj} u_k + \frac{1}{2}(\mathbf{u}, \mathbf{n}_k) u_k - K \frac{\partial}{\partial n_k} u_k)^2}{\int_{\Omega_k} u_k^2} \right) \leq 1.$$

This is true once the following inequality is satisfied:

$$\Delta t \max_k (T_1^k + T_2^k + T_3^k) \leq 1,$$

where T_1^k, T_2^k, T_3^k are given by

$$T_1^k = 3 \max_{\text{degree}(u_k)\le p} \frac{1}{2\alpha_{kj}} \frac{\int_{\partial\Omega_k} (\alpha_{kj} u_k)^2}{\int_{\Omega_k} u_k^2},$$
$$T_2^k = 3 \max_{\text{degree}(u_k)\le p} \frac{1}{2\alpha_{kj}} \frac{\int_{\partial\Omega_k} (\frac{1}{2}(\mathbf{u}, \mathbf{n}_k) u_k)^2}{\int_{\Omega_k} u_k^2},$$
$$T_3^k = 3 \max_{\text{degree}(u_k)\le p} \frac{1}{2\alpha_{kj}} \frac{\int_{\partial\Omega_k} (K \frac{\partial}{\partial n_k} u_k)^2}{\int_{\Omega_k} u_k^2}.$$

Let us introduce the linear transformation F_k that maps the triangular cell Ω_k onto the reference cell $\hat{T}$. Using the regularity of the mesh,

$$T_1^k \le 3 \frac{\alpha_{kj}}{2hc_k} \left(\max_{\text{degree}(\hat{u}_k)\le p} \frac{\int_{\partial\hat{T}} \hat{u}_k^2}{\int_{\hat{T}} \hat{u}_k^2} \right),$$

where c_k depends on transformation F_k. Since the mesh is assumed to be uniformly regular, c_k is uniformly bounded from below. Let us define

$$c^p = \max_{\text{degree}(\hat{u}_k)\le p} \frac{\int_{\partial\hat{T}} \hat{u}_k^2}{\int_{\hat{T}} \hat{u}_k^2} \quad \text{and} \quad C_p^1 = \frac{\min_k c^k}{c^p}. \quad \text{Then} \quad T_1^k \le \frac{3}{2} \frac{\alpha_{kj}}{h} \frac{1}{C_p^1}.$$

Also, one has

$$T_2^k \le \frac{3}{2} \frac{|\mathbf{u}|^2}{4h} \frac{1}{\alpha_{kj} C_p^1}.$$

Using again the regularity of the mesh, we have

$$T_3^k \le \frac{3}{2} \frac{K_k^2}{\alpha_{kj}} \frac{d_k}{h^3} \left(\max_{\text{degree}(u_k)\le p} \frac{\int_{\partial\hat{T}} (\frac{\partial}{\partial\hat{n}_k} u_k)^2}{\int_{\hat{T}} \hat{u}_k^2} \right),$$

where d_k depends on F_k. Since the mesh is assumed to be uniformly regular, d_k is uniformly upper bounded. Let us define

$$e_p = \max_{\text{degree}(\hat{u}_k)\le p} \frac{\int_{\partial\hat{T}} \frac{\partial}{\partial\hat{n}} u_k^2}{\int_{\hat{T}} \hat{u}_k^2} \quad \text{and} \quad C_p^2 = \frac{1}{e_p \max_k d_k}. \quad \text{Then} \quad T_3^k \le \frac{3}{2} \frac{K_k^2}{\alpha_{kj}} \frac{d_k}{h^3} \frac{1}{C_p^2}.$$

Putting this all together, we have

$$\Delta t \max_k (T_1^k + T_2^k + T_3^k) \le \frac{3}{2} \Delta t \max_k \left(\frac{\alpha_{kj}}{C_p^1 h} + \frac{|\mathbf{u}|^2}{4\alpha_{kj} C_p^1 h} + \frac{K_k^2}{\alpha_{kj} C_p^3 h^3} \right).$$

The abstract CFL condition is thus satisfied once we have

$$\frac{3}{2} \Delta t \max_k \left(\frac{\alpha_{kj}}{C_p^1 h} + \frac{|\mathbf{u}|^2}{4\alpha_{kj} C_p^1 h} + \frac{K_k^2}{\alpha_{kj} C_p^3 h^3} \right) \le 1.$$

Assuming K is constant, the optimal value of parameter α is the one that minimizes the multiplicative constant in front of Δt. Since the constant is $a\alpha + \frac{1}{\alpha b}$, where $a > 0$ and $b > 0$ are constants, then the optimal value is the solution of the equation $\frac{d}{d\alpha} \left(a\alpha + \frac{b}{\alpha} \right) = 0$, that is, $\alpha = \sqrt{\frac{b}{a}}$. Expanding with the definition of a and b, it gives (4.2). ∎

4.1. Particular cases. This section discusses the particular cases of the pure advection equation (i.e., $K \equiv 0, \mathbf{u}$ constant) and pure diffusion equation (i.e., $\mathbf{u} \equiv 0$ but $K > 0$).

4.1.1. Pure advection. In this particular case we have ($K \equiv 0, \mathbf{u}$ constant). Notation is still the same as in section 3.1.1. A consequence of Lemma 4.1 is the following.

LEMMA 4.2. *Consider a sequence of triangular and conformal meshes. Assume the sequence of meshes is uniformly regular. Denote by h a characteristic length of the mesh. For all $p \in \mathbf{N}$, there exists a $C_p^1 > 0$ such that if*

$$|\mathbf{u}|\Delta t \leq C_p^1 h, \tag{4.3}$$

then the abstract CFL condition is true.

4.1.2. Pure diffusion. In this particular case we have ($K > 0$ but $\mathbf{u} \equiv 0$). The equation is

$$\partial_t c - \nabla.(K \nabla c) = 0. \tag{4.4}$$

We consider $\alpha_{kj} \equiv \alpha > 0$ for simplicity of notation. Asfwe s in Lemma 4.1 we have the following.

LEMMA 4.3. *Consider a sequence of triangular and conformal meshes. Assume the sequence of meshes is uniformly regular. Denote by h a characteristic length of the mesh. For all $p \in \mathbf{N}$, there exists a $C_p^2 > 0$ such that if*

$$\Delta t \leq \frac{1}{\frac{\alpha}{C_p^1 h} + \frac{K^2}{\alpha C_p^2 h^3}}, \tag{4.5}$$

then the abstract CFL condition is true. Both constants C_p^1, C_p^2 depend only on the mesh and the degree of the polynomials, and not on the parameters of the equations or on α.

For an optimal value for parameter α, we also have the following.

LEMMA 4.4. *Consider the CFL inequality* (4.5)*, with parameter α set to*

$$\alpha = \frac{K}{h}. \tag{4.6}$$

Then inequality (4.5) *is equivalent to the more standard CFL inequality*

$$K\Delta t \leq C_p^3 h^2, \quad \frac{1}{C_p^3} = \frac{1}{C_p^1} + \frac{1}{C_p^2}. \tag{4.7}$$

The proof is left to the reader.

The value (4.6) is optimal, since we recover the classical time step CFL constraint for explicit discretization of diffusion.

Remark. Formula (4.2) is a kind of continuous interpolation between (4.3) and (4.7). More importantly, if $K \equiv 0$, then $\alpha = \frac{|\mathbf{u}|}{2}$ and the scheme defined by (3.4), (3.5), (3.10) is *equal* to the standard DGM for the pure advection case. On the other hand, if $\mathbf{u} \equiv 0$, then the method is *equal* to the DGM defined above for the pure diffusion case. Therefore (4.2) ensures that the scheme for advection-diffusion is a continuous interpolation between the scheme for pure advection and the scheme for pure diffusion.

5. Convergence analysis for the advection case. Let us now state the convergence result. We restrict the analysis to the DGM for advection and leave convergence analysis of diffusion for future studies. Let us define an L^2 projection $\pi_h : \mathcal{H} \to \mathcal{V}_p$,

$$\pi_h(u) = (u_k) \Longleftrightarrow \int_{\Omega_k} u_k(x)v_k(x)dx = \int_{\Omega_k} u(x)v_k(x)dx \quad \forall v_k,\ \forall k. \tag{5.1}$$

The scheme that we analyze in this section is defined by

$$\begin{cases} U_h^0 = \pi_h(u_0), \text{ where } u_0 \text{ is the initial condition,} \\ U_h^1 \text{ is the solution of the first order time scheme (2.4),} \\ U_h^{n+1} \text{ is the solution of the second order time scheme (2.8),} \\ \text{the bilinear forms are } \mathcal{A}_0,\ \mathcal{A}_1,\ \mathcal{A}_2,\ \mathcal{A}_3, \\ \qquad\qquad \text{as defined in section 3.1.2 in the case } K \equiv 0. \end{cases} \tag{5.2}$$

We will use the following approximation property of the projection π_h.

LEMMA 5.1. *Let E be an element (a triangle or a tetrahedron) in $\mathbb{R}^n (n = 2, 3)$ of diameter h_E. Then for any $u \in H^{k+1}(E)$,*

$$\|u - \pi_h u\|_{H^r(E)} \le C h_E^{k+1-r} \|u\|_{H^{k+1}(E)} \qquad r = 0,\ 1,$$

where C is independent of h_E. See [2].

LEMMA 5.2 (trace inequality). *Let E be an element in $\mathbb{R}^n (n = 2, 3)$ of diameter h_E. Let e_k be an edge or a face of E. Then for any f in $H^s(E)$ and for $s \ge 2$,*

$$\|f\|_{L^2(e_k)} \le \hat{C} |e_k|^{\frac{1}{2}} |E|^{-\frac{1}{2}} (\|f\|_{L^2(E)} + h_E \|\nabla f\|_{L^2(E)}).$$

If f is a polynomial of degree $p > 0$ on E,

$$\|f\|_{L^2(e_k)} \le \hat{C} p^2 |e_k|^{\frac{1}{2}} |E|^{-\frac{1}{2}} (\|f\|_{L^2(E)}).$$

Here $\hat{C}$ is a constant independent of h_E and p. See [33].

LEMMA 5.3. *Let $c \in \mathcal{V}$ be the solution of the advection equation and $U_h^n \in \mathcal{V}_p$ be the solution of* (5.2). *Then*

$$\theta_{l+1}^2 - \theta_l^2 \le 6 \Delta t r^{l+1}, \tag{5.3}$$

where

$$\begin{aligned} \theta_l^2 &= (\xi^l, \xi^l) + (2\xi^l - \xi^{l-1}, 2\xi^l - \xi^{l-1}) \quad \forall\ l \ge 1, \\ \xi^l &= \pi_h u(l\Delta t) - U_h^l, \\ 6\chi^l &= \pi_h u(l\Delta t) - u(l\Delta t) \text{ and} \\ r^{l+1} &= \frac{1}{3}\left(\frac{3\chi^{l+1} - 4\chi^l + \chi^{l-1}}{\Delta t}, \xi^{l+1}\right) + \frac{2}{3}\mathcal{A}_0(\chi^{l+1}, \xi^{l+1}) \\ &\quad + \frac{2}{3}\mathcal{A}_1(2\chi^l - \chi^{l-1}, \xi^{l+1}) - \frac{2}{3}\mathcal{A}_2(2\chi^l - \chi^{l-1}, \xi^{l+1}) \\ &\quad + \frac{1}{3}\left(\frac{3u^{l+1} - 4u^l + u^{l-1}}{\Delta t} - 2\partial_t u((l+1)\Delta t), \xi^{l+1}\right) \\ &\qquad + \frac{2}{3}\mathcal{A}_1(2u^l - u^{l-1} - u^{l+1}, \xi^{l+1}) r \\ &\qquad - \frac{2}{3}\mathcal{A}_2(2u^l - u^{l-1} - u^{l+1}, \xi^{l+1}). \end{aligned}$$

Proof. Taking $V_h = \xi^{l+1}$ in (2.8) with U_h^l replaced by $\pi_h u(l\Delta t)$, and subtracting the resulting equation in which $V_h = \xi^{l+1}$, from (2.8), gives

$$\frac{1}{3}\left(\frac{3\xi^{l+1}-4\xi^l+\xi^{l-1}}{\Delta t},\xi^{l+1}\right)+\frac{2}{3}\mathcal{A}_0(\xi^{l+1},\xi^{l+1}) + \frac{2}{3}\mathcal{A}_1(2\xi^l-\xi^{l-1},\xi^{l+1})-\frac{2}{3}\mathcal{A}_2(2\xi^l-\xi^{l-1},\xi^{l+1})=r^{l+1}.$$

Using the lower bounds of $\mathcal{A}_0$ and $\mathcal{A}_2$ given by (2.2) and the symmetry of the bilinear form $\mathcal{A}_1$, we have

$$\frac{1}{3}\left(\frac{3\xi^{l+1}-4\xi^l+\xi^{l-1}}{\Delta t},\xi^{l+1}\right)+\frac{1}{3}\mathcal{A}_1(\xi^{l+1}-2\xi^l+\xi^{l-1},\xi^{l+1}-2\xi^l+\xi^{l-1})\le r^{l+1}.$$

Now applying the abstract CFL condition (2.9), we further obtain

$$\frac{1}{3}\left(\frac{3\xi^{l+1}-4\xi^l+\xi^{l-1}}{\Delta t},\xi^{l+1}\right)-\frac{1}{6\Delta t}(\xi^{l+1}-2\xi^l+\xi^{l-1},\xi^{l+1}-2\xi^l+\xi^{l-1})\le r^{l+1}$$

which, from the equality

$$(\theta_{l+1}^2-\theta_l^2)/(6\Delta t)=\frac{1}{3}\left(\frac{3\xi^{l+1}-4\xi^l+\xi^{l-1}}{\Delta t},\xi^{l+1}\right)-\frac{1}{6\Delta t}(\xi^{l+1}-2\xi^l+\xi^{l-1},\xi^{l+1}-2\xi^l+\xi^{l-1}),$$

reduces to $\theta_{l+1}^2-\theta_l^2 \le 6\Delta t\, r^{l+1}$. This ends the proof. ∎

LEMMA 5.4. *Notation is the same as in Lemma* 5.3. *Let us assume that the solution c is sufficiently smooth. Then there exist two constants, C_1 and C_2 not depending on l, Δt, and h such that*

$$|r^{l+1}| \le (C_1(\Delta t)^2 + C_2 h^{\mu-1})\theta_{l+1}. \tag{5.4}$$

Here $\mu = \min(p+1, s)$ and s is the order of regularity of the solution in Sobolev's spaces.[1]

Proof. The velocity $\mathbf{u}$ is constant. In this proof we denote its module by $c_{vel} = |\mathbf{u}|$. The method consists of estimating all the terms in the right hand side in the definition of r^{l+1} in lemma 5.3. By the definition of the projection π_h, we have

$$\frac{1}{3}\left(\frac{3\chi^{l+1}-4\chi^l+\chi^{l-1}}{\Delta t},\xi^{l+1}\right)=0.$$

Since $\mathbf{u}.\nabla\xi_k^{l+1} \in \mathcal{V}_p$, we have $\int_{\Omega_k}\chi_k^{l+1}\mathbf{u}.\nabla\xi_k^{l+1}dx = 0$. Therefore $\mathcal{A}_0(\chi^{l+1},\xi^{l+1})=0$. Let us estimate $|\mathcal{A}_1(2\chi^l-\chi^{l-1},\xi^{l+1})|$.

$$\begin{aligned}
|\mathcal{A}_1(2\chi^l-\chi^{l-1},\xi^{l+1})| &\le \sum_k \int_{e\in\partial\Omega_k} c_{vel}|(2\chi_k^l-\chi_k^{l-1})||\xi_k^{l+1}| \\
&\le \sum_k c_{vel}h^{-1}\left(\|(2\chi_k^l-\chi_k^{l-1})\|_{L^2(\Omega_k)}\right. \\
&\qquad \left. + h\|\nabla(2\chi_k^l-\chi_k^{l-1})\|_{L^2(\Omega_k)}\right)\|\xi_k^{l+1}\|_{L^2(\Omega_k)} \\
&\le \sum_k c_{vel}c_1 h^{\mu-1}\|\xi_k^{l+1}\|_{L^2(\Omega_k)} \\
&\le Ch^{\mu-1}(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}}.
\end{aligned} \tag{5.5}$$

[1] A requirement of which is that $u \in C^1([0,T];H^s(\Omega))$, $u_{tt} \in L^\infty([0,T];L^\infty(\Omega))$, and $u_{ttt} \in L^\infty([0,T];L^2(\Omega))$.

Similarly,

$$
\begin{aligned}
|\mathcal{A}_2(2\chi^l-\chi^{l-1},\xi^{l+1})| &\le \sum_{k,j}\int_{e\in\partial\Omega_k\cap\partial\Omega_j} c_{vel}|(2\chi_j^l-\chi_j^{l-1})||\xi_k^{l+1}| \\
&\le \sum_{k,j} c_{vel}h^{-1}\left(\|(2\chi_j^l-\chi_j^{l-1})\|_{L^2(\Omega_j)}\right. \\
&\qquad \left. + h\|\nabla(2\chi_j^l-\chi_j^{l-1})\|_{L^2(\Omega_j)}\right)\|\xi_k^{l+1}\|_{L^2(\Omega_k)} \\
&\le Ch^{\mu-1}(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}}.
\end{aligned} \tag{5.6}
$$

The two other terms are

$$
\begin{aligned}
\left|\left(\frac{3u^{l+1}-4u^l+u^{l-1}}{\Delta t}-2(\partial_t u)^{l+1},\xi_k^{l+1}\right)\right| &\le (\Delta t)^2\sum_k\int_{\Omega_k} c_{vel}|\partial_{ttt}u(t^*,x)||\xi_k^{l+1}(x)| \\
&\le C(\Delta t)^2\|\partial_{ttt}u\|_{L^\infty(0,T;L^2(\Omega))}(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}} \\
&\le C(\Delta t)^2(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}}.
\end{aligned}
$$

Also,

$$
\begin{aligned}
|\mathcal{A}_1(2u^l-u^{l-1}-u^{l+1},\xi^{l+1})| &\le (\Delta t)^2\sum_k\int_{e\in\partial\Omega_k} c_{vel}|\partial_{tt}u(t^*,x)||\xi_k^{l+1}(x)| \\
&\le (\Delta t)^2\sum_k c_{vel}\|\partial_{tt}u(t^*)\|_{L^\infty(\Omega_k)}\int_{e\in\partial\Omega_k}|\xi_k^{l+1}(x)| \\
&\le (\Delta t)^2\sum_k c_{vel}\|\partial_{tt}u(t^*)\|_{L^\infty(\Omega_k)}h^{\frac{1}{2}}h^{-\frac{1}{2}}\|\xi_k^{l+1}\|_{L^2(\Omega_k)} \\
&\le C(\Delta t)^2\|\partial_{tt}u\|_{L^\infty(0,T;L^\infty(\Omega))}(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}} \\
&\le C(\Delta t)^2(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}}.
\end{aligned}
$$

Proceeding as above, we have

$$
|\mathcal{A}_2(2u^l-u^{l-1}-u^{l+1},\xi^{l+1})| \le C(\Delta t)^2(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}}.
$$

Now observing that $(\xi^{l+1},\xi^{l+1})^{\frac{1}{2}} \le \theta_{l+1}$, we obtain the result by summing all the above inequalities. ∎

THEOREM 5.5 (L^2 error estimate for pure advection). *Let $c\in\mathcal{V}$ be the solution of* (1.1) *in the advection case ($K\equiv 0$) with initial condition $c_0\in H^s(s\ge 2)$ and $U_h\in\mathcal{V}_p$ the solution of* (2.8)*, with the initial condition given by* (5.2)*. Assume the CFL condition* (2.9)*. Then there exist two constants C_1 and C_2 depending only on T and u such that*

$$
\|(u-U_h)(T)\|_{L^2} \le 3\|\pi_h u(\Delta t)-U_h^1\|_{L^2} + C_1(\Delta t)^2 + C_2h^{\mu-1},
$$

where $\mu=\min(p+1,s)$.

Proof. Using the triangular inequality, we have

$$
\|(u-U_h)(T)\|_{L^2} \le \|(u-\pi_h u)(T)\|_{L^2} + \|(\pi_h u-U_h)(T)\|_{L^2}.
$$

The first term on the R.H.S. is bounded using the classical approximation theory [16] $\|(u-\pi_h u)(T)\|_{L^2}\le c(u)h^\mu$. Observe that by

$$
\|(\pi_h u-U_h)(T)\|_{L^2}^2 = (\xi^N,\xi^N),
$$

it is possible to give an upper bound where N is defined by $T = N\Delta t$. Therefore according to Lemma 5.3, we have

$$\left(\theta_{n+1}^2 - \theta_n^2\right)/6\Delta t \leq r^{n+1}.$$

From Lemma 5.4 there exist two constants C_1 and C_2 such that

$$\theta_{n+1}^2 - \theta_n^2 \leq 6\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1})\theta_{n+1}.$$

We then have $\theta_{n+1}^2 - 6\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1})\theta_{n+1} \leq \theta_n^2$, which can be rewritten as

$$\left(\theta_{n+1} - 3\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1})\right)^2 \leq \theta_n^2 + \left(3\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1})\right)^2.$$

Therefore $\theta_{n+1} - \theta_n \leq 6\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1})$. Summing this inequality over all n from 1 to $N-1$ produces

$$\theta_N \leq \theta_0 + \sum_{n=1}^{n=N-1} 6\Delta t(C_1(\Delta t)^2 + C_2 h^{\mu-1}).$$

Since

$$\begin{aligned}\theta_0^2 &= (\xi^1, \xi^1) + (2\xi^1 - \xi^0, 2\xi^1 - \xi^0)\\ &\leq ((\xi^1, \xi^1)^{\frac{1}{2}} + (2\xi^1 - \xi^0, 2\xi^1 - \xi^0)^{\frac{1}{2}})^2\\ &\leq (3(\xi^1, \xi^1)^{\frac{1}{2}} + (\xi^0, \xi^0)^{\frac{1}{2}})^2,\end{aligned}$$

we have $\theta_0 \leq 3(\xi^1, \xi^1)^{\frac{1}{2}} + (\xi^0, \xi^0)^{\frac{1}{2}}$. By definition of the scheme, initials values are such that

$$\xi^1 = \pi_h u(\Delta t) - U_h^1 \text{ and } \xi^0 = 0.$$

Also one has $N\Delta t = T$ so that $\sum_1^{N-1}(6\Delta t) \leq 6T$. Therefore taking $C_i = C_i 6T, i = 1, 2$, ends the proof. □

Remark.

- The above theorem shows the convergence of the second order time discretization. Note that since it is second order in time, two initial conditions are needed: U_h^0, U_h^1. We have taken U_h^1 as the solution of a particular iteration of the first order scheme. So $\pi_h u(\Delta t) - U_h^1$ can be kept as small as we need.
- One can observe that in the demonstration above, except in Lemma 5.4, we have used only the property of the bilinear forms $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2$. So by just giving an analogous lemma for pure diffusion and for mixed convection-diffusion equations, one obtains the convergence result for those equations. It is possible to guess that, in general, one has

  $$|r^{l+1}| \leq (C_1(\Delta t)^\nu + C_2 h^\mu)(\xi^{l+1}, \xi^{l+1})^{\frac{1}{2}},$$

  where $\nu = 1, 2$ is the order of time discretization and μ is the order of the approximation error seen by the bilinear forms $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2$. Note that μ can be kept optimal by replacing the L^2-projection with a well-chosen projection $\mathbf{R}_h$ related to the Gauss quadrature formula; see [17].

6. Numerical results. This section is devoted to the study of the order of convergence of our method by means of numerical tests and comparison with other methods. The algorithm presented in this work is denoted by the words "new formalism."

6.1. Pure advection. In this example, we consider (1.1) in the case when $K \equiv 0$. The computational domain is $(\Omega = (-0.5, 0.5)^2)$. The initial condition and the inflow boundary condition are taken from the exact solution, which is chosen here to be

$$c(t,x,y) = \exp\left(-\frac{(\hat{x}-x_c)^2 + (\hat{y}-y_c)^2}{2\sigma^2}\right).$$

The velocity field is $u = (-1,1)^T$ and $\hat{x} = x + t, \hat{y} = y - t$. The parameters are $x_c = 0.25, y_c = -0.25, 2\sigma^2 = 0.004$. The time interval for the simulation is $(0, 0.5)$, which is the required time to shift the cone from its initial position to the symmetric position with respect to the center $(0,0)$. The domain is subdivided into an initial mesh consisting of $8\times8\times2 = 138$ uniform regular triangles. We then successively refine the mesh and compute L^2 and L^∞ errors e_h on the mesh of size h and the numerical convergence rates by the ratio $\ln(e_h/e_{h/2})/\ln(2)$. The use of uniform meshes leads to the following values for the parameters in the CFL analysis. In formula (4.3) the value of C_p^1 is

$$C_p^1 = \begin{cases} \frac{1}{4+4\sqrt{2}} & \text{for } \ p = 1, \\ \frac{1}{6+6\sqrt{2}} & \text{for } \ p = 2. \end{cases}$$

For a second order in time discretization the value of C_p^1 is divided by 2. In our computations we divide it by 10, just to stay away from the optimal value. Table 6.1 shows the behavior of our formalism with respect to the order of the polynomial basis and time discretization. In Table 6.2 we compare the new formalism with RKDG (without flux limiting), RKDG (with the Cockburn–Shu flux limiting) that we call TVBMRKDG (total variation bounded modified slope limiter; see [23]), and with a Crank–Nicholson scheme applied to the stabilized DGM formulation of convection-equation introduced by Brezzi, Marini, and Süli [10]. The last one is introduced to compare our results to schemes in which the global matrix is inverted at every time step. We have done an element renumbering in that Crank–Nicholson scheme in order to have a thin band global matrix. We factor the global matrix before entering into loops, which leads to a gain in time compared to a sparse direct resolution of the global algebraic equation at every time step. The time required to do this operation is denoted by R in Table 6.2.

Observations. From Table 6.1, the error at the time T is of the form $C_1(\Delta t)^\alpha + C_2 h^\beta$, where α is the order of the time discretization and β is a real whose optimal value is $\beta = p + 1$ (where p is the degree of the polynomials). Even if constants C_1, C_2 influence the computed convergence rate, one can still observe that when using polynomials of order p with second order time discretization, the L^2 error is at least of order p in space. By comparison with other theoretical results [10] it is possible to conjecture a behavior of the form $O(\Delta t^2) + O(h^{p+\frac{1}{2}})$. But for this test problem the error in time is clearly dominant over the error in space. Therefore it is difficult to clearly identify the asymptotic order of convergence when using the second order in time discretization. At a more general level, it shows the interest of the second order in time discretization. This is seen in Table 6.2, where we observe the same convergence rate with RKDG without flux limiting, which is of order 2 for polynomials of order 1. The same convergence rate is observed for the Crank–Nicholson scheme applied to the formulation of [10]. These three second order formulations produce the same convergence rate for first order polynomials.

TABLE 6.1

Numerical L^2 errors, L^∞ errors, and convergence rate at time $t = 0.5s$, for first and second order in time with first and second order basis polynomials, in the new formalism ((2.4), (2.8)) scheme applied to the pure advection equation.

	First order in time				Second order in time			
h	L^2 error	Rate	L^∞ error	Rate	L^2 error	Rate	L^∞ error	Rate
P_1 basis polynomials								
1/8	$5.47E-02$	–	$7.28E-01$	–	$5.15E-02$	–	$6.42E-01$	–
1/16	$4.08E-02$	0.49	$6.15E-01$	0.31	$3.43E-02$	0.59	$5.07E-01$	0.34
1/32	$2.11E-02$	1.02	$3.54E-01$	0.94	$1.31E-02$	1.39	$2.16E-01$	1.23
1/64	$9.72E-03$	1.16	$1.63E-01$	1.17	$3.08E-03$	2.09	$5.65E-02$	1.93
1/128	$4.78E-03$	1.02	$7.55E-02$	1.11	$5.87E-04$	2.40	$1.13E-02$	2.32
P_2 basis polynomials								
1/8	$4.23E-02$	–	$6.39E-01$	–	$3.14E-02$	–	$4.83E-01$	–
1/16	$2.05E-02$	1.05	$3.21E-01$	0.99	$6.99E-03$	2.17	$1.10E-01$	1.80
1/32	$1.09E-02$	0.91	$1.59E-01$	1.01	$5.44E-04$	3.68	$1.17E-02$	3.23
1/64	$5.90E-03$	0.89	$8.49E-02$	0.91	$4.37E-05$	3.64	$1.87E-03$	2.65
1/128	$3.10E-03$	0.93	$4.49E-02$	0.92	$6.66E-06$	2.73	$2.53E-04$	2.88

TABLE 6.2

*Comparison of numerical errors and convergence rates at time $t = 0.5s$, for second order in time with first order basis polynomials. R is the time spent renumbering the elements and factoring the global matrix. Computational times are for the finest mesh, using a Pentium III/*1.266 *GHZ.*

	New formalism		RKDG		TVBMRKDG		Crank–Nicholson	
h	Error	Rate	Error	Rate	Error	Rate	Error	Rate
L^2 errors								
1/8	$5.15E-02$	–	$5.18E-02$	–	$5.23E-02$	–	$5.15E-02$	–
1/16	$3.43E-02$	0.59	$3.44E-02$	0.59	$3.83E-02$	0.45	$3.43E-02$	0.59
1/32	$1.31E-02$	1.39	$1.31E-02$	1.39	$2.96E-02$	0.37	$1.31E-02$	1.39
1/64	$3.08E-03$	2.09	$3.08E-03$	2.09	$1.39E-02$	1.09	$3.07E-03$	2.09
L^∞ errors								
1/8	$6.42E-01$	–	$6.48E-01$	–	$6.57E-01$	–	$6.43E-01$	–
1/16	$5.07E-01$	0.34	$5.08E-01$	0.35	$5.58E-01$	0.24	$5.05E-01$	0.34
1/32	$2.16E-01$	1.23	$2.16E-01$	1.23	$4.63E-01$	0.27	$2.15E-01$	1.23
1/64	$5.65E-02$	1.93	$5.62E-02$	1.96	$2.79E-01$	0.73	$5.62E-02$	1.93
CPU time								
1/64	81.38		90.34		32400		$553.93 + R$	

6.2. Pure diffusion. In this example we consider the Dirichlet equation (1.1) with $(K \equiv 1, u \equiv 0)$. The computational domain is $\Omega = (0,1)^2$. The boundary condition is homogeneous so that the exact solution is

$$c(t,x,y) = \sin(\pi x)\sin(\pi y)\exp(-2\pi^2 t).$$

The initial condition is taken from this exact solution. The time interval is $(0, 1.510^{-2})$. This is the required time to reduce the maximum of the exact solution by about 25%. The domain is meshed into 16 uniform regular triangles. We successively refine this mesh uniformly. For each mesh of size h we compute the L^2 and L^∞ errors e_h and the numerical convergence rates given by the ratio $\ln(e_h/e_{h/2})/\ln(2)$. The use of uniform meshes leads to the following values of C_p^2 in formula (4.5): $C_p^2 = \frac{1}{12+6\sqrt{2}}$ for $p = 1$ and $C_p^2 = \frac{1}{120+66\sqrt{2}}$ for $p = 2$.

In order to enforce a better interelement continuity for small p, one can choose the parameter α to be of the form $\alpha = \beta\frac{K}{h}$, where $\beta \geq 1$ is a user-defined constant.

The optimal value of β is $\beta = \sqrt{\frac{C_p^1}{C_p^2}}$. Therefore our optimal value for C_p^3 in formula (4.7) is in this case $C_p^3 = \sqrt{C_p^1 C_p^2}$. In Table 6.4 we compare the new formalism for first order in time and second order polynomials with computed solutions obtained by Nonsymmetric Interior Penalty Galerkin (NIPG) and Symmetric Interior Penalty Galerkin (SIPG) GDMs [7, 35, 36]. For this first order in time, we have used an implicit scheme to discretize the SIPG and NIPG methods. We intended to do the same comparison for the second order in time. We tried a θ-scheme (see [34]) to discretize time in both SIPG and NIPG (note that implicit scheme corresponds to a θ-scheme with $\theta = 1$, as in [29], while the Crank–Nicholson scheme corresponds to $\theta = 0$ as described in [34]). But we noticed that using the same time step for the new formalism and for SIPG and NIPG Galerkin methods with the Crank–Nicholson scheme leads to instabilities in SIPG and NIPG. So for that time step, θ must stay in the interval $]0, 1]$, and therefore the θ-scheme is no longer of second order. This is a significant advantage of our formalism over the two others. We have taken the stabilization parameter $\sigma = 1$ for NIPG and $\sigma = 10$ for SIPG; see [35, 36]. The time step has also been multiplied by 10 in SIPG and NIPG, which are implicit methods ($\theta = 1$).

Observations. Here, as in the pure advection case, the error is of the form $C_1(\Delta t)^\alpha + C_2 h^\beta$. Since we have used the optimal CFL condition while refining the mesh, $\Delta t \approx Ch^2$, the convergence rate obtained numerically should be close to

$$\gamma = \min(2\alpha, \beta).$$

Let us discuss the values of α, β, and γ observed in Tables 6.3 and 6.5. For first order time discretization, $\alpha = 1$. Therefore $\gamma = \min(2, \beta)$. It shows that for first order or second order polynomials in conjunction with first order time discretization, we obtain a convergence rate of order 2. This is what we get in Table 6.3. Second order time discretization with first order polynomials gives also a convergence rate of $\gamma = 2$. Hence $\beta = 2$ for first order polynomials, and the convergence in space is optimal in this case. It is also seen in Table 6.3 that when we use second order polynomials with second order time discretization, the convergence rate starts from almost 3 and tends asymptotically to $\gamma = 2$. This shows that $\alpha = 2$ and $\beta = 2$ for second order time discretization with second order polynomials. Hence the convergence in space is suboptimal in this case. However, this is only a matter of worst behavior for even order polynomials. To view that, let us try third order polynomials with a sufficiently small CFL condition so that the error in time is absolutely negligible and we get an accurate value for β. Table 6.5 shows a convergence rate of order $\gamma = \min(2\times 2, 4) = 4$. By inspection of all these results we deduce that the new formalism presented in this paper keeps (on pure diffusion) the optimal space convergence rate for polynomials of odd order. This behavior is similar to other nonsymmetric discretizations like NIPG.

In order to analyze the advantage of our method over NIPG and SIPG for this kind of test problem, let us analyze the ratio accuracy/CPU time of the computation (see the last line of Table 6.4). We see that the error is slightly smaller for our method. But more important is the CPU time required to perform the computation. Due to well-known stability issues, NIPG and SIPG are implicit, which means a certain CPU time is needed to factorize and invert the matrix. This CPU time is denoted as R in the table. It is well known that R can be quite large. In our computations, R is about the same order as the CPU time needed to perform the whole computation. But here the matrix is factorized only once because the coefficients of the problem are constant

TABLE 6.3

Numerical L^2 errors, L^∞ errors, and convergence rates for first and second order in time with first and second order basis polynomials in the new formalism ((2.4), (2.8)) *scheme applied to the pure diffusion equation.*

	First order in time				Second order in time			
h	L^2 error	Rate	L^∞ error	Rate	L^2 error	Rate	L^∞ error	Rate
P_1 basis polynomials								
1/8	$1.00E-02$	–	$3.10E-02$	–	$8.87E-03$	–	$3.17E-02$	–
1/16	$2.50E-03$	2.00	$7.47E-03$	2.05	$2.16E-03$	2.04	$7.50E-03$	2.08
1/32	$6.20E-04$	2.01	$1.83E-03$	2.03	$5.37E-04$	2.00	$1.84E-03$	2.02
1/64	$1.55E-04$	2.00	$4.56E-04$	2.00	$1.34E-04$	2.00	$4.57E-04$	2.00
1/128	$3.87E-05$	2.00	$1.14E-04$	2.00	$3.35E-05$	2.00	$1.14E-04$	2.00
P_2 basis polynomials								
1/8	$8.95E-04$	–	$2.63E-03$	–	$7.53E-04$	–	$3.02E-03$	–
1/16	$2.10E-04$	2.09	$4.63E-04$	2.50	$1.75E-04$	2.11	$4.92E-04$	2.61
1/32	$5.16E-05$	2.02	$1.15E-04$	2.00	$4.27E-05$	2.03	$9.64E-05$	2.35
1/64	$1.28E-05$	2.01	$2.86E-05$	2.00	$1.06E-05$	2.01	$2.20E-05$	2.13
1/128	$3.20E-06$	2.00	$7.14E-06$	2.00	$2.65E-06$	2.00	$5.34E-06$	2.04

TABLE 6.4

Numerical comparison of L^2 errors, L^∞ errors, CPU time, and convergence rate, for first order in time with second order basis polynomials in the new formalism, and implicit scheme for SIPG and NIPG DGM. R is the time spent renumbering the elements and factoring the global matrix. Computational times were evaluated on a Pentium III/1.266 *GH processor.*

	New formalism			NIPG			SIPG		
h	Error	Rate	CPU	Error	Rate	CPU	Error	Rate	CPU
L^2 error									
1/8	$8.95E-04$	–	0.94	$1.94E-02$	–	$0.87+R$	$1.89E-02$	–	$0.86+R$
1/16	$2.10E-04$	2.09	9.29	$4.62E-03$	2.07	$5.91+R$	$4.47E-03$	2.08	$6.18+R$
1/32	$5.16E-05$	2.02	119.8	$1.14E-03$	2.02	$71.25+R$	$1.10E-03$	2.02	$71.21+R$
1/64	$1.28E-05$	2.02	1855	$2.84E-04$	2.00	$1519+R$	$2.75E-04$	2.00	$1334+R$
L^∞ error									
1/8	$2.63E-03$	–	0.94	$3.84E-02$	–	$0.87+R$	$3.75E-02$	–	$0.86+R$
1/16	$4.63E-04$	2.50	9.29	$9.22E-03$	2.06	$5.91+R$	$8.93E-03$	2.07	$6.18+R$
1/32	$1.15E-04$	2.00	119.8	$2.28E-03$	2.02	$71.25+R$	$2.21E-03$	2.01	$71.21+R$
1/64	$2.86E-05$	2.00	1855	$5.69E-04$	2.00	$1519+R$	$5.50E-04$	2.00	$1334+R$

TABLE 6.5

Numerical L^2 errors, L^∞ errors, and convergence rates for second order time discretization with third order basis polynomials in the new formalism scheme (2.8) *applied to pure diffusion equation. Computations are done with a very small CFL condition so as to reduce the time discretization error.*

Second order time scheme with P_3 basis polynomials				
h	L^2 error	Rate	L^∞ error	Rate
1/2	$3.885E-03$	–	$4.517E-02$	–
1/4	$3.663E-04$	3.41	$4.212E-03$	3.42
1/8	$2.636E-05$	3.80	$2.668E-04$	3.98
1/16	$1.757E-06$	3.91	$1.769E-05$	3.91
1/32	$1.139E-07$	3.95	$1.127E-06$	3.97
1/64	$7.246E-09$	3.97	$7.191E-08$	3.97

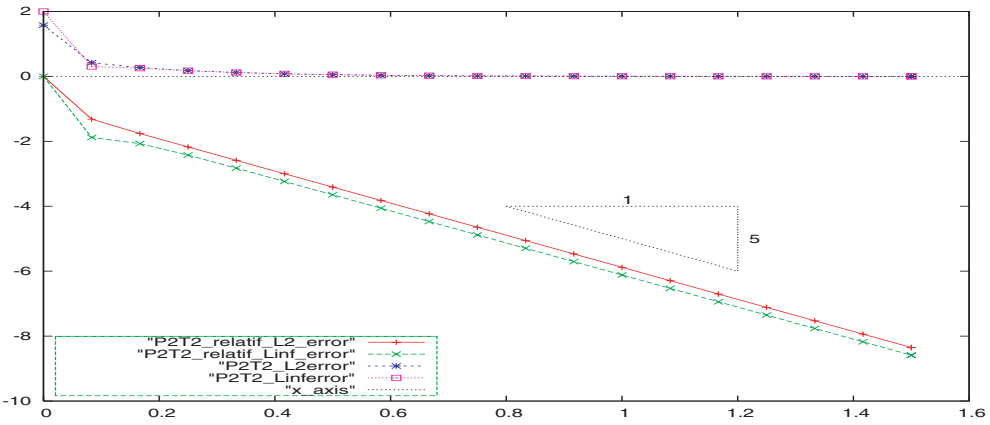


FIG. 6.1. *L^2 and L^∞ convergence errors at different times steps for the pure diffusion equation with nonhomogeneous boundary conditions. The computation is done using the new formalism with polynomials of order 2 in space and second order time discretization. The notation $P2T2$ stands for polynomials of order 2 in space ($P2$) with second order ($T2$) time discretization.*

in time. So if ever one desires to apply NIPG and SIPG to problems with variable coefficients, then R is to be multiplied by the number of iterations. Note that in our calculations, we have adapted the time step for NIPG and SIPG so that the number of time steps is already 10 times smaller for NIPG and SIPG. An even much greater time step is possible for NIPG and SIPG but at the price of a loss of accuracy of the discretization in time. In this case the new method, which is explicit, is much better than NIPG and SIPG.

6.3. An example with a nonhomogeneous Dirichlet boundary condition. Here is an example with a nonhomogeneous boundary Dirichlet condition. Instead of simply writing $\omega_k^+ = R_k^\alpha \omega_k^-$ (see Table 3.1), one uses

$$\omega_k^+ = R_k^\alpha \omega_k^- + \alpha_k (1 - R_k^\alpha) c_d \quad \text{for Dirichlet boundary condition } c = c_d,$$
$$\omega_k^+ = R_k^\alpha \omega_k^- + (1 + R_k^\alpha) g_N \quad \text{for Neumann boundary condition } K \frac{\partial}{\partial n} c = g_N.$$

We now take the same test case as above ($K \equiv 1, u \equiv 0$), with R.H.S. $f(t,x,y) = -4$, and a nonhomogeneous Dirichlet boundary condition $g_D(x,y) = x^2 + y^2$. We know that the limit of the exact solution as time tends to infinity is the solution of the stationary problem. That limit solution is in fact the function we have chosen as the Dirichlet boundary condition. In order to show that the new formalism handles nonhomogeneous boundary conditions, we have computed the solution with the initial condition taken to be $c(t = 0, x, y) = 0$ which is not related to the exact solution. The computational domain is $\Omega = (-1,1)^2$, meshed with nonuniform triangles (with 21 vertices per side) to show that the behavior of the formalism is well suited to the nonuniform mesh. Different steps of the solution are shown in Figure 6.2. Figure 6.1 shows the convergence to the exact solution as L^2 and L^∞ errors (measured by $||u(\infty) - u(t_n)||$) and relative L^2 and L^∞ errors (measured by $\log(||u(\infty) - u(t_n)||/||u(\infty)||)$) at every time step. Here $u(\infty)$ denotes the limit solution.

Observations. In Figure 6.2 the initial solution is zero, and as the time passes the convergence to the exact solution is achieved. It shows that boundary conditions of Dirichlet type are correctly discretized by this method.

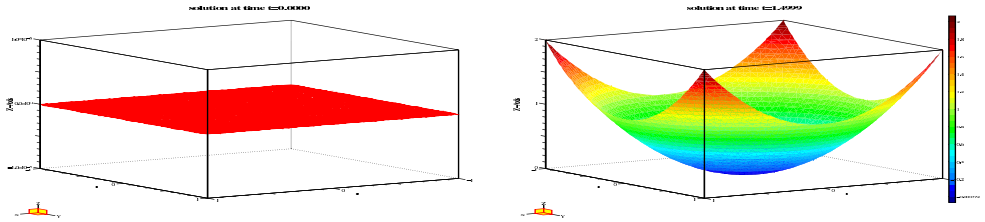


FIG. 6.2. *Asymptotic solution of the pure diffusion equation with nonhomogeneous boundary conditions, on a nonuniform mesh. On the left is the initial solution; on the right is the solution at* $t = 1.5s$. *The computations are done using the new formalism with second order polynomials in space and second order time discretization.*

6.4. A convection-diffusion example. In this section we consider the rotating pulse problem. The spatial domain is $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$, and the rotating field is imposed as $\mathbf{u} = (-4y, 4x)$. The initial condition and Dirichlet boundary condition are taken from the exact solution

$$c(t, x, y) = \frac{2\sigma^2}{2\sigma^2 + 4Kt} \exp(-\frac{(\bar{x} - x_c)^2 + (\bar{y} - y_c)^2}{2\sigma^2 + 4Kt},$$

where $\bar{x} = x\cos(4t) + y\sin(4t)$ and $\bar{y} = -x\sin(4t) + y\cos(4t)$. Here K is the constant diffusion coefficient. The R.H.S. is $f = 0$. This example was considered in [38], where only maxima and minima of many methods were listed. It is also used as a model equation in [4] to compare the L^2 error of a higher order DGM with various other methods on uniform rectangular meshes. Here we consider the same model problem on uniform triangular meshes, and we evaluate the L^2 and L^∞ errors and the convergence rate for the first and second order schemes presented in this paper. We take the same parameters as in [38, 4]: $K = 10^{-4}, x_c = 0.25, y_c = 0$, and $2\sigma^2 = 0.004$. The time interval for the simulation is $[0, T] = [0, \pi/4]$, which is the time for a half rotation. We begin with a uniform mesh of the domain made up of $8 \times 8 \times 2 = 138$ uniform triangles. We then successively refine the mesh and compute the L^2 and L^∞ errors e_h on the mesh of size h and the numerical convergence rates by the ratio $\ln(e_h/e_{h/2})/\ln(2)$. The time step is chosen so that the ratio $\Delta t/h$ is kept constant. The constant value is $1/82$ for first order time discretization and $1/164$ for the second order time scheme. The results obtained are recorded in Table 6.6.

Observations. This numerical test [38] is advection dominant in most parts of the domain and is diffusion dominant in the center of the domain. We solve it with the formalism presented in this paper with a constant ratio $\Delta t/h$. This constant ratio is obtained when we use the optimal parameter α (4.2) to determine the CFL condition (4.1). The second order in time scheme gives good results with higher order polynomials. Table 6.6 shows that using the constant ratio $\Delta t/h$, the convergence rate is greater than 2. Hence in second order time discretization, the time discretization error is small compared to the space discretization error for this test problem. This is a good feature when dealing with a coarse mesh. The second order time discretization is well suited for this kind of problem, where fine meshes are prohibitive due to memory management.

6.5. Conclusion driven from numerical experiments. The theoretical analysis is confirmed by numerical experiments. In particular we have L^2 stability and correct treatment of boundary conditions whatever the order of the polynomials is.

TABLE 6.6

Numerical L^2 errors, L^∞ errors, and convergence rates for first and second order time discretization schemes (2.4), (2.8) *applied to constant diffusion but variable velocity convection-diffusion equation. The convergence rates are obtained by computing the ratio* $\ln(e_h/e_{h/2})/\ln(2)$ *as the mesh is been refined. The polynomial space is of order* $0, 1, 2,$ *and* 3*, and the ratio* $\Delta t/h$ *is kept constant during the mesh refinement. The experimental order is* 1 *for first order in time integration and greater than* 2 *for second order in time integration.*

	First order in time				Second order in time			
h	L^2 error	Rate	L^∞ error	Rate	L^2 error	Rate	L^∞ error	Rate
P_0 basis polynomials								
1/8	$7.28E-02$	–	$3.93E-01$	–	$7.29E-02$	–	$3.93E-01$	–
1/16	$6.77E-02$	0.11	$6.92E-01$	−0.82	$6.78E-02$	0.10	$6.93E-01$	−0.82
1/32	$6.06E-02$	0.16	$7.50E-01$	−0.11	$6.09E-02$	0.16	$7.52E-01$	−0.12
1/64	$5.02E-02$	0.27	$6.77E-01$	0.15	$5.06E-02$	0.27	$6.81E-01$	0.14
1/128	$3.71E-02$	0.44	$5.36E-01$	0.34	$3.76E-02$	0.43	$5.41E-01$	0.33
P_1 basis polynomials								
1/8	$4.94E-02$	–	$5.89E-01$	–	$4.89E-02$	–	$5.76E-01$	–
1/16	$3.28E-02$	0.59	$4.49E-01$	0.39	$3.14E-02$	0.64	$4.30E-01$	0.42
1/32	$1.27E-02$	1.37	$1.86E-01$	1.27	$1.06E-02$	1.56	$1.57E-01$	1.46
1/64	$3.89E-03$	1.71	$5.64E-02$	1.72	$2.27E-03$	2.23	$3.26E-02$	2.27
1/128	$1.31E-03$	1.57	$1.89E-02$	1.58	$4.61E-04$	2.30	$6.09E-03$	2.42
P_2 basis polynomials								
1/8	$3.39E-02$	–	$4.77E-01$	–	$3.03E-02$	–	$4.29E-01$	–
1/16	$1.12E-02$	1.60	$1.55E-01$	1.62	$5.83E-03$	2.38	$7.43E-02$	2.53
1/32	$4.49E-03$	1.32	$6.55E-02$	1.24	$4.91E-04$	3.57	$1.30E-02$	2.51
1/64	$2.17E-03$	1.05	$3.11E-02$	1.07	$5.21E-05$	3.24	$2.32E-03$	2.49
1/128	$1.05E-03$	1.05	$1.48E-02$	1.07	$7.91E-06$	2.72	$3.15E-04$	2.88
P_3 basis polynomials								
1/8	$1.86E-02$	–	$2.53E-01$	–	$1.05E-02$	–	$1.31E-01$	–
1/16	$8.02E-03$	1.21	$1.26E-01$	1.01	$6.11E-04$	4.10	$1.99E-02$	2.72
1/32	$4.15E-03$	0.95	$6.87E-02$	0.87	$2.63E-05$	4.54	$2.25E-03$	3.14
1/64	$2.08E-03$	1.00	$3.20E-02$	1.10	$3.40E-06$	2.95	$1.29E-04$	4.12
1/128	$9.81E-04$	1.08	$1.49E-02$	1.10	$5.97E-07$	2.51	$9.24e-06$	3.80

REFERENCES

[1] V. AZINGER, C. DAWSON, B. COCKBURN, AND P. CASTILLO, *The local discontinuous Galerkin method for contaminant transport*, Adv. Wat. Res., 24 (2001), pp. 73–87.

[2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. DONATELLA MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] P. BASTIAN, *Higher order discontinuous Galerkin methods for flow and transport in porous media*, in Challenges in Scientific Computing—CISC 2002, Lecture Notes Comput. Sci. Eng. 35, Springer, Berlin, 2003, pp. 1–22.

[5] P. BASTIAN AND S. LANG, *Couplex Benchmark Computations with UG*, Computational Geosciences, 8 (2004), pp. 125–147.
Tech. Report 2002-31, IWR (SFB 359), Universität Heidelberg, Germany, 2002; Comput. Geosci., submitted.

[6] P. BASTIAN AND B. RIVIÈRE, *Superconvergence and H(*div*) Projection for discontinuous Galerkin methods*, Int. J. Numer. Methods Fluids, 42 (2003), pp. 1043–1057.

[7] C. E. BAUMAN AND J. T. ODEN, *A discontinuous hp finite element finite element method for convection diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[8] A. BOURGEAT, M. KERN, S. SCHUMACHER, AND J. TALANDIER, *The COUPLEX Test Cases: Nuclear Waste Disposal Simulation*, February, 2002.

[9] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 365–378.

[10] F. Brezzi, L. D. Marini, and E. Süli, *Discontinuous Galerkin methods for first-order hyperbolic problems*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1893–1903.

[11] G. Chavent and B. Cockburn, *The local projection P^0P^1 -discontinuous Galerkin finite element method for scalar conservative laws*, M2AN Math. Model. Anal. Numer., 23 (1989), pp. 565–592.

[12] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[13] H. Chen, *Local error estimates of mixed discontinuous Galerkin methods for elliptic problems*, J. Numer. Math., 12 (2004), pp. 1–21.

[14] H. Chen and Z. Chen, *Stability and convergence of mixed discontinuous finite element methods for second-order elliptic problems*, J. Numer. Math., 11 (2003), pp. 253–324.

[15] H. Chen, Z. Chen, and B. Li, *Numerical study of hp version of mixed discontinuous finite element methods for reaction diffusion problems:* $1D$ *case*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 525–553.

[16] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1975.

[17] B. Cockburn, *Discontinuous Galerkin Methods for Convection Dominated Problems*, School of Mathematics, University of Minnesota, Minneapolis, MN.

[18] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection-dominated problems*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Math. 1697, Springer, Berlin, 1998, pp. 151–268.

[19] B. Cockburn and C. W. Shu, *The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws*, M2AN Math. Model. Anal. Numer., 25 (1991), pp. 337–361.

[20] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[21] B. Cockburn and C. W. Shu, *TVB Runge Kutta local projection discontinuous Galerkin finite element method for conservative laws* II: *General frame-work*, Math. Comp., 52 (1989), pp. 411–435.

[22] B. Cockburn and C. W. Shu, *TVB Runge Kutta local projection discontinuous Galerkin finite element method for conservative laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[23] B. Cockburn and C. W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservative laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[24] B. Cockburn, G. E. Karniadakis, and C. W. Shu, *Discontinuous Galerkin Methods, Theory, Computation and Applications*, Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, Berlin, 2000.

[25] M. Crouzeix and A. L. Mignot, *Analyse numérique des équations différentielles*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1984.

[26] S. Del Pino and O. Pironneau, *Asymptotic analysis and layer decomposition for the couplex exercise*, Comput. Geosci., 8 (2004), pp. 149–162.

[27] H. Hoteit, *Simulation d'écoulements et de transports de polluants en milieu poreux: Apllication à la modélisation de la sureté des dépôts de déchets radioactifs*, Ph.D. thesis, Université de Rennes 1, Rennes, France, 2002.

[28] T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson, *A comparison of discontinuous and continuous Galerkin methods based on error estimates, conservation, robustness and efficiency*, in Discontinuous Galerkin Methods, Lecture Notes in Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 135–146.

[29] W. Hundsdorfer and J. Jaffré, *Implicit-explicit time stepping with spatial discontinuous finite elements*, Appl. Numer. Math., 445 (2003), pp. 231–254.

[30] L. D. Marini, *A survey of DG methods for elliptic problems*, in Proceedings ENUMATH 2001, F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Springer, Berlin, 2003, pp. 805–814.

[31] S. Prudhomme, F. Pascal, J. T. Oden, and A. Romkes, *Review of A Priori Error Estimation for Discontinuous Galerkin Methods*, TICAM Report 00-27, University of Texas at Austin, Austin, TX, October 17, 2000.

[32] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[33] B. Rivière, *Discontinuous Galerkin Methods for Solving the Miscible Displacement Problem in Porous Media*, Ph.D. thesis, University of Texas at Austin, Austin, TX, 2000.

[34] B. Rivière and M. F. Wheeler, *A discontinuous Galerkin method applied to nonlinear Parabolic equations*, in Discontinuous Galerkin Methods, Lecture Notes in Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 231–244.

[35] B. Rivière, M. F. Wheeler, and V. Girault *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems* I, Comput. Geosci., 3 (1999), pp. 337–360.

[36] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.

[37] S. Sun, *Discontinuous Galerkin Methods for Reactive Transport in Porous Media*, Ph.D. thesis, The University of Texas at Austin, Austin, TX, 2003.

[38] H. Wang, H. K. Dahle, R. E. Ewing, M. S. Espedal, R. C. Sharpley, and S. Man, *An ELLAM scheme for advection-diffusion equations in two dimensions*, SIAM J. Sci. Comput., 20 (1999), pp. 2160–2194.

[39] L. Yin, A. Acharya, N. Sobh, R. B. Haber, and D. A. Tortorelli, *A space-time discontinuous Galerkin method for elastodynamic analysis*, in Discontinuous Galerkin Methods, Lecture Notes in Comput. Sci. Eng. 11, Springer, Berlin, 1999, pp. 459–464.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2270–2296

ERROR ANALYSIS OF COARSE-GRAINING FOR STOCHASTIC LATTICE DYNAMICS*

MARKOS A. KATSOULAKIS†, PETR PLECHÁČ‡, AND ALEXANDROS SOPASAKIS†

Abstract. The coarse-grained Monte Carlo (CGMC) algorithm was originally proposed in the series of works [M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos, *J. Comput. Phys.*, 186 (2003), pp. 250–278; M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos, *Proc. Natl. Acad. Sci. USA*, 100 (2003), pp. 782–787; M. A. Katsoulakis and D. G. Vlachos, *J. Chem. Phys.,* 119 (2003), pp. 9412–9427]. In this paper we further investigate the approximation properties of the coarse-graining procedure and provide both analytical and numerical evidence that the hierarchy of the coarse models is built in a systematic way that allows for error control in both transient and long-time simulations. We demonstrate that the numerical accuracy of the CGMC algorithm as an approximation of stochastic lattice spin flip dynamics is of order two in terms of the coarse-graining ratio and that the natural small parameter is the coarse-graining ratio over the range of particle/particle interactions. The error estimate is shown to hold in the weak convergence sense. We employ the derived analytical results to guide CGMC algorithms and demonstrate a CPU speed-up in demanding computational regimes that involve nucleation, phase transitions, and metastability.

Key words. coarse-grained stochastic processes, Monte Carlo simulations, birth-death process, detailed balance, Arrhenius dynamics, Gibbs measures, weak error estimates, microscopic reconstruction

AMS subject classifications. 65C05, 65C20, 82C20, 82C26

DOI. 10.1137/050637339

1. Introduction.

Microscopic computational models for complex systems such as molecular dynamics (MD) and Monte Carlo (MC) algorithms are typically formulated in terms of simple rules describing interactions between individual particles or spin variables. The large number of variables and even larger number of interactions between them present the principal limitation for efficient simulations. Another restricting factor is illustrated by the essentially sequential nature of approximating the time evolution in particle systems that yields a substantial slowdown in the resolution of dynamics, especially in metastable regimes.

In [19, 20, 23] the authors started developing systematic mathematical strategies for the coarse-graining of microscopic models, focusing on the paradigm of stochastic lattice dynamics and the corresponding MC simulators. In principle, coarse-grained models are expected to have fewer observables than the original microscopic system, making them computationally more efficient than the direct numerical simulations. In these papers a hierarchy of coarse-grained stochastic models—referred to as coarse-grained MC (CGMC)—was derived from the microscopic rules through a stochastic closure argument. The CGMC hierarchy is reminiscent of multiresolution analysis approaches to the discretization of operators [3], spanning length/time scales from the microscopic to the mesoscopic. The resulting *stochastic coarse-grained processes*

*Received by the editors August 1, 2005; accepted for publication (in revised form) May 8, 2006; published electronically November 24, 2006.

http://www.siam.org/journals/sinum/44-6/63733.html

†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003–9305 (markos@math.umass.edu, sopas@math.umass.edu). The research of the first author was partially supported by NSF-DMS-0413864 and NSF-ITR-0219211.

‡Mathematics Institute, The University of Warwick, Coventry, CV4 7AL, United Kingdom (plechac@maths.warwick.ac.uk). The research of this author was partially supported by NSF-DMS-0303565.

involve Markovian birth-death and generalized exclusion processes and their combinations, and as demonstrated in [19, 20, 23], they share the same ergodic properties with their microscopic counterparts. The full hierarchy of the coarse-grained stochastic dynamics satisfies detailed balance relations and, as a result, not only yields self-consistent random fluctuation mechanism, but are consistent with the underlying microscopic fluctuations and the unresolved degrees of freedom. From the computational complexity perspective, a comparison of CGMC with conventional MC methods for the same real time shows [19] that the CPU time can decrease approximately as $O(1/q^2)$ or faster, where q is the number of aggregated lattice sites (referred to as the level of coarse-graining), as demonstrated for spin-flip lattice dynamics. Thus, while for macroscopic size systems in the millimeter length scale or larger, microscopic MC simulations are impractical on a single processor, the computational savings of CGMC make it a suitable tool capable of capturing large scale features, while retaining microscopic information on intermolecular forces and particle fluctuations.

In the recent paper [22] the authors rigorously analyzed CGMC models as approximations of conventional MC in *nonequilibrium*, by estimating the *information loss* between microscopic and coarse-grained adsorption/desorption lattice dynamics. In analogy to the numerical analysis for PDEs, an error analysis was carried out between the *exact microscopic process* $\{\sigma_t\}_{t\geq 0}$ and the *approximating coarse-grained process* $\{\eta_t\}_{t\geq 0}$. The key step in this direction was to use, as a quantitative measure for the loss of information in the coarse-graining from finer to coarser scales, the information-theoretic concept of the *relative entropy* between probability measures, [9]. Such relative entropy estimates give a first mathematical reasoning for the parameter regimes, i.e., the degree of coarse-graining versus the interaction range, for which CGMC is expected to give errors within a given tolerance. In this paper using the rigorous results in [22] as a starting point, we focus on carrying out a detailed numerical analysis of the error propagation for spin-flip lattice dynamics. Due to the numerical intractability of the relative entropy for a large particle system, we employ in the numerical error calculations *targeted* coarse observables. The latter point of view necessitates in the use of a weak convergence framework for the study of the error between CGMC and direct numerical simulations of the stochastic lattice dynamics. We demonstrate that the numerical accuracy of the CGMC algorithm is of order two in terms of the ratio of the coarse-graining over the range of particle/particle interactions. We also refer to recent work in [21] on weak error estimates between microscopic MC algorithms and therein derived SDE approximations. Further details about a priori estimates for weak convergence of approximations to SDEs can be found in [2, 34, 26]. Related a posteriori estimates are discussed in [33]. We further employ the derived analytical results to guide CGMC algorithms and we demonstrate a CPU speed-up in demanding computational regimes that involve nucleation, phase transitions, and metastability. We demonstrate computationally that CGMC probes efficiently the energy landscape, yielding *spatial pathwise* agreement with the underlying microscopic lattice dynamics, at least for fairly long but still finite interactions.

The mathematical difficulty in carrying out our error estimates primarily rests with the fact that the projection of the exact microscopic process $\{\sigma_t\}_{t\geq 0}$ on the coarse grid denoted by $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ needs to be compared with the derived approximating process $\{\eta_t\}_{t\geq 0}$. However, $\mathbf{T}\sigma_t$ does not necessarily define a Markov process, while the approximating process $\{\eta_t\}_{t\geq 0}$ is constructed as a Markov process defined by (3.5). To circumvent this technical difficulty the authors in [22] suggested constructing an auxiliary Markov process $\{\gamma_t\}_{t\geq 0}$ as an intermediate step in the estimation of the relative entropy between $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$. We adopt the same

strategy here in order to make a comparison between observables which depend on Markovian processes $\{\sigma_t\}_{t\geq 0}$ and $\{\gamma_t\}_{t\geq 0}$. The *reconstructed microscopic Markov process* $\{\gamma_t\}_{t\geq 0}$ can be directly synthesized from the coarse-grained process $\{\eta_t\}_{t\geq 0}$, and these two processes induce the same probability measure on the coarse-grained path space. Such reconstruction is an inverse procedure to the projection from fine to coarse configuration space and a simple choice of a reconstruction is to distribute particles uniformly on the coarse cells. This action enforces a local equilibrium in each coarse cell, parametrized by the coarse variables. From the technical point of view the reconstruction allows for explicit calculations of averaged quantities in each coarse cell and is crucial in obtaining the second order accuracy of our methods. It is conceivable that the synthetic process $\{\gamma_t\}_{t\geq 0}$ can be used not only as a technical tool but also as a systematic procedure for reconstructing the microscopic process $\{\sigma_t\}_{t\geq 0}$ for the purpose of model refinement or adaptivity since, as shown in Theorem 4.7, the reconstruction is done under rigorous error estimates.

The CGMC algorithms discussed here are related to a number of methods involving coarse-graining at various levels; for instance, fast summation techniques, computational renormalization and simulation, and multiscale computational methods for stochastic systems. One of the sources of the computational complexity of molecular simulations arises in the calculation of particle/particle interactions, especially in the case where long range forces are relevant. The evaluation cost of such pairwise interactions can be significantly reduced by applying well-controlled approximation schemes and/or a hierarchical decomposition of the computation. Typically, once the interaction terms are computed with one of these fast summation methods, they are entered in the microscopic algorithm where a simulation with a large number of individually tracked particles still has to be carried out. The point of view adopted by CGMC is related to these methods in the sense that the interaction potential or operator is approximated in terms of a truncated multiresolution decomposition within a given tolerance. The CGMC is subsequently defined at the coarse level specified by the truncation of the decomposition. However, a notable difference is that CGMC models track much fewer coarse observables instead of simulating every individual particle. The equilibrium setup of CGMC is essentially given by the renormalized Hamiltonian after a single iteration in the renormalization group flow. It is not surprising that such an approach, when applied to near critical temperature simulations, has many limitations. For example, in the nearest-neighbor Ising-type models this fact is manifested in the aforementioned error estimates and the comparative simulations in [19]. On the other hand the focus of CGMC is dynamic simulations usually coupled to a macroscopic system (see, for instance, the hybrid systems in [35, 18]), where criticality may not be as important due to the presence of a time-varying external field. Nevertheless, further corrections to the CGMC dynamics from the renormalization group flow given by RGMC and multigrid MC methods [4, 6, 12] can improve the order of convergence of the CGMC. We refer to [17] for higher order accurate CGMC methods based on cluster expansions, where the coarse-graining procedure described here is the model around which a cluster expansion is carried out with controlled errors. As explained in section 4, in the sense that the CGMC method is of order two accurate in terms of the small parameter q/L, where L is the radius of interaction.

As is the case with most asymptotic results, from a practical point of view a small parameter q/L does not need to be "very small" in order for the asymptotics to work. The case here is no exception even in the phase transitions regime. This observation is further amplified by the higher order estimate $(q/L)^2$ in Theorem 4.7. A typical example of long range interactions is the electrostatic potential; however, the methods

proposed here cannot (yet) handle the singular part of the potential which is close to the origin. However, they can easily handle, with error estimates, the slowly decaying part of the potential (away from the origin), which is a primary computational hurdle for direct numerical simulations with standard MC methods. On the other hand the proposed methods are expected to work well when local averaging gives a good error control in the potentials, as in Lemma 3.2. Such examples include Morse-type potentials as well as oscillating indirect exchange potentials of RKKY type [30] arising in magnetic materials. Furthermore some intermediate-range potentials can be obtained from detailed experiments. Such an example arising in surface processes is described in [31], where a potential with 36 neighbors is obtained. In such a setup the CGMC method would be expected to apply.

In recent years there has been a growing interest in developing and analyzing coarse-graining methods for the purpose of modelling and simulation across scales. Such systems arise in a broad spectrum of scientific disciplines ranging from materials science to macromolecular dynamics, to epidemiology, and to atmosphere/ocean science. Various coarse-graining approaches may yield explicitly derived stochastic coarse models using different coarse approximations, e.g., [13, 15, 28, 32, 8, 36], or can be statistics-based [29] or may rely on on-fly simulations, e.g., the equation-free method [24], the heterogeneous multiscale method [11], or multiscale finite-element methods [14]. A systematic approach to the upscaling of stochastic systems has also been proposed from the multilevel perspective in [7, 1, 5], where the authors proposed algorithms for efficient multiscale simulations using MC methods. Other coarse-graining techniques in the polymer science literature include the bond fluctuation model and its variants [27]. Such coarse-graining methodologies often rely on parametrization, hence at different conditions (e.g., temperature, density, composition) coarse potentials need to be re-parametrized [29].

2. Microscopic lattice models. The presented analysis applies to the class of Ising-type lattice systems. For the sake of simplicity we assume that the computational domain is defined as the discrete periodic lattice $\Lambda_N = \frac{1}{n}\mathbb{Z}^d \cap \mathbb{T}$ which represents discretion of the d-dimensional torus $\mathbb{T} = [0,1)^d$ and d denotes the spatial dimension. We restrict presentation of the results to $d = 1$; nevertheless higher dimensional cases are obtained without significant changes. However, the algorithms can also be implemented on bounded domains with usual boundary conditions. The number of lattice sites $N = n^d$ is fixed. The microscopic degrees of freedom or the microscopic order parameter is given by the spin-like variable $\sigma(x)$ defined at each site $x \in \Lambda_N$. In this paper we discuss only the case of discrete spin variables, i.e., $\sigma(x) \in \Sigma$ with $\Sigma = \{-1, 1\}$, $\Sigma = \{0, 1\}$ (Ising model), or $\Sigma = \{0, 1, \dots s\}$ (Potts models). The case of the spin variable belonging to a compact Riemannien manifold, e.g., $\Sigma = \mathbb{S}^2$ (Heisenberg model), $\Sigma = \mathrm{SU}(2)$ (matrix model), will be studied elsewhere. We denote by $\sigma = \{\sigma(x) \,|\, x \in \Lambda_N\}$ a configuration of spins on the lattice, i.e., an element of the configuration space $\mathcal{S}_N = \Sigma^{\Lambda_N}$. The interactions between spins at a given configuration σ are defined by the microscopic Hamiltonian

$$H(\sigma) = -\frac{1}{2}\sum_{x\in\Lambda_N}\sum_{y\neq x} J(x-y)\sigma(x)\sigma(y) + \sum_{x\in\Lambda_N} h(x)\sigma(x)\,, \tag{2.1}$$

where $h(x)$ denotes the external field at the site x. The two-body interparticle potential J accounts for interactions between individual spins. We consider the class of

potentials with a finite range interaction length L

$$J(x-y) = \frac{1}{L^d} V\left(\frac{n}{L}|x-y|\right), \quad x, y \in \Lambda_N, \tag{2.2}$$

$$V : \mathbb{R} \to \mathbb{R}, \quad V(r) = V(-r), \quad V(r) = 0, \quad \text{if } |r| \geq 1. \tag{2.3}$$

We impose additional assumptions on V which allow us to derive explicit error estimates:

$$V \text{ is smooth on } \mathbb{R} \setminus \{0\}, \tag{2.4}$$

$$\int_{\mathbb{R}} |V(r)|\, dr < \infty, \text{ and } \int_{\mathbb{R}} |\partial_r V(r)|\, dr < \infty. \tag{2.5}$$

Note that the summability condition for V guarantees that the potential J is also summable due to the scaling factor. Hence the Hamiltonian is well defined even for $N, L \to \infty$. The canonical equilibrium state is given in terms of the Gibbs measure

$$\mu_{N,\beta}(d\sigma) = \frac{1}{Z_{N,\beta}} e^{-\beta H(\sigma)} \mathrm{P}_N(d\sigma), \qquad Z_{N,\beta} = \int_{\mathcal{S}_N} e^{-\beta H(\sigma)} \mathrm{P}_N(d\sigma), \tag{2.6}$$

where $\mathrm{P}_N(d\sigma) = \prod_{x \in \Lambda_N} \rho(d\sigma(x))$ is the product measure on $\mathcal{S}_N$ and the spins $\sigma(x)$ are independent identically distributed (i.i.d.) random variables with the common distribution ρ. For example, in the Ising model the prior distribution on $\Sigma = \{0, 1\}$ would typically be $\rho(0) = \rho(1) = 1/2$.

The microscopic dynamics are defined as a continuous-time jump Markov process that defines a change of the spin $\sigma(x)$ with the probability $c(x, \sigma; \xi)\Delta t$ over the time interval $[t, t + \Delta t]$. The function $c : \Lambda_N \times \mathcal{S}_N \times \Sigma \to \mathbb{R}$ is called a rate of the process. The jump process $\{\sigma_t\}_{t \geq 0}$ is constructed in the following way: suppose that at the time t the configuration is σ_t, then the probability of changing the spin at the site $x \in \Lambda_N$ spontaneously from $\sigma_t(x)$ to a new value $\xi \in \Sigma$ over the time interval $[t, t+\Delta t]$ is $c(x, \sigma; \xi)\Delta t + O(\Delta t^2)$. We denote the resulting configuration by $\sigma^{x,\xi}$. In the case of the Ising-type state space and spin-flip dynamics we omit ξ in this notation. The generator $\mathcal{L} : L^\infty(\mathcal{S}_N) \to L^\infty(\mathcal{S}_N)$ of the Markov process acting on a bounded test function $\phi \in L^\infty(\mathcal{S}_N)$ defined on the space of configurations is given by

$$(\mathcal{L}\phi)(\sigma) = \sum_{x \in \Lambda_N} \int_\Sigma c(x, \sigma; \xi) \left(\phi(\sigma^{x,\xi}) - \phi(\sigma)\right) d\xi. \tag{2.7}$$

The evolution of an observable (a test function) ϕ is given by

$$\frac{d}{dt} \mathbb{E}\left[\phi(\sigma_t)\right] = \mathbb{E}\left[\mathcal{L}\phi(\sigma_t)\right], \tag{2.8}$$

where the expectation operator $\mathbb{E}[.]$ is, with respect to a measure, conditioned to the initial configuration $\sigma_{t=0} = \sigma_0$. We require that the dynamics are of a relaxation type such that the invariant measure of this Markov process is the Gibbs measure (2.6). The sufficient condition is known as *detailed balance* (DB) and it imposes a condition on the form of the rate

$$c(x, \sigma; \xi) e^{-\beta H(\sigma)} = c(x, \sigma^{x,\xi}; \sigma(x)) e^{-\beta H(\sigma^{x,\xi})}. \tag{2.9}$$

This condition has a simple interpretation: $c(x, \sigma; \xi)$ is the rate of converting $\sigma(x)$ to the value ξ while $c(x, \sigma^{x,\xi}; \sigma(x))$ is the rate of changing the spin with the value ξ at

the site x back to $\sigma(x)$. The widely used class of Metropolis-type dynamics satisfies (2.9) and has the rate given by

$$c(x,\sigma;\xi) = G(\beta\Delta_{x,\xi}H(\sigma))\,, \text{ where } \Delta_{x,\xi}H(\sigma) = H(\sigma^{x,\xi}) - H(\sigma), \tag{2.10}$$

where G is a continuous function satisfying: $G(r) = G(-r)e^{-r}$ for all $r \in \mathbb{R}$. The most common choices in physics simulations are $G(r) = \frac{1}{1+e^r}$ (Glauber dynamics), $G(r) = e^{-[r]_+}$ (Metropolis dynamics), with $[r]_+ = r$ if $r \geq 0$ and $= 0$ otherwise, or $G(r) = e^{-r/2}$. Such dynamics are often used as samplers from the canonical equilibrium Gibbs measure. However, the kinetic MC method is also used for simulations of nonequilibrium processes. The dynamics in such a case are known as *Arrhenius dynamics*, whose rates are usually derived from transition state theory or obtained from molecular dynamics simulations.

To avoid unnecessary generality we restrict the description to the Ising-type model with $\Sigma = \{0,1\}$ used for modeling adsorption/desorption processes. We also omit ξ in the notation. The Arrhenius rate is defined as follows:

$$c(x,\sigma) = \begin{cases} d_0 & \text{if } \sigma(x) = 0, \\ d_0 e^{-\beta U(x,\sigma)} & \text{if } \sigma(x) = 1, \end{cases} \tag{2.11}$$

where

$$U(x,\sigma) = \sum_{y\in\Lambda_N\,, y\neq x} J(x-y)\sigma(y) - h(x)\,. \tag{2.12}$$

Furthermore, the spin-flip rule is given by

$$\sigma^x(y) = \begin{cases} 1-\sigma(x) & \text{if } y = x, \\ \sigma(y) & \text{if } y \neq x. \end{cases}$$

With the introduced notation the coarse-graining algorithm can be described as an *approximation* of the microscopic dynamics, i.e., of the process $\{\sigma_t\}_{t\geq 0}$ by a coarse-grained process $\{\eta_t\}_{t\geq 0}$, where the approximation is done in a controlled way. We are interested not only in the approximation of the invariant measure $\mu_{N,\beta}(d\sigma)$ (see (2.6)), but also in the approximation of the measure on the path space.

3. Approximation of the coarse-grained process. The coarse-graining is defined in a geometric way by introducing the coarse-grained observables as block-spin variables. This approach follows the standard procedure of real-space renormalization; see, for example, [16]. We remark that although we introduce block-spins our aim is not to approximate the renormalization group flow (either on the space of Gibbs measures or on the path space), but rather to find an approximation that is constructed with low computational cost and with controlled and computable error estimates.

In general terms we define the coarse-graining operator $\mathbf{T} : \mathcal{S}_N \to \mathcal{S}^c_{M,q}$, where the coarse configuration space $\mathcal{S}^c_{M,q}$ is defined on the coarse lattice Λ^c_M, and with the new state space Σ^c, i.e., $\mathcal{S}^c_{M,q} = (\Sigma^c)^{\Lambda^c_M}$. The coarse configuration $\eta = \mathbf{T}\sigma \in \mathcal{S}^c_{M,q}$ is defined on a smaller lattice with M lattice sites and with the coarse state space Σ^c for the new lattice spins $\eta(k)$. The parameter q defines the coarse-graining ratio. The operator $\mathbf{T}$ induces an operator $\mathbf{T}_*$ on the space of probability measures

$$\mathbf{T}_* : \mathcal{P}(\mathcal{S}_N) \to \mathcal{P}(\mathcal{S}^c_{M,q})\,, \quad \mu(\sigma) \mapsto \mu^c(\eta) := \mu\{\sigma \in \mathcal{S}_N \,|\, \mathbf{T}\sigma = \eta\}\,.$$

Ising-type spins. To be more specific we analyze the following case of Ising spin-flip dynamics $\mathcal{S}_N = \{0,1\}^{\Lambda_N}$. Each coarse lattice site $k \in \Lambda_M^c$ represents a cube C_k that contains q sites of the microscopic lattice Λ_N. The projection operator defines the block-spin at the coarse site k to be

$$(\mathbf{T}\sigma)(k) := \sum_{x \in C_k} \sigma(x)\,. \tag{3.1}$$

If the dimension d of the lattice is greater than one, we understand k and x as multi-indices $k = (k_1, \ldots, k_d)$, and we index the corresponding lattice sites in the natural order. Choosing the projection operator in this way defines the coarse state space as $\Sigma^c = \{0, 1, \ldots, q\}$. Given the Markov process $(\{\sigma_t\}_{t\geq 0}, \mathcal{L})$ with the generator $\mathcal{L}$ we obtain a coarse-grained process $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ which is *not*, in general, a Markov process. From the computational point of view this may cause significant difficulties should sampling of such a process be implemented on the computer. Therefore we derive an *approximating* Markov process $(\{\eta_t\}_{t\geq 0}, \bar{\mathcal{L}}^c)$ which can be easily implemented once its generator is given explicitly.

For the model Ising system the projected generator of the coarse-grained process $\{\eta_t\}_{t\geq 0}$ can be evaluated explicitly by rearranging the summations on the lattice Λ_N; given the microscopic state σ and corresponding coarse state $\eta = \mathbf{T}\sigma$,

$$\begin{aligned} \mathcal{L}\psi(\mathbf{T}\sigma) = &\sum_{k\in\Lambda_M^c} \left[\sum_{x\in C_k} c(x,\sigma)(1-\sigma(x))\right] [\psi(\eta+\delta_k) - \psi(\eta)] \\ &+ \sum_{k\in\Lambda_M^c} \left[\sum_{x\in C_k} c(x,\sigma)\sigma(x)\right] [\psi(\eta-\delta_k) - \psi(\eta)]\,. \end{aligned} \tag{3.2}$$

The configuration δ_k defined on the coarse state space is equal to zero at all sites except the site $k \in \Lambda_M^c$ where it is equal 1, i.e., $\delta_k(j) = 1$ for $j = k$ and $= 0$ otherwise. We see from the formula (3.2) that the exact generator for the coarse process can be written in the form

$$\mathcal{L}^c\psi(\eta) = \sum_{k\in\Lambda_M^c} c_a(k)\,[\psi(\eta+\delta_k) - \psi(\eta)] + \sum_{k\in\Lambda_M^c} c_d(k)\,[\psi(\eta-\delta_k) - \psi(\eta)]\,, \tag{3.3}$$

where the new rates

$$c_a(k) = \sum_{x\in C_k} c(x,\sigma)(1-\sigma(x))\,, \qquad c_d(k) = \sum_{x\in C_k} c(x,\sigma)\sigma(x) \tag{3.4}$$

correspond to the adsorption and desorption processes. In this form the rates depend on the microscopic configuration σ and not on the coarse random variable $\mathbf{T}\sigma$. Therefore, it is reasonable to propose an approximating Markov process, which for the case of desorption/adsorption is a *birth-death* process $\{\eta_t\}_{t\geq 0}$ defined on the state space $\Sigma^c = \{0, 1, \ldots q\}$. This process is defined by the generator $\bar{\mathcal{L}}^c$ of the form (3.3) where the rates c_a and c_d are replaced by approximate rates

$$\bar{c}_a(k,\eta) = d_0(q - \eta(k))\,, \qquad \bar{c}_d(k,\eta) = d_0\eta(k)e^{-\beta\bar{U}(k,\eta)}\,. \tag{3.5}$$

For details we refer to [19]. The new rates have a simple interpretation in terms of fluctuations on each cell: $\bar{c}_a(k,\eta)$ describes the rate with which the coarse variable

$\eta(k)$ is increased by one (i.e., adsorption of a single particle in the coarse cell C_k) and $\bar{c}_d(k,\eta)$ defines the rate with which it is decreased by one (desorption in C_k). The new interaction potential $\bar{U}(\eta)$ represents the approximation of the original interaction $U(\sigma)$.

DEFINITION 3.1. *We define the approximation $\bar{U}(k,\eta)$ of the potential $U(x,\sigma)$, (2.12), at the coarse level*

$$\bar{U}(k,\eta) = \sum_{\substack{l\in\Lambda_M^c \\ l\neq k}} \bar{J}(k,l)\eta(l) + \bar{J}(0,0)(\eta(k)-1) - \bar{h}(k)\,. \tag{3.6}$$

The coarse-grained interaction potential $\bar{J}$ is computed as the average of the pairwise interactions between microscopic spins between the coarse cells C_k and C_l,

$$\bar{J}(k,l) = \frac{1}{q^2}\sum_{x\in C_k}\sum_{y\in C_l} J(x-y) \quad \textit{for all } k,l\in\Lambda_M^c, \textit{ such that } k\neq l, \textit{ and} \tag{3.7}$$

$$\bar{J}(k,k) \equiv J(0,0) = \frac{1}{q(q-1)}\sum_{x\in C_k}\sum_{\substack{y\in C_k \\ y\neq x}} J(x-y)\,. \tag{3.8}$$

The error estimate for the projection follows directly from the assumptions on the regularity of J (or V) (2.4)–(2.5). We state it as a separate lemma without the proof, which is obtained by applying the Taylor expansion of the potential J.

LEMMA 3.2. *Assume that J satisfies (2.4)–(2.5); then the coarse-grained interaction potential $\bar{J}$ at the coarse-graining level q approximates the potential J with the error*

$$|J(x-y) - \bar{J}(k,l)| \le \frac{1}{L}c_d \sup_{\substack{x'\in C_k \\ y'\in C_l}} ||\nabla V(x'-y')|| \le \mathrm{O}\left(\frac{q}{L^2}\right), \tag{3.9}$$

$$|J(x-y) - \bar{J}(0,0)| \le \frac{1}{L}c_d \sup_{\substack{x',y'\in C_k \\ y'\neq x'}} ||\nabla V(x'-y')|| \le \mathrm{O}\left(\frac{q}{L^2}\right), \tag{3.10}$$

where $c_d = \max_{k\in\Lambda_M^c}\{\mathrm{diam}\,(C_k)\}$.

From Lemma 3.2 we derive the error bound for the approximation of the coarse-grained potential $\bar{U}$. Note that in the definition of U the principle contribution to the summation involves interactions within the interaction range L and thus we have the following estimate.

COROLLARY 3.3. *The microscopic potential $U(x,\sigma)$ is approximated by $\bar{U}(k,\eta)$, with the error*

$$\Delta_{q,N}(\bar{U},U) \equiv |\bar{U}(k,\mathbf{T}\sigma) - U(x,\sigma)| = \mathrm{O}\left(\frac{q}{L}\right) \quad \textit{for all } x\in C_k. \tag{3.11}$$

Note that this approximation represents the direct projection of the interaction kernel J on the coarse space and the contribution from fine scales are neglected. This procedure differs from the renormalization group approach where fluctuations from the fine scales contribute to the transformed Hamiltonian. However, in the case of finite-range interaction kernels J treated here, the above projection yields approximation of the order $O(q/L)^2$ as we discuss in the next section. The coarse

interaction Hamiltonian is then given explicitly in terms of $\bar{J}$ and $\bar{h}$ as

$$\bar{H}(\eta) = -\frac{1}{2}\sum_{l\in\Lambda_M^c}\sum_{k\neq l}\bar{J}(k,l)\eta(k)\eta(l) - \frac{1}{2}\bar{J}(0,0)\sum_{l\in\Lambda_M^c}\eta(l)(\eta(l)-1) + \sum_{l\in\Lambda_M^c}\bar{h}(l)\eta(l)\,. \tag{3.12}$$

A direct calculation by verifying the condition of detailed balance [19],

$$\begin{aligned}\bar{c}_a(k,\eta)\mu_{M,q,\beta}(\eta) &= \bar{c}_d(k,\eta+\delta_k)\mu_{M,q,\beta}(\eta+\delta_k)\,,\\ \bar{c}_d(k,\eta)\mu_{M,q,\beta}(\eta) &= \bar{c}_a(k,\eta-\delta_k)\mu_{M,q,\beta}(\eta-\delta_k)\,,\end{aligned}$$

shows that the invariant measure of the Markov process $\{\eta_t\}_{t\geq 0}$ generated by $\bar{\mathcal{L}}^c$ is again a canonical Gibbs measure,

$$\mu^c_{M,q,\beta}(d\eta) = \frac{1}{Z_{M,q,\beta}}e^{-\beta\bar{H}(\eta)}\mathrm{P}_{M,q}(d\eta)\,, \tag{3.13}$$

where the product measure $\mathrm{P}_{M,q}(d\eta)$ is the coarse-grained prior distribution. Note that the prior distribution is altered by the coarse-graining procedure and different projection operators $\mathbf{T}$ may yield prior distributions that are computationally intractable. For example, the coarse-grained prior arising from the uniform microscopic prior ($\rho(0)=\rho(1)=1/2$) is the binomial distribution corresponding to q independent sites:

$$\mathrm{P}_{M,q}(d\eta) = \prod_{k\in\Lambda_M^c}\rho_q^c(d\eta(k))\,,\qquad \rho_q^c(\eta(k)=p) = \frac{q!}{p!(q-p)!}\left(\frac{1}{2}\right)^q\,.$$

The coarse-graining procedure described here satisfies basic criteria imposed on an approximating process:

(i) Error control on a finite-time interval $[0,T]$. In particular, the derived coarse-grained stochastic process $\{\eta_t\}_{t\geq 0}$ approximates a prespecified observable on a finite-time interval $[0,T]$, e.g., (3.1). In particular, time-dependent error estimates such as (4.2) can rigorously demonstrate that the process $\{\eta_t\}_{t\geq 0}$ keeps track of fluctuations from the microscopic level. Consequently expected values of certain path-dependent (global) quantities can be properly estimated. We characterize approximation properties of $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ by $\{\eta_t\}_{t\geq 0}$ using a suitable probability metric on the path space.

(ii) Approximation of the invariant (equilibrium) measure. The invariant measure $\mu^c_{M,q,\beta}(d\eta)$ for the process $\{\eta_t\}_{t\geq 0}$ defined on $\mathcal{S}^c_{M,q}$ is close, in a suitable probability metric, to the projection of the microscopic measure $\mathbf{T}_*(\mu_{N,\beta}(d\sigma))$. In particular the error estimates in (4.1) demonstrate that the coarse-grained process can preserve the ergodicity properties of the microscopic process within a prescribed tolerance. We also note that the coarse-graining modifies the microscopic prior $\mathrm{P}_N(d\sigma)$ in (2.6), yielding the coarse prior $\mathrm{P}_{M,q}(d\eta)$.

If the approximating process follows the basic principles (i) and (ii), then we observe as a result of the error estimates presented here and in [22] that both the transient, as well as the long time dynamics, are expected to be captured accurately by the coarse-graining. Although this is not a complete proof of a controlled error for infinite time, it constitutes a first rigorous step in this direction. The approximation properties are also supported by the numerics presented here and in the references.

4. Error analysis and a priori estimates for coarse-grained processes. As described in the previous section we construct a new process which only approximates the projected process $\{\mathbf{T}\sigma_t\}_{t\geq 0}$. We do not attempt to capture the effect of fine scales exactly and incorporate them into the coarse model through the renormalization group transformation. Instead we construct an approximate process $\{\eta_t\}_{t\geq 0}$, with the invariant measure $\mu^c_{M,q,\beta}$. The approximation properties of such construction are quantified in this section.

4.1. Information theory estimates. The first question which needs to be addressed is comparison and an error estimate for the exactly coarse-grained equilibrium measure, i.e., $\mathbf{T}_*\mu_{N,\beta}$, and its approximation $\mu^c_{M,q,\beta}$. We recall that $\mathbf{T}_*$ is the projection operator induced by the fine-to-coarse projection of spin variables. For the comparison of the nonequilibrium processes $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$, we need to carry out a similar a priori analysis on the coarse path space $\mathcal{D}(\mathcal{S}^c_{M,q})$, i.e., on the space of all right-continuous paths $\eta_t : [0,\infty) \to \mathcal{S}^c_{M,q}$. We denote by $Q_{\sigma_0,[0,T]}$ the measure on $\mathcal{D}(\mathcal{S}_N)$ for the process $\{\sigma_t\}_{t\in[0,T]}$ on the interval $[0,T]$ with the initial distribution σ_0. Similarly $Q^c_{\eta_0,[0,T]}$ denotes the measure on the coarse path space $\mathcal{D}(\mathcal{S}^c_{M,q})$. With a slight abuse of notation we also use $\mathbf{T}_*Q$ to denote the projection of the measure Q on the coarse path space, i.e., the exact coarsening of the measure Q.

The principal idea proposed in [22, 23] is to control *the specific loss of information* quantified by the relative entropy.

PROPOSITION 4.1. (i) see [23]: *Let $\mu^c_{M,q,\beta}$ be the approximating measure defined by* (3.13) *and $\mathbf{T}_*\mu_{N,\beta}$ be the exact projection of the microscopic equilibrium measure, then the specific relative entropy is estimated by*

$$\begin{aligned}(4.1)\qquad &\frac{1}{N}\mathcal{R}\left(\mu^c_{M,q,\beta} \,|\, \mathbf{T}_*\mu_{N,\beta}\right)\\ &:= \frac{1}{N}\sum_{\eta\in\mathcal{S}^c_{M,q}} \log\left(\frac{\mu^c_{M,q,\beta}(\eta)}{\mu_{N,\beta}(\{\sigma\in {\mathcal{S}_N}^{\Lambda_N} \,|\, \mathbf{T}\sigma=\eta\})}\right)\mu^c_{M,q,\beta}(\eta) = O\left(\frac{q}{L}\right).\end{aligned}$$

(ii) see [22]: *Suppose the process $\{\eta_t\}_{t\in[0,T]}$, given by the coarse generator $\bar{\mathcal{L}}^c$, defines the coarse approximation of the microscopic process $\{\sigma_t\}_{t\in[0,T]}$ then for any $q<L$ and N, $Mq=N$, the information loss as $q/L\to 0$ is*

$$(4.2)\qquad \frac{1}{N}\mathcal{R}\left(Q^c_{\eta_0,[0,T]} \,|\, \mathbf{T}_*Q_{\mathbf{T}_*\sigma_0,[0,T]}\right) = T\,O\left(\frac{q}{L}\right).$$

We recall that the relative entropy for two probability measures $\pi_1(\sigma)$ and $\pi_2(\sigma)$ on the countable state space $\mathcal{S}$ is defined as

$$(4.3)\qquad \mathcal{R}\left(\pi_1 \,|\, \pi_2\right) = \sum_{\sigma\in\mathcal{S}} \pi_1(\sigma)\log\frac{\pi_1(\sigma)}{\pi_2(\sigma)}.$$

We refer to [9] for a detailed discussion of relative entropy, its properties, and connections to information theory.

Remark. Although the previous estimate is for finite times $[0,T]$ only, and grows with T, in many cases the system nucleates a new phase at the initial stage of its evolution and thus the estimate ensures good approximation of the nucleation phase. It is worth noticing that the relative entropy estimate clearly demonstrates limitations of the coarse-graining method since it gives the error of order one for short-range interactions (the nearest neighbor interaction corresponds to $L=1$). On the other

hand the analysis using the relative entropy (information) distance identifies the small parameter in the asymptotic expansion of the blocking error, the ratio q/L.

In the next estimate we derive a lower bound for the loss of information in terms of coarser observables.

PROPOSITION 4.2 (lower bound). *Suppose the process* $(\{\eta_t\}_{t\in[0,T]}, \bar{\mathcal{L}}^c)$*, defined by the coarse-graining operator* $\mathbf{T}$ *with coarse-graining parameters* $Mq = N$*, is the coarse approximation of the microscopic process* $\{\sigma_t\}_{t\in[0,T]}$*. Let* $\mathbf{T}^{M',q'}$ *be another coarse-graining operator, such that* $M' \le M$*,* $M'q' = Mq = N$*. Then the following estimate for the invariant microscopic measure* $\mu_{N,\beta}$ *and the coarse approximation* $\mu^c_{M,q,\beta}$ *holds:*

$$\mathcal{R}\left(\mu^c_{M,q,\beta} \,|\, \mathbf{T}_*\mu_{N,\beta}\right) \ge \mathcal{R}\left(\mathbf{T}_*^{M',q'}\mu^c_{M,q,\beta} \,|\, \mathbf{T}_*^{M',q'}\mu_{N,\beta}\right). \tag{4.4}$$

Moreover, on any finite-time interval $[0,T]$*,*

$$\mathcal{R}\left(\mathbf{T}_*Q_{\mathbf{T}\sigma_0,[0,T]} \,|\, Q^c_{\eta_0,[0,T]}\right) \ge \mathcal{R}\left(\mathbf{T}_*^{M',q'}Q_{\mathbf{T}\sigma_0,[0,T]} \,|\, \mathbf{T}_*^{M',q'}Q^c_{\eta_0,[0,T]}\right). \tag{4.5}$$

Proof. We first recall the variational formulation for the relative entropy

$$\mathcal{R}\left(\mu \,|\, \nu\right) = \sup_f \left\{ \int f\, d\mu - \log \int e^f\, d\nu \right\}, \tag{4.6}$$

where the supremum is over all bounded functions in the space where the measures are defined. This inequality now readily implies the result since

$$\mathcal{R}\left(\mu \,|\, \nu\right) \ge \sup_{f\circ\mathbf{T}} \left\{ \int f\circ\mathbf{T}\, d\mu - \log \int e^{f\circ\mathbf{T}}\, d\nu \right\} = \mathcal{R}\left(\mathbf{T}_*\mu \,|\, \mathbf{T}_*\nu\right), \tag{4.7}$$

where $\mathbf{T}$ is the projection operator (superscripts omitted) in the statement of the proposition.

Remark. This estimate provides a lower bound for the loss of information in terms of coarser observables, hence the condition $M' \le M$ where $M'q' = Mq = N$. For instance if $M' = 1, q' = N$, then the measures $\mathbf{T}_*^{M',q'}\mu^c_{M,q,\beta}$ and $\mathbf{T}_*^{M',q'}\mu_{N,\beta}$ are the PDFs of the total coverage with respect to the coarse-grained (essentially mean field with a noise) and the microscopic Gibbs states, respectively. At first glance it may appear that such an estimate is hard to implement since it depends on the exact microscopic MC; however, when M' is small, i.e., $M' = 1, 2, 3\ldots$, the PDFs can be calculated as a histogram by MC and subsequently the relative entropy in the lower bound is straightforward to compute.

4.2. Microscopic reconstruction and weak convergence estimates. In many practical MC simulations the main goal is to estimate averages (expected values) of specific observables. Therefore it is natural to analyze the weak approximation properties of the coarse-graining procedure. The weak error is defined as the quantity $e_w \equiv |\mathbb{E}_S\left[\psi(\mathbf{T}\sigma_t)\right] - \mathbb{E}_S\left[\psi(\eta_t)\right]|$, where the expectation $\mathbb{E}_S\left[\cdot\right]$ is defined for the path conditioned on the initial configuration $\eta_0 = \mathbf{T}\sigma_0 = S$. Alternatively we can compare the microscopic process $\{\sigma_t\}_{t\ge0}$ with its synthetic process $\{\gamma_t\}_{t\ge0}$, which is reconstructed from the coarse process $\{\eta_t\}_{t\ge0}$. The weak error is then defined as $e_w \equiv |\mathbb{E}_S\left[\phi(\sigma_t)\right] - \mathbb{E}_S\left[\phi(\gamma_t)\right]|$, where the expectation $\mathbb{E}_S\left[\cdot\right]$ is now defined for the path conditioned on the initial configuration $\sigma_0 = S$. Here and in what follows ϕ denotes a test function (observable) on the fine level while ψ is used for a test function on the

coarse level. Theorem 4.7 and Corollary 4.8 quantify the rate of convergence for the weak error on both levels as $q/L \to 0$. We refer to [21] for error estimates in the weak topology between microscopic MC algorithms and therein derived approximations by stochastic differential equations.

Before we formulate the proposition and proceed with the proof it is worth clarifying the difficulty of comparing the projected process $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ with the approximating process $\{\eta_t\}_{t\geq 0}$. The projection $\mathbf{T}\sigma_t$ of the microscopic process on the coarse grid does not necessarily define a Markov process. On the other hand the approximating process $\{\eta_t\}_{t\geq 0}$ is constructed as a Markov process $(\{\eta_t\}_{t\geq 0}, \bar{\mathcal{L}}^c)$ with the generator $\bar{\mathcal{L}}^c$ defined by (3.5). To circumvent the technical difficulty the authors in [22] suggested constructing an auxiliary process $\{\gamma_t\}_{t\geq 0}$ as an intermediate step in the estimation of the relative entropy between the processes $\{\sigma_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$. We adopt the same strategy in order to make a comparison between observables which depend on Markovian processes $\{\sigma_t\}_{t\geq 0}$ and $\{\gamma_t\}_{t\geq 0}$. The process $\{\gamma_t\}_{t\geq 0}$ can be directly reconstructed from the coarse-grained process $\{\eta_t\}_{t\geq 0}$. Thus we are led to the definition of the *synthetic microscopic (Markov) process* $\{\gamma_t\}_{t\geq 0}$ associated with the process $\{\sigma_t\}_{t\geq 0}$.

DEFINITION 4.3 (synthetic microscopic process). *The auxiliary process $\{\gamma_t\}_{t\geq 0}$ is defined on the microscopic configuration space $\mathcal{S}_N$ by the generator $\mathcal{L}^\gamma : L^\infty(\mathcal{S}_N) \to \mathbb{R}$*

$$(\mathcal{L}^\gamma \phi)(\sigma) = \sum_{x\in\Lambda_N} c_\gamma(x,\sigma)(\phi(\sigma^x) - \phi(\sigma))\,, \tag{4.8}$$

where the rate function $c_\gamma(x,\sigma)$ is defined in terms of the coarse-grained interaction potential

$$c_\gamma(x,\sigma) = d_0(1-\sigma(x)) + d_0\sigma(x)e^{-\beta\bar{U}(k(x),\mathbf{T}\sigma)}\,.$$

The coarse-grained interaction potential $\bar{U}(k,\eta)$ has been defined in (3.6). *The piecewise constant interpolation is used to extend the function $\bar{U}(.,.)$ from the coarse lattice to the fine lattice. We denote $k(x)$ to be the cell index of the cell to which the site x belongs, i.e., $x \in C_{k(x)}$.*

The properties of $\{\gamma_t\}_{t\geq 0}$ were studied in [22] and the following was proved:

(i) The coarse-grained projection $\{\mathbf{T}\gamma_t\}_{t\geq 0}$ of the Markov process $(\{\gamma_t\}_{t\geq 0}, \mathcal{L}^\gamma)$ is still a Markov process.

(ii) The processes $\{\mathbf{T}\gamma_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$ have the same transition rates. Hence, whenever the processes have the same initial distribution they induce the same probability measure on the coarse-grained path space $\mathcal{D}(\mathcal{S}^c_{M,q})$. If we define $Q^c_{\eta_0}(\eta,t)$ and $Q_{\gamma_0}(\gamma,t)$ to be the probability measures of the Markov processes $\{\eta_t\}_{t\geq 0}$ and $\{\gamma_t\}_{t\geq 0}$, respectively (conditioned on the initial condition $\eta_0 = \mathbf{T}\gamma_0$), then for all $t > 0$ we have the projection

$$Q^c_{\eta_0}(\eta,t) = \mathbf{T}_* Q_{\gamma_0}(\gamma,t) \equiv \sum_{\{\gamma\,|\,\mathbf{T}\gamma=\eta_t\}} Q_{\gamma_0}(\gamma,t)\,,$$

provided this relation is satisfied at $t = 0$. Hence this property allows us to compare the processes in a pathwise way.

(iii) The microscopic process $\{\gamma_t\}_{t\geq 0}$ can be reconstructed from the approximating coarse process $\{\eta_t\}_{t\geq 0}$. Such reconstruction is an inverse procedure to the projection from fine to coarse configuration space. In such a way we can compare the original microscopic process with the approximation on the

coarse configuration space. A simple choice of a reconstruction operator is to distribute spins $\gamma_t(x)$ for $x \in C_k$ uniformly so that $\mathbf{T}\gamma_t|_{C_k} = \eta_t(k)$.

Remark. It is conceivable that the synthetic process $\{\gamma_t\}_{t\geq 0}$ can be used not only as a technical tool but also as a systematic procedure for reconstructing the microscopic process $\{\sigma_t\}_{t\geq 0}$ for the purpose of model refinement or adaptivity since, as shown in Theorem 4.7, the reconstruction is done under rigorous error estimates. In the estimates derived below we deal with a specific class of test functions $\phi \in L^\infty(\mathcal{S}_N)$ which depend only on the coarse variable $\eta = \mathbf{T}\sigma$. In other words we impose the assumption

(A1) $\phi(\sigma) = \psi(\mathbf{T}\sigma)\,,$ where $\psi \in L^\infty(\mathcal{S}^c_{M,q})$, and
$$\sum_{x\in\Lambda_N} |\partial_x \phi(\sigma)| \leq C\,, \quad \text{where } C \text{ is a constant independent of } N.$$

Remark. Observables, such as the total coverage used in the numerical simulations, satisfy this assumption.

The principal tool for analyzing the weak error is its representation in terms of solutions to the final value problem on $\mathcal{S}_N$,

$$\partial_t v(t,\sigma) + \mathcal{L}v(t,\sigma) = 0 \qquad v(T,.) = \phi(.) \quad \text{for } t < T\,,$$

where $\mathcal{L}$ is a generator of the Markov semigroup that defines the lattice dynamics. Before we state the main estimate of the weak error and its proof we need several preliminary lemmata that characterize properties of the semigroup generated by the operator $\mathcal{L}$ defined by (2.7). The specific calculations are better presented by introducing an alternative notation for the generator $\mathcal{L}$. We define an operator of discrete differentiation for functions $f \in L^\infty(\mathcal{S}_N)$

$$\partial_x f(\sigma) \equiv f(\sigma^x) - f(\sigma) \quad \text{for all } x \in \Lambda_N, \tag{4.9}$$

and we introduce two vectors indexed by the lattice sites $x \in \Lambda_N$

$$\nabla_\sigma f(\sigma) \equiv (\partial_x f(\sigma))_{x\in\Lambda_N}\,, \quad \mathbf{c}(\sigma) \equiv (c(x,\sigma))_{x\in\Lambda_N}\,.$$

The scalar product is defined in the natural way as $\mathbf{c}(\sigma)\cdot\nabla_\sigma f(\sigma) \equiv \sum_{x\in\Lambda_N} c(x,\sigma)\partial_x f(\sigma)$. Using this notation we write

$$\mathcal{L}f(\sigma) = \mathbf{c}(\sigma)\cdot\nabla_\sigma f(\sigma) \quad \text{for all } \sigma \in \mathcal{S}_N. \tag{4.10}$$

The space of functions defined on the configuration space $\mathcal{S}_N$ is equipped with the strong L^∞ topology given by the norm $||f||_\infty \equiv \sup_\sigma\{f(\sigma)\}$.

To prove the estimate in Theorem 4.7 we need an estimate for the difference operator ∇_σ stated here as a separate lemma.

LEMMA 4.4. *Let $v(t,\sigma)$ be the solution of*

$$\partial_t v + \mathcal{L}v = 0\,, \quad v(T,\sigma) = \phi(\sigma) \ \textit{for } t < T, \tag{4.11}$$

on a given interval $t \leq T$; then

$$\sum_{x\in\Lambda_N} ||\partial_x v(t,.)||_\infty \leq C_T \sum_{x\in\Lambda_N} ||\partial_x \phi||_\infty\,. \tag{4.12}$$

Moreover, the constant C_T depends exponentially on the final time T.

Proof. Using the notation introduced above and the definition of $\mathcal{L}$ we recast the evolution equation (4.11) into a familiar form of a transport equation on the configuration space

$$\partial_t v + \mathbf{c}(\sigma) \cdot \nabla_\sigma v = 0\,, \quad \sigma \in \mathcal{S}_N\,,\ t > 0\,. \tag{4.13}$$

Subtracting (4.13) for $v(t,\sigma^x)$ and $v(t,\sigma)$ we have

$$\partial_t(v(t,\sigma^x) - v(t,\sigma)) + \mathbf{c}(\sigma)\cdot(\nabla_\sigma v(t,\sigma^x) - \nabla_\sigma v(t,\sigma)) + (\mathbf{c}(\sigma^x) - \mathbf{c}(\sigma))\cdot\nabla_\sigma v(t,\sigma^x) = 0\,,$$

which we write as

$$\partial_t\left(\partial_x v(t,\sigma)\right) + \mathbf{c}(\sigma)\cdot\nabla_\sigma\left(\partial_x v(t,\sigma)\right) + \partial_x\mathbf{c}(\sigma)\cdot\nabla_\sigma v(t,\sigma^x) = 0\,. \tag{4.14}$$

Next we derive L^∞-bounds for the discrete derivatives $\partial_x\mathbf{c}(\sigma)$ using the explicit definition of the rates $c(x,\sigma)$ in (2.11). For each component, indexed by $z \in \Lambda_N$, of the vector $\mathbf{c}(\sigma)$ we have

$$\partial_x c(z,\sigma) = c(z,\sigma^x) - c(z,\sigma) = (1-\sigma^x(z)) + \sigma^x(z)e^{-U(z,\sigma^x)} - (1-\sigma(z)) + \sigma(z)e^{-U(z,\sigma)}\,.$$

For the spin-flip dynamics, i.e., $\sigma^x(y) = 1-\sigma(y)$ if $x = y$ and $\sigma^x(y) = \sigma(y)$ otherwise, a straightforward calculation gives $\partial_x U(z,\sigma) \equiv U(z,\sigma^x) - U(z,\sigma) = J(z-x)(1-2\sigma(x))$ if $z \neq x$ and it is equal to zero otherwise. Thus the discrete derivate $\partial_x\mathbf{c}(\sigma)$ is

$$\partial_x c(z,\sigma) = \begin{cases} (2\sigma(x)-1)(1-e^{-U(x,\sigma)}) & \text{for } z = x,\\ \sigma(z)e^{-U(z,\sigma)}\left(1-e^{J(x-z)(1-2\sigma(x))}\right) & \text{if } z \neq x. \end{cases}$$

Recalling the definition (2.3) of the interaction potential J we have that $J(z-x) \sim 1/L$ for $|z-x| \leq L$ and $J = 0$ otherwise. Hence we derived L^∞-bounds for the discrete derivative of the rates

$$\partial_x c(z,\sigma) \sim \begin{cases} \mathrm{O}(1) & \text{for } z = x,\\ \mathrm{O}(1/L) & \text{for } |z-x| < L,\\ 0 & \text{otherwise.} \end{cases} \tag{4.15}$$

Going back to (4.14), we have for all $x \in \Lambda_N$

$$\partial_t\left(\partial_x v(t,\sigma)\right) + \mathcal{L}\partial_x v(t,\sigma) + \sum_{z\in\Lambda_N}\partial_x c(z,\sigma)\partial_z v(t,\sigma^x) = 0\,. \tag{4.16}$$

The estimates in (4.15) imply that

$$\partial_t\partial_x v(t,\sigma) + \mathcal{L}\partial_x v(t,\sigma) + \mathrm{O}(1)\partial_x v(t,\sigma^x) + \mathrm{O}\left(\frac{1}{L}\right)\sum_{\substack{z\in\Lambda_N\\ |z-x|\leq L}}\partial_z v(t,\sigma^x) = 0\,, \tag{4.17}$$

and we have for all $\sigma \in \mathcal{S}_N$ the solution formula

$$\partial_x v(t,\sigma) = e^{t\mathcal{L}}[\partial_x v(0,\sigma)] + \int_t^T e^{(s-t)\mathcal{L}}\left[\mathrm{O}(1)\partial_x v(s,\sigma^x) + \mathrm{O}(1/L)\sum_{|z-x|\leq L}\partial_z v(s,\sigma^x)\right]ds\,.$$

By the contractive property of the semigroup $e^{t\mathcal{L}}$ we have the estimate

$$||\partial_x v(t,\cdot)||_\infty \le ||\partial_x v(0,\cdot)||_\infty + \int_t^T \mathrm{O}(1)||\partial_x v(s,\cdot)||_\infty \, ds + \int_t^T \mathrm{O}(1/L) \sum_{|z-x|\le L} ||\partial_z v(s,\cdot)||_\infty \, ds$$

for all $x \in \Lambda_N$. Thus summing over all $x \in \Lambda_N$, we obtain

$$\sum_{x\in\Lambda_N} ||\partial_x v(t,\cdot)||_\infty \le \sum_{x\in\Lambda_N} ||\partial_x v(0,\cdot)||_\infty + \int_t^T \left(\mathrm{O}(1) \sum_{x\in\Lambda_N} ||\partial_x v(s,\cdot)||_\infty + \mathrm{O}(1/L) \sum_{x\in\Lambda_N} \sum_{|z-x|\le L} ||\partial_z v(s,\cdot)||_\infty \right) ds \,,$$

where the last double sum in the integrand is bounded by $2L \sum_x ||\partial_x v(s,\cdot)||_\infty$. Hence by setting $\theta(t) = \sum_x ||\partial_x v(t,\cdot)||_\infty$ we have

$$\theta(t) \le \theta(0) + \int_t^T \mathrm{O}(1)\theta(s) \, ds \,,$$

from which, by using Gronwall's inequality, we obtain the bound

$$\theta(t) \le e^{c(T-t)}\theta(T) \,,$$

which concludes the proof of (4.4).

Next we establish an L^∞-bound for discrete derivatives of solutions generated by semigroups $e^{t\mathcal{L}}$ and $e^{t\mathcal{L}^\gamma}$.

LEMMA 4.5. *Let $u(t,\sigma)$ be the solution of*

$$\partial_t u + \mathcal{L}u = 0, \quad u(T,.) = \phi \ \text{ for } t < T,$$

and let $v(t,\sigma)$ solve

$$\partial_t v + \mathcal{L}^\gamma v = 0, \quad v(T,.) = \psi \ \text{ for } t < T,;$$

then for any $t \le T$ the following estimate holds:

$$\sum_{x\in\Lambda_N} ||\partial_x u(t,\cdot) - \partial_x v(t,\cdot)||_\infty \le C_1(T) \sum_{x\in\Lambda_N} ||\partial_x\phi - \partial_x\psi||_\infty + C_2(T)\left(\frac{q}{L}\right). \tag{4.18}$$

The constants C_1 and C_2 are independent of q and L but depend exponentially on the final time T.

Proof. We use the same approach and notation as in the proof of Lemma 4.4. Subtracting the evolution equations and defining $w_x(t,\sigma) \equiv \partial_x u(t,\sigma) - \partial_x v(t,\sigma)$, $\mathbf{w}(t,\sigma) \equiv (w_x(t,\sigma))_{x\in\Lambda_N}$, we have

$$\partial_t w_x(t,\sigma) + \mathcal{L}w_x(t,\sigma) \tag{4.19}$$

$$+ (\mathbf{c}_\gamma(\sigma) - \mathbf{c}(\sigma)) \cdot \nabla_\sigma v(t,\sigma^x) \tag{4.20}$$

$$+ \partial_x \mathbf{c}(\sigma) \cdot \mathbf{w}(t,\sigma^x) \tag{4.21}$$

$$+ (\partial_x \mathbf{c}(\sigma) - \partial_x \mathbf{c}_\gamma(\sigma)) \cdot \nabla_\sigma v(t,\sigma^x) = 0 \,. \tag{4.22}$$

From Lemma 4.4 we have estimates for the terms involving $\nabla_\sigma v(t,.)$ (notice that the lemma essentially gives the estimate of $||\nabla_\sigma v(t,.)||_\infty$). Furthermore, from the definition of rates $c(x,\sigma)$ and $c_\gamma(x,\sigma)$ direct calculation (similar to that used in the proof of Lemma 4.4) yields the estimate

$$||\mathbf{c}-\mathbf{c}_\gamma||_\infty = \mathrm{O}\left(\frac{q}{L}\right), \tag{4.23}$$

which allows us to control (4.20) and (4.22). Term (4.21) is treated in the same way as a similar term in the proof of Lemma 4.4. Hence for all $x \in \Lambda_N$ we obtain

$$\partial_t w_x(t,\sigma) + \mathcal{L} w_x(t,\sigma) + \mathrm{O}(1/L) \sum_{|z-x|\le L} w_x(z,\sigma^x) \le \mathrm{O}(q/L)||\partial_x v(t,.)||_\infty .$$

Similarly, as in the proof of Lemma 4.4, we complete the proof by summing over $x \in \Lambda_N$ and applying Gronwall's inequality.

Since we are comparing the process $\{\sigma_t\}_{t\ge 0}$ with the process $\{\gamma_t\}_{t\ge 0}$, which is defined only up to the equivalence given by the projection operator $\mathbf{T}$, we have to establish the uniqueness of solutions for initial data satisfying the assumption (A1).

LEMMA 4.6. *Let $\phi \in L^\infty(\mathcal{S}_N)$, $\psi \in L^\infty(\mathcal{S}^c_{M,q})$ be test functions satisfying* (A1). *Assume that $v(t,\gamma)$ is the solution of the final value problem*

$$\partial_t v + \mathcal{L}^\gamma v = 0, \quad v(T,\gamma) = \phi(\gamma) = \psi(\mathbf{T}\gamma); \tag{4.24}$$

then for all $\gamma, \gamma' \in \mathcal{S}_N$ such that $\mathbf{T}\gamma = \mathbf{T}\gamma'$

$$v(t,\gamma) = v(t,\gamma') \quad \textit{for all } t \le T. \tag{4.25}$$

Proof. For convenience we write $v(t,\gamma) = v(t,\mathbf{T}\gamma)$. Given a configuration $\gamma \in \mathcal{S}_N$ we can reconstruct an arbitrary configuration $\gamma' \in \mathcal{S}_N$ such that $\mathbf{T}\gamma' = \mathbf{T}\gamma$ by considering a permutation $\pi : \Lambda_N \to \Lambda_N$, $\pi = (\pi_1, \dots, \pi_M)$ such that

$$\pi_k : C_k \to C_k, \quad k = 1, \dots, M .$$

The action of π on the configuration space is defined in a natural way $\gamma' = \gamma \circ \pi$, or equivalently $\gamma'(x) = \gamma(\pi x)$. Since the permutation does not change the total spin in the cell we have $\mathbf{T}\gamma \circ \pi = \mathbf{T}\gamma$. Hence we write $v(t,\gamma') = v(t,\gamma\circ\pi)$ and $v(T,\gamma\circ\pi) = v(T,\gamma) = \psi(\mathbf{T}\gamma)$. It is sufficient to show that the function $u(t,\gamma) \equiv v(t,\gamma\circ\pi)$ is a solution of (4.24). From the uniqueness of solutions to (4.24) we conclude immediately that $u(t,\gamma) = v(t,\gamma)$. From the definition of the generator $\mathcal{L}^\gamma$ we have

$$\partial_t v(t,\gamma\circ\pi) + \sum_{k\in\Lambda^c_M} \sum_{x\in C_k} c_\gamma(x,\gamma\circ\pi)(v(t,(\gamma\circ\pi)^x) - v(t,\gamma\circ\pi)) = 0 . \tag{4.26}$$

Recall the definition of the rate c_γ

$$c_\gamma(x,\gamma) = d_0(1-\gamma(x)) + d_0\gamma(x)e^{-\beta\bar{U}(k(x),\mathbf{T}\gamma)},$$

and denote $c_\gamma(x,\gamma)$ by $C_\gamma(\gamma(x),k,\mathbf{T}\gamma)$ to emphasise the dependence on $\gamma(x)$, k, and $\eta = \mathbf{T}\gamma$ only. Thus the inner summation in (4.26) becomes

$$\sum_{x\in C_k} C_\gamma(\gamma\circ\pi,k,\mathbf{T}\gamma)(v(t,(\gamma\circ\pi)^x) - v(t,\gamma\circ\pi)) . \tag{4.27}$$

On the other hand the definition of spin-flip dynamics leads to

(4.28)
$$(\gamma\circ\pi)^x(z)=\begin{cases}\gamma(\pi z), & z\neq x,\\ 1-\gamma(\pi x), & z=x,\end{cases}\quad\text{while }\gamma^{(\pi x)}(\pi z)=\begin{cases}\gamma(\pi z), & z\neq x\,,\\ 1-\gamma(\pi x), & z=x\,.\end{cases}$$

Hence we obtain

$$(\gamma\circ\pi)^x(z)=\gamma^{(\pi x)}(\pi z)=\left(\gamma^{\pi x}\circ\pi\right)(z)\,, \tag{4.29}$$

and substituting to the expression (4.27) leads to

$$\begin{aligned}&\sum_{x\in C_k}C_\gamma(\gamma(\pi x),k,\mathbf{T}\gamma)(v(t,(\gamma\circ\pi)^x)-v(t,\gamma\circ\pi))\\ &=\sum_{x\in C_k}C_\gamma(\gamma(\pi x),k,\mathbf{T}\gamma)(v(t,\gamma^{\pi x}\circ\pi)-v(t,\gamma\circ\pi))\\ &=\sum_{y\in C_k}C_\gamma(\gamma(y),k,\mathbf{T}\gamma)(v(t,\gamma^{y}\circ\pi)-v(t,\gamma\circ\pi))\\ &=\sum_{y\in C_k}C_\gamma(\gamma(y),k,\mathbf{T}\gamma)(u(t,\gamma^{y})-u(t,\gamma))\,.\end{aligned}$$

Thus we have shown that

$$\partial_t u(t,\gamma)+\sum_{k\in\Lambda_M^c}\sum_{x\in C_k}c_\gamma(x,\gamma)(u(t,\gamma^x)-u(t,\gamma))=0\,.$$

Recalling the definition of $u(t,\gamma)$ we obtain that $v(t,\gamma\circ\pi)$ also solves (4.24). The uniqueness of solutions to (4.24) implies that $v(t,\gamma\circ\pi)=v(t,\gamma)$ for all γ or $v(t,\gamma')=v(t,\gamma)$ for all γ' such that $\mathbf{T}\gamma'=\mathbf{T}\gamma$. □

Now we can formulate and prove the weak error estimate that allows us to compare the microscopic process and its coarse-level approximation. We estimate the weak error on the microscopic level by comparing the microscopic process and its synthetic process.

THEOREM 4.7 (weak error). *Let $\phi\in L^\infty(\mathcal{S}_N)$ be a test function (observable) on the microscopic space satisfying* (A1) *and let $(\{\gamma_t\}_{t\geq0},\mathcal{L}^\gamma)$ be the synthetic Markov process (in the sense of Definition* 4.3*) of the microscopic process $(\{\sigma_t\}_{t\geq0},\mathcal{L})$ with the initial condition $\sigma_0=S$; then the weak error satisfies, for $0<T<\infty$,*

$$\left|\mathbb{E}_S\left[\phi(\sigma_T)\right]-\mathbb{E}_S\left[\phi(\gamma_T)\right]\right|\leq C_T\left(\frac{q}{L}\right)^2\,, \tag{4.30}$$

where the constant C_T is independent of q and L but depends on T.

Proof. The two ingredients of the proof, the Feynman–Kac formula and the martingale property, follow from the standard properties of Markov processes (see, for example, [25]). If we define, for the microscopic process $\{\sigma_t\}_{t\geq0}$ defined by the generator $\mathcal{L}$, the function

$$u(t,S)=\mathbb{E}\left[\phi(\sigma_T)\,|\,\sigma_t=S\right],$$

then from the Feynman–Kac formula with the zero potential it follows that the function $u(t,S)$ solves the final value problem

$$\partial_t u+\mathcal{L}u=0\,,\qquad u(T,.)=\phi\,,\quad t<T\,. \tag{4.31}$$

On the other hand the martingale property implies that for any smooth function $v(t,S)$ and the process $\{\gamma_t\}_{t\geq 0}$ with the generator $\mathcal{L}^\gamma$ we have

$$\mathbb{E}_S\left[v(T,\gamma_T)\right] = \mathbb{E}_S\left[v(0,\gamma_0)\right] + \int_0^T \mathbb{E}_S\left[(\partial_s + \mathcal{L}^\gamma)v(s,\gamma_s)\right] ds\,.$$

The definition of $u(t,S)$ leads to the representation of the error $|\mathbb{E}_S\left[\phi(\sigma_T)\right] - \mathbb{E}_S\left[\phi(\gamma_T)\right]|$ by $e_w = |\mathbb{E}_S\left[u(0,S)\right] - \mathbb{E}_S\left[u(T,\gamma_T)\right]|$ and hence

$$e_w = \left| \int_0^T \mathbb{E}_S\left[(\partial_s + \mathcal{L}^\gamma)\, u(s,\gamma_s)\right] ds \right|\,.$$

The function $u(t,S)$ solves the equation $\partial_t u = -\mathcal{L}u$. Thus we obtain

$$\mathbb{E}_S\left[\phi(\sigma_T) - \phi(\gamma_T)\right] = \int_0^T \mathbb{E}_S\left[\mathcal{L}^\gamma u(t,\gamma_t) - \mathcal{L}u(t,\gamma_t)\right]dt$$
$$= \int_0^T \mathbb{E}_S\left[\sum_{x\in\Lambda_N} \left(c(x,\gamma_t) - c_\gamma(x,\gamma_t)\right)\partial_x u(t,\gamma_t)\right] dt\,.$$

We split the summation $\sum_{x\in\Lambda_N}$ which gives us

$$\mathbb{E}_S\left[\phi(\sigma_T) - \phi(\gamma_T)\right] = \int_0^T \mathbb{E}_S\left[\sum_{k\in\Lambda_M^c}\sum_{x\in C_k} \left(c(x,\gamma_t) - c_\gamma(x,\gamma_t)\right)\partial_x u(t,\gamma_t)\right] dt$$
$$= \int_0^T \mathbb{E}_S\left[\sum_{k\in\Lambda_M^c}\sum_{x\in C_k} \gamma_t(x)(e^{-\beta U(x,\gamma_t)} - e^{-\beta\bar{U}(k(x),\mathbf{T}\gamma_t)})(\partial_k v(t,\mathbf{T}\gamma_t) + R_T^{q,L}(x))\right] dt\,.$$

Here we need to replace $\partial_x u$ by $\partial_x v$, where v solves the final value problem (4.31) with $\mathcal{L}$ replaced by $\mathcal{L}^\gamma$. From Lemma 4.5 we know that the error term $R_T^{q,L}(x) = \partial_x u(t,\gamma) - \partial_x v(t,\gamma)$ is controlled by $\mathrm{O}(q/L)$ in $||\cdot||_\infty$. Furthermore, Lemma 4.6 guarantees that with the final condition ϕ which satisfies assumption (A1) the solution depends only on $\mathbf{T}\gamma$ and hence we can replace the discrete difference $\partial_x v$ by the difference $\partial_k v(t,\eta) \equiv v(t,\eta+\delta_k) - v(t,\eta)$, where $\eta = \mathbf{T}\gamma$. Next we expand the exponentials to obtain

$$\Gamma(k,\gamma) \equiv \sum_{x\in C_k} \beta\gamma(x) e^{-\beta\bar{U}(k(x),\mathbf{T}\gamma)} \left(\Delta(\bar{U},U) + \frac{1}{2}\beta^2\Delta^2(\bar{U},U) + \mathrm{O}\left(\beta^3\Delta^3(\bar{U},U)\right)\right)\,,$$

and we recast the error representation into

$$\mathbb{E}_S\left[\phi(\sigma_T) - \phi(\gamma_T)\right]$$
$$= \int_0^T \mathbb{E}_S\left[\sum_{k\in\Lambda_M^c} \Gamma(k,\gamma_t)\partial_k v(t,\mathbf{T}\gamma_t) + \sum_{x\in\Lambda_N} \left(c(x,\gamma_t) - c_\gamma(x,\gamma_t)\right)R_T^{q,L}(x)\right] dt$$
$$= \int_0^T \mathbb{E}_S\left[q\sum_{k\in\Lambda_M^c} \partial_k v(t,\eta_t)\,\mathbb{E}\left[\Gamma(k,\gamma)\big|\mathbf{T}\gamma = \eta_t\right]\right] dt \tag{4.32}$$
$$+ \int_0^T \mathbb{E}_S\left[\sum_{x\in\Lambda_N} \left(c(x,\gamma_t) - c_\gamma(x,\gamma_t)\right)R_T^{q,L}(x)\right] dt\,. \tag{4.33}$$

Assumption (A1) and Lemma 4.4 imply that the term $q\sum_{k\in\Lambda_M^c}\partial_k v(t,\eta_t)$ is bounded. To estimate the conditional expectation we use the property of the reconstruction operator for the process $\{\gamma_t\}_{t\geq 0}$, in particular on each cell $\gamma_t(x)$ is reconstructed from $\eta_t(k)$ by assuming a "local" equilibrium and distributing $\gamma_t(x)$ uniformly in the cell $C_{k(x)}$. Using this property we can compute the conditional expectation explicitly and obtain for $l\neq k$

$$\mathbb{E}\left[\sum_{x\in C_k}\gamma(x)\Delta(\bar{U},U)\big|\mathbf{T}\gamma=\eta\right]=\eta_k\eta_l\sum_{\substack{x\in C_k\\ y\in C_l}}\left(J(x-y)-\bar{J}_{kl}\right)=0\,.$$

Similarly we handle the case $l=k$ and conclude that, after averaging, the first-order term $\Delta(\bar{U},U)$ in $\Gamma(k,\gamma)$ vanishes. We recall (see (3.3)) that

$$\Delta(\bar{U},U)\equiv\bar{U}(k(x),\mathbf{T}\gamma)-U(x,\gamma)=\mathrm{O}\left(\frac{q}{L}\right)\,,$$

and hence we can estimate (4.32) by $O(q^2/L^2)$. For the term (4.33) we use the estimate $\sum_{x\in\Lambda_N}|R_T^{q,L}(x)|\sim\mathrm{O}(q/L)$ from Lemma 4.5 and the Hölder inequality

$$\mathbb{E}_S\left[\sum_{x\in\Lambda_N}(c(x,\gamma_t)-c_\gamma(x,\gamma_t))R_T^{q,L}(x)\right]\leq||\mathbf{c}-\mathbf{c}_\gamma||_\infty\mathbb{E}_S\left[\sum_{x\in\Lambda_N}|R_T^{q,L}(x)|\right].$$

The first term on the right-hand side is estimated from (4.23) by $C(q/L)$ and hence the left-hand side behaves as $\mathrm{O}(q^2/L^2)$. Combining the estimates of (4.32) and (4.33) we conclude the proof.

Using the estimate for the synthetic process and its reconstruction from the coarse-grained process $\{\eta_t\}_{t\geq 0}$ we can compare the projected process $\{\mathbf{T}\sigma_t\}_{t\geq 0}$ and the coarse-grained process $\{\eta_t\}_{t\geq 0}$ also on the coarse level. The weak error for observables on the coarse space is also natural in simulations where we usually project finer simulations on the coarse level and use estimators for the coarse processes.

COROLLARY 4.8. *Let $\psi\in L^\infty(\mathcal{S}_{M,q}^c)$ be a test function on the coarse level such that there exists a test function $\phi\in L^\infty(\mathcal{S}_N)$ satisfying* (A1) *with the property $\psi(\mathbf{T}\sigma)=\phi(\sigma)$. Given the initial configuration σ_0 we define the coarse configuration $\eta_0=\mathbf{T}\sigma_0$. Assume the microscopic process $(\{\sigma_t\}_{t\geq 0},\mathcal{L})$ with the initial condition σ_0 and the approximating coarse process $(\{\eta_t\}_{t\geq 0},\bar{\mathcal{L}}^c)$ with the initial condition $\eta_0=\mathbf{T}\sigma_0$; then the weak error satisfies, for $0<T<\infty$,*

$$\left|\mathbb{E}_S\left[\psi(\mathbf{T}\sigma_T)\right]-\mathbb{E}_S\left[\psi(\eta_T)\right]\right|\leq C_T\left(\frac{q}{L}\right)^2\,, \tag{4.34}$$

where the constant C_T is independent of q and L but depends on T.

We conclude this section with a brief remark on regimes of applicability and limitations of the derived coarse-grained approximation. In the introduction we mentioned a few examples of physically relevant problems where the CGMC method is applicable. However, a closer inspection of the estimate in Theorem 4.7 reveals further regimes of validity for the approximation beyond the smallness of the ratio q/L. More specifically, we note that at high temperatures, i.e., $\beta\ll 1$, the error terms are small. Thus CGMC provides a good approximation even if the involved potentials are of a short range. Furthermore, in the presence of a strong external field, $|h|\gg 1$, the CGMC dynamics also provide a good approximation even for short-range interactions. In such a case we have (almost uniform) clusters of 0's or 1's, hence with large probability $\sigma(\cdot)\approx\eta(\cdot)/q\approx 0$, or 1; therefore, $\mathbb{E}\Delta(\bar{U},U)\ll 1$, and hence the error in Theorem 4.7 is small.

5. Numerical simulations. We use the CGMC described and analyzed in the previous sections for efficient simulations in the spin systems that undergo phase transitions; for the implementation details we refer to [19]. Within the context of spin-flip dynamics a typical example is nucleation of spatial regions of a new phase or a transition from one phase (all spins equal to zero) to another (all spins equal to one). In such simulations the emphasis is on the pathwise properties of the coarse-grained process so that the switching mechanism is simulated efficiently while approximation errors are controlled. We compare simulations on the microscopic level $q = 1$ with those performed on different levels of coarse-graining hierarchy parametrized by q.

The qualitative behavior of the Ising model with a long-range potential can be understood from the mean-field approximation of the equilibrium total coverage $c(\sigma)$. Below the critical temperature the Gibbs measure is not unique (in the thermodynamic limit $N \to \infty$) and two phases can coexist. When the energy landscape is probed by changing the external field h we observe nonuniqueness of the equilibrium coverage. The fluctuations allow for transitions between the equilibrium which leads to nucleation of regions with a different phase. Changing the external field h makes the original phase unstable and a switching occurs—the system transforms into the other equilibrium configuration.

The parameters in the simulations have been chosen as follows: We use a uniform finite range potential for all examples presented. We simulate a finite lattice with a total of $N = 1000$ microscopic nodes and allow a potential interaction range of $2L+1$ for $L = 100$. We choose the constant $d_0 = 1$ so that $c_a = 1$ and $c_d = 1$. Hence in this case the critical value β_c is given by $\beta_c J_0 = 4$. If $\beta J_0 > \beta_c J_0 = 4$, then the system is in the phase transition regime and the two phases can coexist. In this region we typically observe a transition from one phase (e.g., zero (low) coverage) to the other phase (e.g., full coverage). For the phase transition examples we fix $\beta J_0 = 6 > \beta_c J_0$. The simulations become difficult when $\beta \simeq \beta_c$ and there is no external field h applied. We note that the coarse-graining algorithm will not perform well close to the critical point β_c when $h = 0$. In the numerical studies we first investigate approximation properties of the CGMC algorithms for certain global quantities.

Coverage: We define the coverage c_t to be the process computed as the spatial mean

$$c_t(\sigma_t) = \frac{1}{N} \sum_{x \in \Lambda_N} \sigma_t(x)\,, \qquad c_t^q(\eta_t) = \frac{1}{qM} \sum_{l \in \Lambda_M^c} \eta_t(k)\,.$$

First we present, in Figure 5.1, a simulation for the coverage in the absence of phase transitions where we see a remarkable pathwise agreement. Time evolution of the coverage at the phase transition regime, $\beta J_0 = 6$, is depicted in Figure 5.2 for different values of q. Note that the case $q = 1000$, $m = 1$ which corresponds to the mean-field approximation ("over coarse-grained" interactions) does not follow the phase transition path of the other simulations. On the other hand the agreement in the results is extremely good for the remaining values of q. Furthermore, these numerical results indicate pathwise (strong) approximation of the microscopic process by the coarse-grained process. This observation suggests a stronger error control than the relative entropy estimate provided by Proposition 4.1.

To quantify the error behavior we calculate two errors between the exact stochastic process c_t and its coarse approximation c_t^q at the level of coarse-graining q. We define

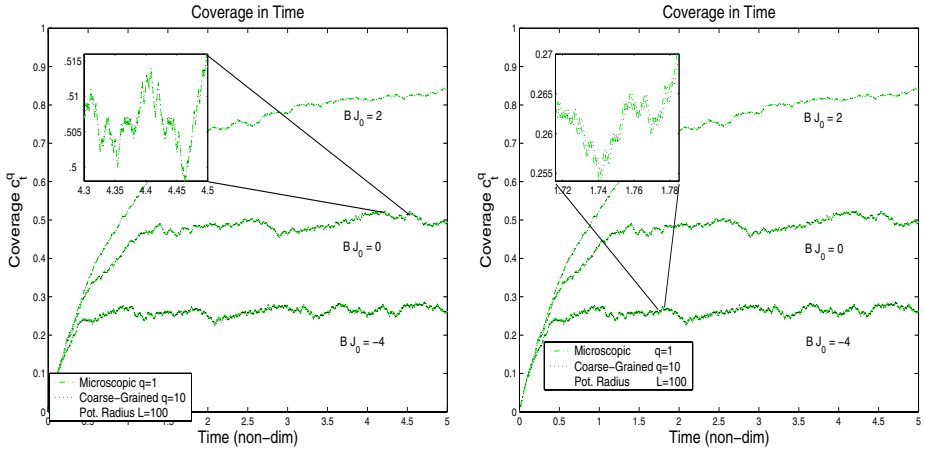


FIG. 5.1. *Relaxation dynamics. Comparison of microscopic* ($q = 1$) *and coarse-grained* ($q = 10$) *simulations. The plot depicts a short time simulation in order to calibrate the code and compare to Figure* 4 *from* [19].

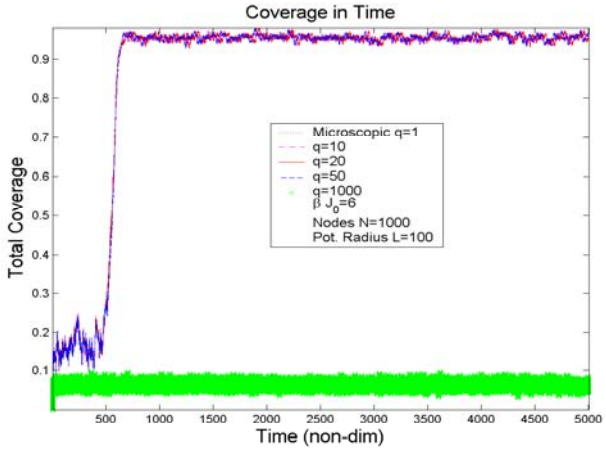


FIG. 5.2. *Time series of the coverage* c_t^q. *Simulations for different coarse-graining ratios are shown in the phase transition regime. The case* $q = 1000$, $m = 1$ *(mean-field approximation) shows significant discrepancy. Parameters used: potential radius length* $L = 100$, $\beta J_0 = 6$, $d_0 = 1$, $c_0 = .072$.

the weak error $e_w[c]$ and the strong error $e_s[c]$, respectively:

$$e_w[c] = \int_0^T |\mathbb{E}[c_t] - \mathbb{E}[c_t^q]|\, dt\,, \quad e_s[c] = \int_0^T \mathbb{E}[|\mathbf{T}c_t - c_t^q|]dt\,.$$

The expected values are estimated by empirical means and the integral in time by the piecewise constant quadrature.

The simulations allow us to estimate the convergence rate for both errors. The rates in the case of fixed parameters $L = 100$, $d_0 = 1.0$, $c_0 = 0.07$, and $\beta J_0 = 6$ on the lattice of the size $N = 1000$ are depicted in Figure 5.3. Note that we need to eliminate the statistical error, arising from approximation of expected values by empirical means. However, as seen in Figure 5.3 the estimator of the rate converges

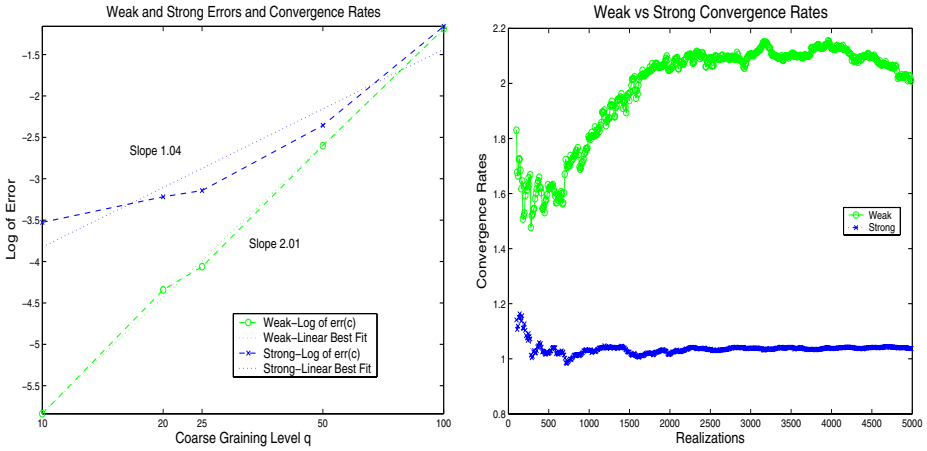


FIG. 5.3. *Estimated weak $e_w[c]$ and strong $e_s[c]$ errors. We compare the exact process c_t, $q = 1$ with coarse approximations c_t^q, $q = 10$, 25, 50, and 100. The simulation parameters were fixed at $L = 100$, $d_0 = 1$, $c_0 = .07$, $\beta J_0 = 6 > \beta_c J_0$, and the lattice size $N = 1000$. The convergence rates depicted are estimated by the linear best fit on the logarithmic scale. The statistical error or dependence of the estimates on the number of realizations is depicted in the right figure.*

TABLE 5.1

Relative strong error $e_s[c]$ in the presence of an external field defined by c_0. Comparisons are made for different values of the interaction radius L and different coarse-graining levels q. Size of the lattice fixed at $N = 1000$.

c_0	L	$q = 5$	$q = 10$	$q = 20$
.07	100	.0591	.0733	.1134
	40	.0820	.0880	.1113
	20	.1508	.2214	.1832
.09	100	.0186	.0563	.0480
	40	.0678	.0749	.1064
	20	.1760	.1767	.1812
1	100	.0010	.0010	.0025
	40	.0036	.0040	.0054
	20	.0016	.0043	.0065

as the number of realizations tends to infinity.

Since the coarse-grained Hamiltonian neglects higher order corrections arising from the fluctuations on fine scales, one may expect that the approximation is poor if q/L is not very small. This is certainly true at the critical point (i.e., $\beta = \beta_c$ and $h = 0$) but further from the critical point the approximation properties are improved. This is demonstrated in Table 5.1, where the simulations were performed in the presence of different (large) external fields. The relative error becomes small even for fairly crude coarse-graining $q = 20$ in the case of shorter interaction radii L.

Mean time to reach phase transition: One quantity of interest that is calculated from the simulations is the mean time $\bar{\tau}_T = \mathbb{E}[\tau_T]$ until the coverage reaches C^+ in its phase transition regime (see Figure 5.2). The random exit time is defined as $\tau_T = \inf\{t > 0 \,|\, c_t \geq C^+\}$. We estimate the probability distributions ρ_τ and ρ_τ^q from the simulations. We record a phase transition at the time $\bar{\tau}_T$ when the coverage exceeds the threshold value $C^+ = 0.9$. The relative error for the estimated mean time $\bar{\tau}_T$ at different levels q is tabulated in Table 5.2 together with estimated relative entropy for the random variable τ_T. In Figure 5.4 we plot approximations of the

TABLE 5.2
Approximation of $\bar{\tau}_T$, $\mathcal{R}\left(\rho_\tau^q \,|\, \mathbf{T}_*\rho_\tau\right)$ *and relative error. Measurements based on averaging over* 10000 *realizations for each* q.

L	q	$\bar{\tau}_T$	$\mathcal{R}\left(\rho_\tau^q \,|\, \mathbf{T}_*\rho_\tau\right)$	Rel. Err.	CPU [s]
100	1	532	0.0	0	309647
100	2	532	0.003	0.01%	132143
100	4	530	0.001	0.22%	86449
100	5	534	0.003	0.38%	58412
100	10	536	0.004	0.82%	38344
100	20	550	0.007	3.42%	16215
100	25	558	0.010	4.91%	7574
100	50	626	0.009	17.69%	4577
100	100	945	0.087	77.73%	345

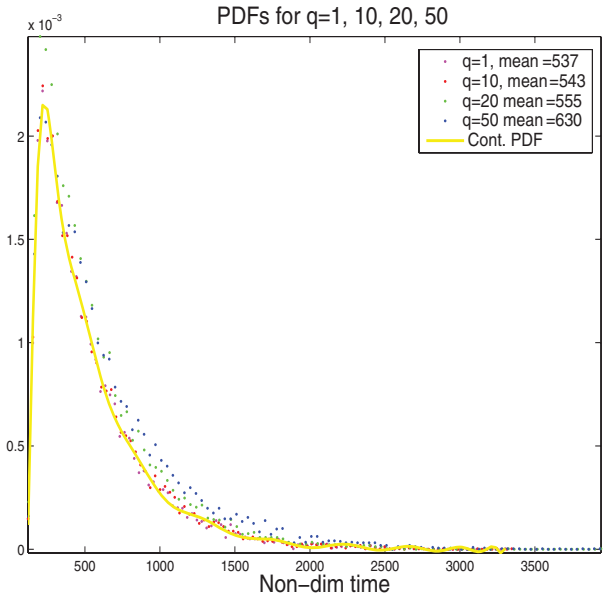


FIG. 5.4. *Probability density function (PDFs) comparisons between different coarse-graining levels* q. *The estimated mean times for each PDF are shown in the figures. All PDFs comprised of* 10000 *samples and the histogram is approximated by* 100 *bins.*

probability density functions (PDFs) of τ_T and compare them for different values of q.

Nucleation: The nucleation of a new phase is a typical phenomenon in the regime where $\beta > \beta_c$. Essentially, there exist two equilibria (phases). Random fluctuations will induce transitions from one state to another by overcoming energy barriers that separate the equilibria. We investigate approximation of the pathwise behavior on the configuration space for nucleation of a new phase. Two different initial configurations are used.

Test case I: The initial state is at the metastable equilibrium where the coverage is zero. The fluctuations will cause the transition to the full coverage equilibrium which is stable due to the external applied field. We present only qualitative comparison in the series of snapshots (Figure 5.5) of the phase transition from the uniform (zero) initial coverage to the full coverage. We observe a striking pathwise agreement on the

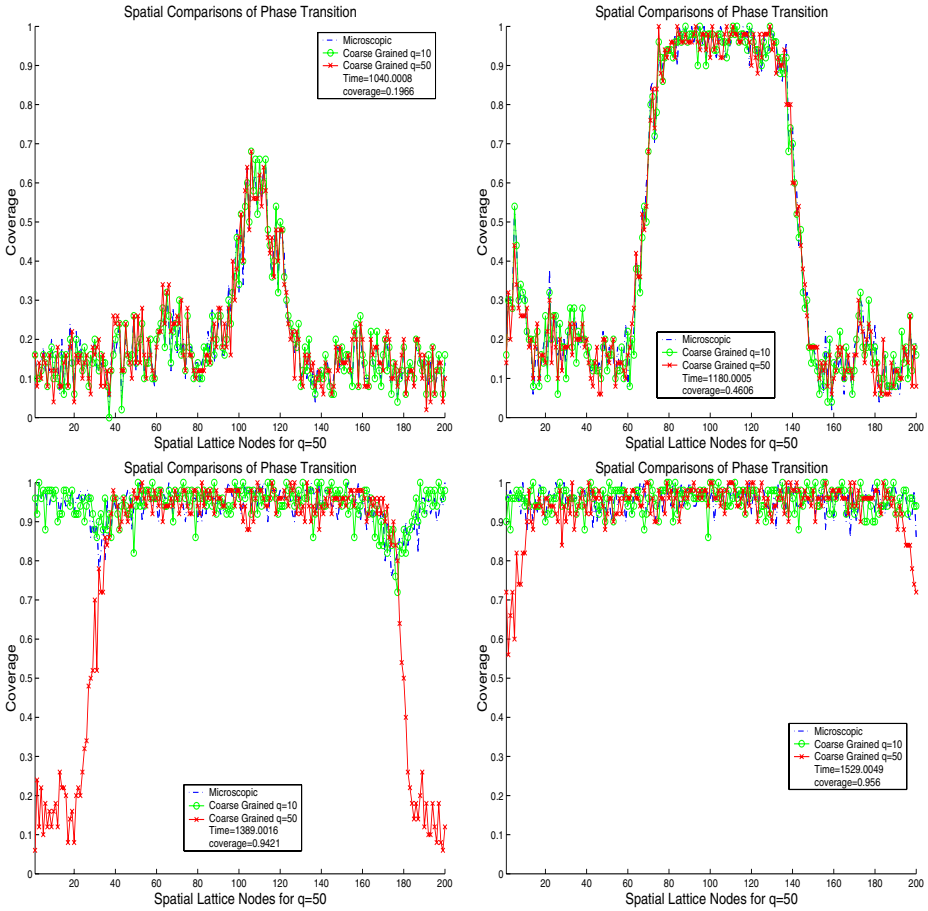


FIG. 5.5. *Snapshots of the transition from zero initial spatial distribution. Comparisons between the microscopic* $q = 1$ *and two coarse-grained simulations* $q = 10$ *and* $q = 50$. *The interaction radius is set to* $L = 200$ *while total nodes are* $N = 10000$.

configuration space for relatively large values of q compared to the interaction radius L. However, as the ratio q/L increases the corresponding coarse-grained process lags behind, which is also demonstrated in the expected values of transition times. Such behavior suggests that fluctuations at regions with uniform states are well-approximated by a highly coarse-grained process while finer resolution is necessary for resolving nucleation of new phases through islands.

Test case II: We have already documented the pathwise agreement of the approximating dynamics under both transition and relaxation cases. In this example we examine the spinodal decomposition phenomenon at the phase transition regime, $\beta J_0 = 6$. We chose the initial state to be at a saddle point of the energy surface, i.e., the mean coverage is set to 0.5. Snapshots of the spatial distribution of spins are presented in Figure 5.6. Under all four dynamics examined, $q = 1, 5, 10$, and 20, we observe complete spatial pathwise agreement. Over time the total coverage may fall towards zero or rise towards one in which case it will remain there since we are at the phase transition regime where these represent stable equilibria. The application of Theorem 4.7 for this case is not immediately obvious as the constant in the error

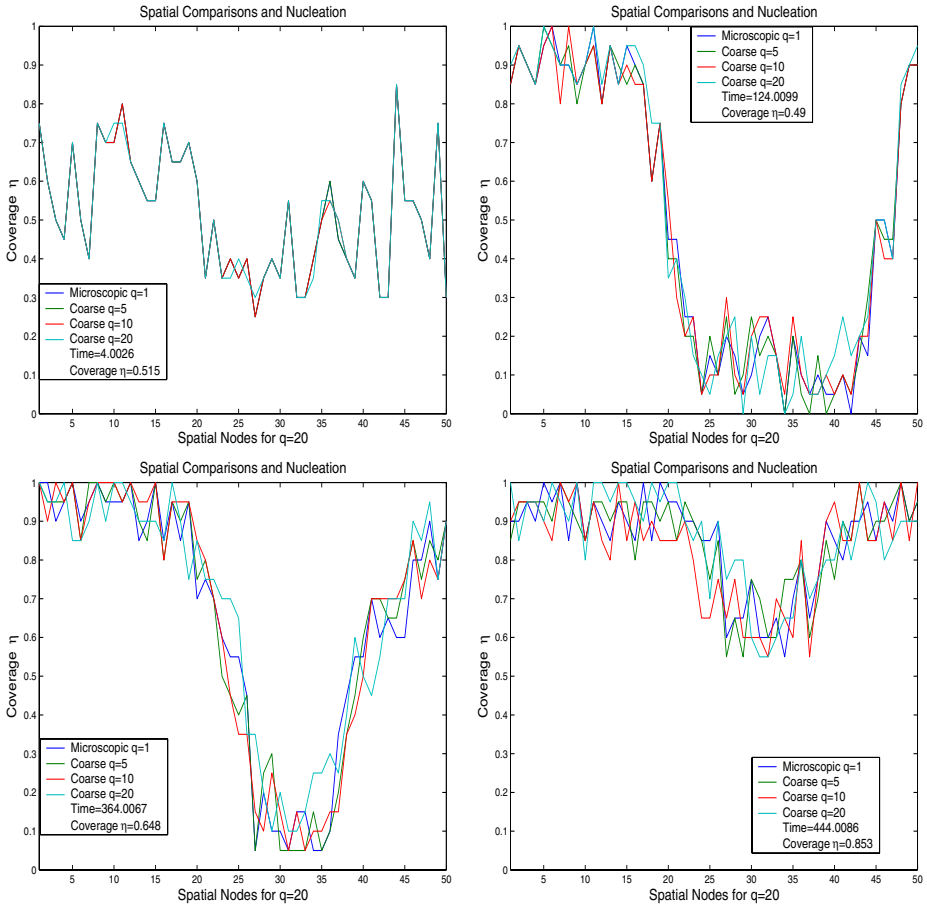


FIG. 5.6. *Snapshots of the transition from the initial state with the mean coverage at* 0.5. *Comparisons between the microscopic* $q = 1$ *and coarse-grained simulations* $q = 5, 10$ *and* $q = 20$. *The interaction radius is set to* $L = 100$, *the external field* $c_0 = 0.0492$, $d_0 = 1$, *and the total number of lattice sites* $N = 1000$.

estimate depends exponentially on the final time. On the other hand it was shown in [10] that in the case of Kac potentials the phases appear at length scales of the order $\log L$ as $L \to \infty$. Thus the error at spinodal decomposition times is controlled by a term of the order $\mathrm{O}(q^2/L)$.

Acknowledgments. M. A. Katsoulakis also acknowledges many valuable conversations with A. Szepessy. The authors would like to thank the Institute for Mathematics and its Applications where part of this work was carried out during the program "Mathematics of Materials and Macromolecules: Multiple Scales, Disorder, and Singularities."

REFERENCES

[1] D. BAI AND A. BRANDT, *Multiscale computation of polymer models*, in Multiscale Computational Methods in Chemistry and Physics, NATO Science Series: Computer and System

Sciences 177, A. Brandt, J. Bernholc, and K. Binder, eds., IOS Press, Amsterdam, 2001, pp. 250–266.

[2] P. Bernard, D. Talay, and L. Tubaro, *Rate of convergence of a stochastic particle method for the Kolmogorov equation with variable coefficients*, Math. Comp., 63 (1994), pp. 555–587, S11–S17.

[3] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms.* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[4] A. Brandt, *Multigrid methods in lattice field computations*, Nucl. Phys. B, 26 (1992), pp. 137–180.

[5] A. Brandt and V. Ilyin, *Multilevel Monte Carlo methods for studying large-scale phenomena in fluids*, J. Molecular Liquids, 105 (2003), pp. 253–256.

[6] A. Brandt and D. Ron, *Renormalization multigrid: Statistically optimal renormalization group flow and coarse-to-fine Monte Carlo acceleration*, J. Stat. Phys., 102 (2001), pp. 231–257.

[7] A. Brandt, D. Ron, and D. J. Amit, *Multilevel approaches to discrete-state and stochastic problems*, in Multigrid Methods, II, W. Hackbusch and U. Trottenberg, eds., Springer-Verlag, Berlin, 1986, pp. 66–99.

[8] A. J. Chorin, A. P. Kast, and R. Kupferman, *Optimal prediction of underresolved dynamics*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 4094–4098.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[10] A. DeMasi, E. Orlandi, E. Presutti, and L. Triolo, *Glauber evolution with Kac potentials.* III. *Spinodal decomposition*, Nonlinearity, 9 (1996), pp. 53–114.

[11] W. E and B. Engquist, *Multiscale modeling and computation*, Notices Amer. Math. Soc., 50 (2003), pp. 1062–1070.

[12] J. Goodman and A. D. Sokal, *Multigrid Monte Carlo methods for lattice field theories*, Phys. Rev. Lett., 56 (1986), pp. 1015–1018.

[13] Q. Hou, N. Goldenfeld, and A. McKane, *Renormalization group and perfect operators for stochastic differential equations*, Phys. Rev. E, 63 (2001), 036125.

[14] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for PDEs with oscillatory coefficients*, in Numerical Treatment of Multi-scale Problems (Kiel, 1997), Notes Numer. Fluid Mech. 70, Vieweg, Braunschweig, Germany, 1999, pp. 58–69.

[15] A. E. Ismail, G. C. Rutledge, and G. Stephanopoulos, *Multiresolution analysis in statistical mechanics.* I. *Using wavelets to calculate thermodynamic properties*, J. Chem. Phys., 118 (2003), pp. 4414–4423.

[16] L. P. Kadanoff, *Statistical Physics: Statics, Dynamics and Renormalization*, World Scientific, River Edge, NJ, 1999.

[17] M. Katsoulakis, P. Plecháč, L. Rey-Bellet, and D. Tsagkarogiannis, *Coarse-graining schemes and a posteriori estimates for stochastic lattice systems*, Math. Model. Numer. Anal., (2006), submitted.

[18] M. A. Katsoulakis, A. J. Majda, and A. Sopasakis, *Multiscale couplings in prototype hybrid deterministic/stochastic systems. Part* I. *Deterministic closures*, Commun. Math. Sci., 2 (2004), pp. 255–294.

[19] M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos, *Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems*, J. Comput. Phys., 186 (2003), pp. 250–278.

[20] M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos, *Coarse-grained stochastic processes for microscopic lattice systems*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 782–787.

[21] M. A. Katsoulakis and A. Szepessy, *Stochastic hydrodynamical limits of particle systems*, Comm. Math. Sci., 4 (2006), pp. 513–549.

[22] M. A. Katsoulakis and J. Trashorras, *Information loss in coarse-graining of stochastic particle dynamics*, J. Stat. Phys., 122 (2006), pp. 115–135.

[23] M. A. Katsoulakis and D. G. Vlachos, *Coarse-grained stochastic processes and kinetic Monte Carlo simulations for diffusion of interacting molecules*, J. Chem. Phys., 119 (2003), pp. 9412–9427.

[24] I. G. Kevrekidis, C. W. Gear, and G. Hummer, *Equation-free: The computer aided analysis of complex multiscale systems*, AIChE J., 50 (2004), pp. 1346–1355.

[25] C. Kipnis and C. Landim, *Scaling Limits of Interacting Particle Systems*, Springer-Verlag, Berlin, 1999.

[26] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, 3rd ed., Appl. Math. 23, Springer-Verlag, Berlin, 1999.

[27] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, Cambridge, UK, 2000.

[28] A. J. Majda, I. Timofeyev, and E. V. Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure Appl. Math., 54 (2001), pp. 891–974.
[29] F. Muller-Plathe, *Coarse-graining in polymer simulation: From the atomistic to the mesoscale and back*, Chem. Phys. Chem., 3 (2002), pp. 754–769.
[30] R. O'Handley, *Modern Magnetic Materials: Principles and Applications*, Wiley-Interscience, New York, 2000.
[31] S. Renisch, R. Schuster, J. Wintterlin, and G. Ertl, *Dynamics of adatom motion under influence of mutual interactions*, Phys. Rev. Lett., 82 (1999), pp. 3839–3842.
[32] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys., 151 (1999), pp. 146–168.
[33] A. Szepessy, R. Tempone, and G. E. Zouraris, *Adaptive weak approximation of stochastic differential equations*, Comm. Pure Appl. Math., 54 (2001), pp. 1169–1214.
[34] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Anal. Appl., 8 (1990), pp. 483–509 (1991).
[35] D. G. Vlachos, L. D. Schmidt, and R. Aris, *The effects of phase transitions, surface diffusion, and defects on surface catalyzed reactions: Fluctuations and oscillations*, J. Chem. Phys., 93 (1990), pp. 8306–8313.
[36] R. Zwanzig, *Nonequlibrium Statistical Mechanics*, Oxford University Press, New York, 2001.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2297–2325

LOCKING-FREE OPTIMAL DISCONTINUOUS GALERKIN METHODS FOR TIMOSHENKO BEAMS*

FATIH CELIKER†, BERNARDO COCKBURN†, AND HENRYK K. STOLARSKI‡

Abstract. In this paper, we consider the so-called *hp*-version of discontinuous Galerkin methods for Timoshenko beams. We prove that, when the numerical traces are properly chosen, the methods display optimal convergence uniformly with respect to the thickness of the beam. These methods are thus free from shear locking. We also prove that, when polynomials of degree p are used, *all* the numerical traces superconverge with a rate of order h^{2p+1}/p^{2p+1}. Numerical experiments verifying the above-mentioned theoretical results are shown.

Key words. discontinuous Galerkin methods, Timoshenko beams, locking, superconvergence

AMS subject classification. 65N30

DOI. 10.1137/050635821

1. Introduction.

In this paper, we present the first rigorous analysis of discontinuous Galerkin (DG) methods for the Timoshenko beam model [24]

$$\begin{aligned}
\frac{d\bar{w}(\bar{x})}{d\bar{x}} &= \bar{\theta}(\bar{x}) - \frac{\bar{T}(\bar{x})}{(\bar{G}\bar{A})(\bar{x})},\\
\frac{d\bar{\theta}(\bar{x})}{d\bar{x}} &= \frac{\bar{M}(\bar{x})}{(\bar{E}\bar{I})(\bar{x})},\\
\frac{d\bar{M}(\bar{x})}{d\bar{x}} &= \bar{T}(\bar{x}),\\
\frac{d\bar{T}(\bar{x})}{d\bar{x}} &= \bar{q}(\bar{x})
\end{aligned}$$

for all $\bar{x} \in \bar{\Omega} := (0, L)$. Here, the unknowns are the transverse displacement $\bar{w}$, the rotation of the transverse cross-section of the beam $\bar{\theta}$, the bending moment $\bar{M}$, and the shear force $\bar{T}$. The material and geometrical properties of the beam are characterized by the shear modulus $\bar{G}$, the cross-section area $\bar{A}$, the Young modulus $\bar{E}$, and the moment of inertia $\bar{I}$. The remaining data of the problem are the transverse load, $\bar{q}$, and the boundary conditions which we take to be

$$\bar{w}(0) = \bar{w}_0, \qquad \bar{\theta}(0) = \bar{\theta}_0, \qquad \bar{w}(L) = \bar{w}_L, \quad \text{and} \quad \bar{\theta}(L) = \bar{\theta}_L.$$

The main motivation for considering this simple, one-dimensional model is that it constitutes a stepping stone towards the more challenging goal of devising DG methods for shells. The construction of numerical methods for shells is delicate because, as the thickness of the shell decreases to zero, the numerical method can exhibit what is

*Received by the editors July 12, 2005; accepted for publication (in revised form) May 26, 2006; published electronically November 24, 2006.

http://www.siam.org/journals/sinum/44-6/63582.html

†School of Mathematics, University of Minnesota, 206 Church Street S.E., Minneapolis, MN 55455 (celiker@math.umn.edu, cockburn@math.umn.edu). The research of the second author was partially supported by the National Science Foundation (grant DMS-0107609) and by the University of Minnesota Supercomputer Institute.

‡Department of Civil Engineering, University of Minnesota, 500 Pillsbury Drive S.E., Minneapolis, MN 55455 (stola001@tc.umn.edu).

known in the engineering literature as *shear* and *membrane locking*. Mathematically, this is reflected in the deterioration of the convergence properties of the method as the thickness becomes small. Since some numerical methods for the Timoshenko beam model exhibit (shear) locking (as the thickness of the beam goes to zero), it is instructive to devise locking-free DG methods for this model before considering shells.

A considerable amount of effort has been devoted to the understanding and resolution of shear and membrane locking in structures. Considering the nature of the problem, it is understandable that such effort originated in engineering applications and was first documented in the engineering literature. The seminal publication in the area, by Zienkiewicz, Taylor, and Too [28], documents the difficulty related to shear effects and uses the so-called "reduced integration" technique to mitigate the problem. The physical understanding of the problem was critical to devise a remedy, and the resulting technique (reduced integration) is to this day widely used in various commercial software. The term "shear locking" appears to have been coined by Hughes, Taylor, and Kanoknukulchai [13] in the context of plate analysis.

In parallel with developments related to shear locking, researchers struggled with similar difficulties caused by membrane effects, manifesting themselves in curved structures, such as arches and shells; see, for example, Ashwell and Sabir [4], Lee and Pian [14], and Parisch [17]. A more thorough explanation of those effects was provided by Stolarski and Belytschko [21], who also introduced the term "membrane locking." They subsequently showed that in some models of curved structures there is a delicate interaction between shear and membrane effects [22].

Over the last two decades or so, there has been a flurry of research activities dealing with shear and membrane locking, and a large number of publications have appeared. Several variations of the known approaches and a number of new ones were developed and described in literature within that time. While related to this work, those approaches address the problem of locking somewhat differently from what we describe here; the interested reader is therefore referred to [23] for a review of many of them. For a locking-free finite element method for shells we refer to Arnold and Brezzi [2], and to Arnold, Brezzi, and Marini [3] for a recently uncovered family of locking-free DG methods for the Reissner–Mindlin plates.

While deeply rooted in physical attributes of the analyzed phenomena, locking is essentially a mathematical problem and its challenge was undertaken by mathematicians early on. Arnold [1] proved that shear locking continuous finite element methods can become locking-free if they are modified by the reduced integration technique. His method of proving error estimates independent of the thickness follows from an equivalence between certain mixed methods and the reduced integration technique; for a discussion of this equivalence in a more general setting we refer to [16]. In [15], Li analyzed the p- and hp-versions of the continuous finite element method and proved error estimates independent of the thickness of the beam. These versions of the method take advantage of the extra degrees of freedom gained by increasing the polynomial degree of the approximation. In [25], [26], and [27], Zhang considered circular arch problems. Here shear locking (and also membrane locking) is again an issue when the arch is thin. Indeed, if the primal form of the method is used where the only unknowns are the displacement and the rotation, both p- and hp-versions exhibit locking. On the other hand, if the shear force is introduced as an additional unknown, along with the membrane forces, and a mixed formulation is employed, then both versions can be made free from locking. Following an approach similar to that of Arnold's, Zhang [25], [26], [27] was able to prove error estimates independent of the thickness of the arch.

In [10], the DG methods for the Timoshenko beams were introduced and sufficient conditions that ensure the existence and uniqueness of their approximate solutions were proved. Moreover, preliminary numerical experiments were obtained which indicated that, when polynomials of degree p are used, the optimal order of convergence of $p+1$ is achieved for the h-version; exponential convergence for the p-version of a DG method was also obtained numerically. Later, in [8], the fact that *all* the numerical traces of the h-version of the DG method superconverge with order $2p+1$ was uncovered, and a local postprocessing resulting in a uniformly accurate solution of order $2p+1$ was devised and numerically tested. These results hold uniformly with respect to the thickness of the beam. In this paper, we put all the above-mentioned numerical results on firm mathematical ground.

The rest of the paper is organized as follows. In section 2, we describe a large class of DG methods and briefly discuss the existence and uniqueness of their approximate solution. Then, we pick a particular DG method for which we state and discuss the main results of our a priori error analysis. This DG method is particularly difficult to analyze due to that fact that it has practically no jump-stabilization terms associated to the interelement boundaries. Section 3 is devoted to the proof of those results and section 4 to some extensions of the main results. Numerical results verifying the theoretical results are presented in section 5. We end in section 6 with some concluding remarks.

2. Main results.

2.1. The dimensionless form of the model. To carry out our analysis, we nondimensionalize the equations of the model. We set $x = \bar{x}/L$, $\Omega = (0,1)$, $w(x) = \bar{w}(\bar{x})/L$, $w_0 = \bar{w}_0/L$, $w_1 = \bar{w}_L/L$, $\theta_0 = \bar{\theta}_0$, $\theta_1 = \bar{\theta}_L$. Suppose that there exist four constants a^*, b^*, E^*, and G^* such that

$$C_1 \le \frac{a^*}{\bar{a}(\bar{x})}, \frac{b^*}{\bar{b}(\bar{x})}, \frac{E^*}{\bar{E}(\bar{x})}, \frac{G^*}{\bar{G}(\bar{x})} \le C_2 \quad \forall\ \bar{x} \in \bar{\Omega}, \tag{2.1}$$

where $\bar{a}(\bar{x})$ and $\bar{b}(\bar{x})$ are the depth and thickness of the beam at the point $\bar{x}$, respectively. We further introduce $A^* := a^* b^*$ and $I^* := a^*(b^*)^3/12$ and then set $M(x) = \bar{M}(\bar{x})L/(E^*I^*)$, $T(x) = \bar{T}(\bar{x})\,L^2/(E^*I^*)$, $(EI)(x) = (\bar{E}\bar{I})(\bar{x})/(E^*I^*)$, $(GA)(x) = (\bar{G}\bar{A})(\bar{x})/(G^*A^*)$, and $q(x) = \bar{q}(\bar{x})\,L^3/(E^*I^*)$. We then rewrite the equations as

$$\frac{dw}{dx} = \theta - d^2\frac{T}{GA}, \qquad \frac{d\theta}{dx} = \frac{M}{EI}, \qquad \frac{dM}{dx} = T, \qquad \frac{dT}{dx} = q \qquad \text{in } \Omega, \tag{2.2}$$

where

$$w(0) = w_0, \qquad w(1) = w_1, \qquad \theta(0) = \theta_0, \quad \text{and} \quad \theta(1) = \theta_1. \tag{2.3}$$

Here,

$$d^2 := \frac{E^*I^*}{G^*A^*L^2} = \frac{E^*}{12\,G^*}\left(\frac{b^*}{L}\right)^2. \tag{2.4}$$

Thus the parameter d is proportional to the thickness of the beam to its length, and for small d the equations (2.2) model a thin beam; if the numerical method is not properly devised, shear locking might occur when the parameter d goes to zero. For further discussion of the locking effects in the finite element method, we refer the reader to Babuška and Suri [5].

2.2. General DG methods. To define the DG methods, we follow [10]. We begin by partitioning the computational domain into intervals. Given the set of *nodes* $\mathscr{E}_h := \{x_i\}_{i=0}^N$, where $0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1$, we set $I_i := (x_{i-1}, x_i)$, $h_i := x_i - x_{i-1}$ and $h := \max_{1\le i\le N} h_i$.

The approximate solution $(T_h, M_h, \theta_h, w_h)$ given by the DG method is sought in the finite dimensional space $V_h^{p_1} \times V_h^{p_2} \times V_h^{p_3} \times V_h^{p_4}$, where

$$V_h^p := \{v : \Omega_h \mapsto \mathbb{R} : \quad v|_{I_j} \in P^p(I_j), \quad j = 1, \dots, N\}$$

and $P^p(K)$ is the set of all polynomials on K of degree not exceeding p. It is determined by requiring that

$$-(w_h, v_1')_{\Omega_h} + \langle \widehat{w}_h, [\![v_1\, n]\!]\rangle_{\mathscr{E}_h} = (\theta_h, v_1)_{\Omega_h} - d^2\,(T_h/GA, v_1)_{\Omega_h}\,, \tag{2.5a}$$

$$-(\theta_h, v_2')_{\Omega_h} + \langle \widehat{\theta}_h, [\![v_2\, n]\!]\rangle_{\mathscr{E}_h} = (M_h/EI, v_2)_{\Omega_h}\,, \tag{2.5b}$$

$$-(M_h, v_3')_{\Omega_h} + \langle \widehat{M}_h, [\![v_3\, n]\!]\rangle_{\mathscr{E}_h} = (T_h, v_3)_{\Omega_h}\,, \tag{2.5c}$$

$$-(T_h, v_4')_{\Omega_h} + \langle \widehat{T}_h, [\![v_4\, n]\!]\rangle_{\mathscr{E}_h} = (q, v_4)_{\Omega_h} \tag{2.5d}$$

hold for all $v_i \in V_h^{p_i}$ for $i = 1, 2, 3, 4$. Here, $\Omega_h = \cup_{j=1,\dots,N} I_j$ and

$$(u, v)_{\Omega_h} := \sum_{j=1}^N (u, v)_{I_j}, \quad \text{where} \quad (u, v)_{I_j} := \int_{I_j} u(x)v(x)\,dx.$$

Moreover, we introduce

$$\langle R, [\![u\, n]\!]\rangle_{\mathscr{E}_h} := \sum_{j=0}^N R(x_j)\,[\![u\, n]\!](x_j),$$

where R is any function defined on the set of nodes $\mathscr{E}_h$ and $[\![u\, n]\!]$ is the *jump* of function u across nodes which is defined as follows:

$$[\![u\, n]\!](x_j) = \begin{cases} -u(0^+) & \text{for } j = 0, \\ -u(x_j^+) + u(x_j^-) & \text{for } 0 < j < N, \\ +u(1^-) & \text{for } j = N. \end{cases}$$

Here, $u(x_j^\pm) := \lim_{\epsilon\downarrow 0} u(x_j \pm \epsilon)$.

To complete the definition of the method, we have to define the numerical traces $(\widehat{T}_h, \widehat{M}_h, \widehat{\theta}_h, \widehat{w}_h)$ at the nodes. We assume that the general form of these traces is as follows. For $x_i \in \mathscr{E}_h^\circ := \{x_1, x_2, \dots, x_{N-1}\}$, we take

$$\begin{aligned}
\widehat{w}_h &= \{\!\{\, w_h \,\}\!\} + C_{11}[\![w_h\, n]\!] + C_{12}[\![\theta_h\, n]\!] + C_{13}[\![M_h\, n]\!] + C_{14}[\![T_h\, n]\!], \\
\widehat{\theta}_h &= \{\!\{\, \theta_h \,\}\!\} + C_{21}[\![w_h\, n]\!] + C_{22}[\![\theta_h\, n]\!] + C_{23}[\![M_h\, n]\!] + C_{24}[\![T_h\, n]\!], \\
\widehat{M}_h &= \{\!\{M_h\}\!\} + C_{31}[\![w_h\, n]\!] + C_{32}[\![\theta_h\, n]\!] + C_{33}[\![M_h\, n]\!] + C_{34}[\![T_h\, n]\!], \\
\widehat{T}_h &= \{\!\{\, T_h \,\}\!\} + C_{41}[\![w_h\, n]\!] + C_{42}[\![\theta_h\, n]\!] + C_{43}[\![M_h\, n]\!] + C_{44}[\![T_h\, n]\!],
\end{aligned} \tag{2.6}$$

where $\{\!\{\varphi\}\!\}(x_i) := \frac{1}{2}(\varphi(x_i^+) + \varphi(x_i^-))$. At $x = 0$, we take

$$\begin{aligned}
\widehat{w}_h(0) &= w_0, \\
\widehat{\theta}_h(0) &= \theta_0, \\
\widehat{M}_h(0) &= M_h(0^+) + C_{31}(0)(w_0 - w_h(0^+)) + C_{32}(0)(\theta_0 - \theta_h(0^+)), \\
\widehat{T}_h(0) &= T_h(0^+) + C_{41}(0)(w_0 - w_h(0^+)) + C_{42}(0)(\theta_0 - \theta_h(0^+)).
\end{aligned} \tag{2.7}$$

And at $x = 1$,

$$
\begin{aligned}
&\widehat{w}_h(1) = w_1,\\
&\widehat{\theta}_h(1) = \theta_1,\\
&\widehat{M_h}(1) = M_h(1^-) + C_{31}(1)(w_h(1^-) - w_1) + C_{32}(1)(\theta_h(1^-) - \theta_1),\\
&\widehat{T_h}(1) = T_h(1^-) \ + C_{41}(1)(w_h(1^-) - w_1) + C_{42}(1)(\theta_h(1^-) - \theta_1).
\end{aligned} \tag{2.8}
$$

The definition of the DG method is now complete.

This method has a unique solution provided the parameters C_{ij}, $i, j = 1, 2, 3, 4$, and the polynomial degrees p_i, $i = 1, 2, 3, 4$, are suitably chosen. The following theorem proven in [10] gives sufficient conditions for this to happen.

THEOREM 2.1 (existence and uniqueness of the DG approximation). *Consider the DG method defined by the weak formulation* (2.5) *and the numerical traces* (2.6), (2.7), *and* (2.8). *Assume that*

$$
C_{21} = C_{43},\ -C_{22} = C_{33},\ C_{24} = C_{13},\ C_{31} = C_{42},\ C_{34} = C_{12},\ -C_{11} = C_{44}, \tag{2.9}
$$

and that

$$
C_{14},\ -C_{23},\ -C_{32},\ C_{41} \geq 0. \tag{2.10}
$$

Then the method has a unique solution in the following cases:

Case 1: $C_{41}, -C_{32} > 0$ *on* $\mathscr{E}_h$, $p_2 \geq p_3 - 1$, *and* $p_1 \geq p_4 - 1$.

Case 2: $C_{ij} = 0$ *on* $\mathscr{E}_h^\circ$, *except* $C_{11} = C_{22} = -C_{33} = -C_{44} = 1/2$, $C_{41}(1) > 0$, $-C_{32}(1) > 0$, $p_2 \geq p_3$, *and* $p_1 \geq p_4$.

Case 3: $p_2 \geq p_3 + 1$ *and* $p_1 \geq p_4 + 1$.

Note that thanks to equations (2.9), only 10 of the 16 coefficients C_{ij} are independent. Moreover, the condition (2.10) states that four of those must have a specific sign. This last condition can be better understood thanks to the so-called discrete energy equality.

PROPOSITION 2.2 (discrete energy identity [10]). *Assume that the hypotheses of Theorem* 2.1 *hold. Then*

$$
(M_h/EI, M_h)_{\Omega_h} + d^2\,(T_h/GA, T_h)_{\Omega_h} + \Theta_{jumps} = (q, w_h)_{\Omega_h} + bc + \Theta_{bc,h}, \tag{2.11}
$$

where

$$
\begin{aligned}
bc &= w_0\,T_h(0^+) - w_1\,T_h(1^-) - \theta_0\,M_h(0^+) + \theta_1\,M_h(1^-),\\
\Theta_{bc,h} &= w_0\left[C_{41}(0)w_h(0^+) - C_{31}(0)\theta_h(0^+)\right] + \theta_0[C_{42}(0)w_h(0^+) - C_{32}(0)\theta_h(0^+)]\\
&\quad + w_1\left[C_{41}(1)w_h(1^-) - C_{31}(1)\theta_h(1^-)\right] + \theta_1[C_{42}(1)w_h(1^-) - C_{32}(1)\theta_h(1^-)]\,,\\
\Theta_{jumps} &= \sum_{x_i \in \mathscr{E}_h} \left(C_{14}[\![T_h\,n]\!]^2 - C_{23}[\![M_h\,n]\!]^2 - C_{32}[\![\theta_h\,n]\!]^2 + C_{41}[\![w_h\,n]\!]^2\right)(x_i),
\end{aligned}
$$

where we set $C_{14} = C_{23} = 0$ *at the boundary nodes.*

We can thus see that the four coefficients that appear in the condition (2.10) are precisely those associated with the energy produced by the jumps of the approximations; they can also be thought of as penalizing the corresponding jumps. As a consequence, if we penalize the jumps "too much," the DG method might behave like a typical continuous method and might lock: It would produce very bad approximations for small values of d. On the contrary, if these penalization parameters are chosen appropriately, the DG method will produce a very good approximation.

We illustrate this phenomenon in Figure 1. Therein, we display the exact solution corresponding to $q(x) = e^x, (EI)(x) = e^x, (GA)(x) = e^{-x}$, together with homogeneous boundary conditions $w_0 = w_1 = \theta_0 = \theta_1 = 0$. We also show approximations by two of the DG methods just described. Both methods take

$$C_{11}(x) = C_{22}(x) = -C_{33}(x) = -C_{44}(x) = 1/2,$$

at all interior nodes $x \in \mathscr{E}_h^\circ$, and all the remaining coefficients equal to zero, except for C_{32} and C_{41}. The first DG method *strongly* penalizes the jumps of the vertical displacement w and the rotation θ, since it takes

$$-C_{32}(x) = C_{41}(x) = 10^6$$

for all nodes. We can see in Figure 1, left column, that, as expected, it locks. The second method, however, does *not* penalize those jumps *at all* since it takes

$$C_{32}(x) = C_{41}(x) = 0$$

at all the nodes except at $x = 1$. At $x = 1$, it takes

$$-C_{32}(1) = C_{41}(1) = 16/h$$

to enforce the Dirichlet boundary condition there. In Figure 1, right column, we can see that the method produces an excellent approximation of the exact solution. In this paper, we study this method in detail; other shear-locking-free DG methods are briefly discussed in section 4.

2.3. A priori error estimates. In this section, we present and briefly discuss a priori error estimates for the DG method obtained by setting

$$C_{11} = C_{22} = -C_{33} = -C_{44} \,=\, 1/2$$

at all interior nodes,

$$-C_{32}(1) = C_{41}(1) \;=\; \mathsf{c}\,\frac{\mathsf{p}}{h_N},$$

and all of the remaining coefficients to zero. Here c is a positive real number and $\mathsf{p} := \max\{1, p\}$. We also assume that $(T_h, M_h, \theta_h, w_h) \in [V_h^p]^4$, where $p \geq 0$. A simple computation gives that, for an interior node $x_j \in \mathscr{E}_h^\circ$,

$$\begin{aligned} &\widehat{w}_h(x_j) = w_h(x_j^-), \quad \widehat{M}_h(x_j) = M_h(x_j^+), \\ &\widehat{\theta}_h(x_j) = \theta_h(x_j^-), \quad \widehat{T}_h(x_j) = T_h(x_j^+), \end{aligned} \tag{2.12}$$

that, at $x = 0$,

$$\begin{aligned} &\widehat{w}_h(0) = w_0, \quad \widehat{M}_h(0) = M_h(0^+), \\ &\widehat{\theta}_h(0) = \theta_0, \quad \widehat{T}_h(0) = T_h(0^+), \end{aligned} \tag{2.13}$$

and that, at $x = 1$,

$$\begin{aligned} &\widehat{w}_h(1) = w_1, \quad \widehat{M}_h(1) = M_h(1^-) - \mathsf{c}\frac{\mathsf{p}}{h_N}(\theta_h(1^-) - \theta_1), \\ &\widehat{\theta}_h(1) = \theta_1, \quad \widehat{T}_h(1) = T_h(1^-) + \mathsf{c}\frac{\mathsf{p}}{h_N}(w_h(1^-) - w_1). \end{aligned} \tag{2.14}$$

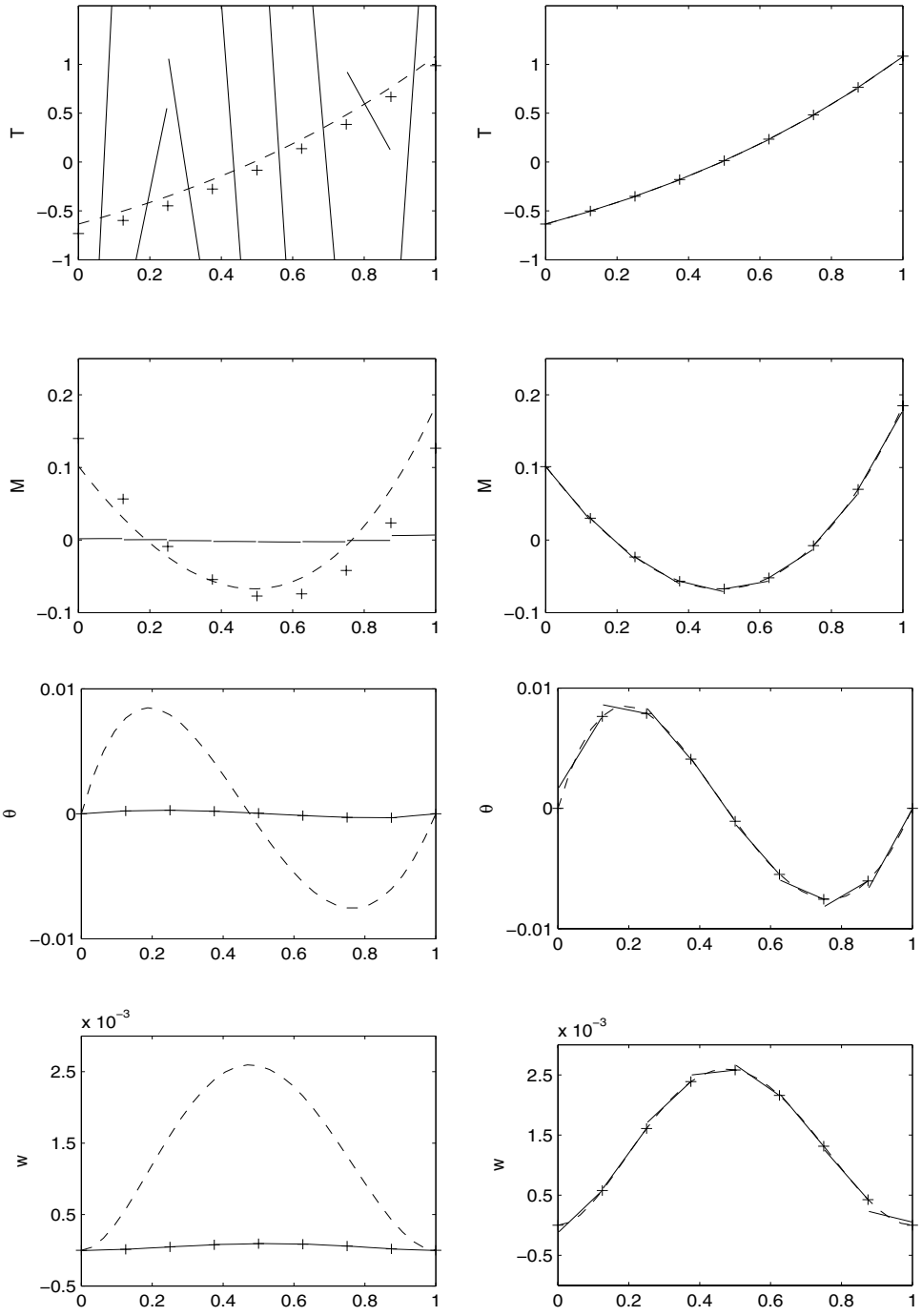


FIG. 1. *The case $d = 10^{-3}$ and $h = 1/8$: Exact (dashed line) and DG approximation (solid line and, for the numerical traces, +). Left column: $-C_{32} = C_{41} = 10^6$ on all the nodes. Right column: $C_{32} = C_{41} = 0$ on all the nodes except at $x = 1$, and $-C_{32}(1) = C_{41}(1) = 16/h = 128$.*

A similar DG method for convection-diffusion problems was introduced in [11]; its *hp*-version was analyzed first in [6] and its superconvergence properties proved later in [9].

We have chosen to analyze this method for two main reasons. The first is that it gives better error estimates in the energy seminorm (see section 4.3), the second is that it is more difficult to analyze. To give a brief idea why it is so, we note that for this DG method the discrete energy associated with the jumps is

$$\Theta_{jumps} = \mathsf{c}\frac{\mathsf{p}}{h_N}\left([\![\theta_h\, n]\!]^2(1) + [\![w_h\, n]\!]^2(1)\right).$$

We thus see that the DG method *penalizes* the jumps of w_h and θ_h only at $x = 1$. This penalization weakly enforces the boundary conditions at $x = 1$ and ensures existence and uniqueness of the approximate solution; see Case 2 of Theorem 2.1. Note, however, that no such penalization is enforced at any other node of the mesh, and in particular at the border $x = 0$. This lack of extra *stabilization* is what renders the analysis of this method more challenging and interesting than that of the other DG methods. Technical details of how the analysis differs and becomes simpler when stabilization is present at all nodes of the mesh are given in Celiker [7].

To state our main results we need to introduce some notation. We begin by setting

$$\begin{aligned}|(u_1,u_2,u_3,u_4)|^2_{\mathscr{A}_h} :=\ & (u_2/EI,u_2)_{\Omega_h} + d^2\,(u_1/GA,u_1)_{\Omega_h}\\ & + \mathsf{c}\frac{\mathsf{p}}{h_N}\left([\![u_3\, n]\!]^2(1) + [\![u_4\, n]\!]^2(1)\right)\end{aligned} \tag{2.15}$$

for all $(u_1,u_2,u_3,u_4) \in [H^1(\Omega_h)]^4$. Since we can rewrite the discrete energy identity of Proposition 2.2 as

$$|(T_h,M_h,\theta_h,w_h)|^2_{\mathscr{A}_h} = (q,w_h)_{\Omega_h} + bc + \Theta_{bc,h},$$

we call this seminorm, the *energy* seminorm. The estimate of the approximation error in this seminorm plays a fundamental role in our error analysis.

Next, we define Green's functions for the problem under consideration. For any superindex $\star = T, M, \theta$ or w, and any point $y \in (0,1)$, we define $(\varphi^\star_{T,y}, \varphi^\star_{M,y}, \varphi^\star_{\theta,y}, \varphi^\star_{w,y})$ as the solution of

$$-\frac{d\varphi^\star_{w,y}}{dx} = \varphi^\star_{\theta,y} - d^2\frac{\varphi^\star_{T,y}}{GA}, \quad -\frac{d\varphi^\star_{\theta,y}}{dx} = \frac{\varphi^\star_{M,y}}{EI}, \quad -\frac{d\varphi^\star_{M,y}}{dx} = \varphi^\star_{T,y}, \quad -\frac{d\varphi^\star_{T,y}}{dx} = 0 \tag{2.16}$$

in $(0,y)\cup(y,1)$ that satisfies the boundary conditions

$$\varphi^\star_{w,y}(0) = \varphi^\star_{w,y}(1) = \varphi^\star_{\theta,y}(0) = \varphi^\star_{\theta,y}(1) = 0 \tag{2.17}$$

and the jump conditions

$$[\![\varphi^\star_{w,y}\, n]\!](y) = \delta_{\star\, T}, \quad [\![\varphi^\star_{\theta,y}\, n]\!](y) = \delta_{\star\, M}, \quad [\![\varphi^\star_{M,y}\, n]\!](y) = \delta_{\star\,\theta}, \quad [\![\varphi^\star_{T,y}\, n]\!](y) = \delta_{\star\, w}. \tag{2.18}$$

Here, $\delta_{ab} = 1$ if $a = b$ and $\delta_{ab} = 0$ otherwise. We also define, for $z \in \{0,1\}$,

$$(\varphi^\star_{T,z}, \varphi^\star_{M,z}, \varphi^\star_{\theta,z}, \varphi^\star_{w,z}) := \lim_{y\to z}(\varphi^\star_{T,y}, \varphi^\star_{M,y}, \varphi^\star_{\theta,y}, \varphi^\star_{w,y}).$$

Note that this implies that the function $(\varphi^\star_{T,z}, \varphi^\star_{M,z}, \varphi^\star_{\theta,z}, \varphi^\star_{w,z})$ is identically equal to zero for $z \in \{0,1\}$ and $\star = \theta, w$.

We denote by $\|\cdot\|_{s,\Omega_h}$ and $|\cdot|_{s,\Omega_h}$ the usual norm and seminorm in $H^s(\Omega_h)$, respectively. We set

$$\begin{aligned}|\,(T,M,\theta,w)\,|_{s,p,\Omega_h} := {} & |\,T\,|_{\min(p,s)+1,\Omega_h} + |\,M\,|_{\min(p,s+1)+1,\Omega_h} \\ & + |\,\theta\,|_{\min(p,s+2)+1,\Omega_h} + |\,w\,|_{\min(p,s+1)+1,\Omega_h}.\end{aligned}$$

The use of this seminorm is motivated by the fact that if the load q is assumed to belong to $H^s(\Omega_h)$ and the functions EI and GA are very smooth in Ω_h, then (2.2), (2.3), and (2.4) indicate that the solution (T,M,θ,w) belongs to the Sobolev space $H^{s+1}(\Omega_h) \times H^{s+2}(\Omega_h) \times H^{s+3}(\Omega_h) \times H^{s+2}(\Omega_h)$.

Finally, for any real number $k \geq 0$, we define

$$|\varphi^\star_{x_i}|_{k,\Omega_h} := \max\{|\varphi^\star_{T,x_i}|_{k,\Omega_h}, |\varphi^\star_{M,x_i}|_{k,\Omega_h}, |\varphi^\star_{\theta,x_i}|_{k,\Omega_h}, |\varphi^\star_{w,x_i}|_{k,\Omega_h}\},$$

where $\star = T$, M, θ or w.

We are ready to state and discuss our main results.

THEOREM 2.3. *Assume that, for some $s \geq 0$, (T,M,θ,w) belong to $H^{s+1}(\Omega_h) \times H^{s+2}(\Omega_h) \times H^{s+3}(\Omega_h) \times H^{s+2}(\Omega_h)$. Set*

$$\boldsymbol{e} := (e_T, e_M, e_\theta, e_w) = (T - T_h, M - M_h, \theta - \theta_h, w - w_h),$$

where $(T_h, M_h, \theta_h, w_h)$ is the approximation given by the DG method (2.5), (2.12), (2.13), *and* (2.14) *with $p \geq 1$. Then, for small enough h or big enough p, we have that*

$$|\boldsymbol{e}|_{\mathscr{A}_h} \leq C_s \, \frac{h^{\min(p,s)+1}}{p^{\min(p,s)+1}} \, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h},$$

and that

$$\begin{aligned}\|e_T\|_{0,\Omega_h} &\leq C_s \, \frac{h^{\min(p,s)+1}}{p^{\min(p,s)+1}} (1 + |\varphi^T_{x_N}|_{p+1,\Omega_h}) \, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h}, \\ \|e_M\|_{0,\Omega_h} + \|e_\theta\|_{0,\Omega_h} + \|e_w\|_{0,\Omega_h} &\leq C_s \, \frac{h^{\min(p,s)+1}}{p^{\min(p,s)+1}} \, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h},\end{aligned}$$

for some constant C_s independent of p, h, and d.

Our second result estimates the errors at each of the nodes.

THEOREM 2.4. *With the same hypotheses as those of Theorem* 2.3, *we have that*

$$|\,\widehat{e_u}(x_i)\,| \leq C_s \, \frac{h^{\min(p,s)+p+1}}{p^{\min(p,s)+p+1}} |\varphi^u_{x_i}|_{p+1,\Omega_h} \, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h}$$

for $u = T, M, \theta$ or w and all the nodes x_i.

Let us emphasize that (not reported) numerical experiments indicate that the condition "small enough h or big enough p" in the results above seems to be unnecessary. It stems from a technicality in our proof and, most probably, could be removed.

Note that all of the estimates appearing in the above theorems show that, for $p \geq 1$, the DG method under consideration is *locking-free* because the constants appearing on the right-hand side of all the estimates do not depend on the parameter

d and because the seminorms appearing on the right-hand side of the estimates can be bounded uniformly with respect to d. Similar results holds for $p = 0$; see section 4.

To see that the seminorms of the exact solution can be uniformly bounded with respect to d, we note that the exact solution is given by

$$
\begin{aligned}
T(x) &= \int_0^x q(r)\,dr + c_1,\\
M(x) &= \int_0^x \int_0^r q(s)\,ds\,dr + c_1 x + c_2,\\
\theta(x) &= \int_0^x \frac{1}{EI(r)} \left[\int_0^r \int_0^s q(t)\,dt\,ds + c_1 r + c_2 \right] dr + c_3,\\
w(x) &= \int_0^x \left[\int_0^r \frac{1}{EI(s)} \left(\int_0^s \int_0^t q(u)\,du\,dt + c_1 s + c_2 \right) ds + c_3 \right] dr\\
&\quad - d^2 \left[\int_0^x \frac{1}{GA(r)} \left(\int_0^r q(s)\,ds + c_1 \right) dr \right] + c_4,
\end{aligned}
$$

where c_1, c_2, c_3, and c_4 are integration constants to be determined by the boundary conditions (2.3). Since EI and GA are piecewise very smooth functions of order one, and d^2 appears only as a factor in $w(x)$, it is easy to see that the seminorms $|T|_{\min(p,s)+1,\Omega_h}$, $|M|_{\min(p,s+1)+1,\Omega_h}$, $|\theta|_{\min(p,s+2)+1,\Omega_h}$, and $|w|_{\min(p,s+1)+1,\Omega_h}$ remain bounded as $d \to 0$. A similar remark is valid for the seminorm $|\varphi^u_{x_i}|_{p+1,\Omega_h}$.

Note also that the above results imply that the h-version of the DG method converges with the optimal order of $p+1$ in the energy norm and in the L^2-norm for *all* variables. They also imply that *all* the numerical traces superconverge with order $2p+1$ at each node. This puts on firm mathematical ground the assumption of superconvergence taken in [8].

In the p-version of the DG method, we have spectral convergence with a rate of order p^{-s-1} for *all* the unknowns. In the case of a piecewise analytic function, by using the approach used in, for example, [20], [18], [19], it is possible to prove that the p-version of the DG method actually achieves exponential rates of convergence,

$$|(e_T, e_M, e_\theta, e_w)|_{\mathscr{A}_h} \le Ce^{-bp},$$

and

$$\|e_u\|_{0,\Omega_h} + \|\widehat{e_u}\|_{L^\infty(\mathscr{E}_h)} \le Ce^{-bp}$$

for $u = T, M, \theta$ or w, for some constants C and b independent of p and d.

Finally, let us point out that if the data E, G, A, and I are constants, then, for $p \ge 3$, Theorem 2.4 implies that, for any node x_i,

$$\widehat{e_u}(x_i) = 0,$$

for $u = T, M, \theta$ or w. Indeed, in this case the Green's functions are piecewise polynomials of degree at most 3 and hence $|\varphi^u_{x_i}|_{p+1,\Omega_h} = 0$.

2.4. Sketch of the proofs. Next, we give a brief outline of the main steps of our proofs. We proceed in three steps.

We begin by estimating the errors in the L^2-norm in terms of the error in the energy seminorm and the error in the numerical traces.

LEMMA 2.5. *We have*

$$\begin{aligned}
\|e_w\|_{0,\Omega_h} &\le C_s \left(|\boldsymbol{e}|_{\mathscr{A}_h} + \|e_\theta\|_{0,\Omega_h} + \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\, (T,M,\theta,w) \,|_{s,p,\Omega_h} \right), \\
\|e_\theta\|_{0,\Omega_h} &\le C_s \left(|\boldsymbol{e}|_{\mathscr{A}_h} + \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\, (T,M,\theta,w) \,|_{s,p,\Omega_h} \right), \\
\|e_M\|_{0,\Omega_h} &\le C_s \quad |\boldsymbol{e}|_{\mathscr{A}_h}, \\
\|e_T\|_{0,\Omega_h} &\le |\widehat{e_T}(1)| + C_s \, \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\, (T,M,\theta,w) \,|_{s,p,\Omega_h}.
\end{aligned}$$

Here C_s is a constant independent of p, h, and d.

Next, we show that the error in the numerical traces can be estimated in terms of the error in the energy seminorm.

LEMMA 2.6. *For sufficiently small mesh-size h or sufficiently big polynomial degree p, we have*

$$|\, \widehat{e_u}(x_i) \,| \le C_s \left(|\boldsymbol{e}|_{\mathscr{A}_h} + \frac{h^{\min(s,p)}}{\mathsf{p}^{\min(s,p)}} \,|\, (T,M,\theta,w) \,|_{s,p,\Omega_h} \right) \frac{h^{p+1}}{\mathsf{p}^{p+1}} \,|\varphi^u_{x_i}|_{p+1,\Omega_h}$$

for $u = T, M, \theta$ or w and for all nodes x_i. Here C_s is a constant independent of p, h, and d.

Finally, we obtain an estimate of the error estimate in the energy seminorm.

LEMMA 2.7. *With the same notation as in Theorem* 2.3, *we have that*

$$|\boldsymbol{e}|_{\mathscr{A}_h} \le C_s \, \frac{h^{\min(s,p)+1}}{\mathsf{p}^{\min(s,p)+1}} \,|\, (T,M,\theta,w) \,|_{s,p,\Omega_h}.$$

Here C_s is a constant independent of p, h, and d.

It is now easy to see that Theorems 2.3 and 2.4 follow from the three above lemmas.

3. Proofs. To prove the lemmas in the previous subsection, we rely, as expected, on the error equations, namely,

$$\begin{aligned}
&-(e_w, v_1')_{\Omega_h} + \langle \widehat{e_w}, [\![v_1\, n]\!] \rangle_{\mathscr{E}_h} = (e_\theta, v_1)_{\Omega_h} - d^2 (e_T/GA, v_1)_{\Omega_h}, && \text{(3.1a)} \\
&-(e_\theta, v_2')_{\Omega_h} + \langle \widehat{e_\theta}, [\![v_2\, n]\!] \rangle_{\mathscr{E}_h} = (e_M/EI, v_2)_{\Omega_h}, && \text{(3.1b)} \\
&-(e_M, v_3')_{\Omega_h} + \langle \widehat{e_M}, [\![v_3\, n]\!] \rangle_{\mathscr{E}_h} = (e_T, v_3)_{\Omega_h}, && \text{(3.1c)} \\
&-(e_T, v_4')_{\Omega_h} + \langle \widehat{e_T}, [\![v_4\, n]\!] \rangle_{\mathscr{E}_h} = 0 && \text{(3.1d)}
\end{aligned}$$

for any $v_i \in V_h^p, i = 1,2,3,4$. They are easily obtained by noting that the exact solution (T, M, θ, w) also satisfies the DG formulation (2.5).

We also rely on interpolation operators naturally associated with the numerical traces of the method. For any $u \in H^1(\Omega_h)$, the function $\pi^\pm u \in V_h^p$ is defined on the element I_j by

$$\begin{aligned}
&(u - \pi^\pm u, v)_{I_j} = 0 \qquad \forall v \in P^{p-1}(I_j) \quad \text{if } p > 0, && \text{(3.2a)} \\
&(\pi^- u)(x_j^-) = u(x_j^-), \qquad (\pi^+ u)(x_{j-1}^+) = u(x_{j-1}^+). && \text{(3.2b)}
\end{aligned}$$

Finally, we use the following notation:

$$\begin{aligned}
\boldsymbol{\pi e} &:= (\pi^+ e_T, \pi^+ e_M, \pi^- e_\theta, \pi^- e_w), && \text{(3.3a)} \\
\boldsymbol{\xi} &:= \boldsymbol{e} - \boldsymbol{\pi e} = (\xi_T^+, \xi_M^+, \xi_\theta^-, \xi_w^-). && \text{(3.3b)}
\end{aligned}$$

3.1. Proof of Lemma 2.5. To prove Lemma 2.5, we begin by obtaining the following representation formulas.

LEMMA 3.1. *Let $\psi \in L^2(\Omega_h)$ and define $\Psi(x) := \int_0^x \psi(s)ds$. Then the following expressions hold:*

$$(e_w,\psi)_{\Omega_h} = -\left((\xi_w^+)',\xi_\Psi^+\right)_{\Omega_h} + e_w(1^-)\xi_\Psi^+(1^-) - \left(e_\theta - d^2\frac{e_T}{GA},\pi^+\Psi\right)_{\Omega_h}, \tag{3.4a}$$

$$(e_\theta,\psi)_{\Omega_h} = -\left((\xi_\theta^+)',\xi_\Psi^+\right)_{\Omega_h} + e_\theta(1^-)\xi_\Psi^+(1^-) - \left(\frac{e_M}{EI},\pi^+\Psi\right)_{\Omega_h}, \tag{3.4b}$$

$$(e_T,\psi)_{\Omega_h} = -\left((\xi_T^-)',\xi_\Psi^-\right)_{\Omega_h} + \widehat{e_T}(1)\Psi(1). \tag{3.4c}$$

Proof. We only prove (3.4a) since the proofs of the other identities are similar. We begin by using the trivial identity

$$(e_w,\psi)_{\Omega_h} = (e_w,\Psi')_{\Omega_h} = (e_w,(\xi_\Psi^+)')_{\Omega_h} + (e_w,(\pi^+\Psi)')_{\Omega_h}.$$

Next, we obtain an expression for $(e_w,(\pi^+\Psi)')_{\Omega_h}$. Taking $v_1 = \pi^+\Psi$ in the first error equation (3.1), we get

$$\begin{aligned}(e_w,\psi)_{\Omega_h} =&(e_w,(\xi_\Psi^+)')_{\Omega_h} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h} - \left(e_\theta - d^2\,\frac{e_T}{GA},\pi^+\Psi\right)_{\Omega_h}\\ =&-\left((\xi_w^+)',\xi_\Psi^+\right)_{\Omega_h} + e_w(1^-)\xi_\Psi^+(1^-) - \left(e_\theta - d^2\frac{e_T}{GA},\pi^+\Psi\right)_{\Omega_h} + \Theta_h,\end{aligned}$$

where

$$\Theta_h := (e_w,(\xi_\Psi^+)')_{\Omega_h} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h} + \left((\xi_w^+)',\xi_\Psi^+\right)_{\Omega_h} - e_w(1^-)\xi_\Psi^+(1^-).$$

It remains to be proven that $\Theta_h = 0$. Integrating by parts the first term of the right-hand side, we get

$$\begin{aligned}\Theta_h =&-((e_w)',\xi_\Psi^+)_{\Omega_h} + \langle 1,[\![e_w\,(\xi_\Psi^+)\,n]\!]\rangle_{\mathscr{E}_h} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h}\\ &+((\xi_w^+)',\xi_\Psi^+)_{\Omega_h} - e_w(1^-)\xi_\Psi^+(1^-)\\ =&-((e_w-\xi_w^+)',\xi_\Psi^+)_{\Omega_h} - e_w(0^+)\xi_\Psi^+(0^+) + \langle 1,[\![e_w\,(\xi_\Psi^+)\,n]\!]\rangle_{\mathscr{E}_h^\circ} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ},\end{aligned}$$

where

$$\langle\eta,\zeta\rangle_{\mathscr{E}_h^\circ} := \sum_{j=1}^{N-1}\eta(x_j)\zeta(x_j).$$

Hence, by the definition of $\pi^\pm$, (3.2),

$$\begin{aligned}\Theta_h &= \langle 1,[\![e_w\,(\xi_\Psi^+)\,n]\!]\rangle_{\mathscr{E}_h^\circ} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ}\\ &= \langle 1,[\![e_w\,(\Psi-\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ} + \langle\widehat{e_w},[\![(\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ} && \text{by definition of } \xi_\Psi^+,\\ &= \langle 1,[\![e_w\,(\Psi-\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ} + \langle\widehat{e_w},[\![(\pi^+\Psi-\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ} && \text{by the continuity of } \Psi,\\ &= \langle 1,[\![(e_w-\widehat{e_w})\,(\Psi-\pi^+\Psi)\,n]\!]\rangle_{\mathscr{E}_h^\circ}\\ &= 0.\end{aligned}$$

Indeed, for any interior node x_j,

$$\begin{aligned}[\![(e_w-\widehat{e_w})\,(\Psi-\pi^+\Psi)\,n]\!](x_j) = &\,(e_w(x_j^-)-\widehat{e_w}(x_j))\,(\Psi(x_j)-\pi^+\Psi(x_j^-))\\ &-(e_w(x_j^+)-\widehat{e_w}(x_j))\,(\Psi(x_j)-\pi^+\Psi(x_j^+)) = 0,\end{aligned}$$

since $\widehat{e_w}(x_j) = e_w(x_j^-)$ and $\pi^+\Psi(x_j^+) = \Psi(x_j)$. This completes the proof. ∎

The proof of Lemma 2.5 now follows from Lemma 3.1 and the following proposition which contains the approximation results of the projection operators $\pi^{\pm}$; see [20] and, for example, [6] or [12].

LEMMA 3.2. *Let $I_j \subset \Omega_h$ be an arbitrary element, and suppose that $u \in H^{t+1}(I_j)$ for some nonnegative real number t. Then*

$$\begin{aligned}\left\|u-\pi^{\pm}u\right\|_{0,I_j}+\frac{h_j}{\mathsf{p}}\left\|(u-\pi^{\pm}u)'\right\|_{0,I_j} &\le C_t\frac{h_j^{\sigma+1}}{\mathsf{p}^{\sigma+1}}|u|_{\sigma+1,I_j},\\ |(u-\pi^{-}u)(x_{j-1}^{+})|+|(u-\pi^{+}u)(x_j^{-})| &\le C_t\frac{h_j^{\sigma+1/2}}{\mathsf{p}^{\sigma+1/2}}|u|_{\sigma+1,I_j},\end{aligned}$$

where $0 \le \sigma \le \min(p,t)$ for some constant C_t depending solely on t.

Proof of Lemma 2.5. Since all the estimates are obtained in a similar way, we prove only the second. To do this, we take $\psi = e_\theta$ in (3.4b) to obtain

$$\|e_\theta\|_{0,\Omega_h}^2 = -\left((\theta-\pi^{+}\theta)',\Psi-\pi^{+}\Psi\right)_{\Omega_h}+e_\theta(1^{-})(\Psi-\pi^{+}\Psi)(1^{-})-(e_M/EI,\pi^{+}\Psi)_{\Omega_h},$$

where $\Psi(x)=\int_0^x e_\theta(s)ds$. Then

$$\begin{aligned}\|e_\theta\|_{0,\Omega_h}^2 \le{}& \left\|(\theta-\pi^{+}\theta)'\right\|_{0,\Omega_h}\left\|\Psi-\pi^{+}\Psi\right\|_{0,\Omega_h}+|\,e_\theta(1^{-})\,|\,|\,(\Psi-\pi^{+}\Psi)(1^{-})\,|\\ &+\|e_M/EI\|_{0,\Omega_h}\left(\|\Psi\|_{0,\Omega_h}+\left\|\Psi-\pi^{+}\Psi\right\|_{0,\Omega_h}\right).\end{aligned}$$

By the approximation results in Lemma 3.2,

$$\begin{aligned}&\left\|(\theta-\pi^{+}\theta)'\right\|_{0,\Omega_h}\le C_t\frac{h^{\sigma}}{\mathsf{p}^{\sigma}}|\theta|_{\sigma+1,\Omega_h},\quad \left\|\Psi-\pi^{+}\Psi\right\|_{0,\Omega_h}\le C_t\frac{h}{\mathsf{p}}|\Psi|_{1,\Omega_h},\\ &|(\Psi-\pi^{+}\Psi)(1^{-})|\le C_t\frac{h^{1/2}}{\mathsf{p}^{1/2}}|\Psi|_{1,\Omega_h},\quad \|\Psi\|_{0,\Omega_h}\le C|\Psi|_{1,\Omega_h},\end{aligned}$$

where $0 \le \sigma \le \min(p,t)$, and by the definition of the seminorm $|\cdot|_{\mathscr{A}_h}$, (2.15),

$$|e_\theta(1^{-})|\le C_t\,\frac{h^{1/2}}{\mathsf{p}^{1/2}}\,|\boldsymbol{e}|_{\mathscr{A}_h},\qquad \|e_M/EI\|_{0,\Omega_h}\le C\,|\boldsymbol{e}|_{\mathscr{A}_h}.$$

Hence, we get

$$\|e_\theta\|_{0,\Omega_h}^2 \le C_t\left(|\boldsymbol{e}|_{\mathscr{A}_h}+\frac{h^{\sigma+1}}{\mathsf{p}^{\sigma+1}}|\theta|_{\sigma+1,\Omega_h}\right)|\Psi|_{1,\Omega_h},$$

and since $|\Psi|_{1,\Omega_h}=\|e_\theta\|_{0,\Omega_h}$,

$$\begin{aligned}\|e_\theta\|_{0,\Omega_h} &\le C_t\left(|\boldsymbol{e}|_{\mathscr{A}_h}+\frac{h^{\sigma+1}}{\mathsf{p}^{\sigma+1}}|\theta|_{\sigma+1,\Omega_h}\right)\\ &\le C_s\left(|\boldsymbol{e}|_{\mathscr{A}_h}+\frac{h^{\min(p,s+2)+1}}{\mathsf{p}^{\min(p,s+2)+1}}|\theta|_{\min(p,s+2)+1,\Omega_h}\right)\qquad \text{for } \sigma=\min(p,s+2),\\ &\le C_s\left(|\boldsymbol{e}|_{\mathscr{A}_h}+\frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}}\,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\right),\end{aligned}$$

and the estimate follows. This completes the proof of Lemma 2.5.

3.2. Proof of Lemma 2.6. To prove this lemma we proceed in two steps. In the first, we establish representation formulas for the errors in the numerical traces. In the second, we use approximation results to estimate them.

Step 1. *The error representation formulas.* We begin by expressing the numerical traces in terms of certain integrals involving the Green's functions.

LEMMA 3.3 (error representation formulas). *Let x_i be an arbitrary node and let φ^u_{w,x_i}, φ^u_{θ,x_i}, φ^u_{M,x_i}, φ^u_{T,x_i}, for $u = T, M, \theta$ or w, be the functions defined by* (2.16), (2.17), *and* (2.18). *Then*

$$\begin{aligned}\widehat{e_u}(x_i) = {} & ((\theta - \pi^-\theta)', \pi^+\varphi^u_{M,x_i} - \varphi^u_{M,x_i})_{\Omega_h} + e_\theta(1^-)\,(\pi^+\varphi^u_{M,x_i} - \varphi^u_{M,x_i})(1^-) \\ & + ((M - \pi^- M)', \pi^+\varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i})_{\Omega_h} + ((T - \pi^- T)', \pi^+\varphi^u_{w,x_i} - \varphi^u_{w,x_i})_{\Omega_h} \\ & + (e_M/EI, \pi^+\varphi^u_{M,x_i} - \varphi^u_{M,x_i})_{\Omega_h} + (e_T, \pi^-\varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i})_{\Omega_h}.\end{aligned}$$

To prove this lemma we need a couple of auxiliary results. The first establishes a relation between the errors in the numerical traces and the Green's functions.

LEMMA 3.4. *Set*

$$\Theta^u_i := \langle \widehat{e_w}, [\![\varphi^u_{T,x_i}\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_\theta}, [\![\varphi^u_{M,x_i}\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_M}, [\![\varphi^u_{\theta,x_i}\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_T}, [\![\varphi^u_{w,x_i}\, n]\!]\rangle_{\mathscr{E}_h}.$$

Then, we have

$$\begin{aligned}\Theta^u_i = {} & (e_w, (v_1 - \varphi^u_{T,x_i})')_{\Omega_h} + \langle \widehat{e_w}, [\![(\varphi^u_{T,x_i} - v_1)\, n]\!]\rangle_{\mathscr{E}_h} \\ & + (e_\theta, (v_2 - \varphi^u_{M,x_i})')_{\Omega_h} + \langle \widehat{e_\theta}, [\![(\varphi^u_{M,x_i} - v_2)\, n]\!]\rangle_{\mathscr{E}_h} \\ & + (e_M, (v_3 - \varphi^u_{\theta,x_i})')_{\Omega_h} + \langle \widehat{e_M}, [\![(\varphi^u_{\theta,x_i} - v_3)\, n]\!]\rangle_{\mathscr{E}_h} \\ & + (e_T, (v_4 - \varphi^u_{w,x_i})')_{\Omega_h}, + \langle \widehat{e_T}, [\![(\varphi^u_{w,x_i} - v_4)\, n]\!]\rangle_{\mathscr{E}_h} \\ & + (e_\theta, v_1 - \varphi^u_{T,x_i})_{\Omega_h} + (e_M/EI, v_2 - \varphi^u_{M,x_i})_{\Omega_h} \\ & + (e_T, v_3 - \varphi^u_{\theta,x_i})_{\Omega_h} - d^2(e_T/GA, v_1 - \varphi^u_{T,x_i})_{\Omega_h}\end{aligned}$$

for any $v_i \in V_h^p$, $i = 1, 2, 3, 4$.

Proof. Since we can write $\Theta^u_i = \Upsilon^u_i + \Delta^u_i$, where

$$\begin{aligned}\Upsilon^u_i := {} & \langle \widehat{e_w}, [\![v_1\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_\theta}, [\![v_2\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_M}, [\![v_3\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_T}, [\![v_4\, n]\!]\rangle_{\mathscr{E}_h}, \\ \Delta^u_i := {} & \langle \widehat{e_w}, [\![(\varphi^u_{T,x_i} - v_1)\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_\theta}, [\![(\varphi^u_{M,x_i} - v_2)\, n]\!]\rangle_{\mathscr{E}_h} \\ & + \langle \widehat{e_M}, [\![(\varphi^u_{\theta,x_i} - v_3)\, n]\!]\rangle_{\mathscr{E}_h} + \langle \widehat{e_T}, [\![(\varphi^u_{w,x_i} - v_4)\, n]\!]\rangle_{\mathscr{E}_h},\end{aligned}$$

we only have to find an expression for Υ^u_i.

To do that, we proceed as follows. First, note that by the definition of Green's functions, we have

$$\begin{aligned} -(e_w, (\varphi^u_{T,x_i})')_{\Omega_h} &= 0, \\ -(e_\theta, (\varphi^u_{M,x_i})')_{\Omega_h} &= (e_\theta, \varphi^u_{T,x_i})_{\Omega_h}, \\ -(e_M, (\varphi^u_{\theta,x_i})')_{\Omega_h} &= (e_M/IE, \varphi^u_{M,x_i})_{\Omega_h}, \\ -(e_T, (\varphi^u_{w,x_i})')_{\Omega_h} &= (e_T, \varphi^u_{\theta,x_i} - d^2\,\varphi^u_{T,x_i}/GA)_{\Omega_h}. \end{aligned}$$

We add these equations and then subtract the result from the addition of all the error equations (3.1). After rearranging terms, we obtain the expression for Υ^u_i we were

seeking, namely,

$$\begin{aligned}\Upsilon_i^u &= \langle \widehat{e_w}, [\![v_1\, n]\!] \rangle_{\mathscr{E}_h} + \langle \widehat{e_\theta}, [\![v_2\, n]\!] \rangle_{\mathscr{E}_h} + \langle \widehat{e_M}, [\![v_3\, n]\!] \rangle_{\mathscr{E}_h} + \langle \widehat{e_T}, [\![v_4\, n]\!] \rangle_{\mathscr{E}_h} \\ &= (e_w, (v_1 - \varphi^u_{T,x_i})')_{\Omega_h} + (e_\theta, (v_2 - \varphi^u_{M,x_i})')_{\Omega_h} \\ &\quad + (e_M, (v_3 - \varphi^u_{\theta,x_i})')_{\Omega_h} + (e_T, (v_4 - \varphi^u_{w,x_i})')_{\Omega_h} \\ &\quad + (e_\theta, v_1 - \varphi^u_{T,x_i})_{\Omega_h} + (e_M/EI, v_2 - \varphi^u_{M,x_i})_{\Omega_h} \\ &\quad + (e_T, v_3 - \varphi^u_{\theta,x_i})_{\Omega_h} - d^2 (e_{T,x_i}/GA, v_1 - \varphi^u_{T,x_i})_{\Omega_h}.\end{aligned}$$

The expression for Θ_i^u immediately follows. This completes the proof. ∎

The second result is needed to evaluate the expression for Θ_i^u.

LEMMA 3.5. *For any functions e, φ in $H^1(\Omega_h)$, set*

$$R = (e, (\pi^\pm \varphi - \varphi)')_{\Omega_h} + \langle \widehat{e}^{\,\mp}, [\![(\varphi - \pi^\pm \varphi)\, n]\!] \rangle_{\mathscr{E}_h},$$

where $\widehat{e}^{\,\mp}(x_j) = e(x_j^\mp)$ for $j = 1, \dots, N-1$. Then

$$R = ((e - \pi^\mp e)', \pi^\pm \varphi - \varphi)_{\Omega_h} - \langle (e - \widehat{e}^{\,\mp}), (\pi^\pm \varphi - \varphi)\, n \rangle_{\partial\Omega^\pm},$$

where

$$\begin{aligned}\langle (e - \widehat{e}^{\,-}), (\pi^+ \varphi - \varphi)\, n \rangle_{\partial\Omega^+} &= (e(1^-) - \widehat{e}^{\,-}(1))\, (\pi^+ \varphi - \varphi)(1^-), \\ \langle (e - \widehat{e}^{\,+}), (\pi^- \varphi - \varphi)\, n \rangle_{\partial\Omega^-} &= -(e(0^+) - \widehat{e}^{\,\mp}(0))\, (\pi^- \varphi - \varphi)(0^+).\end{aligned}$$

Proof. After a simple integration by parts, we get

$$\begin{aligned}R =& - (e', \pi^\pm \varphi - \varphi)_{\Omega_h} + \langle 1, [\![(e - \widehat{e}^{\,\mp})(\pi^\pm \varphi - \varphi)\, n]\!] \rangle_{\mathscr{E}_h} \\ =& ((e - \pi^\mp e)', \pi^\pm \varphi - \varphi)_{\Omega_h} - \langle 1, [\![(e - \widehat{e}^{\,\mp})(\pi^\pm \varphi - \varphi)\, n]\!] \rangle_{\mathscr{E}_h},\end{aligned}$$

by the orthogonality properties of the operators $\pi^\pm$, (3.2a).

Now, by the definition of the operators $\pi^\pm$ at the borders of the intervals, (3.2b), and by the definition of the numerical trace $\widehat{e}^{\,\mp}$, we have

$$-\langle 1, [\![(e - \widehat{e}^{\,\mp})(\pi^\pm \varphi - \varphi)\, n]\!] \rangle_{\mathscr{E}_h} = -\langle (e - \widehat{e}^{\,\mp}), (\pi^\pm \varphi - \varphi)\, n \rangle_{\partial\Omega^\pm},$$

and the result follows. This completes the proof. ∎

We are now ready to prove our representation result.

Proof of Lemma 3.3. We begin by noting that, by the definition of the Green's functions, (2.17) and (2.18), we have

$$\Theta_i^u = \widehat{e_u}(x_i).$$

On the other hand, setting $(v_1, v_2, v_3, v_4) = (\pi^+ \varphi^u_{T,x_i}, \pi^+ \varphi^u_{M,x_i}, \pi^- \varphi^u_{\theta,x_i}, \pi^- \varphi^u_{w,x_i})$ in Lemma 3.4 ,we get

$$\begin{aligned}\widehat{e_u}(x_i) =& (e_w, (\pi^+ \varphi^u_{T,x_i} - \varphi^u_{T,x_i})')_{\Omega_h} + \langle \widehat{e_w}, [\![(\varphi^u_{T,x_i} - \pi^+ \varphi^u_{T,x_i})\, n]\!] \rangle_{\mathscr{E}_h} \\ &+ (e_\theta, (\pi^+ \varphi^u_{M,x_i} - \varphi^u_{M,x_i})')_{\Omega_h} + \langle \widehat{e_\theta}, [\![(\varphi^u_{M,x_i} - \pi^+ \varphi^u_{M,x_i})\, n]\!] \rangle_{\mathscr{E}_h} \\ &+ (e_M, (\pi^- \varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i})')_{\Omega_h} + \langle \widehat{e_M}, [\![(\varphi^u_{\theta,x_i} - \pi^- \varphi^u_{\theta,x_i})\, n]\!] \rangle_{\mathscr{E}_h} \\ &+ (e_T, (\pi^- \varphi^u_{w,x_i} - \varphi^u_{w,x_i})')_{\Omega_h}, + \langle \widehat{e_T}, [\![(\varphi^u_{w,x_i} - \pi^- \varphi^u_{w,x_i})\, n]\!] \rangle_{\mathscr{E}_h} \\ &+ (e_\theta, \pi^+ \varphi^u_{T,x_i} - \varphi^u_{T,x_i})_{\Omega_h} + (e_M/EI, \pi^+ \varphi^u_{M,x_i} - \varphi^u_{M,x_i})_{\Omega_h} \\ &+ (e_T, \pi^- \varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i})_{\Omega_h} - d^2 (e_T/GA, \pi^+ \varphi^u_{T,x_i} - \varphi^u_{T,x_i})_{\Omega_h}.\end{aligned}$$

Now, by Lemma 3.5,

$$\begin{aligned}\widehat{e_u}(x_i) = &((e_w-\pi^- e_w)', \pi^+\varphi^u_{T,x_i}-\varphi^u_{T,x_i})_{\Omega_h} - \langle (e_w-\widehat{e_w}), (\pi^+\varphi^u_{T,x_i}-\varphi^u_{T,x_i})\, n\rangle_{\partial\Omega^+}\\ &((e_\theta-\pi^- e_\theta)', \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} - \langle (e_\theta-\widehat{e_\theta}), (\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})\, n\rangle_{\partial\Omega^+}\\ &((e_M-\pi^+ e_M)', \pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h} - \langle (e_M-\widehat{e_M}), (\pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})\, n\rangle_{\partial\Omega^-}\\ &((e_T-\pi^+ e_T)', \pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})_{\Omega_h} - \langle (e_T-\widehat{e_T}), (\pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})\, n\rangle_{\partial\Omega^-}\\ &+ (e_\theta, \pi^+\varphi^u_{T,x_i}-\varphi^u_{T,x_i})_{\Omega_h} + (e_M/EI, \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h}\\ &+ (e_T, \pi^-\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h} - d^2(e_T/GA, \pi^+\varphi^u_{T,x_i}-\varphi^u_{T,x_i})_{\Omega_h}.\end{aligned}$$

Since the operators $\pi^\pm$ reproduce polynomials of degree p (see (3.2)) and since $\pi^+\varphi^u_{T,x_i}=\varphi^u_{T,x_i}$ (notice that φ^u_{T,x_i} is a constant function), we obtain

$$\begin{aligned}\widehat{e_u}(x_i) = &((\theta-\pi^-\theta)', \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} - \langle (e_\theta-\widehat{e_\theta}), (\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})\, n\rangle_{\partial\Omega^+}\\ &((M-\pi^+ M)', \pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h} - \langle (e_M-\widehat{e_M}), (\pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})\, n\rangle_{\partial\Omega^-}\\ &((T-\pi^+ T)', \pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})_{\Omega_h} - \langle (e_T-\widehat{e_T}), (\pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})\, n\rangle_{\partial\Omega^-}\\ &+ (e_M/EI, \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} + (e_T, \pi^-\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h}.\end{aligned}$$

Finally, by the definition of the numerical traces at $\partial\Omega$, (2.13) and (2.14), we have that

$$\begin{aligned}\langle (e_\theta-\widehat{e_\theta}), (\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})\, n\rangle_{\partial\Omega^+} &= e_\theta(1^-)\,(\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})(1^-),\\ \langle (e_M-\widehat{e_M}), (\pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})\, n\rangle_{\partial\Omega^-} &= 0,\\ \langle (e_T-\widehat{e_T}), (\pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})\, n\rangle_{\partial\Omega^-} &= 0,\end{aligned}$$

and hence,

$$\begin{aligned}\widehat{e_u}(x_i) = &((\theta-\pi^-\theta)', \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} + e_\theta(1^-)\,(\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})(1^-)\\ &+ ((M-\pi^- M)', \pi^+\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h} + ((T-\pi^- T)', \pi^+\varphi^u_{w,x_i}-\varphi^u_{w,x_i})_{\Omega_h}\\ &+ (e_M/EI, \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} + (e_T, \pi^-\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h}.\end{aligned}$$

This completes the proof. □

Step 2. *Estimating the errors in the numerical traces.* Here, we apply the approximation results of Lemma 3.2 to the representation formulas of Lemma 3.3 to prove Lemma 2.6.

Proof of Lemma 2.6. From the representation formulas in Lemma 3.3, we get

$$|\widehat{e_u}(x_i)| \le H_1 + H_2 + H_3,$$

where

$$\begin{aligned}H_1 &= |((\pi^-\theta-\theta)',\ \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h}| + |((\pi^+ M-M)',\ \pi^-\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h}|\\ &\quad + |((\pi^+ T-T)',\ \pi^-\varphi^u_{w,x_i}-\varphi^u_{w,x_i})_{\Omega_h}|,\\ H_2 &= |(e_M/EI,\ \pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})_{\Omega_h} + |e_\theta(1^-)(\pi^+\varphi^u_{M,x_i}-\varphi^u_{M,x_i})(1^-)|,\\ H_3 &= |(e_T,\ \pi^-\varphi^u_{\theta,x_i}-\varphi^u_{\theta,x_i})_{\Omega_h}|.\end{aligned}$$

Let us estimate H_1. By using the approximation inequalities of Lemma 3.2, we immediately obtain that

$$H_1 \le C_s\,\frac{h^{\min(s,p)}}{\mathsf{p}^{\min(s,p)}}\,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\,\frac{h^{p+1}}{\mathsf{p}^{p+1}}\,|\varphi^u_{x_i}|_{p+1,\Omega_h}.$$

To estimate H_2, we use the fact that, from the definition of the seminorm $|\cdot|_{\mathscr{A}_h}$, (2.15),

$$\|e_M/EI\|_{0,\Omega_h} \le C\,|\boldsymbol{e}|_{\mathscr{A}_h}\,, \qquad |e_\theta(1^-)| \le C(h/\mathsf{p})^{1/2}\,|\boldsymbol{e}|_{\mathscr{A}_h}$$

to get

$$H_2 \le C_s\ |\boldsymbol{e}|_{\mathscr{A}_h}\ \frac{h^{p+1}}{\mathsf{p}^{p+1}}\,|\varphi^u_{x_i}|_{p+1,\Omega_h}.$$

Here, we have used once again the approximation results of Lemma 3.2.

The estimate of H_3 term is more delicate. To estimate it, we proceed as follows. We begin by noting that, by (3.4c) of Lemma 3.1, we have that

$$(e_T,\psi)_{\Omega_h} = -\left((\xi_T^-)',\xi_\Psi^-\right)_{\Omega_h} + \widehat{e_T}(1)\Psi(1)$$

for any $\psi \in L^2(\Omega_h)$. Taking $\psi = \pi^-\varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i}$ and using the approximation results in Lemma 3.2 we obtain

$$H_3 \le\ C_s\ \frac{h^\sigma}{\mathsf{p}^\sigma}|T|_{\sigma+1,\Omega_h}\frac{h}{\mathsf{p}}|\Psi|_{1,\Omega_h} + |\widehat{e_T}(1)|\left\|\pi^-\varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i}\right\|_{0,\Omega_h},$$

and since

$$|\Psi|_{1,\Omega_h} = |\Psi'|_{0,\Omega_h} = \left\|\pi^-\varphi^u_{\theta,x_i} - \varphi^u_{\theta,x_i}\right\|_{0,\Omega_h} \le C\frac{h^{p+1}}{\mathsf{p}^{p+1}}|\varphi^u_{\theta,x_i}|_{p+1,\Omega_h}$$

for some constant C independent of p, h, and d, we get

$$H_3 \le C_s\left(\frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}}|T|_{\min(p,s)+1,\Omega_h} + |\widehat{e_T}(1)|\right)\ \frac{h^{p+1}}{\mathsf{p}^{p+1}}\,|\varphi^u_{\theta,x_i}|_{p+1,\Omega_h}$$

with $\sigma = \min(p,s)$ in Lemma 3.2.

Thus, we get

$$|\,\widehat{e_u}(x_i)\,| \le C_s\left(|\boldsymbol{e}|_{\mathscr{A}_h} + |\widehat{e_T}(1)| + \frac{h^{\min(s,p)}}{\mathsf{p}^{\min(s,p)}}\,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\right)\ \frac{h^{p+1}}{\mathsf{p}^{p+1}}\,|\varphi^u_{x_i}|_{p+1,\Omega_h}.$$

We clearly see that the estimates for $u = M,\theta,w$ follow from that of $u = T$. Such an estimate immediately follows if we assume that

$$C_s\frac{h^{p+1}}{\mathsf{p}^{p+1}}\ \sup_{y\in(0,1)}\ |\varphi^T_y|_{p+1,\Omega_h} \le \kappa < 1,$$

that is, if h is sufficiently small or p sufficiently big. This completes the proof. ∎

3.3. Proof of Lemma 2.7. Lemma 2.7 follows immediately from the following auxiliary results.

LEMMA 3.6. *We have* $\boldsymbol{e} = \boldsymbol{\xi} + \boldsymbol{\pi e}$ *and* $|\boldsymbol{\pi e}|^2_{\mathscr{A}_h} = J_1 + J_2 + J_3$, *where*

$$\begin{aligned}
J_1 &= -\left(\xi^+_M/EI,\,\pi^+e_M\right)_{\Omega_h} - d^2\left(\xi^+_T/GA,\,\pi^+e_T\right)_{\Omega_h},\\
J_2 &= \xi^+_M(1^-)(\pi^-e_\theta)(1^-) - \xi^+_T(1^-)(\pi^-e_w)(1^-),\\
J_3 &= \left(\xi^-_\theta,\,\pi^+e_T\right)_{\Omega_h} - \left(\xi^+_T,\,\pi^-e_\theta\right)_{\Omega_h}.
\end{aligned}$$

LEMMA 3.7. *We have*

$$
\begin{aligned}
|\boldsymbol{\xi}|_{\mathscr{A}_h} &\le C_s \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h},\\
|\,J_1\,| &\le C_s \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\ |\boldsymbol{\pi e}|_{\mathscr{A}_h},\\
|\,J_2\,| &\le C_s \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\ |\boldsymbol{\pi e}|_{\mathscr{A}_h},\\
|\,J_3\,| &\le C_s \frac{h^{2\min(p,s)+3}}{\mathsf{p}^{2\min(p,s)+3}} \,|\,(T,M,\theta,w)\,|^2_{s,p,\Omega_h}\\
&\quad + C_s\ |\boldsymbol{\pi e}|_{\mathscr{A}_h}\ \frac{h^{\min(p,s)+2}}{\mathsf{p}^{\min(p,s)+3/2}} \,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}.
\end{aligned}
$$

We prove these results in several steps.

Step 1. *Preliminaries.* We begin by rewriting the method in its classical mixed formulation. Thus, if we add (2.5a) to (2.5d) and subtract (2.5b) and (2.5c) from the resulting expression, we obtain, after rearranging terms,

$$
\begin{aligned}
&(M_h/EI, v_2)_{\Omega_h} + d^2\,(T_h/GA, v_1)_{\Omega_h}\\
&\quad - (w_h, v_1')_{\Omega_h} + \langle \widehat{w}_h, [\![v_1\, n]\!]\rangle_{\mathscr{E}_h} - (T_h, v_4')_{\Omega_h} + \langle \widehat{T}_h, [\![v_4\, n]\!]\rangle_{\mathscr{E}_h}\\
&\quad + (\theta_h, v_2')_{\Omega_h} - \langle \widehat{\theta}_h, [\![v_2\, n]\!]\rangle_{\mathscr{E}_h} + (M_h, v_3')_{\Omega_h} - \langle \widehat{M}_h, [\![v_3\, n]\!]\rangle_{\mathscr{E}_h}\\
&\quad - (\theta_h, v_1)_{\Omega_h} + (T_h, v_3)_{\Omega_h}\\
&= (q, v_4)_{\Omega_h}.
\end{aligned}
$$

Inserting the definition of the numerical traces, (2.12), (2.13), and (2.14), and moving to the right-hand side all the terms containing boundary data, we obtain

$$
\mathscr{A}_h(T_h, M_h, \theta_h, w_h;\, v_1, v_2, v_3, v_4) = b_h(v_1, v_2, v_3, v_4) \quad \forall v_i \in V_h^p, \quad i = 1, 2, 3, 4,
$$

where, writing $\mathscr{A}_h$ for $\mathscr{A}_h(u_1, u_2, u_3, u_4;\, v_1, v_2, v_3, v_4)$,

$$
\begin{aligned}
\mathscr{A}_h &:= (u_2/EI, v_2)_{\Omega_h} + d^2 (u_1/GA, v_1)_{\Omega_h}\\
&\quad + \mathsf{c}\frac{\mathsf{p}}{h_N}\,\big[u_4(1^-)\,v_4(1^-) + u_3(1^-)\,v_3(1^-)\big]\\
&\quad + \mathscr{S}_h(u_1, u_2, u_3, u_4;\, v_1, v_2, v_3, v_4),
\end{aligned}
\tag{3.5}
$$

and, writing $\mathscr{S}_h$ for $\mathscr{S}_h(u_1, u_2, u_3, u_4;\, v_1, v_2, v_3, v_4)$,

$$
\begin{aligned}
\mathscr{S}_h &:= -(u_4, v_1')_{\Omega_h} - (u_1, v_4')_{\Omega_h} + \sum_{j=1}^{N-1} \Big(u_4^-\,[\![v_1\, n]\!] + u_1^+\,[\![v_4\, n]\!]\Big)(x_j)\\
&\quad - u_1(0^+)\,v_4(0^+) + u_1(1^-)\,v_4(1^-)\\
&\quad + (u_3, v_2')_{\Omega_h} + (u_2, v_3')_{\Omega_h} - \sum_{j=1}^{N-1} \Big(u_3^-\,[\![v_2\, n]\!] + u_2^+\,[\![v_3\, n]\!]\Big)(x_j)\\
&\quad - u_2(0^+)\,v_3(0^+) + u_2(1^-)\,v_3(1^-)\\
&\quad - (u_3, v_1)_{\Omega_h} + (u_1, v_3)_{\Omega_h}.
\end{aligned}
$$

Finally,

$$b_h(v_1,v_2,v_3,v_4) := (q,v_4)_{\Omega_h} + \mathsf{c}\frac{\mathsf{p}}{h_N}\left[\, w_1\, v_4(1^-) - \theta_1\, v_3(1^-)\right] + w_0\, v_1(0^+) - w_1\, v_1(1^-) - \theta_0\, v_2(0^+) + \theta_1\, v_2(1^-).$$

The first property we are going to use is the so-called Galerkin orthogonality, namely,

$$\mathscr{A}_h(e_T, e_M, e_\theta, e_w;\, v_1, v_2, v_3, v_4) = 0 \quad \forall v_i \in V_h^p, \tag{3.6}$$

and it follows directly from the error equations (3.1). The second property is that

$$|\boldsymbol{v}|_{\mathscr{A}_h} = \mathscr{A}_h(\boldsymbol{v},\boldsymbol{v})^{1/2} \qquad \text{for any } \boldsymbol{v} \in [V_h^p]^4. \tag{3.7}$$

Indeed, a simple calculation shows that $\mathscr{S}_h(v_1,v_2,v_3,v_4;\, v_1,v_2,v_3,v_4) \equiv 0$.

Step 2. *Proof of the auxiliary Lemma* 3.6. We have

$$\begin{aligned}
|\boldsymbol{\pi e}|^2_{\mathscr{A}_h} &= \mathscr{A}_h(\boldsymbol{\pi e};\, \boldsymbol{\pi e}) && \text{by (3.7)},\\
&= \mathscr{A}_h(\boldsymbol{e}-\boldsymbol{\xi};\, \boldsymbol{\pi e}) && \text{by (3.3b)},\\
&= -\mathscr{A}_h(\boldsymbol{\xi};\, \boldsymbol{\pi e}) && \text{by (3.6)},\\
&= -\left(\xi_M^+/EI,\, \pi^+ e_M\right)_{\Omega_h} - d^2\left(\xi_T^+/GA,\, \pi^+ e_T\right)_{\Omega_h}\\
&\quad + \xi_M^+(1^-)(\pi^- e_\theta)(1^-) - \xi_T^+(1^-)(\pi^- e_w)(1^-)\\
&\quad + \left(\xi_\theta^-,\, \pi^+ e_T\right)_{\Omega_h} - \left(\xi_T^+,\, \pi^- e_\theta\right)_{\Omega_h} && \text{by (3.2)},\\
&= J_1 + J_2 + J_3,
\end{aligned}$$

as claimed. This completes the proof of Lemma 3.6.

Step 3. *Estimate of* $|\boldsymbol{\xi}|_{\mathscr{A}_h}$. We have

$$\begin{aligned}
|\boldsymbol{\xi}|^2_{\mathscr{A}_h} &= \left(\xi_M^+/EI, \xi_M^+\right)_{\Omega_h} + d^2\left(\xi_T^+/GA, \xi_T^+\right)_{\Omega_h} + \mathsf{c}\frac{\mathsf{p}}{h_N}\left(\llbracket \xi_\theta^-\, n\rrbracket^2(1) + \llbracket \xi_w^-\, n\rrbracket^2(1)\right)\\
&= \left(\xi_M^+/EI, \xi_M^+\right)_{\Omega_h} + d^2\left(\xi_T^+/GA, \xi_T^+\right)_{\Omega_h},
\end{aligned}$$

by definition of π^-, (3.2). Hence,

$$\begin{aligned}
|\boldsymbol{\xi}|_{\mathscr{A}_h} &\le C_s\left(\left\|\xi_M^+\right\|_{0,\Omega_h} + \left\|\xi_T^+\right\|_{0,\Omega_h}\right)\\
&\le C_s \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}}\, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h}
\end{aligned}$$

for some constant C_s depending solely on s, by the approximation results of Lemma 3.2 with $\sigma = \min(p,s)$ for T and $\sigma = \min(p,s+1)$ for M. This completes the proof of the estimate.

Step 4. *Estimate of* J_1. To estimate J_1, we proceed as follows:

$$\begin{aligned}
|\,J_1\,| &\le |\left(\xi_M^+/EI,\, \pi^+ e_M\right)_{\Omega_h}| + d^2|\left(\xi_T^+/GA,\, \pi^+ e_T\right)_{\Omega_h}|\\
&\le C\left\|\xi_M^+\right\|_{0,\Omega_h} |\boldsymbol{\pi e}|_{\mathscr{A}_h} + Cd\left\|\xi_T^+\right\|_{0,\Omega_h} |\boldsymbol{\pi e}|_{\mathscr{A}_h}\\
&\le C_s\, \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}}\, |\,(T,M,\theta,w)\,|_{s,p,\Omega_h}\, |\boldsymbol{\pi e}|_{\mathscr{A}_h},
\end{aligned}$$

by the approximation results of Lemma 3.2 with $\sigma = \min(p,s)$ for T and $\sigma = \min(p,s+1)$ for M.

Step 5. *Estimate of* J_2. The estimate of J_2 is obtained in a similar way. We have

$$\begin{aligned}
|\, J_2 \,| &\le |\, \xi_M^+(1^-)(\pi^- e_\theta)(1^-) \,| + |\, \xi_T^+(1^-)(\pi^- e_w)(1^-) \,| \\
&\le \left(|\, \xi_M^+(1^-) \,|^2 + |\, \xi_T^+(1^-) \,|^2\right)^{1/2} \left(|\, (\pi^- e_\theta)(1^-) \,|^2 + |\, (\pi^- e_w)(1^-) \,|^2\right)^{1/2} \\
&= \frac{h_N^{1/2}}{\mathsf{p}^{1/2}} \left(|\, \xi_M^+(1^-) \,|^2 + |\, \xi_T^+(1^-) \,|^2\right)^{1/2} |\boldsymbol{\pi e}|_{\mathscr{A}_h} \quad \text{by definition of } |\cdot|_{\mathscr{A}_h},\ (2.15), \\
&\le C_s \frac{h^{\min(p,s)+1}}{\mathsf{p}^{\min(p,s)+1}} \,|\, (T, M, \theta, w) \,|_{s,p,\Omega_h} \,|\boldsymbol{\pi e}|_{\mathscr{A}_h} \,,
\end{aligned}$$

by the approximation results of Lemma 3.2 with $\sigma = \min(p, s)$ for T and $\sigma = \min(p, s+1)$ for M.

Step 6. *Estimate of* J_3. The estimate of this term requires a very delicate analysis captured in the following auxiliary result.

LEMMA 3.8. *Assume that in the interval* $I = (a, b)$, *we have, for all* $v \in P^p(I)$,

$$\begin{aligned}
&\text{(i)} && -(e, v')_I + \widehat{e}(b)\, v(b) - \widehat{e}(a)\, v(a) = (f, v)_I, \\
&\text{(ii)} && (\xi, v')_I = 0, \\
&\text{(iii)} && \pi e \in P^p(I).
\end{aligned}$$

Then,

$$|\, (\xi, \pi e)_I \,| \le C\, \frac{h^{1/2}}{p^{1/2}} \left(\frac{h}{p} \,\|\, f \,\|_{L^\infty(I)} + \min_{c \in \{a,b\}} |\, \widehat{e}(c) - \pi e(c) \,| \right) \,\|\, \xi \,\|_{0,I},$$

where $h = b - a$ *and* C *is independent of* h, p, *and* e.

Proof. By (iii), we can write $\pi e = \mathbb{P}_{p-1} e + \alpha \ell_p$, where $\mathbb{P}_{p-1}$ is the L^2-projection into $P^{p-1}(I)$, and $\ell_n(x) = P_n\left(\frac{x-(a+b)/2}{h/2}\right)$, where P_n is the Legendre polynomial of degree n. By (ii), this implies that

$$(\xi, \pi e)_I = (\xi, \mathbb{P}_{p-1} e + \alpha \ell_p)_I = \alpha\, (\xi, \ell_p)_I,$$

and, since $\|\, \ell_p \,\|_{0,I}^2 = h/(2\,p+1)$, we get

$$|\, (\xi, \pi e)_I \,| \le |\, \alpha \,| \left(\frac{h}{2\,p+1} \right)^{1/2} \|\, \xi \,\|_{0,I} \le |\, \alpha \,| \, \frac{h^{1/2}}{p^{1/2}} \,\|\, \xi \,\|_{0,I}.$$

It remains to estimate $|\, \alpha \,|$.

Since $|\, \ell_p(c) \,| = 1$, we immediately get that

$$\alpha = \ell_p(c)(-\mathbb{P}_{p-1} e(c) + \pi e(c)).$$

This implies that to estimate α, we need to obtain an expression for $\mathbb{P}_{p-1} e(c)$. We claim that, for $c \in \{a, b\}$,

$$\mathbb{P}_{p-1} e(x) = \widehat{e}(c) - \mathbb{P}_{p-1} g_c(x) \qquad \text{for } x \in I,$$

where $g_c(x) = \int_x^c f(s)\, ds$. Assuming that this is true, we obtain

$$\begin{aligned}
|\, \alpha \,| &\le |\, \mathbb{P}_{p-1} g_c(c) \,| + |\, \pi e(c) - \widehat{e}(c) \,| \\
&\le |\, \mathbb{P}_{p-1} g_c(c) - g_c(c) \,| + |\, \pi e(c) - \widehat{e}(c) \,| \\
&\le C\, \frac{h}{p} \,\|\, f \,\|_{L^\infty(I)} + |\, \pi e(c) - \widehat{e}(c) \,|,
\end{aligned}$$

and the result follows.

It remains to prove the claim. Let us begin with the case $c = a$. Since, by (i),

$$-(e, v')_I + (\widehat{e}(b) - \widehat{e}(a))\, v(b) + \widehat{e}(a)\,(v(b) - v(a)) = (f, v)_I,$$

we get

$$(-e + \widehat{e}(a), v')_I = (f, v)_I - (\widehat{e}(b) - \widehat{e}(a))\, v(b).$$

Setting $v = 1$ in (i), we obtain that

$$\widehat{e}(b) - \widehat{e}(a) = (f, 1)_I,$$

and so

$$(-e + \widehat{e}(a), v')_I = (f, v - v(b))_I = (g_a, v')_I,$$

and the claim follows. The case $c = b$ is proven in a similar way.

This completes the proof. □

We are now ready to estimate $J_3 = (\xi_\theta^-,\, \pi^+ e_T)_{\Omega_h} - (\xi_T^+,\, \pi^- e_\theta)_{\Omega_h}$. Let us begin by noting that

$$(\xi_\theta^-,\, \pi^+ e_T)_{\Omega_h} = 0.$$

To see this, notice that the previous lemma is satisfied with $e = e_T$, $\xi = \xi_\theta^-$, $\pi = \pi^+$, and $f = 0$, thanks to the error equation (3.1d), and that, by the definition of the numerical trace $\widehat{e_T}$ we have that $\widehat{e_T}(x_j) = e_T(x_j^+) = \pi^+ e_T(x_j)$ for $j = 0, \ldots, N-1$.

This implies that

$$|\, J_3 \,| \le \sum_{j=1}^{N} |\, (\xi_T^+,\, \pi^- e_\theta)_{I_j} \,|.$$

To estimate each of the terms of the right-hand side, we use Lemma 3.8 once more. By the error equation (3.1b), the result holds with $e = e_\theta$, $\xi = \xi_T^+$, $\pi = \pi^-$ and $f = e_M/EI$. Hence,

$$|\, (\xi_T^+,\, \pi^- e_\theta)_{I_j} \,| \le C\, \frac{h_j^{1/2}}{p^{1/2}} \left(\frac{h_j}{p} \| \, e_M/EI \, \|_{L^\infty(I_j)} + \Xi_j \right) \| \, \xi_T^+ \, \|_{0,I_j},$$

where $\Xi_j := \min_{c \in \{x_{j-1}, x_j\}} |\, \widehat{e}_\theta(c) - \pi^- e_\theta(c) \,|$. Since $\Xi_j = 0$ for $j = 1, \ldots, N-1$, and $\Xi_N = e_\theta(1^-) = \pi^- e_\theta(1^-)$, and since

$$\begin{aligned} \| \, e_M/EI \, \|_{L^\infty(I_j)} &\le \| \, (M - \pi^+ M)/EI \, \|_{L^\infty(I_j)} + \| \, \pi^+ e_M/EI \, \|_{L^\infty(I_j)} \\ &\le C_s\, \frac{h_j^{\min(p,s)+1/2}}{p^{\min(p,s)+1/2}}\, |\, M \,|_{\min(p,s+1)+1, I_j} \\ &\quad + C_s\, \frac{p}{h_j^{1/2}} \| \, \pi^+ e_M \, \|_{0,I_j}, \end{aligned}$$

by the approximation result of Lemma 3.2 with $\sigma = \min(p, s+1)$, we get

$$\begin{aligned} |\, J_3 \,| &\le C_s\, \frac{h^{\min(p,s)+2}}{p^{\min(p,s)+2}}\, |\, M \,|_{\min(p,s+1)+1, \Omega_h} \| \, \xi_T^+ \, \|_{0,\Omega_h} \\ &\quad + C_s \left(\frac{h}{p^{1/2}} \| \, \pi^+ e_M/EI \, \|_{0,\Omega_h} + \frac{h^{1/2}}{p^{1/2}}\, |\, \pi^- e_\theta(1^-) \,| \right) \| \, \xi_T^+ \, \|_{0,\Omega_h} \\ &\le C_s \left(\frac{h^{\min(p,s)+2}}{p^{\min(p,s)+2}}\, |\, M \,|_{\min(p,s+1)+1, \Omega_h} + \frac{h}{p^{1/2}}\, |\boldsymbol{\pi e}|_{\mathscr{A}_h} \right) \| \, \xi_T^+ \, \|_{0,\Omega_h} \end{aligned}$$

by the definition of the energy seminorm $|\cdot|_{\mathscr{A}_h}$, (2.15). Finally, using the approximation results of Lemma 3.2, we get

$$\begin{aligned}|\,J_3\,| \le & C_s\,\frac{h^{2\min(p,s)+3}}{p^{2\min(p,s)+3}}\,|\,(T,M,\theta,w)\,|^2_{s,p,\Omega_h}\\ &+ C_s\;|\boldsymbol{\pi e}|_{\mathscr{A}_h}\,\frac{h^{\min(p,s)+2}}{p^{\min(p,s)+3/2}}\,|\,(T,M,\theta,w)\,|_{s,p,\Omega_h}.\end{aligned}$$

This completes the proof of Lemma 3.6 and hence that of our main results.

4. Extensions.

4.1. Uniform convergence for piecewise-constant approximations. In the case of piecewise-constant approximation, $p = 0$, we obtain the following result.

THEOREM 4.1. *Assume that, for some $s \ge 0$, (T, M, θ, w) belongs to $H^{s+1}(\Omega_h) \times H^{s+2}(\Omega_h) \times H^{s+3}(\Omega_h) \times H^{s+2}(\Omega_h)$. Assume also that $q \in L^\infty(\Omega_h)$. Set*

$$\boldsymbol{e} := (e_T, e_M, e_\theta, e_w) = (T - T_h, M - M_h, \theta - \theta_h, w - w_h),$$

where $(T_h, M_h, \theta_h, w_h)$ is the approximation given by the DG method (2.5), (2.12), (2.13), *and* (2.14) *with $p = 0$. Then, for small enough h, we have that, for $u = T, M, \theta$, or w,*

$$\|\,e_u\,\|_{L^\infty(\Omega_h)} + \frac{\mathsf{c}}{h}\left(|\,e_\theta(1^-)\,| + |\,e_w(1^-)\,|\right) \le C_s\,h\,\left(\|\,q\,\|_{L^\infty(\Omega_h)} + |\,(T,M,\theta,w)\,|_{s,0,\Omega_h}\right)$$

for some constant C_s independent of h and d.

This result implies that the DG method in its specific form, which we discussed here, does not suffer from shear locking, even if it uses piecewise-constant approximation for all the unknowns. It also implies the unexpected superconvergence of the approximations to w and θ superconverge at the border $x = 1$.

Proof. To prove this result, we only have to slightly modify the proof of our main results. Indeed, we only have to modify the estimate of the terms J_3 in Lemma 3.7. We proceed in several steps.

Step 1. *Estimate of J_3.* We estimate such a term as follows:

$$\begin{aligned}J_3 &= (\xi_\theta^-,\,\pi^+ e_T)_{\Omega_h} - (\xi_T^+,\,\pi^- e_\theta)_{\Omega_h}\\ &= \sum_{j=1}^{N}\left((\xi_\theta^-,\,e_T(x_{j-1}^+))_{I_j} - (\xi_T^+,\,e_\theta(x_j^-))_{I_j}\right) \qquad \text{by definition of } \pi^\pm,\ (3.2),\\ &= \sum_{j=1}^{N}(\xi_\theta^-,1)_{I_j}\,\widehat{e_T}(x_j) - \sum_{j=1}^{N-1}(\xi_T^+,1)_{I_j}\,\widehat{e_\theta}(x_j) - (\xi_T^+,1)_{I_N}\,e_\theta(1^-),\end{aligned}$$

by the definition of the numerical traces $\widehat{T}_h$ and $\widehat{\theta}_h$, (2.12)–(2.14). Hence

$$\begin{aligned}|\,J_3\,| &\le C_s\left(\|\widehat{e_T}\|_{L^\infty(\mathscr{E}_h)} + \|\widehat{e_\theta}\|_{L^\infty(\mathscr{E}_h)} + |\pi\boldsymbol{e}|_{\mathscr{A}_h}\right)\,h\,(|\,\theta\,|_{1,\Omega_h} + |\,T\,|_{1,\Omega_h})\\ &\le C_s\left(\|\widehat{e_T}\|_{L^\infty(\mathscr{E}_h)} + \|\widehat{e_\theta}\|_{L^\infty(\mathscr{E}_h)} + |\pi\boldsymbol{e}|_{\mathscr{A}_h}\right)\,h\,|\,(T,M,\theta,w)\,|_{s,0,\Omega_h},\end{aligned}$$

by the definition of $|\cdot|_{\mathscr{A}_h}$, (2.15), and the approximation results of Lemma 3.2.

Step 2. *Estimate of* $|\,\widehat{e_u}(x_i)\,|$. Combining the above estimate with the results of Lemmas 3.6 and 3.7, we easily get

$$|\boldsymbol{e}|_{\mathscr{A}_h} \leq C_s\, h \left(\|\widehat{e_T}\|_{L^\infty(\mathscr{E}_h)} + \|\widehat{e_\theta}\|_{L^\infty(\mathscr{E}_h)} + |\,(T,M,\theta,w)\,|_{s,0,\Omega_h} \right).$$

Inserting this estimate into the estimate of Lemma 2.6, we get

$$|\,\widehat{e_u}(x_i)\,| \leq C_s\, h \left(\|\widehat{e_T}\|_{L^\infty(\mathscr{E}_h)} + \|\widehat{e_\theta}\|_{L^\infty(\mathscr{E}_h)} + |\,(T,M,\theta,w)\,|_{s,0,\Omega_h} \right),$$

which implies that, for small enough h,

$$|\,\widehat{e_u}(x_i)\,| \leq C_s\, h |\,(T,M,\theta,w)\,|_{s,0,\Omega_h}.$$

Step 3. *Estimate of* $\|\,e_T\,\|_{L^\infty(\Omega_h)}$ *and* $\|\,e_w\,\|_{L^\infty(\Omega_h)}$. Let us begin by estimating the error in T. For $x \in I_j$, we have

$$\begin{aligned} |\,e_T(x)\,| &= |\,T(x) - T(x_{j-1}) + e_T(x_{j-1}^+)\,| \\ &= |\,T(x) - T(x_{j-1}) + \widehat{e_T}(x_{j-1})\,| \\ &\leq C\, h\, \|\,T'\,\|_{L^\infty(I_j)} + |\,\widehat{e_T}(x_{j-1})\,|, \end{aligned}$$

and, since $T' = q$,

$$\|\,e_T\,\|_{L^\infty(\Omega_h)} \leq C\, h\, \left(\|\,q\,\|_{L^\infty(\Omega_h)} + |\,(T,M,\theta,w)\,|_{s,0,\Omega_h} \right).$$

Now, let us estimate the error in w. For $x \in I_j$, $j = 1, \ldots, N-1$,

$$\begin{aligned} |\,e_w(x)\,| &= |\,w(x) - w(x_j) + e_w(x_j^-)\,| \\ &= |\,w(x) - w(x_j) + \widehat{e_w}(x_j)\,| \\ &\leq C\, h\, \|\,w'\,\|_{L^\infty(I_j)} + |\,\widehat{e_T}(x_{j-1})\,|, \end{aligned}$$

and, since $w' = \theta - d^2\, T/GA$,

$$\|\,e_w\,\|_{L^\infty(I_j)} \leq C\, h\, |\,(T,M,\theta,w)\,|_{s,0,\Omega_h}.$$

It remains to consider the interval I_N. It enough to show that $e_w(1^-)$ is of order h^2. To do that, we note that, by the definition of $\widehat{T}_h(1)$,

$$e_w(1^-) = \frac{h}{\mathsf{c}}(\widehat{e_T}(1) - e_T(1^-)),$$

and hence,

$$|\,e_w(1^-)\,| \leq C\, \frac{h^2}{\mathsf{c}} \left(\|\,q\,\|_{L^\infty(\Omega_h)} + |\,(T,M,\theta,w)\,|_{s,0,\Omega_h} \right).$$

This implies that

$$\|\,e_w\,\|_{L^\infty(I_N)} \leq C\, h\, \|\,w'\,\|_{L^\infty(I_j)} + |\,e_w(1^-)\,|,$$

and since $w' = \theta - d^2\, T/GA$,

$$\|\,e_w\,\|_{L^\infty(I_N)} \leq C\, h\, |\,(T,M,\theta,w)\,|_{s,0,\Omega_h}.$$

Step 4. *Conclusion.* The estimates of $\|\,e_M\,\|_{L^\infty(\Omega_h)}$, $\|\,e_\theta\,\|_{L^\infty(\Omega_h)}$, and $|\,e_\theta(1^-)\,|$ can be obtained in a similar way. This completes the proof. ∎

4.2. The *hp*-version of the method. It is straightforward to extend the error analysis we have presented to the DG method that takes the approximation $(T_h, M_h, \theta_h, w_h)$ on the interval I_j in the space $P^{p_j}(I_j) \times P^{p_j}(I_j) \times P^{p_j}(I_j) \times P^{p_j}(I_j)$. It is also possible to choose different polynomial degrees for different variables in each interval according to Theorem 2.1; see [10].

4.3. Two other families of locking-free DG methods. In this paper we have analyzed a specific DG method, namely, the method which corresponds to Case 2 of Theorem 2.1. Reasons for concentrating on this particular case were given in section 2.3. In this subsection we briefly mention theoretical results for DG methods corresponding to Cases 1 and 3 of Theorem 2.1.

Given a polynomial degree $p \geq 0$, let

$$C_{14}(x) = -C_{23}(x) = -C_{32}(x) = C_{41}(x) = \mathsf{c} \quad \forall x \in \mathscr{E}_h$$

for some arbitrary positive number c independent of the mesh-size h or the polynomial degree p. Suppose further that

$$(C_{ii}(x) - 1/2)^2 \leq \mathsf{c} \qquad \text{for } i = 1, 2, 3, 4,$$
$$C_{12}^2(x),\, C_{13}^2(x),\, C_{21}^2(x),\, C_{24}^2(x),\, C_{31}^2(x),\, C_{34}^2(x),\, C_{42}^2(x),\, C_{43}^2 \leq \mathsf{c}$$

for all $x \in \mathscr{E}_h$.

Suppose that the polynomial degrees defining the DG method are such that $p_i = p$ for $i = 1, 2, 3, 4$. Then, the method defines a unique approximate solution by Case 1 of Theorem 2.1. Moreover, the method is free from shear locking. For these DG methods, the order of convergence of the error in the energy seminorm is now smaller by $1/2$, both in h and in p; this is, however, a sharp estimate. The orders of convergence of the errors in the L^2-norm and those of the errors in the numerical traces remain the same.

If we take $p_1 = p_2 = p + 1$ and $p_3 = p_4 = p$, then the existence and uniqueness of the DG approximation is guaranteed by Case 3 of Theorem 2.1. These methods as well are free from shear locking. On the other hand, although they have more degrees of freedom, the orders of convergence remain the same as ones for the methods with $p_i = p$ for $i = 1, 2, 3, 4$. This is simply due to the fact that θ_h and w_h are still being approximated by polynomials of degree p, even though T_h and M_h are being approximated by those of degree $p+1$. Once again, however, these estimates are sharp; see [7].

5. Numerical results. In this section, we display numerical results verifying our theoretical findings. For a set of numerical experiments verifying the superconvergence predicted by Theorem 2.4 we refer to [8]. Therein, it was also shown how to post-process the original DG solution in an element-by-element fashion to obtain a better approximation which converges to the exact solution with order $2\,p + 1$ *uniformly* throughout the domain, rather than just at the nodes of the mesh.

We solve the equations (2.2) with $q(x) = e^x$, $(EI)(x) = e^x$, $(GA)(x) = e^{-x}$ together with the boundary conditions $w_0 = w_1 = \theta_0 = \theta_1 = 0$. To verify that the DG method is locking-free, the thickness of the beam, d, is taken to be 10^{-2} and then decreased to 10^{-16}.

We consider two DG methods. The first is the one analyzed in full detail in this paper; it is defined by the numerical traces (2.12)–(2.14). The second is a particular

example of a class of DG methods analyzed in [7]. Its numerical fluxes are obtained by setting

$$C_{14}(x) = -C_{23}(x) = -C_{32}(x) = C_{41}(x) = 1 \quad \forall x \in \mathscr{E}_h$$

and $C_{ij} = 0$ at all nodes for all the remaining coefficients.

We display our results in Tables 1 through 4. Therein, p indicates the polynomial degree we used to define the DG method, and "mesh = i" means we employed a uniform mesh with 2^i elements. We also display numerical rates of convergence which are computed as follows. Let $e_u(i)$ denote the error where a mesh with 2^i elements has been employed to obtain the DG solution. Then the order of convergence, r_i, at level i is defined as

$$r_i := \frac{\log\left(\frac{e_u(i-1)}{e_u(i)}\right)}{\log 2}.$$

In Tables 1 and 2, we display the numerical results for the first DG method, for $d = 10^{-2}$ and $d = 10^{-16}$, respectively. We see that the optimal rates of convergence predicted by the error estimates given in section 2 are indeed achieved. As predicted by our error estimates the DG method is completely robust with respect to this parameter. In Tables 3 and 4, we display the numerical results for the second DG method. We also see that the predicted orders of convergence are actually achieved. Notice, in particular, that the energy seminorm converges with an order which is smaller by 1/2, as expected.

TABLE 1
History of convergence for $d = 10^{-2}$ for the first DG method.

		$\lvert e\rvert_{\mathscr{A}_h}$		$\lVert e_T\rVert_{0,\Omega_h}$		$\lVert e_M\rVert_{0,\Omega_h}$		$\lVert e_\theta\rVert_{0,\Omega_h}$		$\lVert e_w\rVert_{0,\Omega_h}$	
p	mesh	error	order	error	order	error	order	error	order	error	order
	3	1.21E-01	1.38	8.67E-02	0.81	2.35E-02	1.70	4.48E-03	2.15	1.43E-02	2.42
0	4	4.47E-02	1.44	5.40E-02	0.68	1.01E-02	1.22	1.27E-03	1.82	2.58E-03	2.44
	5	1.63E-02	1.46	2.90E-02	0.90	4.95E-03	1.03	6.47E-04	0.97	4.93E-04	2.38
	6	5.96E-03	1.45	1.48E-02	0.97	2.47E-03	1.00	3.40E-04	0.93	1.25E-04	1.97
	3	3.50E-03	2.36	1.74E-03	2.09	1.66E-03	1.95	2.87E-04	2.54	1.84E-04	3.30
1	4	6.76E-04	2.37	4.27E-04	2.03	4.20E-04	1.98	7.05E-05	2.03	2.00E-05	3.20
	5	1.33E-04	2.34	1.06E-04	2.01	1.06E-04	1.99	1.79E-05	1.98	2.88E-06	2.80
	6	2.74E-05	2.28	2.65E-05	2.00	2.65E-05	2.00	4.51E-06	1.99	5.90E-07	2.29
	3	4.31E-05	3.38	1.69E-05	2.99	1.68E-05	2.98	1.15E-05	3.01	3.62E-06	3.70
2	4	4.08E-06	3.40	2.12E-06	3.00	2.12E-06	2.99	1.44E-06	2.99	3.79E-07	3.25
	5	3.91E-07	3.38	2.66E-07	3.00	2.65E-07	3.00	1.81E-07	2.99	4.56E-08	3.06
	6	3.88E-08	3.33	3.32E-08	3.00	3.32E-08	3.00	2.27E-08	3.00	5.67E-09	3.01
	3	3.81E-07	4.38	1.30E-07	3.99	1.30E-07	3.98	1.68E-07	4.04	8.95E-08	4.08
3	4	1.79E-08	4.41	8.16E-09	4.00	8.14E-09	3.99	1.05E-08	4.01	5.56E-09	4.01
	5	8.45E-10	4.40	5.11E-10	4.00	5.10E-10	4.00	6.53E-10	4.00	3.48E-10	4.00
	6	4.11E-11	4.36	3.19E-11	4.00	3.19E-11	4.00	4.08E-11	4.00	2.18E-11	4.00
	3	2.63E-09	5.39	8.07E-10	4.99	8.04E-10	4.99	1.50E-09	4.97	1.02E-09	4.96
4	4	6.15E-11	5.42	2.53E-11	5.00	2.52E-11	4.99	4.73E-11	4.99	3.22E-11	4.98
	5	1.44E-12	5.42	7.91E-13	5.00	7.90E-13	5.00	1.48E-12	5.00	1.01E-12	4.99
	6	3.45E-14	5.38	2.47E-14	5.00	2.47E-14	5.00	4.64E-14	5.00	3.16E-14	5.00

TABLE 2
History of convergence for $d = 10^{-16}$ for the first DG method.

		$\|e\|_{\mathscr{A}_h}$		$\|e_T\|_{0,\Omega_h}$		$\|e_M\|_{0,\Omega_h}$		$\|e_\theta\|_{0,\Omega_h}$		$\|e_w\|_{0,\Omega_h}$	
p	mesh	error	order	error	order	error	order	error	order	error	order
	3	1.21E-01	1.38	8.67E-02	0.81	2.35E-02	1.70	4.48E-03	2.15	1.43E-02	2.42
0	4	4.47E-02	1.44	5.40E-02	0.68	1.01E-02	1.22	1.27E-03	1.82	2.58E-03	2.44
	5	1.63E-02	1.46	2.90E-02	0.90	4.95E-03	1.03	6.47E-04	0.97	4.93E-04	2.38
	6	5.96E-03	1.45	1.48E-02	0.97	2.47E-03	1.00	3.40E-04	0.93	1.25E-04	1.97
	3	3.50E-03	2.36	1.74E-03	2.09	1.66E-03	1.95	2.87E-04	2.54	1.85E-04	3.30
1	4	6.76E-04	2.37	4.27E-04	2.03	4.20E-04	1.98	7.04E-05	2.03	2.01E-05	3.20
	5	1.33E-04	2.34	1.06E-04	2.01	1.06E-04	1.99	1.79E-05	1.98	2.89E-06	2.80
	6	2.74E-05	2.28	2.65E-05	2.00	2.65E-05	2.00	4.51E-06	1.99	5.91E-07	2.29
	3	4.31E-05	3.38	1.69E-05	2.99	1.68E-05	2.98	1.15E-05	3.01	3.62E-06	3.70
2	4	4.08E-06	3.40	2.12E-06	3.00	2.12E-06	2.99	1.44E-06	2.99	3.80E-07	3.25
	5	3.91E-07	3.38	2.66E-07	3.00	2.65E-07	3.00	1.81E-07	2.99	4.56E-08	3.06
	6	3.88E-08	3.33	3.32E-08	3.00	3.32E-08	3.00	2.27E-08	3.00	5.67E-09	3.01
	3	3.81E-07	4.38	1.30E-07	3.99	1.30E-07	3.98	1.68E-07	4.04	8.95E-08	4.08
3	4	1.79E-08	4.41	8.16E-09	4.00	8.14E-09	3.99	1.04E-08	4.01	5.57E-09	4.01
	5	8.45E-10	4.40	5.11E-10	4.00	5.10E-10	4.00	6.53E-10	4.00	3.49E-10	4.00
	6	4.11E-11	4.36	3.19E-11	4.00	3.19E-11	4.00	4.08E-11	4.00	2.18E-11	4.00
	3	2.63E-09	5.39	8.07E-10	4.99	8.04E-10	4.99	1.50E-09	4.97	1.02E-09	4.96
4	4	6.15E-11	5.42	2.53E-11	5.00	2.52E-11	4.99	4.73E-11	4.99	3.21E-11	4.98
	5	1.44E-12	5.42	7.91E-13	5.00	7.90E-13	5.00	1.48E-12	5.00	1.01E-12	4.99
	6	3.45E-14	5.38	2.47E-14	5.00	2.47E-14	5.00	4.64E-14	5.00	3.15E-14	5.00

TABLE 3
History of convergence for $d = 10^{-2}$ for the second DG method.

		$\|e\|_{\mathscr{A}_h}$		$\|e_T\|_{0,\Omega_h}$		$\|e_M\|_{0,\Omega_h}$		$\|e_\theta\|_{0,\Omega_h}$		$\|e_w\|_{0,\Omega_h}$	
p	mesh	error	order	error	order	error	order	error	order	error	order
	3	6.11E-01	0.40	9.65E-02	0.82	2.87E-02	0.74	4.03E-02	0.28	2.11E-01	0.80
0	4	4.48E-01	0.45	5.92E-02	0.71	1.72E-02	0.74	2.56E-02	0.66	1.13E-01	0.90
	5	3.22E-01	0.47	3.67E-02	0.69	1.07E-02	0.68	1.42E-02	0.84	5.83E-02	0.95
	6	2.30E-01	0.49	2.17E-02	0.76	6.39E-03	0.74	7.49E-03	0.93	2.96E-02	0.98
	3	9.96E-03	1.57	1.06E-03	2.02	1.05E-03	2.00	7.95E-04	2.15	7.77E-04	2.15
1	4	3.42E-03	1.54	2.63E-04	2.01	2.63E-04	2.00	1.87E-04	2.09	1.82E-04	2.10
	5	1.19E-03	1.52	6.54E-05	2.01	6.56E-05	2.00	4.51E-05	2.05	4.38E-05	2.05
	6	4.18E-04	1.51	1.63E-05	2.00	1.64E-05	2.00	1.11E-05	2.03	1.07E-05	2.03
	3	2.43E-04	2.41	1.13E-05	2.99	1.73E-05	2.83	2.60E-05	2.89	2.39E-05	2.85
2	4	4.44E-05	2.45	1.44E-06	2.97	2.36E-06	2.87	3.33E-06	2.97	3.13E-06	2.94
	5	7.99E-06	2.48	1.82E-07	2.98	3.10E-07	2.93	4.20E-07	2.99	3.99E-07	2.97
	6	1.42E-06	2.49	2.30E-08	2.99	3.98E-08	2.96	5.27E-08	2.99	5.04E-08	2.99
	3	1.58E-06	3.57	9.46E-08	4.01	1.12E-07	4.04	1.27E-07	4.07	8.38E-08	4.10
3	4	1.36E-07	3.54	5.86E-09	4.01	6.86E-09	4.03	7.77E-09	4.03	5.03E-09	4.06
	5	1.18E-08	3.52	3.65E-10	4.01	4.23E-10	4.02	4.79E-10	4.02	3.08E-10	4.03
	6	1.04E-09	3.51	2.27E-11	4.00	2.63E-11	4.01	2.98E-11	4.01	1.90E-11	4.02
	3	2.55E-08	4.40	1.40E-09	4.82	1.97E-09	4.80	1.61E-09	4.96	1.39E-09	4.93
4	4	1.16E-09	4.45	4.70E-11	4.89	6.67E-11	4.89	5.03E-11	5.01	4.37E-11	4.99
	5	5.24E-11	4.47	1.53E-12	4.94	2.17E-12	4.94	1.56E-12	5.01	1.36E-12	5.00
	6	2.34E-12	4.49	4.88E-14	4.97	6.93E-14	4.97	4.84E-14	5.01	4.25E-14	5.00

TABLE 4
History of convergence for $d = 10^{-16}$ for the second DG method.

		$\lvert e\rvert_{\mathscr{A}_h}$		$\lVert e_T\rVert_{0,\Omega_h}$		$\lVert e_M\rVert_{0,\Omega_h}$		$\lVert e_\theta\rVert_{0,\Omega_h}$		$\lVert e_w\rVert_{0,\Omega_h}$	
p	mesh	error	order	error	order	error	order	error	order	error	order
	3	6.11E-01	0.40	9.69E-02	0.81	2.88E-02	0.73	4.03E-02	0.28	2.11E-01	0.80
0	4	4.48E-01	0.45	5.96E-02	0.70	1.73E-02	0.74	2.56E-02	0.66	1.13E-01	0.90
	5	3.22E-01	0.47	3.70E-02	0.69	1.08E-02	0.68	1.42E-02	0.84	5.83E-02	0.95
	6	2.30E-01	0.49	2.19E-02	0.76	6.46E-03	0.74	7.49E-03	0.93	2.96E-02	0.98
	3	9.96E-03	1.57	1.06E-03	2.02	1.05E-03	2.00	7.95E-04	2.15	7.77E-04	2.15
1	4	3.42E-03	1.54	2.63E-04	2.01	2.63E-04	2.00	1.87E-04	2.09	1.82E-04	2.10
	5	1.19E-03	1.52	6.54E-05	2.01	6.56E-05	2.00	4.51E-05	2.05	4.38E-05	2.05
	6	4.18E-04	1.51	1.63E-05	2.00	1.64E-05	2.00	1.11E-05	2.03	1.07E-05	2.03
	3	2.43E-04	2.40	1.13E-05	2.99	1.73E-05	2.83	2.60E-05	2.89	2.39E-05	2.85
2	4	4.44E-05	2.45	1.44E-06	2.97	2.36E-06	2.87	3.33E-06	2.97	3.13E-06	2.94
	5	7.99E-06	2.48	1.82E-07	2.98	3.10E-07	2.93	4.20E-07	2.99	3.99E-07	2.97
	6	1.42E-06	2.49	2.30E-08	2.99	3.98E-08	2.96	5.27E-08	2.99	5.04E-08	2.99
	3	1.58E-06	3.57	9.46E-08	4.01	1.12E-07	4.04	1.27E-07	4.07	8.38E-08	4.10
3	4	1.36E-07	3.54	5.86E-09	4.01	6.85E-09	4.03	7.77E-09	4.03	5.03E-09	4.06
	5	1.18E-08	3.52	3.65E-10	4.01	4.23E-10	4.02	4.79E-10	4.02	3.08E-10	4.03
	6	1.04E-09	3.51	2.27E-11	4.00	2.63E-11	4.01	2.97E-11	4.01	1.90E-11	4.02
	3	2.54E-08	4.40	1.39E-09	4.83	1.97E-09	4.80	1.61E-09	4.99	1.39E-09	4.93
4	4	1.16E-09	4.45	4.69E-11	4.89	6.67E-11	4.89	5.02E-11	5.01	4.37E-11	4.99
	5	5.23E-11	4.47	1.53E-12	4.94	2.17E-12	4.94	1.56E-12	5.01	1.36E-12	5.00
	6	2.33E-12	4.49	4.87E-14	4.97	6.93E-14	4.97	4.84E-14	5.01	4.25E-14	5.00

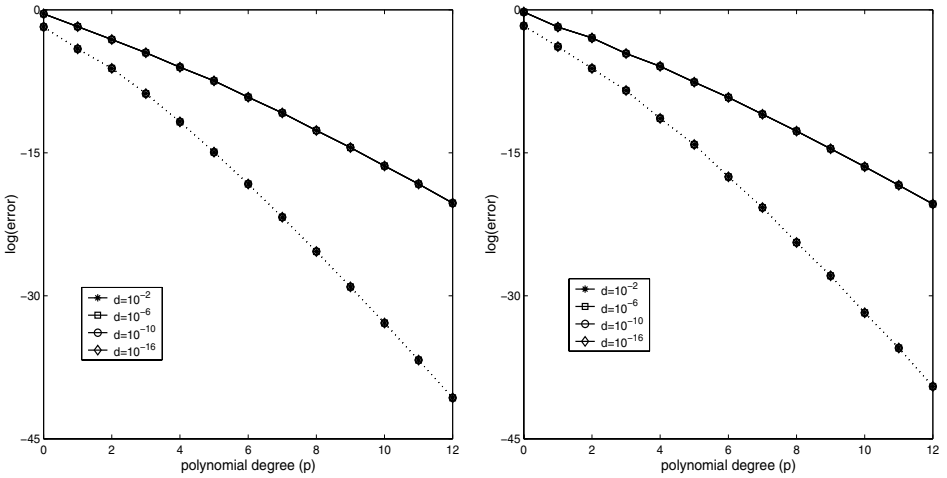


FIG. 2. *Robust exponential convergence for $\lVert e_w\rVert_{0,\Omega_h}$ (solid line) and $\left\lVert \widehat{e_w}\right\rVert_{L^\infty(\mathscr{E}_h)}$ (dotted line). First DG method (left) and second DG method (right).*

In Figure 2, we display the performance of the p-version of both DG methods. On the left we display the semilogarithmic plot of $\|e_w\|_{0,\Omega_h}$ and that of $\|\widehat{e_w}\|_{L^\infty(\mathscr{E}_h)}$ versus the polynomial degree. The DG solution is computed on a uniform mesh with 2 elements, and the polynomial degree is increased from 0 to 12, while the mesh is kept fixed. We display four different plots on the same figure, namely $d = 10^{-2}, d = 10^{-6}, d = 10^{-10}$, and $d = 10^{-16}$. Notice that the curves are practically on top of each other and that the order of convergence of the maximum error in the numerical trace of the displacement is twice as fast as that of the $L^2(\Omega_h)$-norm of its error. Similar results are obtained for the other three variables.

6. Concluding remarks. We have shown that DG methods can be devised which are free from shear locking. We achieved this by a careful study of the relation between the definition of the numerical traces and the corresponding convergence properties of the methods. This provides a powerful approach for devising locking-free DG methods for the much more challenging problems of thin plates and shells which constitutes the subject of ongoing work.

We have not addressed the issue of the actual implementation of the DG methods. In particular, we have not identified DG methods that could be considered extensions of the LDG methods [11] for second-order elliptic problems. This constitutes the subject of ongoing work.

Finally, we note that even though we have carried out our error analysis for the *hp*-version of the methods, we only displayed numerical results for either their h- or p-versions. The study of *hp*-adaptive algorithms for these methods constitutes the subject of ongoing work.

REFERENCES

[1] D. N. Arnold, *Discretization by finite elements of a model parameter dependent problem*, Numer. Math., 37 (1981), pp. 405–421.

[2] D. N. Arnold and F. Brezzi, *Locking-free finite element methods for shells*, Math. Comp., 66 (1997), pp. 1–14.

[3] D. N. Arnold, F. Brezzi, and D. Marini, *A family of locking-free discontinuous Galerkin finite elements for the Reissner-Mindlin plate*, J. Sci. Comput., 22 (2005), pp. 25–45.

[4] D. G. Ashwell and A. B. Sabir, *Limitations of certain curved finite elements when applied to arches*, Internat. J. Mech. Sci., 13 (1971), pp. 133–139.

[5] I. Babuška and M. Suri, *On locking and robustness in the finite element method*, SIAM J. Numer. Anal., 29 (1992), pp. 1261–1293.

[6] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab, *Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems*, Math. Comp., 71 (2002), pp. 455–478.

[7] F. Celiker, *Discontinuous Galerkin Methods for Structural Mechanics*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 2005.

[8] F. Celiker and B. Cockburn, *Element-by-element post-processing of discontinuous Galerkin methods for Timoshenko beams*, J. Sci. Comput., 27 (2006), pp. 177–187.

[9] F. Celiker and B. Cockburn, *Superconvergence of the numerical traces of discontinuous Galerkin and hybridized methods for convection-diffusion problems in one space dimension*, Math. Comp., 76 (2007), 67–96.

[10] F. Celiker, B. Cockburn, S. Güzey, R. Kanapady, S.-C. Soon, H. K. Stolarski, and K. K. Tamma, *Discontinuous Galerkin methods for Timoshenko beams*, in Numerical Mathematics and Advanced Applications, ENUMATH 2003, Springer, 2004, pp. 221–231.

[11] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[12] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[13] T. J. R. Hughes, R. L. Taylor, and W. Kanoknukulchai, *A simple and efficient finite element for plate bending*, Internat. J. Numer. Methods Engrg., 11 (1977), pp. 1529–1543.
[14] S. W. Lee and T. H. H. Pian, *Improvements of plate and shell finite elements by mixed formulations*, AIAA Journal, 16 (1978), pp. 29–34.
[15] L. Li, *Discretization of the Timoshenko beam problem by the p and the h-p versions of the finite element method*, Numer. Math., 57 (1990), pp. 413–420.
[16] D. S. Malkus and T. J. R. Hughes, *Mixed finite element methods—reduced and selective integration techniques: A unification of concepts*, Comput. Methods Appl. Mech. Engrg., 15 (1978), pp. 63–81.
[17] H. Parisch, *A critical survey of the* 9*-node degenerated shell element with special emphasis on thin shell application and reduced integration*, Comput. Methods Appl. Mech. Engrg., 20 (1979), pp. 323–350.
[18] D. Schötzau and C. Schwab, *An hp a priori error analysis of the DG time-stepping method for initial value problems*, Calcolo, 37 (2000), pp. 207–232.
[19] D. Schötzau and C. Schwab, *Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method*, SIAM J. Numer. Anal., 38 (2000), pp. 837–875.
[20] C. Schwab, *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*, Oxford University Press, New York, 1998.
[21] H. Stolarski and T. Belytschko, *Membrane locking and reduced integration for curved elements*, J. Appl. Mech., 49 (1982), pp. 172–176.
[22] H. Stolarski and T. Belytschko, *Shear and membrane locking in curved* C^0 *elements*, Comput. Methods Appl. Mech. Engrg., 41 (1983), pp. 279–296.
[23] H. Stolarski, T. Belytschko, and S.-H. Lee, *A review of shell finite elements and corotational theories*, Comput. Mech. Adv., 2 (1995), pp. 125–212.
[24] S. P. Timoshenko, *On the correction for shear of the differential equation for transverse vibrations of prismatic bars*, Philosophical Magazine, 41 (1921), pp. 744–746.
[25] Z. Zhang, *Arch beam models: Finite element analysis and superconvergence*, Numer. Math., 61 (1992), pp. 117–143.
[26] Z. Zhang, *A note on the hybrid-mixed* C^0 *curved beam elements*, Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 243–252.
[27] Z. Zhang, *Locking and robustness in the finite element method for circular arch problem*, Numer. Math., 69 (1995), pp. 509–522.
[28] O. Zienkiewicz, R. L. Taylor, and J. M. Too, *Reduced integration technique in general analysis of plates and shells*, Internat. J. Numer. Methods Engrg., 3 (1971), pp. 275–290.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2326–2341

EXISTENCE OF SOLUTIONS TO SYSTEMS OF UNDERDETERMINED EQUATIONS AND SPHERICAL DESIGNS*

XIAOJUN CHEN† AND ROBERT S. WOMERSLEY‡

Abstract. This paper is concerned with proving the existence of solutions to an underdetermined system of equations and with the application to existence of spherical t-designs with $(t+1)^2$ points on the unit sphere S^2 in R^3. We show that the construction of spherical designs is equivalent to solution of underdetermined equations. A new verification method for underdetermined equations is derived using Brouwer's fixed point theorem. Application of the method provides spherical t-designs which are close to extremal (maximum determinant) points and have the optimal order $O(t^2)$ for the number of points. An error bound for the computed spherical designs is provided.

Key words. verification, underdetermined system, spherical designs, extremal points, interpolation, numerical integration

AMS subject classifications. 65H10, 65G20, 65D30, 65D05

DOI. 10.1137/050626636

1. Introduction.

Let $c : R^n \to R^m$ be a continuously differentiable function with $m < n$. Suppose that $\hat{x}$ is an approximate solution of the underdetermined system of nonlinear equations

$$c(x) = 0 \tag{1.1}$$

and the Jacobian $c'(x)$ of c at $\hat{x}$ has full row rank. We are interested in the existence of a solution of (1.1) in a neighborhood of $\hat{x}$.

Underdetermined systems of equations arise in constrained optimization problems, continuation methods for underdetermined equations, etc. [3, 12, 14, 21]. This paper gives a verification method for solutions of the underdetermined equations (1.1). The main difficulty in proving the existence of solutions of an underdetermined system of equations is that the Jacobian $c'(x)$ is an $m \times n$ matrix with $m < n$. Let $c'(\hat{x})^+$ be the Moore–Penrose pseudoinverse of $c'(\hat{x})$. A popular method for verifying the existence of solutions of nonlinear equations is to use a Krawczyk-type interval operator [1]. Replacing the inverse by a Moore–Penrose pseudoinverse, we can get a Krawczyk-type interval operator

$$\mathcal{K}(X) = \hat{x} - c'(\hat{x})^+ c(\hat{x}) + (I - c'(\hat{x})^+ C'(X))(X - \hat{x}), \tag{1.2}$$

where X is an interval in R^n defined by

$$X = [\hat{x} - h, \hat{x} + h], \qquad h \in R^n, \quad h \geq 0,$$

and $C'(X)$ is an interval arithmetic evaluation satisfying

$$c'(x) \in C'(X) \qquad \text{for } x \in X.$$

*Received by the editors March 12, 2005; accepted for publication (in revised form) June 2, 2006; published electronically November 24, 2006. This research was supported by the Japan Society of Promotion and Science and the Australian Research Council.
http://www.siam.org/journals/sinum/44-6/62663.html

†Department of Mathematical Sciences, Hirosaki University, Hirosaki 036-8561, Japan (chen@cc.hirosaki-u.ac.jp).

‡School of Mathematics, University of New South Wales, Sydney 2052, Australia (R.Womersley@unsw.edu.au).

It can be shown [1] that there is a solution of (1.1) in X if

$$\mathcal{K}(X) \subseteq X \tag{1.3}$$

and $c'(\hat{x})$ has full row rank. However, the enclosure (1.3) rarely holds due to the equality [8]

$$\|I - c'(\hat{x})^+ c'(\hat{x})\|_2 = \min\{1, n-m\}$$

and the fact that

$$\mathcal{K}(X) \subseteq X \quad \Rightarrow \quad \|I - c'(\hat{x})^+ c'(x)\|_\infty \leq 1 \quad \forall\, x \in X.$$

In section 2 we present a new verification method for underdetermined systems of (1.1) which does not need the generalized inverse $c'(\hat{x})^+$.

A cubature (numerical integration) rule for the unit sphere $S^2 = \{y \in R^3 \ : \ \|y\|_2 = 1\}$ is a set of N points $y_\ell \in S^2$ and weights w_ℓ for $\ell = 1, \ldots, N$ such that

$$\int_{S^2} f(y)dy \approx \sum_{\ell=1}^{N} w_\ell f(y_\ell).$$

Let $\mathbb{P}_t \equiv \mathbb{P}_t(S^2)$ be the linear space of restrictions of polynomials of degree $\leq t$ in 3 variables to S^2. The dimension of the space $\mathbb{P}_t$ is $d_t := (t+1)^2$. Spherical t-designs, introduced in [5], are sets of N points $\{y_1, y_2, \ldots, y_N\} \subset S^2$ such that the equally weighted ($w_\ell = |S^2|/N = 4\pi/N$, $\ell = 1, \ldots, N$) cubature rule is exact for all spherical polynomials of degree at most t, that is,

$$\int_{S^2} p(y)dy = \frac{4\pi}{N} \sum_{\ell=1}^{N} p(y_\ell) \qquad \forall p \in \mathbb{P}_t.$$

For $t \geq 1$, the existence of a spherical t-design was proved in [19]. Commonly, the interest is in the smallest number N_t^* of points required to give a spherical t-design. Lower bounds on N_t^* given in [5] are

$$N_t^* \geq \frac{(t+1)(t+3)}{4} \quad \text{if } t \text{ is odd},$$
$$N_t^* \geq \frac{(t+2)^2}{4} \quad \text{if } t \text{ is even}.$$

A spherical t-design which achieves the lower bounds is called a tight spherical t-design. However, for $t \geq 2$, it is known that tight spherical t-designs do not exist [5]. Hardin and Sloane [7] have extensively investigated spherical designs on S^2 and suggested a sequence of putative spherical t-designs with $\frac{1}{2}t^2 + o(t^2)$ points. A 7-design with 24 points was first found by McLaren in 1963 [13]. Korevaar and Meyers [10] considered the construction for spherical t-designs with $O(t^3)$ points on S^2. An approach for the numerical calculation of spherical designs using multiobjective optimization was studied by Maier [11], and computational proof of the existence of spherical designs using interval methods [9] was investigated by Hardin and Sloane [7].

Extremal (or maximum determinant) points [20] are sets of $(t+1)^2$ points on S^2 which maximize the determinant of a basis matrix for an arbitrary basis of $\mathbb{P}_t$. Sloan and Womersley [20, 22] showed that extremal systems have very nice geometrical

properties as the points are well separated and the computed interpolatory cubature weights are positive ($w_\ell > |S^2|/(2N)$ for $\ell = 1, \ldots, N$ for degrees up to $t = 150$). Also the condition number of the basis matrix grows slowly, giving confidence in the calculated cubature weights. Proving the positivity of the cubature weights for all degrees t for the extremal points is still an open question. Other systems of points, such as minimum energy points, often have basis matrices with such high condition numbers that no confidence can be placed in the calculated cubature weights.

Equal weight cubature rules, or spherical designs, are simpler to implement and there is no question about the positivity of the weights. There are many different characterizations of spherical t-designs [6]. However, these can be very ill conditioned. Extremal points provide excellent starting points for numerically finding solutions to an underdetermined, but highly nonlinear, system of equations which characterize spherical t-designs with $(t+1)^2$ points. Application of the verification method to the system of equations then proves the existence of spherical t-designs which are close to the calculated points and have the optimal order $O(t^2)$ for the number of points. Moreover, spherical designs with $(t+1)^2$ points which also have a basis matrix with a determinant close to the maximum are simultaneously good for cubature and interpolation. Computed spherical t-designs with $(t+1)^2$ points for degrees up to $t = 50$ are available from http://www.maths.unsw.edu.au/∼rsw/Sphere.

The focus here is not on finding a spherical t-design with the minimal number of points, but rather proving the existence of spherical t-designs with $(t+1)^2$ points close to an extremal system. Once existence of a spherical design with $(t+1)^2$ points is established one can then look for *extremal spherical designs*, that is, systems of $(t+1)^2$ points which maximize the determinant of a basis matrix subject to the constraints that they are spherical t-designs.

In section 3 we reformulate the calculation of a spherical t-design with $(t+1)^2$ points as an underdetermined system of nonlinear equations (1.1) with $m = (t+1)^2-1$ equations and $n = 2(t+1)^2 - 3$ variables. We show that a sufficient and necessary condition for the existence of solutions to the system of equations is existence of a spherical t-design with $(t+1)^2$ points. In section 4, we apply the verification method to find new spherical t-designs. The computed spherical designs $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_{d_t}\}$ are compared with the extremal (maximum determinant) points, and error bounds of $\hat{Y}$ to exact spherical designs are given.

For a given $m \times n$ matrix A, let $A_{\mathcal{I}}$ be the submatrix of A whose entries lie in the columns of A indexed by $\mathcal{I}$. For a given vector $x \in R^n$, let $x_{\mathcal{I}}$ be the subvector of x whose entries of x are indexed by $\mathcal{I}$.

2. A verification method. Let $\hat{x}$ be a computed solution of (1.1). Let $\mathcal{B}$ be an index set $\{k_1, k_2, \ldots, k_m\}$ such that $c'_{\mathcal{B}}(\hat{x}) \in R^{m\times m}$ is nonsingular. Define the function $H : R^n \to R^n$ by

$$H_{\mathcal{B}}(x) = x_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1} c(x), \tag{2.1}$$

$$H_{\mathcal{N}}(x) = x_{\mathcal{N}} - \alpha(x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}), \tag{2.2}$$

where $\mathcal{N} = \{1, 2, \ldots, n\}/\mathcal{B}$ and $\alpha \in (0, 1)$ is a constant. Obviously, if $x^* \in R^n$ is a fixed point of H, that is, $H(x^*) = x^*$, then we have $c(x^*) = 0$ with $x^*_{\mathcal{N}} = \hat{x}_{\mathcal{N}}$. Choose two nonnegative numbers r_1 and r_2 and define the convex set

$$X = \{\, x \in R^n : \|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \le r_1, \|x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}\| \le r_2 \,\}.$$

THEOREM 2.1. *Suppose that $c : R^n \to R^m$ is continuously differentiable, c' has full row rank at $\hat{x}$, and*

$$\|c'_{\mathcal{B}}(x) - c'_{\mathcal{B}}(\hat{x})\| \le K\|x - \hat{x}\| \text{ for } x \in X. \tag{2.3}$$

(1) *There is a solution of* (1.1) *in X if*

$$\|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| + \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\left(\frac{1}{2}K(r_1+r_2)r_1 + \max_{x\in X}\|c'_{\mathcal{N}}(x)\|r_2\right) \le r_1. \tag{2.4}$$

(2) *There is no solution of* (1.1) *in X if*

$$\|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| - \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\left(\frac{1}{2}K(r_1+r_2)r_1 + \max_{x\in X}\|c'_{\mathcal{N}}(x)\|r_2\right) > r_1. \tag{2.5}$$

Proof. (1) By the continuity of $c'(x)$ and the mean value theorem, we find

$$\begin{aligned}
H_{\mathcal{B}}(x) &= \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x}) + x_{\mathcal{B}} - \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}(c(x) - c(\hat{x}))\\
&= \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x}) + x_{\mathcal{B}} - \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}\int_0^1 c'(x + t(\hat{x}-x))(x-\hat{x})dt\\
&= \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x}) + x_{\mathcal{B}} - \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}\int_0^1 c'_{\mathcal{B}}(x + t(\hat{x}-x))(x_{\mathcal{B}}-\hat{x}_{\mathcal{B}})dt\\
&\quad - c'_{\mathcal{B}}(\hat{x})^{-1}\int_0^1 c'_{\mathcal{N}}(x + t(\hat{x}-x))(x_{\mathcal{N}}-\hat{x}_{\mathcal{N}})dt\\
&= \hat{x}_{\mathcal{B}} - c'_{\mathcal{B}}(\hat{x})^{-1}\left[c(\hat{x}) + \int_0^1 (c'_{\mathcal{B}}(\hat{x}) - c'_{\mathcal{B}}(x + t(\hat{x}-x)))(x_{\mathcal{B}}-\hat{x}_{\mathcal{B}})dt\right.\\
&\qquad \left. + \int_0^1 c'_{\mathcal{N}}(x + t(\hat{x}-x))(x_{\mathcal{N}}-\hat{x}_{\mathcal{N}})dt\right].
\end{aligned}$$

Therefore, for any $x \in X$, we have

$$\begin{aligned}
&\|H_{\mathcal{B}}(x) - \hat{x}_{\mathcal{B}}\|\\
&\le \|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| + \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\int_0^1 \|c'_{\mathcal{B}}(\hat{x}) - c'_{\mathcal{B}}(x + t(\hat{x}-x))\|\|x_{\mathcal{B}}-\hat{x}_{\mathcal{B}}\|dt\\
&\quad + \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\int_0^1 \|c'_{\mathcal{N}}(x + t(\hat{x}-x))\|\|x_{\mathcal{N}}-\hat{x}_{\mathcal{N}}\|dt\\
&\le \|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| + \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\left(\int_0^1 (1-t)K\|\hat{x}-x\|r_1 dt + \int_0^1 \max_{x\in X}\|c'_{\mathcal{N}}(x)\|r_2 dt\right)\\
&\le \|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| + \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\left(\frac{1}{2}K(r_1+r_2)r_1 + \max_{x\in X}\|c'_{\mathcal{N}}(x)\|r_2\right).
\end{aligned}$$

Here we use the facts that $x + t(\hat{x}-x) \in X$, $\|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \le r_1$, and $\|x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}\| \le r_2$ for all $x \in X$ and $t \in [0,1]$.

This implies that if (2.4) holds, then for any $x \in X$ we have

$$\|H_{\mathcal{B}}(x) - \hat{x}_{\mathcal{B}}\| \le r_1.$$

Moreover, by the definition of H, we always have

$$\|H_{\mathcal{N}}(x) - \hat{x}_{\mathcal{N}}\| = (1-\alpha)\|x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}\| \le r_2.$$

Therefore, (2.4) implies that H maps X into itself; that is,

$$H(x) \in X \quad \text{for any } x \in X. \tag{2.6}$$

Using Brouwer's fixed point theorem, (2.6) implies that there is a fixed point x^* of H in X. From the definition of H, x^* is a solution of (1.1).

(2) Assume that (2.5) holds and there is a solution x^* in X. Following the proof for part (1), we have

$$\begin{aligned}
r_1 &\geq \|x^*_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \\
&= \|H_{\mathcal{B}}(x^*) - \hat{x}_{\mathcal{B}}\| \\
&\geq \|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| - \|c'_{\mathcal{B}}(\hat{x})^{-1}\| \int_0^1 \|c'_{\mathcal{B}}(\hat{x}) - c'_{\mathcal{B}}(x^* + t(\hat{x} - x^*))\|\|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\|dt \\
&\quad -\|c'_{\mathcal{B}}(\hat{x})^{-1}\| \int_0^1 \|c'_{\mathcal{N}}(x^* + t(\hat{x} - x^*))\|\|x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}\|dt \\
&\geq \|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\| - \|c'_{\mathcal{B}}(\hat{x})^{-1}\| \left(\frac{1}{2}K(r_1 + r_2)r_1 + \max_{x\in X}\|c'_{\mathcal{N}}(x)\|r_2\right) > r_1.
\end{aligned}$$

This is a contradiction, which completes the proof. □

Without loss of generality, we assume that $r_1 \neq 0$. Let $\tau \in (0, \frac{1}{2})$. Define a subset of X:

$$X_\tau = \{\, x \mid \|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \leq \tau r_1, \|x_{\mathcal{N}} - \hat{x}_{\mathcal{N}}\| \leq \tau r_2\}.$$

Then we have the following corollary.

COROLLARY 2.2. *Under the assumptions of Theorem* 2.1*, inequality* (2.4) *implies that* $c'_{\mathcal{B}}(x)$ *is nonsingular for all* $x \in X_\tau$ *and the solution* x^* *of* (1.1) *with* $x^*_{\mathcal{N}} = \hat{x}_{\mathcal{N}}$ *is unique in* X_τ.

Proof. For any $x \in X_\tau$ $(x \neq \hat{x})$, inequality (2.4) implies that

$$\begin{aligned}
r_1 &\geq \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\frac{1}{2}K(r_1 + r_2)r_1 \\
&\geq \|c'_{\mathcal{B}}(\hat{x})^{-1}\|\frac{1}{2\tau}K\|x - \hat{x}\|r_1 \\
&> \|c'_{\mathcal{B}}(\hat{x})^{-1}\|K\|x - \hat{x}\|r_1 \\
&\geq r_1\|c'_{\mathcal{B}}(\hat{x})^{-1}\|\|c'_{\mathcal{B}}(\hat{x}) - c'_{\mathcal{B}}(x)\| \\
&\geq r_1\|I - c'_{\mathcal{B}}(\hat{x})^{-1}c'_{\mathcal{B}}(x)\|.
\end{aligned}$$

Dividing r_1 in both sides, we find

$$\|I - c'_{\mathcal{B}}(\hat{x})^{-1}c'_{\mathcal{B}}(x)\| < 1.$$

Hence $c'_{\mathcal{B}}(x)$ is nonsingular. By the implicit function theorem [16], the solution x^* of (1.1) with $x^*_{\mathcal{N}} = \hat{x}_{\mathcal{N}}$ is unique in X_τ. □

Remark 2.1. For the case $m = n$, we have $x = x_{\mathcal{B}}$, $c'_{\mathcal{B}}(x) = c'(x)$, and (2.4) reduces to

$$\|c'(\hat{x})^{-1}c(\hat{x})\| + \frac{1}{2}K\|c'(\hat{x})^{-1}\|r^2 \leq r. \tag{2.7}$$

This is a quadratic inequality in r. If

$$\rho := K\|c'(\hat{x})^{-1}c(\hat{x})\|\|c'(\hat{x})^{-1}\| \le \frac{1}{2}, \tag{2.8}$$

then (2.7) holds for all r satisfying

$$\frac{1-\sqrt{1-2\rho}}{K\|c'(\hat{x})^{-1}\|} \le r \le \frac{1+\sqrt{1-2\rho}}{K\|c'(\hat{x})^{-1}\|}.$$

By Theorem 2.1, there is a solution in $X = \{x \in R^n : \|x - \hat{x}\| \le r\}$. Therefore, Theorem 2.1 is a generalization of the Kantorovich theorem [16] for the existence of the solution.

3. Spherical designs. In this section we describe a method of reformulating construction of spherical t-designs as an underdetermined system of nonlinear equations.

For a given positive integer t, a set of points $Y = \{y_1, \dots, y_{d_t}\} \subset S^2$ is called a fundamental system if the zero polynomial is the only member of $\mathbb{P}_t$ that vanishes at each point y_j, $j = 1, 2, \dots, d_t$. The requirement

$$d_t = (t+1)^2 = \dim \mathbb{P}_t$$

ensures that the basis matrix is square.

Y is called an extremal system if these points maximize the determinant of the interpolation matrix with respect to an arbitrary basis of $\mathbb{P}_t$. An extremal system is obviously a fundamental system. Sloan and Womersley [20] showed that the extremal fundamental systems have excellent geometrical properties and surprisingly good performance for numerical integration. However, it is unknown whether there is always a spherical t-design in a neighborhood of an extremal fundamental system. Our aim is to verify its existence.

Let $L_\ell : [-1, 1] \to R$ be the usual Legendre polynomial [2]. The Rodrigues representation yields

$$L_\ell(z) = \frac{1}{2^\ell} \sum_{k=0}^{[\ell/2]} \frac{(-1)^k (2\ell - 2k)!}{k!(\ell-k)!(\ell-2k)!} z^{\ell-2k}, \tag{3.1}$$

where $[\ell/2]$ is the floor function. Let

$$J_t(z) = \frac{1}{4\pi} \sum_{\ell=0}^{t} (2\ell+1) L_\ell(z), \qquad z \in [-1, 1],$$

which is a normalized Jacobi polynomial. The Gram matrix $G \equiv G(Y)$ is a symmetric positive semidefinite $d_t \times d_t$ matrix with elements

$$G_{i,j} = J_t(y_i^T y_j).$$

The functions

$$g_i(y) = J_t(y_i^T y), \ i = 1, \dots, d_t, \ y \in S^2,$$

belong to $\mathbb{P}_t$. If G is nonsingular, $\{g_1, \ldots, g_{d_t}\}$ is a basis for $\mathbb{P}_t$. For a given arbitrary function $f \in C(S^2)$, the unique polynomial interpolant Λf for the set Y is

$$(\Lambda f)(y) = \sum_{i=1}^{d_t} v_i g_i(y).$$

Here the vector of weights $v = (v_1, \ldots, v_{d_t})$ is the solution of the linear system of equations

$$Gv = b, \tag{3.2}$$

where $b_i = f(y_i), i = 1, 2 \ldots, d_t$.

The cubature rule

$$Q_{d_t}(f) = \sum_{i=1}^{d_t} w_i f(y_i) \approx \int_{S^2} f(y) dy$$

is exact for all polynomials p of degree $\leq t$ if w satisfies the system of linear equations

$$Gw = e, \tag{3.3}$$

where $e = (1, 1, \ldots, 1)^T \in R^{d_t}$. In particular, the cubature rule is exact for the constant polynomial $1 \in \mathbb{P}_t$. Thus

$$\int_{S^2} 1 \, dy = |S^2| = 4\pi = \sum_{i=1}^{d_t} w_i.$$

Hence the average cubature weight is

$$w_{\text{avg}} = \frac{4\pi}{d_t}.$$

Numerical results given in [22] show that the weights defined by (3.3) with the coefficient matrix $G(\bar{Y})$, where

$$\log \det G(\bar{Y}) = \max_{Y \subset S^2} \log \det G(Y), \tag{3.4}$$

are all positive and the scaled weights w_i / w_{avg} lie in $[1/2, 3/2]$.

The set of points $\bar{Y} = \{\bar{y}_1, \ldots, \bar{y}_{d_t}\}$ defined by (3.4) is an extremal fundamental system. It is conjectured that there is a spherical t-design which is very close to an extremal fundamental system; that is, there is a set of points $Y^* = \{y_1^*, y_2^*, \ldots, y_{d_t}^*\}$ in a neighborhood of $\bar{Y} = \{\bar{y}_1, \ldots, \bar{y}_{d_t}\}$ such that

$$\int_{S^2} p(y) dy = \sum_{i=1}^{d_t} w_i p(y_i^*) \qquad \forall p \in \mathbb{P}_t$$

and equal weights

$$w_i = \frac{4\pi}{d_t}, \;\; i = 1, 2, \ldots, d_t. \tag{3.5}$$

To explore this conjecture, we reformulate the problem as an underdetermined system of nonlinear equations. The matrix G is rotationally invariant, so the set of points can be normalized so that the first point is at the north pole and the second is on the prime meridian. Hence a spherical parametrization $\theta_j \in [0, \pi]$ and $\phi_j \in [0, 2\pi)$ of the points y_j, $j = 1, 2, \ldots, d_t$, has $\phi_1 = 0$, $\theta_1 = 0$, and $\phi_2 = 0$, giving a total of $2d_t - 3$ variables.

Let

$$n = 2d_t - 3, \qquad m = d_t - 1,$$

and let

$$\begin{aligned} x_{i-1} &= \theta_i, & i &= 2, 3, \ldots, d_t, \\ x_{d_t+i-3} &= \phi_i, & i &= 3, 4, \ldots, d_t. \end{aligned}$$

The set of points $Y = \{y_1, \ldots, y_{d_t}\}$ and the vector of variables $x \in R^n$ are uniquely related by

$$y_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad y_2 = \begin{bmatrix} \sin x_1 \\ 0 \\ \cos x_1 \end{bmatrix}, \quad y_i = \begin{bmatrix} \sin\theta_i \cos\phi_i \\ \sin\theta_i \sin\phi_i \\ \cos\theta_i \end{bmatrix} = \begin{bmatrix} \sin x_{i-1} \cos x_{d_t+i-3} \\ \sin x_{i-1} \sin x_{d_t+i-3} \\ \cos x_{i-1} \end{bmatrix}.$$

The simple bounds on θ_i and ϕ_i can be ignored due to the periodicity of the sin and cos functions. Hence the matrix G can be regarded as a function of x whose elements are defined by

$$G_{i,j}(x) = J_t(y_i^T y_j).$$

Define the function $c : R^n \to R^m$ by

$$c(x) = EG(x)e, \tag{3.6}$$

where E is the $m \times d_t$ matrix

$$E = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \ldots & 0 & -1 \end{pmatrix}.$$

This is motivated by the simple, but critical, observation that any cubature rule which is exact for constants has $\sum_{i=1}^{d_t} w_i = 4\pi$, so one only requires that $w_1 = w_i$ for $i = 2, \ldots, d_t$ to get (3.5). In fact the system of d_t equations $G(x)e - w_{avg}e = 0$ has a Jacobian with only rank $d_t - 1$.

The following theorem states the relation between a spherical t-design and a zero of the function c defined by (3.6).

THEOREM 3.1. *Suppose that $G(x^*)$ is nonsingular. Then x^* corresponds to a spherical t-design with $(t+1)^2$ points if and only if $c(x^*) = 0$.*

Proof. Let x^* be a solution of $c(x) = 0$, and let $\{y_1^*, y_2^*, \ldots, y_{d_t}^*\}$ be the set of points defined by x^*. First it is shown that $\{y_1^*, y_2^*, \ldots, y_{d_t}^*\}$ is a spherical t-design.

Since $G(x^*)$ is nonsingular, $\{y_1^*, y_2^*, \ldots, y_{d_t}^*\}$ is a fundamental system and the functions

$$g_j(y) = G({y_j^*}^T y), \quad j = 1, 2, \ldots, d_t,$$

form a basis of $\mathbb{P}_t$. Hence for any $p \in \mathbb{P}_t$ there are scalars $\alpha_j, j = 1, \ldots, d_t$, such that

$$p(y) = \sum_{j=1}^{d_t} \alpha_j g_j(y).$$

Note that (see [17] for an example)

$$\int_{S^2} g_j(y)dy = 1 \quad \forall\, j = 1, \ldots, d_t. \tag{3.7}$$

Moreover, $c(x^*) = 0$ implies that all components of $G(x^*)e$ are equal. Hence we can write

$$G(x^*)e = \mu e,$$

where μ is a scalar. Because of the nonsingularity of $G(x^*)$, $\mu \neq 0$. This yields

$$\int_{S^2} g_j(y)dy = 1 = \frac{1}{\mu} \sum_{k=1}^{d_t} G_{j,k}(x^*), \qquad j = 1, 2, \ldots, d_t.$$

We calculate the integral

$$\begin{aligned}
\int_{S^2} p(y)dy &= \sum_{j=1}^{d_t} \alpha_j \int_{S^2} g_j(y)dy \\
&= \frac{1}{\mu} \sum_{j=1}^{d_t} \alpha_j \sum_{k=1}^{d_t} G_{j,k}(x^*) \\
&= \frac{1}{\mu} \sum_{k=1}^{d_t} \sum_{j=1}^{d_t} \alpha_j G_{j,k}(x^*) \\
&= \frac{1}{\mu} \sum_{k=1}^{d_t} \sum_{j=1}^{d_t} \alpha_j g_j(y_k^*) \\
&= \frac{1}{\mu} \sum_{k=1}^{d_t} p(y_k^*).
\end{aligned}$$

In particular, for $p(y) \equiv 1$, the area of the sphere is

$$|S^2| = 4\pi = \int_{S^2} p(y)dy = \frac{1}{\mu} \sum_{k=1}^{d_t} p(y_k^*) = \frac{d_t}{\mu}.$$

Thus $\mu = d_t/4\pi$, and therefore $\{y_1^*, y_2^*, \ldots, y_{d_t}^*\}$ is a spherical t-design.

Now we prove that $c(x^*) = 0$ if x^* corresponds to a spherical t-design with $(t+1)^2$ points. By the definition of a spherical t-design, for any $p \in \mathbb{P}_t$,

$$\int_{S^2} p(y)dy = \frac{4\pi}{d_t} \sum_{k=1}^{d_t} p(y_k^*).$$

In particular, as $g_j \in \mathbb{P}_t$,

$$\int_{S^2} g_j(y)dy = \frac{4\pi}{d_t}\sum_{k=1}^{d_t} g_j(y_k^*), \qquad j = 1, 2, \ldots, d_t.$$

Hence, from the definition of g_j and (3.7), we find

$$\frac{4\pi}{d_t}\sum_{k=1}^{d_t} G_{j,k}(x^*) = \frac{4\pi}{d_t}\sum_{k=1}^{d_t} g_j(y_k^*) = 1.$$

This implies

$$G(x^*)e = \frac{d_t}{4\pi}e,$$

and thus

$$c(x^*) = EG(x^*)e = \frac{d_t}{4\pi}Ee = 0. \qquad \square$$

Let $\hat{x} \in R^n$ correspond to the set of points $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_{d_t}\}$ on the sphere. The condition for the cubature rule

$$Q_{d_t}(f) = \sum_{i=1}^{d_t} w_i f(\hat{y}_i)$$

to be exact for all polynomials in $\mathbb{P}_t$ is that $w = (w_1, \ldots, w_{d_t})^T$ is the solution of

$$G(\hat{x})w = e.$$

From Theorem 3.1, we know that $w = G(\hat{x})^{-1}e = (4\pi/d_t)e$ if and only if $c(\hat{x}) = 0$. The following theorem gives a result of the weights for the case $c(\hat{x}) \neq 0$.

THEOREM 3.2. *Suppose that $G(\hat{x})$ is nonsingular. Let $w = G(\hat{x})^{-1}e$. Then*

$$\max_{1\le i\le d_t} |w_1 - w_i| \le \frac{4}{\|G(\hat{x})e\|_\infty}\|G(\hat{x})^{-1}\|_\infty \|c(\hat{x})\|_\infty. \tag{3.8}$$

Proof. Let $\|\cdot\| = \|\cdot\|_\infty$ and let $|(G(\hat{x})e)_{i_0}| = \|G(\hat{x})e\|$. Then $\mu := (G(\hat{x})e)_{i_0} \neq 0$ and

$$\begin{aligned}\|\mu e - G(\hat{x})e\| &\le \|\mu e - (G(\hat{x})e)_1 e\| + \|(G(\hat{x})e)_1 e - G(\hat{x})e\| \\ &\le 2\|c(\hat{x})\|.\end{aligned}$$

Now, by the definition of the matrix E, we have

$$\begin{aligned}\max_{1\le i\le d_t} |w_1 - w_i| &= \|EG(\hat{x})^{-1}e\| \\ &= \|EG(\hat{x})^{-1}e - \frac{1}{\mu}Ee\| \\ &= \frac{1}{|\mu|}\|\mu EG(\hat{x})^{-1}e - EG(\hat{x})^{-1}G(\hat{x})e\| \\ &= \frac{1}{|\mu|}\|EG(\hat{x})^{-1}(\mu e - G(\hat{x})e)\| \\ &\le \frac{2}{|\mu|}\|E\|\|G(\hat{x})^{-1}\|\|c(\hat{x})\| \\ &= \frac{4}{|\mu|}\|G(\hat{x})^{-1}\|\|c(\hat{x})\|. \qquad \square\end{aligned}$$

4. Numerical verification of spherical *t*-designs. In this section, we use Theorems 2.1 and 3.1 to verify the existence of spherical t-designs. In particular, we use (2.4) to verify the existence of solutions to the system

$$c(x) := EG(x)e = 0. \tag{4.1}$$

Note that the highly nonlinear function $c(\cdot)$ is in $C^\infty(R^n)$ as long as the points are not at the south pole, which can easily be checked. (The first point is always the north pole and is not allowed to vary.) To save computational cost, let $x_{\mathcal{B}} = (x_1, \ldots, x_{d_t-1})^T$ and set $r_2 = 0$. Hence $c'_{\mathcal{B}}(x)$ is the first $(d_t - 1)$ columns of $c'(x)$ for $x \in X$, where

$$X = \{x \mid \|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \le r_1,\ x_{\mathcal{N}} = \hat{x}_{\mathcal{N}}\}.$$

The expansion (3.1) is used to calculate the derivatives of $c_i(x)$. Moreover, we can give an upper bound for the second derivatives. Since for $i, j = 1, \ldots, d_t$, $G_{ij}(x)$ are polynomials of degree t, the function

$$c_i(x) = (G(x)e)_1 - (G(x)e)_{i+1} = \frac{1}{4\pi}\sum_{j=1}^{d_t}\sum_{\ell=0}^{t}(2\ell+1)\left(L_\ell(y_1^T y_j) - L_\ell(y_{i+1}^T y_j)\right)$$

is polynomial of degree $\le t$. The first derivative of c_i is

$$\frac{\partial c_i(x)}{\partial x_k} = \frac{1}{4\pi}\sum_{j=1}^{d_t}\sum_{\ell=0}^{t}(2\ell+1)\left(L'_\ell(y_1^T y_j)\frac{\partial(y_1^T y_j)}{\partial x_k} - L'_\ell(y_{i+1}^T y_j)\frac{\partial(y_{i+1}^T y_j)}{\partial x_k}\right),$$

and the second derivative of c_i is

$$\frac{\partial^2 c_i(x)}{\partial x_k \partial x_\nu} = \frac{1}{4\pi}\sum_{j=1}^{d_t}\sum_{\ell=0}^{t}(2\ell+1)\Bigg(L''_\ell(y_1^T y_j)\frac{\partial(y_1^T y_j)}{\partial x_k}\frac{\partial(y_1^T y_j)}{\partial x_\nu} + L'_\ell(y_1^T y_j)\frac{\partial^2(y_1^T y_j)}{\partial x_k \partial x_\nu}$$

$$-L''_\ell(y_{i+1}^T y_j)\frac{\partial(y_{i+1}^T y_j)}{\partial x_k}\frac{\partial(y_{i+1}^T y_j)}{\partial x_\nu} - L'_\ell(y_{i+1}^T y_j)\frac{\partial^2(y_{i+1}^T y_j)}{\partial x_k \partial x_\nu}\Bigg).$$

Note that we consider only the first $(d_t - 1)$ columns of $c'(x)$ with respect to $x_{\mathcal{B}}$. Let

$$\nabla y_2 = \begin{bmatrix} \cos x_1 \\ 0 \\ -\sin x_1 \end{bmatrix}, \qquad \nabla y_i = \begin{bmatrix} \cos x_{i-1}\cos x_{d_t+i-3} \\ \cos x_{i-1}\sin x_{d_t+i-3} \\ -\sin x_{i-1} \end{bmatrix}.$$

For $k, \nu \le d_t - 1$, we have

$$\frac{\partial(y_1^T y_j)}{\partial x_k} = \begin{cases} y_1^T \nabla y_j & \text{if } k = j-1, \\ 0 & \text{otherwise;} \end{cases} \qquad \frac{\partial^2(y_1^T y_j)}{\partial x_k \partial x_\nu} = \begin{cases} -y_1^T y_j & \text{if } k = \nu = j-1, \\ 0 & \text{otherwise;} \end{cases}$$

$$\frac{\partial(y_{i+1}^T y_j)}{\partial x_k} = \begin{cases} y_{i+1}^T \nabla y_j & \text{if } k = j-1, \\ y_j^T \nabla y_{i+1} & \text{if } k = i, \\ 0 & \text{otherwise;} \end{cases} \qquad \frac{\partial^2(y_{i+1}^T y_j)}{\partial x_k \partial x_\nu} = \begin{cases} -y_{i+1}^T y_j & \text{if } k = \nu = j-1 \\ & \text{or } k = \nu = i, \\ \nabla y_i^T \nabla y_j & \text{if } k = j-1, \nu = i \\ & \text{or } k = i, \nu = j-1, \\ 0 & \text{otherwise.} \end{cases}$$

We use the relations $|y_i^T y_j| \le 1$ and $|\nabla y_i^T y_j| \le 1$ to give an upper bound K for the second derivatives of $c(\cdot)$ with respect to the first $d_t - 1$ variables. This, together with $x_{\mathcal{N}} = \hat{x}_{\mathcal{N}}$, implies

$$\|c'_{\mathcal{B}}(x) - c'_{\mathcal{B}}(\hat{x})\| \le K\|x - \hat{x}\|.$$

The infinity norm was used in the numerical implementation, so in the rest of this section $\|\cdot\|$ denotes $\|\cdot\|_\infty$.

The procedure for verifying the existence of a spherical t-designs is as follows:

1. Find an approximate solution $\hat{x}$ of $c(x) = 0$ starting from $\bar{x}$ corresponding to an extremal fundamental system $\bar{Y}$.
2. Calculate $c'_{\mathcal{B}}(\hat{x})$ and K.
3. Calculate

$$\rho = K\|c'_{\mathcal{B}}(\hat{x})^{-1}c(\hat{x})\|\ \|c'_{\mathcal{B}}(\hat{x})^{-1}\|. \tag{4.2}$$

If $\rho \le \frac{1}{2}$, then there is a solution of (4.1) in the set

$$X = \{x \in R^n : \|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \le r_1, x_{\mathcal{N}} = \hat{x}_{\mathcal{N}}\},$$

where

$$r_1 = \frac{1 - \sqrt{1 - 2\rho}}{K\|c'_{\mathcal{B}}(\hat{x})^{-1}\|}.$$

If $\rho > \frac{1}{2}$, then (4.1) has no solution in

$$X = \{x \in R^n : \|x_{\mathcal{B}} - \hat{x}_{\mathcal{B}}\| \le \gamma_1, x_{\mathcal{N}} = \hat{x}_{\mathcal{N}}\},$$

where

$$\gamma_1 = \frac{\sqrt{1 + 2\rho} - 1}{K\|c'_{\mathcal{B}}(\hat{x})^{-1}\|}.$$

Note that the natural residual $\|c(x)\|_2$ has many local minimizers. To find a good approximate solution of $c(x) = 0$, we choose several starting points around the extremal system and use the Gauss–Newton method with line search. The interest in starting from an extremal system stems from Figure 2 in [20] and Theorem 3.1. The cubature weights for the computed extremal system of [20] are very close to $4\pi/d_t$ and they maximize the determinant $G(x)$. Extremal systems can be downloaded from http://www.maths.unsw.edu.au/∼rsw/Sphere.

Numerical results are given in Table 1, where $\bar{x}$ is the vector corresponding to an extremal fundamental system $\bar{Y}$, $\hat{x}$ is an approximate solution of $c(x) = 0$,

$$\hat{w} = G(\hat{x})^{-1}e$$

is the weight for the cubature rule, and $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_{d_t}\}$ is the set of points corresponding to $\hat{x}$.

As the cubature rule is exact for the constant polynomial $1 \in \mathbb{P}_t$, the average weight is $\hat{w}_{avg} = 4\pi/d_t$. From the last column of Table 1, we see that all weights are positive and

$$\left|\hat{w}_i - \frac{4\pi}{d_t}\right| \le w_{\max} - w_{\min} \approx 0.$$

TABLE 1

Extremal points $\bar{x}$, computed spherical designs $\hat{x}$, exact spherical design x^*, $\|\hat{x} - x^*\| \le r_1$, $x \in R^{2(t+1)^2-3}$.

t	d_t	$\|c(\bar{x})\|$	$\|c(\hat{x})\|$	$\log \det G(\bar{x})$	$\log \det G(\hat{x})$	r_1	$\|\bar{x} - \hat{x}\|$	$\hat{w}_{\max} - \hat{w}_{\min}$
2	9	0.0245	4.44e-16	-3.2134	-3.2157	1.01e-15	0.0255	1.55e-15
3	16	0.4299	2.66e-15	3.3867	2.5779	2.36e-15	0.2742	1.88e-15
4	25	0.3898	7.32e-15	16.1396	15.9337	1.80e-14	0.1002	3.33e-15
5	36	0.6318	7.54e-15	36.1736	35.4829	1.34e-14	0.2595	2.10e-14
6	49	1.1376	2.62e-14	64.0948	62.6443	3.45e-14	0.1918	3.88e-15
7	64	0.9189	6.03e-14	100.6942	100.4167	5.07e-14	0.1277	4.10e-15
8	81	1.3713	1.92e-13	146.1926	144.3611	1.15e-13	0.2974	8.54e-15
9	100	1.4023	4.52e-13	201.5589	186.2265	1.84e-13	0.2526	7.88e-13
10	121	3.7879	8.07e-13	266.3178	265.5019	6.14e-11	0.0358	2.40e-14

Hence the set $\hat{Y}$ can be considered as computed spherical t-designs. These designs are new. Moreover, from Theorem 2.1 and $\|\hat{x} - x^*\| \le r_1$, an error bound for the computed spherical t-designs to an exact spherical design $\{y_1^*, \ldots, y_{d_t}^*\}$ corresponding to the exact solution x^* of $c(x) = 0$ is

$$\max_{1 \le i \le d_t} \|y_i^* - \hat{y}_i\| \le 2\|\hat{x} - x^*\| \le 2r_1,$$

where the first inequality uses the relation between x and y.

The numerical results also give an error bound for the extremal system

$$\begin{aligned} \max_{1 \le i \le d_t} \|y_i^* - \bar{y}_i\| &\le 2\|x^* - \bar{x}\| \\ &\le 2(\|x^* - \hat{x}\| + \|\hat{x} - \bar{x}\|) \\ &\le 2(r_1 + \|\bar{x} - \hat{x}\|). \end{aligned}$$

The interpolatory cubature rule

$$E_t(f) = \frac{4\pi}{d_t} \sum_{j=1}^{d_t} f(\hat{y}_j)$$

associated with $\hat{Y}$ provides high-order numerical integration on the sphere. In particular, by Theorem 4.1 in [20], the worst-case error in a particular Sobolev space is

$$\left| \int_{S^2} f(y) d(y) - E_t(f) \right| = 4\pi D(\hat{Y}) =: e(E_t),$$

where $D(\hat{Y})$ is the Cui–Freeden generalized discrepancy [4]

$$D(\hat{Y}) = \frac{1}{2\sqrt{\pi} d_t} \left[\sum_{j=1}^{d_t} \sum_{i=1}^{d_t} \left(1 - 2\log\left(1 + \sqrt{(1 - \hat{y}_i^T \hat{y}_j)/2} \right) \right) \right]^{1/2}.$$

Table 2 gives the values $D(\hat{Y})$ and $e(E_t)$. These values are better than the values reported by Sloan and Womersley [20]. The values given in [20] use extremal points and are better than the values reported by Cui and Freeden [4].

The computed spherical t-designs with $(t+1)^2$ points are available from http://www.st.hirosaki-u.ac.jp/∼chen/index.html. Computations for these low degrees were

TABLE 2
Worst case for the equal weight rule E_t and generalized discrepancy for computed spherical designs.

t	d_t	$e(E_t)$	$D(\hat{Y})$
2	9	0.349478	0.027811
3	16	0.229009	0.018239
4	25	0.162440	0.012927
5	36	0.123579	0.009834
6	49	0.098188	0.007814
7	64	0.079817	0.006352
8	81	0.067223	0.005349
9	100	0.058809	0.004680
10	121	0.049576	0.003945

performed by using MATLAB 6.1 on an IBM PC with 128MB memory and 500 MHz [15, 18].

Remark 4.1. This paper presents a new verification method for underdetermined systems of equations and uses this method to verify computed spherical t-designs. In comparison the Krawczyk-type interval operator method (1.3) failed for these underdetermined equations. This can be explained as follows.

Consider $\mathcal{K}(X)$ on an interval X which has an interior point $\hat{x}$. For any $x \in X$, $c'(x)$ is singular, and there is an x_b on the boundary of X such that $c'(x)(x_b - \hat{x}) = 0$. This implies that

$$x_b - c'(\hat{x})^+ c(\hat{x}) = \hat{x} - c'(\hat{x})^+ c(\hat{x}) + (I - c'(\hat{x})^+ c'(x))(x_b - \hat{x}) \in \mathcal{K}(X).$$

It is almost impossible to have $x_b - c'(\hat{x})^+ c(\hat{x}) \in X$ for all such boundary points x_b of X with $c'(\hat{x})^+ c(\hat{x}) \neq 0$. Hence $\mathcal{K}(X) \subseteq X$ always fails. On the other hand, the new verification method has no problems with the null space of $c'(x)$. The following example shows the advantage of the new method. Let

$$c(x) = 1 + x_1 + x_2 + x_1 x_2, \quad X = \frac{1}{4}\begin{pmatrix} [-5,-1] \\ [1+h, 3-h] \end{pmatrix}, \quad \hat{x} = \frac{1}{4}\begin{pmatrix} -3 \\ 2 \end{pmatrix},$$

where $h \in [0,1]$. Let $\mathcal{B} = \{1\}$ and $\mathcal{N} = \{2\}$. Straightforward calculation gives

$$c(\hat{x}) = \frac{3}{8}, \quad c'(x) = (1 + x_2, 1 + x_1), \quad c'(\hat{x}) = \frac{1}{4}(6,1), \quad c'_{\mathcal{B}}(\hat{x})^{-1} c(\hat{x}) = \frac{1}{4}.$$

It is easy to show that a Lipschitz constant for $c'_{\mathcal{B}}(x)$ is $K = 1$, and that

$$\max_{x \in N} \|c'_{\mathcal{N}}(x)\| = \frac{3}{4}.$$

Hence statement (1) of Theorem 2.1 holds with

$$\|c'_{\mathcal{B}}(\hat{x})^{-1} c(\hat{x})\| + \|c'_{\mathcal{B}}(\hat{x})^{-1}\| \left(\frac{1}{2} K(r_1 + r_2) r_1 + \max_{x \in N} \|c'_{\mathcal{N}}(x)\| r_2 \right) = \frac{1}{2} - \frac{h}{6} \leq r_1 = \frac{1}{2}$$

for all $h \in [0,1]$. Now we show that $\mathcal{K}(X) \subseteq X$ fails for all $h \in [0,1]$. Interval calculation gives

$$c'(\hat{x})^+ C'(X) = \frac{4}{37}\begin{pmatrix} 6 \\ 1 \end{pmatrix} \left(1 + \frac{1}{4}[1+h, 3-h],\ 1 + \frac{1}{4}[-5,-1] \right),$$

$$(I - c'(\hat{x})^+ C'(X))(X - \hat{x}) = \frac{1}{37 \times 4} \begin{pmatrix} [-80 + 30h,\ 80 - 30h] \\ [-52 + 40h,\ 52 - 40h] \end{pmatrix},$$

and the radii of X and $\mathcal{K}(X)$ satisfy

$$R(X) - R(\mathcal{K}(X)) = \frac{1}{4} \begin{pmatrix} 2 \\ 1 - h \end{pmatrix} - \frac{1}{148} \begin{pmatrix} 80 - 30h \\ 52 - 40h \end{pmatrix} = \frac{1}{148} \begin{pmatrix} -6 + 30h \\ -15 + 3h \end{pmatrix}.$$

Since the second component of the radii $R_2(X) - R_2(\mathcal{K}(X)) < 0$ for all $h \in [0, 1]$, we find that $\mathcal{K}(X) \not\subseteq X$ for all $h \in [0, 1]$.

Acknowledgment. We thank Prof. Andreas Frommer for his encouraging comments on Remark 4.1.

REFERENCES

[1] G. Alefeld, A. Gienger, and F. Potra, *Efficient numerical validation of solutions of nonlinear systems*, SIAM J. Numer. Anal., 31 (1994), pp. 252–260.
[2] G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, Cambridge, UK, 1999.
[3] X. Chen and T. Yamamoto, *Newton-like methods for solving underdetermined nonlinear equations with nondifferentiable terms*, J. Comput. Appl. Math., 55 (1994), pp. 311–324.
[4] J. Cui and W. Freeden, *Equidistribution on the sphere*, SIAM J. Sci. Comput., 18 (1997), pp. 595–609.
[5] P. Delsarte, J. M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geom. Dedicata, 6 (1977), pp. 363–388.
[6] J. M. Goethals and J. J. Seidel, *Spherical designs*, in Relations between Combinatorics and Other Parts of Mathematics, Proc. Sympos. Pure Math. 34, AMS, Providence, RI, 1979, pp. 255–272.
[7] R. H. Hardin and N. J. A. Sloane, *McLaren's improved snub cube and other new spherical designs in three dimensions*, Discrete Comput. Geom., 15 (1996), pp. 429–441.
[8] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[9] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer Academic Publishers, Norwell, MA, 1996.
[10] J. Korevaar and J. L. H. Meyers, *Spherical Faraday cage for the case of equal point charges and Chebyshev-type quadrature on the sphere*, Integral Transform. Spec. Funct., 1 (1993), pp. 105–117.
[11] U. Maier, *Numerical calculation of spherical designs*, in Advances in Multivariate Approximation, Math. Res. 107, W. Haubmann, K. Jetter, and M. Reimer, eds., Wiley-VCH, Berlin, 1999, pp. 213–226.
[12] J. M. Martínez, *Quasi-Newton methods for solving underdetermined nonlinear simultaneous equations*, J. Comput. Appl. Math., 34 (1991), pp. 171–190.
[13] A. D. McLaren, *Optimal numerical integration on a sphere*, Math. Comp., 17 (1963), pp. 361–383.
[14] K. H. Meyn, *Solution of underdetermined nonlinear equations by stationary iteration methods*, Numer. Math., 42 (1983), pp. 161–172.
[15] S. Oishi and S. M. Rump, *Fast verification of solutions of matrix equations*, Numer. Math., 90 (2002), pp. 755–773.
[16] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[17] M. Reimer, *Constructive Theory of Multivariate Functions*, Bibliographisches Institut, Mannheim, Germany, 1990.
[18] S. M. Rump, *INTLAB—INTerval LABoratory, a Matlab Toolbox for Verified Computations, Version* 3.1, 2002; available online from http://www.ti3.tu-harburg.de/rump/intlab/index.html.
[19] P. D. Seymour and T. Zaslavsky, *Averaging sets: A generalization of mean values and spherical designs*, Adv. Math., 52 (1984), pp. 213–240.
[20] I. H. Sloan and R. S. Womersley, *Extremal systems of points and numerical integration on the sphere*, Adv. Comput. Math., 21 (2004), pp. 102–125.

[21] H. F. Walker and L. T. Watson, *Least-change secant update methods for underdetermined systems*, SIAM J. Numer. Anal., 27 (1990), pp. 1227–1262.
[22] R. S. Womersley and I. H. Sloan, *How good can polynomial interpolation on the sphere be?*, Adv. Comput. Math., 14 (2001), pp. 195–226.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2342–2362

OPTIMIZING TALBOT'S CONTOURS FOR THE INVERSION OF THE LAPLACE TRANSFORM*

J. A. C. WEIDEMAN†

Abstract. Talbot's method for the numerical inversion of the Laplace transform consists of numerically integrating the Bromwich integral on a special contour by means of the trapezoidal or midpoint rules. In this paper we address the issue of parameter selection in the method, for the particular situation when parabolic PDEs are solved. In the process the well-known subgeometric convergence rate $O(\exp(-c\sqrt{N}))$ of this method is improved to the geometric rate $O(\exp(-cN))$, with N the number of nodes in the integration rule. The value of the maximum decay rate c is explicitly determined. Numerical results for two versions of the heat equation are presented. With the choice of parameters derived here, the rule of thumb is that to achieve an accuracy of $10^{-\ell}$ at any given time, the associated elliptic problem has to be solved no more than ℓ times.

Key words. Laplace transform, Talbot's method, trapezoidal rule, fractional differential equation

AMS subject classifications. 65D30, 65M70, 65R10

DOI. 10.1137/050625837

1. Introduction.

The Laplace transform is a classical technique for solving linear differential equations. For computational work, however, this approach never really became popular, as for many years numerical analysts tended to focus on discretization methods such as finite differences and finite elements, possibly combined with linear multistep or Runge–Kutta formulas for integration in time. We conjecture that this lack of interest shown by numerical analysts in the Laplace transform is partly due to the following two factors.

First, the Laplace transform restricts one to linear differential equations and in many applications one ultimately aims to solve nonlinear problems. Second, the Laplace transform, particularly its numerical inversion, has a reputation for being a computational challenge. This has to do with the fact that the inverse problem is by nature ill-conditioned when the transform is known only as a real-valued function. When the transform can be sampled in the complex plane the conditioning seems better, but then complex arithmetic is required.

Despite these apparent drawbacks of the Laplace transform, there has been a recent resurgence of the technique, as evidenced by the number of papers on this topic that have appeared since the year 2000; see, for example, [4, 8, 12, 14, 17, 18]. This renewed activity is in part due to recent interest in linear parabolic PDEs of fractional type, which are naturally posed in a transform setting. (These fractional PDEs model phenomena such as anomalous diffusion in several financial and biological applications.) In addition, MATLAB and other modern computational environments make complex arithmetic as easy to work with as real arithmetic and therefore complex inversion formulas become feasible.

*Received by the editors March 3, 2005; accepted for publication (in revised form) June 9, 2006; published electronically November 24, 2006. This work was supported by the National Research Foundation in South Africa under grant FA2005032300018.

http://www.siam.org/journals/sinum/44-6/62583.html

†Department of Applied Mathematics, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa (weideman@dip.sun.ac.za). This work was done at the Oxford University Computing Laboratory, as Visiting Fellow of Exeter College, Oxford, while the author was on sabbatical leave from the University of Stellenbosch, South Africa.

To introduce the problem, consider the linear system of ODEs

$$\frac{d\boldsymbol{f}}{dt} = A\boldsymbol{f}, \qquad \boldsymbol{f}(0) = \boldsymbol{f}_0, \tag{1.1}$$

where A is an $M \times M$ real matrix, $\boldsymbol{f}(t)$ an $M \times 1$ real vector, and $\boldsymbol{f}_0$ the initial condition. We are primarily interested in the case where A is the result of semi-discretization of a parabolic PDE (examples are given in section 5). We assume, therefore, that the eigenvalues of A are real and negative.

The formal solution to (1.1) is

$$\boldsymbol{f}(t) = \exp(At)\,\boldsymbol{f}_0,$$

and this reduces the problem to that of computing the matrix exponential of a (typically) large matrix. To be more precise, we need to compute the product of the matrix exponential and a vector, which can be done without actually computing the matrix exponential itself [15].

The authors of [4, 8, 12, 14, 17, 18] all compute this product by numerically approximating the inverse Laplace transform

$$\boldsymbol{f}(t) = \frac{1}{2\pi i}\int_\Gamma e^{zt}\boldsymbol{F}(z)\,dz, \qquad \boldsymbol{F}(z) \equiv (zI - A)^{-1}\boldsymbol{f}_0. \tag{1.2}$$

In this formula, known as the Bromwich integral, I is the $M \times M$ identity matrix and Γ is the contour of integration. At least initially, Γ is the Bromwich line $\operatorname{Re} z = \sigma$, where the parameter σ should be large enough that all eigenvalues of A lie in the half-plane $\operatorname{Re} z < \sigma$.

The typical approach is to deform the Bromwich line into a curve that begins and ends in the left half-plane, such that $\operatorname{Re} z \to -\infty$ on the contour; see Figure 1.1. Owing to the exponential factor e^{zt}, the integrand decays rapidly on such a contour, and if the contour is smooth this turns the problem into one of the classic situations where the trapezoidal rule converges extraordinarily rapidly [5, 10, 22, 23].

The articles [4, 8, 12, 14, 17, 18] differ with respect to the choice of the integration contour Γ, and how this contour is parameterized. A short summary of contours and convergence rates is given in section 6.

Suprisingly, none of the above references seriously considers Talbot's contour [19], rated in some circles as one of the best methods for inverting the Laplace transform; see [6]. (The method is mentioned in [12, 18], but is neither implemented nor analyzed there.) This contour may not be suitable when part of the spectrum of A is located off the negative real axis, but for pure parabolic problems the method is very accurate, as the numerical results of this paper will testify.

Talbot's contour is parameterized by

$$\Gamma: \quad z(\theta) = \sigma + \mu\,(\theta\cot\theta + \nu\,i\,\theta), \qquad -\pi \le \theta \le \pi, \tag{1.3}$$

where σ, μ, and ν are real parameters that determine the geometry of the curve. Both μ and ν are positive. For the eigenvalues of A to be enclosed by the contour one needs $z(0) > \lambda$, where λ is the largest eigenvalue of A, i.e.,

$$\sigma + \mu > \lambda. \tag{1.4}$$

A typical Talbot contour is shown in Figure 1.1.

A related contour is obtained by replacing the function $\theta\cot\theta$ in (1.3) with the first two terms in its partial fraction expansion,

$$\Gamma: \quad z(\theta) = \sigma + \mu\left(1 + \frac{2\theta^2}{\theta^2-\pi^2} + \nu\, i\,\theta\right), \qquad -\pi \le \theta \le \pi. \tag{1.5}$$

This contour is equivalent to one mentioned in Talbot's original paper [19], from which we quote: "... and indeed such functions can give good results, though their potentialities have not yet been explored."

It will turn out that the contour (1.5) is easier to analyze than (1.3), so for much of the paper we shall focus on the second contour. We shall also show, however, that the first Talbot contour yields superior accuracy.

Using either (1.3) or (1.5), the Bromwich integral (1.2) can be expressed as

$$\boldsymbol{f}(t) = \frac{1}{2\pi i}\int_{-\pi}^{\pi} e^{z(\theta)t}\boldsymbol{F}\big(z(\theta)\big)z'(\theta)\,d\theta, \tag{1.6}$$

where, respectively,

$$z'(\theta) = \mu\,(\cot\theta - \theta\csc^2\theta + \nu\, i) \quad \text{or} \quad z'(\theta) = \mu\Big(-\frac{4\pi^2\theta}{(\theta^2-\pi^2)^2} + \nu\, i\Big).$$

The integral (1.6) is typically approximated by the trapezoidal rule on a uniform partition of $[-\pi,\pi]$. Instead, we prefer to use the equally accurate midpoint rule with an even number of intervals, say $2N$. This is a practical choice that avoids sampling the integrand at the removable singularity at $\theta = 0$, as well as at the essential singularities at $\theta = \pm\pi$.

We hence define the grid

$$\theta_k = (2k+1)\frac{\pi}{2N}, \qquad k = -N, \ldots, N-1, \tag{1.7}$$

and denote the approximation to (1.6) by

$$\boldsymbol{f}_N(t) = \frac{1}{2Ni}\sum_{k=-N}^{N-1} e^{z(\theta_k)t}\, z'(\theta_k)\,\boldsymbol{F}_k,$$

or

$$\boldsymbol{f}_N(t) = \frac{1}{N}\,\mathrm{Im}\left\{\sum_{k=0}^{N-1} e^{z(\theta_k)t}\, z'(\theta_k)\,\boldsymbol{F}_k\right\}, \tag{1.8}$$

if symmetry is used. Here the vectors $\boldsymbol{F}_k \equiv \boldsymbol{F}(z(\theta_k))$ are solved from

$$\Big(z(\theta_k)\,I - A\Big)\boldsymbol{F}_k = \boldsymbol{f}_0, \qquad k = 0, \ldots, N-1. \tag{1.9}$$

Unless A is sparse, the solution of the N linear systems (1.9) represents the bulk of the computational cost of the algorithm. It should, however, be noted that the systems (1.9) can be solved independently and in parallel [10, 17]. In addition, it is possible to solve all N systems (1.9) efficiently using a single Hessenberg or Schur decomposition of A; see Problem P7.4.2 in [9, p. 350].

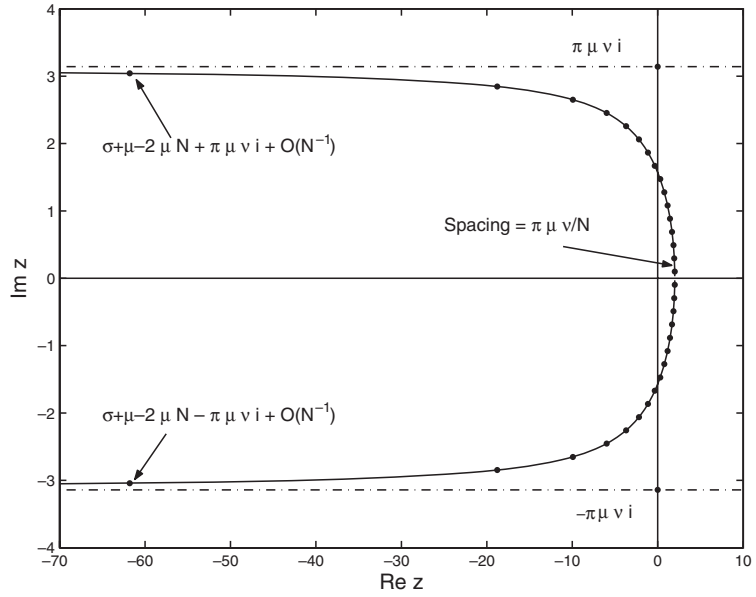


FIG. 1.1. *Talbot's contour* (1.3) *with parameter values* $\sigma = 0$, $\mu = 2$, $\nu = 0.5$. *The dots are the images in the z-plane of the midpoint abscissas* (1.7).

In this paper, we shall aim to optimize the convergence rate $\boldsymbol{f}_N(t) \to \boldsymbol{f}(t)$ as $N \to \infty$, keeping t fixed, by selecting the parameters (σ, μ, ν) in (1.3) and (1.5) to be asymptotically optimal. This is achieved by making σ and μ both proportional to the ratio N/t. By doing so, a geometric convergence rate, $O(e^{-cN})$ as $N \to \infty$, can be obtained. It is well known that Talbot's method with fixed (and therefore suboptimal) parameters converges at a subgeometric rate of $O(e^{-c\sqrt{N}})$; see [19].

To conclude this introduction, we offer Figure 1.1 as a summary of the role of the parameters (σ, μ, ν) in the contour (1.3). The parameter σ represents a shift to the left or right. The parameter μ controls the distance that the two extreme nodes extend into the left half-plane. The parameter ν determines the width of the contour in the sense that the contour approaches two horizontal asymptotes at distance proportional to $\mu\nu$ from the real axis as $\mathrm{Re}\, z \to -\infty$. The factor $\mu\nu$ also determines the spacing of the nodes near the real axis.

The outline of the paper is as follows. In section 2, we indicate how the well-known $O(e^{-c\sqrt{N}})$ convergence rate can be rederived by analyzing the scalar model problem

$$f(t) = e^{\lambda t}, \qquad F(z) = (z - \lambda)^{-1}. \tag{1.10}$$

This analysis suggests the reparameterization that we alluded to above, namely to make σ and μ both proportional to N/t. A saddle point method is used in section 3 to demonstrate that this rescaling of parameters leads to an improved convergence rate $O(e^{-cN})$. In section 4, we determine the value of ν and the proportionality constants in $\sigma \propto N/t$, $\mu \propto N/t$ that will maximize the decay rate, c. We hence obtain the attractive convergence rates $O(e^{-1.90N})$ and $O(e^{-1.73N})$, respectively, for the two versions of the Talbot contour (1.3) and (1.5). A further improvement, which involves omitting those outlying nodes on the Talbot contour that make a negligible

contribution to the midpoint sum, improves these two convergence rates to effectively $O(e^{-2.41N})$ and $O(e^{-2.56N})$. The theory of sections 3 and 4 is tested on two parabolic PDEs in section 5. In section 6, we discuss a few alternate contours, and we also contrast the parameter suggestions of this paper with those made by Talbot in [19].

2. Analysis of the scalar problem. We suppose the matrix A, which may or may not be symmetric, has real and negative eigenvalues, λ_j, corresponding to a complete set of eigenvectors, $\boldsymbol{v}_j$, $j = 1, \ldots, M$. If one expands the initial condition as a linear combination of eigenvectors,

$$\boldsymbol{f}_0 = c_1\boldsymbol{v}_1 + \cdots + c_M\boldsymbol{v}_M,$$

then (1.2) can be expressed as

$$\boldsymbol{f}(t) = \frac{c_1}{2\pi i}\Big(\int_\Gamma \frac{e^{zt}}{z-\lambda_1}\,dz\Big)\boldsymbol{v}_1 + \cdots + \frac{c_M}{2\pi i}\Big(\int_\Gamma \frac{e^{zt}}{z-\lambda_M}\,dz\Big)\boldsymbol{v}_M.$$

Applying Talbot's method to the right-hand side is therefore equivalent to applying it to the scalar problem (1.10), where λ represents a real and negative eigenvalue of A. For any fixed t, we shall restrict attention to λ in the range $|\lambda t| = O(1)$ as $M \to \infty$. We consider this sufficient because in the actual solution,

$$\boldsymbol{f}(t) = c_1 e^{\lambda_1 t}\boldsymbol{v}_1 + \cdots + c_M e^{\lambda_M t}\boldsymbol{v}_M,$$

modes that satisfy $|\lambda t| \gg 1$ are negligible.

Our task is therefore to estimate the error when approximating the integral

$$f(t) = \frac{1}{2\pi i}\int_{-\pi}^{\pi} \frac{e^{z(\theta)t}}{z(\theta)-\lambda} z'(\theta)\,d\theta \tag{2.1}$$

with the midpoint rule, with $z(\theta)$ defined by (1.3) or (1.5), and $\lambda < 0$. To start, we recall an error formula for the midpoint rule.

Consider an integral on $[-\pi, \pi]$ and its midpoint rule approximation

$$I(g) = \int_{-\pi}^{\pi} g(\theta)\,d\theta, \qquad M_N(g) = \frac{\pi}{N}\sum_{k=-N}^{N-1} g(\theta_k), \tag{2.2}$$

where the nodes θ_k are defined by (1.7). Suppose that the function $g(\theta)$ has an absolutely convergent Fourier series expansion

$$g(\theta) = \sum_{k=-\infty}^{\infty} c_k e^{ik\theta}, \qquad \text{with} \qquad c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\theta)e^{-ik\theta}\,d\theta.$$

Then it is possible to insert these formulas into (2.2), followed by termwise integration and summation, to obtain

$$I(g) = 2\pi c_0, \qquad M_N(g) = 2\pi c_0 + 2\pi \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} (-1)^\ell c_{2\ell N}.$$

The error is therefore given by

$$I(g) - M_N(g) = -2\pi \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} (-1)^\ell c_{2\ell N}. \tag{2.3}$$

(The trapezoidal rule error would be similar, except for the absence of the $(-1)^\ell$ factor; see [23].)

When the periodic extension of $g(\theta)$ is infinitely differentiable on $[-\pi, \pi]$, the Fourier coefficients c_k decay rapidly. In fact, repeated integration by parts can then be used to establish $c_k = O(|k|^{-m})$ for each positive integer m. In such cases a good error estimate can be obtained by retaining only the leading two terms in (2.3), as follows:

$$\begin{aligned} I(g) - M_N(g) &\sim 2\pi\Big(c_{-2N} + c_{2N}\Big) \\ &= \int_{-\pi}^{\pi} g(\theta) e^{-2Ni\theta}\, d\theta + \int_{-\pi}^{\pi} g(\theta) e^{+2Ni\theta}\, d\theta. \end{aligned}$$

One may apply this estimate to the special integral (2.1). The factor $e^{z(\theta)t}$ decays sufficiently rapidly as $\theta \to \pm\pi$ to ensure infinite differentiability of the periodic extension of the integrand. We therefore propose to analyze the error estimate,

$$f(t) - f_N(t) \sim E_N^-(t) + E_N^+(t), \qquad N \to \infty,$$

where, using symmetry,

$$E_N^\pm(t) = \frac{1}{\pi} \mathrm{Im} \left\{ \int_{-\pi}^{0} \frac{e^{z(\theta)t \pm 2iN\theta}}{z(\theta) - \lambda} z'(\theta)\, d\theta \right\}. \tag{2.4}$$

We shall keep both $t > 0$ and $\lambda < 0$ fixed, as well as the parameters σ, μ, and ν in the contours (1.3) or (1.5); our interest is the behavior of (2.4) as $N \to \infty$.

We digress for a moment to point out that error estimates such as (2.4) were used to good effect by Lin, to numerically predict optimal parameters for Talbot's contour [11]. A wide range of transforms was considered there, not just the $F(z) = 1/(z - \lambda)$ considered here.

Rather than using numerical optimization, we shall instead use the saddle point method to estimate analytically the two integrals (2.4). Since this analysis is primarily used to justify the form of the rescaling of parameters in section 3, and not in the determination of the actual optimal numbers itself, we omit the details. (A sketch of the derivation can be found in [24].) The result is that, with $E_N(t) \equiv E_N^-(t) + E_N^+(t)$,

$$E_N(t) = O\Big(e^{(\sigma+\mu)t - 2\sqrt{\pi\mu t N}}\Big), \qquad N \to \infty, \tag{2.5}$$

in the case of contour (1.3), and

$$E_N(t) = O\Big(e^{(\sigma+\frac{5}{2}\mu)t - 2\sqrt{\pi\mu t N}}\Big), \qquad N \to \infty, \tag{2.6}$$

in the case of (1.5). Results similar to these were obtained by Talbot [19, eq. (15)], who used a different method to prove that

$$E_N(t) = O\Big(N^2 e^{(\sigma+a\mu)t - b\sqrt{tN}}\Big).$$

The constants a and b depend on the transform and the contour.

The factor 5/2 that appears in (2.6) indicates that the error constant associated with the modified contour (1.5) is larger than that of the original contour (1.3). This was confirmed by the numerical experiments in [24].

The estimates (2.5)–(2.6) suggest the strategy of choosing $\sigma, \mu \propto N/t$. This should improve the subgeometric convergence rate, $O(e^{-c\sqrt{N}})$, to pure geometric convergence, $O(e^{-cN})$. We consider this next.

3. New parameters for the contour. Consider the rescaling

$$\sigma = -s\frac{N}{t}, \qquad \mu = m\frac{N}{t}, \qquad \nu = n, \tag{3.1}$$

where s, m, and n are real parameters to be determined. Both m and n are positive, and in accordance with (1.4) we require that

$$s < m - \frac{\lambda t}{N}. \tag{3.2}$$

The constant λ is defined in (1.10), which we continue to use as the model problem.

Because the parameters become dependent on t, so does the contour and hence also the integration nodes. This means that the N linear systems (1.9) have to be solved for each value of t, which may be inefficient. We therefore intend this rescaling to be used when the solution is required at only a few values of t.

Using the new parameters (3.1) we define $\zeta(\theta) = (t/N)\, z(\theta)$; i.e.,

$$\zeta(\theta) = -s + m\left(\theta\cot\theta + i\,n\,\theta\right) \tag{3.3}$$

in the case of the contour (1.3), and

$$\zeta(\theta) = -s + m\left(1 + \frac{2\theta^2}{\theta^2 - \pi^2} + i\,n\,\theta\right) \tag{3.4}$$

in the case of (1.5). The two error integrals (2.4) therefore become

$$E_N^{\pm}(t) = \frac{1}{\pi}\mathrm{Im}\left\{\int_{-\pi}^{0} \frac{e^{Ng_{\pm}(\theta)}}{\zeta(\theta) - \lambda t/N}\zeta'(\theta)\, d\theta\right\}, \tag{3.5}$$

where

$$g_{\pm}(\theta) = \zeta(\theta) \pm 2\,i\,\theta.$$

We apply the saddle point method to (3.5). (For details of this method, we refer the reader to [1, sect. 6.4; 3, sect. 6.6].) The idea is to deform the interval of integration, $[-\pi, 0]$, to a special contour in the complex θ-plane on which the integral can be estimated accurately. By Cauchy's theorem such a deformed contour will be permissible as long as it starts at $\theta = -\pi$, terminates at $\theta = 0$, and does not cross any singularities of the integrand in between. Suitable contours are steepest descent curves, defined by $\mathrm{Im}\{g_{\pm}(\theta)\} = \text{constant}$, for these remove the oscillations from the integrands in (3.5). The constants are chosen such that the contours pass through the saddle points, $\theta = \theta_+$ and $\theta = \theta_-$, respectively, defined by

$$g'_+(\theta) = 0, \qquad g'_-(\theta) = 0. \tag{3.6}$$

To ensure analyticity of the integrand, one needs to take into consideration the singularities associated with the vanishing of the denominator in (3.5), i.e., the zeros of $\zeta(\theta) = \lambda t/N$. In view of the discussion in the first paragraph of section 2, we shall assume $|\lambda t| \ll N$ and ignore the right-hand side of this equation. We therefore define the critical points, $\theta = \theta_*$, as the zeros of

$$\zeta(\theta) = 0. \tag{3.7}$$

This is the same as setting $\lambda = 0$, and in accordance with (3.2) we shall therefore consider only $m > s$.

To apply the saddle point method, we have to know where the critical points and saddle points are. We restrict ourselves to the modified contour (1.5); i.e., we assume $\zeta(\theta)$ is defined by (3.4). For this contour (3.6) and (3.7) reduce to polynomial equations of degrees 4 and 3, respectively, which can in principle be solved explicitly. The results are unwieldy, however, and we follow a more elementary approach.

In order to work with equations with real coefficients, we introduce the variable ϕ by $\theta = i\phi$. After denominators have been cleared, (3.6) can be factored into

$$\big(mn \pm 2\big)\big(\phi^2 + \pi^2\big)^2 = 4m\pi^2\phi, \tag{3.8}$$

and (3.7) into

$$\big(mn\phi + s - m\big)\big(\phi^2 + \pi^2\big) = 2m\phi^2. \tag{3.9}$$

Assuming $m > s \geq 0$ and working with these two representations, we were able to establish the following properties of the roots of (3.6)–(3.7).

Starting with the cubic equation (3.9), it is readily established that it always has a positive real root. The remaining two roots may either be real as well or occur as a conjugate pair. In the latter case the real part of these roots is positive, and the imaginary part is bounded in absolute value by π. Transplanting this information from the ϕ variable to θ, we deduce that the three critical points θ_* defined by (3.7) are all in the upper half-plane, with real parts in the interval $(-\pi, \pi)$. At least one root lies on the positive imaginary axis. The other two roots may be pure imaginary as well, or they may be located symmetrically with respect to the imaginary axis. (In the next section, we shall conjecture that the optimal configuration occurs when this pair of roots coalesces into a double root on the imaginary axis.)

Turning to (3.8), one notices that it is a quartic equation with real coefficients that is missing its cubic term. The typical configuration of roots is therefore one in each quadrant of the complex θ-plane, at equal distances from the real axis. The exception is when (3.8) admits four real roots, which would mean saddle points on the imaginary θ-axis. Considering the minus sign in (3.8) we see this cannot happen when $mn < 2$, and because of (3.10) below, we disregard this possibility.

Figure 3.1 shows a typical configuration of critical and saddle points. The roots of (3.6) are represented by $+$'s and $\times$'s (corresponding to the $+$ and $-$ signs, respectively), and the roots of (3.7) are plotted as the $*$'s.

We propose a saddle point analysis based on the contours shown in the figure. The $\Gamma_\pm$ are the curves of steepest descent $\mathrm{Im}\big\{g_\pm(\theta)\big\} = \text{constant}$. Writing $\theta = x + yi$, they can be expressed as

$$\Gamma_\pm : \quad \frac{4m\pi^2 xy}{(x^2 - y^2 - \pi^2)^2 + 4x^2y^2} - (mn \pm 2)x = c_\pm.$$

The constants $c_\pm$ are determined by the requirement that each $\Gamma_\pm$ passes through its corresponding saddle point, θ_+ or θ_-, as defined by (3.6).

In the lower half-plane, Γ_- starts at $\theta = -\pi$, passes through θ_-, and continues to $\theta = -i\infty$. This is valid since the integrand in (3.5), with minus sign, approaches zero as $\theta \to -i\infty$, provided that

$$mn < 2. \tag{3.10}$$

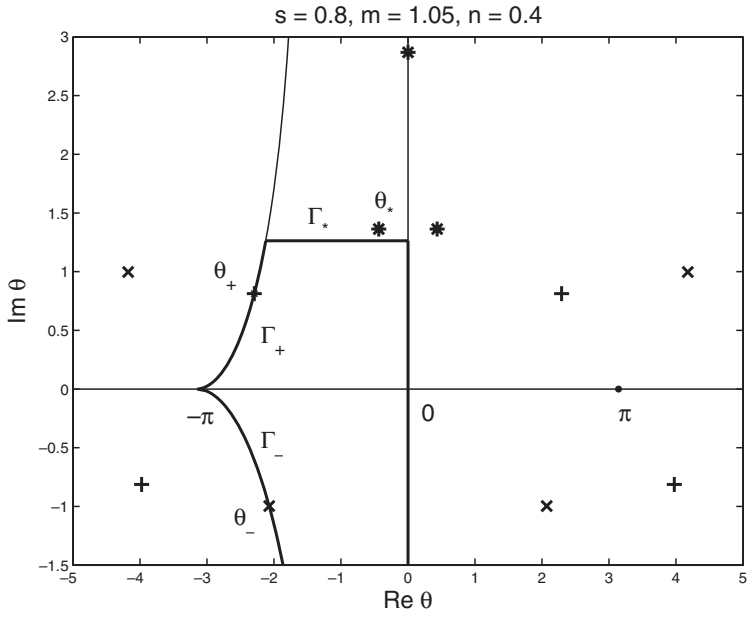


FIG. 3.1. *Saddle points, θ_- and θ_+, critical points θ_*, and steepest descent contours used in deriving the error estimates* (3.11)–(3.13).

(The corresponding restriction for the contour (1.3) is $m(1+n) < 2$.) The contour is then closed at $-i\infty$ and returns to the origin via the negative imaginary θ-axis. On this axis the contribution can be ignored, since the integrand is real. The error $E_N^-(t)$ is therefore solely determined by the saddle point contribution, which can be computed in the usual manner as [1, sect. 6.4; 3, sect. 6.6]

$$E_N^-(t) = O\big(e^{d_- N}\big), \qquad d_- = \mathrm{Re}\big\{g_-(\theta_-)\big\}. \tag{3.11}$$

In the upper half-plane a similar approach is used, except for the fact that the critical points θ_* have to be taken into account. The contour Γ_+ is not continued to $\theta = +i\infty$, as it will not be possible to return to the origin without crossing the singular points $\theta = \theta_*$. To maintain analyticity of the integrand, we introduce a third contour, Γ_*, that branches off from Γ_+ and has a constant imaginary part, say $\mathrm{Im}\{\Gamma_*\} = b$. Typically, the value of b would be determined by the critical point θ_* nearest to the real axis. By letting Γ_* approach such a limiting θ_* from below, it is possible to establish

$$E_N^+(t) = O\big(e^{d_* N}\big), \qquad d_* = \mathrm{Re}\big\{g_+(\theta_*)\big\}. \tag{3.12}$$

If θ_+ lies below Γ_*, a saddle point contribution similar to (3.11),

$$E_N^+(t) = O\big(e^{d_+ N}\big), \qquad d_+ = \mathrm{Re}\big\{g_+(\theta_+)\big\}, \tag{3.13}$$

is to be added to (3.12).

In our numerical experiments, the total error was dominated either by (3.11) or by (3.12). We have not found a set of parameters (s, m, n) for which (3.13) dominates, but neither have we tried to prove that this is impossible.

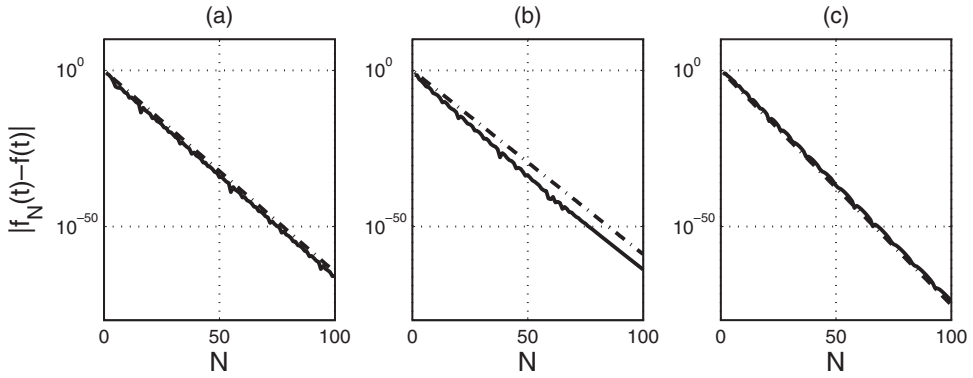


FIG. 3.2. *Convergence curves for Talbot's method applied to the model problem* (1.10), *with* $\lambda = -1$, $t = 1$, *and using contour* (1.5). *The solid curves represent the actual errors, and the dash-dot lines* (*virtually indistinguishable in* (a) *and* (c)) *their theoretical estimates. The plots* (a)–(c) *correspond to three different choices of* (s, m, n), *as summarized in the appendix.*

In Figure 3.2, we offer numerical verification of these error estimates. In the appendix, the corresponding values of saddle points, critical points, and expected convergence rates are summarized. We have picked sets of parameter values (s, m, n) for which (a) the saddle point contribution (3.11) dominates, (b) the critical point contribution (3.12) dominates, and (c) these two contributions are equal (the conjectured optimal situation). Also shown, as the dash-dot curves, are the predicted convergence rates, i.e., the maximum of (3.11) and (3.12). Here we should point out that these estimates are asymptotic, and much information is suppressed by the order notation of (3.11)–(3.13). Therefore, in some cases N has to be large for the estimate to become valid. This can be seen in part (b) of Figure 3.2, for example, where N has to be greater than 70, roughly, before (3.12) becomes evident.

In Figure 3.2, and elsewhere in the paper where multiprecision arithmetic was required, we computed in Maple and exported the numbers to MATLAB for plotting.

4. Computing the optimal parameters. A first attempt at finding optimal parameters (s, m, n) was based on a numerical optimization strategy, involving the objective function

$$F(s, m, n) \equiv \max\left\{d_+, d_-, d_*\right\} = \text{minimum}. \tag{4.1}$$

Here d_+, d_-, d_* are the decay constants in the error estimates (3.13), (3.11), and (3.12). For each set of parameters (s, m, n), the value of F can be computed by first solving (3.6) and (3.7) to obtain θ_+, θ_-, and θ_*. These values are then substituted into $g_\pm(\theta)$, to compute d_+, d_-, d_* as defined by (3.11)–(3.13).

In the case of contour (1.5), (3.6) and (3.7) can be solved with polynomial rootfinding routines, and in this case MATLAB's function `roots` was used. In the case of contour (1.3) a complex Newton process was used. When solving (3.7), one should take care to select the correct root θ_*.

The problem (4.1) was solved using MATLAB's function `fminsearch`, a routine suitable for nonsmooth, unconstrained optimization. Aside from some mild ill-conditioning that will be explained below, this approach worked well.

In the case of contour (1.5), this yielded the parameters presented as case (c) in Figure 3.2, namely

$$s = 0.7556, \quad m = 0.8597, \quad n = 0.3029. \tag{4.2}$$

The corresponding saddle and critical points are summarized in the appendix. The predicted optimal convergence rate is

$$E_N(t) = O\big(e^{-1.7303N}\big), \qquad N \to \infty. \tag{4.3}$$

Applying the same algorithm to the original Talbot contour (1.3), we obtained a better convergence rate, namely

$$E_N(t) = O\big(e^{-1.8975N}\big), \qquad N \to \infty. \tag{4.4}$$

This corresponds to parameter values

$$s = 0.4814, \quad m = 0.6443, \quad n = 0.5653, \tag{4.5}$$

with saddle points and critical point given by

$$\theta_+ = -2.5293 + 0.7435\,i, \quad \theta_- = -2.4158 - 0.9487\,i, \quad \theta_* = 0.9487\,i, \tag{4.6}$$

and decay rates

$$d_+ = -2.5048, \quad d_- = -1.8975, \quad d_* = -1.8975.$$

In Figure 4.1, we show the θ_+, θ_-, and θ_* defined by (4.6) in the top figure, and their images in the z-plane in the bottom figure. Also shown are the nodes of the midpoint approximation, with $N = 16$.

Examining the numerical results (4.5)–(4.6), we conjecture that in the optimal configuration,

(a) θ_* is on the positive imaginary axis,
(b) $\zeta'(\theta_*) = 0$,
(c) $\mathrm{Im}(\theta_*) = -\mathrm{Im}(\theta_-)$, and
(d) $d_+ < d_- = d_*$

for both contours (1.3) and (1.5). All of these properties seem plausible, but we have not pursued rigorous proofs.

Property (b) indicates that θ_* is a double root of (3.7), which is the source of the ill-conditioning mentioned at the beginning of the section. Fortunately, assuming properties (a)–(d) to be true, the problem can be reformulated such that it becomes explicitly solvable. The details are as follows.

Using properties (a) and (c) above, we write

$$\theta_* = y\,i, \qquad \theta_- = x - y\,i,$$

where $x < 0$ and $y > 0$. Because of property (d), we shall ignore θ_+ and try to solve for x, y, s, m, and n from the following five (real) equations: the right-hand equality in (3.6) (two real equations), (3.7), property (b), and the equality in (d).

Using a straightforward but tedious hand calculation this 5×5 system was reduced to a 2×2 system involving x and y. In the case of contour (1.5) this system is

$$\begin{aligned} x(P^2 - Q^2) + 2yPQ &= 0, \\ (5y^2 + \pi^2)(P^2 + Q^2) + P(y^2 + \pi^2)^2 &= 0, \end{aligned} \tag{4.7}$$

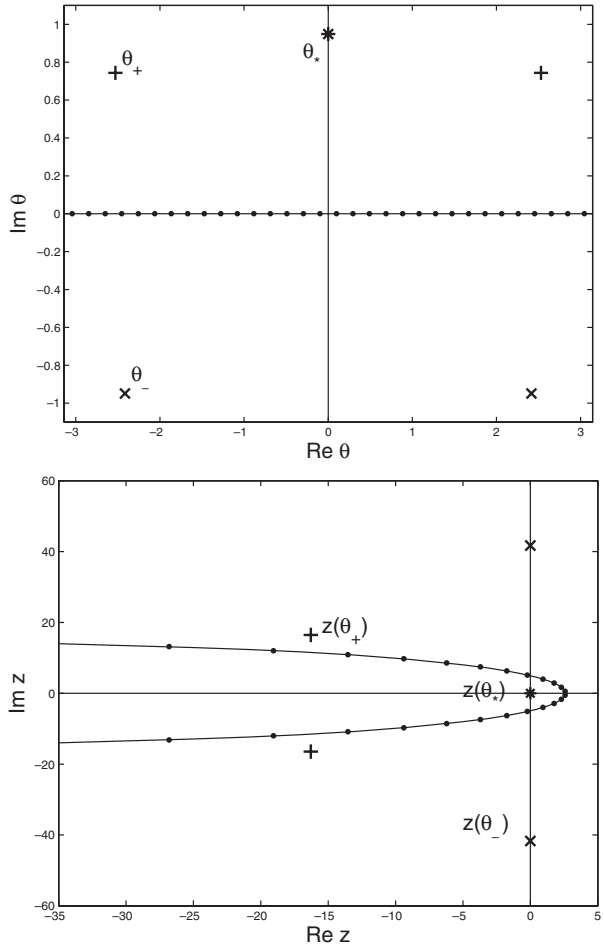


FIG. 4.1. *The optimal configuration of saddle points and critical points in the θ-plane (top) as well as their images in the z-plane (bottom). The dots are the nodes used in the midpoint rule, with* $N = 16$. *Note that in the bottom figure, four sets of nodes lie offscale (towards* $\mathrm{Re}\, z = -\infty$*). The contour is given by* (1.3), *but the figure is qualitatively similar for* (1.5).

where

$$P = x^2 - y^2 - \pi^2, \qquad Q = 2xy.$$

For further simplification we turned to Maple, which produced an explicit solution

$$y = \pi\sqrt{v},$$

where $v \approx 0.07584$ is the smallest positive root of

$$41v^4 - 308v^3 - 98v^2 - 4v + 1 = 0.$$

With this value of v, x is given by

$$x = -\frac{\pi}{4\sqrt{2}}\sqrt{41 - 209v - 1581v^2 + 205v^3}.$$

These formulas yield the values $y \approx 0.8652$ and $x \approx -2.2315$, as obtained above. The values of s, m, and n given in (4.2) follow from

$$n = \frac{4\pi^2 y}{(y^2+\pi^2)^2}, \quad m = \frac{2(R^2+S^2)}{4\pi^2(yR-xS)+n(R^2+S^2)}, \quad s = m\left(\frac{3y^4+\pi^4}{(y^2+\pi^2)^2}\right),$$

where

$$R = P^2 - Q^2, \qquad S = 2PQ.$$

In the case of contour (1.3) the analogue of the system (4.7) is

$$A - x\left(A^2 - B^2 + 1\right) - 2yAB = 0,$$
$$xA + yB + y\left(\coth y - 2\,y\,\mathrm{csch}^2 y\right) = 0,$$

where we have defined $A = \mathrm{Re}\{\cot\theta_-\}$, $B = \mathrm{Im}\{\cot\theta_-\}$, i.e.,

$$A = \frac{\sin x \cos x}{\sin^2 x + \sinh^2 y}, \qquad B = \frac{\sinh y \cosh y}{\sin^2 x + \sinh^2 y}.$$

A numerical solution of this system yields the value of $\theta_- = x - iy$ reported in (4.6). The values of the other parameters can be computed via

$$n = \coth y - y\,\mathrm{csch}^2 y, \quad m = \frac{2}{B + y(A^2 - B^2 + 1) - 2xAB + n}, \quad s = my^2\,\mathrm{csch}^2 y.$$

As verification that the parameters derived here are indeed close to optimal, we offer Figure 4.2. There we show, as the thicker curve, the numerically computed error $E_N(t)$ as a function of N, corresponding to parameters (4.5). Virtually on top of this curve and shown as a dash-dot line is the theoretical error estimate (4.4). To show the near-optimality of these curves, we have computed errors using a uniform sampling of parameter space $(s,m,n) \in (0,1)\times(0,1)\times(0,1)$, with step-size 0.05 in each direction. (That is, $19^3 = 6859$ different parameter sets were used for each value of N.) The vertical line segments in the figure represent the range of these computed errors, with the minima and maxima indicated by the tiny horizontal bars.

We should not neglect to point out that if our sampling of parameter space were finer, some of the lower error bars in this figure could extend further down to 0. This will happen when the two error components, $E_N^+(t)$ and $E_N^-(t)$ in (2.4), are approximately of equal magnitude but of opposite sign. Such instances of fortuitous cancellation will, however, be rare when the matrix as opposed to the scalar problem is solved. We believe that Figure 4.2 represents solid evidence that the suggested parameter values in (4.5) and (4.2) are indeed asymptotically optimal.

To conclude this section, we point out a redundancy in the Talbot contour as noted by Trefethen [21]. Recall Figure 4.1, where the optimal Talbot contour was shown for the case $N = 16$, and recall also that four pairs of nodes were located outside the frame of the figure, towards $\mathrm{Re}\,z = -\infty$. In fact, the contribution of each of these outlying nodes is negligible, as $|e^{z(\theta_k)t}| \le e^{-1.90N}$ when $|k| \ge 3N/4$. It appears that practically no accuracy is lost by including only the middle 75% of nodes and discarding the outlying 25%. A more careful calculation shows that the actual fraction of nodes retained should be about 0.7409. Since $1.8975/0.7409 \approx 2.5611$, the effective convergence rate improves from about $O(e^{-1.90N})$ to $O(e^{-2.56N})$. In the case of the modified Talbot contour (1.5) about 28% of the nodes can be discarded, which increases the effective rate from $O(e^{-1.73N})$ to roughly $O(e^{-2.41N})$.

In the next section we solve two parabolic problems to test some of these convergence estimates.

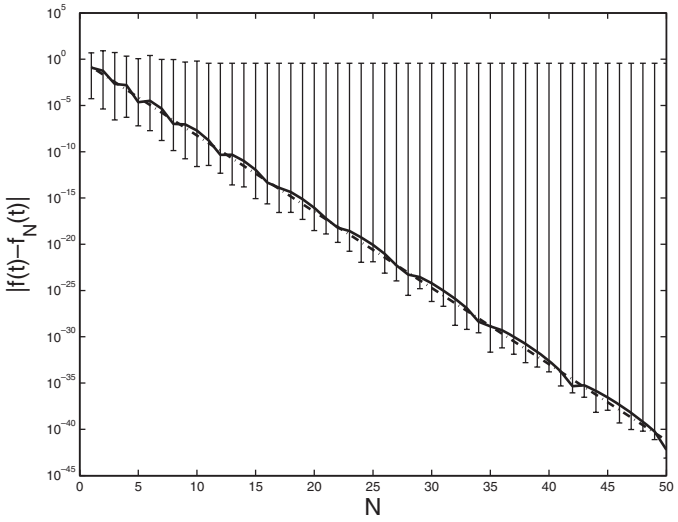


FIG. 4.2. *Absolute errors when Talbot's method is applied to the model function* (1.10), *with* $t = 1, \lambda = -1$, *using the original contour* (1.3). *(The figure is qualitatively similar for the modified contour* (1.5).*) The thicker curve represents the errors computed with the optimal parameters* (4.5). *The dash-dot line (hardly visible) represents the theoretical error estimate* (4.4). *The vertical line segments represent the range of errors computed in a uniform sampling of parameter space.*

5. Application to PDEs. The prototype parabolic PDE is the heat equation

$$u_t = u_{xx}, \qquad 0 \le x \le \pi, \tag{5.1}$$

and here we consider boundary conditions

$$u(0,t) = 0, \qquad u(\pi,t) = 1, \qquad t > 0, \tag{5.2}$$

and an initial condition

$$u(x,0) = 0, \qquad 0 \le x \le \pi. \tag{5.3}$$

The exact solution can be represented either as a Fourier series [2, p. 91], or an infinite series involving the complementary error function [2, p. 93] (efficient for large and small t, respectively).

For numerical work, we let $v(x,t) = u(x,t) - x/\pi$ and rewrite the PDE as

$$v_t = v_{xx}, \tag{5.4}$$

now with homogeneous boundary conditions

$$v(0,t) = 0, \qquad v(\pi,t) = 0, \qquad t > 0, \tag{5.5}$$

but inhomogeneous initial condition

$$v(x,0) = -x/\pi, \qquad 0 \le x \le \pi.$$

To semidiscretize (5.4), a suitable partition $\{x_j\}_{j=1}^{M}$ of $[0,\pi]$ is introduced, along with an $M \times M$ matrix D that represents the approximation to d^2/dx^2 and which

incorporates the boundary conditions (5.5). The approximation to (5.1) is then given by the linear system of ODEs

$$\boldsymbol{v}_t = D\boldsymbol{v}, \qquad \boldsymbol{v}(0) = \boldsymbol{v}_0. \tag{5.6}$$

Here $\boldsymbol{v} = \boldsymbol{v}(t)$ is the $M \times 1$ column vector $[v_1(t), v_2(t), \ldots, v_M(t)]^T$, with $v_j(t)$ representing the approximation to $v(x_j, t)$. Likewise $\boldsymbol{v}_0$ is the vector consisting of samples of $v(x, 0)$ at the grid-points x_j.

Traditionally, the system (5.6) is integrated by a Runge–Kutta or multistep formula (the method of lines). Here we use the transform approach instead. That is, we compute the midpoint sum

$$\boldsymbol{v}_N(t) = \frac{1}{N} \operatorname{Im} \left\{ \sum_{k=0}^{N-1} e^{z(\theta_k)t}\, z'(\theta_k)\, \boldsymbol{F}_k \right\}, \tag{5.7}$$

where $z(\theta)$ is given by (1.3) and θ_k by (1.7). The vectors $\boldsymbol{F}_k$ are solved from

$$\Big(z(\theta_k)\, I - D\Big)\boldsymbol{F}_k = \boldsymbol{v}_0, \qquad k = 0, \ldots, N-1. \tag{5.8}$$

The details of our particular implementation are as follows. Since we have established that the Talbot contour (1.3) is superior to the contour (1.5), we consider only the former. As for the choice of $\{x_j\}_{j=1}^M$ and D, we use the Chebyshev spectral collocation method; i.e., the nodes are the Chebyshev points of the second kind, and D is the corresponding spectral second derivative matrix incorporating the boundary conditions (5.2). The canonical interval for the Chebyshev points is $[-1, 1]$, which we transform to $[0, \pi]$ with $x \mapsto (\pi/2)(x + 1)$. (Codes for computing $\{x_j\}_{j=1}^M$ and D and further details of the spectral method can be found in [7, 20, 25].)

We shall report errors in the L_2-norm, as approximated by the Clenshaw–Curtis rule (the natural quadrature rule for the Chebyshev method). That is, we define as error norm

$$E_N(t) = \sqrt{\frac{\pi}{2} \sum_{j=1}^{M} w_j \Big(v(x_j, t) - v_j(t)\Big)^2}, \tag{5.9}$$

where the w_j are the weights defined in [20, p. 128], and the factor $\pi/2$ comes from the transformation of $[-1, 1]$ to $[0, \pi]$. The exact solution, $v(x, t)$, was computed by the series expansions mentioned below (5.3).

Our first aim is to demonstrate that the convergence estimate (4.4), derived for the model problem (2.1), is also valid for the solution of a PDE. In the latter case, there is of course a spectrum of λ's present, not only the single λ that was assumed in sections 2 and 3. For this reason we chose the side conditions (5.2)–(5.3) to represent a discontinuous solution at $t = 0$. Our interest will therefore be in the regime $t \to 0$, when high frequency modes are relevant.

In Figure 5.1 we show solutions of (5.1)–(5.3) at various values of t. We also show the error, $E_N(t)$, as a function of N, for the corresponding values of t. We have chosen the $M \times M$ Chebyshev matrices D sufficiently large to fully resolve the solution; i.e., the errors reported in the figure are solely due to the Laplace transform quadrature error and not due to inadequate spatial resolution. Owing to the smoothing property of the heat equation the order of D can of course be reduced as t increases, and suitably large values of M were determined by trial and error.

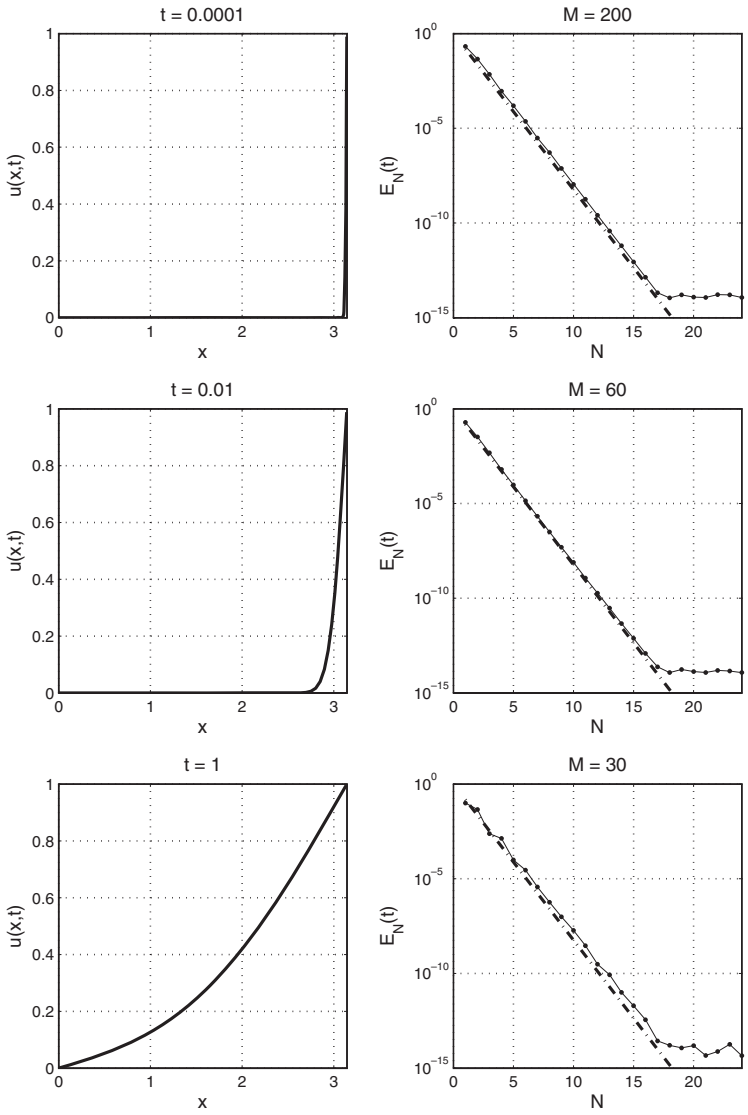


FIG. 5.1. *The left column shows the actual solution of* (5.1)–(5.3) *at various times. The right column shows convergence curves when the solution on the left is approximated with a Chebyshev spectral differentiation matrix of order* $M \times M$ *and Talbot's quadrature rule* (5.7) *using the contour* (1.3) *with optimal parameters* (4.5). *The thinner, dotted curves show the computed errors, and the thicker, dash-dot lines the error model* $\exp(-1.90N)$*; cf.* (4.4). *The error* $E_N(t)$ *is defined by* (5.9).

Assessing these figures, it is clear that the error estimate (4.4) is valid for this problem, even for small t. In addition, one should keep in mind that these results can be achieved by solving effectively only $0.74N$ linear systems (recall the discussion at the end of section 4). This allows us to formulate the rule of thumb stated in the abstract. Suppose an accuracy of $10^{-\ell}$ is required at a particular value of t. By considering

$$e^{-2.56N} = 10^{-\ell} \qquad \Longrightarrow \qquad N \approx 0.9\ell$$

one concludes that this should require no more than ℓ solutions of the system (5.8).

We remark that discretization in space is, strictly speaking, not necessary for (5.1)–(5.3), as the Laplace transform can be obtained explicitly, as follows [2, p. 89]:

$$F(z) = \frac{\sinh(x\sqrt{z})}{z\sinh(\pi\sqrt{z})}.$$

Talbot's method (or any other inversion algorithm) may be applied to this transform for any x in $[0, \pi]$. For problems with nonconstant coefficients or with complicated boundary conditions, however, such explicit representations may not exist.

As a second example, we consider the fractional heat equation

$$D_t^\alpha u = u_{xx}, \tag{5.10}$$

subject to boundary conditions (5.5) and initial condition

$$u(x,0) = \sin\, x, \qquad 0 \le x \le \pi. \tag{5.11}$$

Here D_t^α is the Caputo fractional derivative, defined by [16, p. 79]

$$D_t^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{f'(s)}{(t-s)^\alpha}\, ds \qquad (0 < \alpha < 1).$$

It can be shown [16, p. 79] that if $f(t)$ is twice continuously differentiable, then in the limit $\alpha \to 1$ this formula reproduces the ordinary derivative, in which case (5.10) reduces to the standard heat equation (5.1).

The analytical solution to (5.10)–(5.11) can be written as

$$u(x,t) = M(t)\sin x,$$

where $M(t)$ can be expressed in terms of the Mittag–Leffler function. In the case $\alpha \to 1$, it reduces to $M(t) = e^{-t}$. In the case $\alpha = 1/2$, the function can be expressed in terms of the complementary error function, namely

$$M(t) = e^t \mathrm{erfc}(\sqrt{t}).$$

The qualitative properties of this $\alpha = 1/2$ solution are similar to those of the ordinary heat equation, but steady-state is approached on a longer time scale (subdiffusion).

For the numerical solution of (5.10)–(5.11), one takes a Laplace transform of (5.10), which yields

$$\boldsymbol{F}(z) = \left(zI - z^{1/2}D\right)^{-1}\boldsymbol{u}_0.$$

We shall continue to let D be the Chebyshev second derivative matrix that incorporates the boundary conditions (5.5). The modification to the Talbot method (5.7)–(5.8) is obvious: the scalar $z(\theta_k)^{1/2}$ should be inserted to multiply D in (5.8).

Finding optimal parameters for Talbot's method for the problem (5.10)–(5.11) would mean analyzing $F(z) = 1/(z - z^{1/2}\lambda)$ as a test function. Note that the singularities are no longer isolated, but a branch cut on the negative real axis. Instead of performing such an analysis, we merely demonstrate numerically that Talbot's method with the parameter choices of section 3 is very accurate for this problem as well. The error curves shown in Figure 5.2 confirm that the convergence rate is, to a good approximation, again given by $O(e^{-1.90N})$.

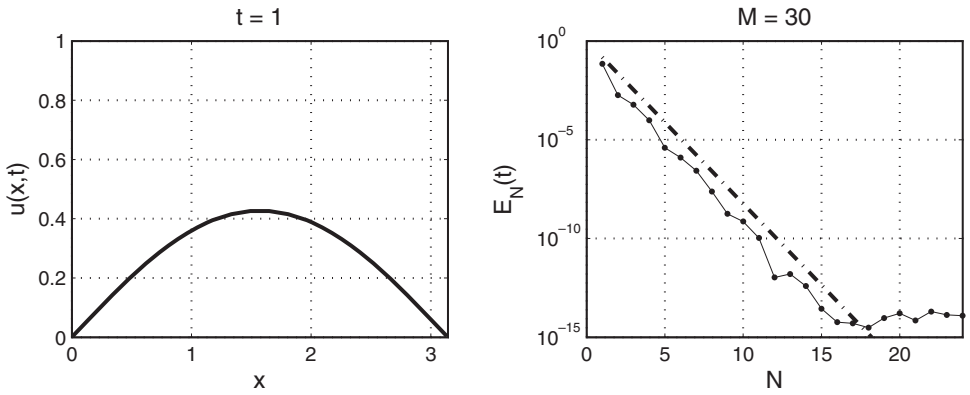


FIG. 5.2. *Same as Figure* 5.1, *but the problem is the fractional PDE* (5.10)–(5.11), *with* $\alpha = 1/2$.

6. Comparisons. Using a combination of asymptotics and heuristics, Talbot made some suggestions for parameter selection in the original paper [19]. In the case of singularities on the real negative axis, the suggested values are

$$\sigma = 0, \qquad \mu = \frac{\omega}{t}, \qquad \nu = 1.$$

The recommended value of ω is 6 (resp., 11) for single (resp., double) precision. In our notation $\omega = mN$, and using $m = 0.6443$ we find that $\omega = 6$ (resp., 11) corresponds to $N \approx 9$ (resp., 17). This is commensurate with our results, as $\exp(-1.90\times 9) \approx 3.7\,10^{-8}$ (approx. single precision) and $\exp(-1.90\times 17) \approx 9.4\,10^{-15}$ (approx. double precision). The recommended values $\sigma = 0$ and $\nu = 1$, however, are suboptimal. Indeed, in the abstract of [19] it is stated that "The required number of points depends on t... and for moderate t is typically 11 for orders of 10^{-6}, 18 for order 10^{-10}, 35 for order 10^{-20}." Fitting a model $E_N = \text{const.} \times e^{-cN}$ to these data yields $c \approx 1.35$, which is not as good as the $c \approx 1.90$ and $c \approx 2.56$ obtained here.

To be fair to Talbot, the aims of the paper [19] were more ambitious than those of the present paper. To begin with, all singularity distributions were taken into account, not just poles on the negative imaginary axis. In addition, Talbot considered finite precision tolerances, and therefore had to deal with the locations of the singularities. By contrast, we let $N \to \infty$, thereby making the errors independent of the singularities, and trusted in the power of asymptotics to make the parameters thus found relevant for finite (indeed, relatively small) values of N as well.

More recently, hyperbolic and parabolic contours have been considered as alternatives to Talbot's contours. Published convergence rates are all subgeometric, namely $O(e^{-cN^{1/2}})$ for the hyperbola of [14], $O(e^{-cN^{2/3}})$ for the parabola of [8], and $O(e^{-cN/\log N})$ for the hyperbola of [12]. The hyperbola has the advantage that it can handle singularities that lie in a sectorial region about the negative real axis; see [12].

Using a rescaling similar to (3.1), the above convergence rates were subsequently improved to the geometric $O(e^{-cN})$; see [13, 26]. In fact, the optimal decay constant c is marginally better for parabolas and hyperbolas than for Talbot contours. With the modification introduced at the end of section 4, however, the Talbot contours regain their superiority.

Appendix. Table A.1 lists saddle points, critical points, and estimated convergence rates corresponding to cases (a)–(c) in Figure 3.2.

TABLE A.1

Case	Parameters	Roots of (3.6)(+)	Roots of (3.6)(−)	Roots of (3.7)	θ_+, θ_-, θ_*	d_+, d_-, d_*	Conv. rate
(a)	$s = 0.8$	$\pm 3.9751 - 0.8123i$	$\pm 4.1761 + 0.9975i$	$2.8686i$	$-2.2930 + 0.8123i$	-2.2380	
	$m = 1.05$	$\pm 2.2930 + 0.8123i$	$\pm 2.0710 - 0.9975i$	$\pm 0.4351 + 1.3633i$	$-2.0710 - 0.9975i$	-1.4859	$\exp(-1.4859N)$
	$n = 0.4$				$-0.4351 + 1.3633i$	-2.7267	
(b)	$s = 0.6$	$\pm 3.8710 - 0.7145i$	$\pm 4.0372 + 0.8700i$	$0.6780i$	$-2.4033 + 0.7145i$	-2.1185	
	$m = 0.8$	$\pm 2.4033 + 0.7145i$	$\pm 2.2258 - 0.8700i$	$\pm 1.9043 + 1.9110i$	$-2.2258 - 0.8700i$	-1.4908	
	$n = 0.5$				$0.6780i$	-1.3560	$\exp(-1.3560N)$
(c)	$s = 0.7556$	$\pm 3.9213 - 0.7620i$	$\pm 4.0319 + 0.8652i$	$5.2713i$	$-2.3503 + 0.7620i$	-2.1524	
	$m = 0.8597$	$\pm 2.3503 + 0.7620i$	$\pm 2.2315 - 0.8652i$	$0.8652i$	$-2.2315 - 0.8652i$	-1.7303	$\exp(-1.7303N)$
	$n = 0.3029$			$0.8652i$	$0.8652i$	-1.7303	

Acknowledgments. Kevin Burrage, Nick Gould, Thomas Schmelzer, and Nick Trefethen all contributed useful suggestions, as did the anonymous referees. The author and Fusen Lin had many discussions on Talbot's method while preparing the thesis [11].

REFERENCES

[1] M. J. Ablowitz and A. S. Fokas, *Complex Variables: Introduction and Applications*, 2nd ed., Cambridge University Press, Cambridge, UK, 2003.

[2] M. Y. Antimirov, A. A. Kolyshkin, and R. Vaillancourt, *Applied Integral Transforms*, AMS, Providence, RI, 1993.

[3] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[4] E. Cuesta and C. Palencia, *A numerical method for an integro-differential equation with memory in Banach spaces: Qualitative properties*, SIAM J. Numer. Anal., 41 (2003), pp. 1232–1241.

[5] P. J. Davis, *On the numerical integration of periodic analytic functions*, in On Numerical Approximation (Proceedings of a Symposium, Madison, WI, 1958), R. E. Langer, ed., The University of Wisconsin Press, Madison, WI, 1959.

[6] D. G. Duffy, *On the numerical inversion of Laplace transforms: Comparison of three new methods on characteristic problems from applications*, ACM Trans. Math. Software, 19 (1993), pp. 333–359.

[7] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, Cambridge, UK, 1996.

[8] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent parallel discretization methods for the first order evolution equations*, Comput. Methods Appl. Math., 1 (2001), pp. 333–355.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[10] E. T. Goodwin, *The evaluation of integrals of the form* $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$, Proc. Cambridge Philos. Soc., 45 (1949), pp. 241–245.

[11] F. Lin, *Numerical Inversion of Laplace Transforms by Trapezoidal-Type Methods*, Ph.D. dissertation, Oregon State University, Corvallis, OR, 2003.

[12] M. López-Fernández and C. Palencia, *On the numerical inversion of the Laplace transform of certain holomorphic mappings*, Appl. Numer. Math., 51 (2004), pp. 289–303.

[13] M. López-Fernández, C. Palencia, and A. Schädle, *A spectral order method for inverting sectorial Laplace transforms*, SIAM J. Numer. Anal., 44 (2006), pp. 1332–1350.

[14] W. McLean and V. Thomée, *Time discretization of an evolution equation via Laplace transforms*, IMA J. Numer. Anal., 24 (2004), pp. 439–463.

[15] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.

[16] I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego, CA, 1999.

[17] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature*, Math. Comp., 69 (2000), pp. 177–195.

[18] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time discretization of parabolic equations based on Laplace transformation and quadrature*, IMA J. Numer. Anal., 23 (2003), pp. 269–299.

[19] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.

[20] L. N. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.

[21] L. N. Trefethen, *private communication*, Oxford University Computing Laboratory, Oxford, UK, 2005.

[22] A. M. Turing, *A method for the calculation of the zeta-function*, Proc. London Math. Soc. (2), 48 (1943), pp. 180–197.

[23] J. A. C. Weideman, *Numerical integration of periodic functions: A few examples*, Amer. Math. Monthly, 109 (2002), pp. 21–36.

[24] J. A. C. Weideman, *Optimizing Talbot's Contours for the Inversion of the Laplace Transform*, Technical report NA 05/05, Oxford University Computing Laboratory, Oxford, UK, 2005.
[25] J. A. C. Weideman and S. C. Reddy, *A MATLAB differentiation matrix suite*, ACM Trans. Math. Software, 26 (2000), pp. 465–519.
[26] J. A. C. Weideman and L. N. Trefethen, *Parabolic and hyperbolic contours for computing the Bromwich integral*, Math. Comp., to appear.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2363–2388

DISCONTINUOUS GALERKIN METHODS FOR FRIEDRICHS' SYSTEMS. PART II. SECOND-ORDER ELLIPTIC PDES*

ALEXANDRE ERN† AND JEAN-LUC GUERMOND‡

Abstract. This paper is the second part of a work attempting to give a unified analysis of discontinuous Galerkin methods. The setting under scrutiny is that of Friedrichs' systems endowed with a particular 2×2 structure in which one unknown can be eliminated to yield a system of second-order elliptic-like PDEs for the remaining unknown. A general discontinuous Galerkin method for approximating such systems is proposed and analyzed. The key feature is that the unknown that can be eliminated at the continuous level can also be eliminated at the discrete level by solving local problems. All the design constraints on the boundary operators that weakly enforce boundary conditions and on the interface operators that penalize interface jumps are fully stated. Examples are given for advection-diffusion-reaction, linear continuum mechanics, and a simplified version of the magneto-hydrodynamics equations. Comparisons with well-known discontinuous Galerkin approximations for the Poisson equation are presented.

Key words. Friedrichs' systems, finite elements, partial differential equations, discontinuous Galerkin method

AMS subject classifications. 65N30, 65M60, 35F15

DOI. 10.1137/05063831X

1. Introduction.

Friedrichs' systems [10] are systems of first-order PDEs endowed with a symmetry and a positivity property. Such systems embrace both elliptic and hyperbolic PDEs; i.e., they include advection-reaction, advection-diffusion-reaction, linear continuum mechanics, and Maxwell's equations in the elliptic regime, to cite a few examples. The analysis of this class of problems and its approximation by means of discontinuous Galerkin (DG) methods has been initiated by Lesaint [13], Lesaint and Raviart [12], and Johnson, Nävert, and Pitkäranta [11]. A thorough systematic analysis generalizing [13, 12, 11] has been undertaken in the first part of this work [9].

In this second part, we specialize the setting to two-field Friedrichs' systems such that (i) the dependent variable z can be partitioned into the form $z = (z^\sigma, z^u)$, and (ii) the σ-component, z^σ, can be eliminated to yield a system of second-order PDEs for the u-component, z^u, which is of elliptic type. To efficiently approximate the above Friedrichs' systems using DG methods, it is desirable to reproduce at the discrete level the possibility of eliminating the σ-component of the discrete unknown *locally* on each mesh element. This feature induces a nontrivial modification of the analysis presented in [9] that constitutes the scope of the present work. In particular, the design of boundary and interface operators has to be revised. The analysis presented herein shows that to recover stability while allowing for the local elimination in question requires an enhanced penalty on the boundary conditions and on the interface jumps of the discrete u-component.

*Received by the editors August 17, 2005; accepted for publication (in revised form) July 3, 2006; published electronically November 24, 2006.

http://www.siam.org/journals/sinum/44-6/63831.html

†CERMICS, Ecole des Ponts, ParisTech, 77455 Marne la Vallée Cedex 2, France (ern@cermics.enpc.fr).

‡Department of Mathematics, Texas A&M, College Station, TX 77843-3368 (guermond@math.tamu.edu); on leave from LIMSI (CNRS-UPR 3251), BP 133, 91403, Orsay, France.

This paper is organized as follows. Section 2 briefly restates the main theoretical results of [9] on the well-posedness of Friedrichs' systems and introduces the above-mentioned two-field structure. Section 3 presents three important examples of two-field Friedrichs' systems, namely advection-diffusion-reaction equations written in mixed form, linear continuum mechanics equations written in the stress-pressure-displacement form, and a simplified form of the magnetohydrodynamics (MHD) equations. Section 4 formulates a general DG method for two-field Friedrichs' systems and describes the technique to locally eliminate the σ-component of the discrete solution. The convergence analysis constitutes the scope of section 5. All the design assumptions on the boundary operators which weakly enforce boundary conditions and on the interface operators which penalize interface jumps are stated. The key results are Theorem 5.8, which contains the main estimate for the σ- and u-component of the approximation error, and Theorem 5.14, which contains an improved estimate for the u-component of the error in the L^2-norm obtained using a duality argument. Finally, section 6 applies the DG method to the PDE systems presented in section 3; in particular, the link with the unified analysis of Arnold et al. [1] for the Poisson equation is explicated to illustrate the fact that various DG methods presented in the literature, e.g., the local discontinuous Galerkin (LDG) method of Cockburn and Shu [7], the interior penalty (IP) method of Baker [3] and Arnold [2], the method of Brezzi et al. [6], and the methods of Bassi and Rebay [5] and Bassi et al. [4], fit into the present framework.

2. Two-field Friedrichs' systems.

Section 2.1 is meant to recall well-posedness results proved in part I, [9]. The reader familiar with this material can jump to section 2.2, where the notion of two-field Friedrichs' systems is introduced.

2.1. Main results on one-field Friedrichs' systems.

Let Ω be a bounded, open, connected, Lipschitz domain in $\mathbb{R}^d$. Let m be a positive integer and set $L = [L^2(\Omega)]^m$ equipped with the canonical L^2-induced inner product $(\cdot,\cdot)_L$. Let $\mathcal{K}$ and $\{\mathcal{A}^k\}_{1\le k\le d}$ be $(d+1)$ functions on Ω with values in $\mathbb{R}^{m,m}$ such that

$$\mathcal{K} \in [L^\infty(\Omega)]^{m,m}, \tag{A1}$$

$$\forall k \in \{1,\dots,d\}, \quad \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m} \quad \text{and} \quad \sum_{k=1}^d \partial_k \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m}, \tag{A2}$$

$$\forall k \in \{1,\dots,d\}, \quad \mathcal{A}^k = (\mathcal{A}^k)^t \ \text{ a.e. in } \Omega, \tag{A3}$$

$$\exists \mu_0 > 0, \quad \mathcal{K} + \mathcal{K}^t - \sum_{k=1}^d \partial_k \mathcal{A}^k \ge 2\mu_0 \mathcal{I}_m \ \text{ a.e. on } \Omega, \tag{A4}$$

where $\mathcal{I}_m$ is the identity matrix in $\mathbb{R}^{m,m}$. To alleviate notation we define the operator $K \in \mathcal{L}(L;L)$ by $K : L \ni z \longmapsto \mathcal{K}z \in L$ and it adjoint $K^* \in \mathcal{L}(L;L)$ by $K^* : L \ni z \longmapsto \mathcal{K}^t z \in L$.

Let $\mathfrak{D}(\Omega)$ be the space of $\mathfrak{C}^\infty$ functions that are compactly supported in Ω. A function z in L is said to have an A-weak derivative in L if the linear form

$$[\mathfrak{D}(\Omega)]^m \ni \phi \longmapsto -\int_\Omega \sum_{k=1}^d z^t \partial_k(\mathcal{A}^k\phi) \in \mathbb{R} \tag{2.1}$$

is bounded on L. In this case, the function in L that can be associated with the above linear form by means of the Riesz representation theorem is denoted by Az. Define the so-called graph space $W = \{z \in L;\, Az \in L\}$ equipped with the graph

norm $\|z\|_W = \|Az\|_L + \|z\|_L$. The space W is endowed with a Hilbert structure when equipped with the scalar product $(z,y)_L + (Az,Ay)_L$. For $z \in W$, the function in L that can be associated with the linear form $[\mathfrak{D}(\Omega)]^m \ni \phi \longmapsto \int_\Omega \sum_{k=1}^d z^t \mathcal{A}^k \partial_k \phi \in \mathbb{R}$ is denoted by $\tilde{A}z$. Clearly, $A \in \mathcal{L}(W;L)$ and $\tilde{A} \in \mathcal{L}(W;L)$ and if z is smooth, e.g., $z \in [\mathfrak{C}^1(\overline{\Omega})]^m$,

$$Az = \sum_{k=1}^d \mathcal{A}^k \partial_k z, \qquad \tilde{A}z = -\sum_{k=1}^d \partial_k(\mathcal{A}^k z). \tag{2.2}$$

Furthermore, we set $T = K + A$, $\tilde{T} = K^* + \tilde{A}$. Note that $\tilde{A}$ and $\tilde{T}$ are the formal adjoints of A and T, respectively, owing to (A3). Assumption (A4) implies

$$\forall z \in W, \quad (Tz,z)_L + (z,\tilde{T}z)_L \geq 2\mu_0 \|z\|_L^2. \tag{2.3}$$

Let $D \in \mathcal{L}(W;W')$ be the operator defined by

$$\forall (z,y) \in W \times W, \quad \langle Dz, y\rangle_{W',W} = (Az,y)_L - (z,\tilde{A}y)_L. \tag{2.4}$$

Observe that D is self-adjoint by construction; moreover, it is a boundary operator in the sense that $\mathrm{Ker}(D)$ is the closure of $[\mathfrak{D}(\Omega)]^m$ in W; see [8] for further results.

Consider the following problem: For $f \in L$, seek $z \in W$ such that $Tz = f$. In general, boundary conditions must be enforced for this problem to be well-posed. In other words, one must find a closed subspace V of W such that the restricted operator $T : V \to L$ is an isomorphism. To achieve this goal, a simple approach inspired from Friedrichs' work [9, 10] consists of introducing an operator $M \in \mathcal{L}(W;W')$ such that

(M1) M is positive, i.e., $\langle Mz, z\rangle_{W',W} \geq 0 \quad \forall z$ in W,

(M2) $W = \mathrm{Ker}(D-M) + \mathrm{Ker}(D+M)$.

Then by setting

$$V = \mathrm{Ker}(D-M) \qquad \text{and} \qquad V^* = \mathrm{Ker}(D+M^*), \tag{2.5}$$

where $M^* \in \mathcal{L}(W;W')$ is the adjoint of M and V and V^* are equipped with the graph norm, the following theorem can be proved (see [8, 9] for a proof).

THEOREM 2.1. *Assume* (A1)–(A4) *and* (M1)–(M2). *Then, the restricted operators* $T : V \to L$ *and* $\tilde{T} : V^* \to L$ *are isomorphisms.*

As a result, for f in L, the following two problems are well-posed:

$$\text{Seek } z \in V \text{ such that } Tz = f, \tag{2.6}$$

$$\text{Seek } z^* \in V^* \text{ such that } \tilde{T}z^* = f. \tag{2.7}$$

A key observation at this point is that the boundary conditions enforced in (2.6) and (2.7) are essential; i.e., they are enforced strongly by seeking the solutions in V and V^*, respectively. The key reason that led us to focus on the theory of Friedrichs' systems is that it yields a way to enforce boundary conditions naturally, thus leading to a suitable framework for developing a DG theory. To see this, we introduce the following bilinear forms on $W \times W$:

$$a(z,y) = (Tz,y)_L + \tfrac{1}{2}\langle (M-D)z, y\rangle_{W',W}, \tag{2.8}$$

$$a^*(z,y) = (\tilde{T}z,y)_L + \tfrac{1}{2}\langle (M^*+D)z, y\rangle_{W',W}. \tag{2.9}$$

It is clear that a and a^* are in $\mathcal{L}(W\times W;\mathbb{R})$. Equipped with these two new bilinear forms, we now consider the following problems: For $f\in L$,

$$\text{Seek } z\in W \text{ such that } a(z,y)=(f,y)_L \quad \forall y\in W, \tag{2.10}$$

$$\text{Seek } z^*\in W \text{ such that } a^*(z^*,y)=(f,y)_L \quad \forall y\in W. \tag{2.11}$$

The key result of this section is the following

THEOREM 2.2. *Assume* (A1)–(A4) *and* (M1)–(M2). *Then,*

(i) *there is a unique solution to* (2.10) *and this solution solves* (2.6)*;*

(ii) *there is a unique solution to* (2.11) *and this solution solves* (2.7).

Theorem 2.2 is proven in [9]. Contrary to (2.6) and (2.7), the boundary conditions in (2.10) and (2.11) are natural; i.e., they are weakly enforced. For this reason, problem (2.10) will constitute our working basis for designing DG methods; see section 4.

2.2. The two-field structure. We now particularize the above setting by assuming that the $(d+1)$ $\mathbb{R}^{m,m}$-valued fields $\mathcal{K}$ and $\{\mathcal{A}^k\}_{1\le k\le d}$ have a 2×2 block structure; i.e., there are two positive integers m_σ and m_u such that $m=m_\sigma+m_u$ and

$$\mathcal{K}=\left[\begin{array}{c:c}\mathcal{K}^{\sigma\sigma}&\mathcal{K}^{\sigma u}\\ \hdashline \mathcal{K}^{u\sigma}&\mathcal{K}^{uu}\end{array}\right],\qquad \mathcal{A}^k=\left[\begin{array}{c:c}0&\mathcal{B}^k\\ \hdashline [\mathcal{B}^k]^t&\mathcal{C}^k\end{array}\right], \tag{2.12}$$

with obvious notation for the blocks of $\mathcal{K}$ and where for all $k\in\{1,\dots,d\}$, $\mathcal{B}^k$ is an $m_\sigma\times m_u$ matrix field and $\mathcal{C}^k$ is a symmetric $m_u\times m_u$ matrix field. To simplify the notation, define the operators $B=\sum_{k=1}^d\mathcal{B}^k\partial_k$, $B^\dagger=\sum_{k=1}^d[\mathcal{B}^k]^t\partial_k$, $\nabla\cdot B=\sum_{k=1}^d\partial_k\mathcal{B}^k$, $C=\sum_{k=1}^d\mathcal{C}^k\partial_k$, $C^\dagger=\sum_{k=1}^d[\mathcal{C}^k]^t\partial_k$, and $\nabla\cdot C=\sum_{k=1}^d\partial_k\mathcal{C}^k$. Set $L_\sigma=[L^2(\Omega)]^{m_\sigma}$ and $L_u=[L^2(\Omega)]^{m_u}$.

The two key hypotheses on which the present work is based are the following:

$$\exists k_0>0\quad \forall\xi\in\mathbb{R}^{m_\sigma},\quad \xi^t\mathcal{K}^{\sigma\sigma}\xi\ge k_0\|\xi\|^2_{\mathbb{R}^{m_\sigma}}\quad \text{a.e. on } \Omega, \tag{A5}$$

$$\forall k\in\{1,\dots,d\},\quad \text{the } m_\sigma\times m_\sigma \text{ upper-left block of } \mathcal{A}^k \text{ is zero.} \tag{A6}$$

Assumption (A5), which means that $\mathcal{K}^{\sigma\sigma}$ is uniformly positive definite, implies that the matrix $\mathcal{K}^{\sigma\sigma}$ is invertible.

Assumptions (A5) and (A6) allow for the elimination of z^σ from the PDE system $Tz=f$. With obvious notation, partition z and f into (z^σ,z^u) and (f^σ,f^u), respectively. Then, z^σ is given by

$$z^\sigma=[\mathcal{K}^{\sigma\sigma}]^{-1}\Big(f^\sigma-\mathcal{K}^{\sigma u}z^u-Bz^u\Big), \tag{2.13}$$

and z^u solves the following second-order PDE:

$$\begin{aligned}-B^\dagger[\mathcal{K}^{\sigma\sigma}]^{-1}Bz^u+(C-B^\dagger[\mathcal{K}^{\sigma\sigma}]^{-1}\mathcal{K}^{\sigma u}-\mathcal{K}^{u\sigma}[\mathcal{K}^{\sigma\sigma}]^{-1}B)z^u\\ +(\mathcal{K}^{uu}-\mathcal{K}^{u\sigma}[\mathcal{K}^{\sigma\sigma}]^{-1}\mathcal{K}^{\sigma u})z^u=f^u-(\mathcal{K}^{u\sigma}+B^\dagger)[\mathcal{K}^{\sigma\sigma}]^{-1}f^\sigma.\end{aligned} \tag{2.14}$$

The objective of the present work is to design DG methods for approximating (2.14). The strategy we are going to follow consists of constructing a DG approximation to (2.10), but at variance with what has been done in [9], the construction is now specialized to the above 2×2 block structure so that the approximate unknown corresponding to z^σ can be eliminated locally on each mesh element by solving local problems.

Remark 2.1. The present study does not cover the DG approximation of the whole realm of second-order PDEs. Indeed, it is clear from (2.14) that the leading-order term in the PDE, namely $B^\dagger[\mathcal{K}^{\sigma\sigma}]^{-1}Bz^u$ (up to first-order terms), has a very particular structure since the matrices $(\mathcal{B}^k)^t[\mathcal{K}^{\sigma\sigma}]^{-1}\mathcal{B}^k$ are positive semidefinite. Hence, the PDEs covered by this work are elliptic-like; see section 3 for various examples.

Remark 2.2. In some applications, K has no local representation; i.e., there is no local field $\mathcal{K}$ to represent K. This is indeed the case for the neutron transport equation, where K is a scattering operator. Everything that is said hereafter is also valid in this case, provided the matrix block representation of $\mathcal{K}$ is replaced by the operator block representation of K and provided $K^{\sigma\sigma}$ has a local representation, i.e., $(K^{\sigma\sigma}z^\sigma, y^\sigma)_{L_\sigma} = \int_\Omega (y^\sigma)^t \mathcal{K}^{\sigma\sigma} z^\sigma$.

2.3. Integral representation of boundary operators. Let $n = (n_1, \dots, n_d)^t$ be the unit outward normal to $\partial\Omega$. Henceforth, we assume that the fields $\{\mathcal{A}^k\}_{1\le k\le d}$ are sufficiently smooth for the matrix $\mathcal{D} = \sum_{k=1}^d n_k \mathcal{A}^k$ to be meaningful at the boundary. Hence, the following representation holds:

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{D} z \tag{2.15}$$

whenever z and y are smooth functions. Owing to (2.12), $\mathcal{D}$ has a 2×2 block structure with $\mathcal{D}^{\sigma u} = \sum_{k=1}^d n_k \mathcal{B}^k$, $\mathcal{D}^{u\sigma} = [\mathcal{D}^{\sigma u}]^t$, $\mathcal{D}^{uu} = \sum_{k=1}^d n_k \mathcal{C}^k$, and

$$\mathcal{D}^{\sigma\sigma} = 0. \tag{2.16}$$

Likewise, we assume that the boundary operator M has an integral representation; i.e., there exists a matrix-valued field $\mathcal{M} : \partial\Omega \longrightarrow \mathbb{R}^{m,m}$ such that

$$\langle Mz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{M} z \tag{2.17}$$

whenever z and y and smooth functions. We denote by $\mathcal{M}^{\sigma u}$, $\mathcal{M}^{u\sigma}$, and $\mathcal{M}^{uu}$ the top-right, bottom-left, and bottom-right blocks of $\mathcal{M}$, respectively. Henceforth, we assume that

$$\mathcal{M}^{\sigma\sigma} = 0. \tag{2.18}$$

This assumption holds for all the two-field Friedrichs' systems presented in section 3. For instance, the Dirichlet-like boundary condition $\mathcal{D}^{\sigma u} z^u = 0$ can be enforced by taking

$$\mathcal{M} = \left[\begin{array}{c|c} 0 & -\mathcal{D}^{\sigma u} \\ \hline \mathcal{D}^{u\sigma} & \mathcal{M}^{uu} \end{array}\right], \tag{2.19}$$

where $\mathcal{M}^{uu}$ is a positive matrix in $\mathbb{R}^{m_u,m_u}$ (this means that for all $\zeta \in \mathbb{R}^{m_u}$, $\zeta^t \mathcal{M}^{uu}\zeta \ge 0$) and is constructed so that $\mathrm{Ker}(\mathcal{D}^{\sigma u}) \subset \mathrm{Ker}(\mathcal{M}^{uu} - \mathcal{D}^{uu})$ (for instance take $\mathcal{M}^{uu} = \mathcal{D}^{uu} + c\,(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}}$ with c large enough for $\mathcal{M}^{uu}$ to be positive). Similarly, taking

$$\mathcal{M} = \left[\begin{array}{c|c} 0 & \mathcal{D}^{\sigma u} \\ \hline -\mathcal{D}^{u\sigma} & \mathcal{M}^{uu} \end{array}\right], \tag{2.20}$$

where $\mathcal{M}^{uu}$ is a positive matrix in $\mathbb{R}^{m_u,m_u}$, yields the Robin boundary condition $2\mathcal{D}^{u\sigma}z^\sigma + (\mathcal{D}^{uu} - \mathcal{M}^{uu})z^u = 0$. The homogeneous Neumann boundary condition is obtained by setting $\mathcal{M}^{uu} = \mathcal{D}^{uu}$ whenever $\mathcal{D}^{uu}$ is positive. See (3.7) and (6.3) for examples.

3. Examples. This section presents three examples of Friedrichs' systems endowed with the 2×2 block structure introduced in section 2.2.

3.1. Advection-diffusion-reaction. Consider the PDE

$$-\nabla\cdot(\kappa\nabla u) + \beta\cdot\nabla u + \mu u = f, \tag{3.1}$$

with $\beta \in [L^\infty(\Omega)]^d$, $\nabla\cdot\beta \in L^\infty(\Omega)$, $\mu \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and where $\kappa = (\kappa_{kl})_{1\le k,l\le d}$ is a symmetric positive definite tensor-valued field defined on Ω whose lowest eigenvalue is uniformly bounded away from zero. Assume also that

$$\mu - \tfrac{1}{2}\nabla\cdot\beta \ge \mu_0 > 0 \quad \text{a.e. in } \Omega. \tag{3.2}$$

The PDE (3.1) can be written as a system of first-order PDEs in the form

$$\begin{cases} \kappa^{-1}\sigma + \nabla u = 0, \\ \mu u + \nabla\cdot\sigma + \beta\cdot\nabla u = f. \end{cases} \tag{3.3}$$

Set $m = d+1$, $m_\sigma = d$, and $m_u = 1$. Then, the mixed formulation (3.3) can be cast into the form of a two-field Friedrichs' system by introducing $(d+1)$ functions with values in $\mathbb{R}^{m,m}$, namely $\mathcal{K}$ and $\{\mathcal{A}^k\}_{1\le k\le d}$ such that

$$\mathcal{K} = \left[\begin{array}{c:c} \kappa^{-1} & 0 \\ \hdashline 0 & \mu \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c:c} 0 & e^k \\ \hdashline (e^k)^t & \beta^k \end{array}\right], \tag{3.4}$$

where e^k is the kth vector in the canonical basis of $\mathbb{R}^d$ and β^k is the kth component of β in this basis. It is clear that hypotheses (A1)–(A6) hold. The graph space is $W = H(\mathrm{div};\Omega) \times H^1(\Omega)$ and for all $(\sigma,u),(\tau,v) \in W$,

$$\langle D(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \sigma\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + \langle \tau\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (\beta\cdot n)uv, \tag{3.5}$$

where $\langle,\rangle_{-\frac{1}{2},\frac{1}{2}}$ denotes the duality pairing between $H^{-\frac{1}{2}}(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$. Note that (3.5) makes sense since functions in $H^1(\Omega)$ have traces in $H^{\frac{1}{2}}(\partial\Omega)$ and vector fields in $H(\mathrm{div};\Omega)$ have normal traces in $H^{-\frac{1}{2}}(\partial\Omega)$.

Homogeneous Dirichlet boundary conditions can be enforced by setting

$$\langle M(\sigma,u),(\tau,v)\rangle_{W',W} = \langle \sigma\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle \tau\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}}. \tag{3.6}$$

With this choice $V = V^* = H(\mathrm{div};\Omega)\times H^1_0(\Omega)$. Let $\varrho \in L^\infty(\partial\Omega)$ be such that $2\varrho + \beta\cdot n \ge 0$ a.e. in $\partial\Omega$. Then, setting

$$\langle M(\sigma,u),(\tau,v)\rangle_{W',W} = -\langle \sigma\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + \langle \tau\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (2\varrho + \beta\cdot n)uv, \tag{3.7}$$

the spaces V and V^* are defined by $V = \{(\sigma,u) \in W;\, (-\sigma\cdot n + \varrho u)|_{\partial\Omega} = 0\}$ and $V^* = \{(\sigma,u) \in W;\, (\sigma\cdot n + (\varrho + \beta\cdot n)u)|_{\partial\Omega} = 0\}$; i.e., a Robin boundary condition is enforced. A Neumann condition corresponds to $\varrho = 0$. We refer the reader to [9] for more details.

Remark 3.1. When κ is not invertible, Friedrichs' formalism can be extended as detailed in [8].

3.2. Linear continuum mechanics. Let α and γ be two positive functions in $L^\infty(\Omega)$ uniformly bounded away from zero by α_0 and γ_0, respectively. Consider the following set of PDEs:

$$\begin{cases} \sigma + p\mathcal{I}_d - \frac{1}{2}(\nabla u + (\nabla u)^t) = 0, \\ \mathrm{tr}(\sigma) + (d+\gamma)p = 0, \\ -\frac{1}{2}\nabla\cdot(\sigma+\sigma^t) + \alpha u = f, \end{cases} \tag{3.8}$$

where σ is $\mathbb{R}^{d,d}$-valued, p is scalar-valued, u is $\mathbb{R}^d$-valued, and $f \in [L^2(\Omega)]^d$. The first and second equations in (3.8) imply $p = -\gamma^{-1}\nabla\cdot u$ and $\sigma = \frac{1}{2}(\nabla u + (\nabla u)^t) + \gamma^{-1}(\nabla\cdot u)\mathcal{I}_d$; γ is a compressibility coefficient, σ is the stress tensor, $\frac{1}{2}(\nabla u + (\nabla u)^t)$ is the strain tensor, and u represents the displacement field in solid mechanics and the velocity field in fluid mechanics. In the usual solid mechanics equations, the function α vanishes identically. The function α has been introduced in (3.8) to ensure that the positivity property (A4) holds; see (3.10). In a forthcoming work, it will be shown that provided mild additional assumptions are made, the positivity property (A4) can be replaced by the weaker assumption (7.1), thus allowing α to vanish identically.

Set $m = d^2 + 1 + d$. The tensor σ in $\mathbb{R}^{d,d}$ is identified with the vector $\overline{\sigma} \in \mathbb{R}^{d^2}$ by setting $\overline{\sigma}_{[ij]} = \sigma_{ij}$ with $1 \le i,j \le d$ and $[ij] = d(j-1)+i$. Then, the mixed formulation (3.8) can be cast into the form of a Friedrichs' system by introducing the $(d+1)$ $\mathbb{R}^{m,m}$-valued fields with the following 3×3 block structure

$$\mathcal{K} = \left[\begin{array}{c|c|c} \mathcal{I}_{d^2} & \mathcal{Z} & 0 \\ \hline (\mathcal{Z})^t & (d+\gamma) & 0 \\ \hline 0 & 0 & \alpha\mathcal{I}_d \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c|c|c} 0 & 0 & \mathcal{E}^k \\ \hline 0 & 0 & 0 \\ \hline (\mathcal{E}^k)^t & 0 & 0 \end{array}\right], \tag{3.9}$$

where $\mathcal{Z} \in \mathbb{R}^{d^2}$ has components given by $\mathcal{Z}_{[ij]} = \delta_{ij}$ with $1 \le i,j \le d$, and for all $k \in \{1,\dots,d\}$, $\mathcal{E}^k \in \mathbb{R}^{d^2,d}$ has components given by $\mathcal{E}^k_{[ij],l} = -\frac{1}{2}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$ with $1 \le i,j,l \le d$; here, the δ's denote Kronecker symbols.

To recover the 2×2 structure introduced in section 2.2, set $m_\sigma = d^2+1$ and $m_u = d$; i.e., the σ-component corresponds to the pair $(\overline{\sigma}, p)$. Then, hypotheses (A1)–(A6) hold. In particular, (A4)–(A5) result from the fact that for all $z = (\overline{\sigma}, p, u) \in \mathbb{R}^m$,

$$z^t\mathcal{K}z \ge \left(1 - \frac{d}{d+\frac{\gamma_0}{2}}\right)\overline{\sigma}^2 + \frac{\gamma_0}{2}p^2 + \frac{d}{d+\frac{\gamma_0}{2}}\left(\overline{\sigma} + \frac{d+\frac{\gamma_0}{2}}{d}p\mathcal{Z}\right)^2 + \alpha_0 u^2 \ge c(\overline{\sigma}^2 + p^2 + u^2), \tag{3.10}$$

where c depends only on d, α_0, and γ_0. Using the second Korn inequality for the variable u, it is readily seen that the graph space is $W = H_{\overline{\sigma}}\times L^2(\Omega)\times[H^1(\Omega)]^d$ with $H_{\overline{\sigma}} = \{\overline{\sigma} \in [L^2(\Omega)]^{d^2};\ \nabla\cdot(\sigma+\sigma^t) \in [L^2(\Omega)]^d\}$. The boundary operator D takes the following form: For all $(\overline{\sigma},p,u), (\overline{\tau},q,v) \in W$,

$$\langle D(\overline{\sigma},p,u), (\overline{\tau},q,v)\rangle_{W',W} = -\langle\tfrac{1}{2}(\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle\tfrac{1}{2}(\sigma+\sigma^t)\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}}, \tag{3.11}$$

where $\langle,\rangle_{-\frac{1}{2},\frac{1}{2}}$ denotes the duality pairing between $[H^{-\frac{1}{2}}(\partial\Omega)]^d$ and $[H^{\frac{1}{2}}(\partial\Omega)]^d$.

To enforce boundary conditions for (3.8), one possibility consists of setting for all $(\overline{\sigma},p,u), (\overline{\tau},q,v) \in W$,

$$\langle M(\overline{\sigma},p,u), (\overline{\tau},q,v)\rangle_{W',W} = \langle\tfrac{1}{2}(\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle\tfrac{1}{2}(\sigma+\sigma^t)\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}}. \tag{3.12}$$

With this choice, the u-component is set to zero at $\partial\Omega$ (i.e., a homogeneous Dirichlet boundary condition on the displacement (in solid mechanics) or on the velocity (in fluid mechanics) is enforced) as shown in the following

LEMMA 3.1. *Let M be given by* (3.12). *Then, $V = V^* = H_{\overline{\sigma}}\times L^2(\Omega)\times[H_0^1(\Omega)]^d$.*

Proof. It is clear that $V = V^*$ since $M + M^* = 0$. Observe that

$$\langle (D-M)(\overline{\sigma},p,u),(\overline{\tau},q,v)\rangle_{W',W} = -\langle (\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}}. \tag{3.13}$$

Hence, it is clear that $H_{\overline{\sigma}}\times L^2(\Omega)\times[H_0^1(\Omega)]^d \subset \mathrm{Ker}(D-M) = V$. Conversely, let $(\overline{\sigma},p,u)\in\mathrm{Ker}(D-M)$. Let $\theta\in[H^{-\frac{1}{2}}(\partial\Omega)]^d$. Consider the following problem: Seek $v_\theta\in[H^1(\Omega)]^d$ such that for all $w\in[H^1(\Omega)]^d$,

$$(v_\theta,w)_{[L^2(\Omega)]^d} + (\nabla v_\theta + (\nabla v_\theta)^t, \nabla w + (\nabla w)^t)_{[L^2(\Omega)]^{d,d}} = \langle\theta,w\rangle_{-\frac{1}{2},\frac{1}{2}}.$$

This problem is well-posed owing to the second Korn inequality and the Lax–Milgram lemma. Set $\tau_\theta = \nabla v_\theta + (\nabla v_\theta)^t$. Since $\overline{\tau}_\theta\in H_{\overline{\sigma}}$, one can take $(\overline{\tau},q,v) = (\overline{\tau}_\theta,0,0)$ in (3.13) yielding $\langle\theta,u\rangle_{-\frac{1}{2},\frac{1}{2}} = 0$. Since θ is arbitrary in $[H^{-\frac{1}{2}}(\partial\Omega)]^d$, it is inferred that $u\in[H_0^1(\Omega)]^d$. ∎

3.3. Simplified MHD. For the sake of simplicity we assume that the space dimension is three, i.e., $d=3$. Let ν, μ, and σ be three functions in $L^\infty(\Omega)$, and let $\beta\in[L^\infty(\Omega)]^3$ be a vector field. A simplified (time-discretized) version of the MHD equations consists of seeking the electric field E and the magnetic field H such that

$$\begin{cases} \nu H + \nabla\times E = 0, \\ \sigma(E+\beta\times(\mu H)) - \nabla\times H = j, \end{cases} \tag{3.14}$$

where $j\in[L^2(\Omega)]^3$ is a given source term. The separation of the electromagnetic field (H,E) into magnetic and electric fields induces a natural partitioning of $[L^2(\Omega)]^6$ into $[L^2(\Omega)]^3\times[L^2(\Omega)]^3$. The PDEs (3.14) are recast into the form of a Friedrichs' system by introducing the following block structured matrices in $\mathbb{R}^{6,6}$:

$$\mathcal{K} = \left[\begin{array}{c:c} \nu\mathcal{I}_3 & 0 \\ \hdashline \sigma\mu\mathcal{V} & \sigma\mathcal{I}_3 \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c:c} 0 & \mathcal{R}^k \\ \hdashline (\mathcal{R}^k)^t & 0 \end{array}\right], \tag{3.15}$$

where $\mathcal{R}^k_{ij} = \epsilon_{ikj}$ is the Levi-Civita permutation tensor, $1\le i,j,k\le 3$, and $\mathcal{V}_{ij} = \sum_{k=1}^d \epsilon_{ikj}\beta^k$. Assume that ν and σ are positive functions on Ω uniformly bounded away from zero and that there is $\alpha_0>0$ such that a.e. in Ω, $2\left(\frac{\nu}{\sigma}\right)^{\frac{1}{2}} - \mu\|\beta\|_{[L^\infty(\Omega)]^d} \ge \alpha_0$. In the above framework, one readily verifies that (A1)–(A6) hold with $m=6$, $m_\sigma=3$, and $m_u=3$. In the full MHD equations, the off-diagonal term induced by β is compensated by a term originating from the conservation of momentum in the Navier–Stokes equations so that the condition for (A4) to hold is simply that ν and σ be uniformly bounded away from zero.

The graph space is $W = H(\mathrm{curl};\Omega)\times H(\mathrm{curl};\Omega)$ and for all $(H,E),(h,e)\in W$,

$$\begin{aligned}\langle D(H,E),(h,e)\rangle_{W',W} = {}& (\nabla\times E,h)_{[L^2(\Omega)]^3} - (E,\nabla\times h)_{[L^2(\Omega)]^3} \\ & + (H,\nabla\times e)_{[L^2(\Omega)]^3} - (\nabla\times H,e)_{[L^2(\Omega)]^3}.\end{aligned} \tag{3.16}$$

When (H,E) and (h,e) are smooth, the above duality product can be interpreted as the boundary integral $\int_{\partial\Omega}[(n\times E)\cdot h + (n\times e)\cdot H]$.

An admissible boundary condition for (3.14) consists of setting

$$\begin{aligned}\langle M(H,E),(h,e)\rangle_{W',W} = &-(\nabla\times E,h)_{[L^2(\Omega)]^3}+(E,\nabla\times h)_{[L^2(\Omega)]^3}\\ &+(H,\nabla\times e)_{[L^2(\Omega)]^3}-(\nabla\times H,e)_{[L^2(\Omega)]^3}\end{aligned}\tag{3.17}$$

for all $(H,E),(h,e)\in W$. Assuming $[H^1(\Omega)]^3$ is dense in $H(\mathrm{curl};\Omega)$, this choice yields $V=V^*=H(\mathrm{curl};\Omega)\times H_0(\mathrm{curl};\Omega)$; i.e., the tangential component of the electric field is set to zero; see [8] for the analysis.

4. Two-field DG approximation. In this section we design a DG method to approximate the two-field Friedrichs' systems introduced in section 2.2. The key feature is that the discrete σ-component can be eliminated locally.

4.1. The discrete setting. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes of Ω. The meshes are assumed to be affine to avoid unnecessary technicalities; i.e., Ω is assumed to be a polyhedron. For $K\in\mathcal{T}_h$, h_K denotes its diameter and we set $h=\max_{K\in\mathcal{T}_h}h_K$. Henceforth, the notation $\xi\lesssim\zeta$ means that there is a positive c, *independent of* h, such that $\xi\le c\zeta$. For any measurable subset E of Ω, we denote by $(\cdot,\cdot)_{L,E}$ the usual scalar product in $[L^2(E)]^m$. We define similarly $(\cdot,\cdot)_{L_u,E}$ and $(\cdot,\cdot)_{L_\sigma,E}$.

We denote by $\mathcal{F}_h^{\mathrm{i}}$ the set of interfaces; i.e., $F\in\mathcal{F}_h^{\mathrm{i}}$ if F is a (d-1)-dimensional manifold and there are $K_1(F)$ and $K_2(F)\in\mathcal{T}_h$ such that $F=K_1(F)\cap K_2(F)$. For $F\in\mathcal{F}_h^{\mathrm{i}}$, we set $\mathcal{T}(F)=K_1(F)\cup K_2(F)$. We denote by $\mathcal{F}_h^{\partial}$ the set of the faces that separate the mesh from the exterior of Ω; i.e., $F\in\mathcal{F}_h^{\partial}$ if F is a (d-1)-dimensional manifold and there is $K(F)\in\mathcal{T}_h$ such that $F=K(F)\cap\partial\Omega$. For $F\in\mathcal{F}_h^{\partial}$, we set $\mathcal{T}(F)=K(F)$. For all $F\in\mathcal{F}_h^{\mathrm{i}}$, we denote by n_F the unit normal vector on F pointing from $K_1(F)$ to $K_2(F)$. For all $F\in\mathcal{F}_h^{\partial}$, we denote by n_F the unit normal vector on F pointing outside Ω. Finally, we set $\mathcal{F}_h=\mathcal{F}_h^{\mathrm{i}}\cup\mathcal{F}_h^{\partial}$. For all $F\in\mathcal{F}_h$, it is assumed that

$$h_{\mathcal{T}(F)}\lesssim h_F,\tag{4.1}$$

where $h_{\mathcal{T}(F)}$ denotes the diameter of $\mathcal{T}(F)$ and h_F that of F. No other assumption than (4.1) is made on the matching of element faces.

For a nonnegative integer p, consider the finite element space of scalar-valued functions

$$P_{h,p}=\{v_h\in L^2(\Omega);\,\forall K\in\mathcal{T}_h,\,v_h|_K\in\mathbb{P}_p\},\tag{4.2}$$

where $\mathbb{P}_p$ denotes the vector space of polynomials with real coefficients and with total degree less than or equal to p. The mesh family $\{\mathcal{T}_h\}_{h>0}$ is assumed to be regular enough for the following inverse and trace inverse inequalities to hold: For all $v_h\in P_{h,p}$,

$$\forall K\in\mathcal{T}_h,\quad \|\nabla v_h\|_{[L^2(K)]^d}\lesssim h_K^{-1}\|v_h\|_{L^2(K)},\tag{4.3}$$

$$\forall F\in\mathcal{F}_h,\quad \|v_h\|_{L^2(F)}\lesssim h_F^{-\frac12}\|v_h\|_{L^2(\mathcal{T}(F))}.\tag{4.4}$$

Let p_u and p_σ be two integers such that

$$1\le p_u\qquad\text{and}\qquad p_u-1\le p_\sigma.\tag{4.5}$$

Define the following vector spaces:

$$U_h=[P_{h,p_u}]^{m_u},\qquad \Sigma_h=[P_{h,p_\sigma}]^{m_\sigma},\qquad W_h=U_h\times\Sigma_h,\tag{4.6}$$

and set $U(h) = [H^1(\Omega)]^{m_u} + U_h$, $\Sigma(h) = [H^1(\Omega)]^{m_\sigma} + \Sigma_h$, and $W(h) = [H^1(\Omega)]^m + W_h$. Obviously, inequalities (4.3) and (4.4) can be applied componentwise to all functions in U_h and in Σ_h. Moreover, since every function v in $U(h)$ has a (possibly two-valued) trace a.e. on $F \in \mathcal{F}_h^{\mathrm{i}}$, we set

$$[\![v]\!] = v^1 - v^2, \qquad \{v\} = \tfrac{1}{2}(v^1 + v^2), \tag{4.7}$$

where for a.e. $x \in F$, $v^\nu(x) = \lim_{y \to x} v(y)|_{K_\nu(F)}$, $\nu \in \{1,2\}$. We define τ^1, τ^2, and $[\![\tau]\!]$ similarly for all τ in $\Sigma(h)$. The arbitrariness in the choice of $K_1(F)$ and $K_2(F)$ could be avoided by choosing intrinsic notations that would, however, unnecessarily complicate the presentation; nothing that is said hereafter depends on this choice. The above mean and jump operators are extended to boundary faces $F \in \mathcal{F}_h^\partial$ by taking the value of the function on that face.

4.2. Boundary and interface operators. For all $F \in \mathcal{F}_h$, we define the matrix-valued field $\mathcal{D}_F : F \to \mathbb{R}^{m,m}$ by

$$\mathcal{D}_F(x) = \sum_{k=1}^d n_{F,k} \mathcal{A}^k(x) \quad \text{a.e. on } F, \tag{4.8}$$

where $n_F = (n_{F,1}, \dots, n_{F,d})^t$. Owing to (2.12), $\mathcal{D}_F$ has a 2×2 block structure with $\mathcal{D}_F^{\sigma u} = \sum_{k=1}^d n_{F,k} \mathcal{B}^k$, $\mathcal{D}_F^{u\sigma} = [\mathcal{D}_F^{\sigma u}]^t$, $\mathcal{D}_F^{uu} = (\mathcal{D}_F^{uu})^t = \sum_{k=1}^d n_{F,k} \mathcal{C}^k$, and

$$\mathcal{D}_F^{\sigma\sigma} = 0. \tag{4.9}$$

The definition (4.8) is clearly compatible with that of $\mathcal{D}$; i.e., if $F \in \mathcal{F}_h^\partial$, $\mathcal{D}_F = \mathcal{D}$. Moreover, observe that for all z, y in $W(h)$ and for all $K \in \mathcal{T}_h$,

$$\sum_{F \subset \partial K} n_F \cdot n_K (\mathcal{D}_F z, y)_{L,F} = (Az, y)_{L,K} - (z, \tilde{A} y)_{L,K}. \tag{4.10}$$

We now extend the matrix-valued field $\mathcal{D}$ to interfaces as follows. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, $\mathcal{D}|_F$ is two-valued, the two values being $n_F \cdot n_{K_1(F)} \mathcal{D}_F$ and $n_F \cdot n_{K_2(F)} \mathcal{D}_F$. Note that $\{\mathcal{D}\} = 0$ a.e. on $\mathcal{F}_h^{\mathrm{i}}$ since $\sum_{k=1}^d \partial_k \mathcal{A}^k$ is bounded owing to (A2).

To weakly enforce boundary conditions, we introduce for all $F \in \mathcal{F}_h^\partial$ a linear operator

$$M_F = \left[\begin{array}{c:c} M_F^{\sigma\sigma} & M_F^{\sigma u} \\ \hdashline M_F^{u\sigma} & M_F^{uu} \end{array}\right] \in \mathcal{L}([L^2(F)]^m; [L^2(F)]^m). \tag{4.11}$$

Note that M_F is not necessarily the restriction of M to functions defined on F; see Remark 5.2 below. Similarly, to penalize interface jumps, we introduce for all $F \in \mathcal{F}_h^{\mathrm{i}}$ a linear operator

$$S_F = \left[\begin{array}{c:c} S_F^{\sigma\sigma} & S_F^{\sigma u} \\ \hdashline S_F^{u\sigma} & S_F^{uu} \end{array}\right] \in \mathcal{L}([L^2(F)]^m; [L^2(F)]^m). \tag{4.12}$$

Star superscripts denote the L^2-adjoint of M_F, S_F, or any block thereof. For instance, $(M_F^{u\sigma})^* \in \mathcal{L}([L^2(F)]^{m_u}; [L^2(F)]^{m_\sigma})$ is defined such that $((M_F^{u\sigma})^*(v), \tau)_{L_\sigma,F} = (M_F^{u\sigma}(\tau), v)_{L_u,F}$ for all $v \in [L^2(F)]^{m_u}$ and for all $\tau \in [L^2(F)]^{m_\sigma}$. Finally, we introduce for all $F \in \mathcal{F}_h$ a linear operator

$$R_F \in \mathcal{L}([L^2(\mathcal{F}_h)]^{m_u}; [L^2(F)]^{m_u}). \tag{4.13}$$

The purpose of this operator is to reduce computational costs when solving the discrete problem for the u-component once the discrete σ-component has been eliminated locally; see section 4.4 and, in particular, (4.31). A simple choice consists of setting $R_F \equiv 0$ for all $F \in \mathcal{F}_h$; an example with nonzero R_F's is the IP method discussed in section 6.1.2.

The operators M_F, S_F, and R_F satisfy various design criteria which are collected in section 5.1. For the time being, we solely mention the important assumption

$$M_F^{\sigma\sigma} = 0 \qquad \text{and} \qquad S_F^{\sigma\sigma} = 0. \tag{4.14}$$

Hence, the jumps across interfaces of the σ-component of the unknown are not controlled. This is the key property that allows for the local elimination of the σ-component of the discrete solution z_h; see section 4.4. This is the most important difference with respect to the DG method analyzed in [9].

4.3. The discrete problem and the notion of fluxes. Drawing inspiration from (2.10), we introduce the bilinear form a_h such that for all z, y in $W(h)$,

$$\begin{aligned} a_h(z,y) = &\sum_{K\in\mathcal{T}_h} (Tz,y)_{L,K} + \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F(z) - \mathcal{D}z, y)_{L,F} - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}z\}, \{y\})_{L,F} \\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F([\![z]\!]), [\![y]\!])_{L,F} + \sum_{F\in\mathcal{F}_h} (R_F([\![z^u]\!]), [\![y^u]\!])_{L_u,F}. \end{aligned} \tag{4.15}$$

The first and second term in the right-hand side come directly from (2.8). The third term is meant to ensure that a_h satisfies a coercivity property on W_h (see Lemma 5.4) in a manner consistent with the continuous setting (this term is zero whenever z is smooth). The fourth term is used to control the jump of the discrete solution across interfaces. The last term is a perturbation (possibly $R_F \equiv 0$) which allows for some modifications of the second and third terms to alleviate computational costs; see the end of section 4.4 and the IP method discussed in section 6.1.2.

The discrete counterpart of (2.10) is the following: For $f = (f^\sigma, f^u) \in L$,

$$\begin{cases} \text{Seek } z_h = (z_h^\sigma, z_h^u) \in W_h \text{ such that} \\ a_h(z_h, y_h) = (f, y_h)_L \quad \forall y_h = (y_h^\sigma, y_h^u) \in W_h. \end{cases} \tag{4.16}$$

As in [9], the discrete problem (4.16) can be localized by using the notion of flux. Let K be a mesh element in $\mathcal{T}_h$ and let $z \in W(h)$. The element flux of z on ∂K, say $\phi_{\partial K}(z) \in [L^2(\partial K)]^m$, is defined by its restriction to the faces F of ∂K as follows:

$$\phi_{\partial K}(z)|_F = \begin{cases} \tfrac{1}{2}(\mathcal{D}z + M_F(z) + 2R'_F(z^u)) & \text{if } F \in \mathcal{F}_h^\partial, \\ n_F{\cdot}n_K(\mathcal{D}_F\{z\} + S_F([\![z]\!]) + R'_F([\![z^u]\!])) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}, \end{cases} \tag{4.17}$$

where $R'_F(z^u) = (0, R_F(z^u)) \in [L^2(F)]^m$.

The discrete problem (4.16) is equivalently reformulated in terms of the following local problems posed for all $K \in \mathcal{T}_h$:

$$\begin{cases} \text{Seek } z_h \in W_h \text{ such that } \forall q = (q^\sigma, q^u) \in [\mathbb{P}_{p_\sigma}(K)]^{m_\sigma} \times [\mathbb{P}_{p_u}(K)]^{m_u}, \\ (Kz_h, q)_{L,K} + (Az_h, q)_{L,K} + (\phi_{\partial K}(z_h) - n_F{\cdot}n_K\mathcal{D}_F z_h|_K, q)_{L,\partial K} = (f,q)_{L,K}, \end{cases} \tag{4.18}$$

or equivalently using the local integration by parts formula (4.10),

$$\begin{cases} \text{Seek } z_h \in W_h \text{ such that } \forall q = (q^\sigma, q^u) \in [\mathbb{P}_{p_\sigma}(K)]^{m_\sigma} \times [\mathbb{P}_{p_u}(K)]^{m_u}, \\ (Kz_h, q)_{L,K} + (z_h, \tilde{A}q)_{L,K} + (\phi_{\partial K}(z_h), q)_{L,\partial K} = (f, q)_{L,K}. \end{cases} \tag{4.19}$$

4.4. Eliminating the σ-component. We now rewrite (4.18) using the 2×2 block structure, and we show how the unknown z_h^σ can be locally eliminated. To simplify, we assume that $f^\sigma \equiv 0$ (this is a natural assumption to define z^σ in physical models). Recall that the σ-component of the element flux is

$$\phi_{\partial K}^\sigma(z^u)|_F = \begin{cases} \frac{1}{2}(\mathcal{D}^{\sigma u} + M_F^{\sigma u})z^u & \text{if } F \in \mathcal{F}_h^\partial, \\ n_F{\cdot}n_K(\mathcal{D}_F^{\sigma u}\{z^u\} + S_F^{\sigma u}([\![z^u]\!])) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}, \end{cases} \tag{4.20}$$

where we stress that $\phi_{\partial K}^\sigma$ solely depends on z^u owing to (4.14). Then, (4.18) implies that z_h^σ solves the following local problems: For all $q^\sigma \in \mathbb{P}_\sigma(K) := [\mathbb{P}_{p_\sigma}(K)]^{m_\sigma}$,

$$(\mathcal{K}^{\sigma\sigma} z_h^\sigma + \mathcal{K}^{\sigma u} z_h^u + B z_h^u, q^\sigma)_{L_\sigma,K} + (\phi_{\partial K}^\sigma(z_h^u) - \mathcal{D}_{\partial K}^{\sigma u} z_h^u|_K, q^\sigma)_{L_\sigma,\partial K} = 0. \tag{4.21}$$

For all $K \in \mathcal{T}_h$, let θ_K^1 be the L^2-orthogonal projection from $[L^2(K)]^{m_\sigma}$ onto $\mathbb{P}_\sigma(K)$ and let $\theta_K^2 : \mathbb{P}_\sigma(K) \to \mathbb{P}_\sigma(K)$ be the mapping such that for all $q^\sigma \in \mathbb{P}_\sigma(K)$, $(\theta_K^2(q^\sigma), r^\sigma)_{L_\sigma,K} = (\mathcal{K}^{\sigma\sigma} q^\sigma, r^\sigma)_{L_\sigma,K}$ for all $r^\sigma \in \mathbb{P}_\sigma(K)$ (note that θ_K^2 is the identity whenever $\mathcal{K}^{\sigma\sigma}$ is the identity matrix in $\mathbb{R}^{m_\sigma,m_\sigma}$). Let $F \in \mathcal{F}_h$. Define the mapping $r_F : [L^2(F)]^{m_\sigma} \longrightarrow \Sigma_h$ so that for all $z^\sigma \in [L^2(F)]^{m_\sigma}$, $r_F(z^\sigma)$ solves

$$(r_F(z^\sigma), y_h^\sigma)_{L_\sigma} = (z^\sigma, \{y_h^\sigma\})_{L_\sigma,F} \quad \forall y_h^\sigma \in \Sigma_h. \tag{4.22}$$

Observe that the support of $r_F(z^\sigma)$ is contained in $\mathcal{T}(F)$. Then, (4.21) yields the local reconstruction formula for the discrete σ-component in the form

$$\forall K \in \mathcal{T}_h, \quad z_h^\sigma|_K = \mathfrak{R}_K(z_h^u) + \mathfrak{R}_{\Delta_K}([\![z_h^u]\!]), \tag{4.23}$$

where

$$\mathfrak{R}_K(z_h^u) = -(\theta_K^2)^{-1}\theta_K^1(\mathcal{K}^{\sigma u} z_h^u + B z_h^u|_K) \tag{4.24}$$

is supported on K, and where

$$\mathfrak{R}_{\Delta_K}([\![z_h^u]\!]) = -(\theta_K^2)^{-1} \sum_{F \subset \partial K} r_F(\psi_{F,K}([\![z_h^u]\!])) \tag{4.25}$$

is supported on $\Delta_K = \{K' \in \mathcal{T}_h;\ \exists F \in \mathcal{F}_h^{\mathrm{i}};\ F = K \cap K'\}$. Here,

$$\psi_{F,K}(v) = \begin{cases} \frac{1}{2}(M_F^{\sigma u} - \mathcal{D}^{\sigma u})v & \text{if } F \in \mathcal{F}_h^\partial, \\ (2n_F{\cdot}n_K S_F^{\sigma u} - \mathcal{D}_F^{\sigma u})v & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}. \end{cases} \tag{4.26}$$

Then, using (4.23) in (4.19) shows that z_h^u solves the following problems: For all $K \in \mathcal{T}_h$ and for all $q^u \in \mathbb{P}_u(K) := [\mathbb{P}_{p_u}(K)]^{m_u}$,

(4.27)

$$\begin{aligned} &((\mathcal{K}^{u\sigma} - (\nabla{\cdot}B)^*)(\mathfrak{R}_K(z_h^u) + \mathfrak{R}_{\Delta_K}([\![z_h^u]\!])) + (\mathcal{K}^{uu} - \nabla{\cdot}C) z_h^u - f^u, q^u)_{L_u,K} \\ &- (z_h^u, C^\dagger q^u)_{L_u,K} - (\mathfrak{R}_K(z_h^u) + \mathfrak{R}_{\Delta_K}([\![z_h^u]\!]), B^\dagger q^u)_{L_u,K} + (\phi_{\partial K}^u(z_h^u), q^u)_{L_u,\partial K} = 0, \end{aligned}$$

where for $F \in \mathcal{F}_h^\partial$,

$$\phi^u_{\partial K}(z^u_h)|_F = \tfrac{1}{2}(M^{u\sigma}_F + \mathcal{D}^{u\sigma})(\mathfrak{R}_K(z^u_h) + \mathfrak{R}_{\Delta_K}([\![z^u_h]\!])) + \tfrac{1}{2}(M^{uu}_F + \mathcal{D}^{uu})z^u_h + R_F([\![z^u_h]\!]), \tag{4.28}$$

and for $F \in \mathcal{F}^{\mathrm{i}}_h$,

$$\phi^u_{\partial K}(z^u_h)|_F = n_F{\cdot}n_K(\mathcal{D}^{u\sigma}_F\{\mathfrak{R}_K(z^u_h) + \mathfrak{R}_{\Delta_K}([\![z^u_h]\!])\} + \mathcal{D}^{uu}_F\{z^u_h\} + S^{u\sigma}_F([\![\mathfrak{R}_K(z^u_h) + \mathfrak{R}_{\Delta_K}([\![z^u_h]\!])]\!]) + S^{uu}_F([\![z^u_h]\!]) + R_F([\![z^u_h]\!])). \tag{4.29}$$

This readily yields the following.

PROPOSITION 4.1. *If the pair* (z^σ_h, z^u_h) *solves* (4.16)*, then* (4.23) *holds and* z^u_h *solves* (4.27)*. Conversely, if* z^u_h *solves* (4.27) *and if* z^σ_h *is defined by* (4.23)*, then the pair* (z^σ_h, z^u_h) *solves* (4.16)*.*

At this point, it is important to observe that owing to the presence of the nonlocal term $\mathfrak{R}_{\Delta_K}$ in the flux $\phi^u_{\partial K}$, the problem (4.27) couples the degrees of freedom for z^u_h in a given element to those in the neighboring elements and also to those in the neighbors of the neighbors. Let us assume that $S^{u\sigma}_F \equiv 0$ and, for simplicity, that Dirichlet boundary conditions are enforced so that $M^{\sigma u}_F = -\mathcal{D}^{\sigma u}$ and $M^{u\sigma}_F = \mathcal{D}^{u\sigma}$ (Neumann/Robin boundary conditions can be treated as well). Then, if R_F is defined so that for all $F \subset \partial K$,

$$R_F([\![z^u_h]\!]) + \mathcal{D}^{u\sigma}_F\{\mathfrak{R}_{\Delta_K}([\![z^u_h]\!])\} = 0, \tag{4.30}$$

the terms involving $\mathfrak{R}_{\Delta_K}([\![z^u_h]\!])$ are eliminated from (4.28)–(4.29). Owing to this elimination, problem (4.27) couples the degrees of freedom for z^u_h in a given element only to those in the neighboring elements. Using (4.25), it is readily verified that (4.30) holds if R_F is designed such that

$$R_F([\![z^u_h]\!]) = \tfrac{1}{2}\mathcal{D}^{u\sigma}_F \sum_{i=1}^{2} (\theta^2_{K_i(F)})^{-1} \sum_{F' \in \partial K_i(F)} r_{F'}(\psi_{F',K_i(F)}([\![z^u_h]\!]))|_F. \tag{4.31}$$

Finally, a further simplification occurs whenever $\mathcal{K}^{u\sigma} - (\nabla{\cdot}B)^* \equiv 0$ since, in this case, the term $\mathfrak{R}_{\Delta_K}([\![z^u_h]\!])$ needs not be evaluated to solve (4.27) for z^u_h; i.e., the reconstruction of z^σ_h from (4.23) can be performed as a postprocessing step.

5. Convergence analysis. In this section, we present the design criteria for the above DG method and perform the error analysis. The main results are Theorem 5.8, which estimates the error in the norm (5.10), and Theorem 5.14, which improves the L_u-estimate of the u-component of the error by means of a duality argument. Throughout this section, we assume the following:

- For all $k \in \{1, \dots, d\}$ and for all $K \in \mathcal{T}_h$, $\mathcal{B}^k \in [\mathcal{C}^{0,1}(K)]^{m_\sigma, m_u}$.
- The mesh family $\{\mathcal{T}_h\}_{h>0}$ is such that (4.1), (4.3), and (4.4) hold.
- The approximation spaces are defined according to (4.2), (4.5), and (4.6).

5.1. The design criteria for the boundary and interface operators. For all $F \in \mathcal{F}_h^\partial$, for all $v, w \in [L^2(F)]^{m_u}$, and for all $\tau \in [L^2(F)]^{m_\sigma}$, we assume that

(DG1) $M_F^{\sigma\sigma} = 0,$

(DG2) $M_F^{\sigma u} + (M_F^{u\sigma})^* = 0,$

(DG3) $(M_F^{uu}(v), v)_{L_u,F} \geq 0,$

(DG4) $|(M_F^{\sigma u}(v) - \mathcal{D}^{\sigma u} v, \tau)_{L_\sigma,F}| \lesssim h_F^{\frac{1}{2}} |v|_{M,F} \|\tau\|_{L_\sigma,F},$

(DG5) $|(M_F^{uu}(v) + \mathcal{D}^{uu} v, w)_{L_u,F}| \lesssim h_F^{-\frac{1}{2}} \|v\|_{L_u,F} |w|_{M,F},$

(DG6) $|(M_F^{uu}(v) - \mathcal{D}^{uu} v, w)_{L_u,F}| \lesssim h_F^{-\frac{1}{2}} |v|_{M,F} \|w\|_{L_u,F},$

(DG7) $\mathrm{Ker}(\mathcal{M} - \mathcal{D}) \subset \mathrm{Ker}(M_F - \mathcal{D}),$

(DG8) $\mathrm{Ker}(\mathcal{M}^t + \mathcal{D}) \subset \mathrm{Ker}(M_F^* + \mathcal{D}),$

where we have introduced the following seminorms:

$$\forall v \in U(h), \qquad |v|_M^2 = \sum_{F \in \mathcal{F}_h^\partial} |v|_{M,F}^2 \qquad \text{with} \qquad |v|_{M,F}^2 = (M_F^{uu}(v), v)_{L_u,F}. \tag{5.1}$$

For all $F \in \mathcal{F}_h^{\mathrm{i}}$, for all $v, w \in [L^2(F)]^{m_u}$, and for all $\tau \in [L^2(F)]^{m_\sigma}$, we assume that

(DG9) $S_F^{\sigma\sigma} = 0,$

(DG10) $S_F^{\sigma u} + (S_F^{u\sigma})^* = 0,$

(DG11) $(S_F^{uu}(v), v)_{L_u,F} \geq 0,$

(DG12) $|(S_F^{uu}(v), w)_{L_u,F}| \lesssim h_F^{-\frac{1}{2}} \|v\|_{L_u,F} |w|_{S,F},$

(DG13) $|(S_F^{uu}(v), w)_{L_u,F}| \lesssim h_F^{-\frac{1}{2}} |v|_{S,F} \|w\|_{L_u,F},$

(DG14) $|(S_F^{\sigma u}(v), \tau)_{L_\sigma,F}| \lesssim h_F^{\frac{1}{2}} |v|_{S,F} \|\tau\|_{L_\sigma,F},$

(DG15) $|(\mathcal{D}^{\sigma u} v, \tau)_{L_\sigma,F}| \lesssim h_F^{\frac{1}{2}} |v|_{S,F} \|\tau\|_{L_\sigma,F},$

(DG16) $|(\mathcal{D}^{uu} v, w)_{L_u,F}| \lesssim h_F^{-\frac{1}{2}} |v|_{S,F} \|w\|_{L_u,F},$

where we have introduced the following seminorms:

$$\forall v \in U(h), \qquad |v|_S^2 = \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} |v|_{S,F}^2 \qquad \text{with} \qquad |v|_{S,F}^2 = (S_F^{uu}(v), v)_{L_u,F}. \tag{5.2}$$

Finally, the design of the operators R_F is based on the following assumptions:

(DG17) $\forall z_h \in W_h, \quad \rho_h([\![z_h^u]\!], [\![z_h^u]\!]) \geq -\frac{1}{4}(|z_h^u|_J^2 + |z_h^u|_M^2),$

(DG18) $\forall (z, y_h) \in W(h) \times W_h, \quad \rho_h([\![z^u]\!], [\![y_h^u]\!]) \lesssim (|z^u|_J + |z^u|_M)(|y_h^u|_J + |y_h^u|_M),$

where $\rho_h([\![z^u]\!], [\![y^u]\!]) := \sum_{F \in \mathcal{F}_h} (R_F([\![z^u]\!]), [\![y^u]\!])_{L_u,F}$ and where for all $z^u \in U(h)$,

$$|z^u|_J^2 = \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} |z^u|_{J,F}^2 \qquad \text{with} \qquad |z^u|_{J,F} = |[\![z^u]\!]|_{S,F}. \tag{5.3}$$

Theorem 5.8 relies only on assumptions (DG1)–(DG5), (DG7), (DG9)–(DG12), (DG14)–(DG15), and (DG17)–(DG18), collectively referred to as (DG$^\flat$). The additional assumptions (DG6), (DG8), (DG13), and (DG16) are needed to prove Theorem 5.14. Assumptions (DG1)–(DG18) are collectively referred to as (DG$^\sharp$).

Remark 5.1. Assumptions (DG1)–(DG6) imply that for all $(\tau, v) \in [L^2(F)]^m$,

$$|v|_{M,F} \lesssim h_F^{-\frac{1}{2}} \|v\|_{L_u,F}, \tag{5.4}$$

$$|(M_F^{\sigma u}(v), \tau)_{L_\sigma,F}| \lesssim \|v\|_{L_u,F} \|\tau\|_{L_\sigma,F}, \tag{5.5}$$

$$|(M_F^{u\sigma}(\tau) + \mathcal{D}^{u\sigma}\tau, v)_{L_u,F}| \lesssim h_F^{\frac{1}{2}} |v|_{M,F} \|\tau\|_{L_\sigma,F}. \tag{5.6}$$

For instance, taking $v = w$ in (DG6) and using the fact that $\mathcal{D}^{uu}$ is bounded yields $|v|_{M,F}^2 \lesssim \|v\|_{L_u,F}^2 + h_F^{-\frac{1}{2}} |v|_{M,F} \|v\|_{L_u,F}$, whence (5.4) readily follows. Properties (5.4)–(5.6) will be used in what follows.

Remark 5.2. Assumptions (DG7) and (DG8) are consistency hypotheses which trivially hold if $M_F(z) = \mathcal{M} z|_F$. However, it is not always possible to make this simple choice because it is sometimes necessary to penalize the boundary values of the u-component of the unknown. For instance, when Dirichlet-like boundary conditions are enforced, i.e., $\mathcal{M}^{\sigma u} = -\mathcal{D}^{\sigma u}$, it may happen that $\mathcal{M}^{uu} = 0$ (see the examples discussed in section 3). In this circumstance, assumptions (DG4)–(DG6) cannot be satisfied if we set $M_F^{uu}(v) = \mathcal{M}^{uu} v|_F = 0$, since $|v|_{M,F} = 0$ for all $v \in [L^2(F)]^{m_u}$. Instead, it is necessary that M_F^{uu} scale like h_F^{-1}. The consistency hypotheses (DG7) and (DG8) then mean that the extra control required by (DG4)–(DG6) is compatible with the way boundary conditions are enforced (see also Remark 6.2 and section 6.1.1, section 6.2, and section 6.3 for examples).

While assumptions (DG$^\sharp$) are just what it takes to prove Theorems 5.8 and 5.14, it is simpler in practice to work with a simplified set of assumptions. These are summarized in the following lemmas. Lemma 5.1 is tailored for the case when Dirichlet-like boundary conditions are enforced, while Lemma 5.2 is tailored for the case when Neumann or Robin boundary conditions are enforced. For brevity, only the proof of Lemma 5.1 is detailed, the other two proofs being similar.

LEMMA 5.1 (Dirichlet-like BCs). *Assume $M_F^{\sigma\sigma} = 0$, $M_F^{\sigma u}(v) = -\mathcal{D}^{\sigma u} v$ for all $v \in [L^2(F)]^{m_u}$, $M_F^{u\sigma} = -(M_F^{\sigma u})^*$, M_F^{uu} is self-adjoint, and*

$$h_F |\mathcal{D}^{uu}| + h_F^{-1} (\mathcal{D}^{u\sigma} \mathcal{D}^{\sigma u})^{\frac{1}{2}} \lesssim M_F^{uu} \lesssim h_F^{-1} \mathcal{I}_{m_u}, \tag{5.7}$$

where $\mathcal{I}_{m_u}$ is the identity matrix in $\mathbb{R}^{m_u, m_u}$. Then, (DG1)–(DG6) *hold.*

Proof. Assumptions (DG1)–(DG3) are evident. To prove (DG4), observe that for every positive semidefinite matrix $\mathcal{Z} \in \mathbb{R}^{m_u, m_u}$ and for all $x \in \mathbb{R}^{m_u}$, $(\mathcal{Z}x, x) \le \|\mathcal{Z}^{1/2}\| (\mathcal{Z}^{1/2} x, x)$. Let $v \in [L^2(F)]^{m_u}$; upon observing that $\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u}$ is positive semidefinite, we apply the above result to derive

$$\begin{aligned} \|\mathcal{D}^{\sigma u} v\|_{L_\sigma,F} &= (\mathcal{D}^{\sigma u} v, \mathcal{D}^{\sigma u} v)_{L_\sigma,F}^{\frac{1}{2}} = (\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u} v, v)_{L_u,F}^{\frac{1}{2}} \\ &\lesssim ((\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}} v, v)_{L_u,F}^{\frac{1}{2}} \lesssim h_F^{\frac{1}{2}} |v|_{M,F}, \end{aligned}$$

whence (DG4) is readily inferred. To prove (DG5)–(DG6), let $v, w \in [L^2(F)]^{m_u}$. Then, $|(M_F^{uu}(v), w)_{L_u,F}| \lesssim |v|_{M,F} |w|_{M,F}$ and since $(\mathcal{D}^{uu})^2$ is positive semidefinite,

$$\|\mathcal{D}^{uu} v\|_{L_u,F} \lesssim (|\mathcal{D}^{uu}| v, v)_{L_u,F}^{\frac{1}{2}} \lesssim h_F^{-\frac{1}{2}} |v|_{M,F},$$

whence (DG5)–(DG6) are readily deduced. ∎

LEMMA 5.2 (Neumann–Robin BCs). *Assume $M_F^{\sigma\sigma} = 0$, $M_F^{\sigma u}(v) = \mathcal{D}^{\sigma u} v$ for all $v \in [L^2(F)]^{m_u}$, $M_F^{u\sigma} = -(M_F^{\sigma u})^*$, M_F^{uu} is self-adjoint, and*

$$h_F |\mathcal{D}^{uu}| \lesssim M_F^{uu} \lesssim h_F^{-1} \mathcal{I}_{m_u}. \tag{5.8}$$

Then, (DG1)–(DG6) *hold.*

LEMMA 5.3 (interface operator). *Assume* $S_F^{\sigma\sigma}=0$, $S_F^{u\sigma}=0$, $S_F^{\sigma u}=0$, S_F^{uu} *is self-adjoint, and*

$$h_F|\mathcal{D}^{uu}|+h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12}\lesssim S_F^{uu}\lesssim h_F^{-1}\mathcal{I}_{m_u}. \tag{5.9}$$

Then, (DG9)–(DG16) *hold.*

Remark 5.3. Conditions (5.7) and (5.9) generally imply that S_F^{uu} and M_F^{uu} are of order h_F^{-1}; this differs from the condition derived in [9], where S_F and M_F are of order 1. Roughly speaking, to be able to eliminate the discrete σ-component, it is necessary to have a stronger control of the interface jumps and of the boundary values of the discrete u-component.

5.2. The direct argument. To perform the error analysis we introduce the following two discrete norms on $W(h)$:

$$\|z\|_{h,A}^2=\|z^\sigma\|_{L_\sigma}^2+\|z^u\|_{L_u}^2+|z^u|_J^2+|z^u|_M^2+\sum_{K\in\mathcal{T}_h}\|Bz^u\|_{L_\sigma,K}^2, \tag{5.10}$$

$$\|z\|_{h,1}^2=\|z\|_{h,A}^2+\sum_{K\in\mathcal{T}_h}[h_K^{-2}\|z^u\|_{L_u,K}^2+h_K^{-1}\|z^u\|_{L_u,\partial K}^2+h_K\|z^\sigma\|_{L_\sigma,\partial K}^2]. \tag{5.11}$$

The norm $\|\cdot\|_{h,A}$ is used to measure the approximation error, and the norm $\|\cdot\|_{h,1}$ serves to measure the interpolation properties of the discrete space W_h. In this section, it is implicitly assumed that (DG$^\flat$) holds.

LEMMA 5.4 (L-coercivity). *For all h and for all $z_h=(z_h^\sigma,z_h^u)$ in W_h,*

$$\|z_h^\sigma\|_{L_\sigma}^2+\|z_h^u\|_{L_u}^2+|z_h^u|_J^2+|z_h^u|_M^2\lesssim a_h(z_h,z_h). \tag{5.12}$$

Proof. Proceeding as in the proof of Lemma 4.1 in [9] and using the skew-symmetry assumptions (DG2) and (DG10) yields for all $z_h\in W_h$,

$$\|z_h^\sigma\|_{L_\sigma}^2+\|z_h^u\|_{L_u}^2+|z_h^u|_J^2+\tfrac12|z_h^u|_M^2+\rho_h([\![z_h^u]\!],[\![z_h^u]\!])\lesssim a_h(z_h,z_h).$$

Then, the desired result follows from (DG17). ∎

LEMMA 5.5 (stability). *The following holds:*

$$\forall z_h\in W_h,\quad \|z_h\|_{h,A}\lesssim\sup_{y_h\in W_h\backslash\{0\}}\frac{a_h(z_h,y_h)}{\|y_h\|_{h,A}}. \tag{5.13}$$

Proof. Let $z_h=(z_h^\sigma,z_h^u)\in W_h\backslash\{0\}$ and set $\mathbb{S}=\sup_{y_h\in W_h\backslash\{0\}}\frac{a_h(z_h,y_h)}{\|y_h\|_{h,A}}$.

(1) Owing to Lemma 5.4, it is inferred that

$$\|z^\sigma\|_{L_\sigma}^2+\|z^u\|_{L_u}^2+|z^u|_J^2+|z^u|_M^2\lesssim a_h(z_h,z_h)\le\mathbb{S}\,\|z_h\|_{h,A}.$$

(2) Control of Bz_h^u. Let $K\in\mathcal{T}_h$. Denote by $\overline{\mathcal{B}_K^k}$ the mean-value of $\mathcal{B}^k$ over K; then,

$$\|\mathcal{B}^k-\overline{\mathcal{B}_K^k}\|_{[L^\infty(K)]^{m_\sigma,m_u}}\le h_K\|\mathcal{B}^k\|_{[\mathcal{C}^{0,1}(K)]^{m_\sigma,m_u}}. \tag{5.14}$$

Define the field π_h such that $\pi_h|_K=\sum_{k=1}^d\overline{\mathcal{B}_K^k}\partial_k z_h^u$. Set $\varpi_h=(\pi_h,0)$. It is clear that $\pi_h\in\Sigma_h$ since $p_u-1\le p_\sigma$; hence, $\varpi_h\in W_h$. Using (5.14), together with the inverse inequalities (4.3) and (4.4), leads, for all $F\subset\partial K$, to

$$\begin{cases}\|\pi_h\|_{L_\sigma,F}\lesssim h_F^{-\frac12}\|\pi_h\|_{L_\sigma,\mathcal{T}(F)}, & \text{if } F\in\mathcal{F}_h^\partial,\\ \|\{\pi_h\}\|_{L_\sigma,F}+\|[\![\pi_h]\!]\|_{L_\sigma,F}\lesssim h_F^{-\frac12}\|\pi_h\|_{L_\sigma,\mathcal{T}(F)} & \text{if } F\in\mathcal{F}_h^{\mathrm{i}},\end{cases} \tag{5.15}$$

$$\|\pi_h\|_{L_\sigma,K}\lesssim\|Bz_h^u\|_{L_\sigma,K}+\|z_h^u\|_{L_u,K}, \tag{5.16}$$

whence it is readily inferred that

$$\|\varpi_h\|_{h,A} = \|\pi_h\|_{L_\sigma} \lesssim \|z_h\|_{h,A}.$$

Furthermore, from the definition of a_h it follows that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \|Bz_h^u\|_{L_\sigma,K}^2 &= a_h(z_h,\varpi_h) + \sum_{K\in\mathcal{T}_h} (Bz_h^u, Bz_h^u - \pi_h)_{L_\sigma,K} \\ &\quad - (\mathcal{K}^{\sigma\sigma} z_h^\sigma + \mathcal{K}^{\sigma u} z_h^u, \pi_h)_{L_\sigma} - \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F^{\sigma u}(z_h^u) - \mathcal{D}^{\sigma u} z_h^u, \pi_h)_{L_\sigma,F} \\ &\quad + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}^{\sigma u} z_h^u\}, \{\pi_h\})_{L_\sigma,F} - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F^{\sigma u}([\![z_h^u]\!]), [\![\pi_h]\!])_{L_\sigma,F} \\ &= a_h(z_h,\varpi_h) + R_1 + R_2 + R_3 + R_4 + R_5,\end{aligned}$$

where R_1 to R_5 denote the second to sixth terms in the right-hand side. Proceeding as in the proof of Lemma 4.3 in [9] and using (DG4), (DG14), (DG15), the terms R_1–R_5 are bounded from above as follows:

$$\sum_{i=1}^5 |R_i| \lesssim (\|z_h^\sigma\|_{L_\sigma}^2 + \|z_h^u\|_{L_u}^2 + |z_h^u|_M^2 + |z_h^u|_J^2) + \gamma \sum_{K\in\mathcal{T}_h} \|Bz_h^u\|_{L_\sigma,K}^2,$$

where $\gamma > 0$ can be chosen as small as needed. Hence,

$$\sum_{K\in\mathcal{T}_h} \|Bz_h^u\|_{L_\sigma,K}^2 \lesssim a_h(z_h,\varpi_h) + a_h(z_h,z_h) \lesssim \mathbb{S}\,\|z_h\|_{h,A}.$$

(3) Collecting the above bounds yields $\|z_h\|_{h,A}^2 \lesssim \mathbb{S}\,\|z_h\|_{h,A}$, thereby completing the proof. ∎

LEMMA 5.6 (continuity). *The following holds:*

$$\forall (z,y_h) \in W(h)\times W_h, \quad a_h(z,y_h) \lesssim \|z\|_{h,1}\|y_h\|_{h,A}. \tag{5.17}$$

Proof. The main idea is to integrate by parts $a_h(z,y_h)$ by using the formal adjoint $\tilde{A}$. Proceeding as in the proof of Lemma 4.4 in [9] leads to

$$\begin{aligned}a_h(z,y_h) = &\sum_{K\in\mathcal{T}_h} [(Kz,z)_{L,K} + (z,\tilde{A}y_h)_{L,K}] + \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F(z) + \mathcal{D}z, y_h)_{L,F} \\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \tfrac{1}{2}([\![\mathcal{D}z]\!], [\![y_h]\!])_{L,F} + \rho_h([\![z^u]\!], [\![y_h^u]\!]) + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F([\![z]\!]), [\![y_h]\!])_{L,F}.\end{aligned} \tag{5.18}$$

Let R_1 to R_5 be the five terms in the right-hand side.

(1) Using the Cauchy–Schwarz inequality and inverse inequalities, we obtain

$$|R_1| \lesssim \sum_{K\in\mathcal{T}_h} \|z\|_{L,K}\|y_h\|_{L,K} + \|z^\sigma\|_{L_\sigma,K}\|By_h^u\|_{L_\sigma,K} + h_K^{-1}\|z^u\|_{L_u,K}\|y_h\|_{L,K}.$$

Hence, $|R_1| \lesssim \|z\|_{h,1}\|y_h\|_{h,A}$.

(2) For the second term, we have

$$\begin{aligned}|R_2| \le \tfrac{1}{2} \sum_{F\in\mathcal{F}_h^\partial} |&(M_F^{\sigma u}(z^u) + \mathcal{D}^{\sigma u} z^u, y_h^\sigma)_{L_\sigma,F} + (M_F^{uu}(z^u) + \mathcal{D}^{uu} z^u, y_h^u)_{L_u,F} \\ &+ (M_F^{u\sigma}(z^\sigma) + \mathcal{D}^{u\sigma} z^\sigma, y_h^u)_{L_u,F}|.\end{aligned}$$

Using (5.5), (DG5), the boundedness of $\mathcal{D}$, (5.6), and the inverse inequality (4.4), each term in the above equality is bounded as follows:

$$|(M_F^{\sigma u}(z^u)+\mathcal{D}^{\sigma u}z^u,y_h^\sigma)_{L_\sigma,F}|\lesssim\|z^u\|_{L_u,F}\|y_h^\sigma\|_{L_\sigma,F}\lesssim h_F^{-\frac12}\|z^u\|_{L_u,F}\|y_h^\sigma\|_{L_\sigma,\mathcal{T}(F)},$$
$$|(M_F^{uu}(z^u)+\mathcal{D}^{uu}z^u,y_h^u)_{L_u,F}|\lesssim h_F^{-\frac12}\|z^u\|_{L_u,F}|y_h^u|_{M,F},$$
$$|(M_F^{u\sigma}(z^\sigma)+\mathcal{D}^{u\sigma}z^\sigma,y_h^u)_{L_u,F}|\lesssim h_F^{\frac12}\|z^\sigma\|_{L_\sigma,F}|y_h^u|_{M,F}.$$

As a result, $|R_2|\lesssim\|z\|_{h,1}\|y_h\|_{h,A}$.

(3) For the third term, we have

$$|R_3|\le\tfrac12\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}|(\llbracket\mathcal{D}^{\sigma u}z^u\rrbracket,\llbracket y_h^\sigma\rrbracket)_{L_\sigma,F}+(\llbracket\mathcal{D}^{uu}z^u\rrbracket,\llbracket y_h^u\rrbracket)_{L_u,F}+(\llbracket\mathcal{D}^{u\sigma}z^\sigma\rrbracket,\llbracket y_h^u\rrbracket)_{L_u,F}|.$$

Using the boundedness of $\mathcal{D}$, the inverse inequality (4.4), and (DG15), each term in the above equality is bounded as follows:

$$|(\llbracket\mathcal{D}^{\sigma u}z^u\rrbracket,\llbracket y_h^\sigma\rrbracket)_{L_\sigma,F}|\lesssim\|\{z^u\}\|_{L_u,F}\|\llbracket y_h^\sigma\rrbracket\|_{L_\sigma,F}\lesssim h_F^{-\frac12}\|\{z^u\}\|_{L_u,F}\|y_h^\sigma\|_{L_\sigma,\mathcal{T}(F)},$$
$$|(\llbracket\mathcal{D}^{uu}z^u\rrbracket,\llbracket y_h^u\rrbracket)_{L_u,F}|\lesssim\|\{z^u\}\|_{L_u,F}\|\llbracket y_h^u\rrbracket\|_{L_u,F}\lesssim h_F^{-\frac12}\|\{z^u\}\|_{L_u,F}\|y_h^u\|_{L_u,\mathcal{T}(F)},$$
$$|(\llbracket\mathcal{D}^{u\sigma}z^\sigma\rrbracket,\llbracket y_h^u\rrbracket)_{L_u,F}|=|(\{z^\sigma\},\mathcal{D}_F^{\sigma u}\llbracket y_h^u\rrbracket)_{L_\sigma,F}|\lesssim h_F^{\frac12}\|\{z^\sigma\}\|_{L_\sigma,F}|y_h^u|_{J,F}.$$

As a result, $|R_3|\lesssim\|z\|_{h,1}\|y_h\|_{h,A}$.

(4) The fourth term is controlled using (DG18).

(5) For the fifth term, we have

$$|R_5|\le\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}|(S_F^{\sigma u}(\llbracket z^u\rrbracket),\llbracket y_h^\sigma\rrbracket)_{L_\sigma,F}+(S_F^{uu}(\llbracket z^u\rrbracket),\llbracket y_h^u\rrbracket)_{L_u,F}+(S_F^{u\sigma}(\llbracket z^\sigma\rrbracket),\llbracket y_h^u\rrbracket)_{L_u,F}|.$$

Using (DG12) and (DG14), together with the inverse inequality (4.4), each term in the above equality is bounded as follows:

$$|(S_F^{\sigma u}(\llbracket z^u\rrbracket),\llbracket y_h^\sigma\rrbracket)_{L_\sigma,F}|\lesssim h_F^{\frac12}|z^u|_{J,F}\|\llbracket y_h^\sigma\rrbracket\|_{L_\sigma,F}\lesssim|z^u|_{J,F}\|y_h^\sigma\|_{L_\sigma,\mathcal{T}(F)},$$
$$|(S_F^{uu}(\llbracket z^u\rrbracket),\llbracket y_h^u\rrbracket)_{L_u,F}|\lesssim h_F^{-\frac12}\|\llbracket z^u\rrbracket\|_{L_u,F}|y_h^u|_{J,F},$$
$$|(S_F^{u\sigma}(\llbracket z^\sigma\rrbracket),\llbracket y_h^u\rrbracket)_{L_u,F}|\lesssim h_F^{\frac12}\|\llbracket z^\sigma\rrbracket\|_{L_\sigma,F}|y_h^u|_{J,F}.$$

As a result, $|R_5|\lesssim\|z\|_{h,1}\|y_h\|_{h,A}$. The proof is complete. ∎

LEMMA 5.7 (consistency). *Let $z\in V\cap[H^1(\Omega)]^m$ solve (2.6) and let z_h solve (4.16). Then,*

$$\forall y_h\in W_h,\quad a_h(z-z_h,y_h)=0. \tag{5.19}$$

Proof. Let $y_h\in W_h$ and use (4.15) to evaluate $a_h(z,y_h)$. Since z solves (2.6), the first term in the right-hand side of (4.15) is equal to $(f,y_h)_L$. Owing to the consistency assumption (DG7), the second term in the right-hand side of (4.15) vanishes. Furthermore, since for all $F\in\mathcal{F}_h^{\mathrm{i}}$, $\{\mathcal{D}z\}=\mathcal{D}_F\llbracket z\rrbracket=0$ and $\llbracket z\rrbracket=0$ because $z\in[H^1(\Omega)]^m$, the third, fourth, and fifth terms in (4.15) are also zero. As a result, $a_h(z,y_h)=(f,y_h)_L=a_h(z_h,y_h)$, completing the proof. ∎

THEOREM 5.8 (convergence). *Let $z \in V \cap [H^1(\Omega)]^m$ solve (2.6) and let z_h solve (4.16). Then,*

$$\|z - z_h\|_{h,A} \lesssim \inf_{y_h \in W_h} \|z - y_h\|_{h,1}. \tag{5.20}$$

Proof. The proof follows from the second Strang lemma. ∎

Owing to the regularity of the mesh family $\{\mathcal{T}_h\}_{h>0}$, the following interpolation property holds: For all $z \in [H^{p_\sigma+1}(\Omega)]^{m_\sigma} \times [H^{p_u+1}(\Omega)]^{m_u}$, there is $y_h \in W_h$ satisfying

$$\|z - y_h\|_{h,1} \lesssim (h^{p_\sigma+1} + h^{p_u})\big(\|z^\sigma\|_{[H^{p_\sigma+1}(\Omega)]^{m_\sigma}} + \|z^u\|_{[H^{p_u+1}(\Omega)]^{m_u}}\big). \tag{5.21}$$

Since $p_u - 1 \le p_\sigma$, the above interpolation error is of order h^{p_u}.

COROLLARY 5.9. *Let $z \in [H^{p_\sigma+1}(\Omega)]^{m_\sigma} \times [H^{p_u+1}(\Omega)]^{m_u}$ solve (2.6) and let z_h solve (4.16). Then,*

$$\|z - z_h\|_{h,A} \lesssim h^{p_u}\big(\|z^\sigma\|_{[H^{p_\sigma+1}(\Omega)]^{m_\sigma}} + \|z^u\|_{[H^{p_u+1}(\Omega)]^{m_u}}\big). \tag{5.22}$$

Remark 5.4. For both the σ- and the u-component of the solution, the error estimate in the L^2-norm is $\mathcal{O}(h^{p_u})$. If $p_\sigma = p_u := p$, this result is suboptimal when compared with that obtained using the DG method analyzed in [9], which yields $\mathcal{O}(h^{p+\frac{1}{2}})$ error estimates. The reason for this slight optimality loss is that in the present method the interface jumps of the σ-component are not controlled to allow for this component to be locally eliminated, the consequence being that the jumps on the u-component must be penalized with an $\mathcal{O}(h^{-1})$ weight. If $p_\sigma = p_u - 1$, (5.22) is still suboptimal for the u-component but is optimal in the L^2-norm for the σ-component.

Finally, when the exact solution z is only in the graph space W, i.e., when z is not in $[H^1(\Omega)]^m$ so that $a_h(z,\cdot)$ may not be meaningful, we use a density argument to infer the convergence of the DG approximation. For $z \in W + W_h$, define the norm

$$\|z\|_{W^-} = \|z\|_L + \left(\sum_{K \in \mathcal{T}_h} \|Bz^u\|^2_{L_\sigma,K}\right)^{\frac{1}{2}}. \tag{5.23}$$

Observe that $\|z\|_{W^-} \le \|z\|_{h,A}$.

COROLLARY 5.10. *Assume that there is $\gamma > 0$ such that $[H^{\gamma+1}(\Omega)]^m \cap V$ is dense in V. Let z solve (2.6) and let z_h solve (4.16). Then,*

$$\lim_{h\to 0} \|z - z_h\|_{W^-} = 0. \tag{5.24}$$

Proof. Let $\epsilon > 0$. There is $z_\epsilon \in [H^{\gamma+1}(\Omega)]^m \cap V$ such that $\|z - z_\epsilon\|_W \le \frac{\epsilon}{2}$. Let $z_{\epsilon h}$ be the unique solution in W_h such that $a_h(z_{\epsilon h}, y_h) = (Tz_\epsilon, y_h)_L$ for all $y_h \in W_h$. From the regularity of z_ϵ together with Theorem 5.8 and Corollary 5.9, it is inferred that $\lim_{h\to 0} \|z_{\epsilon h} - z_\epsilon\|_{h,A} = 0$. Furthermore, using the discrete inf-sup condition (5.13) yields

$$\begin{aligned}\|z_{\epsilon h} - z_h\|_{W^-} &\lesssim \sup_{y_h \in W_h \setminus \{0\}} \frac{a_h(z_{\epsilon h}, y_h) - a_h(z_h, y_h)}{\|y_h\|_{h,A}} = \sup_{y_h \in W_h \setminus \{0\}} \frac{(T(z_\epsilon - z), y_h)_L}{\|y_h\|_{h,A}} \\ &\le \|T(z_\epsilon - z)\|_L \sup_{y_h \in W_h \setminus \{0\}} \frac{\|y_h\|_L}{\|y_h\|_{h,A}} \le \|z - z_\epsilon\|_W \le \frac{\epsilon}{2},\end{aligned}$$

where we have used the fact that for all $y_h \in W_h$, $a_h(z_h, y_h) = (Tz, y_h)_L$. Finally, using the triangle inequality $\|z - z_h\|_{W^-} \le \|z - z_\epsilon\|_{W^-} + \|z_\epsilon - z_{\epsilon h}\|_{W^-} + \|z_{\epsilon h} - z_h\|_{W^-}$, we deduce that $\limsup_{h\to 0} \|z - z_h\|_{W^-} \le \epsilon$. ∎

5.3. The duality argument. We now improve the error estimate on the L^2-norm of the u-component of the solution by using a duality argument. In this section, it is implicitly assumed that (DG$^\sharp$) holds.

Let z solve (2.6) and let z_h solve (4.16). Let $\psi := (\psi^\sigma, \psi^u) \in V^*$ solve

$$\tilde{T}\psi = (0, z^u - z_h^u). \tag{5.25}$$

We assume that the above problem yields (elliptic) regularity; i.e., ψ^u is in $[H^2(\Omega)]^{m_u}$, ψ^σ is in $[H^1(\Omega)]^{m_\sigma}$, and the following uniform bound holds:

$$\|\psi^u\|_{[H^2(\Omega)]^{m_u}} + \|\psi^\sigma\|_{[H^1(\Omega)]^{m_\sigma}} \lesssim \|z^u - z_h^u\|_{L_u}. \tag{5.26}$$

LEMMA 5.11. *Under the above hypotheses, the following holds:*

$$a_h(y, \psi) = (y^u, z^u - z_h^u)_{L_u} \quad \forall y \in W(h). \tag{5.27}$$

Proof. Let $y \in W(h)$. By integrating by parts (i.e., using (5.18)) and using the fact that ψ is continuous across interfaces, we obtain

$$a_h(y, \psi) = \sum_{K \in \mathcal{T}_h} (y, \tilde{T}\psi)_{L,K} + \sum_{F \in \mathcal{F}_h^\partial} \tfrac{1}{2}(M_F(y) + \mathcal{D}y, \psi)_{L,F}.$$

Since $\psi \in V^* \cap [H^1(\Omega)]^m$, (DG8) implies $(M_F(y) + \mathcal{D}y, \psi)_{L,F} = 0$ for all $F \in \mathcal{F}_h^\partial$. The conclusion is straightforward since ψ solves (5.25). ∎

To avoid lengthy technicalities, we introduce the following norms:

$$\|y^\sigma\|_{h,\widetilde{1}} = \left(\sum_{K \in \mathcal{T}_h} [h_K^2 \|y^\sigma\|^2_{[H^1(K)]^{m_\sigma}} + h_K \|y^\sigma\|^2_{L_\sigma, \partial K}] \right)^{\frac{1}{2}}, \tag{5.28}$$

$$\|y\|_{h,A^+} = \|y\|_{h,A} + \|y^\sigma\|_{h,\widetilde{1}}, \tag{5.29}$$

$$\|y\|_{h,1^+} = \|y\|_{h,1} + \|y^\sigma\|_{h,\widetilde{1}}. \tag{5.30}$$

The DG method converges optimally in the $\|\cdot\|_{h,A^+}$-norm as stated in the following.

COROLLARY 5.12. *Let $z \in V \cap [H^1(\Omega)]^m$ solve* (2.6) *and let z_h solve* (4.16). *Then,*

$$\|z - z_h\|_{h,A^+} \lesssim \inf_{y_h \in W_h} \|z - y_h\|_{h,1^+}. \tag{5.31}$$

Proof. Let y_h be an arbitrary element in W_h. Using inverse inequalities yields

$$\begin{aligned}
\|z^\sigma - z_h^\sigma\|_{h,\widetilde{1}} &\le \|z^\sigma - y_h^\sigma\|_{h,\widetilde{1}} + \|y_h^\sigma - z_h^\sigma\|_{h,\widetilde{1}} \lesssim \|z^\sigma - y_h^\sigma\|_{h,\widetilde{1}} + \|y_h^\sigma - z_h^\sigma\|_{L_\sigma} \\
&\le \|z^\sigma - y_h^\sigma\|_{h,\widetilde{1}} + \|y_h^\sigma - z^\sigma\|_{L_\sigma} + \|z^\sigma - z_h^\sigma\|_{L_\sigma} \\
&\le \|z^\sigma - y_h^\sigma\|_{h,\widetilde{1}} + \|z - y_h\|_{h,A} + \|z - z_h\|_{h,A} \\
&\lesssim \|z - y_h\|_{h,A^+} + \|z - z_h\|_{h,A}.
\end{aligned}$$

Hence, using the above inequality along with (5.20) leads to

$$\|z - z_h\|_{h,A^+} \lesssim \|z - y_h\|_{h,A^+} + \|z - y_h\|_{h,1} \lesssim \|z - y_h\|_{h,1^+}.$$

That concludes the proof since y_h is arbitrary in W_h. ∎

LEMMA 5.13 (continuity). *Assume that for all $K \in \mathcal{T}_h$ and for all $y \in W(h)$,*

$$\|Cy^u\|_{L_u,K} \lesssim \|By^u\|_{L_\sigma,K} + \|y^u\|_{L_u,K}. \tag{5.32}$$

Then, the following holds:

$$\forall (r,y)\in W(h)\times W(h),\quad a_h(r,y)\lesssim \|r\|_{h,A^+}\|y\|_{h,1}. \tag{5.33}$$

Proof. Let us bound all the terms in the right-hand side of (4.15).

(1) For the first term, say R_1, we proceed as follows:

$$\begin{aligned}|(Tr,y)_{L,K}| &\le |(Kr,y)_{L,K}| + |(Br^u,y^\sigma)_{L_\sigma,K}| + |(B^\dagger r^\sigma + Cr^u,y^u)_{L_u,K}|\\ &\lesssim \|r\|_{L,K}\|y\|_{L,K} + \|Br^u\|_{L_\sigma,K}\|y\|_{L,K} + \|r^\sigma\|_{[H^1(K)]^{m_\sigma}}\|y^u\|_{L_u,K}\\ &\lesssim (\|r\|_{L,K}^2 + \|Br^u\|_{L_\sigma,K}^2 + h_K^2\|r^\sigma\|_{[H^1(K)]^{m_\sigma}}^2)^{\frac12}(\|y\|_{L,K}^2 + h_K^{-2}\|y^u\|_{L_u,K}^2)^{\frac12},\end{aligned}$$

where (5.32) has been used to bound $\|Cr^u\|$. Hence, $|R_1|\lesssim \|r\|_{h,A^+}\|y\|_{h,1}$.

(2) To bound the second term, say R_2, use (DG4), (DG6), (5.5), and the boundedness of $\mathcal{D}$ to infer

$$\begin{aligned}&|(M_F^{\sigma u}(r^u) - \mathcal{D}^{\sigma u}r^u, y^\sigma)_{L_\sigma,F}| \lesssim |r^u|_{M,F}\, h_F^{\frac12}\|y^\sigma\|_{L_\sigma,F},\\ &|(M_F^{uu}(r^u) - \mathcal{D}^{uu}r^u, y^u)_{L_u,F}| \lesssim |r^u|_{M,F}\, h_F^{-\frac12}\|y^u\|_{L_u,F},\\ &|(M_F^{u\sigma}(r^\sigma) - \mathcal{D}^{u\sigma}r^\sigma, y^u)_{L_u,F}| \lesssim \|r^\sigma\|_{L_\sigma,F}\|y^u\|_{L_u,F} \lesssim h_F^{\frac12}\|r^\sigma\|_{L_\sigma,F}\, h_F^{-\frac12}\|y^u\|_{L_u,F}.\end{aligned}$$

As a result, $|R_2|\lesssim \|r\|_{h,A^+}\|y\|_{h,1}$.

(3) To bound the third term, say R_3, use (DG15), (DG16), and the boundedness of $\mathcal{D}$ to infer

$$\begin{aligned}&|(\{\mathcal{D}^{\sigma u}r^u\},\{y^\sigma\})_{L_\sigma,F}| = |2(\mathcal{D}^{\sigma u}_{\partial K_1(F)}[\![r^u]\!],\{y^\sigma\})_{L_\sigma,F}| \lesssim |r^u|_{J,F}\, h_F^{\frac12}\|\{y^\sigma\}\|_{L_\sigma,F},\\ &|(\{\mathcal{D}^{uu}r^u\},\{y^u\})_{L_u,F}| = |2(\mathcal{D}^{uu}_{\partial K_1(F)}[\![r^u]\!],\{y^u\})_{L_u,F}| \lesssim |r^u|_{J,F}\, h_F^{-\frac12}\|\{y^u\}\|_{L_u,F},\\ &|(\{\mathcal{D}^{u\sigma}r^\sigma\},\{y^u\})_{L_u,F}| \lesssim \|[\![r^\sigma]\!]\|_{L_\sigma,F}\|\{y^u\}\|_{L_u,F} \lesssim h_F^{\frac12}\|[\![r^\sigma]\!]\|_{L_\sigma,F}\, h_F^{-\frac12}\|\{y^u\}\|_{L_u,F}.\end{aligned}$$

These bounds yield $|R_3|\lesssim \|r\|_{h,A^+}\|y\|_{h,1}$.

(4) To bound the fourth term, use (DG18).

(5) To bound the fifth term, say R_5, use (DG10), (DG13), and (DG14) to infer

$$\begin{aligned}&|(S_F^{\sigma u}([\![r^u]\!]),[\![y^\sigma]\!])_{L_\sigma,F}| \lesssim |r^u|_{J,F}\, h_F^{\frac12}\|[\![y^\sigma]\!]\|_{L_\sigma,F},\\ &|(S_F^{uu}([\![r^u]\!]),[\![y^u]\!])_{L_u,F}| \lesssim |r^u|_{J,F}\, h_F^{-\frac12}\|[\![y^u]\!]\|_{L_u,F},\\ &|(S_F^{u\sigma}([\![r^\sigma]\!]),[\![y^u]\!])_{L_u,F}| \lesssim h_F^{\frac12}\|[\![r^\sigma]\!]\|_{L_\sigma,F}\,|y^u|_{J,F}.\end{aligned}$$

Hence, $|R_5|\lesssim \|r\|_{h,A^+}\|y\|_{h,1}$. The proof is complete. ∎

THEOREM 5.14 (convergence). *Let $z\in V\cap[H^1(\Omega)]^m$ solve* (2.6) *and let z_h solve* (4.16). *Assume elliptic regularity, i.e.,* (5.26)*, and that* (5.32) *holds. Then,*

$$\|z^u - z_h^u\|_{L_u} \lesssim h \inf_{y_h\in W_h}\|z-y_h\|_{h,1^+}. \tag{5.34}$$

Proof. Using $z-z_h$ as test function in (5.27) we infer $a_h(z-z_h,\psi) = \|z^u - z_h^u\|_{L_u}^2$. Then, using the consistency property stated in Lemma 5.7, this yields for all $\psi_h\in W_h$, $a_h(z-z_h,\psi-\psi_h) = \|z^u - z_h^u\|_{L_u}^2$. Lemma 5.13 in turn implies

$$\|z^u - z_h^u\|_{L_u}^2 \lesssim \|z-z_h\|_{h,A^+}\|\psi-\psi_h\|_{h,1}\quad \forall \psi_h\in W_h.$$

Then, using the elliptic regularity (5.26) and the fact that $p_u \geq 1$ leads to

$$\begin{aligned}\|z^u - z_h^u\|_{L_u}^2 &\lesssim \|z - z_h\|_{h,A^+} \inf_{\psi_h \in W_h} \|\psi - \psi_h\|_{h,1} \\ &\lesssim h\|z - z_h\|_{h,A^+}(\|\psi^u\|_{[H^2(\Omega)]^{m_u}} + \|\psi^\sigma\|_{[H^1(\Omega)]^{m_\sigma}}) \\ &\lesssim h\|z - z_h\|_{h,A^+}\|z^u - z_h^u\|_{L_u}.\end{aligned}$$

The conclusion follows readily using Corollary 5.12. ∎

Remark 5.5. Stability and convergence in the $\|\cdot\|_{h,A^+}$-norm could have been proved directly by adding the quantity $(\sum_{K\in\mathcal{T}_h} h_K^2\|B^\dagger y^\sigma + Cy^u\|_{L_u,K}^2)^{\frac{1}{2}}$ in the definition of the $\|\cdot\|_{h,A}$-norm, but this significantly lengthens the proof of Lemma 5.5. With this modification of the $\|\cdot\|_{h,A}$-norm, hypothesis (5.32) can be removed. However, this appears to be a minor issue since (5.32) holds for all the two-field Friedrichs' systems presented in section 3.

6. Applications. In this section we apply the DG method designed in section 4 and analyzed in section 5 to the Friedrichs' systems presented in section 3.

6.1. Advection-diffusion-reaction. We describe various DG methods that can be used to approximate the advection-diffusion-reaction equation introduced in section 3.1 and in which the σ-component of the unknown can be eliminated locally. Comparisons with the unified approached developed by Arnold et al. [1] are presented to illustrate the fact that the present DG method generalizes some of the DG methods that have been previously developed in the literature for the Poisson equation.

6.1.1. A first example: The LDG method. Consider first Dirichlet boundary conditions. Owing to (3.5) and (3.6), the integral representations (2.15) and (2.17) hold with the $\mathbb{R}^{d+1,d+1}$-valued boundary fields

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & n \\ \hdashline n^t & \beta\cdot n \end{array}\right] \qquad \text{and} \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^t & 0 \end{array}\right], \tag{6.1}$$

where n is the unit outward normal to $\partial\Omega$. Let $\varsigma > 0$ and $\eta > 0$ (these design parameters can vary from face to face). For all $F \in \mathcal{F}_h$, set $R_F \equiv 0$ and

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -n_F \\ \hdashline n_F^t & \varsigma h_F^{-1} \end{array}\right], \qquad \mathcal{S}_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & \eta h_F^{-1} \end{array}\right], \tag{6.2}$$

and define for all $y \in [L^2(F)]^{d+1}$, $M_F(y) = \mathcal{M}_F y$ and $S_F(y) = \mathcal{S}_F y$.

LEMMA 6.1. *Let M_F, S_F, and R_F be defined as above. Then, properties* (DG$^\sharp$) *hold.*

Proof. The consistency properties (DG7) and (DG8) are readily verified. Properties (DG17)–(DG18) are evident. The remaining properties are direct consequences of Lemmata 5.1 and 5.3. ∎

Remark 6.1. Let $\delta \in \mathbb{R}^d$. A slightly more general choice for the interface operator consists of setting for all $F \in \mathcal{F}_h^{\mathrm{i}}$, $\mathcal{S}_F^{\sigma u} = (\delta\cdot n_F)n_F$, where n_F is any of the two unit normal vectors to F. This choice leads to the so-called LDG method of Cockburn and Shu [7] as considered in the unified approach of [1] for the Poisson equation.

When Neumann and Robin boundary conditions are enforced, the integral representation (2.17) holds for the $\mathbb{R}^{d+1,d+1}$-valued boundary field

$$\mathcal{M} = \left[\begin{array}{c:c} 0 & n \\ \hdashline -n^t & 2\varrho + \beta\cdot n \end{array}\right]. \tag{6.3}$$

Assume that $\varrho \geq (\beta{\cdot}n)^-$, the negative part of $\beta{\cdot}n$ (this is not restrictive in practice since the usual Robin condition at an inflow boundary uses $\varrho = -\beta{\cdot}n \geq 0$). For all $F \in \mathcal{F}_h$, set $R_F \equiv 0$ and

$$\mathcal{M}_F = \left[\begin{array}{c|c} 0 & n_F \\ \hline -n_F^t & 2\varrho + \beta{\cdot}n_F \end{array}\right], \qquad \mathcal{S}_F = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \eta h_F^{-1} \end{array}\right], \tag{6.4}$$

and for all $y \in [L^2(F)]^{d+1}$, define $M_F(y) = \mathcal{M}_F y$ and $S_F(y) = \mathcal{S}_F y$. Then, it is easily verified that (5.8) holds. Hence, Lemma 5.2 implies that assumptions (DG1)–(DG6) hold. Moreover, the consistency assumptions (DG7) and (DG8) trivially hold. Of course, (DG9)–(DG16) hold since the definition of $\mathcal{S}_F$ is independent of the type of boundary condition. Finally, (DG17)–(DG18) are evident since $R_F \equiv 0$.

Remark 6.2. Observe that the scalings of the block $\mathcal{M}_F^{uu}$ are radically different whether Dirichlet or Robin/Neumann boundary conditions are enforced.

6.1.2. Comparison with other methods. In this section we restrict the setting to the equation $u - \Delta u = f$ and to homogeneous Dirichlet boundary conditions so as to make comparisons with the unified approach developed in [1], where it is shown that most of the DG methods amount to solving the following problem:

$$\begin{cases} \text{Seek } z_h = (\sigma_h, u_h) \in W_h \text{ such that } \forall y_h \in [\mathbb{P}_{p_\sigma}(K)]^d \times \mathbb{P}_{p_u}(K), \\ (z_h, \tilde{T} y_h)_{L,K} + (\widehat{\phi}_{\partial K}(z_h), y_h)_{L,\partial K} = (f, y_h)_{L,K}, \end{cases} \tag{6.5}$$

where the so-called numerical fluxes $\widehat{\phi}_{\partial K}(z_h)$ depend on the method under consideration. In view of (4.17) and (4.19), the link between the present formalism and that of [1] is based on the identification $\widehat{\phi}_{\partial K}(z_h)|_F = \phi_{\partial K}(z_h)|_F$. For the purpose of comparison, we restrict ourselves to boundary and interface operators such that for all $F \in \mathcal{F}_h$, for all $v \in L^2(F)$, and for all $\tau \in [L^2(F)]^d$,

$$M_F^{\sigma u}(v) = -n_F v, \qquad M_F^{u\sigma}(\tau) = \tau{\cdot}n_F, \tag{6.6}$$

$$S_F^{\sigma u}(v) = 0, \qquad S_F^{u\sigma}(\tau) = 0. \tag{6.7}$$

Therefore, the methods that can be constructed from this set of assumptions differ only in the design of M_F^{uu}, S_F^{uu}, and R_F. We set $\widehat{\phi}_{\partial K}(z_h) = (\widehat{u}_K n_K, \widehat{\sigma}_K{\cdot}n_K)$ (note that $\widehat{u}_K$ is $\mathbb{R}$-valued, $\widehat{\sigma}_K$ is $\mathbb{R}^d$-valued, and the sign convention we use herein for σ_h and $\widehat{\sigma}_K$ is opposite to that in [1]). Then, the above identification of the fluxes is possible if the DG method under consideration is such that

$$\widehat{\phi}_{\partial K}(z_h) = \begin{cases} (0, \sigma_h{\cdot}n_F + \frac{1}{2} M_F^{uu}(u_h) + R_F(u_h)) & \text{if } F \in \mathcal{F}_h^\partial, \\ (\{u_h\}\, n_K, \{\sigma_h\}\cdot n_K + n_F{\cdot}n_K (S_F^{uu}([\![u_h]\!]) + R_F([\![u_h]\!]))) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}. \end{cases} \tag{6.8}$$

The DG methods that belong to this class are those from [3, 5, 4, 6] together with that of [7] already discussed above. Observe that in this setting, the local flux reconstruction formula (4.23) takes the form

$$\forall K \in \mathcal{T}_h, \quad z_h^\sigma|_K = -\nabla z_h^u|_K + \sum_{F \subset \partial K} r_F([\![z_h^u]\!] n_F). \tag{6.9}$$

Comparison with the method of Brezzi et al. The method described by Brezzi et al. [6] (see also [1]) is such that

$$\widehat{\phi}_{\partial K}(z_h) = \begin{cases} (0, \sigma_h{\cdot}n_F + \frac{1}{2}\varsigma r_F(u_h n_F){\cdot}n_F) & \text{if } F \in \mathcal{F}_h^\partial, \\ (\{u_h\}\, n_K, \{\sigma_h\}\cdot n_K + \eta\, \{r_F([\![u_h]\!] n_F)\}\cdot n_K) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}, \end{cases} \tag{6.10}$$

where ς and η are positive constants. This amounts to specifying M_F^{uu}, S_F^{uu}, and R_F such that for all $v \in L^2(F)$,

$$M_F^{uu}(v) = \varsigma r_F(vn_F)\cdot n_F, \qquad S_F^{uu}(v) = \eta \{r_F(vn_F)\}\cdot n_F, \qquad R_F(v) \equiv 0. \tag{6.11}$$

The operator r_F is endowed with the following property.

LEMMA 6.2. *For all $F \in \mathcal{F}_h$ and for all $\tau_h \in [\mathbb{P}_{p_\sigma}(F)]^d$,*

$$h_F^{-\frac{1}{2}} \|\tau_h\|_{L_\sigma,F} \lesssim \|r_F(\tau_h)\|_{L_\sigma,\mathcal{T}(F)} \lesssim h_F^{-\frac{1}{2}} \|\tau_h\|_{L_\sigma,F}. \tag{6.12}$$

This lemma and the definition of r_F imply that for all $F \in \mathcal{F}_h$ and for all $v_h \in \mathbb{P}_{p_u}(F)$,

$$h_F^{-1}\|v_h\|_{L_u,F}^2 \lesssim (\{r_F(v_h n_F)\}\cdot n_F, v_h)_{L_u,F} \lesssim h_F^{-1}\|v_h\|_{L_u,F}^2. \tag{6.13}$$

These inequalities are just what is takes to prove that if the boundary and interface operators are defined using (6.6), (6.7), and (6.11), properties (DG$^\sharp$) hold. Therefore, the conclusions of Theorems 5.8 and 5.14 hold.

Comparison with the IP method. Let ς and η be two positive constants. The IP method of Baker [3] (see also Arnold [2]) is such that the flux is defined by

$$\widehat{\phi}_{\partial K}(z_h) = \begin{cases} (0, \sigma_h\cdot n_F + \frac{1}{2}\frac{\varsigma}{h_F} u_h + \rho_F(\llbracket u_h \rrbracket)\cdot n_F) & \text{if } F \in \mathcal{F}_h^\partial, \\ (\{u_h\} n_K, \{\sigma_h\}\cdot n_K + \frac{\eta}{h_F}\llbracket u_h \rrbracket n_F\cdot n_K + \rho_F(\llbracket u_h \rrbracket)\cdot n_K) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}, \end{cases} \tag{6.14}$$

where the operator $\rho_F : L^2(\Delta_F) \longrightarrow L^2(F)$ is defined by

$$\rho_F(v) = -\sum_{F' \in \Delta_F} \{r_{F'}(vn_{F'})\}, \tag{6.15}$$

and $\Delta_F = \{F' \in \mathcal{F}_h;\ \exists K' \in \mathcal{T}_h,\ F \cup F' \subset \partial K'\}$. This method fits the present framework if we set

$$M_F^{uu}(v) = \varsigma h_F^{-1} v, \qquad S_F^{uu}(v) = \eta h_F^{-1} v, \qquad R_F(v) = \rho_F(v)\cdot n_F. \tag{6.16}$$

Using Lemma 6.2, it is readily seen that (DG18) holds and that (DG17) holds if the design parameters ς and η are large enough. Therefore, the conclusions of Theorems 5.8 and 5.14 hold for the IP method. Note that the expression (4.31) derived for R_F in the general setting of two-field Friedrichs' systems reduces to (6.16) for the Poisson problem with Dirichlet boundary conditions.

Comparison with the methods of Bassi et al. The method proposed by Bassi and Rebay [5] corresponds to the choice of $M_F^{uu} \equiv 0$, $S_F^{uu} \equiv 0$, and $R_F \equiv 0$. Our analysis needs to be revised to account for this situation. Obviously, the L^2-coercivity still holds in the form $\|y\|_L^2 \lesssim a_h(y,y)$ for all $y \in W(h)$. Moreover, one easily derives the following continuity estimate: For all $(y, y_h) \in W(h) \times W_h$,

$$|a_h(y, y_h)| \lesssim \left(\sum_{K \in \mathcal{T}_h} [\|Ty\|_{L,K}^2 + h_K^{-1}\|y\|_{L,\partial K}^2] \right)^{\frac{1}{2}} \|y_h\|_L. \tag{6.17}$$

Then, provided $p_\sigma = p_u := p$, the second Strang lemma implies $\|z - z_h\|_L \lesssim h^p \|z\|_{[H^{p+1}(\Omega)]^m}$. Although this estimate is not optimal, it shows that the method of Bassi and Rebay is (possibly nonoptimally) convergent. Finally, the method proposed by Bassi et al. [4] fits the present framework by defining the operators

$$M_F^{uu}(v) = \varsigma r_F(vn_F)\cdot n_F, \qquad S_F^{uu}(v) = \eta \{r_F(vn_F)\}\cdot n_F, \tag{6.18}$$

and the operator R_F as in the IP method, i.e., (6.16). By using what has been shown above for the method of Brezzi et al. and the IP method, it is clear that the conclusions of Theorems 5.8 and 5.14 hold in this case also, provided ς and η are large enough.

6.2. Linear continuum mechanics. Consider the linear continuum mechanics equations introduced in section 3.2 and let us describe a DG method where the $(\overline{\sigma}, p)$-component of the unknown can be eliminated locally. Owing to (3.11) and (3.12), the integral representations (2.15) and (2.17) hold with the $\mathbb{R}^{m,m}$-valued boundary fields (recall that $m = d^2 + 1 + d$)

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & \mathcal{H} \\ \hdashline \mathcal{H}^t & 0 \end{array}\right] \qquad \text{and} \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & -\mathcal{H} \\ \hdashline \mathcal{H}^t & 0 \end{array}\right], \tag{6.19}$$

where $\mathcal{H} = \sum_{k=1}^{d} n_k(\mathcal{E}^k, 0)^t \in \mathbb{R}^{d^2+1,d}$. Observe that for all $\xi \in \mathbb{R}^d$, $\mathcal{H}\xi = (-\frac{1}{2}(n\otimes\xi + \xi\otimes n), 0)$. Let $\varsigma > 0$ and $\eta > 0$ (these design parameters can vary from face to face). For all $F \in \mathcal{F}_h$, set $R_F \equiv 0$ and

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -\mathcal{H}_F \\ \hdashline \mathcal{H}_F^t & \varsigma h_F^{-1}\mathcal{I}_d \end{array}\right], \qquad \mathcal{S}_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & \eta h_F^{-1}\mathcal{I}_d \end{array}\right], \tag{6.20}$$

where $\mathcal{H}_F$ is defined as $\mathcal{H}$ with n_F substituting for n. Define, for all $y \in [L^2(F)]^m$, $M_F(y) = \mathcal{M}_F y$ and $S_F(y) = \mathcal{S}_F y$. Then, using Lemmata 5.1 and 5.3, one readily verifies that properties (DG$^\sharp$) hold. An IP-like method can be derived as well.

6.3. Simplified MHD. Consider the simplified MHD equations introduced in section 3.3 and let us describe a DG method where the H-component of the unknown can be eliminated locally (the derivation of a DG method where the E-component of the unknown can be eliminated locally is similar). To recover the notation of section 5, set $\sigma \equiv H$ and $u \equiv E$. Owing to (3.16) and (3.17), the integral representations (2.15) and (2.17) hold with the $\mathbb{R}^{6,6}$-valued boundary fields

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & \mathcal{N} \\ \hdashline \mathcal{N}^t & 0 \end{array}\right] \qquad \text{and} \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & -\mathcal{N} \\ \hdashline \mathcal{N}^t & 0 \end{array}\right], \tag{6.21}$$

where $\mathcal{N} = \sum_{k=1}^{3} n_k\mathcal{R}^k$, and the $\mathbb{R}^{3,3}$-valued fields $\mathcal{R}^1$, $\mathcal{R}^2$, and $\mathcal{R}^3$ are defined in section 3.3. Observe that for all $\xi \in \mathbb{R}^3$, $\mathcal{N}\xi = n\times\xi$. Let $\varsigma > 0$ and $\eta > 0$ (these design parameters can vary from face to face). For all $F \in \mathcal{F}_h$, set $R_F \equiv 0$ and

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -\mathcal{N}_F \\ \hdashline \mathcal{N}_F^t & \varsigma h_F^{-1}\mathcal{N}_F^t\mathcal{N}_F \end{array}\right], \qquad \mathcal{S}_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & \eta h_F^{-1}\mathcal{N}_F^t\mathcal{N}_F \end{array}\right], \tag{6.22}$$

where $\mathcal{N}_F$ is defined as $\mathcal{N}$ by using n_F instead of n. For all $y \in [L^2(F)]^6$, let $M_F(y) = \mathcal{M}_F y$ and $S_F(y) = \mathcal{S}_F y$. Then, using Lemmata 5.1 and 5.3, one readily verifies that properties (DG$^\sharp$) hold. An IP-like method can be derived as well.

Remark 6.3. As opposed to advection-diffusion-reaction equations, the upper bounds in (5.7) and (5.9) are not sharp for the simplified MHD equations since the operators M_F and S_F do not need to control the whole L^2-norm of the electric field.

7. Conclusions. It happens sometimes that (A4) does not hold; instead, the following weaker inequality holds:

$$\exists \mu_0 > 0 \quad \forall z \in W, \qquad (Tz, z)_L + (z, \tilde{T}z)_L \geq 2\mu_0 \|\pi z^\sigma\|_{L_\sigma}^2, \tag{7.1}$$

where $\pi \in \mathcal{L}(L_\sigma; L_\sigma)$ may not be injective. In other words, coercivity no longer holds on the u-component of the unknown but holds only on a piece of the σ-component,

namely πz^σ. The equation $-\Delta u = f$ corresponds to this situation with π equal to the identity. The linear continuum mechanics equations in the incompressible limit, e.g., the Stokes equations, also fall in this framework with a nontrivial noninjective operator π. It will be shown in a forthcoming third part that, provided additional mild assumptions are made on the differential operators and on the DG setting, all that has been said herein in the fully L-coercive case remains valid in the situation with partial coercivity.

Acknowledgment. Fruitful discussions with Daniele Di Pietro (University of Bergamo and CERMICS, ENPC, ParisTech) are gratefully acknowledged.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[4] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery, Fluid Dynamics and Thermodynamics, R. Decuypere and G. Dibelius, eds., Technologisch Instituut, Antwerpen, Belgium, 1997, pp. 99–108.

[5] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[6] F. Brezzi, M. Manzini, L. D. Marini, P. Pietra, and A. Russo, *Discontinuous finite elements for diffusion problems*, in Atti Convegno in Onore di F. Brioschi, Instituto Lombardo, Accademia di Scienze e Lettere, Milan, Italy, 1999. pp. 197–217.

[7] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[8] A. Ern, J.-L. Guermond, and G. Caplain, *An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems*, Comm. Partial Differential Equations, to appear.

[9] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems.* I. *General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.

[10] K. O. Friedrichs, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[11] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[12] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Math. Res. 33, Academic Press, New York, 1974, pp. 89–123.

[13] P. Lesaint, *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*, Ph.D. thesis, University of Paris VI, Paris, France, 1975.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2389–2407

ON THE LOCAL LINEAR INDEPENDENCE OF GENERALIZED SUBDIVISION FUNCTIONS*

JÖRG PETERS† AND XIAOBIN WU†

Abstract. Characterizing the linear and local linear independence of the functions that span a linear space is a key task if the space is to be used computationally. Given a control net, the spanning functions of one spatial coordinate of a generalized subdivision surface are called nodal functions. They are the limit, under subdivision, of associating the value one with one control net node and zero with all others. No characterization of independence of nodal functions has been published to date, even for the two most popular generalized subdivision algorithms, Catmull–Clark subdivision and Loop's subdivision. This paper provides a road map for the verification of linear and local linear independence of generalized subdivision functions. It proves the conjectured global independence of the nodal functions of both algorithms, disproves local linear independence (for higher valences), and establishes linear independence on every surface region corresponding to a facet of the control net. Subtle exceptions, even to global independence, underscore the need for a detailed analysis to provide a sound basis for a number of recently developed computational approaches.

Key words. linear independence, nodal functions, subdivision surfaces, basis, Loop's scheme, Catmull–Clark scheme, local linear independence, condition number

AMS subject classifications. 41A15, 41A05, 41A63, 68U05, 68U07, 65D17, 65D18

DOI. 10.1137/050627496

1. Introduction.

Subdivision algorithms create an ever-tighter approximation of a smooth free-form surface by recursively refining and smoothing a polyhedral input mesh, known as control net, or mesh (see Figure 1.1). The two most popular subdivision algorithms, Catmull–Clark and Loop, replace at each step one facet with four. Catmull–Clark subdivision meshes consist of four-sided facets [3] and Loop meshes consist of triangles [11]. The two refinement methods are very popular in graphics, and, although the resulting surfaces are not "fair" enough for high-end industrial styling purposes [14], they are increasingly considered for computational purposes [9, 4]. It is therefore important to know whether the functions, associated with each control node of the mesh, are linearly independent, or what dependence exists. Locally, on a fixed domain Ω, the pieces of the subdivision surfaces have the form $\sum_i \mathbf{a}_i \nu_i$, $\mathbf{a}_i \in \mathbb{R}^3$, i.e., they are a linear combination of *nodal functions* ν_i.

A number of publications have tacitly assumed that the nodal functions ν_i are linearly independent. Without proof, [9, 8, 4] call the nodal functions subdivision basis functions, [12] uses the nodal functions as scaling functions to form a "basis" of the coarsest level of a multiresolution hierarchy, [13] fits subdivision surfaces by allowing one interpolation condition for each mesh node, and [17, 16, 19] call certain eigenfunctions (a set closely related to the set of nodal functions) an "eigenbasis"; and the analysis via universal surfaces, e.g., [19], is error prone unless the eigenfunctions form a basis. In fact, while generically true, for Catmull–Clark subdivision the assumption of linear independence of the nodal functions is false in some cases. The eight nodal functions of the simplest quadrilateral control mesh, a cube, are globally

*Received by the editors March 24, 2005; accepted for publication (in revised form) February 22, 2006; published electronically December 1, 2006. This work was supported in part by NSF grants DMI-0400214 and CCF-0430891.

http://www.siam.org/journals/sinum/44-6/62749.html

†Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611 (jorg@cise.ufl.edu, xwu@cise.ufl.edu).

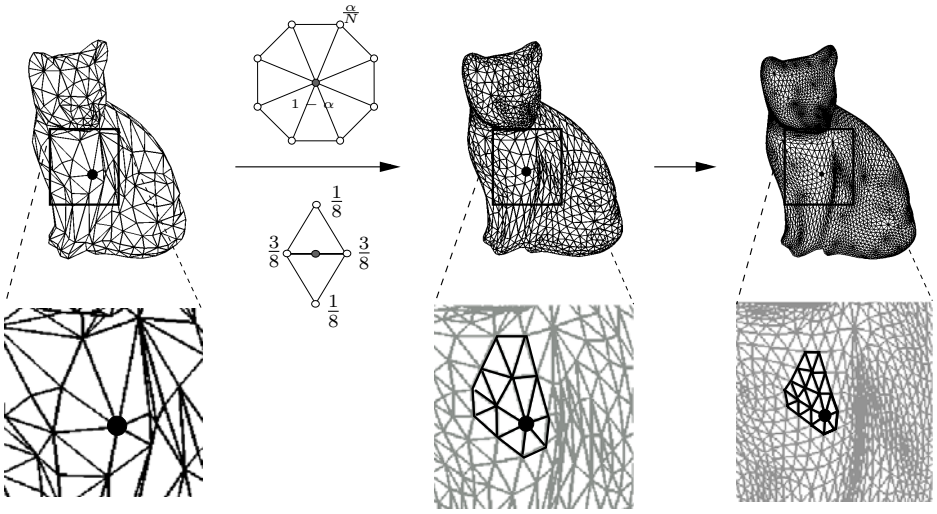


FIG. 1.1. *Modeling with Loop's subdivision. Top: A triangulated cat sculpture refined by Loop subdivision. Two stencils give the weights for averaging new nodes from old ones. The "vertex rule" (above the arrow) is centrally symmetric with* $64\alpha := 40 - (3 + 2\cos(2\pi/N))^2$ *and preserves the number N of neighbors, i.e., the node's valence. The "edge rule" creates nodes of valence six for each edge. Bottom, solid lines: Sequence of submeshes defining one of* $N = 7$ *triangular surface pieces that surround the limit point of the sequence of extraordinary nodes* $\bullet$.

linearly dependent (see Lemma 4.1); in general, we cannot fit eight arbitrary data points by adjusting the coefficients $\mathbf{a}_i$ of the corresponding surface $\sum_{i=1}^{8} \mathbf{a}_i \nu_i$.

For the well-known tensor-product spline functions, (global) linear independence may be interpreted as linear independence over the checkerboard grid of the union of domain rectangles delineated by the knot lines and joined by identifying edges of the rectangles in the natural fashion. generalizes to subdivision surfaces This definition as follows. Let Ω be a unit square if the kth facet of the control mesh has four vertices; let Ω be a unit triangle if it has three vertices. Let Γ be the union of all domains (Ω, k), indexed by their control mesh facet index, with edges topologically identified if the facets share edges. This gives Γ the structure of a 2-manifold homeomorphic to the control mesh. Global linear independence is linear independence with respect to Γ.

DEFINITION 1.1 (global linear independence). *A set of nodal functions are* globally linearly independent *if they are independent over the domain manifold* Γ. *That is, if* $\forall \mathbf{u} \in \Gamma : \sum_i \mathbf{a}_i \nu_i(\mathbf{u}) = \mathbf{0}$, *then* $\mathbf{a}_i = 0$.

While some numerical methods require only standard (global) linear independence, others, such as local Hermite interpolation and localized multiresolution, rely on stronger notions of independence. We need to analyze independence on certain ring-shaped annuli $\mathcal{A}$ and on subsets Ω_i of Ω. A stronger, subtle notion of independence is local linear independence.

DEFINITION 1.2 (local linear independence). *A set of nodal functions are* locally linearly independent *if for any bounded open* $G \subseteq \Gamma$, *all the nodal functions having some support in G are linearly independent on G.*

Remarkably, for box-splines and B–splines, the standard notion of (global) linear independence is equivalent to local linear independence [6, (II.57) Theorem, p. 51]. That is, if all coefficients $\mathbf{a}_i$ have to vanish in order that $\sum_i \mathbf{a}_i \nu_i \equiv \mathbf{0}$ (global linear

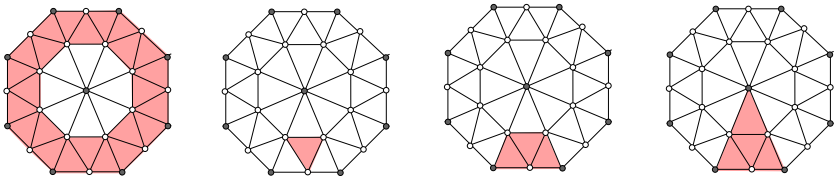


FIG. 1.2. *Summary of findings for Loop subdivision. Domains G (shaded) and valence N for which the nodal functions with support on G are linearly independent.*

independence) then the coefficients of all nodal functions that are nonzero on any open set G have to vanish if $\sum_i \mathbf{a}_i \nu_i$ vanishes on G (local linear independence). Since G can be arbitrarily small, local linear independence is a stricter requirement on the nodal functions than global linear independence. By contrast, local and global independence are not equivalent for subdivision nodal functions. This observation provides rare insight into the structural difference between subdivision and spline surfaces. Specifically, we show that for Catmull–Clark and Loop subdivision,

(i) the nodal functions are globally linearly independent[1];

(ii) the nodal functions are linearly independent over an annulus[1] such as in Figure 1.2, see Lemma 2.2;

(iii) for valence N higher than the "regular" valence, the nodal functions are not locally linearly independent;

(iv) the nodal functions are linearly independent on each domain Ω naturally associated with one facet of the control net[1].

Points (i)–(iv) have direct implications on interpolation (to be compared with the Schoenberg–Whitney theorem of spline interpolation). Consider interpolation with Loop subdivision surfaces. If all three vertices of a facet have valence six, interpolating 12 data points on a domain Ω, by adjusting 12 control points, is a well-posed problem with a unique solution. However, if one of the vertices has valence $N < 6$, then matching 12 data points represents too many constraints; if $N > 6$, then the fitting problem is underconstrained. If we choose the number of interpolation conditions to equal the number of nodal functions that are nonzero on Ω, i.e., if we specify $N + 6$ interpolation points, we find that, if the points belong to a subregion Ω_1 the problem is overconstrained for $N > 6$. (Ω_1 is the shaded area in Figure 1.2, labeled Conjecture 1.) Interpolation with Catmull–Clark subdivision follows a similar pattern with an additional complication for $N = 3$.

The analysis is made easier by the fact that the component functions of most popular subdivision schemes, and in particular of both Catmull–Clark and Loop subdivision, are variations of the well-understood box-spline subdivision [6]; much of the subdivision limit surfaces, corresponding to quads with 4-valent vertices, respectively, triangles with 6-valent vertices are "regular," i.e., are spline surfaces generated by box-splines. This box-spline connection should make us cautious since the shifts of box-splines are, in general, not linearly independent. For example, the four-direction

[1]There is one exception: Catmull–Clark subdivision applied to nodes with valence $N = 3$; see Lemma 4.1.

(quincunx) subdivision, which gives rise to 4–8 subdivision [18], has dependent nodal functions. Catmull–Clark subdivision rules generalize the two-direction box-spline rules, i.e., the rules of the bicubic tensor-product spline; and Loop subdivision generalizes a three-direction box-spline, the convolution of the linear "hat" function with itself. For both splines we know [6] that the nodal functions form a basis and therefore are independent. This means we can focus on submeshes that define the neighborhood of extraordinary nodes, where the connectivity of the control mesh differs from the regular connectivity of the box-spline, namely submeshes surrounding nodes of valence $N \neq 4$ for Catmull–Clark meshes and of valence $N \neq 6$ for Loop meshes. We do not assume that all direct neighbors of these extraordinary nodes are of regular valence.

We first discuss Loop's subdivision, proposed for computational purposes in [9, 8, 4], then Catmull–Clark subdivision, and then generalize the key results.

2. Loop subdivision. A subdivision algorithm states how a new node is computed from a (small local) submesh of old nodes, and how this new node is to be connected to other new nodes. In particular, for Loop subdivision, there are only two rules: to compute new nodes, corresponding to edges of the old mesh, and to compute new nodes, corresponding to old nodes. These rules are expressed by the two stencils (weighted neighborhood graphs) in Figure 1.1, above and below the arrow. A node of a Loop mesh is *extraordinary* if it does not have six neighbors.

Due to the small footprint of the rules, a submesh consisting of one triangle and all triangles attached to it defines, by going to the limit, a triangular piece of the surface adjacent to the limit of the extraordinary node. If all nodes of the central triangle are of valence six, the surface is a polynomial piece of a three-direction box-spline and its properties are well understood. Since new edge nodes have valence six, extraordinary nodes are more isolated under refinement, and we can focus on triangles with one extraordinary node of valence $N \neq 6$. In the following, the subscript 0 refers to a mesh where any two extraordinary nodes are separated by at least one node of valence six. This may be the result of one subdivision applied to the original mesh.

While the typical application of Loop subdivision creates a parametrized surface in $\mathbb{R}^3$, *for the analysis it is sufficient to look at one spatial coordinate since the coordinates do not interact.* The nodal function ν_i may then be defined by choosing an association of the control points a_j with the domain, setting one scalar control point a_i to 1 and all others to 0 and applying subdivision. The relevant submesh defining the triangular surface piece consists of $K := N + 6$ nodes that can be labeled as in Figure 2.1 (top left). We store the submesh as a vector

$$\mathbf{c}_0 := (\mathbf{c}_{0,1}, \dots \mathbf{c}_{0,K}) \in \mathbb{R}^K.$$

Subdivision generates a new set of $M := K+6$ control vertices as shown in Figure 2.1 (top right). We store those control vertices in a new vector

$$\mathbf{c}_1 := (\mathbf{c}_{1,1}, \dots, \mathbf{c}_{1,K}, \mathbf{c}_{1,K+1}, \dots \mathbf{c}_{1,M}).$$

If we represent the averaging rules as rows of a $M \times K$ matrix $\mathbf{A}$ (with row sum one), then the subdivision rules to compute the vector $\mathbf{c}_1$ from $\mathbf{c}_0$ are

$$\mathbf{c}_1 = \mathbf{A}\mathbf{c}_0, \quad \text{where } \mathbf{A} := \begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \\ \mathbf{A}_{31} & \mathbf{A}_{32} \end{pmatrix}.$$

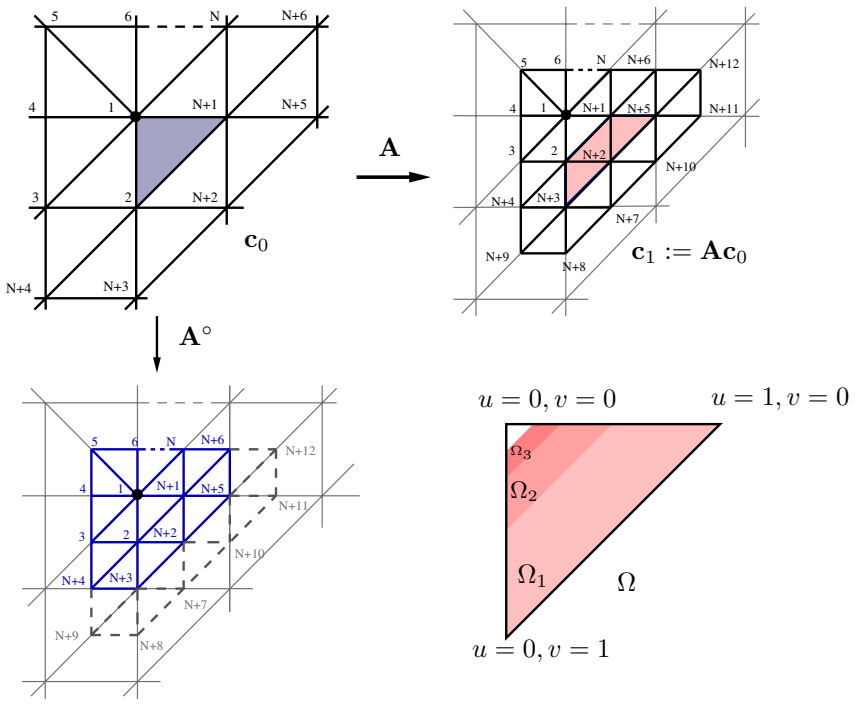


FIG. 2.1. *Top left: Labeling of the submesh that defines a triangular surface piece (schematically represented by the shaded area) near an extraordinary node (label 1). Top right: Refined submesh,* $\mathbf{A}\mathbf{c}_0$. *Bottom left: Refined submesh,* $\mathbf{A}^\circ\mathbf{c}_0$, *used to evaluate the next spline ring. Bottom right: The domain* Ω *of the composite triangular surface piece consists of an infinite sequence of quadrilateral (chopped triangle) subdomains. The first three such subdomains,* Ω_1, Ω_2, Ω_3, *are shaded.*

Here $\mathbf{A}_{11}$ is an $(N+1)\times(N+1)$-matrix that computes the new extraordinary node and the vertices adjacent to it (note that this also holds for an optional initial refinement to generate $\mathbf{c}_0$ from a mesh that has neighboring extraordinary nodes); $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ determine the five vertices with indices $N+4, N+3, N+2, N+5, N+6$ of the next layer; and $\mathbf{A}_{31}$ and $\mathbf{A}_{32}$ define the six outermost nodes. The sizes of $\mathbf{A}_{22}$ and $\mathbf{A}_{32}$ are 5×5 and 6×5, respectively. Leaving out the direct neighbors of the extraordinary node, $\mathbf{c}_{1,4}, \mathbf{c}_{1,5}, \ldots, \mathbf{c}_{1,N-1}$, the remaining control points

$$\mathbf{c}_1^{\text{box}} := (\mathbf{c}_{1,1}, \mathbf{c}_{1,2}, \mathbf{c}_{1,3}, \mathbf{c}_{1,N}, \ldots \mathbf{c}_{1,M})$$

define three triangular polynomial pieces shown as shaded in Figure 2.1 (top right).

To compute the nodes of the next subdivision step, we need only the first K control points of $\mathbf{c}_1$ (see Figure 2.1 bottom left),

$$(\mathbf{c}_{1,1}, \mathbf{c}_{1,2}, \ldots, \mathbf{c}_{1,K}) = \mathbf{A}^\circ\mathbf{c}_0 := \begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \mathbf{c}_0.$$

By repeating the process, an infinite sequence of piecewise polynomial rings is generated. We can choose their domains Ω_ℓ so that their union fills out the triangular domain Ω:

$$\Omega := \{(u,v) | u+v+w = 1, u, v, w \geq 0\} = \cup_{\ell=1}^{\infty}\Omega_\ell \qquad \Omega_1 := \Omega \backslash \frac{1}{2}\Omega, \quad \Omega_{\ell+1} := \frac{1}{2}\Omega_\ell.$$

The control vertices $\mathbf{c}_n$ after n subdivision steps that determine the function on Ω_n are

$$\mathbf{c}_n = \mathbf{A}(\mathbf{A}^\circ)^{n-1}\mathbf{c}_0, \qquad n \geq 1. \tag{2.1}$$

From the recursion in 2.1, it is evident that the eigenstructure of $\mathbf{A}^\circ$ plays a crucial rule in determining the properties of the subdivision surfaces such as the computation of the limit position, tangent plane, and shape analysis [7, 1, 15, 14, 10].

Using Fourier transform, it is easy to derive the vector of eigenvalues Λ_{11} of $\mathbf{A}_{11}$,

$$\Lambda_{11} := \left[1, \frac{5}{8} - \alpha(N), f(1), \ldots, f(N-1)\right],$$

where

$$f(k) := \frac{3 + 2\cos(2\pi k/N)}{8}, \qquad \alpha(N) := \frac{5}{8} - f(1)f(1),$$

and Λ_{22} of $\mathbf{A}_{22}$,

$$\Lambda_{22} := \left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right].$$

Except for the case $N = 3$, $\mathbf{A}^\circ$ can be diagonalized by the matrix $\mathbf{V}$ of its eigenvectors (details of the eigenanalysis of $\mathbf{A}^\circ$ can be found, e.g., in [16]):

$$\mathbf{A}^\circ = \mathbf{V}\Lambda\mathbf{V}^{-1}, \quad \Lambda = \operatorname{diag}(\Lambda_{11}, \Lambda_{22}), \mathbf{V} = \begin{pmatrix} \mathbf{U}_0 & 0 \\ \mathbf{U}_1 & \mathbf{W}_1 \end{pmatrix}, \tag{2.2}$$

where the submatrices $\mathbf{U}_0$ and $\mathbf{W}_1$ are the eigenvectors of $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$, respectively. For $N > 3$, the columns of $\mathbf{V}$ are linearly independent vectors in $\mathbb{R}^K$.

Now let the initial submesh $\mathbf{c}_0 := \mathbf{v}_i$ be an eigenvector associated with eigenvalue λ_i and φ_i the corresponding linear combination of nodal functions. Then, after n steps of subdivision,

$$\mathbf{c}_n|_{\mathbf{c}_0=\mathbf{v}_i} = \mathbf{A}(\mathbf{A}^\circ)^{n-1}\mathbf{v}_i = \mathbf{A}\lambda_i^{n-1}\mathbf{v}_i = \lambda_i^{n-1}\mathbf{A}\mathbf{v}_i = \lambda_i^{n-1}\mathbf{c}_1|_{\mathbf{c}_0=\mathbf{v}_i}.$$

Therefore $\varphi_i(\Omega_{n+1})$ is a scaled multiple of $\varphi_i(\Omega_1)$. Precisely,

$$\forall (u,v) \in \Omega \text{ and } \forall n \geq 1, \varphi_i\left(\frac{u}{2^n}, \frac{v}{2^n}\right) = \lambda_i^n \varphi_i(u,v). \tag{2.3}$$

In [16] these K functions φ_i are called eigenbasis. However, adjacent to an extraordinary node, each φ_i consists of an infinite union of polynomial pieces. The subtle but important point to be settled here is that, even if the columns of $\mathbf{V}$ are independent, the corresponding functions can be dependent. We therefore call the functions φ_i *eigenfunctions*. We note that, to be scalable, the control net of an eigenfunction are only well defined if the extraordinary node is surrounded by regular nodes. For the proofs of independence of *nodal functions*, we will be allowed to assume that the extraordinary node is isolated. An *extraordinary node is isolated* if it is surrounded by regular nodes. For, if $\mathbf{c}_0$ is isolated as the result of one refinement, and we show that $\mathbf{c}_0$ is zero, then the values associated with the unrefined nodes must also be zero since the matrix $\mathbf{A}_{11}$ that maps the original nodes to the refined nodes is of full rank for all valences. To see that $\mathbf{A}_{11}$ is of full rank also for $N = 3$, we need only

observe that $\Lambda_{11} = [1, f(1)f(1), f(1), f(1)]$, $f(1) = 1/4$ and the eigenvectors of $f(1)$ are independent.

LEMMA 2.1. $\mathbf{A}_{11}$ *is of full rank for all valences.*

To show that subdivision near extraordinary nodes is similar to but different from spline representations, we will show that the eigenfunctions are linearly independent over Ω, but linearly dependent on certain subsets of Ω.

To show that the nodal functions of Loop subdivision are (globally) linearly independent, we focus on subdomains that form an annulus surrounding the preimage of a sequence of extraordinary nodes. With the natural topological identification of edges to induce the structure of a 2-manifold with two boundaries, we define an annulus as N copies of Ω_1,

$$\mathcal{A} := \{1, \ldots, N\} \times \Omega_1.$$

Note that this requires that the extraordinary node is isolated.

LEMMA 2.2. *The nodal functions of Loop subdivision with support on* $\mathcal{A}$ *are linearly independent over* $\mathcal{A}$.

Proof. For $\mathcal{A}$ to be well defined, the extraordinary node must be isolated. Let $f := \sum_i \mathbf{c}_{0,i}\nu_i$ be zero on all of $\mathcal{A}$. Then the subset of nodes $\mathbf{c}_1^{\text{box}}$ can be interpreted as a regular three-direction box-spline control net defining three polynomial pieces near the extraordinary node. Since the box-splines are locally linearly independent [6], all box-spline control points defining f on $\mathcal{A}$ are zero. Since all eigenvalues of $\mathbf{A}^\circ$ are positive for $N > 3$, $\mathbf{A}^\circ$ is of full rank and for $N = 3$, Lemma 3.5 shows that the matrix M is of full rank. Therefore all $\mathbf{c}_{0,i}$ must be zero. ∎

Now consider *all* nodal functions of a once-refined control net. Lemma 2.2 proves linear independence of these nodal functions on the union of all annuli $\mathcal{A}$. By Lemma 2.1, the original control nodes must also be zero if the function vanishes on all annuli.

COROLLARY 2.3. *The nodal functions of Loop subdivision are globally linearly independent.*

3. Local linear independence of Loop subdivision. In this section, we characterize the local linear independence of Loop subdivision nodal functions.

LEMMA 3.1. *For general* N*, the nodal functions of Loop subdivision are* not locally linearly independent. *Specifically, for any* k *there exists a valence* N *so that the nodal functions of Loop subdivision with support on* Ω_k*,* $\nu_i, i = 1 \ldots N + 6$*, are* locally linearly dependent *on* Ω_k *and even on* $\cup_{\ell=1}^{k}\Omega_\ell$.

Proof. All nodal functions corresponding to $\mathbf{c}_0$ have support (are nonzero) on each subdomain Ω_k. Each vector $\mathbf{c}_k^{\text{box}}$ corresponding to Ω_k has 16 entries. For sufficiently large valence, the nodal functions on Ω_k must therefore be dependent. By the same reasoning, for sufficiently large valence, the nodal functions on $\cup_{\ell=1}^{k}\Omega_\ell$, for finite k, must be dependent. ∎

As could be hoped by the failure of the above counting argument, nodal functions are locally linearly independent for sufficiently low valence N.

LEMMA 3.2. *For* $N \leq 6$ *the nodal functions* $\nu_i, i = 1 \ldots N+6$*, are locally linearly independent.*

Proof. Denote by $\mathbf{P}_i, i = 1, 2, 3$ the three $12 \times (N + 12)$ picking matrices that select the box-spline coefficients of each of the three triangular domain parts. Since $\mathbf{P}_i\mathbf{A}, i = 1, 2, 3$ is of full rank $\mathbf{c}_0$ must be zero if the 12 box-spline control points are zero. Then local linear independence of the three-direction box-spline implies the claim on Ω_1. Since the control points on Ω_ℓ are computed from $\mathbf{c}_0$ by applying

$$\mathbf{P}_i\mathbf{A}(\mathbf{A}^\circ)^{\ell-1}$$

local linear independence on Ω_1 implies local linear independence on Ω_ℓ for $N > 3$. For $N = 3$, the claim follows from Lemma 3.5. ∎

For $N = 7$, the nodal functions are independent on Ω_1 and hence on $\cup_{\ell=1}^{k}\Omega_\ell$, but, due to the dimension of the three polynomial pieces corresponding to Ω_1, they are not linearly independent on subsets of Ω_1 that do not straddle all three piecewise polynomial domains. As the valence N increases, a subtle pattern emerges.

CONJECTURE 1. *For $k := \lfloor (N-6)/2 \rfloor + 1 > 1$, the nodal functions of Loop subdivision $\nu_i, i = 1 \dots N+6$ are linearly independent on $\cup_{i=1}^{k}\Omega_i$ and linearly dependent on $\cup_{i=1}^{k-1}\Omega_i$.*

We verified the conjecture for isolated extraordinary nodes by symbolic calculation up to $N = 30$, which should cover most cases of practical interest.

To investigate linear independence on the natural domains corresponding to control facets, namely on $\Omega = \cup_{\ell=1}^{\infty}\Omega_\ell$, we need a better strategy. We use the eigenproperty of the eigenfunctions, that additional layers are scaled copies of the earlier layers.

LEMMA 3.3. *For valence $N > 3$, the* eigenfunctions φ_i, $i = 1, \dots N+6$, *of Loop subdivision are linearly independent on Ω.*

Proof. The proof is by contradiction. All eigenvalues λ_i of $\mathbf{A}^\circ$ are positive. We sort the eigenfunctions φ_i $(i = 1, \dots N+6)$ so that their associated eigenvalues λ_i descend from the largest to the smallest. Suppose there exist scalars $a_1, a_2, \dots a_{N+6}$, not all zero, such that

$$\sum_{i=1}^{N+6} a_i \varphi_i = 0.$$

Let λ_j be the largest eigenvalue such that $a_j \neq 0$. Then with $w_i := -a_i/a_j$, we write

$$\varphi_j = \sum_{i=j+1}^{N+6} w_i \varphi_i.$$

For $\forall (u,v) \in \Omega$ and $\forall n \geq 1$, $(\frac{u}{2^n}, \frac{v}{2^n}) \in \Omega$, with the above equation and (2.3), we have

$$\begin{aligned}
\varphi_j\left(\frac{u}{2^n}, \frac{v}{2^n}\right) &= \sum_{i=j+1}^{N+6} w_i \varphi_i\left(\frac{u}{2^n}, \frac{v}{2^n}\right), \\
\Rightarrow \lambda_j^n \varphi_j(u,v) &= \sum_{i=j+1}^{N+6} w_i \lambda_i^n \varphi_i(u,v), \\
\Rightarrow \varphi_j(u,v) &= \sum_{i=j+1}^{N+6} w_i \left(\frac{\lambda_i}{\lambda_j}\right)^n \varphi_i(u,v).
\end{aligned}$$

Since $(\frac{\lambda_i}{\lambda_j})^n \to 0$ as $n \to \infty$ unless $\lambda_i = \lambda_j$,

$$\varphi_j(u,v) = \sum_{i \in \{i | \lambda_i = \lambda_j\}} w_i \varphi_i(u,v)$$

must hold. Therefore the eigenfunctions associated with λ_j must be linearly dependent. In the remainder of the proof, we show this to be false. In other words, the

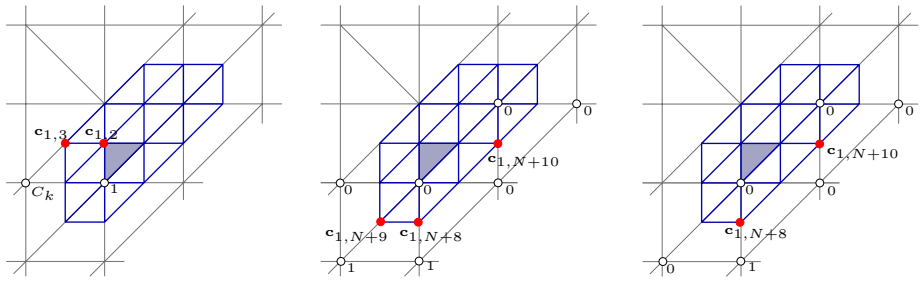


FIG. 3.1. *The box-spline control points* $\mathbf{c}_{1,i}$ *(solid dots) used to certify that pairs and triples of eigenfunctions are independent.*

problem of proving the linear independence of all eigenfunctions has been reduced to the independence of the eigenfunctions with the same eigenvalue.

Because of the eigenstructure of $\mathbf{A}^\circ$, the multiplicities of its eigenvalues are small (at most four) and do not increase with N. Recall that the eigenvalues of $\mathbf{A}^\circ$ are

$$\left[1, \frac{5}{8} - \alpha(N), f(1), \ldots, f(N-1), \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right],$$

where

$$\alpha(N) := \frac{5}{8} - \frac{(3 + 2\cos(2\pi/N))^2}{64}, f(k) := \frac{3 + 2\cos(2\pi k/N)}{8}.$$

To find the repeated eigenvalues, we observe that for $k \in \{1 \ldots N-1\}$,

1. $f(k) = f(N-k)$,
2. if N is even and $k = N/2$, $f(k) = \frac{1}{8}$; otherwise $f(k) \notin \{\frac{1}{8}, \frac{1}{16}\}$,
3. $f(k) \neq 1$ and $f(k) \neq \frac{5}{8} - \alpha(N)$.

That is, if λ is an eigenvalue of $\mathbf{A}$ with multiplicity greater than one, then $\lambda = f(k) \neq \frac{1}{8}$, or $\lambda = \frac{1}{8}$, or $\lambda = \frac{1}{16}$. In particular, all relevant eigenvalues are nonzero. We look at each case individually.

- *Case* 1. $\lambda = f(k) \neq \frac{1}{8}$
  In this case λ has multiplicity 2 and the associated eigenvectors $\mathbf{u}_k$ and $\mathbf{w}_k$ are given in [16]:

$$\mathbf{u}_k^T = (0, 1, C_k, C_{2k}, \ldots C_{(N-1)k}, \ldots) \qquad \text{and}$$
$$\mathbf{w}_k^T = (0, 0, S_k, S_{2k}, \ldots S_{(N-1)k}, \ldots),$$

  where $C_k := \cos(2\pi k/N)$ and $S_k := \sin(2\pi k/N)$. To show the two eigenfunctions defined by $\mathbf{u}_k$ and $\mathbf{w}_k$ are linearly independent, we consider the two box-spline entries of $\mathbf{c}_{1,2}$ and $\mathbf{c}_{1,3}$ (solid dots in Figure 3.1, left) after one step of subdivision applied to the mesh $\mathbf{c}_0 := \mathbf{u}_k$ and one step applied with $\mathbf{c}_0 := \mathbf{w}_k$. The two corresponding eigenfunctions are independent because

$$\det \begin{pmatrix} \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{u}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{u}_k} \\ \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{w}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{w}_k} \end{pmatrix} = \lambda^2 \det \begin{pmatrix} 1 & C_k \\ 0 & S_k \end{pmatrix} \neq 0,$$

  since $S(k) \neq 0$ because $f(k) \neq \frac{1}{8}$ and hence $k \neq \frac{N}{2}$.

- *Case* 2. $\lambda = \frac{1}{8}$
  In this case λ can have multiplicity of 3 or 4. We first show that the eigenfunctions corresponding to the first three columns $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ of $\binom{0}{\mathbf{W}_1}$ are independent. The eigendecomposition (2.2) of $\mathbf{A}_{22}$ is (see [16])

$$\mathbf{W}_1 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

  The independence of the eigenfunctions follows from the independence of the three box-spline control points $\mathbf{c}_{1,N+8}, \mathbf{c}_{1,N+9}, \mathbf{c}_{1,N+10}$ (solid dots in Figure 3.1, middle) after one subdivision:

$$\det \begin{pmatrix} \mathbf{c}_{1,N+8}|_{\mathbf{c}_0=\mathbf{y}_1} & \mathbf{c}_{1,N+9}|_{\mathbf{c}_0=\mathbf{y}_1} & \mathbf{c}_{1,N+10}|_{\mathbf{c}_0=\mathbf{y}_1} \\ \mathbf{c}_{1,N+8}|_{\mathbf{c}_0=\mathbf{y}_2} & \mathbf{c}_{1,N+9}|_{\mathbf{c}_0=\mathbf{y}_2} & \mathbf{c}_{1,N+10}|_{\mathbf{c}_0=\mathbf{y}_2} \\ \mathbf{c}_{1,N+8}|_{\mathbf{c}_0=\mathbf{y}_3} & \mathbf{c}_{1,N+9}|_{\mathbf{c}_0=\mathbf{y}_3} & \mathbf{c}_{1,N+10}|_{\mathbf{c}_0=\mathbf{y}_3} \end{pmatrix} = \frac{1}{8^3} \det \begin{pmatrix} 4 & 4 & 0 \\ 0 & 0 & 1 \\ 1 & 4 & 4 \end{pmatrix} \neq 0.$$

  If the multiplicity of $\frac{1}{8}$ is three, then we are done. Otherwise, $\lambda = f(N/2)$ for N is even and we have one additional eigenvector $\mathbf{u}_k$ from $\binom{\mathbf{U}_0}{\mathbf{U}_1}$. After one subdivision, the box-spline control point $\mathbf{c}_{1,3}$ is zero for $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ and nonzero for $\mathbf{u}_k$. This proves independence of all four eigenfunctions.
- *Case* 3. $\lambda = \frac{1}{16}$
  The eigenvectors of the two eigenfunctions associated with $\frac{1}{16}$ correspond to the last two columns $\mathbf{z}_1$ and $\mathbf{z}_2$ of $\binom{0}{\mathbf{W}_1}$. Pairwise independence follows from the independence of the two box-spline control points $\mathbf{c}_{1,N+8}, \mathbf{c}_{1,N+10}$ (solid dots in Figure 3.1, right)

$$\det \begin{pmatrix} \mathbf{c}_{1,N+8}|_{\mathbf{c}_0=\mathbf{z}_1} & \mathbf{c}_{1,N+10}|_{\mathbf{c}_0=\mathbf{z}_1} \\ \mathbf{c}_{1,N+8}|_{\mathbf{c}_0=\mathbf{z}_2} & \mathbf{c}_{1,N+10}|_{\mathbf{c}_0=\mathbf{z}_2} \end{pmatrix} = \frac{1}{8^2} \det \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix} \neq 0.$$

This completes the proof of Lemma 3.3. ∎

We can now address our original goal of showing that the nodal functions ν_i are linearly independent.

COROLLARY 3.4. *For $N > 3$, the nodal functions of Loop subdivision, $\nu_i, i = 1 \ldots N+6$, are linearly independent on Ω.*

Proof. Each nodal function ν_i is generated by subdivision when setting control point i to 1 and all others to 0. If the extraordinary node is surrounded by regular nodes,

$$[\varphi_1, \ldots, \varphi_K] = \mathbf{V}[\nu_1, \ldots, \nu_K], \qquad K = N + 6,$$

and independence follows since, for $N > 3$, the matrix $\mathbf{V}$ of eigenvectors is an invertible matrix. The general case follows by one step of subdivision and the full rank of $\mathbf{A}_{11}$. ∎

For the special case $N = 3$, the matrix $\mathbf{A}^\circ$ has a nontrivial Jordan block and cannot be diagonalized. However, since the number of nodal functions is small, namely nine, we need not decompose into the eigenspace.

LEMMA 3.5. *For $N = 3$, the nodal functions of Loop subdivision, $\nu_i, i = 1 \ldots N+6$, are linearly independent on Ω.*

Proof. If the extraordinary node is isolated, we explicitly determine the $(N + 12) \times 9$ matrix $\mathbf{M}$ that maps $\mathbf{c}_0$ to the box-spline control points $\mathbf{c}_1^{\text{box}}$.

$$\frac{1}{16}\begin{bmatrix}
7 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0\\
6 & 6 & 2 & 2 & 0 & 0 & 0 & 0 & 0\\
6 & 2 & 6 & 2 & 0 & 0 & 0 & 0 & 0\\
6 & 2 & 2 & 6 & 0 & 0 & 0 & 0 & 0\\
2 & 6 & 0 & 6 & 2 & 0 & 0 & 0 & 0\\
1 & 10 & 1 & 1 & 1 & 1 & 1 & 0 & 0\\
2 & 6 & 6 & 0 & 0 & 0 & 2 & 0 & 0\\
1 & 1 & 1 & 10 & 1 & 0 & 0 & 1 & 1\\
2 & 0 & 6 & 6 & 0 & 0 & 0 & 0 & 2\\
0 & 6 & 0 & 2 & 6 & 2 & 0 & 0 & 0\\
0 & 6 & 0 & 0 & 2 & 6 & 2 & 0 & 0\\
0 & 6 & 2 & 0 & 0 & 2 & 6 & 0 & 0\\
0 & 2 & 0 & 6 & 6 & 0 & 0 & 2 & 0\\
0 & 0 & 0 & 6 & 2 & 0 & 0 & 6 & 2\\
0 & 0 & 2 & 6 & 0 & 0 & 0 & 2 & 6
\end{bmatrix}.$$

Since $\mathbf{M}$ has full rank and since the box-splines associated with each of $\mathbf{c}_{1,1}, \mathbf{c}_{1,2} \ldots \mathbf{c}_{1,N+12}$ are linearly independent, the $\nu_i, i = 1 \ldots 9$ are also linearly independent. The general case follows by one step of subdivision and the full rank of $\mathbf{A}_{11}$. ∎

Together, Corollary 3.4 and Lemma 3.5 prove the main Theorem 3.6.

THEOREM 3.6. *The nodal functions of Loop subdivision, $\nu_i, i = 1 \ldots N + 6$, are linearly independent over Ω.*

The theorem sharply characterizes the locality of linear independence. On any finite union of Ω_ℓ the nodal functions are linearly dependent for sufficiently high valence. Only once we take the union to the limit Ω, do we obtain linear independence of the nodal functions for all possible valences.

Lemmas 3.3 and 3.5 imply the analogous result for eigenfunctions.

COROLLARY 3.7. *For all N, the* eigenfunctions *φ_i, $i = 1, \ldots N + 6$, of Loop subdivision are linearly independent and form a basis for the Loop subdivision functions over Ω.*

In particular, we can now call the Loop eigenfunctions an *eigenbasis*.

4. Catmull–Clark subdivision. In this section, we investigate another widely used subdivision scheme, Catmull–Clark subdivision. The Catmull–Clark algorithm [3] accepts input meshes that have m-sided facets and vertices with N neighbors. However, all m-sided facets are split into m quadrilaterals in the first step as follows. A new face node is computed as the average of the facet vertices, a new edge node as the average of the edge endpoints and the two new face nodes of the faces joined by the edge, and a new vertex node of valence N is computed as

$$(Q + 2R + (N - 3)S)/N,$$

where Q is the average of the new face nodes of all faces adjacent to the old vertex, R is the average of the midpoints of all old edges incident on the old vertex point, and S is the old vertex point. A new quadrilateral facet then consists of consecutive edge node, vertex node, edge node, and the face node. The rules are consistent with the Catmull–Clark stencils for quadrilateral meshes listed in Figure 4.1. The standard Catmull–Clark choices for α, β, and γ are

$$\alpha := 1 - \frac{7}{4N}, \quad \beta := \frac{3}{2N^2}, \quad \gamma := \frac{1}{4N^2}.$$

If each node of a quadrilateral mesh facet has valence $N = 4$, Catmull–Clark subdivision amounts to tensor product bicubic spline subdivision. In this case, the nodal functions are the standard tensor product uniform B-spline basis functions whose independence is well documented. Since the extraordinary nodes (with valence $N \neq 4$) are always isolated after two subdivision steps, i.e., any two extraordinary

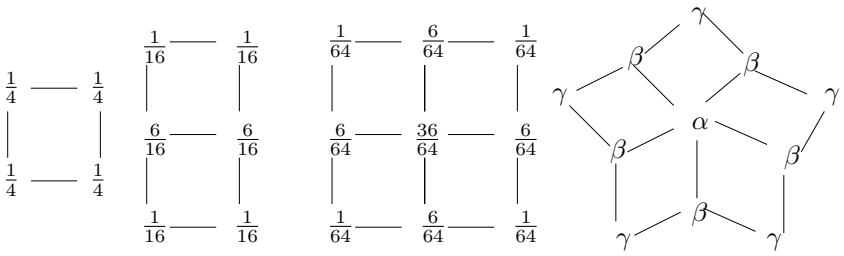


FIG. 4.1. *Refinement stencils of generalized Catmull–Clark subdivision.*

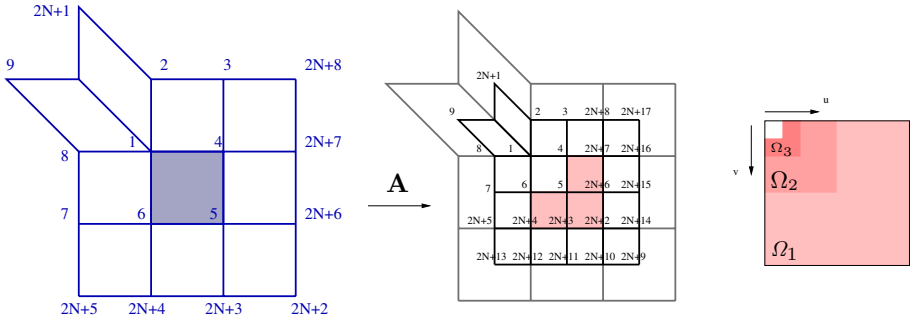


FIG. 4.2. *Left. Indices of Catmull–Clark nodes near a facet with one extraordinary node ($N = 5$). Middle. The indices of the new control points after one subdivision. Three quarters of the domain now have well-defined tensor product B-spline structure. Right. The complete rectangular domain is composed of an infinite number of L shaped regions Ω_ℓ.*

nodes are separated by at least one node of valence four, we can focus our local analysis on surface parts adjacent to a single extraordinary node. That is, the subscript 0 refers to a mesh with isolated extraordinary nodes.

The indices of the $K := 2N + 8$ adjacent control points are stored in $\mathbf{c}_0$ as in Figure 4.2, left:

$$\mathbf{c}_0 := (\mathbf{c}_{0,1}, \dots \mathbf{c}_{0,K}).$$

After subdivision, the new set of $M := K + 9$ control vertices is ordered as shown in Figure 4.2, middle and stored in the vector

$$\mathbf{c}_1 := (\mathbf{c}_{1,1}, \dots, \mathbf{c}_{1,K}, \mathbf{c}_{1,K+1}, \dots \mathbf{c}_{1,M}).$$

The subdivision rules are again denoted by

$$\mathbf{c}_1 = \mathbf{A}\mathbf{c}_0, \quad \text{where } \mathbf{A} := \begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \\ \mathbf{A}_{31} & \mathbf{A}_{32} \end{pmatrix}.$$

Here $\mathbf{A}_{11}$ is an $(2N + 1) \times (2N + 1)$-matrix that computes the new extraordinary node and the vertices adjacent to it; $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ determine the seven vertices with indices $2N + 2, \dots, 2N + 8$, in the middle vertex ring; and $\mathbf{A}_{31}$ and $\mathbf{A}_{32}$ compute the last nine vertices with indices $2N + 9, \dots, 2N + 17$. We have enough control points in $\mathbf{c}_1$ to evaluate three regular patches (see shaded area in Figure 4.2, middle). The first K control points of $\mathbf{c}_1$,

$$(\mathbf{c}_{1,1}, \mathbf{c}_{1,2}, \dots, \mathbf{c}_{1,K}) = \mathbf{A}^\circ \mathbf{c}_0 := \begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \mathbf{c}_0,$$

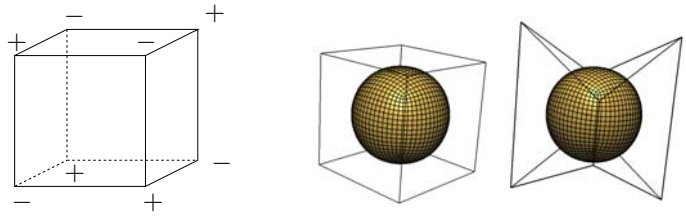


FIG. 4.3. *Global linear dependence of Catmull–Clark subdivision. Left: An alternative representation of the zero function with + indication any nonzero number and - its negative value. Right: Two control nets with the connectivity of a cube but different node positions. They generate the same Catmull–Clark surface!*

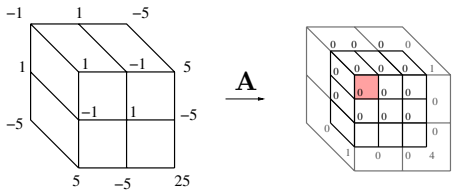


FIG. 4.4. *Nonzero input coefficients of the eigenfunction corresponding to the eigenvalue zero generating the zero function on* $\cup_{i=2}^{\infty}\Omega_i$ *(shaded area).*

are used as the control points for the next subdivision step. $\mathbf{A}^\circ$ can always be diagonalized by its eigenvectors $\mathbf{V}$,

$$\mathbf{A}^\circ = \mathbf{V}\Lambda\mathbf{V}^{-1}.$$

All eigenvalues are nonzero, except for $N = 3$ when one eigenvalue is zero. (The second eigenvalue of the zero Fourier block.) For $N > 3$, the linear independence of the nodal functions on $\mathcal{A}$ follows, just as in the case of Loop subdivision, from the local linear independence of tensor-product splines and the full rank of $\mathbf{A}^\circ$. The full rank of $\mathbf{A}_{11}$ implies *global linear independence* for $N > 3$.

The case $N = 3$ merits closer scrutiny.

LEMMA 4.1. *The nodal functions of Catmull–Clark subdivision corresponding to the graph in Figure* 4.3 *are (globally) linearly dependent.*

Proof. Given the displayed choice of nonzero values at the vertices, all new face nodes have value 0 and all averages of two old nodes connected by an edge have value 0. Therefore all new edge nodes have value zero and so do the new vertex nodes: $(Q + 2R + (N - 3)S)/N = (0 + 0 + 0S)/3 = 0$. □

Figure 4.3, right, illustrates dependence as the nonuniqueness of the control net for a given surface. Interestingly, an early version of the Catmull–Clark subdivision algorithm, quoted by Doo and Sabin [7], can be shown to be locally linearly independent for $N = 3$. Here a new vertex node of valence N is computed as $(Q+R+2S)/4$.

In general, on $\cup_{i=2}^{\infty}\Omega_i$, the nodal functions associated with the mesh for $N = 3$ are locally linearly dependent as illustrated in Figure 4.4. On Ω_1, if at least one node involved has valence $N \neq 3$, then $\mathbf{A}$ has full rank and the nodal functions are linearly independent. This implies global linear independence.

Therefore we have:

LEMMA 4.2. *The nodal functions of Catmull–Clark subdivision with support on* $\mathcal{A}$ *are linearly independent over* $\mathcal{A}$ *unless their nodes all have valence* $N = 3$.

5. Local linear independence of Catmull–Clark subdivision. Since the valence N can be arbitrary but each layer of the subdivision function corresponding to a region Ω_ℓ is defined by a finite number of B-spline control points, the nodal functions of Catmull–Clark subdivision cannot in general be locally linearly independent over any subset of Ω.

LEMMA 5.1. *The nodal functions of Catmull–Clark subdivision are locally linearly independent if and only if* $N = 4$.

Proof. For $N = 3$, the nodal functions are not linearly independent over $\cup_{i=2}^{\infty}\Omega_i$ due to the example given in Figure 4.4.

If $N = 4$, the local linearly independence follows from the local linearly independence of tensor product B-splines.

For $N = 5$, the nodal functions are independent on Ω_1, (which implies linear independence on $\cup_{\ell=1}^{k}\Omega_\ell$ since all eigenvalues are positive) and on any subset of Ω_1 that straddles at least two of the three quad subdomains of Ω_1 on which the subdivision surface is a single polynomial. However, due to the dimension of a polynomial piece (16), on any single one of the subdomains, the $2N+8$ nodal functions must be linearly dependent.

For $N > 5$, the nodal functions are linearly dependent on Ω_1. ∎

Just as for Loop subdivision, for any k there exists a valence N so that the nodal functions ν_i of Catmull–Clark subdivision with support on Ω_k are locally linearly dependent on Ω_k and even on $\cup_{\ell=1}^{k}\Omega_\ell$. The pattern, verified by symbolic calculation for isolated extraordinary nodes up to $N = 20$, is as follows.

CONJECTURE 2. *For* $k := N - 4 > 0$, *the nodal functions of Catmull–Clark subdivision* $\nu_i, i = 1 \dots 2N + 8$ *are linearly independent on* $\cup_{i=1}^{k}\Omega_i$ *but linearly dependent on* $\cup_{i=1}^{k-1}\Omega_i$.

Next, we show that this characterization of the localness of linear independence is sharp: once we take the union of regions to the limit Ω, the nodal functions are linearly independent regardless of valence. As before, we first prove independence over Ω of the eigenfunctions defined by the column vectors in $\mathbf{V}$. Then we conclude independence of the nodal functions for Catmull–Clark subdivision over Ω. As always, to be scalable, the control net of an eigenfunction is only well defined if the extraordinary node is isolated.

LEMMA 5.2. *The eigenfunctions of Catmull–Clark subdivision are linearly independent over* Ω.

Proof. For $N > 3$, analogous to the proof of Lemma 3.3, we can reduce the problem to the independence of the eigenfunctions associated with the same eigenvalue.

According to [1, 2, 10], the eigenvalues of $\mathbf{A}_{11}$,

$$\lambda_k := \frac{1}{16}(C_k + 5 \pm \sqrt{(C_k + 9)(C_k + 1)}), \quad k = 1, \dots N - 1,$$

each have a multiplicity of two. We have $\lambda_k \neq 0$ since $(C_k + 5)^2 \neq (C_k + 9)(C_k + 1)$.

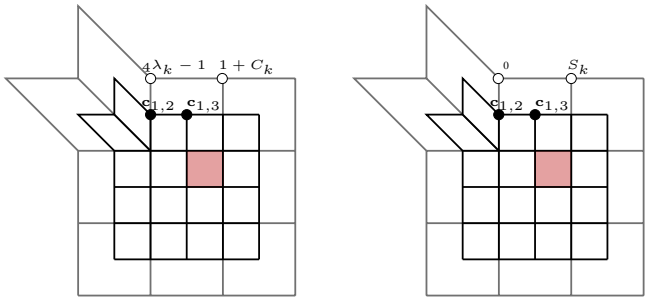


FIG. 5.1. *The B-spline control points* $\mathbf{c}_{1,2}, \mathbf{c}_{1,3}$ *(red points) used to certify that the eigenfunctions associated with* $\mathbf{u}_k$ *and* $\mathbf{w}_k$ *are independent.*

When $k \neq N/2$, the associated eigenvectors $\mathbf{u}_k$ and $\mathbf{w}_k$ are (see [17])

$$\mathbf{u}_k = \begin{pmatrix} 0 \\ 4\lambda_k - 1 \\ 1 + C_k \\ (4\lambda_k - 1)C_k \\ C_k + C_{2k} \\ \vdots \\ (4\lambda_k - 1)C_{(N-1)k} \\ C_{(N-1)k} + 1 \end{pmatrix} \quad \text{and} \quad \mathbf{w}_k = \begin{pmatrix} 0 \\ 0 \\ S_k \\ (4\lambda_k - 1)S_k \\ S_k + S_{2k} \\ \vdots \\ (4\lambda_k - 1)S_{(N-1)k} \\ S_{(N-1)k} \end{pmatrix},$$

where $C_k := \cos(2\pi k/N)$ and $S_k := \sin(2\pi k/N)$. To show that the two eigenfunctions defined by $\mathbf{u}_k$ and $\mathbf{w}_k$ are linearly independent, we consider the tensor-product B-spline entries $\mathbf{c}_{1,2}$ and $\mathbf{c}_{1,3}$ of $\mathbf{c}_1|_{\mathbf{c}_0=\mathbf{u}_k}$ and $\mathbf{c}_1|_{\mathbf{c}_0=\mathbf{w}_k}$ (solid dots in Figure 5.1). The two eigenfunctions are linearly independent over the shaded region if they generate independent B-spline control points $\mathbf{c}_{1,2}$ and $\mathbf{c}_{1,3}$, i.e., if

$$\det \begin{pmatrix} \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{u}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{u}_k} \\ \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{w}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{w}_k} \end{pmatrix} = \lambda_k^2 \det \begin{pmatrix} 4\lambda_k - 1 & 1 + C_k \\ 0 & S_k \end{pmatrix} \neq 0.$$

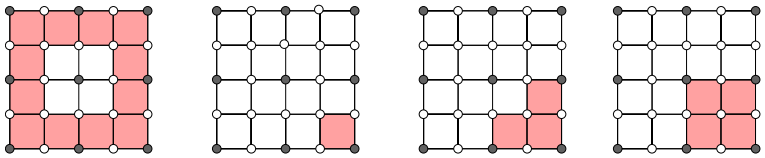
In fact,

$$\begin{aligned} & 4\lambda_k - 1 \neq 0 \\ \iff & \frac{1}{4}(C_k + 5 \pm \sqrt{(C_k + 9)(C_k + 1)}) \neq 1, \\ \iff & (C_k + 5) \pm \sqrt{(C_k + 9)(C_k + 1)}) \neq 4, \\ \iff & (\pm\sqrt{(C_k + 9)(C_k + 1)}) \neq -1 - C_k, \\ \iff & (C_k + 9)(C_k + 1) \neq (-1 - C_k)^2, \\ \iff & 8C_k + 8 \neq 0, \\ \iff & C_k \neq -1. \end{aligned}$$

$C_k \neq -1$ and $S_k \neq 0$ follows from $k \neq N/2$.

When $k = N/2$, the eigenvectors of $\lambda_k = \frac{1}{4}$ are

$$\begin{aligned} \mathbf{u}_k^T &= (0, 1, 0, -1, 0, 1, 0, \ldots, -1, 0, \ldots) \quad \text{and} \\ \mathbf{w}_k^T &= (0, 0, 1, 0, -1, 0, 1, \ldots, 0, -1, \ldots), \end{aligned}$$



$\forall N$ (Lemma 4.2) | $N \le 4$ (Lemma 5.1) | $N \le 5$ (Conjecture 2) | $\forall N$ (Theorem 5.3)

FIG. 5.2. *Summary of findings for Catmull–Clark subdivision. Domains G (shaded) and valence N for which the nodal functions with support on G are linearly independent.*

then

$$\det \begin{pmatrix} \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{u}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{u}_k} \\ \mathbf{c}_{1,2}|_{\mathbf{c}_0=\mathbf{w}_k} & \mathbf{c}_{1,3}|_{\mathbf{c}_0=\mathbf{w}_k} \end{pmatrix} = \lambda_k^2 \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \neq 0.$$

For the eigenvalues of $\mathbf{A}_{22}$, $\{\frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{64}\}$, the eigenfunctions are the tensor-product power basis functions [17]

$$\{u^3, v^3, u^3v, uv^3, u^3v^2, u^2v^3, u^3v^3\}$$

whose pairwise independence is well known.

For the special case $N = 3$, there is a zero eigenvalue and, as illustrated in Figure 4.4, the associated eigenfunction has zero values on $\cup_{i=2}^{\infty}\Omega_i$. But the remaining eigenfunctions are linearly independent on $\cup_{i=2}^{\infty}\Omega_i$ and, if at least one neighbor of the central node has valence $N \neq 3$, the zero eigenfunction has nonzero values on Ω_1. ∎

Since the transformation between the eigenfunctions and nodal functions are invertible, provided the extraordinary node is surrounded by regular nodes, all such nodal functions are linearly independent and form a basis when the eigenfunctions do. If the nodal functions were the result of one refinement, then full rank of $\mathbf{A}_{11}$ (unless all relevant nodes have valence $N = 3$) establishes the main result.

THEOREM 5.3. *The nodal functions of Catmull–Clark subdivision that have support on Ω are linearly independent over Ω unless their nodes all have valence $N = 3$. The findings are summarized in Figure* 5.2.

6. Primal schemes. We now collect the key ideas of the preceding analyses. An extraordinary node is a control net node of valence different from the "regular" majority. A subdivision scheme is called primal if there exists a sequence of extraordinary nodes converging towards each extraordinary point. While the typical application of subdivision creates a parametrized surface in $\mathbb{R}^3$, for the analysis, it is sufficient to look at one spatial coordinate since the coordinates do not interact.

Near extraordinary points, subdivision surfaces can be understood as an infinite union of nested surface rings [15]. An extraordinary node is *isolated* if at least one surface ring separates it from any other extraordinary node. If the extraordinary node is isolated, then $\mathcal{A} := \{1, \ldots, N\} \times \Omega_1$, the domain of the first surface ring, is well defined. We denote by $\{g_i\}$ the nodal functions corresponding to regular nodes that generate the first subdivision surface ring. For Loop's subdivision, the g_i are three-direction box-splines. For Catmull–Clark subdivision, they are tensor-product bicubic B-splines. Denote by A_* the refinement matrix from the input control net to a control net with isolated extraordinary node. This may be the identity if all extraordinary nodes are already isolated. To show that the nodal functions of a subdivision scheme are (globally) linearly independent, we focus on $\mathcal{A}$.

LEMMA 6.1 (global linear independence). *If the functions g_i are linearly independent and the matrix A_* is of full rank, then the nodal functions ν_j of the subdivision are globally linearly independent.*

Proof. By assumption, the nodal functions g_i of the once-refined control net are linearly independent on the union of all local domain rings $\mathcal{A}$. Since A_* is of full rank and, as a refinement matrix, its rank is less than the number of g_i, the coefficients of the input control net are zero. ∎

For example, if the functions g_i are four-direction box-splines, then the nodal functions are not globally independent since this space of splines has dependent generating g_i [6]. The example of Figure 4.3 underscores the need for the assumption on A_*—a counting argument shows that typical schemes are not locally linearly independent.

LEMMA 6.2 (local dependence for high valence). *Let ν_0 be the nodal function associated with an extraordinary node of valence N and ν_ℓ, $\ell = 1, \dots, N$ the nodal functions of its direct neighbor nodes. Denote by x_i^m the ith segment of the mth surface ring and assume that x_i^m belongs to a space of fixed finite dimension that is independent of N. If each ν_j has support on $\Omega_k \, \forall \, k \geq k_0$, then there exists $m \in \mathbb{Z}$ so that ν_ℓ, $\ell = 0, \dots, n$ are* locally linearly dependent *on Ω_m.*

Proof. For sufficiently large valence N, the number of ν_j with support on the domain of x_i^m exceeds the fixed dimension of the space from which x_i^m is drawn. ∎

The assumptions of the lemma hold in particular for symmetric C^1 subdivision schemes derived from box-splines since each x_i^m is finitely generated and the ν_j must all interact at the extraordinary point to guarantee smoothness and partition of unity. For such schemes, nodal functions contribute to the whole spline ring x^m of high valence only after a number of subdivision steps k_0.

As the example in Figure 4.4 shows, failure of the counting argument for low valences does not imply that the nodal functions are locally linearly independent. To establish local linear independence and also to find the m in Lemma 6.2 beyond which the nodal functions are locally linearly dependent, requires an analysis specific to each scheme.

To investigate linear independence on the domains Ω, corresponding to control facets surrounding an extraordinary node, we use eigenfunctions.

LEMMA 6.3 (linear independence of eigenfunctions). *The eigenfunctions φ_i, $i = 1, \dots, k$ are linearly independent on Ω if the eigenfunctions corresponding to each eigenvalue, separately, are independent.*

Proof. Let λ_i be sorted by absolute value and $\lambda_0 = 1$. Suppose there exist scalars $a_1, a_2, \dots a_k$, not all zero, such that $\sum_{i=1}^k a_i \varphi_i = 0$. Let λ_j be an absolute largest eigenvalue such that $a_j \neq 0$. If $\lambda_j = 0$, then the eigenfunctions to the eigenvalue 0 are dependent contradicting the assumptions. If $\lambda_j \neq 0$, then, with $w_i := -a_i/a_j$, we can write $\varphi_j = \sum_{i=j+1}^k w_i \varphi_i$. and, by the defining property of eigenfunctions, for any $(u, v) \in \mathcal{A}$,

$$\varphi_j\left(\frac{u}{2^m}, \frac{v}{2^m}\right) = \sum_{i=j+1}^{k} w_i \varphi_i\left(\frac{u}{2^m}, \frac{v}{2^m}\right) \quad \Rightarrow \quad \varphi_j(u, v) = \sum_{i=j+1}^{k} w_i \left(\frac{\lambda_i}{\lambda_j}\right)^m \varphi_i(u, v).$$

Since the equality has to hold for all m and $(\frac{\lambda_i}{\lambda_j})^m$ goes to zero or repeatedly changes sign, this implies $\varphi_j = \sum_{i:\lambda_i = \lambda_j} w_i \varphi_i$. That is, the eigenfunctions associated with the same eigenvalue λ_j must be linearly dependent in contradiction to the assumption. ∎

While the, possibly generalized, eigenvectors of an eigenvalue are independent, checking that the corresponding eigenfunctions are independent requires an analysis

specific to each scheme.

We can now characterize when the nodal functions ν_i are linearly independent.

COROLLARY 6.4. *If the refinement matrix A_* from the input mesh to a mesh with isolated extraordinary nodes is of full rank and the scheme has only effective and nonzero eigenvalues and the eigenfunctions corresponding to each eigenvalue, separately, are independent, then the nodal functions $\nu_i, i = 1 \dots K$ are linearly independent on Ω.*

Proof. If the extraordinary node is isolated, then nodal functions and eigenfunctions are related by the invertible matrix V of the generalized eigenvectors of the Jordan decomposition of the subdivision matrix:

$$[\varphi_1, \dots, \varphi_k] = V[\nu_1, \dots, \nu_k].$$

The general case follows by subdivision and the full rank of A_*. ∎

Corollary 6.4 and Lemma 6.2 characterize the locality of linear independence of subdivision schemes. On any finite union of Ω_ℓ, the nodal functions are linearly dependent for sufficiently high valence. On the infinite union Ω, the nodal functions are typically linearly independent provided the underlying regular spline functions g_i are. Subtle exceptions, illustrated by the examples in Figures 4.3 and 4.4, underscore the need for the assumption on A_*.

7. Summary and related open issues. The characterization of linear independence is a vital part of the foundations of generalized subdivision and illuminates the numerical properties of the nodal functions. It implies a cautionary note for computational use: near vertices of high valence, the nodal functions may not be linearly independent.

If the nodal functions are linearly independent, hence form a basis, we can additionally analyze the *condition* of the basis. The condition number of a nodal basis can be defined analogous to the condition of the B-spline basis (see, e.g., [5]). This concept is not to be confused with the condition number of the subdivision matrix but *quantifies* linear independence of the basis. At present, the notion of condition has not been explored in the context of subdivision surfaces although some implications are familiar: the poor condition of the Loop nodal basis for low valences explains, for example, why designers need to strongly exaggerate some features in the Loop control net.

Acknowledgments. The presentation benefited from the kind suggestions of the reviewers and the editor.

REFERENCES

[1] A. A. BALL AND D. J. T. STORRY, *Conditions for tangent plane continuity over recursively generated B-spline surfaces*, ACM Trans. Graphs., 7 (1988), pp. 83–102.

[2] A. A. BALL AND D. J. T. STORRY, *An investigation of curvature variations over recursively generated B-spline surfaces*, ACM Trans. Graphs., 9 (1990), pp. 424–437.

[3] E. CATMULL AND J. CLARK, *Recursively generated B-spline surfaces on arbitrary topological meshes*, Comput. Aided Design, 10 (1978), pp. 350–355.

[4] F. CIRAK, M. ORTIZ, AND P. SCHRÖDER, *Subdivision surfaces: A new paradigm for thin-shell finite-element analysis*, Internat. J. Numer. Methods Engrg., 47 (2000), pp. 2039–2072.

[5] C. DE BOOR, *The exact condition of the B-spline basis may be hard to determine*, J. Approx. Theory, 60 (1990), pp. 344–359.

[6] C. DE BOOR, K. HÖLLIG, AND S. RIEMENSCHNEIDER, *Box Splines*, Appl. Math. Sci. 98, Springer-Verlag, New York, 1993.

[7] D. Doo and M. Sabin, *Behavior of recursive division surfaces near extraordinary points*, Comput. Aided Design, 10 (1978), pp. 356–360.

[8] E. Grinspun, *The Basis Refinement Method*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2003, p. 39.

[9] E. Grinspun, P. Krysl, and P. Schröder, *CHARMS: A simple framework for adaptive simulation*, in SIGGRAPH 2002 Conference Proceedings, J. Hughes, ed., Annual Conference Series, ACM Press/ACM SIGGRAPH, New York, 2002, pp. 281–290.

[10] K. Karciauskas, J. Peters, and U. Reif, *Shape characterization of subdivision surfaces—case studies*, 21 (2004), pp. 601–614; available online from http://authors.elsevier.com/sd/article/ S0167839604000627.

[11] C. T. Loop, *Smooth Subdivision Surfaces Based on Triangles*, Master's thesis, Department of Mathematics, University of Utah, Salt Lake City, UT, 1987.

[12] M. Lounsbery, T. D. DeRose, and J. Warren, *Multiresolution analysis for surfaces of arbitrary topological type*, ACM Trans. Graphs., 16 (1997), pp. 34–73.

[13] A. H. Nasri, *Polyhedral subdivision methods on free-form surfaces*, ACM Trans. Graphs., 6 (1987), pp. 29–73.

[14] J. Peters and U. Reif, *Shape characterization of subdivision surfaces—basic principles*, Comput. Aided Geometric Design, 21 (2004), pp. 585–599.

[15] U. Reif, *A unified approach to subdivision algorithms near extraordinary vertices*, Comput. Aided Geometric Design, 12 (1995), pp. 153–174.

[16] J. Stam, *Evaluation of Loop subdivision surfaces*, SIGGRAPH 1998 Proceedings Notes, Orlando, FL, 1998.

[17] J. Stam, *Exact evaluation of Catmull-Clark subdivision surfaces at arbitrary parameter values*, in SIGGRAPH 1998 Conference Proceedings, Orlando, FL, Michael Cohen, ed., Addison Wesley, Reading, MA, 1998, pp. 395–404.

[18] L. Velho and D. Zorin, 4-8 *subdivision*, Comput. Aided Geometric Design, 18 (2001), pp. 397–427.

[19] D. Zorin, *Smoothness of stationary subdivision on irregular meshes*, Constr. Approx., 16 (2000), pp. 359–397.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2408–2431

DISCONTINUOUS GALERKIN FINITE ELEMENT METHOD FOR THE WAVE EQUATION*

MARCUS J. GROTE†, ANNA SCHNEEBELI†, AND DOMINIK SCHÖTZAU‡

Abstract. The symmetric interior penalty discontinuous Galerkin finite element method is presented for the numerical discretization of the second-order wave equation. The resulting stiffness matrix is symmetric positive definite, and the mass matrix is essentially diagonal; hence, the method is inherently parallel and leads to fully explicit time integration when coupled with an explicit time-stepping scheme. Optimal a priori error bounds are derived in the energy norm and the L^2-norm for the semidiscrete formulation. In particular, the error in the energy norm is shown to converge with the optimal order $\mathcal{O}(h^{\min\{s,\ell\}})$ with respect to the mesh size h, the polynomial degree ℓ, and the regularity exponent s of the continuous solution. Under additional regularity assumptions, the L^2-error is shown to converge with the optimal order $\mathcal{O}(h^{\ell+1})$. Numerical results confirm the expected convergence rates and illustrate the versatility of the method.

Key words. discontinuous Galerkin finite element methods, wave equation, acoustic waves, second-order hyperbolic problems, a priori error analysis, explicit time integration

AMS subject classification. 65N30

DOI. 10.1137/05063194X

1. Introduction.

The numerical solution of the wave equation is of fundamental importance to the simulation of time dependent acoustic, electromagnetic, or elastic waves. For such wave phenomena the scalar second-order wave equation often serves as a model problem. Finite element methods (FEMs) can easily handle inhomogeneous media or complex geometry. However, if explicit time-stepping is subsequently employed, the mass matrix arising from the spatial discretization by standard continuous finite elements must be inverted at each time step: a major drawback in terms of efficiency. For low-order Lagrange ($\mathcal{P}^1$) elements, so-called mass lumping overcomes this problem [6, 15], but for higher-order elements this procedure can lead to unstable schemes unless particular finite elements and quadrature rules are used [11]. In addition, continuous Galerkin methods impose significant restrictions on the underlying mesh and discretization; in particular, they do not easily accommodate hanging nodes.

To avoid these difficulties, we consider instead discontinuous Galerkin (DG) methods. Based on discontinuous finite element spaces, these methods easily handle elements of various types and shapes, irregular nonmatching grids, and even locally varying polynomial order; thus, they are ideally suited for *hp*-adaptivity. Here continuity is weakly enforced across mesh interfaces by adding suitable bilinear forms, so-called numerical fluxes, to standard variational formulations. These fluxes are easily included within an existing conforming finite element code.

*Received by the editors May 19, 2005; accepted for publication (in revised form) May 7, 2006; published electronically December 1, 2006.

http://www.siam.org/journals/sinum/44-6/63194.html

†Department of Mathematics, University of Basel, Rheinsprung 21, 4051 Basel, Switzerland (Marcus.Grote@unibas.ch, anna.schneebeli@unibas.ch). The second author was supported by the Swiss National Science Foundation.

‡Mathematics Department, University of British Columbia, 121–1984 Mathematics Road, Vancouver V6T 1Z2, BC, Canada (schoetzau@math.ubc.ca). This author was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Because individual elements decouple, DGFEMs are also inherently parallel; see [8, 9, 10, 7] for further details and recent reviews. Moreover, the mass matrix arising from the spatial DG discretization is block-diagonal, with block size equal to the number of degrees of freedom per element; it can therefore be inverted at very low computational cost. In fact, for a judicious choice of (locally orthogonal) shape functions, the mass matrix is diagonal. When combined with explicit time integration, the resulting time marching scheme will be fully explicit.

The origins of DG methods can be traced back to the 1970s, when they were proposed for the numerical solution of hyperbolic neutron transport equations, as well as for the weak enforcement of continuity in Galerkin methods for elliptic and parabolic problems; see Cockburn, Karniadakis, and Shu [8] for a review of the development of DG methods. When applied to second-order hyperbolic problems, most DG methods first require the problem to be reformulated as a first-order hyperbolic system, for which various DG methods are available. In [9], for instance, Cockburn and Shu used a DGFEM in space combined with a Runge–Kutta scheme in time to discretize hyperbolic conservation laws. Hesthaven and Warburton [13] used the same approach to implement high-order methods for Maxwell's equations in first-order hyperbolic form. Space-time DG methods for linear symmetric first-order hyperbolic systems were presented by Falk and Richter in [12] and later generalized by Monk and Richter in [17] and by Houston, Jensen, and Süli in [14]. A first DG method for the acoustic wave equation in its original second-order formulation was recently proposed by Rivière and Wheeler [21]; it is based on a *nonsymmetric* interior penalty formulation and requires additional stabilization terms for optimal convergence in the L^2-norm [20].

Here we propose and analyze the *symmetric* interior penalty DG method for the spatial discretization of the second-order scalar wave equation. In particular, we shall derive optimal a priori error bounds in the energy norm and the L^2-norm for the semidiscrete formulation. Besides the well-known advantages of DG methods mentioned above, a symmetric discretization of the wave equation in its second-order form offers an additional advantage, which also pertains to the classical continuous Galerkin formulation: since the stiffness matrix is positive definite, the semidiscrete formulation conserves (a discrete version of) the energy for all time; thus, it is free of any (unnecessary) damping. The dispersive properties of the symmetric interior penalty DG method were recently analyzed by Ainsworth, Monk, and Muniz [1].

The outline of our paper is as follows. In section 2 we describe the setting of our model problem. Next, we present in section 3 the symmetric interior penalty DG method for the wave equation. Our two main results, optimal error bounds in the energy norm and the L^2-norm for the semidiscrete scheme, are stated at the beginning of section 4 and proved subsequently. The analysis relies on an idea suggested by Arnold et al. [2] together with the approach presented by Perugia and Schötzau in [18] to extend the DG bilinear forms by suitable lifting operators. In section 5, we demonstrate the sharpness of our theoretical error estimates by a series of numerical experiments. By combining our DG method with the second-order Newmark scheme, we obtain a fully discrete method. To illustrate the versatility of our method, we also propagate a wave across an inhomogenous medium with discontinuity, where the underlying finite element mesh contains hanging nodes. Finally, we conclude with some remarks on possible extensions of our DG method to electromagnetic and elastic waves.

2. Model problem. We consider the (second-order) scalar wave equation

$$u_{tt} - \nabla \cdot (\, c\, \nabla u) = f \qquad \text{in } J \times \Omega, \tag{2.1}$$

$$u = 0 \qquad \text{on } J \times \partial\Omega, \tag{2.2}$$

$$u|_{t=0} = u_0 \qquad \text{in } \Omega, \tag{2.3}$$

$$u_t|_{t=0} = v_0 \qquad \text{in } \Omega, \tag{2.4}$$

where $J = (0, T)$ is a finite time interval and Ω is a bounded domain in $\mathbb{R}^d$, $d = 2, 3$. For simplicity, we assume that Ω is a polygon ($d = 2$) or a polyhedron ($d = 3$). The (known) source term f lies in $L^2(J; L^2(\Omega))$, while $u_0 \in H_0^1(\Omega)$ and $v_0 \in L^2(\Omega)$ are prescribed initial conditions. We assume that the speed of propagation, $\sqrt{c(x)}$, is piecewise smooth and satisfies the bounds

$$0 < c_\star \le c(x) \le c^\star < \infty, \qquad x \in \overline{\Omega}. \tag{2.5}$$

The standard variational form of (2.1)–(2.4) is to find $u \in L^2(J; H_0^1(\Omega))$, with $u_t \in L^2(J; L^2(\Omega))$ and $u_{tt} \in L^2(J; H^{-1}(\Omega))$, such that $u|_{t=0} = u_0$, $u_t|_{t=0} = v_0$, and

$$\langle u_{tt}, v\rangle + a(u, v) = (f, v) \qquad \forall v \in H_0^1(\Omega) \quad \text{a.e. in } J. \tag{2.6}$$

Here, the time derivatives are understood in a distributional sense, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, $(\cdot, \cdot)$ is the inner product in $L^2(\Omega)$, and $a(\cdot, \cdot)$ is the elliptic bilinear form given by

$$a(u, v) = (c\, \nabla u, \nabla v). \tag{2.7}$$

It is well known that problem (2.6) is well posed [16]. Moreover, the weak solution u can be shown to be continuous in time; that is,

$$u \in C^0(\overline{J}; H_0^1(\Omega)), \qquad u_t \in C^0(\overline{J}; L^2(\Omega)); \tag{2.8}$$

see [16, Chapter III, Theorems 8.1 and 8.2] for details. In particular, this result implies that the initial conditions in (2.3) and (2.4) are well defined.

3. Discontinuous Galerkin discretization. We shall now discretize the wave equation (2.1)–(2.4) by using the interior penalty discontinuous Galerkin finite element method in space, while leaving the time dependence continuous.

3.1. Preliminaries. We consider shape-regular meshes $\mathcal{T}_h$ that partition the domain Ω into disjoint elements $\{K\}$ such that $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} \overline{K}$. For simplicity, we assume that the elements are triangles or parallelograms in two space dimensions, and tetrahedra or parallelepipeds in three dimensions, respectively. The diameter of element K is denoted by h_K, and the mesh size h is given by $h = \max_{K \in \mathcal{T}_h} h_K$. We assume that the partition is aligned with the discontinuities of the wave speed $\sqrt{c}$. Generally, we allow for irregular meshes with hanging nodes. However, we assume that the local mesh sizes are of bounded variation; that is, there is a positive constant κ, depending only on the shape-regularity of the mesh, such that

$$\kappa h_K \le h_{K'} \le \kappa^{-1} h_K \tag{3.1}$$

for all neighboring elements K and K'.

An interior face of $\mathcal{T}_h$ is the (nonempty) interior of $\partial K^+ \cap \partial K^-$, where K^+ and K^- are two adjacent elements of $\mathcal{T}_h$. Similarly, a boundary face of $\mathcal{T}_h$ is the (nonempty)

interior of $\partial K \cap \partial\Omega$, which consists of entire faces of ∂K. We denote by $\mathcal{F}_h^{\mathcal{I}}$ the set of all interior faces of $\mathcal{T}_h$ and by $\mathcal{F}_h^{\mathcal{B}}$ the set of all boundary faces and set $\mathcal{F}_h = \mathcal{F}_h^{\mathcal{I}} \cup \mathcal{F}_h^{\mathcal{B}}$. Here we generically refer to any element of $\mathcal{F}_h$ as a "face," in both two and three dimensions.

For any piecewise smooth function v we now introduce the following trace operators. Let $F \in \mathcal{F}_h^{\mathcal{I}}$ be an interior face shared by two neighboring elements K^+ and K^- and let $x \in F$; we write $\mathbf{n}^{\pm}$ to denote the unit outward normal vectors on the boundaries $\partial K^{\pm}$. Denoting by $v^{\pm}$ the trace of v taken from within $K^{\pm}$, we define the jump and average of v at $x \in F$ by

$$[\![v]\!] := v^+\mathbf{n}^+ + v^-\mathbf{n}^-, \qquad \{\!\{v\}\!\} := (v^+ + v^-)/2,$$

respectively. On every boundary face $F \in \mathcal{F}_h^{\mathcal{B}}$, we set $[\![v]\!] := v\mathbf{n}$ and $\{\!\{v\}\!\} := v$. Here, $\mathbf{n}$ is the unit outward normal vector on $\partial\Omega$.

For a piecewise smooth vector-valued function $\mathbf{q}$, we analogously define the average across interior faces by $\{\!\{\mathbf{q}\}\!\} := (\mathbf{q}^+ + \mathbf{q}^-)/2$, and on boundary faces we set $\{\!\{\mathbf{q}\}\!\} := \mathbf{q}$. The jump of a vector-valued function will not be used. For a vector-valued function $\mathbf{q}$ with *continuous* normal components across a face f, the trace identity

$$v^+(\mathbf{n}^+ \cdot \mathbf{q}^+) + v^-(\mathbf{n}^- \cdot \mathbf{q}^-) = [\![v]\!] \cdot \{\!\{\mathbf{q}\}\!\} \quad \text{on } f$$

immediately follows from the definitions.

3.2. Discretization in space. For a given partition $\mathcal{T}_h$ of Ω and an approximation order $\ell \geq 1$, we wish to approximate the solution $u(t,\cdot)$ of (2.1)–(2.4) in the finite element space

$$V^h := \{v \in L^2(\Omega) : \, v|_K \in \mathcal{S}^\ell(K) \ \ \forall K \in \mathcal{T}_h\}, \tag{3.2}$$

where $\mathcal{S}^\ell(K)$ is the space $\mathcal{P}^\ell(K)$ of polynomials of total degree at most ℓ on K if K is a triangle or a tetrahedra, or the space $\mathcal{Q}^\ell(K)$ of polynomials of degree at most ℓ in each variable on K if K is a parallelogram or a parallelepiped.

Then, we consider the following (semidiscrete) DG approximation of (2.1)–(2.4): find $u^h : \overline{J} \times V^h \to \mathbb{R}$ such that

$$(u^h_{tt}, v) + a_h(u^h, v) = (f, v) \qquad \forall v \in V^h, \quad t \in J, \tag{3.3}$$

$$u^h|_{t=0} = \Pi_h u_0, \tag{3.4}$$

$$u^h_t|_{t=0} = \Pi_h v_0. \tag{3.5}$$

Here, Π_h denotes the L^2-projection onto V^h, and the discrete bilinear form a_h on $V^h \times V^h$ is given by

$$\begin{aligned} a_h(u,v) := & \sum_{K\in\mathcal{T}_h} \int_K c\,\nabla u \cdot \nabla v \, dx - \sum_{F\in\mathcal{F}_h} \int_F [\![u]\!] \cdot \{\!\{c\,\nabla v\}\!\}\, dA \\ & - \sum_{F\in\mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{c\,\nabla u\}\!\}\, dA + \sum_{F\in\mathcal{F}_h} \int_F \mathtt{a}\,[\![u]\!] \cdot [\![v]\!]\, dA. \end{aligned} \tag{3.6}$$

The last three terms in (3.6) correspond to jump and flux terms at element boundaries; they vanish when $u, v \in H_0^1(\Omega) \cap H^{1+\sigma}(\Omega)$ for $\sigma > \frac{1}{2}$. Hence the above semidiscrete DG formulation (3.3) is consistent with the original continuous problem (2.6).

In (3.6) the function $\mathtt{a}$ penalizes the jumps of u and v over the faces of $\mathcal{T}_h$. It is referred to as the interior penalty stabilization function and is defined as follows. We first introduce the function $\mathtt{h}$ by

$$\mathtt{h}|_F = \begin{cases} \min\{h_K, h_{K'}\}, & F \in \mathcal{F}_h^{\mathcal{I}},\ F = \partial K \cap \partial K', \\ h_K, & F \in \mathcal{F}_h^{\mathcal{B}},\ F = \partial K \cap \partial\Omega. \end{cases}$$

For $x \in F$, we further define $\mathtt{c}$ by

$$\mathtt{c}|_F(x) = \begin{cases} \max\{c|_K(x), c|_{K'}(x)\}, & F \in \mathcal{F}_h^{\mathcal{I}},\ F = \partial K \cap \partial K', \\ c|_K(x), & F \in \mathcal{F}_h^{\mathcal{B}},\ F = \partial K \cap \partial\Omega. \end{cases}$$

Then, on each $F \in \mathcal{F}_h$, we set

$$\mathtt{a}|_F := \alpha\, \mathtt{ch}^{-1}, \tag{3.7}$$

where α is a positive parameter independent of the local mesh sizes and the coefficient c.

To conclude this section we recall the following stability result for the DG form a_h.

LEMMA 3.1. *There exists a threshold value $\alpha_{\min} > 0$ which depends only on the shape-regularity of the mesh, the approximation order ℓ, the dimension d, and the bounds in* (2.5)*, such that for $\alpha \geq \alpha_{\min}$*

$$a_h(v,v) \geq C_{\mathrm{coer}} \left(\sum_{K \in \mathcal{T}_h} \|c^{\frac{1}{2}} \nabla v\|_{0,K}^2 + \sum_{F \in \mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}} [\![v]\!]\|_{0,F}^2 \right), \qquad v \in V^h,$$

where the constant C_{coer} is independent of c and h.

The proof of this lemma follows readily from the arguments in [2]. However, to make explicit the dependence of $\alpha_{\min}$ on the bounds in (2.5), we present the proof of a slightly more general stability result in Lemma 4.4 below. Throughout the rest of the paper we shall assume that $\alpha \geq \alpha_{\min}$, so that by Lemma 3.1 the semidiscrete problem (3.3)–(3.5) has a unique solution.

We remark that the condition $\alpha \geq \alpha_{\min}$ can be omitted by using other symmetric DG discretizations of the div-grad operator, such as the local discontinuous Galerkin (LDG) method; see, e.g., [2] for details. It can also be avoided by using the nonsymmetric interior penalty method proposed in [20]. However, since the symmetry of a_h is crucial in the analysis below, our error estimates (section 4) do not hold for the nonsymmetric DG method in [20].

Remark 3.2. Because the bilinear form a_h is symmetric and coercive, for $\alpha \geq \alpha_{\min}$, the semidiscrete DG formulation (3.3)–(3.5) with $f = 0$ conserves the (discrete) energy

$$E_h(t) := \frac{1}{2}\|u_t^h(t)\|_0^2 + \frac{1}{2} a_h(u^h(t), u^h(t)).$$

4. A priori error estimates. We shall now derive optimal a priori error bounds for the DG method (3.3)–(3.5), first with respect to the DG energy norm and then with respect to the L^2-norm. These two key results are stated immediately below, while their proofs are postponed to subsequent sections.

4.1. Main results. To state our a priori error bounds, we define the space

$$V(h) = H_0^1(\Omega) + V^h.$$

On $V(h)$, we define the DG energy norm

$$\|v\|_h^2 := \sum_{K\in\mathcal{T}_h} \|c^{\frac{1}{2}}\nabla v\|_{0,K}^2 + \sum_{F\in\mathcal{F}_h} \|\mathsf{a}^{\frac{1}{2}}[\![v]\!]\|_{0,F}^2.$$

Furthermore, for $1 \le p \le \infty$ we will make use of the Bochner space $L^p(J;V(h))$, endowed with the norm

$$\|v\|_{L^p(J;V(h))} = \begin{cases} \big(\int_J \|v\|_h^p\,dt\big)^{1/p}, & 1\le p<\infty,\\ \operatorname{ess\,sup}_{t\in J}\|v\|_h, & p=\infty. \end{cases}$$

Our first main result establishes an optimal error estimate of the energy norm $\|\cdot\|_h$ of the error. It also gives a bound in the $L^2(\Omega)$-norm on the error in the first time derivative.

THEOREM 4.1. *Let the analytical solution u of* (2.1)–(2.4) *satisfy*

$$u \in L^\infty(J;H^{1+\sigma}(\Omega)),\qquad u_t \in L^\infty(J;H^{1+\sigma}(\Omega)),\qquad u_{tt}\in L^1(J;H^\sigma(\Omega))$$

for a regularity exponent $\sigma > \frac{1}{2}$, and let u^h be the semidiscrete discontinuous Galerkin approximation obtained by (3.3)–(3.5), *with $\alpha \ge \alpha_{\min}$. Then, the error $e = u - u^h$ satisfies the estimate*

$$\begin{aligned}\|e_t\|_{L^\infty(J;L^2(\Omega))} + \|e\|_{L^\infty(J;V(h))} &\le C\Big[\|e_t(0)\|_0 + \|e(0)\|_h\Big]\\ &\quad + Ch^{\min\{\sigma,\ell\}}\Big[\|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + T\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + \|u_{tt}\|_{L^1(J;H^\sigma(\Omega))}\Big],\end{aligned}$$

with a constant C that is independent of T and h.

We remark that the fact that $u_t \in L^\infty(J;H^{1+\sigma}(\Omega))$ implies that u is continuous in time on $\overline{J}$ with values in $H^{1+\sigma}(\Omega)$. Similarly, $u_{tt} \in L^1(J;H^\sigma(\Omega))$ implies the continuity of u_t on $\overline{J}$ with values in $H^\sigma(\Omega)$. In Theorem 4.1 we thus implicitly assume that the initial conditions satisfy $u_0 \in H^{1+\sigma}(\Omega)$ and $v_0 \in H^\sigma(\Omega)$. Hence, standard approximation properties imply that

$$\|e_t(0)\|_0 = \|v_0 - \Pi_h v_0\|_0 \le C\,h^{\min\{\sigma,\ell+1\}}\|v_0\|_\sigma,$$

$$\|e(0)\|_h = \|u_0 - \Pi_h u_0\|_h \le C\,h^{\min\{\sigma,\ell\}}\|u_0\|_{1+\sigma};$$

see also Lemma 4.6 below. As a consequence, Theorem 4.1 yields optimal convergence in the (DG) energy norm

$$\|e_t\|_{L^\infty(J;L^2(\Omega))} + \|e\|_{L^\infty(J;V(h))} \le Ch^{\min\{\sigma,\ell\}},$$

with a constant $C = C(T)$ that is independent of h.

Next, we state an optimal error estimate with respect to the L^2-norm (in space). To do so, we need to assume elliptic regularity; that is, we assume that there is a stability constant C_S such that for any $\lambda \in L^2(\Omega)$ the solution of the problem

$$-\nabla\cdot(c\nabla z) = \lambda \quad \text{in } \Omega,\qquad z = 0 \quad \text{on } \Gamma, \tag{4.1}$$

belongs to $H^2(\Omega)$ and satisfies the stability bound

$$\|z\|_2 \le C_S \|\lambda\|_0. \tag{4.2}$$

This condition is certainly satisfied for convex domains and smooth coefficients. Then, the following L^2-error bound holds.

THEOREM 4.2. *Assume elliptic regularity as in* (4.1)–(4.2)*, and let the analytical solution* u *of* (2.1)–(2.4) *satisfy*

$$u \in L^\infty(J; H^{1+\sigma}(\Omega)), \qquad u_t \in L^\infty(J; H^{1+\sigma}(\Omega)), \qquad u_{tt} \in L^1(J; H^\sigma(\Omega))$$

for a regularity exponent $\sigma > \frac{1}{2}$*. Let* u^h *be the semidiscrete DG approximation obtained by* (3.3)–(3.5) *with* $\alpha \ge \alpha_{\min}$*. Then, the error* $e = u - u^h$ *satisfies the estimate*

$$\|e\|_{L^\infty(J;L^2(\Omega))} \le C h^{\min\{\sigma,\ell\}+1} \big[\|u_0\|_{1+\sigma} + \|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + T\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))} \big],$$

with a constant C *that is independent of* T *and the mesh size.*

For smooth solutions, Theorem 4.2 thus yields optimal convergence rates in the L^2-norm:

$$\|e\|_{L^\infty(J;L^2(\Omega))} \le C h^{\ell+1},$$

with a constant C that is independent of h.

The rest of this section is devoted to the proofs of Theorems 4.1 and 4.2. We shall first collect preliminary results in section 4.2. In section 4.3, we present the proof of Theorem 4.1. Following an argument by Baker [3] for conforming finite element approximations, we shall then derive the estimate of Theorem 4.2 in section 4.4.

4.2. Preliminaries.

Extension of the DG form a_h. The DG form a_h in (3.6) does not extend in a standard way to a continuous form on the (larger) space $V(h) \times V(h)$. Indeed the average $\{\!\{c\nabla v\}\!\}$ on a face $F \in \mathcal{F}_h$ is not well defined in general for $v \in H^1(\Omega)$. To circumvent this difficulty, we shall extend the form a_h in a nonstandard and nonconsistent way to the space $V(h) \times V(h)$ by using the lifting operators from [2] and the approach in [18]. Thus, for $v \in V(h)$ we define the lifted function, $\mathcal{L}_c(v) \in \big(V^h\big)^d$, $d = 2, 3$, by requiring that

$$\int_\Omega \mathcal{L}_c(v) \cdot \mathbf{w}\, dx = \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{c\mathbf{w}\}\!\}\, dA, \qquad \mathbf{w} \in \big(V^h\big)^d,$$

where c is the material coefficient from (2.1). We shall now show that the lifting operator $\mathcal{L}_c$ is stable in the DG norm; see [18] for a similar result for the LDG method.

LEMMA 4.3. *There exists a constant* C_{inv} *which depends only on the shape-regularity of the mesh, the approximation order* ℓ*, and the dimension* d *such that*

$$\|\mathcal{L}_c(v)\|_0^2 \le \alpha^{-1} c^\star\, C_{\mathrm{inv}}^2 \sum_{F \in \mathcal{F}_h} \|\mathsf{a}^{\frac{1}{2}} [\![v]\!]\|_{0,F}^2$$

for any $v \in V(h)$.

Moreover, if the speed of propagation $c^{\frac{1}{2}}$ is piecewise constant, with discontinuities aligned with the finite element mesh $\mathcal{T}_h$, then

$$\|c^{-\frac{1}{2}}\mathcal{L}_c(v)\|_0^2 \le \alpha^{-1}C_{\mathrm{inv}}^2 \sum_{F\in\mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}}[\![v]\!]\|_{0,F}^2.$$

Proof. We have

$$\begin{aligned}
\|\mathcal{L}_c(v)\|_0 &= \max_{\mathbf{w}\in(V^h)^d} \frac{\sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\cdot\{\!\{c\mathbf{w}\}\!\}\,dA}{\|\mathbf{w}\|_0}\\
&\le \max_{\mathbf{w}\in(V^h)^d} \frac{\left(\sum_{F\in\mathcal{F}_h}\int_F \mathtt{a}[\![v]\!]^2\,dA\right)^{\frac{1}{2}}\left(\sum_{F\in\mathcal{F}_h}\int_F \mathtt{a}^{-1}|\{\!\{c\mathbf{w}\}\!\}|^2\,dA\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}\\
&\le \alpha^{-\frac{1}{2}} \max_{\mathbf{w}\in(V^h)^d} \frac{\left(\sum_{F\in\mathcal{F}_h}\int_F \mathtt{a}[\![v]\!]^2\,dA\right)^{\frac{1}{2}}\left(\sum_{F\in\mathcal{F}_h}\int_F \mathtt{hc}^{-1}|\{\!\{c\mathbf{w}\}\!\}|^2\,dA\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}\\
&\le \alpha^{-\frac{1}{2}}\,(c^\star)^{\frac{1}{2}} \max_{\mathbf{w}\in(V^h)^d} \frac{\left(\sum_{F\in\mathcal{F}_h}\int_F \mathtt{a}[\![v]\!]^2\,dA\right)^{\frac{1}{2}}\left(\sum_{K\in\mathcal{T}_h} h_K\int_{\partial K}|\mathbf{w}|^2\,dA\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}.
\end{aligned}$$

Here, we have used the Cauchy–Schwarz inequality, the definition of $\mathtt{a}$ in (3.7), and the upper bound for c in (2.5). We recall the inverse inequality

$$\|\mathbf{w}\|_{0,\partial K}^2 \le C_{\mathrm{inv}}^2 h_K^{-1}\|\mathbf{w}\|_{0,K}^2, \qquad \mathbf{w}\in\left(\mathcal{S}^\ell(K)\right)^d, \tag{4.3}$$

with a constant C_{inv} that depends only on the shape-regularity of the mesh, the approximation order ℓ, and the dimension d. Using this bound, we obtain

$$\left(\sum_{K\in\mathcal{T}_h} h_K\int_{\partial K}|\mathbf{w}|^2\,dA\right)^{\frac{1}{2}} \le C_{\mathrm{inv}}\|\mathbf{w}\|_0,$$

which shows the first statement.

With $c^{\frac{1}{2}}$ piecewise constant, we have $c^{-\frac{1}{2}}z\in\left(V^h\right)^d$ for all $z\in\left(V^h\right)^d$. Hence, we obtain as before

$$\begin{aligned}
\|c^{-\frac{1}{2}}\mathcal{L}_c(v)\|_0 &= \max_{\mathbf{w}\in(V^h)^d} \frac{\sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\cdot\{\!\{c^{\frac{1}{2}}\mathbf{w}\}\!\}\,dA}{\|\mathbf{w}\|_0}\\
&\le \alpha^{-1}C_{\mathrm{inv}}^2 \sum_{F\in\mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}}[\![v]\!]\|_{0,F}^2,
\end{aligned}$$

which completes the proof. ∎

Next, we introduce the auxiliary bilinear form

$$\begin{aligned}
\widetilde{a}_h(u,v) := &\sum_{K\in\mathcal{T}_h}\int_K c\,\nabla u\cdot\nabla v\,dx - \sum_{K\in\mathcal{T}_h}\int_K \mathcal{L}_c(u)\cdot\nabla v\,dx\\
&- \sum_{K\in\mathcal{T}_h}\int_K \mathcal{L}_c(v)\cdot\nabla u\,dx + \sum_{F\in\mathcal{F}_h}\int_F \mathtt{a}\,[\![u]\!]\cdot[\![v]\!]\,dA.
\end{aligned} \tag{4.4}$$

The following result establishes that $\widetilde{a}_h$ is continuous and coercive on the entire space $V(h)\times V(h)$; hence it is well defined. Furthermore, since

$$\widetilde{a}_h = a_h \quad \text{on } V^h\times V^h, \qquad \widetilde{a}_h = a \quad \text{on } H_0^1(\Omega)\times H_0^1(\Omega), \tag{4.5}$$

the form $\widetilde{a}_h$ can be viewed as an extension of the two forms a_h and a to the space $V(h) \times V(h)$.

LEMMA 4.4. *Let the interior penalty parameter* $\mathtt{a}$ *be defined as in* (3.7)*, and set*

$$\alpha_{\min} = 4\, c_\star^{-1} c^\star\, C_{\mathrm{inv}}^2$$

for a general piecewise smooth c*, and*

$$\alpha_{\min} = 4\, C_{\mathrm{inv}}^2$$

for a piecewise constant c*, with discontinuities aligned with the finite element mesh* $\mathcal{T}_h$*.* C_{inv} *is the constant from Lemma* 4.3*.*

Setting $C_{\mathrm{cont}} = 2$ *and* $C_{\mathrm{coer}} = 1/2$*, we have for* $\alpha \geq \alpha_{\min}$

$$\begin{aligned} |\widetilde{a}_h(u,v)| &\leq C_{\mathrm{cont}} \|u\|_h \|v\|_h, \qquad && u, v \in V(h), \\ \widetilde{a}_h(u,u) &\geq C_{\mathrm{coer}} \|u\|_h^2, && u \in V(h). \end{aligned}$$

In particular, the coercivity bound implies the result in Lemma 3.1*.*

Proof. By taking into account the bounds in (2.5) and Lemma 4.3, application of the Cauchy–Schwarz inequality readily gives in the general case

$$|\widetilde{a}_h(u,v)| \leq \max\{2, \alpha^{-1} c_\star^{-1} c^\star\, C_{\mathrm{inv}}^2 + 1\} \|u\|_h \|v\|_h.$$

For $\alpha \geq \alpha_{\min}$, the continuity of $\widetilde{a}_h$ immediately follows. The case of piecewise constant c follows analogously.

To show the coercivity of the form $\widetilde{a}_h$, we note that

$$\widetilde{a}_h(u,u) = \sum_{K \in \mathcal{T}_h} \|c^{\frac{1}{2}} \nabla u\|_{0,K}^2 - 2 \sum_{K \in \mathcal{T}_h} \int_K \mathcal{L}_c(u) \cdot \nabla u \, dx + \sum_{F \in \mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}} [\![u]\!]\|_{0,F}^2.$$

By using the weighted Cauchy–Schwarz inequality, the geometric-arithmetic inequality $ab \leq \frac{\varepsilon a^2}{2} + \frac{b^2}{2\varepsilon}$, valid for any $\varepsilon > 0$, the bounds in (2.5), and the stability bound for the lifting operator in Lemma 4.3, we obtain for general c

$$\begin{aligned} 2 \sum_{K \in \mathcal{T}_h} \int_K \mathcal{L}_c(u) \cdot \nabla u \, dx &= 2 \sum_{K \in \mathcal{T}_h} \int_K c^{-\frac{1}{2}} \mathcal{L}_c(u) \cdot c^{\frac{1}{2}} \nabla u \, dx \\ &\leq 2 \sum_{K \in \mathcal{T}_h} \|c^{-\frac{1}{2}} \mathcal{L}_c(u)\|_{0,K} \|c^{\frac{1}{2}} \nabla u\|_{0,K} \\ &\leq \varepsilon \sum_{K \in \mathcal{T}_h} \|c^{\frac{1}{2}} \nabla u\|_{0,K}^2 + \varepsilon^{-1} c_\star^{-1} \sum_{K \in \mathcal{T}_h} \|\mathcal{L}_c(u)\|_{0,K}^2 \\ &\leq \varepsilon \sum_{K \in \mathcal{T}_h} \|c^{\frac{1}{2}} \nabla u\|_{0,K}^2 + \varepsilon^{-1} \alpha^{-1} c_\star^{-1} c^\star\, C_{\mathrm{inv}}^2 \sum_{F \in \mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}} [\![u]\!]\|_{0,F}^2 \end{aligned}$$

for a parameter $\varepsilon > 0$ still at our disposal. We conclude that

$$\widetilde{a}_h(u,u) \geq (1-\varepsilon) \sum_{K \in \mathcal{T}_h} \|c^{\frac{1}{2}} \nabla u\|_{0,K}^2 + \big(1 - \varepsilon^{-1} \alpha^{-1} c_\star^{-1} c^\star\, C_{\mathrm{inv}}^2\big) \sum_{F \in \mathcal{F}_h} \|\mathtt{a}^{\frac{1}{2}} [\![u]\!]\|_{0,F}^2.$$

For $\varepsilon = \frac{1}{2}$ and $\alpha \geq \alpha_{\min}$, we obtain the desired coercivity bound.

For a piecewise constant c we use the bound for $\|c^{-\frac{1}{2}} \mathcal{L}_c(u)\|_0^2$ from Lemma 4.4 and proceed analogously. ∎

Error equation. Because $\widetilde{a}_h$ coincides with a_h on $V^h \times V^h$, the semidiscrete scheme in (3.3)–(3.5) is equivalent to the following:

Find $u^h : \overline{J} \times V^h \to \mathbb{R}$ such that $u^h|_{t=0} = \Pi_h u_0$, $u_t^h|_{t=0} = \Pi_h v_0$, and

$$(u_{tt}^h, v) + \widetilde{a}_h(u^h, v) = (f, v) \qquad \forall v \in V^h. \tag{4.6}$$

We shall use the formulation in (4.6) as the basis of our error analysis.

To derive an error equation, we first define for $u \in H^{1+\sigma}(\Omega)$ with $\sigma > 1/2$

$$r_h(u; v) = \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{ c \nabla u - c \Pi_h(\nabla u) \}\!\} \, dA, \qquad v \in V(h). \tag{4.7}$$

Here Π_h denotes the L^2-projection onto $(V^h)^d$. The assumption $u \in H^{1+\sigma}(\Omega)$ ensures that $r_h(u; v)$ is well defined. From the definition in (4.7) it is immediate that $r_h(u; v) = 0$ when $v \in H_0^1(\Omega)$.

LEMMA 4.5. *Let the analytical solution u of* (2.1)–(2.4) *satisfy*

$$u \in L^\infty(J; H^{1+\sigma}(\Omega)), \qquad u_{tt} \in L^1(J; L^2(\Omega)).$$

Let u^h be the semidiscrete DG approximation obtained by (4.6). *Then, the error $e = u - u^h$ satisfies*

$$(e_{tt}, v) + \widetilde{a}_h(e, v) = r_h(u; v) \qquad \forall v \in V^h \text{ a.e. in } J,$$

with $r_h(u; v)$ given in (4.7).

Proof. Let $v \in V^h$. Since $u_{tt} \in L^1(J; L^2(\Omega))$, we have $\langle u_{tt}, v \rangle = (u_{tt}, v)$ almost everywhere in J. Hence, using the discrete formulation in (3.3)–(3.5), we obtain that

$$(e_{tt}, v) + \widetilde{a}_h(e, v) = (u_{tt}, v) + \widetilde{a}_h(u, v) - (f, v) \qquad \text{a.e. in } J.$$

Now, by definition of $\widetilde{a}_h$, the fact that $\mathcal{L}_c(u) = 0$ and that $[\![u]\!] = 0$ on all faces, the defining properties of the L^2-projection Π_h, and the definition of the lifted element $\mathcal{L}_c(v)$, we obtain

$$\widetilde{a}_h(u, v) = \sum_{K \in \mathcal{T}_h} \int_K c \nabla u \cdot \nabla v \, dx - \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{ c \Pi_h(\nabla u) \}\!\} \, dA.$$

Since $u_{tt} \in L^1(J; L^2(\Omega))$ and $f \in L^2(J; L^2(\Omega))$, we have that $\nabla \cdot (c \nabla u) \in L^2(\Omega)$ almost everywhere in J, which implies that $c \nabla u$ has continuous normal components across all interior faces. Therefore, elementwise integration by parts combined with the trace operators defined in section 3.1 yields

$$\begin{aligned} \widetilde{a}_h(u, v) = &- \sum_{K \in \mathcal{T}_h} \int_K \nabla \cdot (c \nabla u) \, v \, dx + \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{ c \nabla u \}\!\} \, dA \\ &- \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{ c \Pi_h(\nabla u) \}\!\} \, dA. \end{aligned}$$

From the definition of $r_h(u, v)$ in (4.7), we therefore conclude that

$$(u_{tt}, v) + \widetilde{a}_h(u, v) = (u_{tt} - \nabla \cdot (c \nabla u), v) + r_h(u; v)$$

and obtain

$$(e_{tt}, v) + \widetilde{a}_h(e, v) = (u_{tt} - \nabla \cdot (c \nabla u) - f, v) + r_h(u; v) = r_h(u; v),$$

where we have used the differential equation (2.1). ∎

Approximation properties. Let Π_h and Π_h denote the L^2-projections onto V^h and $(V^h)^d$, respectively. We recall the following approximation properties; see [6].

LEMMA 4.6. *Let $K \in \mathcal{T}_h$. Then the following hold:*

(i) *For $v \in H^t(K)$, $t \geq 0$, we have*

$$\|v - \Pi_h v\|_{0,K} \leq C h_K^{\min\{t,\ell+1\}} \|v\|_{t,K},$$

with a constant C that is independent of the local mesh size h_K and depends only on the shape-regularity of the mesh, the approximation order ℓ, the dimension d, and the regularity exponent t.

(ii) *For $v \in H^{1+\sigma}(K)$, $\sigma > \frac{1}{2}$, we have*

$$\begin{aligned} \|\nabla v - \nabla(\Pi_h v)\|_{0,K} &\leq C h_K^{\min\{\sigma,\ell\}} \|v\|_{1+\sigma,K}, \\ \|v - \Pi_h v\|_{0,\partial K} &\leq C h_K^{\min\{\sigma,\ell\}+\frac{1}{2}} \|v\|_{1+\sigma,K}, \\ \|\nabla v - \Pi_h(\nabla v)\|_{0,\partial K} &\leq C h_K^{\min\{\sigma,\ell+1\}-\frac{1}{2}} \|v\|_{1+\sigma,K}, \end{aligned}$$

with a constant C that is independent of the local mesh size h_K and depends only on the shape-regularity of the mesh, the approximation order ℓ, the dimension d, and the regularity exponent σ.

As a consequence of the approximation properties in Lemma 4.6, we obtain the following results.

LEMMA 4.7. *Let $u \in H^{1+\sigma}(\Omega)$, $\sigma > \frac{1}{2}$. Then the following hold:*

(i) *We have*

$$\|u - \Pi_h u\|_h \leq C_A h^{\min\{\sigma,\ell\}} \|u\|_{1+\sigma},$$

with a constant C_A that is independent of the mesh size and depends only on α, the constant κ in (3.1), *the bounds in* (2.5), *and the constants in Lemma* 4.6.

(ii) *For $v \in V(h)$, the form $r_h(u;v)$ in* (4.7) *can be bounded by*

$$|r_h(u;v)| \leq C_R h^{\min\{\sigma,\ell\}} \left(\sum_{F \in \mathcal{F}_h} \|\mathsf{a}^{\frac{1}{2}} [\![v]\!]\|_{0,F}^2 \right)^{\frac{1}{2}} \|u\|_{1+\sigma},$$

with a constant C_R independent of h, which depends only on α, the bounds in (2.5), *and the constants in Lemma* 4.6.

Proof. The estimate in (i) is an immediate consequence of Lemma 4.6, the definition of a, and the bounded variation property (3.1). To show the bound in (ii), we apply the Cauchy–Schwarz inequality and obtain

$$\begin{aligned} |r_h(u;v)| &\leq \left(\sum_{F \in \mathcal{F}_h} \int_F \mathsf{a} [\![v]\!]^2 \, ds \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_h} \int_F \mathsf{a}^{-1} |\{\!\{ c \nabla u - c \Pi_h(\nabla u) \}\!\}|^2 \, ds \right)^{\frac{1}{2}} \\ &\leq \alpha^{-\frac{1}{2}} c_\star^{-\frac{1}{2}} c^\star \left(\sum_{F \in \mathcal{F}_h} \|\mathsf{a}^{\frac{1}{2}} [\![v]\!]\|_{0,F}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} h_K \|\nabla u - \Pi_h(\nabla u)\|_{0,\partial K}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Applying the approximation properties in Lemma 4.6 completes the proof. ∎

4.3. Proof of Theorem 4.1. We are now ready to complete the proof of Theorem 4.1. We begin by proving the following auxiliary result.

LEMMA 4.8. *Let the analytical solution u of (2.1)–(2.4) satisfy*

$$u \in L^\infty(J; H^{1+\sigma}(\Omega)), \qquad u_t \in L^\infty(J; H^{1+\sigma}(\Omega))$$

for $\sigma > \frac{1}{2}$. Let $v \in C^0(\overline{J}; V(h))$ and $v_t \in L^1(J; V(h))$. Then we have

$$\int_J |r_h(u; v_t)|\, dt \le C_R\, h^{\min\{\sigma,\ell\}}\, \|v\|_{L^\infty(J;V(h))} \cdot \Big[2\,\|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + T\,\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}\Big],$$

where C_R is the constant from the bound (ii) *in Lemma* 4.7.

Proof. From the definition of r_h in (4.7) and integration by parts, we obtain

$$\begin{aligned}
\int_J r_h(u; v_t)\, dt &= \int_J \sum_{F\in\mathcal{F}_h} \int_F [\![v_t]\!] \cdot \{\!\{c\,\nabla u - c\,\Pi_h(\nabla u)\}\!\}\, dA\, dt \\
&= -\int_J \sum_{F\in\mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{c\,\nabla u_t - c\,\Pi_h(\nabla u_t)\}\!\}\, dA\, dt \\
&\quad + \Big[\sum_{F\in\mathcal{F}_h} \int_F [\![v]\!] \cdot \{\!\{c\,\nabla u - c\,\Pi_h(\nabla u)\}\!\}\, dA\Big]_{t=0}^{t=T} \\
&= -\int_J r_h(u_t; v)\, dt + \Big[r_h(u; v)\Big]_{t=0}^{t=T}.
\end{aligned}$$

Lemma 4.7 then implies the two estimates

$$\left|\int_J r_h(u_t; v)\, dt\right| \le C_R h^{\min\{\sigma,\ell\}}\, T\, \|v\|_{L^\infty(J;V(h))} \|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}$$

and

$$\left|\Big[r_h(u; v)\Big]_{t=0}^{t=T}\right| \le 2C_R h^{\min\{\sigma,\ell\}} \|v\|_{L^\infty(J;V(h))} \|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))},$$

which concludes the proof of the lemma. ∎

To complete the proof of Theorem 4.1, we now set $e = u - u^h$ and recall that Π_h is the L^2-projection onto V^h. Because of (2.8), we have

$$e \in C^0(\overline{J}; V(h)) \cap C^1(\overline{J}; L^2(\Omega)).$$

Next, we use the symmetry of $\widetilde{a}_h$ and the error equation in Lemma 4.5 to obtain

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\big[\|e_t\|_0^2 + \widetilde{a}_h(e,e)\big] &= (e_{tt}, e_t) + \widetilde{a}_h(e, e_t) \\
&= (e_{tt}, (u - \Pi_h u)_t) + \widetilde{a}_h(e, (u - \Pi_h u)_t) \\
&\quad + r_h(u; (\Pi_h u - u^h)_t).
\end{aligned} \tag{4.8}$$

We fix $s \in J$ and integrate (4.8) over the time interval $(0, s)$. This yields

$$\begin{aligned}
\frac{1}{2}\|e_t(s)\|_0^2 + \frac{1}{2}\widetilde{a}_h(e(s), e(s)) &= \frac{1}{2}\|e_t(0)\|_0^2 + \frac{1}{2}\widetilde{a}_h(e(0), e(0)) \\
&\quad + \int_0^s (e_{tt}, (u - \Pi_h u)_t)\, dt + \int_0^s \widetilde{a}_h(e, (u - \Pi_h u)_t)\, dt \\
&\quad + \int_0^s r_h(u; (\Pi_h u - u^h)_t)\, dt.
\end{aligned}$$

Integration by parts of the third term on the right-hand side yields

$$\int_0^s (e_{tt}, (u - \Pi_h u)_t)\, dt = -\int_0^s (e_t, (u - \Pi_h u)_{tt})\, dt + \Big[(e_t, (u - \Pi_h u)_t)\Big]_{t=0}^{t=s}.$$

From the stability properties of $\widetilde{a}_h$ in Lemma 4.4 and standard Hölder's inequalities, we conclude that

$$\begin{aligned}
\frac{1}{2}\|e_t(s)\|_0^2 + \frac{1}{2}C_{\rm coer}\|e(s)\|_h^2 \le\ & \frac{1}{2}\|e_t(0)\|_0^2 + \frac{1}{2}C_{\rm cont}\|e(0)\|_h^2 \\
&+ \|e_t\|_{L^\infty(J;L^2(\Omega))}\Big(\|(u - \Pi_h u)_{tt}\|_{L^1(J;L^2(\Omega))} + 2\|(u - \Pi_h u)_t\|_{L^\infty(J;L^2(\Omega))}\Big) \\
&+ C_{\rm cont}T\|e\|_{L^\infty(J;V(h))}\, \|(u - \Pi_h u)_t\|_{L^\infty(J;V(h))} \\
&+ \left| \int_J r_h(u; (\Pi_h u - u^h)_t)\, dt \right|.
\end{aligned}$$

Since this inequality holds for any $s \in J$, it also holds for the maximum over J, that is

$$\|e_t\|_{L^\infty(J;L^2(\Omega))}^2 + C_{\rm coer}\|e\|_{L^\infty(J;V(h))}^2 \le \|e_t(0)\|_0^2 + C_{\rm cont}\|e(0)\|_h^2 + T_1 + T_2 + T_3,$$

with

$$\begin{aligned}
T_1 &= 2\|e_t\|_{L^\infty(J;L^2(\Omega))}\Big(\|(u - \Pi_h u)_{tt}\|_{L^1(J;L^2(\Omega))} + 2\|(u - \Pi_h u)_t\|_{L^\infty(J;L^2(\Omega))}\Big), \\
T_2 &= 2C_{\rm cont}T\|e\|_{L^\infty(J;V(h))}\, \|(u - \Pi_h u)_t\|_{L^\infty(J;V(h))}, \\
T_3 &= 2\left| \int_J r_h(u; (\Pi_h u - u^h)_t)\, dt \right|.
\end{aligned}$$

Using the geometric-arithmetic mean inequality $|ab| \le \frac{1}{2\varepsilon}a^2 + \frac{\varepsilon}{2}b^2$, valid for any $\varepsilon > 0$, and the approximation results in Lemma 4.6, we conclude that

$$\begin{aligned}
T_1 &\le \frac{1}{2}\|e_t\|_{L^\infty(J;L^2(\Omega))}^2 + 2\Big(\|(u - \Pi_h u)_{tt}\|_{L^1(J;L^2(\Omega))} + 2\|(u - \Pi_h u)_t\|_{L^\infty(J;L^2(\Omega))}\Big)^2 \\
&\le \frac{1}{2}\|e_t\|_{L^\infty(J;L^2(\Omega))}^2 + 4\|(u - \Pi_h u)_{tt}\|_{L^1(J;L^2(\Omega))}^2 + 16\|(u - \Pi_h u)_t\|_{L^\infty(J;L^2(\Omega))}^2, \\
&\le \frac{1}{2}\|e_t\|_{L^\infty(J;L^2(\Omega))}^2 + Ch^{2\min\{\sigma,\ell\}}\left(\|u_{tt}\|_{L^1(J;H^\sigma(\Omega))}^2 + h^2\, \|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}^2 \right),
\end{aligned}$$

with a constant C that depends only on the constants in Lemma 4.6. Similarly,

$$\begin{aligned}
T_2 &\le \frac{1}{4}C_{\rm coer}\|e\|_{L^\infty(J;V(h))}^2 + 4\frac{C_{\rm cont}^2}{C_{\rm coer}}T^2\|(u - \Pi_h u)_t\|_{L^\infty(J;V(h))}^2 \\
&\le \frac{1}{4}C_{\rm coer}\|e\|_{L^\infty(J;V(h))}^2 + T^2Ch^{2\min\{\sigma,\ell\}}\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}^2,
\end{aligned}$$

where the constant C depends on $C_{\rm coer}$, $C_{\rm cont}$, and the constant C_A in Lemma 4.7.

It remains to bound the term T_3. To do so, we use Lemma 4.8 to obtain

$$T_3 \le 2C_R\mathcal{R}\, h^{\min\{\sigma,\ell\}}\|\Pi_h u - u^h\|_{L^\infty(J;V(h))},$$

with

$$\mathcal{R} := \Big[2\|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + T\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}\Big].$$

The triangle inequality, the geometric-arithmetic mean, and the approximation properties of Π_h in Lemma 4.7 then yield

$$\begin{aligned} T_3 &\le 2C_R\mathcal{R}\, h^{\min\{\sigma,\ell\}} \left[\|e\|_{L^\infty(J;V(h))} + \|u-\Pi_h u\|_{L^\infty(J;V(h))}\right] \\ &\le \frac{1}{4}C_{\mathrm{coer}}\|e\|^2_{L^\infty(J;V(h))} + Ch^{2\min\{\sigma,\ell\}}\left[\|u\|^2_{L^\infty(J;H^{1+\sigma}(\Omega))} + \mathcal{R}^2\right], \end{aligned}$$

with a constant C that depends only on C_{coer}, C_R, and C_A. Combining the above estimates for T_1, T_2, and T_3 then shows that

$$\begin{aligned} &\frac{1}{2}\|e_t\|^2_{L^\infty(J;L^2(\Omega))} + \frac{1}{2}C_{\mathrm{coer}}\|e\|^2_{L^\infty(J;V(h))} \le \|e_t(0)\|_0^2 + C_{\mathrm{cont}}\|e(0)\|_h^2 \\ &\qquad + Ch^{2\min\{\sigma,\ell\}}\left[\|u_{tt}\|^2_{L^1(J;H^\sigma(\Omega))} + T^2\|u_t\|^2_{L^\infty(J;H^{1+\sigma}(\Omega))} + \|u\|^2_{L^\infty(J;H^{1+\sigma}(\Omega))}\right], \end{aligned}$$

with a constant that is independent of T and the mesh size. This concludes the proof of Theorem 4.1.

4.4. Proof of Theorem 4.2. To prove the error estimate in Theorem 4.2, we first establish the following variant of [3, Lemma 2.1].

LEMMA 4.9. *For $u \in H^{1+\sigma}(\Omega)$ with $\sigma > \frac{1}{2}$, let $w^h \in V^h$ be the solution of*

$$\widetilde{a}_h(w^h, v) = \widetilde{a}_h(u, v) - r_h(u; v) \qquad \forall v \in V^h.$$

Then, we have

$$\|u - w^h\|_h \le C_E\, h^{\min\{\sigma,\ell\}}\|u\|_{1+\sigma},$$

with a constant C_E that is independent of h and depends only on C_{coer}, C_{cont} in Lemma 4.4 and C_A, C_R in Lemma 4.7.

Moreover, if the elliptic regularity defined in (4.1) and (4.2) holds, we have the L^2-bound

$$\|u - w^h\|_0 \le C_L\, h^{\min\{\sigma,\ell\}+1}\|u\|_{1+\sigma},$$

with a constant C_L that is independent of h and depends only on the stability constant C_S in (4.2); C_{coer}, C_{cont} in Lemma 4.4; and C_A, C_R in Lemma 4.7.

Proof. We first remark that the approximation w^h is well defined because of the stability properties in Lemma 4.4 and the estimates in Lemma 4.7. To prove the estimate for $\|u - w^h\|_h$, we first use the triangle inequality,

$$\|u - w^h\|_h \le \|u - \Pi_h u\|_h + \|\Pi_h u - w^h\|_h. \tag{4.9}$$

From the approximation properties of Π_h in Lemma 4.7, we immediately infer that

$$\|u - \Pi_h u\|_h \le C_A h^{\min\{\sigma,\ell\}}\|u\|_{1+\sigma}.$$

It remains to bound $\|\Pi_h u - w^h\|_h$. From the coercivity and continuity of $\widetilde{a}_h$ in Lemma 4.4, the definition of w^h, and the bound in Lemma 4.7, we conclude that

$$\begin{aligned} C_{\mathrm{coer}}\|\Pi_h u - w^h\|_h^2 &\le \widetilde{a}_h(\Pi_h u - w^h, \Pi_h u - w^h) \\ &= \widetilde{a}_h(\Pi_h u - u, \Pi_h u - w^h) + \widetilde{a}_h(u - w^h, \Pi_h u - w^h) \\ &= \widetilde{a}_h(\Pi_h u - u, \Pi_h u - w^h) + r_h(u; \Pi_h u - w^h) \\ &\le C_{\mathrm{cont}}\|\Pi_h u - u\|_h\|\Pi_h u - w^h\|_h + C_R h^{\min\{\sigma,\ell\}}\|u\|_{1+\sigma}\|\Pi_h u - w^h\|_h. \end{aligned}$$

Thus,

$$\|\Pi_h u - w^h\|_h \le \left(\frac{C_{\rm cont} C_A + C_R}{C_{\rm coer}}\right) h^{\min\{\sigma,\ell\}} \|u\|_{1+\sigma},$$

which proves the bound for $\|u - w^h\|_h$.

We shall now prove the L^2-bound. To do so, let $z \in H_0^1(\Omega)$ be the solution of

$$-\nabla \cdot (c\,\nabla z) = u - w^h \quad \text{in } \Omega, \qquad z = 0 \quad \text{on } \Gamma. \tag{4.10}$$

Then, the elliptic regularity assumption in (4.1) and (4.2) implies that

$$z \in H^2(\Omega), \qquad \|z\|_2 \le C_S \|u - w^h\|_0. \tag{4.11}$$

Next, we multiply (4.10) by $u - w^h$ and integrate the resulting expression by parts. Since $c\,\nabla z$ has continuous normal components across all interior faces, we have

$$\begin{aligned}\|u - w^h\|_0^2 &= \sum_{K\in\mathcal{T}_h} \left[\int_K c\nabla z \cdot \nabla(u - w^h)\,dx - \int_{\partial K} c\nabla z \cdot \mathbf{n}_K (u - w^h)\,dA\right] \\ &= \sum_{K\in\mathcal{T}_h} \int_K c\nabla z \cdot \nabla(u - w^h)\,dx - \sum_{F\in\mathcal{F}_h} \int_F \{\!\{c\nabla z\}\!\} \cdot [\![u - w^h]\!]\,dA,\end{aligned}$$

with $\mathbf{n}_K$ denoting the unit outward normal on ∂K. By definition of $\widetilde{a}_h$ and r_h, we immediately find that

$$\|u - w^h\|_0^2 = \widetilde{a}_h(z, u - w^h) - r_h(z; u - w^h).$$

From the symmetry of $\widetilde{a}_h$, the definition of w^h, and the fact that $[\![z]\!] = 0$ on all faces, we conclude that

$$\begin{aligned}\|u - w^h\|_0^2 &= \widetilde{a}_h(u - w^h, z - \Pi_h z) - r_h(u; z - \Pi_h z) - r_h(z; u - w^h) \\ &=: T_1 + T_2 + T_3.\end{aligned} \tag{4.12}$$

We shall now derive upper bounds for each individual term T_1, T_2, and T_3 in (4.12).

To estimate the term T_1, we use the continuity of $\widetilde{a}_h$, the approximation result in Lemma 4.7 with $\sigma = 1$, and the bound in (4.11). Thus,

$$\begin{aligned}T_1 &\le C_{\rm cont} \|u - w^h\|_h \|z - \Pi_h z\|_h \\ &\le C_{\rm cont} C_A h \|u - w^h\|_h \|z\|_2 \\ &\le C_{\rm cont} C_A C_S h \|u - w^h\|_h \|u - w^h\|_0.\end{aligned}$$

By using Lemma 4.7 and the stability bound in (4.11), we can estimate T_2 by

$$\begin{aligned}T_2 &\le C_R h^{\min\{\sigma,\ell\}} \|z - \Pi_h z\|_h \|u\|_{1+\sigma} \\ &\le C_R C_A h^{\min\{\sigma,\ell\}+1} \|z\|_2 \|u\|_{1+\sigma} \\ &\le C_R C_A C_S h^{\min\{\sigma,\ell\}+1} \|u - w^h\|_0 \|u\|_{1+\sigma}.\end{aligned}$$

Similarly,

$$T_3 \le C_R h \|z\|_2 \|u - w^h\|_h \le C_R C_S h \|u - w^h\|_0 \|u - w^h\|_h.$$

The use of these bounds for T_1, T_2, and T_3 in (4.12) then leads to

$$\|u - w^h\|_0 \le Ch\|u - w^h\|_h + Ch^{\min\{\sigma,\ell\}+1}\|u\|_{1+\sigma},$$

which completes the proof of the lemma, since $\|u - w^h\|_h \le Ch^{\min\{\sigma,\ell\}}\|u\|_{1+\sigma}$. ∎

Now, let u be defined by the exact solution of (2.1)–(2.4). We may define $w^h(t,\cdot) \in V^h$ almost everywhere in J by

$$\widetilde{a}_h(w^h(t,\cdot), v) = \widetilde{a}_h(u(t,\cdot), v) - r_h(u(t,\cdot); v) \qquad \forall v \in V^h. \tag{4.13}$$

If $u \in L^\infty(J; H^{1+\sigma}(\Omega))$, it can be readily seen that $w^h \in L^\infty(J; V(h))$. Moreover, if we also have $u_t \in L^\infty(J; H^{1+\sigma}(\Omega))$, then $w^h_t \in L^\infty(J; V(h))$ and

$$\widetilde{a}_h(w^h_t, v) = \widetilde{a}_h(u_t, v) - r_h(u_t; v), \qquad v \in V^h \text{ a.e. in } J,$$

as well as

$$\widetilde{a}_h(w^h(0), v) = \widetilde{a}_h(u_0, v) - r_h(u_0; v), \qquad v \in V^h.$$

Therefore Lemma 4.9 immediately implies the following estimates.

LEMMA 4.10. *Let w^h be defined by* (4.13). *Under the regularity assumptions of Theorem* 4.2, *we have*

$$\begin{aligned} \|(u - w^h)_t\|_{L^\infty(J;V(h))} &\le C_E\, h^{\min\{\sigma,\ell\}}\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}, \\ \|(u - w^h)(0)\|_h &\le C_E\, h^{\min\{\sigma,\ell\}}\|u_0\|_{1+\sigma}. \end{aligned}$$

Moreover, if elliptic regularity as defined in (4.1) *and* (4.2) *holds, we have the L^2-bounds*

$$\begin{aligned} \|(u - w^h)_t\|_{L^\infty(J;L^2(\Omega))} &\le C_L\, h^{\min\{\sigma,\ell\}+1}\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}, \\ \|(u - w^h)(0)\|_0 &\le C_L\, h^{\min\{\sigma,\ell\}+1}\|u_0\|_{1+\sigma}. \end{aligned}$$

The constants C_E and C_L are as in Lemma 4.9.

To complete the proof of Theorem 4.2, let $w^h \in L^\infty(J; V(h))$ be defined by (4.13) and consider

$$\|e\|^2_{L^\infty(J;L^2(\Omega))} \le 2\|u - w^h\|^2_{L^\infty(J;L^2(\Omega))} + 2\|w^h - u^h\|^2_{L^\infty(J;L^2(\Omega))}. \tag{4.14}$$

The first term can be estimated from the L^2-bounds in Lemma 4.9. We shall now derive an estimate for the second term. First, we fix $v \in L^\infty(J; V^h)$ and assume that $v_t \in L^\infty(J; V^h)$. From the definition of w^h in (4.13) and the error equation in Lemma 4.5, we have

$$\begin{aligned} ((u^h - w^h)_{tt}, v) + \widetilde{a}_h(u^h - w^h, v) &= (u^h_{tt}, v) + \widetilde{a}_h(u^h, v) - \widetilde{a}_h(w^h, v) - (w^h_{tt}, v) \\ &= (u^h_{tt}, v) + \widetilde{a}_h(u^h - u, v) + r_h(u; v) - (w^h_{tt}, v) \\ &= (u_{tt}, v) - (w^h_{tt}, v). \end{aligned}$$

We rewrite this identity as

$$\frac{d}{dt}((u^h - w^h)_t, v) - ((u^h - w^h)_t, v_t) + \widetilde{a}_h(u^h - w^h, v) = \frac{d}{dt}((u - w^h)_t, v) - ((u - w^h)_t, v_t),$$

which yields

$$-((u^h - w^h)_t, v_t) + \widetilde{a}_h(u^h - w^h, v) = \frac{d}{dt}((u - u^h)_t, v) - ((u - w^h)_t, v_t). \tag{4.15}$$

Let $\tau \in (0, T]$ be fixed, and consider the function

$$\widehat{v}(t, \cdot) = \int_t^\tau (u^h - w^h)(s, \cdot)\, ds, \qquad t \in \overline{J}.$$

Note that

$$\widehat{v}(\tau, \cdot) = 0, \qquad \widehat{v}_t(t, \cdot) = -(u^h - w^h)(t, \cdot) \qquad \text{a.e. } t \in \overline{J}.$$

Next, choose $v = \widehat{v}$ in (4.15) which yields

$$((u^h - w^h)_t, u^h - w^h) - \widetilde{a}_h(\widehat{v}_t, \widehat{v}) = \frac{d}{dt}((u - u^h)_t, \widehat{v}) + ((u - w^h)_t, u^h - w^h).$$

Since the DG form $\widetilde{a}_h(\cdot, \cdot)$ is symmetric, we obtain

$$\frac{1}{2}\frac{d}{dt}\|u^h - w^h\|_0^2 - \frac{1}{2}\frac{d}{dt}\widetilde{a}_h(\widehat{v}, \widehat{v}) = \frac{d}{dt}((u - u^h)_t, \widehat{v}) + ((u - w^h)_t, u^h - w^h).$$

Integration over $(0, \tau)$ and using that $\widehat{v}(\tau, \cdot) = 0$ then yield

$$\begin{aligned} &\|(u^h - w^h)(\tau)\|_0^2 - \|(u^h - w^h)(0)\|_0^2 + \widetilde{a}_h(\widehat{v}(0), \widehat{v}(0)) \\ &\qquad = -2((u - u^h)_t(0), \widehat{v}(0)) + 2\int_0^\tau ((u - w^h)_t, u^h - w^h)\, dt. \end{aligned} \tag{4.16}$$

Since $u_t(0) = v_0$, $u_t^h(0) = \Pi_h v_0$, and $\widehat{v}(0)$ belongs to V^h, we conclude that

$$((u - u^h)_t(0), \widehat{v}(0)) = (v_0 - \Pi_h v_0, \widehat{v}(0)) = 0.$$

Hence, the first term on the right-hand side of (4.16) vanishes. Moreover, the coercivity of the form $\widetilde{a}_h$ in Lemma 4.4 ensures that $\widetilde{a}_h(\widehat{v}(0), \widehat{v}(0)) \geq 0$. This leads to the inequality

$$\|(u^h - w^h)(\tau)\|_0^2 \leq \|(u^h - w^h)(0)\|_0^2 + 2\int_0^\tau \|(u - w^h)_t\|_0 \|u^h - w^h\|_0\, dt. \tag{4.17}$$

By using the Cauchy–Schwarz inequality and the geometric-arithmetic mean inequality, we obtain

$$\begin{aligned} 2\int_0^\tau \|(u - w^h)_t\|_0 \|u^h - w^h\|_0\, dt &\leq 2T\, \|(u - w^h)_t\|_{L^\infty(J;L^2(\Omega))} \|u^h - w^h\|_{L^\infty(J;L^2(\Omega))} \\ &\leq \frac{1}{2}\|u^h - w^h\|_{L^\infty(J;L^2(\Omega))}^2 + 2T^2\, \|(u - w^h)_t\|_{L^\infty(J;L^2(\Omega))}^2. \end{aligned}$$

Because this upper bound is independent of τ, it also holds for the supremum over $\tau \in J$, which yields the estimate

$$\begin{aligned} \frac{1}{2}\|u^h - w^h\|_{L^\infty(J;L^2(\Omega))}^2 &\leq \|(u^h - w^h)(0)\|_0^2 + 2T^2 \|(u - w^h)_t\|_{L^\infty(J;L^2(\Omega))}^2 \\ &\leq 2\,\|(u^h - u)(0)\|_0^2 + 2\|(u - w^h)(0)\|_0^2 + 2T^2 \|(u - w^h)_t\|_{L^\infty(J;L^2(\Omega))}^2. \end{aligned}$$

Next, we use this estimate in (4.14) to obtain

$$\|e\|^2_{L^\infty(J;L^2(\Omega))} \le 2\|u - w^h\|^2_{L^\infty(J;L^2(\Omega))} + 8\,\|u_0 - \Pi_h u_0\|_0^2 + 8\|u_0 - w^h(0)\|_0^2 + 8T^2\|(u-w^h)_t\|^2_{L^\infty(J;L^2(\Omega))}.$$

From the L^2-approximation properties in Lemma 4.6, Lemma 4.9, and Lemma 4.10, we finally conclude that

$$\|e\|^2_{L^\infty(J;L^2(\Omega))} \le h^{2\min\{\sigma,\ell\}+2}\Big[\max\{8C, 8C_L^2\}\|u_0\|^2_{1+\sigma} + 2C_L^2\|u\|^2_{L^\infty(J;H^{1+\sigma}(\Omega))} + 8C_L T^2\|u_t\|^2_{L^\infty(J;H^{1+\sigma}(\Omega))}\Big].$$

Here, C is the constant from Lemma 4.6. This completes the proof of Theorem 4.2.

5. Numerical results. We shall now present a series of numerical experiments which verify the sharpness of the theoretical error bounds stated in Theorems 4.1 and 4.2. Furthermore, we shall demonstrate the robustness and flexibility of our DG method by propagating a pulse through an inhomogeneous medium with discontinuity on a finite element mesh with hanging nodes.

To obtain a fully discrete discretization of the wave equation, we choose to augment our DG spatial discretization with the second-order Newmark scheme in time; see, e.g., [19, sections 8.5–8.7]. The resulting scheme has been implemented using the general purpose finite element library `deal.II`,[1] which provides powerful C^{++} classes to handle the finite element mesh and the degrees of freedom, and to solve the linear system of equations; see [5, 4]. In all our examples, the DG stabilization parameter is set to $\alpha = 20$.

5.1. Time discretization. The discretization of (2.1)–(2.4) in space by the DG method (3.3)–(3.5) leads to the linear second-order system of ordinary differential equations

$$\mathbf{M}\ddot{\mathbf{u}}^h(t) + \mathbf{A}\mathbf{u}^h(t) = \mathbf{f}^h(t), \qquad t \in J, \tag{5.1}$$

with initial conditions

$$\mathbf{M}\mathbf{u}^h(0) = \mathbf{u}_0^h, \quad \mathbf{M}\dot{\mathbf{u}}^h(0) = \mathbf{v}_0^h. \tag{5.2}$$

Here, $\mathbf{M}$ denotes the mass matrix and $\mathbf{A}$ the stiffness matrix. To discretize (5.1) in time, we employ the Newmark time-stepping scheme; see, e.g., [19]. We let k denote the time step and set $t_n = n \cdot k$. Then the Newmark method consists in finding approximations $\{\mathbf{u}_n^h\}_n$ to $\mathbf{u}^h(t_n)$ such that

$$(\mathbf{M} + k^2\beta\mathbf{A})\mathbf{u}_1^h = \left[\mathbf{M} - k^2\left(\frac{1}{2} - \beta\right)\mathbf{A}\right]\mathbf{u}_0^h + k\mathbf{M}\mathbf{v}_0^h + k^2\left[\beta\mathbf{f}_1^h + \left(\frac{1}{2} - \beta\right)\mathbf{f}_0^h\right] \tag{5.3}$$

and

$$\begin{aligned}(\mathbf{M} + k^2\beta\mathbf{A})\mathbf{u}_{n+1}^h = {} & \left[2\mathbf{M} - k^2\left(\frac{1}{2} - 2\beta + \gamma\right)\mathbf{A}\right]\mathbf{u}_n^h - \left[\mathbf{M} + k^2\left(\frac{1}{2} + \beta - \gamma\right)\mathbf{A}\right]\mathbf{u}_{n-1}^h \\ & + k^2\left[\beta\mathbf{f}_{n+1}^h + \left(\frac{1}{2} - 2\beta + \gamma\right)\mathbf{f}_n + \left(\frac{1}{2} + \beta - \gamma\right)\mathbf{f}_{n-1}\right]\end{aligned} \tag{5.4}$$

[1] See www.dealii.org.

for $n = 1, \ldots, N-1$. Here, $\mathbf{f}_n := \mathbf{f}(t_n)$, while $\beta \geq 0$ and $\gamma \geq 1/2$ are free parameters that still can be chosen. We recall that for $\gamma = 1/2$ the Newmark scheme is second-order accurate in time, whereas it is only first-order accurate for $\gamma > 1/2$. For $\beta = 0$, the Newmark scheme (5.3)–(5.4) requires at each time step the solution of a linear system with the mass matrix $\mathbf{M}$. However, because individual elements decouple, $\mathbf{M}$ is block-diagonal with a block size equal to the number of degrees of freedom per element. It can be inverted at very low computational cost, and the scheme is essentially fully explicit. In fact, if the basis functions are chosen mutually orthogonal, $\mathbf{M}$ reduces to the identity; see [8] and the references therein. Then, with $\gamma = 1/2$, the explicit Newmark method corresponds to the standard leap-frog scheme.

For $\beta > 0$, the resulting scheme is implicit and involves the solution of a linear system with the symmetric positive definite stiffness matrix $\mathbf{A}$ at each time step. We finally note that the second-order Newmark scheme with $\gamma = 1/2$ is unconditionally stable for $\beta \geq 1/4$, whereas for $1/4 > \beta \geq 0$ the time step k has to be restricted by a CFL condition. In the case $\beta = 0$, that condition is $k^2 \lambda_{\max}(\mathbf{A}) \leq 4(1-\varepsilon)$, $\varepsilon \in (0,1)$, where $\lambda_{\max}(\mathbf{A})$ is the maximal eigenvalue of the DG stiffness matrix $\mathbf{A}$ (which is of the order $\mathcal{O}(h^{-2})$ and also depends on α).

In all our tests, we will employ the explicit second-order Newmark scheme, setting $\gamma = 1/2$ and $\beta = 0$ in (5.3)–(5.4).

5.2. Example 1: Smooth solution. First, we consider the two-dimensional wave equation (2.1)–(2.4) in $J \times \Omega = (0,1) \times (0,1)^2$, with $c \equiv 1$ and data f, u_0, and v_0 chosen such that the analytical solution is given by

$$u(x_1, x_2, t) = t^2 \sin(\pi x_1) \sin(\pi x_2). \tag{5.5}$$

This solution is arbitrarily smooth so that all our theoretical regularity assumptions are satisfied. We discretize this problem using the polynomial spaces $\mathcal{Q}^\ell(K)$, $\ell = 1, 2, 3$, on a sequence $\{\mathcal{T}_h\}_{i \geq 1}$ of square meshes of size $h_i = 2^{-i}$. With increasing polynomial degree ℓ and decreasing mesh size h_i, smaller time steps k_i are necessary to ensure stability. We found that the choice $k_i = h_i/20$ provides a stable time discretization on every mesh. Because our numerical scheme is second-order accurate in time, the time integration of (5.5) is exact so that the spatial error is the only error component in the discrete solution.

In Figure 5.1 we show the relative errors at time $T = 1$ in the energy norm and in the L^2-norm, as we decrease the mesh size h_i. The numerical results corroborate with the expected theoretical rates of $\mathcal{O}(h^\ell)$ for the energy norm and of $\mathcal{O}(h^{\ell+1})$ for the L^2-norm; see Theorems 4.1 and 4.2.

Next, we modify the data so that the analytical solution u is given by

$$u(x_1, x_2, t) = \sin(t^2) \sin(\pi x_1) \sin(\pi x_2). \tag{5.6}$$

Although u remains arbitrarily smooth, it is no longer integrated exactly in time by (5.3)–(5.4). Since the Newmark scheme is only second-order accurate, we repeat the above experiment only for the lowest order spatial discretization, $\ell = 1$. Again, we set $k_i = h_i/20$. In Figure 5.2, the relative errors for the fully discrete approximation of (5.6) show convergence rates of order h in the energy norm and order h^2 in the L^2-norm, thereby confirming the theoretical estimates of Theorems 4.1 and 4.2.

5.3. Example 2: Singular solution. Here, we consider the two-dimensional wave equation (2.1)–(2.4) on the L-shaped domain $\Omega = (-1,1)^2 \backslash [0,1)^2$. We set $c = 1$

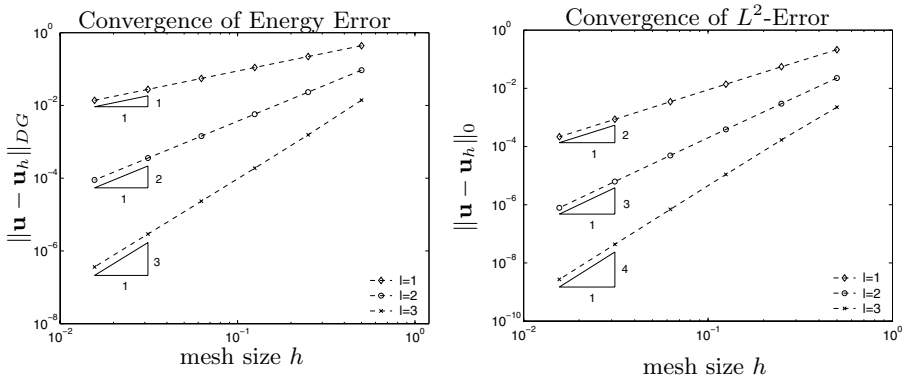


FIG. 5.1. *Example* 1: *Convergence of the error at time* $T = 1$ *in the energy norm and the* L^2*-norm for* $\ell = 1, 2, 3$.

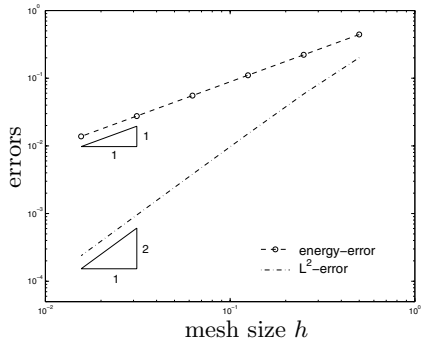


FIG. 5.2. *Example* 1: *Convergence of the error at time* $T = 1$ *in the energy norm and the* L^2*-norm for* $\ell = 1$.

everywhere and choose the data f, u_0, and v_0 such that the analytical solution u is given by

$$u(r, \phi, t) = t^2\, r^{2/3} \sin(2/3\, \phi) \tag{5.7}$$

in polar coordinates (r, ϕ). Although u is smooth in time (and can even be integrated exactly in time), it has a spatial singularity at the origin, such that $u \in C^\infty(\overline{J}; H^{5/3}(\Omega))$. Hence, this example is well suited to establishing the sharpness of the regularity assumptions in our theoretical results. Since u is inhomogeneous at the boundary of Ω, we need to impose inhomogeneous Dirichlet conditions within our DG discretization. We do so in straightforward fashion by modifying the semidiscrete formulation as follows: find $u^h(t, \cdot) : J \to V^h$ such that

$$(u^h_{tt}, v) + a_h(u^h, v) = (f, v) + \sum_{F \in \mathcal{F}_h^{\mathcal{B}}} \int_F g\, (\mathbf{a} v - c \nabla v \cdot \mathbf{n})\, dA. \tag{5.8}$$

Here, g is the boundary data and $\mathbf{n}$ is the outward unit normal vector on $\partial\Omega$.

We discretize (5.8) by using bilinear polynomials ($\ell = 1$) on the same sequence of meshes as before. Again, we set $k_i = h_i/20$ and integrate the problem up to $T = 1$. For the analytical solution u in (5.7), the regularity assumptions in Theorem 4.1 hold

TABLE 5.1
Example 2: *Relative errors at time* $T = 1$ *in the energy norm and* L^2*-norm, and corresponding numerical convergence rates.*

i	Cells	Energy-error		L^2-error	
1	12	1.11e-01	-	1.61e-02	-
2	48	7.18e-02	0.62	5.96e-03	1.43
3	192	4.61e-02	0.64	2.27e-03	1.40
4	768	2.94e-02	0.65	8.72e-04	1.38
5	3072	1.87e-02	0.66	3.38e-04	1.37
6	12288	1.18e-02	0.66	1.32e-04	1.36

with $\sigma = 2/3$. Thus, Theorem 4.1 predicts numerical convergence rates of $2/3$ in the energy norm, as confirmed by our numerical results in Table 5.1.

As the elliptic regularity assumptions (4.1)–(4.2) from Theorem 4.2 are violated, we do not expect L^2-error rates of the order $1+\sigma$ for this problem. Indeed, in Table 5.1 we observe convergence rates close to $4/3$. To explain this behavior, let us consider the following weaker elliptic regularity assumption: for any $\lambda \in L^2(\Omega)$ we assume that the solution of the problem

$$-\nabla \cdot (c\nabla z) = \lambda \quad \text{in } \Omega, \qquad z = 0 \quad \text{on } \partial\Omega, \tag{5.9}$$

belongs to $H^{1+s}(\Omega)$ for a parameter $s \in (1/2, 1]$ and satisfies the stability bound

$$\|z\|_{1+s} \le C_S \|\lambda\|_0 \tag{5.10}$$

for a stability constant C_S. The results from Lemmas 4.9 and 4.10 can be easily adapted to this case. As a consequence, the L^2-bound for $e = u - u^h$ from Theorem 4.2 can then be generalized to this weaker setting as

$$\|e\|_{L^\infty(J;L^2(\Omega))} \le Ch^{\min\{\sigma,\ell\}+s}\big[\|u_0\|_{1+\sigma} + \|u\|_{L^\infty(J;H^{1+\sigma}(\Omega))} + T\|u_t\|_{L^\infty(J;H^{1+\sigma}(\Omega))}\big].$$

For the L-shaped domain Ω and $c \equiv 1$, the (weaker) regularity assumption in (5.9)–(5.10) holds with $s = 2/3$, which underpins the rate $\sigma + s = 4/3$ observed in Table 5.1.

5.4. Example 3: Inhomogeneous medium. Finally, we consider (2.1)–(2.4) on the rectangular domain $\Omega = (-1, 2)\times(-1, 1)$, with homogeneous initial and boundary conditions and the piecewise constant material coefficient

$$c(x_1, x_2) = \begin{cases} 0.1, & x_1 \le 0, \\ 1, & \text{else.} \end{cases}$$

The wave is locally excited until $t = 0.2$ by the source term

$$f(x_1, x_2, t) = \begin{cases} 1, & 0.2 < x_1 < 0.4 \text{ and } t < 0.2, \\ 0, & \text{else.} \end{cases}$$

We discretize the problem by the DG method (3.3)–(3.5) on a fixed mesh $\mathcal{T}_h$ that consists of *nonmatching components*, which are adapted to the discontinuity c; see Figure 5.3. The mesh $\mathcal{T}_h$ is composed of 9312 nonuniform squares, where the smallest local mesh size is given by $h_{min} \approx 0.016$. The hanging nodes are naturally incorporated in the DG method without any difficulty. Here, the time step $k = 0.002$,

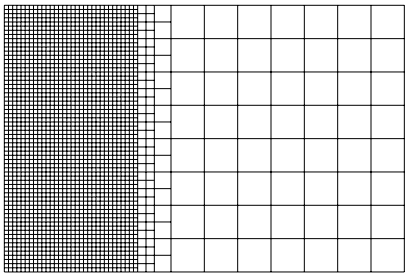


FIG. 5.3. *Example* 3: *Domain* Ω *with a finite element mesh* $\mathcal{T}_h$ *adapted to the values of the piecewise constant wave speed* $\sqrt{c} = 0.1$ *(left),* $\sqrt{c} = 1$ *(right).*

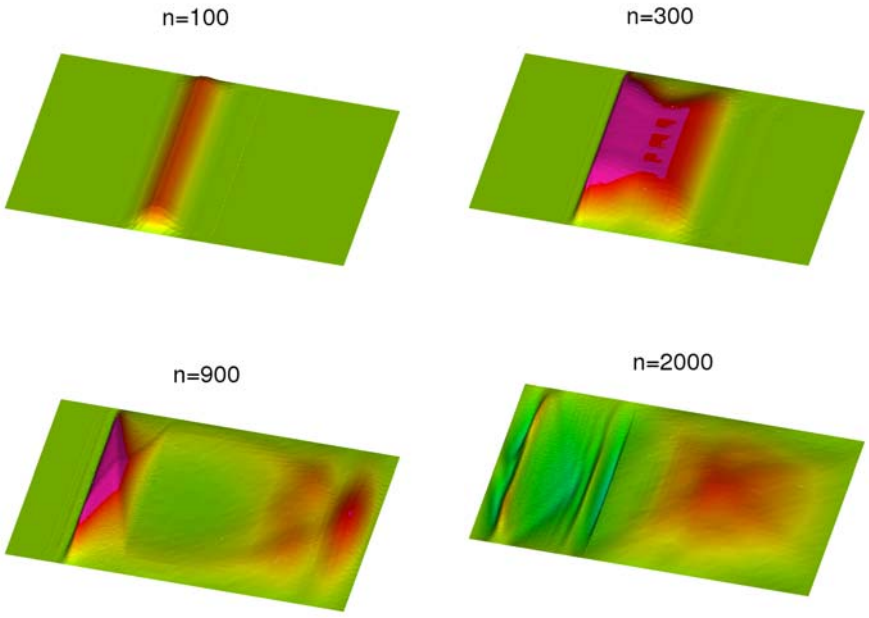


FIG. 5.4. *Example* 3: *The approximate DG solutions* u_n^h *shown at times* $t_n = 0.2, 0.6, 1.8, 4$ *display the behavior of a wave propagating through an inhomogeneous medium with homogeneous Dirichlet boundary.*

that is, $k \approx h_{\min}/8$, proved to be sufficiently small to ensure the stability of the explicit Newmark method ($\beta = 0$).

In Figure 5.4, the numerical solution is shown after $n = 100, 300, 900$, and 2000 time steps. The initial pulse splits into two planar wave fronts propagating in opposite directions to either side of the domain. After $n = 300$ time steps, the left-moving wave hits the much slower medium in the region $x_1 \leq 0$, resulting in a much steeper and narrower wave front. Meanwhile, the right-moving wave rapidly reaches the boundary at $x_1 = 2$, where it is reflected and eventually enters the slow medium, too. The discontinuous interface at $x_1 = 0$ generates multiple reflections, which interact with each other at later times.

6. Conclusion. We have presented and analyzed the symmetric interior penalty discontinuous Galerkin finite element method (DGFEM) for the numerical solution of the (second-order) scalar wave equation. Taking advantage of the symmetry of the method, we have carried out an a priori error analysis of the semidiscrete method and derived optimal error bounds in the energy norm and, under additional regularity assumptions, optimal error bounds in the L^2-norm. Our numerical results confirm the expected convergence rates and demonstrate the versatility of the method. The error analysis of the fully discrete scheme is the subject of ongoing work.

Based on discontinuous finite element spaces, the proposed DG method easily handles elements of various types and shapes, irregular nonmatching grids, and even locally varying polynomial order. As continuity is only weakly enforced across mesh interfaces, domain decomposition techniques immediately apply. Since the resulting mass matrix is essentially diagonal, the method is inherently parallel and leads to fully explicit time integration schemes. Moreover, as the stiffness matrix is symmetric positive definite, the DG method shares the following two important properties with the classical continuous Galerkin approach. First, the semidiscrete formulation conserves (a discrete version of) the energy for all time and therefore is nondissipative. Second, if implicit time integration is used to overcome CFL constraints, the resulting linear system to be solved at each time step will also be symmetric positive definite.

The symmetric interior penalty DGFEM, applied here to the scalar wave equation, can also be utilized for other second-order hyperbolic equations, such as in electromagnetics or elasticity. In fact, our error analysis for the semidiscrete (scalar) case readily extends to the second-order (vector) wave equation for time dependent elastic waves. The same DG approach also extends to Maxwell's equations in second-order form:

$$\varepsilon\,\mathbf{u}_{tt} + \sigma\mathbf{u}_t + \nabla \times (\mu^{-1}\,\nabla \times \mathbf{u}) = \mathbf{f}, \qquad \sigma \geq 0.$$

Here, DG discretizations with standard discontinuous piecewise polynomials indeed offer an alternative to edge elements, as typically used in conforming discretizations of Maxwell's equations. The corresponding error analysis, however, is more involved and will be reported elsewhere in the near future.

REFERENCES

[1] M. Ainsworth, P. Monk, and W. Muniz, *Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second order wave equation*, J. Sci. Comput., 27 (2006), pp. 5–40.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] G. A. Baker, *Error estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564–576.

[4] W. Bangerth, R. Hartmann, and G. Kanschat, `deal.II` *Differential Equations Analysis Library, Technical Reference*, IWR, Universität Heidelberg, Heidelberg, Germany, http://www.dealii.org.

[5] W. Bangerth and G. Kanschat, *Concepts for Object-Oriented Finite Element Software—The* `deal.II` *Library*, Report 99-43, Sonderforschungsbereich 3-59, IWR, Universität Heidelberg, Heidelberg, Germany, 1999.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[7] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, T. Barth and H. Deconink, eds., Springer-Verlag, Berlin, 1999, pp. 69–224.

[8] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 3–50.

[9] B. Cockburn and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin method for conservation laws.* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[10] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[11] G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal., 38 (2001), pp. 2047–2078.

[12] R. S. Falk and G. R. Richter, *Explicit finite element methods for symmetric hyperbolic equations*, SIAM J. Numer. Anal., 36 (1999), pp. 935–952.

[13] J. S. Hesthaven and T. Warburton, *Nodal high-order methods on unstructured grids.* I. *Time-domain solution of Maxwell's equations*, J. Comput. Phys., 181 (2002), pp. 186–221.

[14] P. Houston, M. Jensen, and E. Süli, *hp-discontinuous Galerkin finite element methods with least-squares stabilization*, J. Sci. Comput., 17 (2002), pp. 3–25.

[15] T. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[16] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, New York, 1972.

[17] P. Monk and G. R. Richter, *A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media*, J. Sci. Comput., 22–23 (2005), pp. 443–477.

[18] I. Perugia and D. Schötzau, *An hp-analysis of the local discontinuous Galerkin method for diffusion problems*, J. Sci. Comput., 17 (2002), pp. 561–571.

[19] P. A. Raviart and J.-M. Thomas, *Introduction à l'Analyse Numérique des Équations aux Dérivées Partielles*, Masson, Paris, 1983.

[20] B. Rivière and M. F. Wheeler, *Discontinuous Finite Element Methods for Acoustic and Elastic Wave Problems*, I: *Semidiscrete Error Estimates*, Technical report 01-02, TICAM, University of Texas, Austin, TX, 2001.

[21] B. Rivière and M. F. Wheeler, *Discontinuous finite element methods for acoustic and elastic wave problems*, in ICM2002-Beijing Satellite Conference on Scientific Computing, Contemp. Math. 329, AMS, Providence, RI, 2003, pp. 271–282.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2432–2455

BDDC ALGORITHMS FOR INCOMPRESSIBLE STOKES EQUATIONS*

JING LI† AND OLOF WIDLUND‡

Abstract. The purpose of this paper is to extend the balancing domain decomposition by constraints (BDDC) algorithm to saddle-point problems that arise when mixed finite element methods are used to approximate the system of incompressible Stokes equations. The BDDC algorithms are iterative substructuring methods which form a class of domain decomposition methods based on the decomposition of the domain of the differential equations into nonoverlapping subdomains. They are defined in terms of a set of primal continuity constraints which are enforced across the interface between the subdomains and which provide a coarse space component of the preconditioner. Sets of such constraints are identified for which bounds on the rate of convergence can be established that are just as strong as previously known bounds for the elliptic case. In fact, the preconditioned operator is effectively positive definite, which makes the use of a conjugate gradient method possible. A close connection is also established between the BDDC and dual-primal finite element tearing and interconnecting (FETI-DP) algorithms for the Stokes case.

Key words. domain decomposition, incompressible Stokes, preconditioners, mixed finite elements

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/050628556

1. Introduction.

The balancing domain decomposition by constraints (BDDC) algorithms are domain decomposition methods based on nonoverlapping subdomains into which the domain of a given partial differential equation has been divided. Introduced by Dohrmann [4] and analyzed in the elliptic case by Dohrmann, Mandel, and Tezaur [30, 31], these methods represent an important advance over the balancing Neumann–Neumann methods that have been used extensively in the past to solve large finite element problems; cf. [37, section 6.2] where references to earlier work can also be found. Just as the classical balancing methods have much in common with the original one-level finite element tearing and interconnecting (FETI) methods, BDDC is closely related to the more recent dual-primal FETI (FETI-DP) methods. Each BDDC and FETI-DP method is defined in terms of a set of *primal* continuity constraints across the interface Γ formed by the parts of the subdomain boundaries which are common to at least two subdomains. In addition to, or instead of, point constraints, it is important to make certain averages over edges or faces of the interface the same. In some applications, we also should have certain first order moments, over edges, with common values; see [23, 18] for a discussion of such fully primal edges for three-dimensional elasticity.

In an important contribution to the theory, Mandel, Dohrmann, and Tezaur [31] established that the preconditioned operators of a pair of BDDC and FETI-DP algo-

*Received by the editors April 5, 2005; accepted for publication (in revised form) June 13, 2006; published electronically December 1, 2006.

http://www.siam.org/journals/sinum/44-6/62855.html

†Department of Mathematical Sciences, Kent State University, Kent, OH 44242 (li@math.kent.edu, http://www.math.kent.edu/∼li/). This author's work was supported in part by National Science Foundation contract DMS-0612574.

‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (widlund@cs.nyu.edu, http://www.cs.nyu.edu/cs/faculty/widlund/). This author's work was supported in part by U.S. Department of Energy contracts DE-FG02-06ER25718 and DE-FC02-01ER25482 and in part by National Science Foundation contract DMS-0513251.

rithms, with the same primal constraints, have the same nonzero eigenvalues, except possibly for an eigenvalue equal to 1. We note that this fact was first observed experimentally by Fragakis and Papadrakakis [11] for pairs of balancing Neumann–Neumann and one-level FETI methods; these authors also discussed primal iterative substructuring methods which are close counterparts to various FETI algorithms. An important consequence of the results in [11, 31] is that these algorithms, which can be built from the same set of subprograms, have very similar performance. The choice of algorithm can therefore be based on other considerations. We believe that it is easier to introduce inexact subdomain and coarse level problem solvers in the BDDC algorithms than in FETI-DP algorithms; cf. [7, 16, 28, 38, 41].

In a recent paper [29], the authors rederived the BDDC and FETI-DP algorithms for elliptic problems and also gave a short proof of the main result in [31]. A key to these simplifications is a change of variables so that, e.g., a primal constraint on the average over an interface edge or face is represented by a single primal variable in the new coordinate system. Simultaneously, a complementary set of dual displacement variables is introduced, for each of which the edge or face averages vanish; an illustrative example of how the change of variables can be carried out is given in [29]. This leads to a clear separation of the different sets of variables, and the description and analysis of the algorithm are simplified considerably. This approach has also been the basis for a successful and highly accurate implementation of FETI-DP algorithms; cf. [23, 17, 18, 16].

Brenner and Sung [3] have also recently established that any such common eigenvalue of the FETI-DP and BDDC algorithms, not equal to 0 or 1, has the same multiplicity. In addition, they give an example for which the eigenvalues of the FETI-DP operator all exceed 1 while the BDDC operator has an eigenvalue equal to 1.

In this paper, a BDDC algorithm is developed for mixed finite element approximations of the incompressible Stokes equations in a way very similar to [29]. If the set of primal constraints on the velocity across the interface satisfies a certain assumption, we are then able to show that the preconditioned saddle-point problem is positive definite when restricted to the subspace that satisfies the primal constraints and that the iterates stay in this subspace. We are then able to use a preconditioned conjugate gradient method and we can, if an additional assumption is satisfied, also prove as strong a bound on the convergence rate as for the standard elliptic case.

We note that the new algorithm has much in common with relatively recent extensions of the classical balancing Neumann–Neumann method to the Stokes equations and almost incompressible elasticity by Pavarino, Goldfeld, and the second author (see [33, 13, 12]), and extensions of the FETI-DP methods developed by the first author in [25, 26, 27]. We note that, in our experience, all these methods converge quite rapidly. Just as in our earlier work, we will work with *benign subspaces*, i.e., subspaces of the mixed finite element spaces on which the saddle-point problem is positive definite. (We note that the same space of functions is called *balanced* in [37, section 9.4.2].) We are also able to prove that any two BDDC and FETI-DP methods, with the same set of primal constraints and which satisfy our first assumption, have the same set of nonzero eigenvalues with the possible exception of 1; this is the same result as given in [31, 29, 3] for the elliptic case.

We note that Dohrmann [5, 6] has recently developed and tested a BDDC method for the related problem of almost incompressible elasticity. We will comment further on his work in section 7. Results for FETI-DP are given by Klawonn, Rheinbach, and Wohlmuth [20]. Recent results on BDDC algorithms for flow in porous media, discretized by mixed finite element methods, are given by Tu [42, 39, 40]. The BDDC

algorithms have also been extended to mortar finite element methods by Kim, Dryja, and the second author [15]. For older references to domain decomposition algorithms for mixed finite element approximations, see [37, Chapter 9].

In addition to deriving and analyzing the algorithms, we also report on some numerical experiments in the final section.

2. Discretization of a saddle-point problem. Let us consider the incompressible Stokes problem on a bounded, polyhedral domain Ω, in two or three dimensions. We denote the boundary of the domain by $\partial\Omega$; for simplicity a homogeneous Dirichlet boundary condition is enforced. (Generally, in order for a divergence free extension of the boundary values to exist, the integral of the normal component of the velocity over the boundary of the region must vanish.) The weak solution satisfies the following saddle-point problem: find $\mathbf{u} \in (H_0^1(\Omega))^d = \{\mathbf{v} \in (H^1(\Omega))^d \mid \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\}$, $d = 2$ or 3, and $p \in L_0^2(\Omega) = \{q \in L^2(\Omega) \mid \int_\Omega q = 0\}$ such that

$$\begin{cases} a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = (\mathbf{f},\mathbf{v}) & \forall \mathbf{v} \in \left(H_0^1(\Omega)\right)^d, \\ b(\mathbf{u},q) \qquad\qquad = 0 & \forall q \in L_0^2(\Omega)\ , \end{cases} \tag{1}$$

where $b(\mathbf{u},q) = -\int_\Omega (\nabla\cdot\mathbf{u})q$, and $a(\mathbf{u},\mathbf{v}) = \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}$, or $a(\mathbf{u},\mathbf{v}) = 2\int_\Omega \epsilon(\mathbf{u}) : \epsilon(\mathbf{v})$. Here the strain tensor $\epsilon(\mathbf{u})$ is defined by $\epsilon_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$, and

$$\nabla\mathbf{u} : \nabla\mathbf{v} = \sum_{i,j=1}^{d} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j} \quad \text{and} \quad \epsilon(\mathbf{u}) : \epsilon(\mathbf{v}) = \sum_{i,j=1}^{d} \epsilon_{ij}(\mathbf{u})\epsilon_{ij}(\mathbf{v}).$$

The operator form of the Stokes problem with Dirichlet boundary conditions is the same for either choice of the bilinear form $a(\cdot,\cdot)$, but we will adopt the second, which gives rise to a natural boundary condition of the form

$$2\sum_{j=1}^{d} \epsilon_{ij} n_j - p n_i = g_i \ \text{ on } \partial\Omega, \quad i = 1,\dots,d. \tag{2}$$

This is the normal component of the stress field. We note that this approach is consistent with the derivation of a physically relevant interface condition in Batchelor's book [1] and also with the discussion in Quarteroni and Valli [35, section 5.3]. There is the further advantage that we will develop a theory which is equally valid for almost incompressible elasticity and that we can draw very directly on some recent results on compressible elasticity by Klawonn and the second author [23]. The following lemma (see [21, Lemma 4], [12, Lemma 1.3], and [23, Lemma 6.4]) shows the equivalence between the chosen bilinear form and that of H^1. Essentially, it is a variant of Korn's second inequality; here $\|\epsilon(\mathbf{u})\|^2_{L^2(\Omega)} = \int_\Omega \epsilon(\mathbf{u}) : \epsilon(\mathbf{u})$.

LEMMA 2.1. *There exists a constant $c > 0$ such that*

$$c\|\nabla\mathbf{u}\|_{L^2(\Omega)} \le \|\epsilon(\mathbf{u})\|_{L^2(\Omega)} \le \|\nabla\mathbf{u}\|_{L^2(\Omega)} \quad \forall \mathbf{u} \in (H^1(\Omega))^d,\ \mathbf{u} \perp \ker(\epsilon),$$

where $\ker(\epsilon)$ *is the space of rigid body motions of the elasticity problem.*

In our mixed finite element methods for solving the saddle-point problem (1), the velocity space (or the space of displacements for the elasticity problems) will be denoted by $\widehat{\mathbf{W}}$. It consists of vector-valued, low order piecewise polynomial functions which are continuous across element boundaries. The pressure space $Q \subset L_0^2(\Omega)$

will consist of scalar, discontinuous, piecewise polynomial functions. A characteristic diameter of the elements of the underlying triangulation is denoted by h. The finite element approximation $(\mathbf{u}, p)$ of the variational problem (1) satisfies

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}, \tag{3}$$

where the matrices A and B represent the restrictions of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to the finite-dimensional space $\widehat{\mathbf{W}} \times Q$. (We will use the same notation for vectors of nodal values and the corresponding finite element functions.)

We will always assume that the chosen mixed finite element space $\widehat{\mathbf{W}} \times Q$ is inf-sup stable, i.e., that there exists a positive constant β, independent of h, such that

$$\sup_{\mathbf{w} \in \widehat{\mathbf{W}}} \frac{b(\mathbf{w}, q)}{\|\mathbf{w}\|_{H^1}} \geq \beta \|q\|_{L^2} \qquad \forall q \in Q. \tag{4}$$

We note that we will only need this estimate for the subdomains, which we will introduce in the next section. This assumption will guarantee that the local subdomain problems, as well as the global one, are well posed.

3. Reduced subdomain interface problem. The domain Ω is decomposed into N nonoverlapping polyhedral subdomains Ω_i, $i = 1, 2, \ldots, N$, of characteristic diameter H. We assume that each of them is a union of a number of shape-regular tetrahedra (or triangles), that there is a uniform bound on these numbers, and that the faces of the subdomains are all convex. The nodes on the boundaries of neighboring subdomains match across the interface $\Gamma = (\cup \partial\Omega_i) \backslash \partial\Omega$. The interface of an individual subdomain Ω_i is defined by $\Gamma_i = \partial\Omega_i \cap \Gamma$. We will denote the set of nodes on Γ_i by $\Gamma_{i,h}$, etc. We assume, as is customary in domain decomposition theory, that the triangulation of each subdomain is quasi-uniform. Our algorithms are also well defined for more irregular subdomains such as those that result from a mesh partitioner, but our theory does not fully cover such cases. The requirements on the subdomains in our full theory are discussed systematically in [37, section 4.2].

We decompose the discrete velocity and pressure spaces $\widehat{\mathbf{W}}$ and Q into

$$\widehat{\mathbf{W}} = \mathbf{W}_I \bigoplus \widehat{\mathbf{W}}_\Gamma, \quad Q = Q_I \bigoplus Q_0. \tag{5}$$

$\mathbf{W}_I$ and Q_I are products of subdomain interior velocity spaces $\mathbf{W}_I^{(i)}$ and subdomain interior pressure spaces $Q_I^{(i)}$, respectively; i.e.,

$$\mathbf{W}_I = \prod_{i=1}^{N} \mathbf{W}_I^{(i)}, \qquad Q_I = \prod_{i=1}^{N} Q_I^{(i)}.$$

The elements of $\mathbf{W}_I^{(i)}$ are supported in the subdomain Ω_i and vanish on its interface Γ_i, while the elements of $Q_I^{(i)}$ are restrictions of elements in Q to Ω_i which satisfy $\int_{\Omega_i} q_I^{(i)} = 0$. $\widehat{\mathbf{W}}_\Gamma$ is the space of traces on Γ of functions in $\widehat{\mathbf{W}}$ and Q_0 is the subspace of Q with constant values $q_0^{(i)}$ in the subdomain Ω_i that satisfy $\int_\Omega q_0 dx = \sum_{i=1}^{N} q_0^{(i)} m(\Omega_i) = 0$, where $m(\Omega_i)$ is the measure of the subdomain Ω_i.

We denote the space of interface velocity variables of the subdomain Ω_i by $\mathbf{W}_\Gamma^{(i)}$, and the associated product space by $\mathbf{W}_\Gamma = \prod_{i=1}^{N} \mathbf{W}_\Gamma^{(i)}$; generally functions in $\mathbf{W}_\Gamma$ are

discontinuous across the interface. $R_\Gamma^{(i)}$ is the restriction operator which maps functions in the continuous interface velocity space $\widehat{\mathbf{W}}_\Gamma$ to their subdomain components in the space $\mathbf{W}_\Gamma^{(i)}$. The direct sum of the $R_\Gamma^{(i)}$ is denoted by R_Γ.

With this decomposition of the solution space as in (5), the global saddle-point problem (3) can be written as follows: find $(\mathbf{u}_I, p_I, \mathbf{u}_\Gamma, p_0) \in (\mathbf{W}_I, Q_I, \widehat{\mathbf{W}}_\Gamma, Q_0)$ such that

$$\begin{bmatrix} A_{II} & B_{II}^T & \widehat{A}_{\Gamma I}^T & 0 \\ B_{II} & 0 & \widehat{B}_{I\Gamma} & 0 \\ \widehat{A}_{\Gamma I} & \widehat{B}_{I\Gamma}^T & \widehat{A}_{\Gamma\Gamma} & \widehat{B}_{0\Gamma}^T \\ 0 & 0 & \widehat{B}_{0\Gamma} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ p_I \\ \mathbf{u}_\Gamma \\ p_0 \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ 0 \\ \mathbf{f}_\Gamma \\ 0 \end{bmatrix}. \tag{6}$$

The leading two-by-two block of this matrix can, by a symmetric permutation, be made into a block diagonal matrix with blocks corresponding to independent subdomain problems. The lower left block in (6) is zero since the bilinear form $b(\mathbf{v}_I, q_0)$ always vanishes for any $\mathbf{v}_I \in \mathbf{W}_I$ and $q_0 \in Q_0$. The blocks related to the continuous interface velocity are assembled from the corresponding subdomain submatrices, e.g., $\widehat{A}_{\Gamma\Gamma} = \sum_{i=1}^N {R_\Gamma^{(i)}}^T A_{\Gamma\Gamma}^{(i)} R_\Gamma^{(i)}$, $\widehat{B}_{0\Gamma} = \sum_{i=1}^N B_{0\Gamma}^{(i)} R_\Gamma^{(i)}$. Correspondingly, the right-hand-side vector $\mathbf{f}_I$ consists of subdomain vectors $\mathbf{f}_I^{(i)}$, and $\mathbf{f}_\Gamma$ is assembled from the subdomain components $\mathbf{f}_\Gamma^{(i)}$. We denote the spaces of right-hand-side vectors $\mathbf{f}_I$ and $\mathbf{f}_\Gamma$ by $\mathbf{F}_I$ and $\mathbf{F}_\Gamma$, respectively; we will also use $\widetilde{\mathbf{F}}_\Gamma$, $\widehat{\mathbf{F}}_\Gamma$, $\widehat{\mathbf{F}}_\Pi$, $\mathbf{F}_\Delta^{(i)}$, and F_0 to represent different spaces of right-hand-side vectors in this paper without providing much detail.

Eliminating the independent subdomain interior variables $(\mathbf{u}_I, p_I)$ from the global problem (6), we have the global interface problem

$$\begin{bmatrix} \widehat{S}_\Gamma & \widehat{B}_{0\Gamma}^T \\ \widehat{B}_{0\Gamma} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_\Gamma \\ p_0 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_\Gamma \\ 0 \end{bmatrix}, \tag{7}$$

where the right-hand-side vector $\mathbf{g}_\Gamma$ is given by

$$\mathbf{g}_\Gamma = \sum_{i=1}^N {R_\Gamma^{(i)}}^T \left\{ \mathbf{f}_\Gamma^{(i)} - \begin{bmatrix} A_{\Gamma I}^{(i)} & B_{I\Gamma}^{(i)^T} \end{bmatrix} \begin{bmatrix} A_{II}^{(i)} & B_{II}^{(i)^T} \\ B_{II}^{(i)} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_I^{(i)} \\ 0 \end{bmatrix} \right\}.$$

$\widehat{S}_\Gamma$ is assembled from subdomain Stokes–Schur complements $S_\Gamma^{(i)}$, which are defined by the following: given $\mathbf{w}_\Gamma^{(i)} \in \mathbf{W}_\Gamma^{(i)}$, determine $S_\Gamma^{(i)} \mathbf{w}_\Gamma^{(i)} \in \mathbf{F}_\Gamma^{(i)}$ such that

$$\begin{bmatrix} A_{II}^{(i)} & B_{II}^{(i)^T} & A_{\Gamma I}^{(i)^T} \\ B_{II}^{(i)} & 0 & B_{I\Gamma}^{(i)} \\ A_{\Gamma I}^{(i)} & B_{I\Gamma}^{(i)^T} & A_{\Gamma\Gamma}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(i)} \\ p_I^{(i)} \\ \mathbf{w}_\Gamma^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ S_\Gamma^{(i)} \mathbf{w}_\Gamma^{(i)} \end{bmatrix}. \tag{8}$$

The leading two-by-two block of (8) corresponds to a Dirichlet problem on the subdomain Ω_i and it is always nonsingular; this is a direct consequence of the assumption of inf-sup stability. We see from (8) that the action of $S_\Gamma^{(i)}$ on a vector can be evaluated at a cost of solving a Dirichlet problem on Ω_i and a few matrix-vector multiplies. We denote the direct sum of the $S_\Gamma^{(i)}$ by S_Γ. Then $\widehat{S}_\Gamma$ is given by

$$\widehat{S}_\Gamma = R_\Gamma^T S_\Gamma R_\Gamma = \sum_{i=1}^N {R_\Gamma^{(i)}}^T S_\Gamma^{(i)} R_\Gamma^{(i)}. \tag{9}$$

We denote the operator of the interface problem (7) by $\widehat{S}$. Since $\widehat{S}$ is symmetric and indefinite, we could use the minimal residual method, possibly with a positive definite block preconditioner, as in [37, section 9.2], to solve problem (7). We will instead propose a different type of preconditioner and show that the preconditioned operator is positive definite, provided that a suitable set of primal constraints are chosen; cf. Assumption 1. A preconditioned conjugate gradient method can then be used.

4. A BDDC preconditioner for Stokes equations. When using a BDDC or FETI-DP method, we relax most, but not all, of the continuity constraints on the velocity across the interface; we will always retain sufficiently many *primal* continuity constraints to assure that we will never encounter a need for solving any singular linear systems of algebraic equations. In a BDDC algorithm, full continuity is restored, at the end of each iteration step, by using an averaging operator, while in a FETI-DP algorithm, continuity will not be fully satisfied until the algorithm has converged. The primal constraints should also be chosen so that the rate of convergence of the iterative method is enhanced.

For our purposes, we introduce a partially assembled interface velocity space $\widetilde{\mathbf{W}}_\Gamma$:

$$\widetilde{\mathbf{W}}_\Gamma = \widehat{\mathbf{W}}_\Pi \bigoplus \mathbf{W}_\Delta = \widehat{\mathbf{W}}_\Pi \bigoplus \left(\prod_{i=1}^{N} \mathbf{W}_\Delta^{(i)} \right).$$

Here, $\widehat{\mathbf{W}}_\Pi$ is the continuous, coarse level, primal interface velocity space which is typically spanned by subdomain vertex nodal basis functions, and/or by interface edge and/or face basis functions with constant values, or with values of positive weight functions, on these edges or faces. These basis functions correspond to the primal interface velocity continuity constraints, which will be discussed in section 7. We will always assume that the basis has been changed so that each primal basis function corresponds to an explicit degree of freedom. In other words, we will have explicit primal unknowns corresponding to the primal continuity constraints on edges or faces as indicated in section 1, and further described in [29], [23, section 4], and [17]. The primal, coarse level degrees of freedom are shared by neighboring subdomains. The complementary space $\mathbf{W}_\Delta$ is the product of the subdomain dual interface velocity spaces $\mathbf{W}_\Delta^{(i)}$, which correspond to the remaining interface velocity degrees of freedom and are spanned by basis functions which vanish at the primal degrees of freedom. Thus, an element in the space $\widetilde{\mathbf{W}}_\Gamma$ has a continuous primal velocity component and typically a discontinuous dual velocity component.

We need to introduce several restriction, extension, and scaling operators between a variety of spaces. As in section 3, $R_\Gamma^{(i)}$ is the restriction operator which maps a function in the space $\widehat{\mathbf{W}}_\Gamma$ to its component in $\mathbf{W}_\Gamma^{(i)}$. We define $R_\Delta^{(i)}$ as the operator which maps functions in the space $\widehat{\mathbf{W}}_\Gamma$ to its dual component in the space $\mathbf{W}_\Delta^{(i)}$. $R_{\Gamma\Pi}$ is the restriction operator from the space $\widehat{\mathbf{W}}_\Gamma$ to its subspace $\widehat{\mathbf{W}}_\Pi$; $R_\Pi^{(i)}$ is the operator which maps $\widehat{\mathbf{W}}_\Pi$ into its Γ_i-component. $\widetilde{R}_\Gamma$ is the direct sum of $R_{\Gamma\Pi}$ and the $R_\Delta^{(i)}$, and it is a map from $\widehat{\mathbf{W}}_\Gamma$ into $\widetilde{\mathbf{W}}_\Gamma$.

In order to define certain scaling operators, which will be used in the definition of the BDDC preconditioner, we introduce a positive scaling factor $\delta_i^\dagger(x)$ for the nodes x on the interface Γ_i of each subdomain Ω_i. For the incompressible Stokes problems, with $\mathcal{I}_x$ the set of indices of the subdomains which have x on their boundaries, we will only need to use inverse counting functions defined by $\delta_i^\dagger(x) = 1/card(\mathcal{I}_x), x \in \Gamma_{i,h}$,

where $card(\mathcal{I}_x)$ is the number of the subdomain boundaries to which x belongs. It is then easy to see that $\sum_{j\in\mathcal{I}_x}\delta_j^\dagger(x)=1$ for any $x\in\Gamma_{i,h}$. Given the scaling factors at the subdomain interface nodes, we can define scaled restriction operators $R_{D,\Delta}^{(i)}$. We first note that each row of $R_\Delta^{(i)}$ has only one nonzero entry, which corresponds to a node $x\in\Gamma_{i,h}$. Multiplying each such element with the scaling factor $\delta_i^\dagger(x)$ gives us $R_{D,\Delta}^{(i)}$. The scaled operator $\widetilde{R}_{D,\Gamma}$ is the direct sum of $R_{\Gamma\Pi}$ and the $R_{D,\Delta}^{(i)}$. (For elasticity problems, these scaling factors should depend on the first Lamé constant μ, which can be allowed to change across the interface between neighboring subdomains; see [37, section 8.5.1] and [23].)

The interface velocity Schur complement $\widetilde{S}_\Gamma$ is defined on the partially assembled interface velocity space $\widetilde{\mathbf{W}}_\Gamma$ by the following: given $\mathbf{w}_\Gamma\in\widetilde{\mathbf{W}}_\Gamma$, $\widetilde{S}_\Gamma\mathbf{w}_\Gamma\in\widetilde{\mathbf{F}}_\Gamma$ satisfies

$$\begin{bmatrix} A_{II}^{(1)} & B_{II}^{(1)^T} & A_{\Delta I}^{(1)^T} & & \widetilde{A}_{\Pi I}^{(1)^T} \\ B_{II}^{(1)} & 0 & B_{I\Delta}^{(1)} & & \widetilde{B}_{I\Pi}^{(1)} \\ A_{\Delta I}^{(1)} & B_{I\Delta}^{(1)^T} & A_{\Delta\Delta}^{(1)} & & \widetilde{A}_{\Pi\Delta}^{(1)^T} \\ & & & \ddots & \vdots \\ \widetilde{A}_{\Pi I}^{(1)} & \widetilde{B}_{I\Pi}^{(1)^T} & \widetilde{A}_{\Pi\Delta}^{(1)} & \dots & \widetilde{A}_{\Pi\Pi} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ p_I^{(1)} \\ \mathbf{w}_\Delta^{(1)} \\ \vdots \\ \mathbf{w}_\Pi \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ (\widetilde{S}_\Gamma\mathbf{w}_\Gamma)_\Delta^{(1)} \\ \vdots \\ (\widetilde{S}_\Gamma\mathbf{w}_\Gamma)_\Pi \end{bmatrix}. \tag{10}$$

Here $\widetilde{A}_{\Pi\Pi}=\sum_{i=1}^N R_\Pi^{(i)^T}A_{\Pi\Pi}^{(i)}R_\Pi^{(i)}$, $\widetilde{A}_{\Pi I}^{(i)}=R_\Pi^{(i)^T}A_{\Pi I}^{(i)}$, $\widetilde{A}_{\Pi\Delta}^{(i)}=R_\Pi^{(i)^T}A_{\Pi\Delta}^{(i)}$, and $\widetilde{B}_{I\Pi}^{(i)}=B_{I\Pi}^{(i)}R_\Pi^{(i)}$.

From the definition of $\widetilde{S}_\Gamma$, we see that it can be obtained from the subdomain Schur complements $S_\Gamma^{(i)}$ by assembling only with respect to the primal interface velocity part, i.e., as

$$\widetilde{S}_\Gamma=\overline{R}_\Gamma^T S_\Gamma\overline{R}_\Gamma. \tag{11}$$

Here $\overline{R}_\Gamma$ is the restriction from the space $\widetilde{\mathbf{W}}_\Gamma$ into the product space $\mathbf{W}_\Gamma$ associated with the set of subdomains. We recall that the global interface Schur operator $\widehat{S}_\Gamma$ is obtained by fully assembling the $S_\Gamma^{(i)}$ across the subdomain interface; cf. (9). $\widehat{S}_\Gamma$ can therefore also be obtained from $\widetilde{S}_\Gamma$ by further assembling with respect to the dual interface velocity part; i.e., we have $\widehat{S}_\Gamma=\widetilde{R}_\Gamma^T\widetilde{S}_\Gamma\widetilde{R}_\Gamma$. Correspondingly, we define an operator $\widetilde{B}_{0\Gamma}$, which maps the partially assembled interface velocity space $\widetilde{\mathbf{W}}_\Gamma$ into F_0, the space of right-hand sides corresponding to Q_0. $\widetilde{B}_{0\Gamma}$ is obtained from the subdomain operators $B_{0\Gamma}^{(i)}$ by assembling with respect to the primal interface velocity part. The operator $\widehat{B}_{0\Gamma}$ can then be obtained from $\widetilde{B}_{0\Gamma}$ by assembling with respect to the dual interface velocity part on the subdomain interfaces, i.e., $\widehat{B}_{0\Gamma}=\widetilde{B}_{0\Gamma}\widetilde{R}_\Gamma$. We can then write $\widehat{S}$, the operator of the global interface problem (7), as

$$\widehat{S}=\begin{bmatrix} \widehat{S}_\Gamma & \widehat{B}_{0\Gamma}^T \\ \widehat{B}_{0\Gamma} & 0 \end{bmatrix}=\begin{bmatrix} \widetilde{R}_\Gamma^T\widetilde{S}_\Gamma\widetilde{R}_\Gamma & \widetilde{R}_\Gamma^T\widetilde{B}_{0\Gamma}^T \\ \widetilde{B}_{0\Gamma}\widetilde{R}_\Gamma & 0 \end{bmatrix}=\widetilde{R}^T\widetilde{S}\widetilde{R}, \tag{12}$$

where we use the notation

$$\widetilde{R}=\begin{bmatrix} \widetilde{R}_\Gamma & \\ & I \end{bmatrix},\quad \widetilde{S}=\begin{bmatrix} \widetilde{S}_\Gamma & \widetilde{B}_{0\Gamma}^T \\ \widetilde{B}_{0\Gamma} & 0 \end{bmatrix}. \tag{13}$$

The preconditioner for solving the global interface saddle-point problem (7) is

$$M^{-1} = \widetilde{R}_D^T \widetilde{S}^{-1} \widetilde{R}_D. \tag{14}$$

Here $\widetilde{R}_D$ is of the same form as $\widetilde{R}$ in (13), except that $\widetilde{R}_\Gamma$ is replaced by the scaled operator $\widetilde{R}_{D,\Gamma}$. It is easy to see that $\widetilde{R}_{D,\Gamma}$ is of full rank and that the preconditioner is nonsingular. To determine $\widetilde{S}^{-1}\mathbf{g}$ for any given $\mathbf{g} = (\mathbf{g}_\Gamma, g_0) \in \widetilde{\mathbf{F}}_\Gamma \times F_0$, we need to solve the linear system

$$\begin{bmatrix} \widetilde{S}_\Gamma & \widetilde{B}_{0\Gamma}^T \\ \widetilde{B}_{0\Gamma} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_\Gamma \\ p_0 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_\Gamma \\ g_0 \end{bmatrix}. \tag{15}$$

Given the definition of $\widetilde{S}_\Gamma$ in (10), we know that solving (15) is equivalent to solving

$$\begin{bmatrix} A_{II}^{(1)} & B_{II}^{(1)^T} & A_{\Delta I}^{(1)^T} & & \widetilde{A}_{\Pi I}^{(1)^T} & \\ B_{II}^{(1)} & 0 & B_{I\Delta}^{(1)} & & \widetilde{B}_{I\Pi}^{(1)} & \\ A_{\Delta I}^{(1)} & B_{I\Delta}^{(1)^T} & A_{\Delta\Delta}^{(1)} & & \widetilde{A}_{\Pi\Delta}^{(1)^T} & B_{0\Delta}^{(1)^T} \\ & & & \ddots & \vdots & \\ \widetilde{A}_{\Pi I}^{(1)} & \widetilde{B}_{I\Pi}^{(1)^T} & \widetilde{A}_{\Pi\Delta}^{(1)} & \dots & \widetilde{A}_{\Pi\Pi} & \widetilde{B}_{0\Pi}^T \\ & & B_{0\Delta}^{(1)} & & \widetilde{B}_{0\Pi} & \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ p_I^{(1)} \\ \mathbf{u}_\Delta^{(1)} \\ \vdots \\ \mathbf{u}_\Pi \\ p_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{g}_\Delta^{(1)} \\ \vdots \\ \mathbf{g}_\Pi \\ g_0 \end{bmatrix}, \tag{16}$$

where $\widetilde{B}_{0\Pi} = \sum_{i=1}^N B_{0\Pi}^{(i)} R_\Pi^{(i)}$. As in [29], a block factorization can be used to solve (16). The leading diagonal subdomain matrix blocks are eliminated to form a coarse level Schur complement problem for $(\mathbf{u}_\Pi, p_0)$. After solving the coarse level problem, the subdomain variables $(\mathbf{u}_I^{(i)}, p_I^{(i)}, \mathbf{u}_\Delta^{(i)})$ are computed by solving independent subdomain problems.

5. Benign subspaces and quotient spaces. The subdomain Schur complements $S_\Gamma^{(i)}$ are symmetric, positive semidefinite. This is a consequence of a well-known result on the inertia of Schur complements. We know, e.g., that the number of negative eigenvalues of a symmetric, two-by-two block matrix equals the sum of the number of negative eigenvalues of the leading block and those of the Schur complement formed by eliminating the variables of the leading block. Thus, we have the following lemma.

LEMMA 5.1. *The subdomain Schur complements $S_\Gamma^{(i)}$, defined in* (8)*, are symmetric, positive semidefinite, and singular for any subdomain with a boundary that does not intersect $\partial\Omega$.*

The $S_\Gamma^{(i)}$- and S_Γ-seminorms are defined by $|\mathbf{w}_\Gamma^{(i)}|^2_{S_\Gamma^{(i)}} = \mathbf{w}_\Gamma^{(i)^T} S_\Gamma^{(i)} \mathbf{w}_\Gamma^{(i)}$ and $|\mathbf{w}_\Gamma|^2_{S_\Gamma} = \mathbf{w}_\Gamma^T S_\Gamma \mathbf{w}_\Gamma = \sum_{i=1}^N |\mathbf{w}_\Gamma^{(i)}|^2_{S_\Gamma^{(i)}}$. The $|\cdot|_{\mathbf{E}(\Gamma_i)}$-seminorm is defined on the space $\mathbf{W}_\Gamma^{(i)}$ by

$$|\mathbf{w}_\Gamma^{(i)}|_{\mathbf{E}(\Gamma_i)} = \inf_{\substack{\mathbf{v}^{(i)} \in (H^1(\Omega_i))^d \\ \mathbf{v}^{(i)}|_{\Gamma_i} = \mathbf{w}_\Gamma^{(i)}}} \|\epsilon(\mathbf{v}^{(i)})\|_{L^2(\Omega_i)},$$

and a seminorm on $\mathbf{W}_\Gamma$ by $|\mathbf{w}_\Gamma|^2_{\mathbf{E}(\Gamma)} = \sum_{i=1}^N |\mathbf{w}_\Gamma^{(i)}|^2_{\mathbf{E}(\Gamma_i)}$.

The following lemma shows the equivalence of the $|\cdot|_{S_\Gamma}$- and $|\cdot|_{\mathbf{E}(\Gamma)}$-seminorms. It can be found essentially in Bramble and Pasciak [2, Theorem 4.1], or Pavarino and

Widlund [33, Lemma 3.1], for incompressible Stokes problems. This same result is also valid for the incompressible elasticity problem and with the underlying bilinear form $a(\cdot,\cdot)$ given in terms of the strain tensor; cf. Lemma 2.1 and [23].

LEMMA 5.2. *There exists a positive constant c, which is independent of H and h, such that*

$$c\beta^2|\mathbf{w}_\Gamma^{(i)}|^2_{S_\Gamma^{(i)}} \leq |\mathbf{w}_\Gamma^{(i)}|^2_{\mathbf{E}(\Gamma_i)} \leq |\mathbf{w}_\Gamma^{(i)}|^2_{S_\Gamma^{(i)}} \quad \forall \mathbf{w}_\Gamma^{(i)} \in \mathbf{W}_\Gamma^{(i)},$$

where β is the inf-sup stability constant defined in (4).

The operators $\widehat{S}_\Gamma$ and $\widetilde{S}_\Gamma$, given in (9) and (11), are both symmetric, positive definite because of the Dirichlet boundary conditions on $\partial\Omega$ and the fact that sufficiently many primal continuity constraints are always chosen. We can then define the $\widehat{S}_\Gamma$- and $\widetilde{S}_\Gamma$-norms on the spaces $\widehat{\mathbf{W}}_\Gamma$ and $\widetilde{\mathbf{W}}_\Gamma$, respectively, by

$$\|\mathbf{w}_\Gamma\|^2_{\widehat{S}_\Gamma} = \mathbf{w}_\Gamma^T R_\Gamma^T S_\Gamma R_\Gamma \mathbf{w}_\Gamma = |R_\Gamma \mathbf{w}_\Gamma|^2_{S_\Gamma} \quad \forall \mathbf{w}_\Gamma \in \widehat{\mathbf{W}}_\Gamma, \tag{17}$$

$$\|\mathbf{w}_\Gamma\|^2_{\widetilde{S}_\Gamma} = \mathbf{w}_\Gamma^T \overline{R}_\Gamma^T S_\Gamma \overline{R}_\Gamma \mathbf{w}_\Gamma = |\overline{R}_\Gamma \mathbf{w}_\Gamma|^2_{S_\Gamma} \quad \forall \mathbf{w}_\Gamma \in \widetilde{\mathbf{W}}_\Gamma. \tag{18}$$

Two subspaces of $\widehat{\mathbf{W}}_\Gamma$ and $\widetilde{\mathbf{W}}_\Gamma$ are defined as follows.

DEFINITION 1.

$$\widehat{\mathbf{W}}_{\Gamma,B} = \{\mathbf{w}_\Gamma \in \widehat{\mathbf{W}}_\Gamma \mid \widehat{B}_{0\Gamma}\mathbf{w}_\Gamma = 0\} \text{ and } \widetilde{\mathbf{W}}_{\Gamma,B} = \{\mathbf{w}_\Gamma \in \widetilde{\mathbf{W}}_\Gamma \mid \widetilde{B}_{0\Gamma}\mathbf{w}_\Gamma = 0\}.$$

We will call $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$ and $\widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$ the *benign subspaces* of $\widehat{\mathbf{W}}_\Gamma \times Q_0$ and $\widetilde{\mathbf{W}}_\Gamma \times Q_0$, respectively. The interface problem operator $\widehat{S}$ of (7) is indefinite on the space $\widehat{\mathbf{W}}_\Gamma \times Q_0$. But restricted to the subspace $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, it is positive semidefinite, which follows from the fact that, for any $\mathbf{w} = (\mathbf{w}_\Gamma, q_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$,

$$\mathbf{w}^T \widehat{S}\mathbf{w} = \begin{bmatrix} \mathbf{w}_\Gamma^T & q_0^T \end{bmatrix} \begin{bmatrix} \widehat{S}_\Gamma & \widehat{B}_{0\Gamma}^T \\ \widehat{B}_{0\Gamma} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_\Gamma \\ q_0 \end{bmatrix} = \mathbf{w}_\Gamma^T \widehat{S}_\Gamma \mathbf{w}_\Gamma = \|\mathbf{w}_\Gamma\|^2_{\widehat{S}_\Gamma} \geq 0. \tag{19}$$

The same is also true for the operator $\widetilde{S}$ on the space $\widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$. Thus, $\widehat{S}$- and $\widetilde{S}$-seminorms can be defined on the benign subspaces by

$$|\mathbf{w}|^2_{\widehat{S}} = \mathbf{w}^T \widehat{S}\mathbf{w} = \|\mathbf{w}_\Gamma\|^2_{\widehat{S}_\Gamma} \qquad \forall \mathbf{w} = (\mathbf{w}_\Gamma,\ q_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0, \tag{20}$$

$$|\mathbf{w}|^2_{\widetilde{S}} = \mathbf{w}^T \widetilde{S}\mathbf{w} = \|\mathbf{w}_\Gamma\|^2_{\widetilde{S}_\Gamma} \qquad \forall \mathbf{w} = (\mathbf{w}_\Gamma,\ q_0) \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0. \tag{21}$$

For elements $\mathbf{w} = (\mathbf{0}, q_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, $\mathbf{w}^T \widehat{S}\mathbf{w} = 0$; we denote the space of such elements by $\widehat{Y}$ and introduce the quotient space $(\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0)/\widehat{Y}$. An element of this space, $\{\mathbf{w}_\Gamma,\ q_0\}$, is the congruence class containing the vector $(\mathbf{w}_\Gamma,\ q_0)$. Two elements of $(\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0)/\widehat{Y}$, $\{\mathbf{v}_\Gamma, p_0\}$ and $\{\mathbf{w}_\Gamma, q_0\}$, are the same if $\mathbf{v}_\Gamma = \mathbf{w}_\Gamma$ and the zero element of the space can be represented by any $(\mathbf{0}, q_0)$.

We see that even though $\widehat{S}$ is only positive semidefinite on the space $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, it becomes positive definite when considered on the quotient space $(\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0)/\widehat{Y}$. In the next section, we will show that, under an assumption on the choice of the primal velocity continuity constraints of the BDDC algorithm, the preconditioned BDDC operator $M^{-1}\widehat{S}$ is positive definite on the quotient space and that, correspondingly, a preconditioned conjugate gradient iteration restricted to the quotient space will be successful.

6. Condition number bounds. We first define an averaging operator $E_D = \widetilde{R}\widetilde{R}_D^T$, which maps $\widetilde{\mathbf{W}}_\Gamma \times Q_0$, with generally discontinuous interface velocities, to elements with continuous interface velocities in the same space: for any $\mathbf{w} = (\mathbf{w}_\Gamma, q_0) \in \widetilde{\mathbf{W}}_\Gamma \times Q_0$,

(22)
$$E_D \begin{bmatrix} \mathbf{w}_\Gamma \\ q_0 \end{bmatrix} = \begin{bmatrix} \widetilde{R}_\Gamma & \\ & I \end{bmatrix} \begin{bmatrix} \widetilde{R}_{D,\Gamma}^T & \\ & I \end{bmatrix} \begin{bmatrix} \mathbf{w}_\Gamma \\ q_0 \end{bmatrix} = \begin{bmatrix} E_{D,\Gamma}\mathbf{w}_\Gamma \\ q_0 \end{bmatrix} \in \widetilde{\mathbf{W}}_\Gamma \times Q_0,$$

where $E_{D,\Gamma} = \widetilde{R}_\Gamma \widetilde{R}_{D,\Gamma}^T$ provides the average of the interface velocities across the interface Γ.

Two assumptions will be needed for the condition number bound of the preconditioned operator; recipes for which these assumptions hold will be provided in section 7. The first assumption is a requirement on the primal velocity continuity constraints of the BDDC algorithm.

Assumption 1. The primal interface velocity continuity constraints in the BDDC algorithm are chosen such that $\int_{\partial\Omega_i} (R_\Delta^{(i)}\mathbf{w}_\Gamma) \cdot \mathbf{n}_i = 0 \ \forall \mathbf{w}_\Gamma \in \widehat{\mathbf{W}}_\Gamma$, with $\mathbf{n}_i$ the unit outward normal of $\partial\Omega_i$. Equivalently, $B_{0\Delta}^{(i)}(R_\Delta^{(i)}\mathbf{w}_\Gamma) = 0$.

The following results follow from Assumption 1.

LEMMA 6.1. *Let Assumption* 1 *hold. Then all matrices* $B_{0\Delta}^{(i)}$, $i = 1, \ldots, N$, *vanish.*

Proof. Since $R_\Delta^{(i)}$ is a mapping onto the space $\mathbf{W}_\Delta^{(i)}$, on which $B_{0\Delta}^{(i)}$ is defined, it follows from Assumption 1 that $B_{0\Delta}^{(i)}$ must vanish. ∎

LEMMA 6.2. *Let Assumption* 1 *hold. Then* $\widetilde{R}_D^T\mathbf{w} \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$ *for any* $\mathbf{w} \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$.

Proof. We need to show that given $\mathbf{w}_\Gamma = \mathbf{w}_\Pi + \mathbf{w}_\Delta \in \widetilde{\mathbf{W}}_{\Gamma,B}$, $\widehat{B}_{0\Gamma}\widetilde{R}_{D,\Gamma}^T\mathbf{w}_\Gamma = 0$. Since $\widetilde{B}_{0\Gamma}\mathbf{w}_\Gamma = 0$, we have from Lemma 6.1 that $\widetilde{B}_{0\Gamma}\mathbf{w}_\Gamma = \widetilde{B}_{0\Pi}R_{\Gamma\Pi}\widetilde{R}_\Gamma^T\mathbf{w}_\Gamma = \widetilde{B}_{0\Pi}\mathbf{w}_\Pi = 0$. From Lemma 6.1, we also know that $\widehat{B}_{0\Gamma}\widetilde{R}_{D,\Gamma}^T\mathbf{w}_\Gamma = \widetilde{B}_{0\Pi}R_{\Gamma\Pi}\widetilde{R}_{D,\Gamma}^T\mathbf{w}_\Gamma = \widetilde{B}_{0\Pi}\mathbf{w}_\Pi$. Therefore, $\widehat{B}_{0\Gamma}\widetilde{R}_{D,\Gamma}^T\mathbf{w}_\Gamma = 0$. ∎

LEMMA 6.3. *Let Assumption* 1 *hold. Then* $\widehat{S}$ *is an isomorphism from the space* $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$ *to* $\widehat{\mathbf{F}} \times \{0\}$ *and* M^{-1} *an isomorphism from the space* $\widehat{\mathbf{F}} \times \{0\}$ *to* $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$. *For any* $q_0 \in Q_0$,

$$M^{-1} \begin{bmatrix} \widehat{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ q_0 \end{bmatrix}.$$

Proof. We know from (7) that $\widehat{S}$ maps the space $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$ into $\widehat{\mathbf{F}} \times \{0\}$, and from (14), (15), and Lemma 6.2 that M^{-1} maps $\widehat{\mathbf{F}} \times \{0\}$ into $\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$. The first part of the lemma then follows since we have established that $\widehat{S}$ and M^{-1} are invertible.

To prove the second part, we observe, using Lemma 6.1, that for any $q_0 \in Q_0$, $\widehat{B}_{0\Gamma}^T q_0 = R_{\Gamma\Pi}^T \widetilde{B}_{0\Pi}^T q_0$, and $\widetilde{B}_{0\Gamma}^T q_0 = \widetilde{R}_\Gamma R_{\Gamma\Pi}^T \widetilde{B}_{0\Pi}^T q_0$, which equals $\widetilde{R}_{D,\Gamma} R_{\Gamma\Pi}^T \widetilde{B}_{0\Pi}^T q_0$ since $\widetilde{R}_{D,\Gamma}$ does not change the primal part of any element. We then have, from the definition of M^{-1}, that

$$M^{-1} \begin{bmatrix} \widehat{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix} = \widetilde{R}_D^T \widetilde{S}^{-1} \widetilde{R}_D \begin{bmatrix} R_{\Gamma\Pi}^T \widetilde{B}_{0\Pi}^T q_0 \\ 0 \end{bmatrix} = \widetilde{R}_D^T \widetilde{S}^{-1} \begin{bmatrix} \widetilde{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix}.$$

From the definition of $\widetilde{S}$ in (15), we know that the right-hand side equals $(\mathbf{0}, q_0)$. ∎

LEMMA 6.4. *Let Assumption* 1 *hold. Then any vector of the form* $\mathbf{u} = (\mathbf{0}, q_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$ *is an eigenvector of the preconditioned operator* $M^{-1}\widehat{S}$ *with an eigenvalue equal to* 1.

Proof. Given any vector $\mathbf{u} = \left[\begin{smallmatrix} \mathbf{0} \\ q_0 \end{smallmatrix}\right] \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$,

$$\widehat{S}\mathbf{u} = \begin{bmatrix} \widehat{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix}.$$

The lemma then follows from Lemma 6.3. ∎

The eigenvalues of the preconditioned operator $M^{-1}\widehat{S}$, when restricted to the quotient space $(\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0)/\widehat{Y}$, are determined by the eigenvalues of $M^{-1}\widehat{S}$ with eigenvectors with nonzero velocity components. We first prove a lower bound on the eigenvalues of $M^{-1}\widehat{S}$, when restricted to the quotient space.

LEMMA 6.5. *For any* $\mathbf{u} = (\mathbf{u}_\Gamma, p_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$,

$$\langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}} \leq \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}.$$

Proof. Given $\mathbf{u} \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, let $\mathbf{w} = \widetilde{S}^{-1}\widetilde{R}_D\widehat{S}\mathbf{u} \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$. We have, from the fact that $\widetilde{R}^T\widetilde{R}_D = \widetilde{R}_D^T\widetilde{R} = I$,

$$\langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}} = \mathbf{u}^T\widehat{S}\widetilde{R}_D^T\widetilde{R}\mathbf{u} = \mathbf{u}^T\widehat{S}\widetilde{R}_D^T\widetilde{S}^{-1}\widetilde{S}\widetilde{R}\mathbf{u} = \left\langle \mathbf{w}, \widetilde{R}\mathbf{u} \right\rangle_{\widetilde{\mathrm{S}}}. \tag{23}$$

Using the Cauchy–Schwarz inequality and the fact that $\widehat{S} = \widetilde{R}^T\widetilde{S}\widetilde{R}$, we find that

$$\left\langle \mathbf{w}, \widetilde{R}\mathbf{u} \right\rangle_{\widetilde{\mathrm{S}}} \leq \langle \mathbf{w}, \mathbf{w} \rangle_{\widetilde{\mathrm{S}}}^{1/2} \left\langle \widetilde{R}\mathbf{u}, \widetilde{R}\mathbf{u} \right\rangle_{\widetilde{\mathrm{S}}}^{1/2} = \langle \mathbf{w}, \mathbf{w} \rangle_{\widetilde{\mathrm{S}}}^{1/2} \langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}}^{1/2}. \tag{24}$$

Therefore, from (23) and (24), we have $\langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}} \leq \langle \mathbf{w}, \mathbf{w} \rangle_{\widetilde{\mathrm{S}}}$. Since

$$\langle \mathbf{w}, \mathbf{w} \rangle_{\widetilde{\mathrm{S}}} = \mathbf{u}^T\widehat{S}\widetilde{R}_D^T\widetilde{S}^{-1}\widetilde{S}\widetilde{S}^{-1}\widetilde{R}_D\widehat{S}\mathbf{u} = \left\langle \mathbf{u}, \widetilde{R}_D^T\widetilde{S}^{-1}\widetilde{R}_D\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}} = \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}, \tag{25}$$

we have $\langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}} \leq \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}$. ∎

In order to obtain a scalable upper eigenvalue bound, we need a second assumption which concerns the stability of the averaging operator $E_{D,\Gamma}$ on the space $\widetilde{\mathbf{W}}_\Gamma$; it is quite similar to those introduced in [32, 24, 23], for standard elliptic problems.

Assumption 2. There exists a positive constant C, which is independent of H, h, and the number of subdomains, such that

$$|\overline{R}_\Gamma \left(E_{D,\Gamma}\mathbf{w}_\Gamma\right)|_{\mathbf{E}(\Gamma)} \leq C\left(1 + \log\frac{H}{h}\right)|\overline{R}_\Gamma\mathbf{w}_\Gamma|_{\mathbf{E}(\Gamma)} \quad \forall \mathbf{w}_\Gamma \in \widetilde{\mathbf{W}}_\Gamma.$$

The following lemma can be proved by using Assumptions 1 and 2.

LEMMA 6.6. *Let Assumptions* 1 *and* 2 *hold. There then exists a positive constant* C*, which is independent of* H*,* h*, and the number of subdomains, such that*

$$|E_D\mathbf{w}|_{\widetilde{\mathrm{S}}} \leq C\frac{1}{\beta}\left(1 + \log\frac{H}{h}\right)|\mathbf{w}|_{\widetilde{\mathrm{S}}} \quad \forall \mathbf{w} = (\mathbf{w}_\Gamma, q_0) \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0,$$

where β is the inf-sup stability constant of (4).

Proof. Given any vector $\mathbf{w} = (\mathbf{w}_\Gamma, q_0) \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$, we know, from Lemma 6.2, that $\widetilde{R}_D^T \mathbf{w} \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$. Therefore, $E_D \mathbf{w} = \widetilde{R}\widetilde{R}_D^T \mathbf{w} \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$. From the definition of the $\widetilde{S}$-seminorm in (21), we have

$$|E_D \mathbf{w}|_{\widetilde{\mathrm{S}}}^2 = \|E_{D,\Gamma} \mathbf{w}_\Gamma\|_{\widetilde{\mathrm{S}}_\Gamma}^2 = |\overline{R}_\Gamma \left(E_{D,\Gamma} \mathbf{w}_\Gamma\right)|_{S_\Gamma}^2 \leq C \frac{1}{\beta^2} |\overline{R}_\Gamma \left(E_{D,\Gamma} \mathbf{w}_\Gamma\right)|_{\mathbf{E}(\Gamma)}^2, \tag{26}$$

where the last inequality follows from Lemma 5.2.

We have, from Assumption 2, Lemma 5.2, and (18),

$$\begin{aligned} |\overline{R}_\Gamma \left(E_{D,\Gamma} \mathbf{w}_\Gamma\right)|_{\mathbf{E}(\Gamma)}^2 &\leq C \left(1 + \log \frac{H}{h}\right)^2 |\overline{R}_\Gamma \mathbf{w}_\Gamma|_{\mathbf{E}(\Gamma)}^2 \\ &\leq C \left(1 + \log \frac{H}{h}\right)^2 |\overline{R}_\Gamma \mathbf{w}_\Gamma|_{S_\Gamma}^2 = C \left(1 + \log \frac{H}{h}\right)^2 \|\mathbf{w}_\Gamma\|_{\widetilde{\mathrm{S}}_\Gamma}^2. \end{aligned} \tag{27}$$

Then from (26), (27), and (21), we have

$$|E_D \mathbf{w}|_{\widetilde{\mathrm{S}}}^2 \leq C \frac{1}{\beta^2} \left(1 + \log \frac{H}{h}\right)^2 \|\mathbf{w}_\Gamma\|_{\widetilde{\mathrm{S}}_\Gamma}^2 = C \frac{1}{\beta^2} \left(1 + \log \frac{H}{h}\right)^2 |\mathbf{w}|_{\widetilde{\mathrm{S}}}^2. \qquad \square$$

THEOREM 6.7. *Let Assumptions* 1 *and* 2 *hold. The preconditioned operator $M^{-1}\widehat{S}$ is then symmetric, positive definite with respect to the bilinear form $\langle \cdot, \cdot \rangle_{\widehat{\mathrm{S}}}$ on the space $(\widehat{\mathbf{W}}_{\Gamma,B} \times Q_0)/\widehat{Y}$. Its eigenvalues are bounded from below by* 1 *and from above by $C \frac{1}{\beta^2} \left(1 + \log(H/h)\right)^2$, where C is a constant which is independent of H, h, and the number of subdomains and β is the inf-sup stability constant defined in* (4).

Proof. It is sufficient to prove that, for any $\mathbf{u} = (\mathbf{u}_\Gamma, p_0) \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, with $\mathbf{u}_\Gamma \neq \mathbf{0}$,

$$\langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}} \leq \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}} \leq C \frac{1}{\beta^2} \left(1 + \log \frac{H}{h}\right)^2 \langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}}.$$

The lower bound has already been established in Lemma 6.5. For the upper bound, given $\mathbf{u} \in \widehat{\mathbf{W}}_{\Gamma,B} \times Q_0$, let $\mathbf{w} = \widetilde{S}^{-1}\widetilde{R}_D \widehat{S}\mathbf{u} \in \widetilde{\mathbf{W}}_{\Gamma,B} \times Q_0$, the same element as in the proof of the lower bound in Lemma 6.5. We have $\widetilde{R}_D^T \mathbf{w} = M^{-1}\widehat{S}\mathbf{u}$. Since $\widehat{S} = \widetilde{R}^T \widetilde{S} \widetilde{R}$, we have, by using Lemma 6.6,

$$\begin{aligned} \left\langle M^{-1}\widehat{S}\mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}} = \left\langle \widetilde{R}_D^T \mathbf{w}, \widetilde{R}_D^T \mathbf{w} \right\rangle_{\widehat{\mathrm{S}}} &= \left\langle \widetilde{R}\widetilde{R}_D^T \mathbf{w}, \widetilde{R}\widetilde{R}_D^T \mathbf{w} \right\rangle_{\widetilde{\mathrm{S}}} \\ = |E_D \mathbf{w}|_{\widetilde{\mathrm{S}}}^2 &\leq C \frac{1}{\beta^2} \left(1 + \log \frac{H}{h}\right)^2 |\mathbf{w}|_{\widetilde{\mathrm{S}}}^2. \end{aligned}$$

Therefore, from (25), we have

$$\left\langle M^{-1}\widehat{S}\mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}} \leq C \frac{1}{\beta^2} \left(1 + \log \frac{H}{h}\right)^2 \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}. \tag{28}$$

Using the Cauchy–Schwarz inequality and (28), we have

$$\begin{aligned} \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}} &\leq \langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}}^{1/2} \left\langle M^{-1}\widehat{S}\mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}^{1/2} \\ &\leq C \frac{1}{\beta} \left(1 + \log \frac{H}{h}\right) \langle \mathbf{u}, \mathbf{u} \rangle_{\widehat{\mathrm{S}}}^{1/2} \left\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u} \right\rangle_{\widehat{\mathrm{S}}}^{1/2}. \end{aligned}$$

This gives $\langle \mathbf{u}, M^{-1}\widehat{S}\mathbf{u}\rangle_{\widehat{\mathbb{S}}} \leq C\frac{1}{\beta^2}\left(1+\log(H/h)\right)^2 \langle \mathbf{u},\mathbf{u}\rangle_{\widehat{\mathbb{S}}}$ and the upper bound of the eigenvalues. ∎

The preconditioned conjugate gradient iteration for $M^{-1}\widehat{S}$ is given in Algorithm 1. It is a preconditioned conjugate gradient iteration restricted to the quotient space $(\widehat{\mathbf{W}}_{\Gamma,B}\times Q_0)/\widehat{Y}$. Its convergence will be established in Theorem 6.10, and its convergence rate is determined by the eigenvalue bounds of $M^{-1}\widehat{S}$ given in Theorem 6.7.

Algorithm 1. (Preconditioned Conjugate Gradient Algorithm For Solving (7))

1. *Initialization:* $u^0=0$, $r^0=b=\begin{bmatrix}\mathbf{g}_\Gamma\\ 0\end{bmatrix}\left(=\begin{bmatrix}\widehat{S}_\Gamma\mathbf{u}_\Gamma+\widehat{B}_{0\Gamma}^T p_0\\ 0\end{bmatrix}\right)$, $k=1$.
2. *while* $\langle r^{k-1}, r^{k-1}\rangle_{M^{-1}} \geq$ *tolerance*

$$\begin{aligned} z^{k-1} &= M^{-1}r^{k-1}\\ \beta^k &= \langle z^{k-1}, r^{k-1}\rangle \,/\, \langle z^{k-2}, r^{k-2}\rangle \;; \quad [\ \beta^1=0\ ]\\ d^k &= z^{k-1}+\beta^k d^{k-1}\;; \quad [\ d^1=z^0\ ]\\ \alpha^k &= \langle z^{k-1}, r^{k-1}\rangle \,/\, \langle d^k, d^k\rangle_{\widehat{S}}\\ u^k &= u^{k-1}+\alpha^k d^k\\ r^k &= r^{k-1}-\alpha^k \widehat{S}d^k\\ k &= k+1 \end{aligned}$$

3. $u = u^{k-1}+M^{-1}r^{k-1}$

The following lemma will be used in the proof of Theorem 6.10 to show that the denominators in Algorithm 1 cannot vanish in the iteration.

Lemma 6.8. *Let Assumption* 1 *hold. Then any vector of the form*

$$\mathbf{f}=\widehat{S}\mathbf{w}=\begin{bmatrix}\widehat{S}_\Gamma\mathbf{w}_\Gamma+\widehat{B}_{0\Gamma}^T q_0\\ 0\end{bmatrix}, \quad \textit{where} \quad \mathbf{w}=\begin{bmatrix}\mathbf{w}_\Gamma\\ q_0\end{bmatrix}\in\widehat{\mathbf{W}}_{\Gamma,B}\times Q_0, \tag{29}$$

satisfies $\langle \mathbf{f},\mathbf{f}\rangle_{M^{-1}}=0$ *if and only if* $\mathbf{w}_\Gamma=\mathbf{0}$. *If* $\mathbf{w}_\Gamma\neq\mathbf{0}$ *in* (29), *then the velocity component of* $M^{-1}\mathbf{f}$ *is also nonzero.*

Proof. We know from Lemma 6.3 that if $\mathbf{w}_\Gamma=\mathbf{0}$ in (29), then

$$\mathbf{f}^T M^{-1}\mathbf{f}=\begin{bmatrix}\widehat{B}_{0\Gamma}^T q_0\\ 0\end{bmatrix}^T M^{-1}\begin{bmatrix}\widehat{B}_{0\Gamma}^T q_0\\ 0\end{bmatrix}=\begin{bmatrix}\widehat{B}_{0\Gamma}^T q_0\\ 0\end{bmatrix}^T\begin{bmatrix}\mathbf{0}\\ q_0\end{bmatrix}=0.$$

On the other hand, if $\langle \mathbf{f},\mathbf{f}\rangle_{M^{-1}}=0$, then $\langle \mathbf{w}, M^{-1}\widehat{S}\mathbf{w}\rangle_{\widehat{\mathbb{S}}}=0$. By Lemma 6.5 $\langle \mathbf{w},\mathbf{w}\rangle_{\widehat{\mathbb{S}}}=0$, and by (19) $\mathbf{w}_\Gamma$ must vanish.

To prove the second part of this lemma, we know, from Lemma 6.3, that M^{-1} is a one-to-one map from the space of vectors $\mathbf{f}$ with $\mathbf{w}_\Gamma=\mathbf{0}$ in (29) to the space $\{\mathbf{0}\}\times Q_0$. Since for any nonzero $\mathbf{w}_\Gamma\in\widehat{\mathbf{W}}_{\Gamma,B}$, $\widehat{S}_\Gamma\mathbf{w}_\Gamma\notin\mathbf{range}\,(\widehat{B}_{0\Gamma}^T)$ (otherwise the system matrix of (7) would be singular), we know that for any vector $\mathbf{f}$ with $\mathbf{w}_\Gamma\neq\mathbf{0}$ in (29) the image $M^{-1}\mathbf{f}$ cannot be in the space $\{\mathbf{0}\}\times Q_0$; therefore, the velocity component of $M^{-1}\mathbf{f}$ must be nonzero. ∎

The following lemma can be found in [14, equation (10.3.4)].

Lemma 6.9. *The residuals* r^k *in Algorithm* 1 *are* M^{-1}*-orthogonal; i.e., for any* $i\neq j$, $\langle r^i, r^j\rangle_{M^{-1}}=0$.

THEOREM 6.10. *Let Assumption* 1 *hold. The preconditioned conjugate gradient algorithm then converges and at convergence* u *is the solution of* (7).

Proof. We will show that, for any $k = 1, 2, \ldots$, if $\left\langle r^{k-1}, r^{k-1}\right\rangle_{M^{-1}} \neq 0$, then the denominator $\left\langle d^k, d^k\right\rangle_{\widehat{\mathrm{S}}}$ cannot vanish in the iteration. (It is easy to see that the other denominator $\left\langle z^{k-2}, r^{k-2}\right\rangle = \left\langle r^{k-2}, r^{k-2}\right\rangle_{M^{-1}} > 0$.) We see from Algorithm 1 that

$$\begin{aligned} span\{r^0, r^1, \ldots, r^{k-1}\} &= span\{r^0,\ \widehat{S}M^{-1}r^0, \ldots,\ (\widehat{S}M^{-1})^{k-1}r^0\}, \\ span\{d^1, d^2, \ldots, d^k\} &= span\{M^{-1}r^0,\ M^{-1}r^1, \ldots,\ M^{-1}r^{k-1}\}. \end{aligned}$$

The vectors r^i, for $i = 0, 1, \ldots, k-1$, and d^k can be written as

$$r^i = \begin{bmatrix} \widehat{S}_\Gamma \mathbf{w}_\Gamma^i + \widehat{B}_{0\Gamma}^T q_0^i \\ 0 \end{bmatrix},\ i = 0, 1, \ldots, k-1, \tag{30}$$

$$d^k = M^{-1} \begin{bmatrix} \widehat{S}_\Gamma(\mathbf{w}_\Gamma^{k-1} + \sum_{i=0}^{k-2} s_i \mathbf{w}_\Gamma^i) + \widehat{B}_{0\Gamma}^T(q_0^{k-1} + \sum_{i=0}^{k-2} t_i q_0^i) \\ 0 \end{bmatrix}, \tag{31}$$

where $\mathbf{w}_\Gamma^i \in \widehat{W}_{\Gamma,B}$, $q_0^i \in Q_0$, for $i = 0, 1, \ldots, k-1$, and s_i and t_i are scalar factors.

For $k = 1$, and if $\left\langle r^0, r^0\right\rangle_{M^{-1}} > 0$, then by Lemma 6.8, $\mathbf{w}_\Gamma^0 \neq \mathbf{0}$ in the formula for r^0 in (30). We then see from Lemma 6.8 that the velocity component of $d^1 = M^{-1}r^0$ must be nonzero; therefore, by (19), $\left\langle d^1, d^1\right\rangle_{\widehat{\mathrm{S}}} > 0$.

For any $k > 1$, if $\left\langle r^{k-1}, r^{k-1}\right\rangle_{M^{-1}} > 0$, then, by Lemma 6.8, $\mathbf{w}_\Gamma^{k-1} \neq \mathbf{0}$ in the formula for r^{k-1} in (30). We will show that $\mathbf{w}_\Gamma^{k-1}$ cannot be written as a linear combination of $\mathbf{w}_\Gamma^i$, $i = 0, \ldots, k-2$. Assuming $\mathbf{w}_\Gamma^{k-1} = \sum_{i=0}^{k-2} c_i \mathbf{w}_\Gamma^i$, we would have

$$r^{k-1} = \begin{bmatrix} \widehat{S}_\Gamma \mathbf{w}_\Gamma^{k-1} + \widehat{B}_{0\Gamma}^T q_0^{k-1} \\ 0 \end{bmatrix} = \sum_{i=0}^{k-2} c_i r^i + \begin{bmatrix} \widehat{B}_{0\Gamma}^T(q_0^{k-1} - \sum_{i=0}^{k-2} c_i q_0^i) \\ 0 \end{bmatrix}.$$

Then, from Lemmas 6.9 and 6.3, we would have

$$\begin{aligned} 0 = \left\langle \sum_{i=0}^{k-2} c_i r^i, r^{k-1} \right\rangle_{M^{-1}} &= \left(\sum_{i=0}^{k-2} c_i r^{i^T} \right) M^{-1} \left\{ \sum_{i=0}^{k-2} c_i r^i + \begin{bmatrix} \widehat{B}_{0\Gamma}^T(q_0^{k-1} - \sum_{i=0}^{k-2} c_i q_0^i) \\ 0 \end{bmatrix} \right\} \\ &= \left\langle \sum_{i=0}^{k-2} c_i r^i, \sum_{i=0}^{k-2} c_i r^i \right\rangle_{M^{-1}} + \left(\sum_{i=0}^{k-2} c_i r^{i^T} \right) \begin{bmatrix} 0 \\ q_0^{k-1} - \sum_{i=0}^{k-2} c_i q_0^i \end{bmatrix} \\ &= \left\langle \sum_{i=0}^{k-2} c_i r^i, \sum_{i=0}^{k-2} c_i r^i \right\rangle_{M^{-1}} = \sum_{i=0}^{k-2} |c_i|^2 \left\langle r^i, r^i \right\rangle_{M^{-1}} > 0, \end{aligned}$$

which is a contradiction. Therefore, $\mathbf{w}_\Gamma^{k-1} + \sum_{i=0}^{k-2} s_i \mathbf{w}_\Gamma^i$ cannot vanish and by Lemma 6.8, $\left\langle d^k, d^k\right\rangle_{\widehat{\mathrm{S}}} > 0$.

At full convergence (*tolerance* $= 0$) of Algorithm 1, we have $\left\langle r^{k-1}, r^{k-1}\right\rangle_{M^{-1}} = 0$, and

$$b - \widehat{S}u = b - \widehat{S}\left(u^{k-1} + M^{-1}r^{k-1}\right) = r^{k-1} - \widehat{S}M^{-1}r^{k-1}.$$

Since $\left\langle r^{k-1}, r^{k-1}\right\rangle_{M^{-1}} = 0$, we know from Lemma 6.8 that r^{k-1} is of the form

$$r^{k-1} = \begin{bmatrix} \widehat{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix}.$$

Therefore, from Lemma 6.3,

$$\widehat{S}M^{-1}r^{k-1} = \widehat{S}\begin{bmatrix} 0 \\ q_0 \end{bmatrix} = \begin{bmatrix} \widehat{B}_{0\Gamma}^T q_0 \\ 0 \end{bmatrix} = r^{k-1},$$

and hence $b - \widehat{S}u = 0$; i.e., u is the solution of (7). ∎

7. Satisfying the assumptions: Choosing primal constraints. Assumptions 1 and 2 can be satisfied with appropriate choices of the primal continuity constraints on the interface velocity variables. We first describe a recipe for two-dimensional problems, and then one for the more complicated three-dimensional case.

For two-dimensional problems, it is natural to make all subdomain vertices primal, i.e., make both components of the velocity continuous at those nodes. The vertex constraints by themselves are able to control the subdomain rigid body modes. It is straightforward to modify the concept of *fully primal* faces, introduced for three-dimensional elasticity in [23] and outlined later in this section, and to prove that any edge, with two primal variables at each of its end points, is fully primal. It is then easy to prove that Assumption 2 holds.

In order to satisfy Assumption 1, additional constraints are necessary. For each interface edge Γ^{ij}, which is shared by a pair of subdomains Ω_i and Ω_j, we enforce

$$\int_{\Gamma^{ij}} \mathbf{w}_\Gamma^{(i)} \cdot \mathbf{n}_{ij} = \int_{\Gamma^{ij}} \mathbf{w}_\Gamma^{(j)} \cdot \mathbf{n}_{ij}, \tag{32}$$

with a fixed selection of the normal $\mathbf{n}_{ij}$ to Γ^{ij}. We implement a change of basis choosing an edge basis element vector with components $\int_{\Gamma^{ij}} \varphi_k^{\Gamma^{ij}} \cdot \mathbf{n}_{ij}$ at the interior nodes k of Γ^{ij} which vanishes at the end points of the edge. Here $\varphi_k^{\Gamma^{ij}}$ is the nodal finite element velocity basis function of the node k. All the other, complementary velocity element vectors of this edge are chosen to be orthogonal to the special primal edge basis function just introduced and the integrals of their normal components will therefore vanish; i.e., $\int_{\Gamma^{ij}} \mathbf{w}_\Delta^{(i)} \cdot \mathbf{n}_{ij} = 0$ for any $\mathbf{w}_\Delta^{(i)} \in \mathbf{W}_\Delta^{(i)}$. Since $\partial\Omega_i$ is a union of edges, Assumption 1 is satisfied. For details on the implementation of the change of basis, see [29], [23, section 4], and [17].

For three-dimensional problems, the interface Γ is composed of subdomain faces, denoted by $\mathcal{F}^l$, shared by two subdomains, edges $\mathcal{E}^k$, which make up parts of the boundaries of faces, and are often shared by more than two subdomains, and vertices which are the end points of the edges. We will use a partition of unity to separate contributions from the faces, edges, and vertices; cf. [37, section 4.6]. For each face $\mathcal{F}^l$, we denote by $\theta_{\mathcal{F}^l}$ the finite element cut-off function which equals 1 at the interior nodes of the face $\mathcal{F}^l$ and vanishes at all other nodes on the interface. For each edge $\mathcal{E}^k$, we denote by $\theta_{\mathcal{E}^k}$ the finite element cut-off function which equals 1 at all the interior nodes of $\mathcal{E}^k$ and vanishes at all other nodes. We denote the set of faces which share the edge $\mathcal{E}^k$ by $\mathcal{M}_{\mathcal{E}^k}$. We also select a normal $\mathbf{n}_l$ for each face $\mathcal{F}^l$.

Our recipe for satisfying Assumption 1 in three dimensions is similar to that of the two-dimensional case. We will consider each face of the interface separately and also make all velocity components at all subdomain vertices primal. We can then use a partition of unity based only on face and edge functions and find that for any dual velocity element $\mathbf{w}_\Delta^{(i)} \in \mathbf{W}_\Delta^{(i)}$,

$$\int_{\mathcal{F}^l} \mathbf{w}_\Delta^{(i)} \cdot \mathbf{n}_l = \int_{\mathcal{F}^l} I^h\left(\theta_{\mathcal{F}^l}\mathbf{w}_\Delta^{(i)}\right) \cdot \mathbf{n}_l + \sum_{\mathcal{E}^k \subset \partial\mathcal{F}^l} \int_{\mathcal{F}^l} I^h\left(\theta_{\mathcal{E}^k}\mathbf{w}_\Delta^{(i)}\right) \cdot \mathbf{n}_l. \tag{33}$$

Here I^h is the interpolation operator into the velocity finite element space.

On the face $\mathcal{F}^l$, shared by a pair of subdomains Ω_i and Ω_j, we enforce

$$\int_{\mathcal{F}^l} I^h\left(\theta_{\mathcal{F}^l}\mathbf{w}_\Gamma^{(i)}\right)\cdot\mathbf{n}_l = \int_{\mathcal{F}^l} I^h\left(\theta_{\mathcal{F}^l}\mathbf{w}_\Gamma^{(j)}\right)\cdot\mathbf{n}_l, \tag{34}$$

in the same way as for the edge constraints in two dimensions. This face average corresponds to a face average basis element, which is made primal, and the dual interface velocities $\mathbf{w}_\Delta^{(i)}$ always satisfy $\int_{\mathcal{F}^l} I^h(\theta_{\mathcal{F}^l}\mathbf{w}_\Delta^{(i)})\cdot\mathbf{n}_l = 0$.

The discussion of the edge terms is complicated by the fact that the nodal basis functions associated with an edge node will differ from zero in narrow strips on the boundaries of all subdomains that share that edge. For an edge $\mathcal{E}^k$ and on each face $\mathcal{F}^m \in \mathcal{M}_{\mathcal{E}^k}$ which shares this edge, and for all pairs of subdomains Ω_i and Ω_j which share this edge, we will enforce

$$\int_{\mathcal{F}^m} I^h\left(\theta_{\mathcal{E}^k}\mathbf{w}_\Gamma^{(i)}\right)\cdot\mathbf{n}_m = \int_{\mathcal{F}^m} I^h\left(\theta_{\mathcal{E}^k}\mathbf{w}_\Gamma^{(j)}\right)\cdot\mathbf{n}_m. \tag{35}$$

The number of faces in the set $\mathcal{M}_{\mathcal{E}^k}$ is denoted by m_k; this many primal degrees of freedom are required to satisfy the constraints (35). The corresponding primal basis element vectors are determined by the integrals of the normal components of the edge nodal finite element basis functions over the relevant faces. As in two dimensions, the dual velocity basis element vectors are made orthogonal to all the primal basis vectors. Therefore, the dual interface velocities $\mathbf{w}_\Delta^{(i)}$ will always satisfy $\int_{\mathcal{F}^m} I^h\big(\theta_{\mathcal{E}^k}\mathbf{w}_\Delta^{(i)}\big)\cdot\mathbf{n}_m = 0$. By enforcing the constraints (34) and (35), the integral (33) is always zero and Assumption 1 is satisfied.

It can easily happen that the m_k primal basis vectors, for the edge $\mathcal{E}^k$, are linearly dependent. This happens, e.g., in the case when the subdomains are cubes and a uniform mesh is used. In general, we must make sure that the primal basis functions maintain linear independence for each edge separately. We can use a singular value decomposition, in a preprocessing step of the algorithm, to single out only those that are numerically linearly independent and should be retained. This device for eliminating linearly dependent coarse level primal constraints has previously been applied for both FETI-DPH (a variant for Helmholtz's equation) and BDDC algorithms; see [10, 5, 6].

Remark 1. A different BDDC algorithm was introduced in [5, 6] by Dohrmann for solving nearly incompressible elasticity problems. Zero divergence constraints were used for the substructure corrections to keep the volume change of each substructure small for nearly incompressible materials. For two-dimensional problems, our constraints (32) are the same as those in [5, 6]; the same type of constraints have also been used previously in FETI-DP algorithms for Stokes problems; cf. [27]. For three-dimensional problems, our vertex and face constraints are the same as Dohrmann's. But for each edge $\mathcal{E}^k$, Dohrmann requires, on each subdomain Ω_i which shares this edge, that the integrals

$$\sum_{\mathcal{F}^l\subset\partial\Omega_i}\int_{\mathcal{F}^l} I^h\left(\theta_{\mathcal{E}^k}\mathbf{w}_\Gamma^{(m)}\right)\cdot\mathbf{n}_l \tag{36}$$

be the same for all $m \in \mathcal{N}_{\mathcal{E}^k}$; here $\mathcal{N}_{\mathcal{E}^k}$ is the set of indices of the subdomains which have the edge $\mathcal{E}^k$ in common. While either set of edge constraints, (35) or

(36), together with the face and vertex constraints will satisfy Assumption 1, we have adopted the form (35) to facilitate the analysis.

We also have to make sure that we have a set of constraints which guarantees a stable $E_{D,\Gamma}$ operator, as in Assumption 2, for three-dimensional problems. The continuity constraints (34) and (35), together with the vertex constraints, developed for Assumption 1, are not always sufficient for Assumption 2 to hold. We will show that some additional edge tangential continuity constraints are sometimes needed. Such a constraint is introduced on an edge $\mathcal{E}^k$, by requiring that $\int_{\mathcal{E}^k} I^h(\theta_{\mathcal{E}^k}\mathbf{w}_\Gamma^{(i)})\cdot\mathbf{t}_{\mathcal{E}^k}$ takes on a common value. Here $\mathbf{t}_{\mathcal{E}^k}$ is the unit vector tangent to $\mathcal{E}^k$. By selecting a primal degree of freedom corresponding to this tangential edge integral, we can make the resulting dual interface variables satisfy $\int_{\mathcal{E}^k} I^h(\theta_{\mathcal{E}^k}\mathbf{w}_\Delta^{(i)})\cdot\mathbf{t}_{\mathcal{E}^k}=0$. We note that only one extra primal variable will be introduced for each such edge.

We recall that the space of rigid body modes on each subdomain Ω_i is spanned by the three translations

$$\mathbf{r}_1 := \begin{bmatrix} 1\\0\\0\end{bmatrix},\quad \mathbf{r}_2 := \begin{bmatrix} 0\\1\\0\end{bmatrix},\quad \mathbf{r}_3 := \begin{bmatrix} 0\\0\\1\end{bmatrix}, \tag{37}$$

and the three rotations

$$\mathbf{r}_4 := \frac{1}{H_i}\begin{bmatrix} x_2-\hat{x}_2\\ -x_1+\hat{x}_1\\ 0\end{bmatrix},\quad \mathbf{r}_5 := \frac{1}{H_i}\begin{bmatrix} -x_3+\hat{x}_3\\ 0\\ x_1-\hat{x}_1\end{bmatrix},\quad \mathbf{r}_6 := \frac{1}{H_i}\begin{bmatrix} 0\\ x_3-\hat{x}_3\\ -x_2+\hat{x}_2\end{bmatrix}. \tag{38}$$

Here $\hat{\mathbf{x}}\in\Omega_i$ and H_i denotes the diameter of Ω_i. (The shift of the origin makes the basis for the space of rigid body modes well conditioned, and the scaling and shift make the $L^2(\Omega_i)$-norms of these six functions scale similarly with H_i.) For each subdomain face $\mathcal{F}^l$, we represent the primal continuity constraints enforced on its edges (excluding the vertex constraints) in terms of a set of linear functionals $f_m^l(\cdot)$, $m=1,2,\ldots,M_l$. All the $f_m^l(\cdot)$ vanish when applied to the dual velocity component. We will use the idea of *fully primal* faces, which has been developed in [23] for compressible elasticity; see Definition 2 below. We will use our functionals, which define our edge constraints, to create a basis, $\{g_k\}_1^6$, which, when restricted to the six-dimensional space of rigid body modes, is a dual basis. Each g_k will be a linear combination of the constraint functionals for the face in question. We will also require certain bounds for the linear functionals; these bounds enter into the bound for the condition number of our algorithm. We note that the quality of these bounds is better than what can be accomplished with vertex constraints. It is in fact known that the exclusive use of point constraints leads to less satisfactory performance; see, e.g., [9, 34, 19].

DEFINITION 2. *A face $\mathcal{F}^l$ of subdomain Ω_i is fully primal if a set of six continuity constraints $g_k(\cdot), k=1,2,\ldots,6$, are enforced on the edges of $\mathcal{F}^l$ and satisfy*

$$|g_k(\mathbf{w}^{(i)})|^2 \le CH^{-1}\left(1+\log\left(\frac{H}{h}\right)\right)\left\{|\mathbf{w}^{(i)}|^2_{H^{1/2}(\mathcal{F}^{ij})}+\frac{1}{H}\|\mathbf{w}^{(i)}\|^2_{L^2(\mathcal{F}^{ij})}\right\}, \tag{39}$$

$$g_k(\mathbf{r}_j)=\delta_{kj}\quad \forall k,j=1,\ldots,6. \tag{40}$$

Essentially the same proof as for [23, Lemma 8.4] can be used to prove the following lemma; establishing the inequalities (39) is a key part in any such proof.

LEMMA 7.1. *Assumption* 2 *is satisfied for the three-dimensional problems if all the faces* $\mathcal{F}^l$ *of the interface* Γ *are fully primal and all the subdomain vertices are primal.*

We will now consider a set of sufficient conditions for the conditions of Definition 2 to hold. Many details can again be found in [23], e.g., the bounds for the functionals associated with the tangential components of the velocity. We only indicate how to bound the functionals associated with the constraints on the normal components and the edges; the constraint given by (34) will play no role in establishing that faces are fully primal. We have the following result.

LEMMA 7.2. *Let the* $f_m^l(\cdot)$, $m = 1, 2, \ldots, M_l$, *be given by the continuity constraints* (35) *and a tangential edge constraint for each of the edges of a face* $\mathcal{F}^l$. *Then* $\mathcal{F}^l$ *is fully primal.*

Proof. Using a normalization as in [23], the constraint (35) associated with an edge $\mathcal{E}^k$ and a normal component is represented by a functional

$$f_m^l(\mathbf{w}^{(i)}) = \frac{\int_{\mathcal{E}^{k+}} I^h(\theta_{\mathcal{E}^k}\mathbf{w}^{(i)}) \cdot \mathbf{n}_l \; ds}{\int_{\mathcal{E}^{k+}} 1 ds},$$

where $\mathbf{w}^{(i)} \in \mathbf{W}^{(i)}$, $\mathbf{n}_l$ is a unit normal to the face $\mathcal{F}^l$, and $\mathcal{E}^{k+}$ represents the strip of elements next to the edge $\mathcal{E}^k$ on $\mathcal{F}^l$. By using the Cauchy–Schwarz inequality, we have $|f_m^l(\mathbf{w}^{(i)})|^2 \leq CH^{-1}\|\theta_{\mathcal{E}^k}\mathbf{w}^{(i)}\|^2_{L^2(\mathcal{E}^{ij})} \leq CH^{-1}\|\mathbf{w}^{(i)}\|^2_{L^2(\mathcal{E}^{ij})}$. Using [23, Lemma 7.4] or [8, Lemma 3.3], we have

$$|f_m^l(\mathbf{w}^{(i)})|^2 \leq CH^{-1}\left(1 + \log\left(\frac{H}{h}\right)\right)\left\{|\mathbf{w}^{(i)}|^2_{H^{1/2}(\mathcal{F}^{ij})} + \frac{1}{H}\|\mathbf{w}^{(i)}\|^2_{L^2(\mathcal{F}^{ij})}\right\}. \tag{41}$$

What is now left is to show that among the given edge normal constraints (35) and tangential edge constraints for the edges of a face $\mathcal{F}^l$, we can choose six functionals, denoted by $f_m^l(\cdot)$, $m = 1, 2, \ldots, 6$, such that if $f_m^l(\mathbf{r}) = 0, m = 1, 2, \ldots, 6$, for a rigid body mode $\mathbf{r}$, then $\mathbf{r}$ must vanish. Once this has been established, the g_k of Definition 2 can be chosen as linear combinations of these f_m^l.

Let us, without loss of generality, consider a face $\mathcal{F}^l$ which is part of the $x_1 - x_2$ plane and let $\hat{\mathbf{x}} = 0$. Since we have weighted edge average constraints in (35) for the third component over all, i.e., at least three edges of the face, we can conclude that the third component of the rigid body mode $\mathbf{r}$ must vanish at three or more points which are not colinear; we recall that we have assumed that all faces are convex. We denote the three corresponding edge normal continuity functionals by $f_m^l(\cdot)$, $m = 3, 5, 6$. Since this third component of $\mathbf{r}$ is a linear combination of the third component of the three basis elements $\mathbf{r}_3$, $\mathbf{r}_5$, and $\mathbf{r}_6$, the rigid body mode $\mathbf{r}$ cannot have any component involving these three basis elements. The remaining part is a linear combination of $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_4$, i.e., effectively a rigid body mode in two dimensions. It has the form of a first order Nédélec element on the face,

$$\mathbf{r} = \begin{bmatrix} a_1 + bx_2 \\ a_2 - bx_1 \\ 0 \end{bmatrix}, \tag{42}$$

where a_1, a_2, and b are the three remaining degrees of freedom of the rigid body mode for this two-dimensional surface.

We will now show that the tangential edge constraints will make it possible to conclude that $a_1 = a_2 = b = 0$. It is known, and easy to show, that the first order

Nédélec elements have a constant tangential component on each edge. It is then easy to see that any three edge tangential constraints will make it possible to conclude that the remaining rigid body components must vanish; these three selected edge tangential continuity functionals, $f_m^l(\cdot)$, $m = 1, 2, 4$, complete the set of functionals. ∎

We note that one can also check numerically if the normal face and edge constraints (34) and (35) will, in themselves, make a face fully primal. What matters is if a full dual basis can be constructed when these functionals are restricted to the space of rigid body motions.

We end this section by discussing the need for edge tangential constraints; this discussion concerns only the final three constraints. By a relatively simple computation, we can show that only two tangential edge constraints on two adjacent edges are needed for each face $\mathcal{F}^l$ to make it fully primal; the third one can be replaced by an already existing edge normal continuity constraint. One can also show that these two edge tangential continuity constraints are not always necessary and that one will suffice at times. To understand why tangential constraints appear to be required in some cases, we consider a face with three edges only and with constant weights. Then, by the divergence theorem and the fact that the rigid body modes are divergence free, we have linear dependence since the integral of the normal component over one edge equals the negative of the sum over the integrals over the other two. A simple computation reveals that the rank is also two for a rectangular face. In such a case, at least one tangential continuity constraint will be needed to make such a face fully primal.

8. Connections with the FETI-DP algorithms. In the FETI-DP algorithms developed in [27] for incompressible Stokes equations, the subdomain problems are also assembled only at the coarse level, primal velocity degrees of freedom, which are shared by neighboring subdomains. Lagrange multipliers are then introduced on the interface to enforce the continuity of the dual velocity variables, by requiring that $B_\Delta \mathbf{u}_\Delta = \sum_{i=1}^N B_\Delta^{(i)} \mathbf{u}_\Delta^{(i)} = 0$. Here, the subdomain matrices $B_\Delta^{(i)}$ have elements chosen from the set $\{0, 1, -1\}$. The original problem is then reduced to a linear system for the Lagrange multipliers by eliminating the other variables; cf. [27]. The FETI-DP operator for the Lagrange multipliers is $B_\Delta \widetilde{S}_\Delta^{-1} B_\Delta^T$, where the operator $\widetilde{S}_\Delta$ is defined by $\widetilde{S}_\Delta^{-1} = R_\Delta \widetilde{S}^{-1} R_\Delta^T$ and R_Δ is the restriction map from $\widetilde{\mathbf{W}}_\Gamma \times Q_0$ to $\mathbf{W}_\Delta$.

The preconditioner used in [27] for the FETI-DP algorithm is $B_{D,\Delta} S_\Delta B_{D,\Delta}^T$, where $B_{D,\Delta}$ is constructed from the subdomain operators $B_{D,\Delta}^{(i)}$ in the same way as B_Δ from the $B_\Delta^{(i)}$; a constant scaling factor was used for $B_{D,\Delta}^{(i)}$ for two-dimensional problems. In general, $B_{D,\Delta}^{(i)}$ is defined as follows: each nonzero element of $B_\Delta^{(i)}$ corresponds to a Lagrange multiplier connecting the subdomain Ω_i to a neighboring subdomain Ω_j at a point $x \in \partial\Omega_{i,h} \cap \partial\Omega_{j,h}$. Multiplying each such element with the positive scaling factor $\delta_j^\dagger(x)$ gives us $B_{D,\Delta}^{(i)}$. S_Δ is the direct sum of subdomain Schur operators $S_\Delta^{(i)}$, which are defined on the dual subdomain velocity space $\mathbf{W}_\Delta^{(i)}$ as the $S_\Gamma^{(i)}$ in (8), except that the operator is restricted to the dual interface velocity variables; S_Δ can be written as the restriction of the operator $\widetilde{S}$ to the space $\mathbf{W}_\Delta$, i.e., $S_\Delta = R_\Delta \widetilde{S} R_\Delta^T$.

Therefore, the preconditioned FETI-DP operator can be written as

$$B_{D,\Delta} R_\Delta \widetilde{S} R_\Delta^T B_{D,\Delta}^T B_\Delta R_\Delta \widetilde{S}^{-1} R_\Delta^T B_\Delta^T. \tag{43}$$

Since the diagonal blocks corresponding to the dual interface velocity part in $\mathbf{W}_\Delta$ of the matrices $\widetilde{S}$ and $\widetilde{S}^{-1}$ are positive definite, both $R_\Delta \widetilde{S} R_\Delta^T$ and $R_\Delta \widetilde{S}^{-1} R_\Delta^T$ are positive definite. When nonredundant Lagrange multipliers are used, the matrices B_Δ^T

and $B_{D,\Delta}^T$ are of full rank and the FETI-DP operator (43) is therefore a product of two positive definite matrices; cf. [37, section 6.4]. If redundant Lagrange multipliers are used, as in [22, 37], then B_Δ^T will not be of full rank. But this does not matter; the Lagrange multiplier is always restricted to $\mathbf{range}\,(B_\Delta)$, which is orthogonal to the null space of B_Δ^T (cf. [22]).

We now introduce the operator $P_D = R_\Delta^T B_{D,\Delta}^T B_\Delta R_\Delta$, which maps the space $\widetilde{\mathbf{W}}_\Gamma \times Q_0$ into itself. It computes the jump across the subdomain interface of the dual interface velocity component, and maps any element in the primal space $\widehat{\mathbf{W}}_\Pi \times Q_0$ to zero; cf. [29]. It can then be verified that E_D and P_D are complementary projectors with $E_D + P_D = I$ and $E_D P_D = P_D E_D = 0$; cf. [29, Lemma 1].

Since, for any two matrices Z and T, the nonzero eigenvalues of the two products ZT and TZ—assuming that they both exist—are the same, we know that the preconditioned FETI-DP operator (43) has the same nonzero eigenvalues as the operator $P_D^T \widetilde{S} P_D \widetilde{S}^{-1}$, where we have moved the last factor $R_\Delta^T B_\Delta^T$ of (43) to the front. The preconditioned BDDC operator $M^{-1}\widehat{S}$, which is $\widetilde{R}_D^T \widetilde{S}^{-1} \widetilde{R}_D \widetilde{R}^T \widetilde{S} \widetilde{R}$, has the same nonzero eigenvalues as $E_D \widetilde{S}^{-1} E_D^T \widetilde{S}$, where we have moved the last factor $\widetilde{R}$ to the front. We can then prove, just as in the elliptic case (see [29]) that $P_D^T \widetilde{S} P_D \widetilde{S}^{-1}$ and $E_D \widetilde{S}^{-1} E_D^T \widetilde{S}$ have the same nonzero eigenvalues with the possible exception of 1. We obtain the following theorem.

THEOREM 8.1. *Let Assumption* 1 *hold. The preconditioned FETI-DP and BDDC operators, given by* (43) *and* $M^{-1}\widehat{S}$*, respectively, have the same nonzero eigenvalues with the possible exception of* 1*.*

This is the same result as for the positive definite elliptic problems; cf. [31, 11, 29, 3].

9. Numerical experiments. We solve a lid-driven-cavity problem on the domain $\Omega = [0,1] \times [0,1]$ with the Dirichlet boundary condition, where the velocity is $(1,0)$ on the upper side and vanishes on the other three sides. We use a uniform mesh, as in Figure 1. The mixed finite elements are also indicated in Figure 1; the velocity is continuous and linear in each element, and the pressure is constant on macroelements which are unions of four triangles. The inf-sup stability of these mixed finite elements can easily be proved by using the macroelement technique developed in [36].

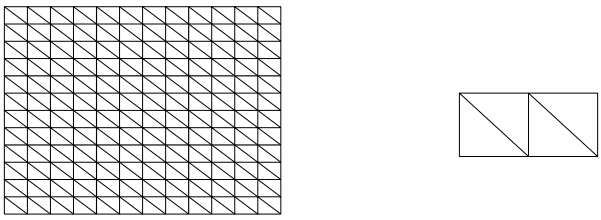


FIG. 1. *The mesh and the mixed finite elements.*

Both the BDDC and FETI-DP algorithms have been tested. The preconditioned conjugate gradient method is used and the iteration is halted when the L^2-norm of the residual has been reduced by a factor 10^{-6}. In our experiments, we have used three different sets of primal constraints. The first two satisfy both Assumptions 1 and 2 and we see that both the BDDC and FETI-DP operators are positive definite and that the results are fully consistent with our theory. Our third choice violates Assumption 1 and the BDDC operator is then no longer positive definite.

TABLE 1

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different numbers of subdomains, for $H/h = 8$ and a primal space spanned by both corner and normal edge basis functions.

Num. of subs	BDDC			FETI-DP		
$n_x \times n_y$	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4×4	1.00	3.14	11	1.00	3.14	11
8×8	1.00	3.88	12	1.00	3.88	12
12×12	1.00	4.02	12	1.00	4.02	13
16×16	1.00	4.06	12	1.00	4.07	13
20×20	1.00	4.08	12	1.00	4.08	13

TABLE 2

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different H/h, for 4×4 subdomains and a primal space spanned by both corner and normal edge basis functions.

	BDDC			FETI-DP		
H/h	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4	1.00	2.17	8	1.00	2.17	9
8	1.00	3.14	11	1.00	3.14	11
16	1.00	4.22	13	1.00	4.22	12
32	1.00	5.42	14	1.00	5.42	14

TABLE 3

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different numbers of subdomains, for $H/h = 8$ and a primal space spanned by both corner and two edge basis functions for each edge.

Num. of subs	BDDC			FETI-DP		
$n_x \times n_y$	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4×4	1.00	2.32	8	1.00	2.32	9
8×8	1.00	2.58	9	1.00	2.58	9
12×12	1.00	2.63	9	1.00	2.63	10
16×16	1.00	2.65	9	1.00	2.65	10
20×20	1.00	2.65	9	1.00	2.65	10

In the first case, the primal velocity space is spanned by the subdomain vertex nodal basis functions for both components and by a constant vector in the direction normal to the edge for each interface edge as in (32). From Tables 1 and 2, we see that the preconditioned BDDC and FETI-DP operators are both positive definite and quite well conditioned as established in Theorems 6.7 and 8.1. We observe that the extreme eigenvalues and the iteration counts of the BDDC and FETI-DP algorithms match very well, and that the condition numbers of both algorithms are independent of the number of subdomains and increase only slowly with the number of elements across each subdomain, all as predicted by the theory. In our experiments, the extreme eigenvalues are estimated by using the tridiagonal Lanczos matrix generated by the preconditioned conjugate gradient method.

In the experiments of Tables 3 and 4, the integrals of both velocity components are required to have common values across each interface edge. The subdomain corner degrees of freedom are also chosen as primal variables as in the first case. Both Assumptions 1 and 2 are again satisfied and we observe similar, slightly faster convergence compared with the first experiments since the primal, coarse level problem has been enlarged.

In Tables 5 and 6, the primal velocity space is spanned only by the corner basis

TABLE 4

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different H/h, for 4×4 subdomains and a primal space spanned by both corner and two edge basis functions for each edge.

	BDDC			FETI-DP		
H/h	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4	1.00	1.66	7	1.00	1.65	7
8	1.00	2.32	8	1.00	2.32	9
16	1.00	3.07	10	1.00	3.07	10
32	1.00	3.93	11	1.00	3.93	12

TABLE 5

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different numbers of subdomains, for $H/h = 8$ and a primal space spanned only by the corner basis functions.

Num. of subs	BDDC			FETI-DP		
$n_x \times n_y$	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4×4			17	0.49	3.61	16
8×8			21	0.37	4.01	21
12×12	N/A	N/A	21	0.33	4.08	23
16×16			21	0.31	4.10	22
20×20			22	0.29	4.10	24

TABLE 6

Spectral bounds and iteration counts for a pair of BDDC and FETI-DP algorithms, with different H/h, for 4×4 subdomains and for a primal space spanned only by the corner basis functions.

	BDDC			FETI-DP		
H/h	λ_{min}	λ_{max}	Iter.	λ_{min}	λ_{max}	Iter.
4			13	0.51	2.34	13
8	N/A	N/A	17	0.49	3.61	16
16			19	0.48	5.13	19
32			21	0.48	6.99	21

functions; Assumption 1, then, does not hold. In this case, the preconditioned BDDC operator is no longer positive definite and the iterates will no longer stay in the benign space of the saddle-point problem. However, the FETI-DP operator (43) is still positive definite. The interface problems of both the BDDC and the FETI-DP algorithms are solved by a preconditioned conjugate gradient method, but the residual norm of the BDDC methods is no longer strictly decreasing. We see that the iteration counts of the BDDC and FETI-DP algorithms still match very well, but that for both algorithms this count will now depend on the number of subdomains as well as on the number of elements across each subdomain. These results are less satisfactory than those of the previous two choices of primal constraints.

REFERENCES

[1] G. K. BATCHELOR, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1967.

[2] J. H. BRAMBLE AND J. E. PASCIAK, *A domain decomposition technique for Stokes problems*, Appl. Numer. Math., 6 (1990), pp. 251–261.

[3] S. C. BRENNER AND L. SUNG, *BDDC and FETI-DP without matrices or vectors*, Comput. Methods Appl. Mech. Engrg., to appear.

[4] C. R. DOHRMANN, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.

[5] C. R. Dohrmann, *Preconditioning of saddle point systems by substructuring and a penalty approach*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, O. B. Widlund and D. E. Keyes, eds., Springer-Verlag, Berlin, 2006, pp. 53–64.

[6] C. R. Dohrmann, *A Substructuring Preconditioner for Nearly Incompressible Elasticity Problems*, Technical report SAND2004-5393, Sandia National Laboratories, Albuquerque, NM, 2004.

[7] C. R. Dohrmann, *An Approximate BDDC Preconditioner*, Technical report SAND2005-5424, Sandia National Laboratories, Albuquerque, NM, 2005.

[8] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[9] C. Farhat, M. Lesoinne, P. Le Tallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method—part* I: *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[10] C. Farhat and J. Li, *An iterative domain decomposition method for the solution of a class of indefinite problems in computational structural dynamics*, Appl. Numer. Math., 54 (2005), pp. 150–166.

[11] Y. Fragakis and M. Papadrakakis, *The mosaic of high performance domain decomposition methods for structural mechanics: Formulation, interrelation and numerical efficiency of primal and dual methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 3799–3830.

[12] P. Goldfeld, *Balancing Neumann-Neumann Preconditioners for Mixed Formulation of Almost-Incompressible Linear Elasticity*, Ph.D. thesis, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2003.

[13] P. Goldfeld, L. F. Pavarino, and O. B. Widlund, *Balancing Neumann-Neumann preconditioners for mixed approximations of heterogeneous problems in linear elasticity*, Numer. Math., 95 (2003), pp. 283–324.

[14] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[15] H. H. Kim, M. Dryja, and O. B. Widlund, *A BDDC Algorithm for Problems with Mortar Discretization*, Technical report TR2005-873, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2005.

[16] A. Klawonn and O. Rheinbach, *Inexact FETI-DP methods*, Internat. J. Numer. Methods Engrg., to appear.

[17] A. Klawonn and O. Rheinbach, *A parallel implementation of dual-primal FETI methods for three dimensional linear elasticity using a transformation of basis*, SIAM J. Sci. Comput., 28 (2006), pp. 1886–1906.

[18] A. Klawonn and O. Rheinbach, *Robust FETI-DP methods for heterogeneous three dimensional elasticity problems*, Comput. Methods Appl. Mech. Engrg., to appear.

[19] A. Klawonn, O. Rheinbach, and O. B. Widlund, *Some computational results for dual-primal FETI methods for three dimensional elliptic problems*, in Domain Decomposition Methods in Science and Engineering, Lect. Notes Comput. Sci. Eng. 40, R. Kornhuber, R. H. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, eds., Springer-Verlag, Berlin, 2004, pp. 361–368.

[20] A. Klawonn, O. Rheinbach, and B. Wohlmuth, *Dual-primal iterative substructuring for almost incompressible elasticity*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, O. B. Widlund and D. E. Keyes, eds., Springer-Verlag, Berlin, 2006, pp. 399–406.

[21] A. Klawonn and O. B. Widlund, *A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity*, SIAM J. Sci. Comput., 22 (2000), pp. 1199–1219.

[22] A. Klawonn and O. B. Widlund, *FETI and Neumann-Neumann iterative substructuring methods: Connections and new results*, Comm. Pure Appl. Math, 54 (2001), pp. 57–90.

[23] A. Klawonn and O. B. Widlund, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.

[24] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.

[25] J. Li, *A Dual-Primal FETI Method for Incompressible Stokes and Linearized Navier-Stokes Equations*, Technical report TR-828, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2002.

[26] J. Li, *Dual-Primal FETI Methods for Stationary Stokes and Navier-Stokes Equations*, Ph.D. thesis, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2002.

[27] J. Li, *A dual-primal FETI method for incompressible Stokes equations*, Numer. Math., 102 (2005), pp. 257–275.
[28] J. Li and O. B. Widlund, *On the use of inexact subdomain solvers for BDDC algorithms*, Comput. Methods Appl. Mech. Engrg., to appear.
[29] J. Li and O. B. Widlund, *FETI–DP, BDDC, and block Cholesky methods*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 250–271.
[30] J. Mandel and C. R. Dohrmann, *Convergence of a balancing domain decomposition by constraints and energy minimization*, Numer. Linear Algebra Appl., 10 (2003), pp. 639–659.
[31] J. Mandel, C. R. Dohrmann, and R. Tezaur, *An algebraic theory for primal and dual substructuring methods by constraints*, Appl. Numer. Math., 54 (2005), pp. 167–193.
[32] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.
[33] L. F. Pavarino and O. B. Widlund, *Balancing Neumann-Neumann methods for incompressible Stokes equations*, Comm. Pure Appl. Math., 55 (2002), pp. 302–335.
[34] K. H. Pierson, *A Family of Domain Decomposition Methods for the Massively Parallel Solution of Computational Mechanics Problems*, Ph.D. thesis, University of Colorado, Boulder, CO, 2000.
[35] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Clarendon Press, Oxford, UK, 1999.
[36] R. Stenberg, *A technique for analysing finite element methods for viscous incompressible flow*, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 935–948.
[37] A. Toselli and O. B. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer-Verlag, New York, 2005.
[38] X. Tu, *Three-level BDDC in two dimensions*, Internat. J. Numer. Methods Engrg., to appear.
[39] X. Tu, *A BDDC algorithm for a mixed formulation of flows in porous media*, Electron. Trans. Numer. Anal., 20 (2005), pp. 164–179.
[40] X. Tu, *A BDDC Algorithm for Flow in Porous Media with a Hybrid Finite Element Discretization*, Technical report TR2005-865, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2005.
[41] X. Tu, *Three-level BDDC in Three Dimensions*, Technical report TR2005-862, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2005.
[42] X. Tu, *BDDC Domain Decomposition Algorithms: Methods with Three Levels and for Flow in Porous Media*, Ph.D. thesis, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, 2006.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2456–2480

CONVERGENCE ANALYSIS OF THE ENERGY AND HELICITY PRESERVING SCHEME FOR AXISYMMETRIC FLOWS*

JIAN-GUO LIU† AND WEI-CHENG WANG‡

Abstract. We give an error estimate for the energy and helicity preserving scheme (EHPS) in second order finite difference setting on axisymmetric incompressible flows with swirling velocity. This is accomplished by a weighted energy estimate, along with careful and nonstandard local truncation error analysis near the geometric singularity and a far field decay estimate for the stream function. A key ingredient in our a priori estimate is the permutation identities associated with the Jacobians, which are also a unique feature that distinguishes EHPS from standard finite difference schemes.

Key words. incompressible viscous flow, Navier–Stokes equation, pole singularity, conservative scheme, Jacobian, permutation identity, geometric singularity

AMS subject classifications. 65M06, 65M12, 65M15, 76D05, 35Q30

DOI. 10.1137/050639314

1. Introduction.

Axisymmetric flow is an important subject in fluid dynamics and has become standard textbook material (e.g., [2]) as a starting point of theoretical study for complicated flow patterns. Although the number of independent spatial variables is reduced by symmetry, some of the essential features and complexities of generic three-dimensional (3D) flows remain. For example, when the swirling velocity is nonzero, there is a vorticity stretching term present. This is widely believed to account for possible singularity formation for Navier–Stokes and Euler flows. For general smooth initial data, it is well known that the solution remains smooth for a short time in Euler [8] and Navier–Stokes flows [9]. A fundamental regularity result concerning the solution of the Navier–Stokes equation (NSE) is given in the pioneering work of Caffarelli, Kohn, and Nirenberg [3]: The 1D Hausdorff measure of the singular set is zero. As a consequence, the only possible singularity for axisymmetric Navier–Stokes flows would be on the axis of rotation. This result has motivated subsequent research activities concerning the regularity of axisymmetric solutions of the NSE. Some regularity and partial regularity results for axisymmetric Euler and Navier–Stokes flows can be found, for example, in [4] and the references therein. To date, the regularity of the Navier–Stokes and Euler flows, whether axisymmetric or not, remains a challenging open problem. For a comprehensive review of the regularity of the NSE, see [10] and the references therein.

Due to the subtle regularity issue, the numerical simulation of axisymmetric flows is also a challenging subject for computational fluid dynamicists. The earliest attempt at a numerical search for potential singularities of axisymmetric flows dates back to the 90s [5, 6]. In a recent work [11], the authors have developed a class of energy and helicity preserving schemes (EHPS) for incompressible Navier–Stokes and MHD

*Received by the editors August 31, 2005; accepted for publication (in revised form) June 13, 2006; published electronically December 1, 2006.
http://www.siam.org/journals/sinum/44-6/63931.html
†Institute for Physical Science and Technology and Department of Mathematics, University of Maryland, College Park, MD 20742 (jliu@math.umd.edu). The research of this author was sponsored in part by NSF grant DMS 05-12176.
‡Department of Mathematics, National Tsing Hua University, HsinChu, Taiwan 300 (wangwc@math.nthu.edu.tw). The research of this author was sponsored in part by NSC of Taiwan grant 92-2115-M-007-022.

equations. There the authors extended the vorticity-stream formulation of axisymmetric flows given in [5] and proposed a generalized vorticity-stream formulation for 3D Navier–Stokes and MHD flows with coordinate symmetry. In the case of axisymmetric flows, the major difference between EHPS and the formulation in [5] is the expression and numerical discretization of the nonlinear terms. It is shown in [11] that all the nonlinear terms in the Navier–Stokes and MHD equation, including convection, vorticity stretching, geometric source, Lorentz force, and electro-motive force, can be written as Jacobians. Associated with the Jacobians is a set of permutation identities which leads naturally to the conservation laws for first and second moments. The primary feature of the EHPS is the numerical realization of these conservation laws. In addition to preserving physically relevant quantities, the discrete form of conservation laws provides numerical advantages as well. In particular, the conservation of energy automatically enforces nonlinear stability of EHPS. For 2D flows, EHPS is equivalent to the energy and enstrophy preserving scheme of Arakawa [1], who first pointed out the importance of discrete conservation laws in long time numerical simulations.

Other than the Jacobian approach, most of the energy conserving finite difference schemes for standard flows (without geometric singularity) are based on discretization of the fluid equation in primitive variables. A well-known trick that dates back to the 70s is to take the average of conservative and nonconservative discretizations of convection term (Piacsek and Williams [16]). In [14], Morinishi et al. further explored and compared various combinations among conservative, nonconservative, and rotation forms of the convection term. More recently in [18], Verstappen and Veldman proposed a discretization for the convection term that resulted in a skew-symmetric difference operator and therefore the conservation of energy could be achieved.

A potential difficulty associated with axisymmetric flows is the appearance of a $\frac{1}{r}$ factor which becomes infinite at the axis of rotation, and therefore sensitive to inconsistent or low order numerical treatment near this "pole singularity." In [11], the authors proposed a second order finite difference scheme and handled the pole singularity by shifting the grids a half-grid length away from the origin. Remarkably, the permutation identities and therefore the energy and helicity identities remain valid in this case. There are alternative numerical treatments proposed in literatures (e.g., [6]) to handle this coordinate singularity. However, rigorous justifications for various pole conditions are yet to be established.

The purpose of this paper is to give a rigorous error estimate of EHPS for axisymmetric flows. To focus on the pole singularity and avoid complication caused by physical boundary conditions, we consider here only the whole space problem with the swirling components of velocity and vorticity decaying fast enough at infinity. The error analysis of numerical methods for NSE with nonslip physical boundary condition has been well studied. We refer the works of Hou and Wetton [7] and Wang and Liu [19] to interested readers. Our proof is based on a weighted energy estimate along with a careful and detailed pointwise local truncation error analysis. A major ingredient in our energy estimate is the permutation identities associated with the Jacobians (4.17). These identities are key to the energy and helicity preserving property of EHPS for general symmetric flows. Here the same identities enable us to obtain a priori estimate even in the presence of the pole singularity; see section 5 for details. To our knowledge, this is the first rigorous convergence proof for finite difference schemes devised for axisymmetric flows.

In our pointwise local truncation error estimate, a fundamental issue is the identification of smooth flows in the vicinity of the pole. Using a symmetry argument,

it can be shown [12] that if the swirling component is even in r (or more precisely, is the restriction of an even function on $r > 0$), the vector field is in fact singular. See Example 1 in section 2 for details. This is an easily overlooked mistake that even appeared in some research papers targeted at numerical search for potential formation of finite time singularities. In addition to the regularity issue at the axis of symmetry, a refined decay estimate for the stream function also plays an important role in our analysis. In general, the stream function only decays as $O((x^2 + r^2)^{-1})$ at infinity. Accordingly, we have selected an appropriate combination of weight functions that constitute an r-homogeneous norm. As a result, the slow decay of the stream function is compensated by the fast decay of velocity and vorticity. Overall, we obtained a second order error estimate on axisymmetric flows.

The rest of this paper is organized as follows: In section 2, we give a brief review of the regularity results developed in [12], including the characterization of pole regularity for general axisymmetric solenoidal vector fields and solutions of the axisymmetric NSE (2.2). In section 3, we formulate a regularity assumption on the solution of NSE at infinity. We basically assume that the swirling components of velocity and vorticity decay fast enough at infinity, and use this to analyze the decay rate of the stream function. In section 4, we briefly review the energy and helicity preserving property for EHPS and use it to prove our main theorem by energy estimate in section 5. The proof of some technical lemmas is given in the Appendix.

2. Generalized vorticity-stream formulation for axisymmetric flows. In this section, we review the generalized vorticity-stream formulation of axisymmetric NSE

$$\begin{aligned} &\partial_t \boldsymbol{u} + (\nabla \times \boldsymbol{u}) \times \boldsymbol{u} + \nabla p = -\nu \nabla \times \nabla \times \boldsymbol{u} \\ &\nabla \cdot \boldsymbol{u} = 0 \end{aligned} \tag{2.1}$$

and related regularity issues.

Denoting by the x-axis the axis of symmetry, the axisymmetric NSE in the cylindrical coordinate system $x = x$, $y = r\cos\theta$, $z = r\sin\theta$ can be written as [11]

$$\begin{aligned} &u_t + \tfrac{1}{r^2} J\left(ru, r\psi\right) = \nu(\nabla^2 - \tfrac{1}{r^2})u\,, \\ &\omega_t + J\left(\tfrac{\omega}{r}, r\psi\right) = \nu(\nabla^2 - \tfrac{1}{r^2})\omega + J\left(\tfrac{u}{r}, ru\right)\,, \\ &\omega = -(\nabla^2 - \tfrac{1}{r^2})\psi\,, \end{aligned} \tag{2.2}$$

where $J(a,b) = (\partial_x a)(\partial_r b) - (\partial_r a)(\partial_x b)$.

In (2.2), $u(t;x,r)$, $\omega(t;x,r)$, and $\psi(t;x,r)$ represent the swirling components of velocity, vorticity, and stream function, respectively. The quantity $r\psi$ is also known as Stokes' stream function and the formal correspondence between the solutions of (2.1) and (2.2) is given by

$$\boldsymbol{u} = u\boldsymbol{e}_\theta + \nabla \times (\psi \boldsymbol{e}_\theta) = \frac{\partial_r(r\psi)}{r}\boldsymbol{e}_x - \partial_x \psi \boldsymbol{e}_r + u\boldsymbol{e}_\theta, \tag{2.3}$$

where $\boldsymbol{e}_x$, $\boldsymbol{e}_r$, and $\boldsymbol{e}_\theta$ are the unit vectors in the x, r, and θ directions, respectively. The vorticity-stream formulation (2.2) has appeared in [5] with an alternative expression for the nonlinear terms. In [11], the authors have generalized the vorticity formulation to general symmetric flows with the nonlinear terms recast in Jacobians as in (2.2) and proposed a class of EHPS based on discretizing (2.2). In sections 4 and 5, we will review EHPS for (2.2) and give a rigorous error estimate in second order finite

difference setting. The error bound certainly depends on the regularity of the solution to (2.2). Although (2.2) can be derived formally from (2.1), the equivalence between the two expressions in terms of regularity of solutions is not quite obvious. An essential prerequisite to our analysis is to characterize the proper meaning of "smoothness" of solutions to (2.2). This turns out to be a subtle issue.

Example 1. Take

$$u(x,r) = r^2 e^{-r}, \qquad \omega = \psi \equiv 0. \tag{2.4}$$

It is easy to verify that (2.4) is an *exact* stationary solution of the Euler equation ($\nu = 0$ in (2.2)). Note that $u = O(r^2)$ near the axis and $\partial_r^2 u(x, 0^+) \neq 0$. Similar functions can be found in literatures as initial data in numerical search for finite time singularities. Although $u \in C^\infty(R \times \overline{R^+})$, the following regularity lemma for general axisymmetric solenoidal vector fields shows that $\boldsymbol{u} = u\boldsymbol{e}_\theta$ is not even in $C^2(R^3, R^3)$.

LEMMA 1 (see [12]). *Denote the axisymmetric divergence free subspace of C^k vector fields by*

$$\mathcal{C}_s^k \overset{def}{=} \{\boldsymbol{u} \in C^k(R^3, R^3), \quad \partial_\theta u_x = \partial_\theta u_r = \partial_\theta u_\theta = 0, \quad \nabla \cdot \boldsymbol{u} = 0\}. \tag{2.5}$$

Then

(a) *for any $\boldsymbol{u} \in \mathcal{C}_s^k$, there exists a unique (u, ψ) such that*

$$\boldsymbol{u} = u\boldsymbol{e}_\theta + \nabla \times (\psi \boldsymbol{e}_\theta) = \frac{\partial_r(r\psi)}{r}\boldsymbol{e}_x - \partial_x \psi \boldsymbol{e}_r + u\boldsymbol{e}_\theta, \qquad r > 0, \tag{2.6}$$

with

$$u(x,r) \in C^k(R \times \overline{R^+}), \quad \partial_r^{2\ell} u(x, 0^+) = 0 \text{ for } 0 \leq 2\ell \leq k, \tag{2.7}$$

and

$$\psi(x,r) \in C^{k+1}(R \times \overline{R^+}), \quad \partial_r^{2\ell} \psi(x, 0^+) = 0 \text{ for } 0 \leq 2\ell \leq k+1. \tag{2.8}$$

(b) *If (u, ψ) satisfies* (2.7), (2.8) *and $\boldsymbol{u}$ is given by* (2.6) *for $r > 0$, then $\boldsymbol{u} \in \mathcal{C}_s^k$ with a removable singularity at $r = 0$.*

Here in (2.5) and throughout this paper, the subscripts of u are used to denote components rather than partial derivatives. The proof of Lemma 1 is based on the observation that $\boldsymbol{e}_\theta$ changes direction across the axis of symmetry; therefore $u = u_\theta$ must admit an odd extension in order to compensate for this discontinuity. The details can be found in [12].

For simplicity of presentation, we recast Lemma 1 as follows.

LEMMA 1′.

$$\mathcal{C}_s^k = \{u\boldsymbol{e}_\theta + \nabla \times (\psi \boldsymbol{e}_\theta) \,|\, u \in C_s^k(R \times \overline{R^+}),\ \psi \in C_s^{k+1}(R \times \overline{R^+})\}, \tag{2.9}$$

where

$$C_s^k\left(R \times \overline{R^+}\right) \overset{def}{=} \left\{ f(x,r) \in C^k\left(R \times \overline{R^+}\right), \quad \partial_r^{2j} f(x, 0^+) = 0,\ 0 \leq 2j \leq k \right\}. \tag{2.10}$$

From Lemma 1 and Example 1, it is clear that the proper meaning of the smooth solution to (2.2) should be supplemented by the pole conditions (2.7), (2.8). In the case of NSE ($\nu > 0$), our main concern in this paper, (2.2) is an elliptic-parabolic

system on a semibounded region ($r > 0$). From standard PDE theory, we need to assign one and only one boundary condition for each of the variables ψ, u, and ω. An obvious choice is the zeroth order part of the pole conditions (2.7), (2.8):

$$\psi(x,0) = u(x,0) = \omega(x,0) = 0. \tag{2.11}$$

It is therefore a natural question to ask whether a smooth solution of (2.2), (2.11) in the class

$$\begin{aligned} &\psi(t;x,r) \in C^1\left(0,T;C^{k+1}(R\times\overline{R^+})\right),\\ &u(t;x,r) \in C^1\left(0,T;C^{k}(R\times\overline{R^+})\right),\\ &\omega(t;x,r) \in C^1\left(0,T;C^{k-1}(R\times\overline{R^+})\right) \end{aligned} \tag{2.12}$$

will give rise to a smooth solution of (2.2). In other words, is the pole condition (2.7), (2.8) automatically satisfied if only the zeroth order part (2.11) is imposed?

The answer to this question is affirmative.

THEOREM 1 (see [12]).

(a) *If $(\boldsymbol{u},p)$ is an axisymmetric solution to (2.1) with $\boldsymbol{u} \in C^1(0,T;\mathcal{C}_s^k)$, $p \in C^0(0,T;C^{k-1}(R^3))$, and $k \geq 3$, then there is a solution (ψ,u,ω) to (2.2) in the class*

$$\begin{aligned} &\psi(t;x,r) \in C^1\left(0,T;C_s^{k+1}(R\times\overline{R^+})\right),\\ &u(t;x,r) \in C^1\left(0,T;C_s^{k}(R\times\overline{R^+})\right),\\ &\omega(t;x,r) \in C^1\left(0,T;C_s^{k-1}(R\times\overline{R^+})\right), \end{aligned} \tag{2.13}$$

and $\boldsymbol{u} = u\boldsymbol{e}_\theta + \nabla\times(\psi\boldsymbol{e}_\theta)$.

(b) *If (ψ,u,ω) is a solution to (2.2), (2.11) in the class (2.12) with $k \geq 3$, then (ψ,u,ω) is in the class (2.13), $\boldsymbol{u} \overset{def}{=} u\boldsymbol{e}_\theta + \nabla\times(\psi\boldsymbol{e}_\theta) \in C^1(0,T;\mathcal{C}_s^k)$, and there is an axisymmetric scalar function $p \in C^0(0,T;C^{k-1}(R^3))$ such that $(\boldsymbol{u},p)$ is a solution to (2.1).*

The proof of Theorem 1 can be found in [12]. We remark here that Theorem 1 not only establishes the equivalence between (2.1) and (2.2) for classical solutions; the fact that smooth solutions to (2.2) automatically satisfy the pole condition (2.13) is also crucial to our local truncation error analysis. See the appendix for details.

3. Regularity assumption on solutions of NSE at infinity. The focus of this paper is the convergence rate of EHPS in the presence of the pole singularity. To separate difficulties and avoid complications introduced by physical boundaries, we only consider the whole space problems with solutions decaying rapidly at infinity.

To be more specific, we restrict our attention to the case where the supports of the initial data $u(x,0)$ and $\omega(x,0)$ are essentially compact. Since (2.2) is a transport diffusion equation for u and ω with initially finite speed of propagation, we expect u and ω to be essentially compactly supported, at least for short time. In the case of *linear* transport diffusion equations, the solution together with its derivatives will then decay faster than polynomials at infinity for $t > 0$. Some rigorous results concerning the spatial decay rate for the solutions of axisymmetric flows can be found in [4] and the references therein. In particular, it is shown in [4] that both u and ω decay algebraically at infinity as long as this is the case initially. Here we make a stronger

yet plausible assumption along this direction. The precise form of our assumption is formulated in terms of weighted norms and is less stringent than the analogy we draw from linear transport diffusion equations; see Assumption 1 below.

To quantify our assumption, we first introduce a family of r-homogeneous composite norms and corresponding function spaces which turn out to be natural for our pointwise energy estimate.

DEFINITION 1.

$$\|a\|_{\ell,\alpha,\beta} = \sum_{\ell_1+\ell_2=\ell} \| (1+r)^{\alpha}(1+|x|)^{\beta}|\partial_x^{\ell_1}\partial_r^{\ell_2}\left(\frac{a}{r}\right)| \|_{L^{\infty}(R\times\overline{R^+})}, \tag{3.1}$$

$$|\!|\!| a |\!|\!|_{k,\alpha,\beta} = \sum_{0\le\ell\le k} \|a\|_{k-\ell,\alpha-\ell,\beta}. \tag{3.2}$$

Note that the norms (3.1), (3.2) are well defined for functions in $C_s^k(R\times\overline{R^+})$ that decay properly at infinity. We denote them by

$$C_s^{k,\alpha,\beta} = \left\{ a(x,r) \in C_s^k\left(R\times\overline{R^+}\right), |\!|\!| a |\!|\!|_{k,\alpha,\beta} < \infty \right\}. \tag{3.3}$$

In section 5, we will show that EHPS is second order accurate provided the solution satisfies

$$\begin{cases} (\psi,\omega) \in C^1\left(0,T; C_s^{4,\alpha+\frac{7}{2},\beta} \cap C_s^{4,2\alpha+2,2\beta}\right), \\ u \in C^1\left(0,T; C_s^{4,2\alpha+2,2\beta} \cap C_s^{1,2,0}\right), \end{cases} \qquad \alpha > \frac{1}{2},\ \beta > \frac{1}{4}. \tag{3.4}$$

In view of (3.4), we formulate our regularity assumption as follows.

Assumption 1.

$$(\psi,\omega) \in C^1\left(0,T; C_s^{4,\gamma,\delta}\right), \quad u \in C^1\left(0,T; C_s^{4,5,\delta}\right), \quad \gamma > 4,\ \delta > \frac{1}{2}. \tag{3.5}$$

Although we expect u, ω and their derivatives to decay faster than any polynomial at infinity, the same expectation is not realizable for ψ. As we will see, generically ψ only decays like $O((x^2+r^2)^{-1})$ at infinity. Nevertheless, we will show that Assumption 1 is still realizable if ω decays fast enough.

To analyze the decay rate of ψ, we start with the integral expression for ψ. From the vorticity-stream relation

$$\nabla\times\nabla\times\boldsymbol{\psi} = \boldsymbol{\omega}$$

and the identification

$$\psi(x,r) = \psi_z(x,y,z)|_{y=r,z=0}, \qquad \omega(x,r) = \omega_z(x,y,z)|_{y=r,z=0},$$

one can derive the following integral formula for ψ [17]:

$$\psi(x,r) = \int_0^{\infty}\int_{-\infty}^{\infty} \omega(x',r')K(x-x',r,r')dx'dr', \tag{3.6}$$

where

$$\begin{aligned} K(x-x',r,r') &= r'\tfrac{1}{4\pi}\int_0^{2\pi} \frac{\cos\theta}{\sqrt{(x-x')^2+(r-r'\cos\theta)^2+(r'\sin\theta)^2}}d\theta \\ &= r'^2\tfrac{2}{\pi}\int_0^{\frac{\pi}{2}} \frac{r\cos^2\theta}{\rho_+\rho_-(\rho_++\rho_-)}d\theta \end{aligned} \tag{3.7}$$

and

$$\rho_\pm^2 = (x - x')^2 + (r \pm r' \cos\theta)^2 + (r' \sin\theta)^2.$$

As a consequence, we have the following far field estimate for K.

LEMMA 2.

$$|\partial_x^\ell \partial_r^m K(x - x', r, r')| \le C_{\ell,m}(x', r') \left(\sqrt{x^2 + r^2}\right)^{-2-\ell-m} \quad as \quad x^2 + r^2 \to \infty.$$

Proof. We will derive a far field estimate for the integrand in (3.7). We first consider a typical term

$$\lim_{x^2+r^2\to\infty} |\partial_x^\ell \partial_r^m \rho|$$

with

$$\rho^2 = (x - x_0)^2 + (r - r_0)^2 + c_0^2,$$

where x_0, r_0, and c_0 are some constants.

With the change of variables

$$\begin{aligned} r - r_0 &= \sigma \cos\lambda, \\ x - x_0 &= \sigma \sin\lambda, \end{aligned}$$

we can rewrite the x and r derivatives by

$$\begin{aligned} \partial_r \rho = \partial_r \sqrt{\sigma^2 + c_0^2} &= (\partial_r \sigma)\partial_\sigma \sqrt{\sigma^2 + c_0^2} + (\partial_r \lambda)\partial_\lambda \sqrt{\sigma^2 + c_0^2} = \tfrac{\sigma}{\rho} \cos\lambda, \\ \partial_x \rho = \partial_x \sqrt{\sigma^2 + c_0^2} &= (\partial_x \sigma)\partial_\sigma \sqrt{\sigma^2 + c_0^2} + (\partial_x \lambda)\partial_\lambda \sqrt{\sigma^2 + c_0^2} = \tfrac{\sigma}{\rho} \sin\lambda. \end{aligned}$$

Therefore by induction

$$\partial_x^\ell \partial_r^m \rho = P^{\ell,m}(\cos\lambda, \sin\lambda) Q^{\ell,m}(\sigma, \rho),$$

where $P^{\ell,m}(\cos\lambda, \sin\lambda)$ is a polynomial of degree $\ell+m$ in its arguments and $Q^{\ell,m}(\sigma, \rho)$ a rational function of σ and ρ of degree $1 - \ell - m$. By degree of a rational function we mean the degree of the numerator subtracting the degree of the denominator.

Since $\sigma = O(\sqrt{x^2 + r^2})$ and $\rho = O(\sqrt{x^2 + r^2})$, we conclude that

$$|\partial_x^\ell \partial_r^m \rho| = O\left(\sqrt{x^2 + r^2}^{1-\ell-m}\right).$$

We can now apply the argument above and Leibniz's rule to get

$$\partial_x^\ell \partial_r^m \frac{r}{\rho_+ \rho_- (\rho_+ + \rho_-)} = \sum_j^{J_{\ell,m}} \tilde{P}_j^{\ell,m}(\cos\lambda_+, \sin\lambda_+, \cos\lambda_-, \sin\lambda_-) \tilde{Q}_j^{\ell,m}(\sigma_+, \rho_+, \sigma_-, \rho_-, r),$$

where $J_{\ell,m}$ is a finite integer, $\sigma_\pm$ and $\rho_\pm$ are defined by

$$\begin{aligned} r \pm r' \cos\theta &= \sigma_\pm \cos\lambda_\pm, \\ x - x_0 &= \sigma_\pm \sin\lambda_\pm, \end{aligned}$$

and $\tilde{P}_j^{\ell,m}$, $\tilde{Q}_j^{\ell,m}$ are polynomials and rational functions of degrees $\ell + m$, $-2 - \ell - m$ in their arguments, respectively. The lemma follows by integrating θ over $(0, \frac{\pi}{2})$ in (3.7). ∎

We close this section by noting that ψ exhibits slow decay rate at infinity as a consequence of (3.6) and Lemma 2. More precisely, $\psi(x,r) \sim O((x^2+r^2)^{-1})$ in general. This may seem to raise the question whether Assumption 1 is realizable at all.

Indeed, using a similar calculation as in the proof of Lemma 2, one can derive the following.

PROPOSITION 1. *If $\gamma+\delta < k+2$ and $\omega \in C_s^{k,\gamma',\delta'}$ for sufficiently large γ' and δ', then $\psi \in C_s^{k,\gamma,\delta}$.*

As a consequence, we see that the range of γ and δ in (3.5) is not void provided ω decays fast enough at infinity. This justifies Assumption 1.

4. Energy and helicity preserving scheme. In this section, we outline the derivation of the discrete energy and helicity identities for EHPS. A key ingredient in the derivation is the reformulation of nonlinear terms into Jacobians. The details can be found in [11].

We introduce the standard notations:

$$D_x\phi(x,r) = \frac{\phi(x+\frac{\Delta x}{2},r)-\phi(x-\frac{\Delta x}{2},r)}{\Delta x}, \quad D_r\phi(x,r) = \frac{\phi(x,r+\frac{\Delta r}{2})-\phi(x,r-\frac{\Delta r}{2})}{\Delta r},$$

$$\tilde{D}_x\phi(x,r) = \frac{\phi(x+\Delta x,r)-\phi(x-\Delta x,r)}{2\Delta x}, \quad \tilde{D}_r\phi(x,r) = \frac{\phi(x,r+\Delta r)-\phi(x,r-\Delta r)}{2\Delta r},$$

and

$$\tilde{\nabla}_h = (\tilde{D}_x, \tilde{D}_r), \quad \tilde{\nabla}_h^\perp = (-\tilde{D}_r, \tilde{D}_x).$$

The finite difference approximation of ∇^2 and the Jacobians are given by

$$\nabla_h^2\psi = D_x\left(D_x\psi\right) + \frac{1}{r}\left(D_r(rD_r\psi)\right)$$

and

$$J_h(f,g) = \frac{1}{3}\left\{\tilde{\nabla}_h^\perp f\cdot\tilde{\nabla}_h g + \tilde{\nabla}_h^\perp\cdot(f\tilde{\nabla}_h g) + \tilde{\nabla}_h\cdot(g\tilde{\nabla}_h^\perp f)\right\}. \tag{4.1}$$

Altogether, the second order finite difference version of EHPS is

$$\begin{aligned} &\partial_t u_h + \tfrac{1}{r^2}J_h\left(ru_h, r\psi_h\right) = \nu(\nabla_h^2 - \tfrac{1}{r^2})u_h,\\ &\partial_t\omega_h + J_h\left(\tfrac{\omega_h}{r}, r\psi_h\right) = \nu(\nabla_h^2-\tfrac{1}{r^2})\omega_h + J_h\left(\tfrac{u_h}{r}, ru_h\right),\\ &\omega_h = (-\nabla_h^2+\tfrac{1}{r^2})\psi_h. \end{aligned} \tag{4.2}$$

To derive the discrete energy and helicity identity, we first introduce the discrete analogue of weighted inner products

$$\langle a,b\rangle_h = \sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}(rab)_{i,j}\,\Delta x\Delta r, \tag{4.3}$$

$$\begin{aligned}[a,b]_h = \Bigg(&\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\left(r(D_xa)(D_xb)\right)_{i-\frac{1}{2},j}\\ &+\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}{}'\left(r(D_ra)(D_rb)\right)_{i,j-\frac{1}{2}}\Bigg)\Delta x\Delta r + \langle\tfrac{a}{r},\tfrac{b}{r}\rangle_h, \end{aligned} \tag{4.4}$$

and the corresponding norms

$$\|a\|_{0,h}^2 = \langle a, a\rangle_h, \qquad \|a\|_{1,h}^2 = [a, a]_h, \tag{4.5}$$

where the grids have been shifted [13] to avoid placing the grid points on the axis of rotation:

$$x_i = i\Delta x, \quad i = 0, \pm 1, \pm 2, \ldots, \quad r_j = \left(j - \frac{1}{2}\right)\Delta r, \quad j = 1, 2, \ldots, \tag{4.6}$$

and

$$\sum_{j=1}^{\infty}{}' f_{j-\frac{1}{2}} = \frac{1}{2} f_{\frac{1}{2}} + \sum_{j=2}^{\infty} f_{j-\frac{1}{2}}. \tag{4.7}$$

The evaluation of the $\tilde{D}_r$ and ∇_h^2 terms in (4.2) at $j = 1$ involves the dependent variables u_h, ψ_h, ω_h and the stretching factor $h_3 = |\nabla\theta|^{-1} = r$ at the ghost points $j = 0$. In view of Lemma 1, we impose the following reflection boundary condition across the axis of rotation:

$$u_h(i, 0) = -u_h(i, 1), \quad \psi_h(i, 0) = -\psi_h(i, 1), \quad \omega_h(i, 0) = -\omega_h(i, 1). \tag{4.8}$$

Furthermore, we take even extension for the coordinate stretching factor $h_3 = |\nabla\theta|^{-1} = r$ which appears in the evaluation of the Jacobians at $j = 1$:

$$h_3(i, 0) = h_3(i, 1). \tag{4.9}$$

We will show in the remaining sections that the extensions (4.8) and (4.9) indeed give rise to a discrete version of energy and helicity identity and optimal local truncation error. As a consequence, second order accuracy of EHPS is justified for axisymmetric flows.

Remark 1. At first glance, the extension (4.9) may seem to contradict (4.6) on the ghost points $j = 0$. A less ambiguous restatement of (4.9) is to incorporate it into (4.2) as

(4.10)

$$\partial_t u_h + \tfrac{1}{r^2} J_h\left(|r|u_h, |r|\psi_h\right) = \nu(\nabla_h^2 - \tfrac{1}{r^2})u_h,$$

$$\partial_t \omega_h + J_h\left(\tfrac{\omega_h}{|r|}, |r|\psi_h\right) = \nu(\nabla_h^2 - \tfrac{1}{r^2})\omega_h + J_h\left(\tfrac{u_h}{|r|}, |r|u_h\right) \quad \text{on } (x_i, r_j), \quad j \geq 1,$$

$$\omega_h = (-\nabla_h^2 + \tfrac{1}{r^2})\psi_h.$$

The following identities are essential to the discrete energy and helicity identity and the error estimate.

LEMMA 3. *Suppose* (a, b, c) *satisfies the reflection boundary condition*

$$a(i, 0) = -a(i, 1), \quad b(i, 0) = -b(i, 1), \quad c(i, 0) = -c(i, 1)$$

and define

$$T_h(a, b, c) := \frac{1}{3}\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\left(c\tilde{\nabla}_h^{\perp} a \cdot \tilde{\nabla}_h b + a\tilde{\nabla}_h^{\perp} b \cdot \tilde{\nabla}_h c + b\tilde{\nabla}_h^{\perp} c \cdot \tilde{\nabla}_h a\right)_{i,j} \Delta x \Delta r. \tag{4.11}$$

Then

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty} c_{i,j}J_h(a,b)_{i,j}\Delta x\Delta r = T_h(a,b,c), \tag{4.12}$$

and

$$\left\langle a, \left(-\nabla_h^2 + \frac{1}{r^2}\right) b\right\rangle_h = [a,b]_h. \tag{4.13}$$

Proof. We first derive (4.12). In view of (4.1) and (4.11), it suffices to show that

$$\sum_j\sum_i c\tilde{\nabla}_h^{\perp}\cdot(a\tilde{\nabla}_h b) = -\sum_{i,j} a\tilde{\nabla}_h^{\perp}c\cdot\tilde{\nabla}_h b, \tag{4.14}$$

$$\sum_i\sum_j c\tilde{\nabla}_h\cdot(b\tilde{\nabla}_h^{\perp}a) = -\sum_{i,j} b\tilde{\nabla}_h c\cdot\tilde{\nabla}_h^{\perp}a \tag{4.15}$$

or, since there is no boundary terms in the x direction, simply

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}(f\tilde{D}_r g)_{i,j} = -\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}(g\tilde{D}_r f)_{i,j} \tag{4.16}$$

with $f = c$ and $g = b\tilde{D}_x a - a\tilde{D}_x b$.

Using the summation-by-parts identity (see, for example, [15] or [11]), it is straightforward to verify that

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}(f\tilde{D}_r g)_{i,j} = -\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}(g\tilde{D}_r f)_{i,j} - \sum_{i=-\infty}^{\infty}(f_{i,0}g_{i,1} + g_{i,0}f_{i,1}).$$

In the derivation of the discrete energy and helicity identities (see (4.18)–(4.20) below), a typical triplet (a,b,c) is given by, say, $a = r\psi_h$, $b = ru_h$, and $c = \frac{u_h}{r}$. From the reflection boundary condition (4.8) and (4.9), we see that

$$f_{i,0} = -f_{i,1}, \quad g_{i,0} = g_{i,1}.$$

This gives (4.16), and therefore (4.14), (4.15), and (4.12).

Next we derive (4.13). From the identity

$$\sum_{j=1}^{\infty} f_j(g_{j+\frac{1}{2}} - g_{j-\frac{1}{2}}) = -\sum_{j=1}^{\infty}{}'(f_j - f_{j-1})g_{j-\frac{1}{2}} - \frac{1}{2}(f_1 + f_0)g_{\frac{1}{2}}$$

and $r_{\frac{1}{2}} = 0$, it is easy to show that

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty} a_{i,j}D_r(rD_r b)_{i,j} = -\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}{}'(D_r a)_{i,j-\frac{1}{2}}r_{j-\frac{1}{2}}(D_r b)_{i,j-\frac{1}{2}}\,.$$

Therefore (4.13) follows. ∎

From (4.11), we can easily derive the permutation identities

$$T_h(a,b,c) = T_h(b,c,a) = T_h(c,a,b), \qquad T_h(a,b,c) = -T_h(b,a,c)\,. \tag{4.17}$$

Moreover, from (4.12), (4.13), it follows that

$$\langle \upsilon, \partial_t u_h\rangle_h + T_h(ru_h, r\psi_h, \tfrac{\upsilon}{r}) = \nu\langle \upsilon, (\nabla_h^2 - \tfrac{1}{r^2})u_h\rangle_h,$$

$$[\varphi, \partial_t \psi_h]_h + T_h(\tfrac{\omega_h}{r}, r\psi_h, r\varphi) = \nu\langle \varphi, (\nabla_h^2 - \tfrac{1}{r^2})\omega_h\rangle_h + T_h(\tfrac{u_h}{r}, ru_h, r\varphi), \tag{4.18}$$

$$\langle \xi, \omega_h\rangle_h = [\xi, \psi_h]_h$$

for all υ, φ, and ξ satisfying

$$\upsilon(i,0) = -\upsilon(i,1), \quad \varphi(i,0) = -\varphi(i,1), \quad \xi(i,0) = -\xi(i,1).$$

As a direct consequence of the permutation identity (4.17), we take $(\upsilon, \varphi) = (u_h, \psi_h)$ in (4.18) and recover the discrete energy identity

$$\frac{d}{dt}\frac{1}{2}(\langle u_h, u_h\rangle_h + [\psi_h, \psi_h]_h) + \nu([u_h, u_h]_h + \langle \omega_h, \omega_h\rangle_h) = 0. \tag{4.19}$$

Similarly, the discrete helicity identity

$$\frac{d}{dt}\langle u_h, \omega_h\rangle_h + \nu\left([u_h, \omega_h]_h - \left\langle \omega_h, \left(\nabla_h^2 - \frac{1}{r^2}\right)u_h\right\rangle_h\right) = 0 \tag{4.20}$$

follows by taking $(\upsilon, \varphi) = (\omega_h, u_h)$ in (4.18).

Remark 2. In the presence of physical boundaries, the no-slip boundary condition gives

$$\boldsymbol{u}\cdot\boldsymbol{n} = \partial_\tau(r\psi) = 0, \quad \boldsymbol{u}\cdot\boldsymbol{\tau} = \partial_n(r\psi) = 0, \quad \boldsymbol{u}\cdot\boldsymbol{e}_\theta = u = 0, \tag{4.21}$$

where $\boldsymbol{\tau} = \boldsymbol{n}\times\boldsymbol{e}_\theta$ and $\boldsymbol{e}_\theta$ is the unit vector in θ direction. When the cross section Ω is simply connected, (4.21) reads as follows:

$$u = 0, \quad \psi = 0, \quad \partial_n(r\psi) = 0 \quad \text{on} \quad \partial\Omega\,. \tag{4.22}$$

It can be shown that the energy and helicity identities (4.19), (4.20) remain valid in the presence of physical boundary conditions [11]. The numerical realization of the no-slip condition (4.22) introduced in [11] is second order accurate and seems to be new even for usual 2D flows. The convergence proof for this new boundary condition will be reported elsewhere.

5. Energy estimate and the main theorem. In this section, we proceed with the main theorem of the error estimate. We denote by $(\psi_h,\ u_h,\ \omega_h)$ the numerical solution satisfying

$$\begin{aligned}
&\partial_t u_h + \tfrac{1}{r^2}J_h(ru_h, r\psi_h) = \nu(\nabla_h^2 - \tfrac{1}{r^2})u_h,\\
&\partial_t \omega_h + J_h\left(\tfrac{\omega_h}{r}, r\psi_h\right) = \nu(\nabla_h^2 - \tfrac{1}{r^2})\omega_h + J_h\left(\tfrac{u_h}{r}, ru_h\right),\\
&\omega_h = (-\nabla_h^2 + \tfrac{1}{r^2})\psi_h,
\end{aligned} \tag{5.1}$$

and $(\psi,\ u,\ \omega)$ the exact solution to (2.2),

$$\begin{aligned}
&\partial_t u + \tfrac{1}{r^2}J_h(ru, r\psi) = \nu(\nabla_h^2 - \tfrac{1}{r^2})u + \mathcal{E}_1,\\
&\partial_t \omega + J_h\left(\tfrac{\omega}{r}, r\psi\right) = \nu(\nabla_h^2 - \tfrac{1}{r^2})\omega + J_h\left(\tfrac{u}{r}, ru\right) + \mathcal{E}_2,\\
&\omega = (-\nabla_h^2 + \tfrac{1}{r^2})\psi + \mathcal{E}_3,
\end{aligned} \tag{5.2}$$

where the local truncation errors $\mathcal{E}_j$ can be derived by subtracting (2.2) from (5.2):

$$\begin{aligned}
&\mathcal{E}_1 = \tfrac{1}{r^2}(J_h - J)(ru, r\psi) - \nu(\nabla_h^2 - \nabla^2)u,\\
&\mathcal{E}_2 = (J_h - J)\left(\tfrac{\omega}{r}, r\psi\right) - \nu(\nabla_h^2 - \nabla^2)\omega - (J_h - J)\left(\tfrac{u}{r}, ru\right),\\
&\mathcal{E}_3 = (\nabla_h^2 - \nabla^2)\psi.
\end{aligned} \tag{5.3}$$

From (5.1) and (5.2), we see that

$$\partial_t(u - u_h) + \frac{1}{r^2}\left(J_h(ru, r\psi) - J_h(ru_h, r\psi_h)\right) = \nu\left(\nabla_h^2 - \frac{1}{r^2}\right)(u - u_h) + \mathcal{E}_1, \tag{5.4}$$

$$\begin{aligned}
&\partial_t(\omega - \omega_h) + \left(J_h\left(\tfrac{\omega}{r}, r\psi\right) - J_h\left(\tfrac{\omega_h}{r}, r\psi_h\right)\right)\\
=\;& \nu(\nabla_h^2 - \tfrac{1}{r^2})(\omega - \omega_h) + \left(J_h\left(\tfrac{u}{r}, ru\right) - J_h\left(\tfrac{u_h}{r}, ru_h\right)\right) + \mathcal{E}_2,
\end{aligned} \tag{5.5}$$

$$(\omega - \omega_h) = \left(-\nabla_h^2 + \frac{1}{r^2}\right)(\psi - \psi_h) + \mathcal{E}_3. \tag{5.6}$$

Lemmas 4 and 5 below are key to our error estimate. The permutation identities (4.17) associated with EHPS result in exact cancellation among the nonlinear terms and lead to an exact identity (5.7). The estimates for the trilinear form in (5.13), (5.14) then furnish necessary inequalities for our a priori error estimate. The proof for Lemma 5 and the local truncation error analysis, Lemma 6, is given in the appendix.

LEMMA 4.

$$\begin{aligned}
&\tfrac{1}{2}\partial_t(\|u - u_h\|_{0,h}^2 + \|\psi - \psi_h\|_{1,h}^2) + \nu(\|u - u_h\|_{1,h}^2 + \|\omega - \omega_h\|_{0,h}^2)\\
&= \langle u - u_h, \mathcal{E}_1\rangle_h + \langle \psi - \psi_h, \mathcal{E}_2 - \partial_t\mathcal{E}_3\rangle_h + \nu\langle \omega - \omega_h, \mathcal{E}_3\rangle_h - T_h\left(\tfrac{u - u_h}{r}, r(u - u_h), r\psi\right)\\
&\quad - T_h\left(r(\psi - \psi_h), \tfrac{(\omega - \omega_h)}{r}, r\psi\right) + T_h\left(r(\psi - \psi_h), \tfrac{u}{r}, r(u - u_h)\right).
\end{aligned} \tag{5.7}$$

Proof. We take the weighted inner product of $u - u_h$ with (5.4) to get

$$\begin{aligned}
&\tfrac{1}{2}\partial_t\|u - u_h\|_{0,h}^2 + \langle u - u_h, \tfrac{1}{r^2}\left(J_h(ru, r\psi) - J_h(ru_h, r\psi_h)\right)\rangle_h\\
&= \nu\langle u - u_h, (\nabla_h^2 - \tfrac{1}{r^2})(u - u_h)\rangle_h + \langle u - u_h, \mathcal{E}_1\rangle_h.
\end{aligned} \tag{5.8}$$

The second term on the left-hand side of (5.8) can be rewritten as

$$\begin{aligned}
&\langle u - u_h, \tfrac{1}{r^2}\left(J_h(ru, r\psi) - J_h(ru_h, r\psi_h)\right)\rangle_h\\
&= T_h\left(\tfrac{u - u_h}{r}, ru, r\psi\right) - T_h\left(\tfrac{u - u_h}{r}, ru_h, r\psi_h\right)\\
&= -T_h\left(\tfrac{u - u_h}{r}, r(u - u_h), r(\psi - \psi_h)\right) + T_h\left(\tfrac{u - u_h}{r}, r(u - u_h), r\psi\right)\\
&\qquad + T_h\left(\tfrac{u - u_h}{r}, ru, r(\psi - \psi_h)\right).
\end{aligned} \tag{5.9}$$

In addition, from (4.13) we have

$$\nu\left\langle u - u_h, \left(\nabla_h^2 - \frac{1}{r^2}\right)(u - u_h)\right\rangle_h = -\nu[u - u_h, u - u_h]_h = -\nu\|u - u_h\|_{1,h}^2.$$

Thus

$$
\begin{aligned}
&\tfrac{1}{2}\partial_t\|u-u_h\|_{0,h}^2 - T_h\left(\tfrac{u-u_h}{r}, r(u-u_h), r(\psi-\psi_h)\right) + \nu\|u-u_h\|_{1,h}^2 \\
&= \langle u-u_h, \mathcal{E}_1\rangle_h - T_h\left(\tfrac{u-u_h}{r}, r(u-u_h), r\psi\right) - T_h\left(\tfrac{u-u_h}{r}, ru, r(\psi-\psi_h)\right).
\end{aligned} \tag{5.10}
$$

Similarly, we take the weighted inner product of $\psi-\psi_h$ with (5.5) and proceed as (5.9)–(5.10) to get

$$
\begin{aligned}
&\tfrac{1}{2}\partial_t\|\psi-\psi_h\|_{1,h}^2 + T_h\left(r(\psi-\psi_h), \tfrac{(\omega-\omega_h)}{r}, r\psi\right) \\
&= -T_h\left(r(\psi-\psi_h), \tfrac{(u-u_h)}{r}, r(u-u_h)\right) + T_h\left(r(\psi-\psi_h), \tfrac{u}{r}, r(u-u_h)\right) \\
&\quad + T_h\left(r(\psi-\psi_h), \tfrac{(u-u_h)}{r}, ru\right) + \langle \psi-\psi_h, \mathcal{E}_2 - \partial_t\mathcal{E}_3\rangle_h \\
&\quad + \nu\langle(\psi-\psi_h), (\nabla_h^2 - \tfrac{1}{r^2})(\omega-\omega_h)\rangle_h .
\end{aligned} \tag{5.11}
$$

Next, we apply (4.13) twice to get

$$
\begin{aligned}
\nu\left\langle(\psi-\psi_h), \left(\nabla_h^2 - \frac{1}{r^2}\right)(\omega-\omega_h)\right\rangle_h &= \nu\left\langle\left(\nabla_h^2 - \frac{1}{r^2}\right)(\psi-\psi_h), \omega-\omega_h\right\rangle_h \\
&= -\nu\|\omega-\omega_h\|_{0,h}^2 + \nu\langle\omega-\omega_h, \mathcal{E}_3\rangle_h,
\end{aligned} \tag{5.12}
$$

and (5.7) follows. This completes the proof of this lemma. ∎

We now proceed with the estimate for the trilinear form T_h.

LEMMA 5. *For a, b, and $c \in C_s^2(R\times\overline{R^+})$, we have*

$$
|T_h\left(ra, rb, \frac{c}{r}\right)| \le C\|a\|_{1,h}\|b\|_{1,h}|\!|\!|c|\!|\!|_{1,2,0} \tag{5.13}
$$

and

$$
|T_h\left(\frac{a}{r}, rb, rc\right)| \le C\|a\|_{0,h}\|b\|_{1,h}|\!|\!|c|\!|\!|_{2,2,0}. \tag{5.14}
$$

Proof. See section A.1. ∎

From Lemmas 4 and 5, we can therefore derive

$$
\begin{aligned}
&\tfrac{1}{2}\partial_t(\|u-u_h\|_{0,h}^2 + \|\psi-\psi_h\|_{1,h}^2) + \nu(\|u-u_h\|_{1,h}^2 + \|\omega-\omega_h\|_{0,h}^2) \\
\le\ & |\langle u-u_h, \mathcal{E}_1\rangle_h| + |\langle\psi-\psi_h, \mathcal{E}_2-\partial_t\mathcal{E}_3\rangle_h| + \nu|\langle\omega-\omega_h, \mathcal{E}_3\rangle_h| \\
&+ C\|u-u_h\|_{0,h}\|u-u_h\|_{1,h}|\!|\!|\psi|\!|\!|_{2,2,0} + C\|\omega-\omega_h\|_{0,h}\|\psi-\psi_h\|_{1,h}|\!|\!|\psi|\!|\!|_{2,2,0} \\
&+ C\|\psi-\psi_h\|_{1,h}\|u-u_h\|_{1,h}|\!|\!|u|\!|\!|_{1,2,0}.
\end{aligned} \tag{5.15}
$$

Since

$$
\|\frac{a}{r}\|_{0,h} \le \|a\|_{1,h},
$$

we can further estimate the first few terms on the right-hand side of (5.15) by

$$
|\langle u-u_h, \mathcal{E}_1\rangle_h| = |\left\langle\frac{u-u_h}{r}, r\mathcal{E}_1\right\rangle_h| \le \frac{\nu}{4}\|u-u_h\|_{1,h}^2 + \frac{1}{\nu}\|r\mathcal{E}_1\|_{0,h}^2,
$$

$$
|\langle\psi-\psi_h, \mathcal{E}_2-\partial_t\mathcal{E}_3\rangle_h| \le \|\psi-\psi_h\|_{1,h}^2 + \|r(\mathcal{E}_2-\partial_t\mathcal{E}_3)\|_{0,h}^2,
$$

and

$$|\langle \omega - \omega_h, \mathcal{E}_3 \rangle_h| \le \frac{1}{2}\|\omega - \omega_h\|_{0,h}^2 + \frac{1}{2}\|\mathcal{E}_3\|_{0,h}^2.$$

Applying Hölder's inequality to the remaining terms of (5.15), we have derived the following proposition.

PROPOSITION 2.

$$\begin{aligned}
&\tfrac{1}{2}\partial_t(\|u-u_h\|_{0,h}^2 + \|\psi-\psi_h\|_{1,h}^2) + \tfrac{\nu}{4}(\|u-u_h\|_{1,h}^2 + \|\omega-\omega_h\|_{0,h}^2) \\
&\le \|\psi-\psi_h\|_{1,h}^2 + \tfrac{C}{\nu}\|r\mathcal{E}_1\|_{0,h}^2 + \|r\mathcal{E}_2\|_{0,h}^2 + \|r\partial_t\mathcal{E}_3\|_{0,h}^2 \\
&\quad + \nu\|\mathcal{E}_3\|_{0,h}^2 + \tfrac{C}{\nu}\|u-u_h\|_{0,h}^2 |\!|\!| \psi |\!|\!|_{2,2,0}^2 \\
&\quad + \tfrac{C}{\nu}\|\psi-\psi_h\|_{1,h}^2 |\!|\!| \psi |\!|\!|_{2,2,0}^2 + \tfrac{C}{\nu}\|\psi-\psi_h\|_{1,h}^2 |\!|\!| u |\!|\!|_{1,2,0}^2.
\end{aligned} \tag{5.16}$$

With Proposition 2, it remains to estimate $\|r\mathcal{E}_1\|_{0,h}$, $\|r\mathcal{E}_2\|_{0,h}$, $\|r\partial_t\mathcal{E}_3\|_{0,h}$, and $\|\mathcal{E}_3\|_{0,h}$. We summarize the results in the following lemma.

LEMMA 6. *Let $(\psi, u, \omega) \in C^1(0,T;C_s^4)$ be a solution of the axisymmetric NSE* (2.2) *and $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ be defined by* (5.2). *Then we have the following pointwise local truncation error estimate for $\alpha, \beta \in R$:*

$$r|\mathcal{E}_1| \le C\frac{\Delta x^2 + \Delta r^2}{(1+r)^{2\alpha}(1+|x|)^{2\beta}} \left(|\!|\!| \psi |\!|\!|_{4,\alpha+\frac{7}{2},\beta} |\!|\!| u |\!|\!|_{4,\alpha+\frac{7}{2},\beta} + |\!|\!| u |\!|\!|_{4,2\alpha+2,2\beta} \right), \tag{5.17}$$

$$r|\mathcal{E}_2| \le C\frac{\Delta x^2 + \Delta r^2}{(1+r)^{2\alpha}(1+|x|)^{2\beta}} \left(|\!|\!| \psi |\!|\!|_{4,\alpha+\frac{7}{2},\beta} |\!|\!| \omega |\!|\!|_{4,\alpha+\frac{7}{2},\beta} + |\!|\!| u |\!|\!|_{4,\alpha+\frac{7}{2},\beta}^2 + |\!|\!| \omega |\!|\!|_{4,2\alpha+2,2\beta} \right), \tag{5.18}$$

$$r|\partial_t\mathcal{E}_3| \le C\frac{\Delta x^2 + \Delta r^2}{(1+r)^{2\alpha}(1+|x|)^{2\beta}} |\!|\!| \partial_t\psi |\!|\!|_{4,2\alpha+2,2\beta}, \tag{5.19}$$

and

$$|\mathcal{E}_3| \le C\frac{\Delta x^2 + \Delta r^2}{r(1+r)^{2\alpha}(1+|x|)^{2\beta}} |\!|\!| \psi |\!|\!|_{4,2\alpha+2,2\beta}. \tag{5.20}$$

Proof. See section A.2. ∎

From Lemma 4 to 6, our main result follows.

THEOREM 2. *Let (ψ, u, ω) be a solution of the axisymmetric NSE* (2.2) *satisfying*

$$(\psi, \omega) \in C^1\left(0,T;C_s^{4,\gamma,\delta}\right), \quad u \in C^1\left(0,T;C_s^{4,5,\delta}\right), \qquad \gamma > 4,\ \delta > \frac{1}{2}. \tag{5.21}$$

Then

$$\begin{aligned}
&\sup_{[0,T]} \left(\|u-u_h\|_{0,h}^2 + \|\psi-\psi_h\|_{1,h}^2\right) \\
&\qquad + \int_0^T (\|u-u_h\|_{1,h}^2 + \|\omega-\omega_h\|_{0,h}^2)dt \le C(\Delta x^4 + \Delta r^4)|\log \Delta r|,
\end{aligned} \tag{5.22}$$

where $C = C(\psi, u, \nu, T)$.

Proof. From Lemma 6, we have

$$\begin{aligned}&\|r\mathcal{E}_1\|_{0,h}^2+\|r\mathcal{E}_2\|_{0,h}^2+\|r\partial_t\mathcal{E}_3\|_{0,h}^2\\ &\le C(\Delta x^4+\Delta r^4)\left(\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\frac{r_j\Delta r\Delta x}{(1+r_j)^{4\alpha}(1+|x_i|)^{4\beta}}\right)\\ &\quad\times\left(|\!|\!|(\psi,u,\omega)|\!|\!|_{4,\alpha+\frac{7}{2},\beta}^4+|\!|\!|(u,\omega,\partial_t\psi)|\!|\!|_{4,2\alpha+2,2\beta}^2\right).\end{aligned}$$

Similarly,

$$\|\mathcal{E}_3\|_{0,h}^2\le C(\Delta x^4+\Delta r^4)\left(\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\frac{\Delta r\Delta x}{r_j(1+r_j)^{4\alpha}(1+|x_i|)^{4\beta}}\right)|\!|\!|\psi|\!|\!|_{4,2\alpha+2,2\beta}^2.$$

Since

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\frac{r_j\Delta r\Delta x}{(1+r_j)^{4\alpha}(1+|x_i|)^{4\beta}}\le C\qquad\text{for }\alpha>\frac{1}{2},\ \beta>\frac{1}{4}$$

and

$$\sum_{i=-\infty}^{\infty}\sum_{j=1}^{\infty}\frac{\Delta r\Delta x}{r_j(1+r_j)^{4\alpha}(1+|x_i|)^{4\beta}}\le C|\log\Delta r|\qquad\text{for }\alpha>0,\ \beta>\frac{1}{4},$$

it follows that

(5.23)
$$\|r\mathcal{E}_1\|_{0,h}^2+\|r\mathcal{E}_2\|_{0,h}^2+\|r\partial_t\mathcal{E}_3\|_{0,h}^2\le C(\Delta x^4+\Delta r^4)\left(|\!|\!|(\psi,u,\omega)|\!|\!|_{4,\gamma,\delta}^4+|\!|\!|(u,\omega,\partial_t\psi)|\!|\!|_{4,\gamma,\delta}^2\right)$$

and

$$\|\mathcal{E}_3\|_{0,h}^2\le C(\Delta x^4+\Delta r^4)|\log\Delta r||\!|\!|\psi|\!|\!|_{4,\gamma,\delta}^2 \tag{5.24}$$

provided $\gamma>4$, $\delta>\frac{1}{2}$.

Under assumption (5.21), we have, in particular, $\psi\in C_s^{2,2,0}$, $u\in C_s^{1,2,0}$. It follows from Proposition 2 and (5.23), (5.24) that

$$\begin{aligned}&\tfrac{1}{2}\partial_t(\|u-u_h\|_{0,h}^2+\|\psi-\psi_h\|_{1,h}^2)+\tfrac{\nu}{4}(\|u-u_h\|_{1,h}^2+\|\omega-\omega_h\|_{0,h}^2)\\ \le\ & C\|u-u_h\|_{0,h}^2+C\|\psi-\psi_h\|_{1,h}^2+C(\Delta x^4+\Delta r^4)|\log\Delta r|.\end{aligned}$$

The error estimate (5.22) then follows from Gronwall's inequality. ∎

6. Conclusion. The importance and subtlety of the pole singularity has been a major difficulty in theoretical analysis and algorithm design for axisymmetric flows. The numerical analysis near the pole singularity is much more complicated than that of standard smooth flows. The principal ingredients of our error analysis are as follows:

(a) The fact that smooth solutions to (2.2) automatically satisfy the pole condition and thus belong to the class (2.13). This symmetry property plays an essential role in the local truncation error analysis.

(b) Proper formulation and discretization of the nonlinear terms. Here the Jacobian formulation along with the distinctive discretization (4.1) result in exact cancellation among the nonlinear terms in the energy estimate and therefore lead to conservation identities in discrete setting.

These ingredients may also serve as a guideline of algorithm design for axisymmetric flows.

In addition, the slow decay of the stream function at infinity poses extra technical difficulties in analyzing the whole space problem. This difficulty is carefully resolved by choosing a properly weighted r-homogeneous norm (3.2). On the one hand, (3.2) takes into account the local behavior of the swirling components near the pole singularity. On the other hand, it incorporates free parameters so that the slow decay of the stream function can be properly compensated by tuning the parameters through careful analysis.

Appendix A. Proof of technical lemmas.

A.1. Estimate for the trilinear form T_h—proof of Lemma 5. We start with the following basic identities.

PROPOSITION 3. *Define*

$$(\widetilde{A}_x f)_{i,j} = \frac{1}{2}(f_{i+1,j} + f_{i-1,j}), \qquad (\widetilde{A}_r f)_{i,j} = \frac{1}{2}(f_{i,j+1} + f_{i,j-1}).$$

Then the following estimates hold for $j \geq 1$:

$$|\widetilde{D}_r(ra)| \leq C|\widetilde{A}_r a| + Cr|\widetilde{D}_r a|, \tag{A.1}$$

$$|\widetilde{D}_r(\frac{a}{r})| \leq C\frac{|\widetilde{A}_r a|}{r^2} + C\frac{|\widetilde{D}_r a|}{r}, \tag{A.2}$$

$$|\widetilde{A}_r(ra)| \leq Cr\widetilde{A}_r|a|, \tag{A.3}$$

$$|\Delta r\widetilde{D}_r a| \leq \widetilde{A}_r|a|, \quad |\Delta x\widetilde{D}_x a| \leq \widetilde{A}_x|a|. \tag{A.4}$$

Remark 3. As in Remark 1, the stretching factor r in the arguments of the left-hand side of (A.1)–(A.3) satisfy the even extension (4.9). A more precise statement for, say, (A.1) is given by

$$|\widetilde{D}_r(|r|a)|_{i,j} \leq C|\widetilde{A}_r a|_{i,j} + Cr_j|\widetilde{D}_r a|_{i,j}, \qquad j \geq 1.$$

For simplicity of presentation, we will adopt the expression as in (A.1)–(A.3) through the rest of the paper.

Proof of Proposition 3. It is easy to verify that

$$\widetilde{D}_r(fg) = (\widetilde{A}_r f)(\widetilde{D}_r g) + (\widetilde{A}_r g)(\widetilde{D}_r f), \quad \widetilde{D}_x(fg) = (\widetilde{A}_x f)(\widetilde{D}_x g) + (\widetilde{A}_x g)(\widetilde{D}_x f).$$

A straightforward calculation shows that

$$(\tilde{A}_r|r|)_j \leq Cr_j, \quad |\,\tilde{D}_r|r|\,|_j \leq C$$

and

$$\tilde{A}_r\left(\frac{1}{|r|}\right)_j \leq C\frac{1}{r_j}, \quad |\tilde{D}_r\left(\frac{1}{|r|}\right)|_j \leq C\frac{1}{r_j^2}$$

for $j \geq 1$. The estimates (A.1)–(A.3) then follow. The proof for (A.4) is also straightforward. ∎

Proof of Lemma 5. We begin with the proof of (5.13). We expand the left-hand side as

$$\begin{aligned} T_h\left(ra, rb, \tfrac{c}{r}\right) &= \tfrac{1}{3}\Big(\langle \tfrac{c}{r^2}, \tilde{\nabla}_h^{\perp}(ra)\cdot\tilde{\nabla}_h(rb)\rangle_h + \langle a, \tilde{\nabla}_h^{\perp}(rb)\cdot\tilde{\nabla}_h(\tfrac{c}{r})\rangle_h \\ &\qquad + \langle b, \tilde{\nabla}_h^{\perp}(\tfrac{c}{r})\cdot\tilde{\nabla}_h(ra)\rangle_h\Big) \\ &= \tfrac{1}{3}(I_1 + I_2 + I_3) \end{aligned}$$

and estimate the I_j's term by term. First, we have

$$|I_1| = |\left\langle \frac{c}{r^2}, \tilde{\nabla}_h^{\perp}(ra)\cdot\tilde{\nabla}_h(rb)\right\rangle_h| = |\left\langle c, -\frac{1}{r}\tilde{D}_r(ra)\tilde{D}_x(b) + \tilde{D}_x(a)\frac{1}{r}\tilde{D}_r(rb)\right\rangle_h|;$$

therefore the estimate

$$\begin{aligned} |I_1| &\leq C\left\langle |c|, \left(|\frac{\tilde{A}_r(a)}{r}| + |\tilde{D}_r(a)|\right)|\tilde{D}_x(b)| + \left(|\frac{\tilde{A}_r(b)}{r}| + |\tilde{D}_r(b)|\right)|\tilde{D}_x(a)|\right\rangle_h \\ &\leq C\|a\|_{1,h}\|b\|_{1,h}\|c\|_{0,1,0} \end{aligned}$$

follows from (A.1), Hölder's inequality, and the inequality $|c| = |r\frac{c}{r}| \leq \|c\|_{0,1,0}$.

Second, we have

$$\begin{aligned} |I_2| &\leq C\langle |a|, |\tfrac{\tilde{A}_r(b)}{r}| + |\tilde{D}_r(b)||\tilde{D}_x(c)|\rangle_h + C\langle |a|, |\tilde{D}_r(c)||\tilde{D}_x(b)|\rangle_h \\ &\qquad + C\langle \tfrac{|a|}{r}, |A_r(c)||\tilde{D}_x(b)|\rangle_h \\ &= C\langle \tfrac{|a|}{r}, |\tfrac{\tilde{A}_r(b)}{r}| + |\tilde{D}_r(b)||r\tilde{D}_x(c)|\rangle_h + C\langle \tfrac{|a|}{r}, |r\tilde{D}_r(c)||\tilde{D}_x(b)|\rangle_h \\ &\qquad + C\langle \tfrac{|a|}{r}, |A_r(c)||\tilde{D}_x(b)|\rangle_h \\ &\leq C\|a\|_{1,h}\|b\|_{1,h}(\|c\|_{0,1,0} + \|c\|_{1,2,0}) \leq C\|a\|_{1,h}\|b\|_{1,h}\|c\|_{1,2,0}. \end{aligned}$$

The estimate for I_3 is similar and (5.13) follows.

Next we proceed with (5.14). Since

$$|T_h\left(\frac{a}{r}, rb, rc\right)| = |\left\langle a, \frac{1}{r^2}J_h(rb, rc)\right\rangle_h| \leq \|a\|_{0,h}\|\frac{1}{r^2}J_h(rb, rc)\|_{0,h},$$

it suffices to give a pointwise estimate for the integrand $J_h(rb, rc)$ as follows:

(A.5)

$$\begin{aligned} -3J_h(rb, rc) &= \widetilde{D}_r(rb)\widetilde{D}_x(rc) - \widetilde{D}_x(rb)\widetilde{D}_r(rc) + \widetilde{D}_r\big(rb\widetilde{D}_x(rc)\big) - \widetilde{D}_x\big(rb\widetilde{D}_r(rc)\big) \\ &\quad + \widetilde{D}_x\big(rc\widetilde{D}_r(rb)\big) - \widetilde{D}_r\big(rc\widetilde{D}_x(rb)\big) \\ &= \widetilde{D}_r(rb)(I + \widetilde{A}_r)\widetilde{D}_x(rc) - \widetilde{D}_x(rb)(I + \widetilde{A}_x)\widetilde{D}_r(rc) \\ &\quad + (\widetilde{A}_r - \widetilde{A}_x)(rb)\widetilde{D}_r\widetilde{D}_x(rc) + (\tilde{A}_x - \tilde{A}_r)(rc)\tilde{D}_x\tilde{D}_r(rb) \\ &\quad + \widetilde{D}_x(rc)\widetilde{A}_x\widetilde{D}_r(rb) - \widetilde{D}_r(rc)\widetilde{A}_r\widetilde{D}_x(rb) \\ &= \widetilde{D}_r(rb)(I + \widetilde{A}_r)\widetilde{D}_x(rc) - \widetilde{D}_x(rb)(I + \widetilde{A}_x)\widetilde{D}_r(rc) \\ &\quad + (\widetilde{A}_r - \widetilde{A}_x)(rb)\widetilde{D}_r\widetilde{D}_x(rc) \\ &\quad + \tfrac{1}{2}\Delta x^2\widetilde{D}_r\widetilde{D}_x(rb)D_x^2(rc) - \tfrac{1}{2}\Delta r^2\widetilde{D}_r\widetilde{D}_x(rb)D_r^2(rc) \\ &\quad + \widetilde{D}_x(rc)\widetilde{A}_x\widetilde{D}_r(rb) - \widetilde{D}_r(rc)\widetilde{A}_r\widetilde{D}_x(rb). \end{aligned}$$

Here I is the identity operator and we have used the identities

$$\tilde{A}_x = \frac{1}{2}\Delta x^2 D_x^2 + I, \quad \tilde{A}_r = \frac{1}{2}\Delta r^2 D_r^2 + I$$

in the second equality of (A.5).

From (A.1), the first two terms on the right-hand side of (A.5) can be estimated by

$$|\widetilde{D}_r(rb)(I+\widetilde{A}_r)\widetilde{D}_x(rc)| \le Cr^2\left(|\widetilde{D}_r b| + \frac{|\widetilde{A}_r b|}{r}\right)\|\partial_x c\|_{L^\infty}, \tag{A.6}$$

$$|\widetilde{D}_x(rb)(I+\widetilde{A}_x)\widetilde{D}_r(rc)| \le Cr^2|\widetilde{D}_x b|\|\partial_r c + \frac{c}{r}\|_{L^\infty} \le Cr^2|\widetilde{D}_x b|(\|c\|_{0,0,0} + \|c\|_{1,1,0}). \tag{A.7}$$

From (A.3) and (A.4), we can similarly estimate the remaining terms in (A.5):

$$\begin{aligned}|(\widetilde{A}_r - \widetilde{A}_x)(rb)\widetilde{D}_r\widetilde{D}_x(rc)| &\le Cr^2\frac{(\widetilde{A}_r+\widetilde{A}_x)|b|}{r}\|\partial_x\partial_r(rc)\|_{L^\infty}\\ &\le Cr^2\frac{(\widetilde{A}_r+\widetilde{A}_x)|b|}{r}(\|c\|_{1,1,0} + \|c\|_{2,2,0}),\end{aligned} \tag{A.8}$$

$$|\frac{1}{2}\Delta x^2\widetilde{D}_r\widetilde{D}_x(rb)D_x^2(rc)| \le C\frac{(\Delta x)^2}{\Delta r}|\widetilde{A}_r(r\widetilde{D}_x(b))D_x^2(rc)| \le Cr^2\frac{\Delta r}{r}\widetilde{A}_r|\widetilde{D}_x b|\|c\|_{2,2,0}, \tag{A.9}$$

$$|\frac{1}{2}\Delta r^2\widetilde{D}_r\widetilde{D}_x(rb)D_r^2(rc)| \le C\Delta r|\widetilde{A}_r\widetilde{D}_x(rb)|\|\partial_r^2(rc)\|_{L^\infty} \le Cr^2\frac{\Delta r}{r}\widetilde{A}_r|\widetilde{D}_x b|\|\!|c|\!\|_{2,2,0}, \tag{A.10}$$

$$|\widetilde{A}_x\widetilde{D}_r(rb)\widetilde{D}_x(rc)| \le Cr^2|\widetilde{A}_x\left(\frac{1}{r}\widetilde{D}_r(rb)\right)|\|\partial_x c\|_{L^\infty} \le Cr^2\widetilde{A}_x\left(|\widetilde{D}_r b| + \frac{1}{r}\widetilde{A}_r|b|\right)\|c\|_{1,1,0}, \tag{A.11}$$

and

$$|\widetilde{A}_r\widetilde{D}_x(rb)\widetilde{D}_r(rc)| \le Cr^2\widetilde{A}_r|\widetilde{D}_x b|\|\!|c|\!\|_{1,1,0}. \tag{A.12}$$

From (A.6)–(A.12), we can estimate the weighted L^2 norm of $\frac{1}{r^2}J_h(rb,rc)$ by

$$\|\frac{1}{r^2}J_h(rb,rc)\|_{0,h} \le C\|\left(|\widetilde{D}_x b| + |\widetilde{D}_r b| + \frac{|b|}{r}\right)\|_{0,h}\|\!|c|\!\|_{2,2,0} \le C\|b\|_{1,h}\|\!|c|\!\|_{2,2,0}$$

and (5.14) follows. ∎

A.2. Local truncation error analysis—proof of Lemma 6. In this subsection, we proceed with the local truncation error estimate. All the assertions in Lemmas 7 to 10 are pointwise estimates on the grid points (x_i, r_j), $j \ge 1$. For brevity, we omit the indices (i,j) whenever it is obvious.

We start with the estimates of the diffusion terms in (5.3).

LEMMA 7. *If $a \in C_s^4(R\times\overline{R^+})$ and $\alpha_0, \beta_0 \in R$, we have*

$$r|(\nabla_h^2 - \nabla^2)a| \le C\left(\Delta x^2 + \Delta r^2\right)\frac{1}{(1+r)^{\alpha_0}(1+|x|)^{\beta_0}}\|\!|a|\!\|_{4,\alpha_0+2,\beta_0} \tag{A.13}$$

and

$$|(\nabla_h^2 - \nabla^2)a| \le C\left(\Delta x^2 + \Delta r^2\right)\frac{1}{r}\frac{1}{(1+r)^{\alpha_0}(1+|x|)^{\beta_0}}\|\!|a|\!\|_{4,\alpha_0+2,\beta_0}. \tag{A.14}$$

Proof. Since $a \in C_s^4(R \times \overline{R^+})$, the odd extension of a given by

$$\tilde{a}(x,r) = \begin{cases} a(x,r), & \text{if } r \geq 0, \\ -a(x,-r), & \text{if } r < 0, \end{cases}$$

is in $C^4(R^2)$. It follows that

$$\begin{aligned} \nabla_h^2 a &= \left(D_x^2 + D_r^2 + \frac{\widetilde{D}_r}{r} \right) a \\ &= \nabla^2 a + \frac{1}{12}\Delta x^2 \partial_x^4 a|_{(\xi,r)} + \Delta r^2 \left(\frac{1}{12}\partial_r^4 a|_{(x,\eta_1)} + \frac{1}{6}\frac{1}{r}(\partial_r^3 a)_{(x,\eta_2)} \right) \end{aligned} \tag{A.15}$$

is valid for all $j \geq 1$ with $\xi \in (x - \Delta x, x + \Delta x)$ and $\eta_1, \eta_2 \in (r - \Delta r, r + \Delta r)$.

Thus

$$\begin{aligned} & r|(\nabla_h^2 - \nabla^2)a| \\ \leq\ & C\left(\Delta x^2 + \Delta r^2\right)\left(r|\partial_x^4(r\tfrac{a}{r})|_{(\xi,r)}| + r|\partial_r^4(r\tfrac{a}{r})|_{(x,\eta_1)}| + |\partial_r^3(r\tfrac{a}{r})|_{(x,\eta_2)}| \right) \\ \leq\ & C\left(\Delta x^2 + \Delta r^2\right)\Big(\tfrac{r\|a\|_{4,\alpha_0+2,\beta_0}}{(1+r)^{\alpha_0+1}(1+|\xi|)^{\beta_0}} + \tfrac{r(\|a\|_{4,\alpha_0+2,\beta_0}+\|a\|_{3,\alpha_0+1,\beta_0})}{(1+\eta_1)^{\alpha_0+1}(1+|x|)^{\beta_0}} \\ & \qquad + \tfrac{\|a\|_{3,\alpha_0+1,\beta_0}+\|a\|_{2,\alpha_0,\beta_0}}{(1+\eta_2)^{\alpha_0}(1+|x|)^{\beta_0}} \Big) \\ \leq\ & C\left(\Delta x^2 + \Delta r^2\right) \tfrac{1}{(1+r)^{\alpha_0}(1+|x|)^{\beta_0}} \|\!|a|\!\|_{4,\alpha_0+2,\beta_0}. \end{aligned}$$

This gives (A.13), together with (A.14) as a direct consequence. ∎

Next we proceed with the estimates for the Jacobians, starting with their typical factors.

LEMMA 8. *For $a \in C_s^4(R \times \overline{R^+})$, $\alpha, \beta \in R$, we have*

$$\widetilde{D}_x\left(\frac{a}{r}\right) = \partial_x\left(\frac{a}{r}\right) + O(1)\Delta x^2 \frac{1}{(1+r)^\alpha (1+|x|)^\beta} \|\!|a|\!\|_{3,\alpha,\beta}, \tag{A.16}$$

$$\widetilde{D}_x(ra) = \partial_x(ra) + O(1) r^2 \Delta x^2 \frac{1}{(1+r)^\alpha (1+|x|)^\beta} \|\!|a|\!\|_{3,\alpha,\beta}, \tag{A.17}$$

$$\widetilde{D}_r\left(\frac{a}{r}\right) = \partial_r\left(\frac{a}{r}\right) + O(1)\frac{\Delta r^2}{r^3} \frac{1}{(1+r)^\alpha (1+|x|)^\beta} \|\!|a|\!\|_{3,\alpha+3,\beta}, \tag{A.18}$$

$$\widetilde{D}_r(ra) = \partial_r(ra) + O(1)\frac{\Delta r^2}{r} \frac{1}{(1+r)^\alpha (1+|x|)^\beta} \|\!|a|\!\|_{3,\alpha+3,\beta}. \tag{A.19}$$

Proof. We begin with (A.16) and (A.17).

Since

$$(\widetilde{D}_x - \partial_x) f = \frac{\Delta x^2}{6} \partial_x^3 f_{|(\xi,r)}, \quad \xi \in (x - \Delta x, x + \Delta x),$$

it follows that

$$\left|(\widetilde{D}_x - \partial_x)\left(\frac{a}{r}\right)\right| = \frac{\Delta x^2}{6}\left|\partial_x^3\left(\frac{a}{r}\right)\right|_{|(\xi,r)} \leq C\Delta x^2 \frac{1}{(1+r)^\alpha (1+|x|)^\beta} \|a\|_{3,\alpha,\beta}$$

and

$$|(\widetilde{D}_x - \partial_x)(ra)| = \frac{\Delta x^2}{6}\left|\partial_x^3\left(r^2\frac{a}{r}\right)_{|(\xi,r)}\right| \le Cr^2\Delta x^2\frac{1}{(1+r)^\alpha(1+|x|)^\beta}\|a\|_{3,\alpha,\beta}.$$

For (A.18) and (A.19), the estimate is more complicated due to our reflection boundary condition (4.8) and (4.9). We estimate for $j > 1$ and $j = 1$ separately.

When $j > 1$, we have

$$(\widetilde{D}_r - \partial_r)f = \frac{1}{6}\Delta r^2\partial_r^3 f|_{(x,\eta)}, \quad \eta \in (r - \Delta r, r + \Delta r).$$

Therefore

$$\left|(\widetilde{D}_r - \partial_r)\left(\frac{a}{r}\right)\right| = \frac{\Delta r^2}{6}\left|\partial_r^3\left(\frac{a}{r}\right)\right|_{(x,\eta)} \le C\frac{\Delta r^2}{r^3}\frac{1}{(1+r)^\alpha(1+|x|)^\beta}\|a\|_{3,\alpha+3,\beta}$$

and

$$\begin{aligned}|(\widetilde{D}_r - \partial_r)(ra)| &\le C\Delta r^2\left|\partial_r^3(r^2\tfrac{a}{r})\right|_{(x,\eta)}\\ &\le C\tfrac{\Delta r^2}{r}\tfrac{1}{(1+r)^\alpha(1+|x|)^\beta}(\|a\|_{3,\alpha+3,\beta} + \|a\|_{2,\alpha+2,\beta} + \|a\|_{1,\alpha+1,\beta}).\end{aligned}$$

When $j = 1$, we have

$$\left|\partial_r\left(\frac{a}{r}\right)\right|_{j=1} = C\frac{\Delta r^2}{r_1^3}r_1\left|\partial_r\left(\frac{a}{r}\right)\right|_{j=1} \le C\frac{\Delta r^2}{r_1^3}\frac{1}{(1+r_1)^\alpha(1+|x|)^\beta}\|a\|_{1,\alpha+1,\beta}.$$

In addition, since $r_1 = \frac{\Delta r}{2}$, we apply (4.9) to get

$$\left|\widetilde{D}_r\left(\frac{a}{r}\right)\right|_{j=1} = \left|\frac{\frac{a_2}{r_2} + \frac{a_1}{r_1}}{2\Delta r}\right| = \left|C\frac{\Delta r^2}{r_1^3}\left(\frac{a_2}{r_2} + \frac{a_1}{r_1}\right)\right| \le C\frac{\Delta r^2}{r_1^3}\frac{1}{(1+r_1)^\alpha(1+|x|)^\beta}\|a\|_{0,\alpha,\beta},$$

and (A.18) follows.

(A.19) can be proved similarly, as follows:

$$\widetilde{D}_r(ra)_{j=1} = \frac{\frac{3}{2}\Delta r a_2 + \frac{1}{2}\Delta r a_1}{2\Delta r} = \frac{3}{4}a_2 + \frac{1}{4}a_1,$$
$$|a_1| \le C\frac{\Delta r^2}{r_1}\left|\frac{a_1}{r_1}\right| \le C\frac{\Delta r^2}{r_1}\frac{1}{(1+r_1)^\alpha(1+|x|)^\beta}\|a\|_{0,\alpha,\beta},$$

and

$$|a_2| \le C\frac{\Delta r^2}{r_1}\left|\frac{a_2}{r_2}\right| \le C\frac{\Delta r^2}{r_1}\frac{1}{(1+r_1)^\alpha(1+|x|)^\beta}\|a\|_{0,\alpha,\beta}.$$

Therefore

$$\left|\widetilde{D}_r(ra)\right|_{j=1} \le C\frac{\Delta r^2}{r_1}\frac{1}{(1+r_1)^\alpha(1+|x|)^\beta}\|a\|_{0,\alpha,\beta}.$$

In addition,

$$|\partial_r(ra)|_{j=1} \le \left(r^2\left|\partial_r\left(\frac{a}{r}\right)\right| + 2r\left|\frac{a}{r}\right|\right)_{j=1} \le C\frac{\Delta r^2}{r_1}\frac{\|a\|_{1,\alpha+1,\beta} + \|a\|_{0,\alpha,\beta}}{(1+r_1)^\alpha(1+|x|)^\beta},$$

and (A.19) follows. ∎

We now continue with the pointwise estimate for the Jacobi terms $\frac{1}{r}|J_h(ra,rb) - J(ra,rb)|$ and $r|J_h\left(\frac{a}{r},rb\right) - J\left(\frac{a}{r},rb\right)|$. Since

(A.20)
$$\begin{aligned}\tfrac{3}{r^2}J_h(ra,rb) = \;& \widetilde{D}_x(\tfrac{a}{r})\widetilde{D}_r(rb) - \widetilde{D}_r(ra)\widetilde{D}_x(\tfrac{b}{r}) + \widetilde{D}_x\left(\tfrac{a}{r}\widetilde{D}_r(rb) - \tfrac{b}{r}\widetilde{D}_r(ra)\right)\\ & + \tfrac{1}{r^2}\widetilde{D}_r\left(r^2 b\widetilde{D}_x a - r^2 a\widetilde{D}_x b\right),\end{aligned}$$

(A.21)
$$\begin{aligned}3J_h\left(\tfrac{a}{r},rb\right) = \;& \widetilde{D}_x(\tfrac{a}{r})\widetilde{D}_r(rb) - \widetilde{D}_r(\tfrac{a}{r})\widetilde{D}_x(rb) + \widetilde{D}_x\left(\tfrac{a}{r}\widetilde{D}_r(rb) - rb\widetilde{D}_r(\tfrac{a}{r})\right)\\ & + \widetilde{D}_r\left(b\widetilde{D}_x a - a\widetilde{D}_x b\right),\end{aligned}$$

it suffices to estimate the terms in (A.20) and (A.21) individually. We summarize them as the following lemma.

LEMMA 9. *If $a, b \in C_s^4(R \times \overline{R^+})$ and $\alpha_1, \alpha_2, \beta_1, \beta_2 \in R$, then*

(A.22)
$$\begin{aligned}& r|\widetilde{D}_r(\tfrac{a}{r})\widetilde{D}_x(rb) - \partial_r(\tfrac{a}{r})\partial_x(rb)| + \tfrac{1}{r}|\widetilde{D}_r(rb)\widetilde{D}_x(ra) - \partial_r(rb)\partial_x(ra)|\\ & \le C(\Delta x^2 + \Delta r^2)\tfrac{1}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{3,\alpha_2+\frac{5}{2},\beta_2},\end{aligned}$$

(A.23)
$$\begin{aligned}& r|\widetilde{D}_x(\tfrac{a}{r}\widetilde{D}_r(rb)) - \partial_x(\tfrac{a}{r}\partial_r(rb))| + r|\widetilde{D}_x(ra\widetilde{D}_r(\tfrac{b}{r})) - \partial_x(ra\partial_r(\tfrac{b}{r}))|\\ & \le C(\Delta x^2 + \Delta r^2)\tfrac{1}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2},\end{aligned}$$

(A.24)
$$\begin{aligned}& r|\widetilde{D}_r(a\widetilde{D}_x b) - \partial_r(a\partial_x b)| + \tfrac{1}{r}|\widetilde{D}_r(r^2 a\widetilde{D}_x b) - \partial_r(r^2 a\partial_x b)|\\ & \le C(\Delta x^2 + \Delta r^2)\tfrac{1}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}.\end{aligned}$$

Proof. Since (A.16)–(A.19) are valid for any α, $\beta \in R$, we have

(A.25)
$$\widetilde{D}_x\left(\frac{a}{r}\right) = \partial_x\left(\frac{a}{r}\right) + O(1)\Delta x^2\frac{1}{(1+r)^{\alpha_1+\lambda}(1+|x|)^{\beta_1}}\|\!|a|\!\|_{3,\alpha_1+\lambda,\beta_1},$$

(A.26)
$$\widetilde{D}_x(ra) = \partial_x(ra) + O(1)r^2\Delta x^2\frac{1}{(1+r)^{\alpha_1+\lambda}(1+|x|)^{\beta_1}}\|\!|a|\!\|_{3,\alpha_1+\lambda,\beta_2},$$

(A.27)
$$\widetilde{D}_r\left(\frac{a}{r}\right) = \partial_r\left(\frac{a}{r}\right) + O(1)\frac{\Delta r^2}{r^3}\frac{1}{(1+r)^{\alpha_1+\lambda}(1+|x|)^{\beta_1}}\|\!|a|\!\|_{3,\alpha_1+\lambda+3,\beta_1},$$

(A.28)
$$\widetilde{D}_r(ra) = \partial_r(ra) + O(1)\frac{\Delta r^2}{r}\frac{1}{(1+r)^{\alpha_1+\lambda}(1+|x|)^{\beta_1}}\|\!|a|\!\|_{3,\alpha_1+\lambda+3,\beta_1},$$

and

(A.29)
$$\widetilde{D}_x\left(\frac{b}{r}\right) = \partial_x\left(\frac{b}{r}\right) + O(1)\Delta x^2\frac{1}{(1+r)^{\alpha_2+\mu}(1+|x|)^{\beta_2}}\|\!|b|\!\|_{3,\alpha_2+\mu,\beta_2},$$

(A.30)
$$\widetilde{D}_x(rb) = \partial_x(rb) + O(1)r^2\Delta x^2\frac{1}{(1+r)^{\alpha_2+\mu}(1+|x|)^{\beta_2}}\|\!|b|\!\|_{3,\alpha_2+\mu,\beta_2},$$

(A.31)
$$\widetilde{D}_r\left(\frac{b}{r}\right) = \partial_r\left(\frac{b}{r}\right) + O(1)\frac{\Delta r^2}{r^3}\frac{1}{(1+r)^{\alpha_2+\mu}(1+|x|)^{\beta_2}}\|\!|b|\!\|_{3,\alpha_2+\mu+3,\beta_2},$$

(A.32)
$$\widetilde{D}_r(rb) = \partial_r(rb) + O(1)\frac{\Delta r^2}{r}\frac{1}{(1+r)^{\alpha_2+\mu}(1+|x|)^{\beta_2}}\|\!|b|\!\|_{3,\alpha_2+\mu+3,\beta_2}$$

for any λ, $\mu \in R$. We apply (A.27), (A.30) with $\lambda = -\frac{1}{2}$, $\mu = \frac{5}{2}$ to get

$$\begin{aligned}
& r|\widetilde{D}_r(\tfrac{a}{r})\widetilde{D}_x(rb) - \partial_r(\tfrac{a}{r})\partial_x(rb)| \\
= \;& r|\widetilde{D}_r(\tfrac{a}{r})\widetilde{D}_x(rb) - \partial_r(\tfrac{a}{r})\widetilde{D}_x(rb) + \partial_r(\tfrac{a}{r})\widetilde{D}_x(rb) - \partial_r(\tfrac{a}{r})\partial_x(rb)| \\
= \;& O(|\widetilde{D}_x(rb)|)\tfrac{\Delta r^2}{r^2}\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}}{(1+r)^{\alpha_1-\frac{1}{2}}(1+|x|)^{\beta_1}} + O(|\partial_r(\tfrac{a}{r})|)r^3\Delta x^2\frac{\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_2+\frac{5}{2}}(1+|x|)^{\beta_2}}.
\end{aligned}$$

Moreover, since

$$r^3|\partial_r\left(\frac{a}{r}\right)| \le \frac{1}{(1+r)^{\alpha_1-\frac{5}{2}}(1+|x|)^{\beta_1}}\|a\|_{1,\alpha_1+\frac{1}{2},\beta_1}$$

and

$$|\widetilde{D}_x(rb)| = |\partial_x(rb)(\xi,r)| \le r^2\frac{1}{(1+r)^{\alpha_2+\frac{1}{2}}(1+|x|)^{\beta_2}}\|b\|_{1,\alpha_2+\frac{1}{2},\beta_2},$$

it follows that

$$\begin{aligned}
& r|\widetilde{D}_r(\tfrac{a}{r})\widetilde{D}_x(rb) - \partial_r(\tfrac{a}{r})\partial_x(rb)| \\
\le \;& C(\Delta x^2+\Delta r^2)\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{1,\alpha_2+\frac{1}{2},\beta_2}+\|a\|_{1,\alpha_1+\frac{1}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \qquad \text{(A.33)} \\
\le \;& C(\Delta x^2+\Delta r^2)\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}.
\end{aligned}$$

Similarly, from (A.32) and (A.25), we have

$$\begin{aligned}
& r|\widetilde{D}_x(\tfrac{a}{r})\widetilde{D}_r(rb) - \partial_x(\tfrac{a}{r})\partial_r(rb)| \\
= \;& r|\widetilde{D}_x(\tfrac{a}{r})\widetilde{D}_r(rb) - \widetilde{D}_x(\tfrac{a}{r})\partial_r(rb) + \widetilde{D}_x(\tfrac{a}{r})\partial_r(rb) - \partial_x(\tfrac{a}{r})\partial_r(rb)| \\
= \;& O(|\widetilde{D}_x(\tfrac{a}{r})|)\Delta r^2\frac{\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_2-\frac{1}{2}}(1+|x|)^{\beta_2}} + O(|\partial_r(rb)|)r\Delta x^2\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}}{(1+r)^{\alpha_1+\frac{5}{2}}(1+|x|)^{\beta_1}} \qquad \text{(A.34)} \\
\le \;& C\Delta r^2\frac{\|a\|_{1,\alpha_1+\frac{1}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} + C\Delta x^2\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{1,\alpha_2+\frac{1}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \\
\le \;& C(\Delta x^2+\Delta r^2)\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}.
\end{aligned}$$

The estimate (A.22) then follows from (A.33) and (A.34).

For (A.23), we have

$$\begin{aligned}
& \widetilde{D}_x(f\widetilde{D}_r g) - \partial_x(f\partial_r g) \qquad \text{(A.35)} \\
= \;& \widetilde{D}_x(f(\widetilde{D}_r-\partial_r)g) + (\widetilde{D}_x-\partial_x)(f\partial_r g) \\
= \;& \partial_x(f(\widetilde{D}_r-\partial_r)g)_{|(\xi_1,r)} + \tfrac{1}{6}\Delta x^2\partial_x^3(f\partial_r g)_{|(\xi_2,\eta)} \\
= \;& (\partial_x f)((\widetilde{D}_r-\partial_r)g)_{|(\xi_1,r)} + f((\widetilde{D}_r-\partial_r)\partial_x g)_{|(\xi_1,r)} + \tfrac{1}{6}\Delta x^2\partial_x^3(f\partial_r g)_{|(\xi_2,\eta)}.
\end{aligned}$$

We proceed with individual terms in (A.35), taking $f = \frac{a}{r}$ and $g = rb$. From (A.30) with $\mu = -\frac{1}{2}$, we have

$$\begin{aligned} r\left|(\partial_x \tfrac{a}{r})(\widetilde{D}_r - \partial_r)(rb)\right| &\le C|\partial_x(\tfrac{a}{r})|\Delta r^2 \tfrac{1}{(1+r)^{\alpha_2-\frac{1}{2}}(1+|x|)^{\beta_2}} \|\!|b|\!\|_{3,\alpha_2+\frac{5}{2},\beta_2} \\ &\le C\Delta r^2 \tfrac{1}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \|\!|a|\!\|_{1,\alpha_1+\frac{1}{2},\beta_1} \|\!|b|\!\|_{3,\alpha_2+\frac{5}{2},\beta_2}. \end{aligned}$$

Similarly, from (A.32)

$$\begin{aligned} & r\left|\tfrac{a}{r}(\widetilde{D}_r - \partial_r)\partial_x(rb)\right| \\ \le\ & C\Delta r^2 |\tfrac{a}{r}| \tfrac{1}{(1+r)^{\alpha_2+\frac{1}{2}}(1+|x|)^{\beta_2}} \|\!|\partial_x b|\!\|_{3,\alpha_2+\frac{7}{2},\beta_2} \\ \le\ & C\Delta r^2 \tfrac{1}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \|a\|_{0,\alpha_1-\frac{1}{2},\beta_1} \|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}, \\ & r\left|\Delta x^2 \partial_x^3 (\tfrac{a}{r}\partial_r(rb))_{|(x,\eta)}\right| \\ \le\ & C\Delta x^2 \left|r\partial_x^3\left(\tfrac{a}{r}\right)\partial_r(rb) + r\left(\tfrac{a}{r}\right)\partial_x^3\partial_r(rb)\right| \\ \le\ & C\Delta x^2 \tfrac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{1,\alpha_2+\frac{1}{2},\beta_2} + \|a\|_{0,\alpha_1-\frac{1}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \\ \le\ & C\Delta x^2 \tfrac{\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \end{aligned}$$

Therefore

$$r|\widetilde{D}_x\left(\frac{a}{r}\widetilde{D}_r(rb)\right) - \partial_x\left(\frac{a}{r}\partial_r(rb)\right)| \le C(\Delta x^2 + \Delta r^2)\frac{\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}.$$

Using the same argument as above, one can derive

$$r|\widetilde{D}_x\left(ra\widetilde{D}_r\left(\frac{b}{r}\right)\right) - \partial_x\left(ra\partial_r\left(\frac{b}{r}\right)\right)| \le C(\Delta x^2 + \Delta r^2)\frac{\|\!|a|\!\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}$$

and therefore (A.23) is proved.

We continue with (A.24). For the first term, we can write

$$\widetilde{D}_r(a\widetilde{D}_x b) - \partial_r(a\partial_x b) = \widetilde{D}_r(a(\widetilde{D}_x - \partial_x)b) + (\widetilde{D}_r - \partial_r)(a\partial_x b).$$

Since $a, b \in C_s^4(R \times \overline{R^+})$, by extending a, b to odd functions across $r = 0$, we see that the extended $a\widetilde{D}_x b$ is in $C^4(R^2)$; thus

$$\begin{aligned} \widetilde{D}_r(a(\widetilde{D}_x - \partial_x)b) &= \partial_r(a(\widetilde{D}_x - \partial_x)b)|_{(x,\eta)} \\ &= \left(\partial_r a(\widetilde{D}_x - \partial_x)b + a(\widetilde{D}_x - \partial_x)(\partial_r b)\right)|_{(x,\eta)} \\ &= \tfrac{\Delta x^2}{6}\left(\partial_r a|_{(x,\eta)}\partial_x^3 b|_{(\xi_1,\eta)} + a|_{(x,\eta)}\partial_x^3\partial_r b|_{(\xi_2,\eta)}\right) \end{aligned}$$

and therefore

$$r|\widetilde{D}_r(a(\widetilde{D}_x - \partial_x)b)| \le C\Delta x^2 \frac{\|\!|a|\!\|_{1,\alpha_1+\frac{1}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \tag{A.36}$$

Similarly, the extended $a\partial_x b$ is in $C^3(R^2)$, and thus we have

$$r|(\widetilde{D}_r-\partial_r)(a\partial_x b)| = r\frac{\Delta r^2}{6}\partial_r^3(a\partial_x b)|_{(x,\eta)} \le C\Delta r^2 \frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \tag{A.37}$$

From (A.36) and (A.37), we have the following estimate for the first term of (A.24):

$$r|\widetilde{D}_r(a\widetilde{D}_x b) - \partial_r(a\partial_x b)| \le C(\Delta x^2+\Delta r^2)\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \tag{A.38}$$

The second term in (A.24) can be treated similarly, as follows:

$$\frac{1}{r}\widetilde{D}_r(r^2a\widetilde{D}_x b) - \frac{1}{r}\partial_r(r^2a\partial_x b) = \frac{1}{r}\widetilde{D}_r(r^2a(\widetilde{D}_x-\partial_x)b) + \frac{1}{r}(\widetilde{D}_r-\partial_r)(r^2a\partial_x b). \tag{A.39}$$

Again, since the extensions of $r^2a(\widetilde{D}_x-\partial_x)b$ and $r^2a\partial_x b$ are both in $C^3(R^2)$, we can directly estimate these two terms by

(A.40)

$$\begin{aligned}
\tfrac{1}{r}\widetilde{D}_r(r^2a(\widetilde{D}_x-\partial_x)b) &= \tfrac{1}{r}\partial_r(r^2a(\widetilde{D}_x-\partial_x)b)_{(x,\eta)} \\
&= \tfrac{1}{r}\left(\Big(\partial_r(r^2a)(\widetilde{D}_x-\partial_x)b\Big)_{(x,\eta)} + \Big(r^2a(\widetilde{D}_x-\partial_x)(\partial_r b)\Big)_{(x,\eta)}\right) \\
&= C\Delta x^2\left(\big((r\partial_r a+2a)\partial_x^3 b\big)_{(\xi_1,\eta)} + \big(ra\partial_x^3(\partial_r b)\big)_{(\xi_2,\eta)}\right)
\end{aligned}$$

and

$$\frac{1}{r}(\widetilde{D}_r-\partial_r)(r^2a\partial_x b) = \frac{\Delta r^2}{r}\partial_r^3(r^2a\partial_x b)_{(x,\eta)} = \frac{\Delta r^2}{r}\partial_r^3\left(r^4\frac{a}{r}\partial_x(\frac{b}{r})\right)_{(x,\eta)}. \tag{A.41}$$

From (A.40) and (A.41), we have

$$\begin{aligned}
&|\tfrac{1}{r}\widetilde{D}_r(r^2a(\widetilde{D}_x-\partial_x)b)| \\
&\le C\Delta x^2\frac{\|a\|_{1,\alpha_1+\frac{1}{2},\beta_1}\|b\|_{3,\alpha_2+\frac{5}{2},\beta_2}+\|a\|_{0,\alpha_1-\frac{1}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}} \qquad \text{(A.42)} \\
&\le C\Delta x^2\frac{\|a\|_{1,\alpha_1+\frac{1}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}
\end{aligned}$$

and

$$\frac{1}{r}\left|(\widetilde{D}_r-\partial_r)(r^2a\partial_x b)\right| \le C\Delta r^2\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \tag{A.43}$$

From (A.39), (A.42), and (A.43), we conclude that

$$\left|\frac{1}{r}\widetilde{D}_r(r^2a\widetilde{D}_x b) - \frac{1}{r}\partial_r(r^2a\partial_x b)\right| \le C(\Delta x^2+\Delta r^2)\frac{\|a\|_{3,\alpha_1+\frac{5}{2},\beta_1}\|b\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}. \tag{A.44}$$

The estimates (A.38) and (A.44) imply (A.24). Thus the proof of Lemma 9 is completed. ∎

As a direct consequence of Lemma 9, we have the following pointwise estimate for the Jacobians.

LEMMA 10. *If* $a, b \in C_s^4(R \times \overline{R^+})$, *then*

$$\frac{1}{r}|J_h(ra,rb) - J(ra,rb)| \le C(\Delta x^2 + \Delta r^2)\frac{\|\!|a|\!\|_{4,\alpha_1+\frac{7}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}},$$

$$r|J_h\left(\frac{a}{r},rb\right) - J\left(\frac{a}{r},rb\right)| \le C(\Delta x^2 + \Delta r^2)\frac{\|\!|a|\!\|_{4,\alpha_1+\frac{7}{2},\beta_1}\|\!|b|\!\|_{4,\alpha_2+\frac{7}{2},\beta_2}}{(1+r)^{\alpha_1+\alpha_2}(1+|x|)^{\beta_1+\beta_2}}$$

for any $\alpha_1, \alpha_2, \beta_1, \beta_2 \in R$.

From (5.3), Lemma 7, and Lemma 10, we can easily derive (5.17)–(5.20). This completes the proof of Lemma 6. □

Acknowledgments. The authors would like to thank the anonymous referees and Dr. YinLiang Huang for their valuable suggestions that helped to improve this paper.

REFERENCES

[1] A. ARAKAWA, *Computational design for long-term numerical integration of the equations of fluid motion: Two dimensional incompressible flow. Part* I, J. Comput. Phys., 1 (1966), pp. 119–143.

[2] G. K. BATCHELOR, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1999.

[3] L. CAFFARELLI, R. KOHN, AND L. NIRENBERG, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[4] D. CHAE AND J. LEE, *On the regularity of axisymmetric solutions of the Navier-Stokes equations*, Math. Z., 239 (2002), pp. 645–671.

[5] R. GRAUER AND T. C. SIDERIS, *Numerical computation of* 3*D incompressible ideal fluids with swirl*, Phys. Rev. Lett., 67 (1991), pp. 3511–3514.

[6] R. GRAUER AND T. C. SIDERIS, *Finite time singularities in ideal fluids with swirl*, Phys. D, 88 (1995), pp. 116–132.

[7] T. Y. HOU AND B. T. R. WETTON, *Convergence of a finite difference scheme for the Navier–Stokes equations using vorticity boundary conditions*, SIAM J. Numer. Anal., 29 (1992), pp. 615–639.

[8] T. KATO, *Nonstationary flows of viscous and ideal fluids in* R^3, J. Funct. Anal., 9 (1972), pp. 296–305.

[9] O. LADYZHENSKAYA, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[10] P. L. LIONS, *Mathematical Topics in Fluid Mechanics, Volume 1: Incompressible Models*, Oxford Lecture Ser. Math. Appl., Oxford University Press, New York, 1996.

[11] J.-G. LIU AND W. C. WANG, *Energy and helicity preserving schemes for hydro- and magnetohydro-dynamics flows with symmetry*, J. Comput. Phys., 200 (2004), pp. 8–33.

[12] J.-G. LIU AND W. C. WANG, *Characterization and Regularity of Axisymmetric Solenoidal Vector Fields with Application to Navier-Stokes Equation*, preprint, 2006.

[13] P. E. MERILEES, *The pseudospectral approximation applied to the shallow water equations on a sphere*, Atmosphere, 11 (1973), pp. 13–20.

[14] Y. MORINISHI, T. S. LUND, O. V. VASILYEV, AND P. MOIN, *Fully conservative higher order finite difference schemes for incompressible flow*, J. Comput. Phys., 143 (1998), pp. 90–124.

[15] P. OLSSON, *Summation by parts, projections, and stability.* I, Math. Comp., 64 (1995), pp. 1035–1065.

[16] S. A. PIACSEK AND G. P. WILLIAMS, *Conservation properties of convection difference schemes*, J. Comput. Phys., 6 (1970), pp. 392–405.

[17] P. SAFFMAN, *Vortex Dynamics*, Cambridge University Press, Cambridge, UK, 1992.

[18] R. W. C. P. VERSTAPPEN AND A. E. P. VELDMAN, *Symmetry-preserving discretization of turbulent flow*, J. Comput. Phys., 187 (2003), pp. 343–368.

[19] C. WANG AND J.-G. LIU, *Analysis of finite difference schemes for unsteady Navier-Stokes equations in vorticity formulation*, Numer. Math., 91 (2002), pp. 543–576.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2481–2504

A RESTARTED KRYLOV SUBSPACE METHOD FOR THE EVALUATION OF MATRIX FUNCTIONS*

MICHAEL EIERMANN† AND OLIVER G. ERNST†

Abstract. We show how the Arnoldi algorithm for approximating a function of a matrix times a vector can be restarted in a manner analogous to restarted Krylov subspace methods for solving linear systems of equations. The resulting restarted algorithm reduces to other known algorithms for the reciprocal and the exponential functions. We further show that the restarted algorithm inherits the superlinear convergence property of its unrestarted counterpart for entire functions and present the results of numerical experiments.

Key words. matrix function, Krylov subspace approximation, Krylov projection method, restarted Krylov subspace method, linear system of equations, initial value problem

AMS subject classifications. 65F10, 65F99, 65M20

DOI. 10.1137/050633846

1. Introduction.

The evaluation of

$$f(A)\boldsymbol{b}, \qquad \text{where } A \in \mathbb{C}^{n\times n}, \boldsymbol{b} \in \mathbb{C}^n, \tag{1.1}$$

and $f : \mathbb{C} \supset D \to \mathbb{C}$ is a function for which $f(A)$ is defined, is a common computational task. Besides the solution of linear systems of equations, which involves the reciprocal function $f(\lambda) = 1/\lambda$, by far the most important application is the time evolution of a system under a linear operator, in which case $f(\lambda) = f_t(\lambda) = e^{t\lambda}$ and time acts as a parameter t. Other applications involving differential equations require the evaluation of (1.1) for the square root and trigonometric functions (see [8, 1]). Further applications include identification problems for semigroups involving the logarithm (see, e.g., [29]) and lattice quantum chromodynamics simulations requiring the evaluation of the matrix sign function (see [34] and the references therein).

In many of the applications mentioned above the matrix A is large and sparse or structured, typically resulting from discretization of an infinite-dimensional operator. In this case evaluating (1.1) by first computing $f(A)$ is usually unfeasible, so that most of the algorithms for the latter task (see, e.g., [18, 5]) cannot be used. The standard approach for approximating (1.1) directly is based on a Krylov subspace of A with initial vector $\boldsymbol{b}$ [8, 9, 28, 14, 17, 4, 19]. The advantage of this approach is that it requires A only for computing matrix-vector products and that, for smooth functions such as the exponential, it converges superlinearly [8, 28, 31, 17].

One shortcoming of the Krylov subspace approximation, however, lies in the fact that computing an approximation of (1.1) from a Krylov subspace $\mathcal{K}_m(A, \boldsymbol{b})$ of dimension m involves all the basis vectors of $\mathcal{K}_m(A, \boldsymbol{b})$, and hence these need to be stored. Memory constraints therefore often limit the size of the problem that can be solved, which is an issue especially when A is the discrete representation of a partial differential operator in three space dimensions. When A is Hermitian, the Hermitian Lanczos process allows the basis of $\mathcal{K}_m(A, \boldsymbol{b})$ to be constructed by a three-term recurrence. When solving linear systems of equations, this recurrence for the basis vectors

*Received by the editors June 17, 2005; accepted for publication (in revised form) June 13, 2006; published electronically December 1, 2006.
http://www.siam.org/journals/sinum/44-6/63384.html
†Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, D-09596 Freiberg, Germany (eiermann@math.tu-freiberg.de, ernst@math.tu-freiberg.de).

immediately translates to efficient update formulas for the approximation. This, however, is a consequence of the simple form of the reciprocal function and such update formulas are not available for general (nonrational) functions.

When solving non-Hermitian linear systems of equations by Krylov subspace approximation, a common remedy is to limit storage requirements by restarting the algorithm each time the Krylov space has reached a certain maximal dimension [26, 10]. The subject of this work is the extension of this restarting approach to general functions. How such a generalization may be accomplished is not immediately obvious, since the restarting approach for linear systems is based on solving successive residual equations to obtain corrections to the most recent approximation. The availability of a residual, however, is another property specific to problems where $f(A)\boldsymbol{b}$ solves an (algebraic or differential) equation.

The remainder of this paper is organized as follows: Section 2 recalls the definition and properties of matrix functions and their approximation in Krylov spaces, emphasizing the role of Hermite interpolation, and closes with an error representation formula for Krylov subspace approximations. In section 3 we introduce a new restarted Krylov subspace algorithm for the approximation of (1.1) for general functions f. We derive two mathematically equivalent formulations of the restarted algorithm, the second of which, while slightly more expensive, was found to be more stable in the presence of rounding errors.

In section 4 we show that, for the reciprocal and exponential functions, our restarted method reduces to the restarted full orthogonalization method (FOM; see [27]) and is closely related to an algorithm by Celledoni and Moret [4], respectively. We further establish that, for entire functions of order one (such as the exponential function), the superlinear convergence property of the Arnoldi/Lanczos approximation of (1.1) is retained by our restarted method. In section 5 we demonstrate the performance of the restarted method for several test problems.

2. Matrix functions and their Krylov subspace approximation. In this section we fix notation, provide some background material on functions of matrices and their approximation using Krylov subspaces, highlight the connection with Hermite interpolation, and derive a new representation formula for the error of Krylov subspace approximations of $f(A)\boldsymbol{b}$.

2.1. Functions of matrices. We recall the definition of functions of matrices (as given, e.g., in Gantmacher [15, Chapter 5]): Let $\Lambda(A) = \{\lambda_1, \lambda_2, \dots, \lambda_k\}$ denote the k distinct eigenvalues of $A \in \mathbb{C}^{n\times n}$ and let the minimal polynomial of A be given by

$$m_A(\lambda) = \prod_{j=1}^{k} (\lambda - \lambda_j)^{n_j} \in \mathcal{P}_K, \qquad \text{where } K = \sum_{j=1}^{k} n_j.$$

Given a complex-valued function f, the matrix $f(A)$ is defined if $f^{(r)}(\lambda_j)$ exists for $r = 0, 1, \dots, n_j - 1$; $j = 1, 2, \dots, k$. In this case $f(A) := q_{f,A}(A)$, where $q_{f,A} \in \mathcal{P}_{K-1}$ denotes the unique polynomial of degree at most $K-1$ which satisfies the K Hermite interpolation conditions

$$q_{f,A}^{(r)}(\lambda_j) = f^{(r)}(\lambda_j), \qquad r = 0, 1, \dots, n_j - 1,\ j = 1, 2, \dots, k. \tag{2.1}$$

In the remainder of the paper, we denote the unique polynomial q which interpolates f in the Hermite sense at a set of nodes $\{\vartheta_j\}_{j=1}^{k}$ with multiplicities n_j by $I_p f \in \mathcal{P}_{K-1}$,

$K = \sum_j n_j$, where $p \in \mathcal{P}_K$ is a (not necessarily monic) nodal polynomial with zeros ϑ_j of multiplicities n_j. In this notation, (2.1) reads

$$q_{f,A} = I_{m_A} f.$$

Our objective is the evaluation of $f(A)\boldsymbol{b}$ rather than $f(A)$, and this can possibly be achieved with polynomials of lower degree than $q_{f,A}$. To this end, let the minimal polynomial of $\boldsymbol{b} \in \mathbb{C}^n$ with respect to A be given by

$$m_{A,\boldsymbol{b}}(\lambda) = \prod_{j=1}^{\ell} (\lambda - \lambda_j)^{m_j} \in \mathcal{P}_L, \qquad \text{where } L = L(A,\boldsymbol{b}) = \sum_{j=1}^{\ell} m_j. \tag{2.2}$$

PROPOSITION 2.1. *Given a function f, a matrix $A \in \mathbb{C}^{n\times n}$ such that $f(A)$ is defined, and a vector $\boldsymbol{b} \in \mathbb{C}^n$ whose minimal polynomial with respect to A is given by* (2.2)*, there holds $f(A)\boldsymbol{b} = q_{f,A,\boldsymbol{b}}(A)\boldsymbol{b}$, where $q_{f,A,\boldsymbol{b}} := I_{m_{A,\boldsymbol{b}}} f \in \mathcal{P}_{L-1}$ denotes the unique Hermite interpolating polynomial determined by the conditions*

$$q_{f,A,\boldsymbol{b}}^{(r)}(\lambda_j) = f^{(r)}(\lambda_j), \qquad r = 0, 1, \ldots, m_j - 1,\ j = 1, 2, \ldots, \ell.$$

2.2. Krylov subspace approximations. We recall the definition of the mth Krylov subspace of $A \in \mathbb{C}^{n\times n}$ and $\mathbf{0} \neq \boldsymbol{b} \in \mathbb{C}^n$ given by

$$\mathcal{K}_m(A,\boldsymbol{b}) := \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, ..., A^{m-1}\boldsymbol{b}\} = \{q(A)\boldsymbol{b} \,:\, q \in \mathcal{P}_{m-1}\}.$$

By Proposition 2.1, $f(A)\boldsymbol{b}$ lies in $\mathcal{K}_L(A,\boldsymbol{b})$. The index $L = L(A,\boldsymbol{b}) \in \mathbb{N}$ (cf. (2.2)) is the smallest number for which $\mathcal{K}_L(A,\boldsymbol{b}) = \mathcal{K}_{L+1}(A,\boldsymbol{b})$. Note that for certain functions such as $f(\lambda) = 1/\lambda$, we have $f(A)\boldsymbol{b} \in \mathcal{K}_L(A,\boldsymbol{b}) \setminus \mathcal{K}_{L-1}(A,\boldsymbol{b})$; in general, however, $f(A)\boldsymbol{b}$ may lie in a space $\mathcal{K}_m(A,\boldsymbol{b})$ with $m < L$.[1]

In what follows, we consider a sequence of approximations $\boldsymbol{y}_m := q(A)\boldsymbol{b} \in \mathcal{K}_m(A,\boldsymbol{b})$ to $f(A)\boldsymbol{b}$ with polynomials $q \in \mathcal{P}_{m-1}$ which in some sense approximate f. The most popular of these approaches (see [28, 14, 17]), to which we shall refer as the *Arnoldi approximation*, is based on the Arnoldi decomposition of $\mathcal{K}_m(A,\boldsymbol{b})$,

$$AV_m = V_{m+1}\widetilde{H}_m = V_m H_m + \eta_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^\top. \tag{2.3}$$

Here, the columns of $V_m = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$ form an orthonormal basis of $\mathcal{K}_m(A,\boldsymbol{b})$ with $\boldsymbol{v}_1 = \boldsymbol{b}/\|\boldsymbol{b}\|$, $\widetilde{H}_m = [\eta_{j,\ell}] \in \mathbb{C}^{(m+1)\times m}$ as well as $H_m := [I_m, \mathbf{0}]\,\widetilde{H}_m \in \mathbb{C}^{m\times m}$ are unreduced upper Hessenberg matrices, and $\boldsymbol{e}_m \in \mathbb{R}^m$ denotes the mth unit coordinate vector. The Arnoldi approximation to $f(A)\boldsymbol{b}$ is then defined by

$$\boldsymbol{f}_m := \beta V_m f(H_m)\boldsymbol{e}_1, \qquad \text{where } \beta = \|\boldsymbol{b}\|.$$

The rationale behind this approximation is that H_m represents the compression of A onto $\mathcal{K}_m(A,\boldsymbol{b})$ with respect to the basis V_m and that $\boldsymbol{b} = \beta V_m \boldsymbol{e}_1$.

The non-Hermitian (or two-sided) Lanczos algorithm is another procedure for generating a decomposition of the form (2.3). In that case the columns of V_m still form a basis of $\mathcal{K}_m(A,\boldsymbol{b})$, albeit one that is, in general, not orthogonal, and the upper Hessenberg matrices $\widetilde{H}_m$ are tridiagonal (or block tridiagonal if a look-ahead

[1] For the exponential function it was shown in [28, Theorem 3.6] that $e^{tA}\boldsymbol{b} \in \mathcal{K}_m(A,\boldsymbol{b})$ for all $t \in \mathbb{R}$ if and only if $m \geq L$.

technique is employed). The associated approximation to $f(A)\boldsymbol{b}$ is again defined by $\boldsymbol{f}_m := \beta V_m f(H_m)\boldsymbol{e}_1$ (see, e.g., [14, 28, 17]). Both approximations $\boldsymbol{f}_m$, based either on the Arnoldi or Lanczos decomposition, result from an interpolation procedure: If $q \in \mathcal{P}_{m-1}$ denotes the polynomial which interpolates f in the Hermite sense on the spectrum of H_m (counting multiplicities), then

$$\boldsymbol{f}_m = \beta V_m f(H_m)\boldsymbol{e}_1 = \beta V_m q(H_m)\boldsymbol{e}_1 = q(A)\boldsymbol{b}, \qquad m = 1, 2, \ldots, L$$

(see [28, Theorem 3.3]).

For later applications, next we show that similar results hold true for more general decompositions of $\mathcal{K}_m(A, \boldsymbol{b})$. To this end, we introduce a sequence of ascending (not necessarily orthonormal) basis vectors $\{\boldsymbol{w}_m\}_{m=1}^L$ such that

$$\mathcal{K}_m(A, \boldsymbol{b}) = \operatorname{span}\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}, \qquad m = 1, 2, \ldots, L. \tag{2.4}$$

As is well known, there exists a unique unreduced upper Hessenberg matrix $H = [\eta_{j,m}] \in \mathbb{C}^{L\times L}$ such that, with $W := [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_L] \in \mathbb{C}^{n\times L}$, there holds $AW = WH$ and, for $m = 1, 2, \ldots, L-1$, we have

$$AW_m = W_{m+1}\widetilde{H}_m = W_m H_m + \eta_{m+1,m}\boldsymbol{w}_{m+1}\boldsymbol{e}_m^\top, \tag{2.5}$$

where $\widetilde{H}_m$ is the $(m+1)\times m$ leading submatrix of H, $H_m := [I_m, \boldsymbol{0}]\widetilde{H}_m$, and $W_m = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m]$. We shall refer to (2.5) as an *Arnoldi-like decomposition*[2] to distinguish it from a proper Arnoldi decomposition (2.3). We shall require the following lemma, which is a simple generalization of the corresponding result for (proper) Arnoldi decompositions (cf. [28, 23]).

LEMMA 2.2. *For any polynomial $q(\lambda) = \alpha_m\lambda^m + \cdots + \alpha_1\lambda + \alpha_0 \in \mathcal{P}_m$, the vector $q(A)\boldsymbol{b}$ may be represented in terms of the Arnoldi-like decomposition* (2.5) *as*

$$q(A)\boldsymbol{b} = \begin{cases} \beta\big[W_m q(H_m)\boldsymbol{e}_1 + \alpha_m\gamma_m\boldsymbol{w}_{m+1}\big], & m < L, \\ \beta W_L q(H_L)\boldsymbol{e}_1, & m \geq L, \end{cases} \tag{2.6}$$

where $\gamma_m := \prod_{j=1}^m \eta_{j+1,j}$ and $\beta\boldsymbol{w}_1 = \boldsymbol{b}$. In particular, for any $q \in \mathcal{P}_{m-1}$ there holds $q(A)\boldsymbol{b} = \beta W_m q(H_m)\boldsymbol{e}_1$.

The proof follows by verifying the assertion for monomials, taking account of the sparsity pattern of powers of a Hessenberg matrix (see, e.g., [12]).

We next introduce polynomial notation to describe Krylov subspaces. To each vector $\boldsymbol{w}_m$ of the nested basis (2.4) there corresponds a unique polynomial $w_{m-1} \in \mathcal{P}_{m-1}$ such that $\boldsymbol{w}_m = w_{m-1}(A)\boldsymbol{b}$. Via this correspondence, the Arnoldi-like recurrence (2.5) becomes

$$\begin{aligned} \lambda[w_0(\lambda), w_1(\lambda), \ldots, w_{m-1}(\lambda)] = {} & [w_0(\lambda), w_1(\lambda), \ldots, w_{m-1}(\lambda)]H_m \\ & + \eta_{m+1,m}[0, 0, \ldots, 0, w_m(\lambda)]. \end{aligned} \tag{2.7}$$

From this equation it is evident that each zero of w_m is an eigenvalue of H_m. Moreover, by differentiating (2.7), one observes that zeros of multiplicity ℓ are eigenvalues of H_m with Jordan blocks of dimension ℓ. Since H_m is an unreduced Hessenberg matrix and

[2] We mention that the related term *Krylov decomposition* introduced by Stewart in [32] refers to a decomposition of the form (2.5) without the restriction that the basis be ascending and, consequently, to a matrix H which is not necessarily Hessenberg.

hence nonderogatory, we conclude that the zeros of w_m coincide with the eigenvalues of H_m counting multiplicity.

LEMMA 2.3. *Let H_m be the unreduced upper Hessenberg matrix in* (2.5) *and* (2.7) *and let f be a function such that $f(H_m)$ is defined. Then a polynomial $q_{m-1} \in \mathcal{P}_{m-1}$ satisfies*

$$q_{m-1}(H_m) = f(H_m)$$

if and only if $q_{m-1} = I_{w_m} f$, i.e., if q_{m-1} interpolates f in the Hermite sense at the eigenvalues of H_m.

Proof. The proof follows directly from the definition of $f(H_m)$ and the fact that the zeros of w_m are the eigenvalues of H_m with multiplicity. ∎

We summarize the contents of Lemmata 2.2 and 2.3 as follows.

THEOREM 2.4. *Given the Arnoldi-like decomposition* (2.5) *and a function f such that $f(A)$ as well as $f(H_m)$ are defined, we denote by $q \in \mathcal{P}_{m-1}$ the unique polynomial which interpolates f at the eigenvalues of H_m. Then there holds*

$$\boldsymbol{f}_m := \beta V_m f(H_m)\boldsymbol{e}_1 = \beta V_m q(H_m)\boldsymbol{e}_1 = q(A)\boldsymbol{b}. \tag{2.8}$$

We shall refer to (2.8) as *the Krylov subspace approximation of $f(A)\boldsymbol{b}$ associated with the Arnoldi-like decomposition* (2.5). Note that (2.8) is merely a computational device for generating the Krylov subspace approximation of $f(A)\boldsymbol{b}$ without explicitly carrying out the interpolation process. This is an advantage whenever $f(H_m)\boldsymbol{e}_1$ for $m \ll n$ can be evaluated efficiently.

Remark 2.5. We also point out the following—somewhat academic—detail regarding finite termination: While Krylov subspace approximations $q(A)\boldsymbol{b}$ are defined for polynomials q of any degree, Arnoldi-like decompositions, and hence (2.8), are only available for $1 \le m \le L = L(A, \boldsymbol{b})$. At index $m = L$, the characteristic polynomial of $H_L = H$ coincides with the minimal polynomial $m_{A,\boldsymbol{b}}$ of $\boldsymbol{b}$ with respect to A (see (2.2)). In view of (2.8) and Proposition 2.1, we then have $\boldsymbol{f}_L = f(A)\boldsymbol{b}$. In this sense, Krylov subspace approximations of the form (2.8) respect the spectral distribution of A relevant for $\boldsymbol{b}$ and, in exact arithmetic, possess the finite termination property. This is in contrast to other approaches such as those based on Chebyshev or Faber expansions (see below).

Besides those generated by the Arnoldi or Lanczos processes, any ascending basis $\{\boldsymbol{w}_m\}_{m=1}^L$ of $\mathcal{K}_L(A, \boldsymbol{b})$ or, equivalently, any sequence of polynomials $\{w_{m-1}\}_{m=1}^L$ of exact degree $m-1$ may be used in the Arnoldi-like decomposition (2.5) or its polynomial counterpart (2.7), provided a means for obtaining the matrix $\widetilde{H}_L$ of recurrence coefficients is available. One such example is the sequence of kernel/quasi-kernel polynomials associated with the Arnoldi/Lanczos decomposition (see [13]), where the corresponding Hessenberg matrix is easily constructed from that of the original decomposition. Approximations based on quasi-kernel polynomials are discussed in [20]. Yet another approach—one which emphasizes the interpolation aspect of the Krylov subspace approximation—fixes a sequence of nodes

$$\begin{array}{lll} \vartheta_1^{(1)} & & \\ \vartheta_1^{(2)} & \vartheta_2^{(2)} & \\ \vartheta_1^{(3)} & \vartheta_2^{(3)} & \vartheta_3^{(3)} \\ \vdots & \vdots & \ddots \end{array}$$

and chooses the basis vectors $\boldsymbol{w}_m = w_{m-1}(A)\boldsymbol{b}$ as the associated nodal polynomials

$$w_{m-1}(\lambda) = \omega_{m-1}(\lambda - \vartheta_1^{(m)})(\lambda - \vartheta_2^{(m)}) \dots (\lambda - \vartheta_m^{(m)}), \qquad \omega_m \neq 0.$$

One possible choice of such a node sequence is the zeros of Chebyshev polynomials, in which case the nested basis vectors correspond to Chebyshev polynomials. Other choices of node sequences are explored in [19, 22, 20]. Note that, in view of Remark 2.5, such a basis choice, which is independent of A and $\boldsymbol{b}$, will generally destroy the finite termination property.

We also point out that, when $f(A)$ is defined, this need not be so for $f(H_m)$ for $m < L$. For the Arnoldi approximation, a sufficient condition ensuring this is that f, as a scalar function, be analytic in a neighborhood of the field of values of A. As a case in point, consider the FOM for solving a nonsingular system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$. The solution is $f(A)\boldsymbol{b}$ with $f(\lambda) = 1/\lambda$ and, if the initial approximation is $\boldsymbol{x}_0 = \boldsymbol{0}$, the mth FOM iterate is simply the Arnoldi approximation $\boldsymbol{f}_m = \beta V_m H_m^{-1} \boldsymbol{e}_1$. There are well-known examples [3] in which $f(A)$, i.e., A^{-1}, is defined but for which H_m is singular for one or more of the indices $m = 1, \dots, L-1$, a phenomenon sometimes called a *Galerkin breakdown*.

2.3. An error representation. We conclude this section with a representation of the error of the Krylov subspace approximation of $f(A)\boldsymbol{b}$ based on any Arnoldi-like decomposition or, equivalently, any interpolatory approximation. We shall need the following notation: Given a function f and a set of nodes $\vartheta_1, \dots, \vartheta_m$ with associated nodal polynomial

$$p(\lambda) = (\lambda - \vartheta_1)(\lambda - \vartheta_2) \cdots (\lambda - \vartheta_m), \tag{2.9}$$

we denote the mth order divided difference of f with respect to the nodes $\{\vartheta_j\}_{j=1}^m$ by[3]

$$\mathbb{\Delta}_p f := \frac{f - I_p f}{p}. \tag{2.10}$$

THEOREM 2.6. *Given $A \in \mathbb{C}^{n\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$, and a function f, let* (2.5) *be an Arnoldi-like decomposition of $\mathcal{K}_m(A, \boldsymbol{b})$ and let $w_m \in \mathcal{P}_{m-1}$ be the associated polynomial; cf.* (2.7)*. Then there holds*

$$f(A)\boldsymbol{b} - \beta W_m f(H_m)\boldsymbol{e}_1 = \beta\gamma_m [\mathbb{\Delta}_{w_m} f](A)\boldsymbol{w}_{m+1} \tag{2.11}$$

with γ_m as in Lemma 2.2*.*

Proof. We consider first an arbitrary set of nodes $\vartheta_1, \dots, \vartheta_m$ with associated nodal polynomial p as in (2.9). From the definition (2.10), there holds $f(\lambda) = [I_p f](\lambda) + [\mathbb{\Delta}_p f](\lambda)p(\lambda)$. Inserting A for λ in this identity and multiplying by $\boldsymbol{b}$, we obtain

$$f(A)\boldsymbol{b} = [I_p f](A)\boldsymbol{b} + [\mathbb{\Delta}_p f](A)p(A)\boldsymbol{b}.$$

Since $I_p f \in \mathcal{P}_{m-1}$, Lemma 2.2 yields $[I_p f](A)\boldsymbol{b} = \beta W_m [I_p f](H_m)\boldsymbol{e}_1$ and, since $p \in \mathcal{P}_m$ is monic, $p(A)\boldsymbol{b} = \beta W_m p(H_m)\boldsymbol{e}_1 + \beta\gamma_m \boldsymbol{w}_{m+1}$, giving

$$f(A)\boldsymbol{b} - \beta W_m [I_p f](H_m)\boldsymbol{e}_1 = \beta [\mathbb{\Delta}_p f](A)\big(W_m p(H_m)\boldsymbol{e}_1 + \gamma_m \boldsymbol{w}_{m+1}\big).$$

[3] The source of and justification for this notation can be found in [7].

Choosing p as the characteristic polynomial w_m of H_m, it follows that $w_m(H_m) = O$ by the Cayley–Hamilton theorem and, since $I_{w_m} f$ interpolates f at the eigenvalues of H_m, there also holds $[I_{w_m} f](H_m) = f(H_m)$ by Lemma 2.3. ∎

We interpret (2.11) as follows: Further improvement of a Krylov approximation $\boldsymbol{f}_m = \beta V_m f(H_m)\boldsymbol{e}_1$ could be achieved by approximating the error term

$$f(A)\boldsymbol{b} - \boldsymbol{f}_m = \widetilde{f}(A)\widetilde{\boldsymbol{b}}$$

with $\widetilde{f}(\lambda) := [\Delta_{w_m} f](\lambda)$ and $\widetilde{\boldsymbol{b}} := \beta\gamma_m \boldsymbol{w}_{m+1}$. Note that the modified function $\widetilde{f}$ has the same domain of analyticity as f and the vector $\widetilde{\boldsymbol{b}}$ points in the direction of the last vector in the Arnoldi-like decomposition.

3. A restarted Arnoldi approximation. For the remainder of the paper we shall restrict the discussion to the Arnoldi approximation of $f(A)\boldsymbol{b}$. To set this apart from a general Krylov subspace approximation (2.8) we denote the (orthonormal) Arnoldi basis vectors by $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L$ and the Arnoldi decomposition by

$$AV_m = V_m H_m + \eta_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^\top, \qquad V_m = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$$

(cf. (2.3)). Our results apply to other Krylov subspace approximations with obvious modifications, some of which we shall point out.

3.1. Short recurrences are not enough. Besides the evaluation of $f(H_m)$, the computation of the mth Arnoldi approximation $\boldsymbol{f}_m = \beta V_m f(H_m)\boldsymbol{e}_1$ requires the Arnoldi basis V_m, which consists of m vectors of size n. As a consequence, even if the evaluation of $f(H_m)$ can be accomplished inexpensively, work and storage requirements incurred by V_m make this method impractical for moderate to large values of m. For $f(\lambda) = 1/\lambda$, i.e., when solving linear systems of equations, one can take advantage of the fact that the Arnoldi process reduces to the Hermitian Lanczos process when A is Hermitian. In this case the matrices H_m are Hermitian, hence tridiagonal, and three-term recurrence formulas can be derived for their characteristic polynomials w_m. (The same is true even in the non-Hermitian case when employing the non-Hermitian Lanczos process, possibly with look-ahead techniques.) If we interpolate $f(\lambda) = 1/\lambda$ at the zeros of the mth basis polynomial w_m, the resulting interpolating polynomial $q_{m-1} = I_{w_m} f$ satisfies

$$q_{m-1}(\lambda) = \frac{w_m(0) - w_m(\lambda)}{\lambda w_m(0)}, \tag{3.1}$$

and therefore q_{m-1} and hence also the approximation $\boldsymbol{f}_m$ obey a similar three-term recurrence. The relation (3.1) between the nodal and the interpolation polynomials can therefore be viewed as the basis for the efficiency of the conjugate gradient method and other polynomial acceleration methods such as Chebyshev iteration for solving linear systems of equations.

A relation analogous to (3.1) fails to hold for more complicated (nonrational) functions f such as the exponential function, and therefore short recurrences for the nodal polynomials do not translate into short recurrences for the interpolation polynomials. The computation of $\boldsymbol{f}_m$ therefore necessitates storing the full Arnoldi basis V_m also when A is Hermitian. It is therefore of interest to modify the Arnoldi approximation in a way that allows the construction of successively better approximations of $f(A)\boldsymbol{b}$

based on a sequence of Krylov spaces of small dimension.[4] Such *restarted Krylov subspace methods* are well known for the solution of linear systems of equations; see [26, 10].

3.2. Krylov approximation after Arnoldi restart. Consider two Krylov spaces of order m with Arnoldi decompositions

$$AV_m^{(1)} = V_m^{(1)} H_m^{(1)} + \eta_{m+1,m}^{(1)} \boldsymbol{v}_{m+1}^{(1)} \boldsymbol{e}_m^\top, \tag{3.2a}$$

$$AV_m^{(2)} = V_m^{(2)} H_m^{(2)} + \eta_{m+1,m}^{(2)} \boldsymbol{v}_{m+1}^{(2)} \boldsymbol{e}_m^\top, \tag{3.2b}$$

where $\boldsymbol{v}_1^{(1)} = \boldsymbol{b}/\beta$ and $\boldsymbol{v}_1^{(2)} = \boldsymbol{v}_{m+1}^{(1)}$, i.e., obtained from two cycles of the Arnoldi process applied to A, beginning with initial vector $\boldsymbol{b}$ and restarted after m steps with the last Arnoldi basis vector $\boldsymbol{v}_{m+1}^{(1)}$ from the first cycle. We note that the columns of $W_{2m} := [V_m^{(1)}, V_m^{(2)}]$ form a basis of $\mathcal{K}_{2m}(A, \boldsymbol{b})$, albeit not an orthonormal one, and we may combine the two proper Arnoldi decompositions (3.2) to the Arnoldi-like decomposition

$$AW_{2m} = W_{2m} H_{2m} + \eta_{m+1,m}^{(2)} \boldsymbol{v}_{m+1}^{(2)} \boldsymbol{e}_{2m}^\top, \tag{3.3}$$

where H_{2m} is the Hessenberg matrix

$$H_{2m} := \begin{bmatrix} H_m^{(1)} & O \\ \eta_{m+1,m}^{(1)} \boldsymbol{e}_1 \boldsymbol{e}_m^\top & H_m^{(2)} \end{bmatrix}. \tag{3.4}$$

Remark 3.1. We restart the Arnoldi process with $\boldsymbol{v}_{m+1}^{(1)}$, which is the most natural choice. We could, however, restart with any vector of the form

$$\widehat{\boldsymbol{v}}_{m+1} = V_m^{(1)} \boldsymbol{y} + y_{m+1} \boldsymbol{v}_{m+1}^{(1)} \in \mathcal{K}_{m+1}(A, \boldsymbol{b}) \setminus \mathcal{K}_m(A, \boldsymbol{b})$$

with a coefficient vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^\top \in \mathbb{C}^m$. In this case we must replace $H_m^{(1)}$ in (3.4) by the rank-one modification $H_m^{(1)} - (\eta_{m+1,m}^{(1)}/y_{m+1}) \boldsymbol{y} \boldsymbol{e}_m^\top$ and $\eta_{m+1,m}^{(1)}$ by $\eta_{m+1,m}^{(1)}/y_{m+1}$. It is conceivable that this could be used to emphasize certain directions such as Ritz approximations of certain eigenvectors as is done in popular restarting techniques for linear systems of equations [21] and eigenvalue calculations [30], but we shall not pursue this here.

Our objective is to compute the Krylov subspace approximation associated with (3.3) without reference to $V_m^{(1)}$. The former is defined as

$$\boldsymbol{f}_{2m} = [I_{w_{2m}} f](A)\boldsymbol{b} = \beta W_{2m} [I_{w_{2m}} f](H_{2m}) \boldsymbol{e}_1 = \beta W_{2m} f(H_{2m}) \boldsymbol{e}_1, \tag{3.5}$$

where w_{2m} is the nodal polynomial with zeros $\Lambda(H_m^{(1)}) \cup \Lambda(H_m^{(2)})$ with multiplicity. To evaluate the approximation (3.5), we note that $f(H_{2m})$ is of the form

$$f(H_{2m}) = \begin{bmatrix} f\big(H_m^{(1)}\big) & O \\ X_{2,1} & f\big(H_m^{(2)}\big) \end{bmatrix}, \qquad X_{2,1} \in \mathbb{C}^{m \times m}, \tag{3.6}$$

[4] Another remedy, well known from Lanczos-based eigenvalue computations (see [24, Chapter 13]), is to discard the basis vectors no longer needed in the recurrence and either recompute these or retrieve them from secondary storage when forming the approximation.

a consequence of the block triangular structure of H_{2m}, whereby (3.5) becomes

$$\boldsymbol{f}_{2m} = \beta V_m^{(1)} f(H_m^{(1)})\boldsymbol{e}_1 + \beta V_m^{(2)} X_{2,1}\boldsymbol{e}_1. \tag{3.7}$$

The first term on the right is the Arnoldi approximation with respect to $\mathcal{K}_m(A, \boldsymbol{b})$. If $X_{2,1}\boldsymbol{e}_1$ were computable, one could discard the basis vectors $V_m^{(1)}$ and use (3.7) to update the Arnoldi approximation, thus yielding the basis for a restarting scheme.

One conceivable approach is to observe that $X_{2,1}$ satisfies the Sylvester equation

$$H_m^{(2)}X_{2,1} - X_{2,1}H_m^{(1)} = \eta_{m+1,m}^{(1)}[f(H_m^{(2)})\boldsymbol{e}_1\boldsymbol{e}_m^\top - \boldsymbol{e}_1\boldsymbol{e}_m^\top f(H_m^{(1)})], \tag{3.8}$$

which follows from comparing the $(2,1)$ blocks of the identity $H_{2m}f(H_{2m}) = f(H_{2m})$ H_{2m}, and one could therefore proceed by solving (3.8). This approach, however, suffers from the shortcoming that the Sylvester equation (3.8) is only well conditioned if the spectra of $H_m^{(1)}$ and $H_m^{(2)}$ are well separated (cf. [16, section 15.3]). Since $H_m^{(1)}$ and $H_m^{(2)}$ are both compressions of the same matrix A, it is to be expected that at least some of their eigenvalues match very closely.

We shall instead derive a computable expression for $X_{2,1}\boldsymbol{e}_1$ directly by way of interpolation.

LEMMA 3.2. *Given two successive Arnoldi decompositions as in* (3.2)*, let* $w_m^{(1)}$*,* $w_m^{(2)}$*, and* w_{2m} *denote the monic nodal polynomials associated with* $\Lambda(H_m^{(1)})$*,* $\Lambda(H_m^{(2)})$*, and* $\Lambda(H_{2m}) = \Lambda(H_m^{(1)}) \cup \Lambda(H_m^{(2)})$*, respectively, with* H_{2m} *the upper Hessenberg matrix of the combined Arnoldi-like decomposition* (3.3)*. Then there holds*

$$[I_{w_{2m}}f](H_{2m})\boldsymbol{e}_1 = \begin{bmatrix} [I_{w_m^{(1)}}f]\big(H_m^{(1)}\big)\boldsymbol{e}_1 \\ \gamma_m^{(1)}[I_{w_m^{(2)}}(\Delta_{w_m^{(1)}}f)]\big(H_m^{(2)}\big)\boldsymbol{e}_1 \end{bmatrix}, \tag{3.9}$$

where $\gamma_m^{(1)} = \prod_{j=1}^m \eta_{j+1,j}^{(1)}$ *(cf. Lemma* 2.2*).*

Proof. Due to the block triangular structure of H_{2m} as given in (3.6), there holds

$$[I_{w_{2m}}f]\left(\begin{bmatrix} H_m^{(1)} & O \\ \eta_{m+1,m}^{(1)}\boldsymbol{e}_1\boldsymbol{e}_m^\top & H_m^{(2)} \end{bmatrix}\right) = \begin{bmatrix} [I_{w_{2m}}f]\big(H_m^{(1)}\big) & O \\ X_{2,1} & [I_{w_{2m}}f]\big(H_m^{(2)}\big) \end{bmatrix} \tag{3.10}$$

with $X_{2,1}$ as in (3.6). Next, we establish the polynomial identity

$$[I_{w_{2m}}f] = I_{w_m^{(1)}}f + I_{w_m^{(2)}}(\Delta_{w_m^{(1)}}f)\, w_m^{(1)}, \tag{3.11}$$

which can be seen by noting that both polynomials have the same degree $2m-1$ and interpolate f in the Hermite sense at the nodes $\Lambda(H_m^{(1)}) \cup \Lambda(H_m^{(1)})$. For nodes $\vartheta \in \Lambda(H_m^{(1)})$ this is so because $w_m^{(1)}(\vartheta) = 0$ and therefore

$$[I_{w_m^{(1)}}f](\vartheta) + [I_{w_m^{(2)}}(\Delta_{w_m^{(1)}}f)](\vartheta)\, w_m^{(1)}(\vartheta) = [I_{w_m^{(1)}}f](\vartheta) = f(\vartheta) = [I_{w_{2m}}f](\vartheta).$$

For nodes $\vartheta \in \Lambda(H_m^{(2)})$ we have

$$\begin{aligned}[I_{w_m^{(1)}}f](\vartheta) + [I_{w_m^{(2)}}(\Delta_{w_m^{(1)}}f)](\vartheta)\, w_m^{(1)}(\vartheta) &= [I_{w_m^{(1)}}f](\vartheta) + [\Delta_{w_m^{(1)}}f](\vartheta)\, w_m^{(1)}(\vartheta) \\ &= f(\vartheta) = [I_{w_{2m}}f](\vartheta)\end{aligned}$$

with the second equality following from the definition (2.10), and (3.11) is established. The assertion on the first block of the vector (3.9) is now verified by inserting the

matrix $H_m^{(1)}$ into the polynomials on either side of (3.11), noting that $w_m^{(1)}(H_m^{(1)}) = O$, and multiplying both sides of (3.10) by $\boldsymbol{e}_1$.

To verify the second block of (3.9), we use identity (3.11) to write

$$[I_{w_{2m}} f](H_{2m}) = M^{(1)} + M^{(2)} M^{(3)},$$

where

$$M^{(1)} := [I_{w_m^{(1)}} f](H_{2m}), \quad M^{(2)} := [I_{w_m^{(2)}}(\Delta_{w_m^{(1)}} f)](H_{2m}), \quad M^{(3)} := w_m^{(1)}(H_{2m}).$$

The block lower triangular structure of H_{2m} carries over to functions of H_{2m}, giving

$$M^{(i)} = \begin{bmatrix} M_{1,1}^{(i)} & O \\ M_{2,1}^{(i)} & M_{2,2}^{(i)} \end{bmatrix}, \qquad i = 1, 2, 3,$$

where in addition $M_{1,1}^{(3)} = w_m^{(1)}(H_m^{(1)}) = O$. In this notation the second block of (3.9) is given by

$$X_{2,1}\boldsymbol{e}_1 = M_{2,1}^{(1)}\boldsymbol{e}_1 + M_{2,2}^{(2)} M_{2,1}^{(3)}\boldsymbol{e}_1. \tag{3.12}$$

For the first term on the right, we have $M_{2,1}^{(1)}\boldsymbol{e}_1 = \mathbf{0}$ because, as the $(2,1)$-block of $M^{(1)} = [I_{w_m^{(1)}} f](H_{2m})$, a polynomial of degree $m-1$ in the Hessenberg matrix H_{2m}, $M_{2,1}^{(1)}$ has a zero first column. Next, again by the block lower triangular structure of H_{2m}, there holds $M_{2,2}^{(2)} = [I_{w_m^{(2)}}(\Delta_{w_m^{(1)}} f)](H_m^{(2)})$. Finally, we note that $M_{2,1}^{(3)}\boldsymbol{e}_1 = \gamma_m^{(1)}\boldsymbol{e}_1$. This follows in a similar way as the evaluation of $M_{2,1}^{(1)}\boldsymbol{e}_1$, but here $M^{(3)} = w_m^{(1)}(H_{2m})$ is a polynomial of degree m in the $2m \times 2m$ upper Hessenberg matrix H_{2m}. Again by the sparsity structure of powers of Hessenberg matrices, the first column of $M_{2,1}^{(3)}$ is a multiple of $\boldsymbol{e}_1$. Comparing coefficients reveals this multiple to be $\gamma_m^{(1)}$. Inserting these quantities in (3.12) establishes the second block of identity (3.9), and the proof is complete. ∎

Remark 3.3. We note that the same proof applies when the two Krylov spaces are of different dimensions m_1 and m_2.

Comparing coefficients in (3.7) and (3.9) reveals that $X_{2,1}\boldsymbol{e}_1 = \gamma_m^{(1)}[\Delta_{w_m^{(1)}} f]\big(H_m^{(2)}\big)\boldsymbol{e}_1$, and we summarize the resulting basic restart step in the following theorem.

THEOREM 3.4. *The Krylov subspace approximation* (3.5) *based on the Arnoldi-like decomposition* (3.3) *is given by*

$$\boldsymbol{f}_{2m} = \beta V_m^{(1)} f\big(H_m^{(1)}\big)\boldsymbol{e}_1 + \beta\gamma_m^{(1)} V_m^{(2)}[\Delta_{w_m^{(1)}} f]\big(H_m^{(2)}\big)\boldsymbol{e}_1. \tag{3.13}$$

Proof. The proof follows immediately from (3.5) upon inserting the representation for $[I_{w_{2m}} f](H_{2m})$ given in Lemma 3.2: Starting with (3.5), we obtain

$$\begin{aligned} \boldsymbol{f}_{2m} &= \beta W_{2m} f(H_{2m})\boldsymbol{e}_1 \\ &= \beta\left(V_m^{(1)}[I_{w_m^{(1)}} f]\big(H_m^{(1)}\big)\boldsymbol{e}_1 + V_m^{(2)}\gamma_m^{(1)}[I_{w_m^{(2)}}(\Delta_{w_m^{(1)}} f)]\big(H_m^{(2)}\big)\boldsymbol{e}_1\right) \\ &= \beta V_m^{(1)} f\big(H_m^{(1)}\big)\boldsymbol{e}_1 + \beta\gamma_m^{(1)} V_m^{(2)}[\Delta_{w_m^{(1)}} f]\big(H_m^{(2)}\big)\boldsymbol{e}_1, \end{aligned}$$

where the last equality follows from the interpolation properties of $I_{w_m^{(1)}}$ and $I_{w_m^{(2)}}$. ∎

3.3. The restarting algorithm. Theorem 3.4 suggests the following scheme for a Krylov approximation of $f(A)\boldsymbol{b}$ based on the restarted Arnoldi process with cycle length m: The first approximation $\boldsymbol{f}^{(1)}$ is simply the usual Arnoldi approximation with respect to the first Krylov space $\mathcal{K}_m(A, \boldsymbol{b})$, i.e., $\boldsymbol{f}^{(1)} = \beta V_m^{(1)} f(H_m^{(1)})\boldsymbol{e}_1$. The next Krylov space is generated with the initial vector $\boldsymbol{v}_{m+1}^{(1)}$ and, according to (3.13), the correction to $\boldsymbol{f}^{(1)}$ required to obtain the Krylov subspace approximation of $f(A)\boldsymbol{b}$ with respect to the Arnoldi-like decomposition (3.3) is given by

$$\boldsymbol{f}^{(2)} = \boldsymbol{f}^{(1)} + \beta\gamma_m^{(1)} V_m^{(2)} [\Delta_{w_m^{(1)}} f]\big(H_m^{(2)}\big)\boldsymbol{e}_1.$$

The effect of restarting is seen to be a modification of the function f to $\Delta_{w_m}^{(1)} f$ and a replacement of the vector $\boldsymbol{b}$ by $\beta\gamma_m^{(1)}\boldsymbol{v}_{m+1}^{(1)}$. Note that this is in line with the error representation (2.3) in that, after restarting, we are in fact approximating the error term and using this approximation as a correction. The computation of this update requires storing a representation of $\Delta_{w_m^{(1)}} f$ as well as the current approximation $\boldsymbol{f}^{(1)}$, but the Arnoldi basis $V_m^{(1)}$ can be discarded. Proceeding in this fashion, we arrive at the restarting scheme given in Algorithm 1.

ALGORITHM 1: RESTARTED ARNOLDI APPROXIMATION FOR $f(A)\boldsymbol{b}$

Given: $A, \boldsymbol{b}, f$
$f^{(0)} := f$, $\boldsymbol{f}^{(0)} := \boldsymbol{0}$, $\boldsymbol{b}^{(0)} := \boldsymbol{b}$, $\gamma^{(0)} := \|\boldsymbol{b}\|$.
for $k = 1, 2, \ldots$ *until convergence* **do**
  Compute the Arnoldi decomposition $AV_m^{(k)} = V_m^{(k)} H^{(k)} + \eta_{m+1,m}^{(k)} \boldsymbol{b}^{(k)} \boldsymbol{e}_m^\top$ of $\mathcal{K}_m(A, \boldsymbol{b}^{(k-1)})$.
  Update the approximation $\boldsymbol{f}^{(k)} := \boldsymbol{f}^{(k-1)} + \gamma^{(k-1)} V_m^{(k)} f^{(k-1)}(H_m^{(k)})\boldsymbol{e}_1$.
  $\gamma^{(k)} := \gamma^{(k-1)} \prod_{j=1}^{m} \eta_{j+1,j}^{(k)}$
  $f^{(k)} := \Delta_{w_m^{(k)}} f^{(k-1)}$, where $w_m^{(k)}$ is the characteristic polynomial of $H_m^{(k)}$.

Remark 3.5. Algorithm 1 is formulated for Krylov spaces of constant dimension m in each restart cycle, but this dimension can vary from cycle to cycle.

Although Algorithm 1 appears very attractive from a computational point of view, numerical experiments with a MATLAB implementation have revealed it to be afflicted with severe stability problems. The cause of this seems to be the difficulty of numerically computing interpolation polynomials of high degree (see also [33]).

We therefore turn to a slightly less efficient variant of our restarting scheme, which our numerical tests indicate to be free from these stability problems. The generic step of this second variant of the restarted Arnoldi algorithm proceeds as follows: After $k-1$ cycles of the algorithm, we may collect the entirety of Arnoldi decompositions in the $(k-1)$-fold Arnoldi-like decomposition

$$AW_{(k-1)m} = W_{(k-1)m} H_{(k-1)m} + \eta_{m+1,m}^{(k-1)} \boldsymbol{v}_{m+1}^{(k-1)} \boldsymbol{e}_{(k-1)m}^\top$$

with $W_{(k-1)m} = [V_m^{(1)} V_m^{(2)} \ldots V_m^{(k-1)}]$. Combining this with the Arnoldi decomposition

$$AV_m^{(k)} = V_m^{(k)} H_m^{(k)} + \eta_{m+1,m}^{(k)} \boldsymbol{v}_{m+1}^{(k)} \boldsymbol{e}_m^\top$$

of the next Krylov space $\mathcal{K}_m(A, \boldsymbol{v}_{m+1}^{(k-1)})$, we obtain the next Arnoldi-like decomposi-

tion

$$AW_{km} = W_{km}H_{km} + \eta^{(k)}_{m+1,m}\boldsymbol{v}^{(k)}_{m+1}\boldsymbol{e}_{km}^\top$$

with $W_{km} = [W_{(k-1)m}, V^{(k)}_m]$ and

$$H_{km} = \begin{bmatrix} H_{(k-1)m} & O \\ \eta^{(k-1)}_{m+1,m}\boldsymbol{e}_1\boldsymbol{e}^\top_{(k-1)m} & H^{(k)}_m \end{bmatrix}.$$

Denoting by w_{km} the characteristic polynomial of H_{km}, formula (2.11) for the Krylov subspace approximation with respect to an Arnoldi-like decomposition gives

$$\boldsymbol{f}^{(k)} = \beta W_{km} f(H_{km})\boldsymbol{e}_1 = \boldsymbol{f}^{(k-1)} + \beta V^{(k)}_m [f(H_{km})\boldsymbol{e}_1]_{(k-1)m+1:km}, \tag{3.14}$$

where the subscript in the last term is meant to refer to the vector with the last m components of $f(H_{km})\boldsymbol{e}_1$. (3.14) provides an alternative update formula for the restarted Arnoldi approximation. It is somewhat less efficient than that given in Algorithm 1 in that it requires storing H_{km} and the evaluation of $f(H_{km})$, but we have found it to be much stabler than the former. The second variant is summarized in Algorithm 2.

ALGORITHM 2: RESTARTED ARNOLDI APPROXIMATION FOR $f(A)\boldsymbol{b}$ (VARIANT 2)

Given: $A, \boldsymbol{b}, f$
$f^{(0)} := f$, $\boldsymbol{f}^{(0)} := \boldsymbol{0}$, $\boldsymbol{b}^{(0)} := \boldsymbol{b}$, $\beta := \|\boldsymbol{b}\|$.
for $k = 1, 2, \ldots$ *until convergence* **do**
  Compute the Arnoldi decomposition $AV^{(k)}_m = V^{(k)}_m H^{(k)}_m + \eta^{(k)}_{m+1,m}\boldsymbol{b}^{(k)}\boldsymbol{e}_m^\top$ of $\mathcal{K}_m(A, \boldsymbol{b}^{(k-1)})$.
  if $k = 1$ **then**
    $H_{km} := H^{(1)}_m$
  else
    $H_{km} := \begin{bmatrix} H_{(k-1)m} & O \\ \eta^{(k-1)}_{m+1,m}\boldsymbol{e}_1\boldsymbol{e}^\top_{(k-1)m} & H^{(k)}_m \end{bmatrix}$.
  Update the approximation $\boldsymbol{f}^{(k)} := \boldsymbol{f}^{(k-1)} + \beta V^{(k)}_m [f(H_{km})\boldsymbol{e}_1]_{(k-1)m+1:km}$.

4. Properties of the restarted Arnoldi algorithm.

4.1. Special cases. In this section we recover some known algorithms as special cases of the restarted Arnoldi approximation.

4.1.1. Linear systems of equations. We begin by showing that for $f(\lambda) = 1/\lambda$ we recover the well-known restarted FOM for solving linear systems of equations [27]. With $I_{w_m} f$ for this case given in (3.1), there results

$$[\mathbb{\Delta}_{w_m} f](\lambda) = \frac{1}{w_m(0)}\frac{1}{\lambda},$$

so that the representation (2.11) becomes

$$A^{-1}\boldsymbol{b} = \beta V_m H_m^{-1}\boldsymbol{e}_1 + \frac{\beta\gamma_m}{w_m(0)}A^{-1}\boldsymbol{v}_{m+1} = \boldsymbol{f}_m + \frac{\beta\gamma_m}{w_m(0)}A^{-1}\boldsymbol{v}_{m+1},$$

where $\boldsymbol{f}_m$ denotes the mth FOM iterate. The associated residual is therefore

$$\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{f}_m = \frac{\beta\gamma_m}{w_m(0)}\boldsymbol{v}_{m+1}, \tag{4.1}$$

which leads to

$$A^{-1}\boldsymbol{b} = \boldsymbol{f}_m + A^{-1}\boldsymbol{r}_m.$$

We conclude that in this case the exact correction $\boldsymbol{c}$ to the Arnoldi approximation $\boldsymbol{f}_m$ is the solution of the residual equation $A\boldsymbol{c} = \boldsymbol{r}_m$, leading to the problem of approximating $f(A)\boldsymbol{r}_m$, which in restarted FOM is carried out using a new Krylov space with initial vector $\boldsymbol{r}_m$. As an aside, we observe that (4.1) implies that the FOM residual norm can be expressed as

$$\|\boldsymbol{r}_m\| = \frac{\beta\gamma_m}{|w_m(0)|} = \frac{\beta\prod_{j=1}^{m}\eta_{j+1,j}}{|\det H_m|},$$

an expression first given in [4].

4.2. Initial value problems. We consider the initial value problem

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t), \quad \boldsymbol{y}(0) = \boldsymbol{b} \tag{4.2}$$

with $A \in \mathbb{C}^{n\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$ (independent of t) with solution

$$\boldsymbol{y}(t) = f_t(A)\boldsymbol{b}, \qquad f_t(\lambda) = e^{t\lambda}. \tag{4.3}$$

The Arnoldi approximation of (4.3) with respect to (2.3) is given by

$$\boldsymbol{y}_m(t) = V_m\boldsymbol{u}(t), \qquad \boldsymbol{u}(t) = \beta e^{tH_m}\boldsymbol{e}_1, \quad \beta = \|\boldsymbol{b}\|. \tag{4.4}$$

As is easily verified, the associated approximation error $\boldsymbol{d}_m(t) := \boldsymbol{y}(t) - \boldsymbol{y}_m(t)$ as a function of t satisfies the initial value problem

$$(\partial_t - A)\boldsymbol{d}_m(t) = \boldsymbol{r}_m(t), \qquad \boldsymbol{d}_m(0) = \boldsymbol{0}, \tag{4.5}$$

in which the forcing term $\boldsymbol{r}_m(t)$, which plays the role of a residual, is given by

$$\boldsymbol{r}_m(t) := \eta_{m+1,m}\boldsymbol{e}_1^\top\boldsymbol{u}(t)\boldsymbol{v}_{m+1} = \beta\eta_{m+1,m}\boldsymbol{e}_m^\top e^{tH_m}\boldsymbol{e}_1\boldsymbol{v}_{m+1} =: \rho_m(t)\boldsymbol{v}_{m+1}. \tag{4.6}$$

In [4] (see also [19]) Celledoni and Moret propose a restarted Krylov subspace scheme for solving (4.2) based on the variation of constants formula

$$\boldsymbol{d}_m(t) = F_t(A)\boldsymbol{v}_{m+1}, \qquad F_t(\lambda) := \int_0^t e^{(t-s)\lambda}\rho_m(s)\,ds, \tag{4.7}$$

for the solution of the residual equation (4.5) using repeated Arnoldi approximations of $F_t(A)\boldsymbol{v}_{m+1}$ in a manner similar to Algorithm 1. We note that, in contrast to Algorithm 1, their method requires a time-stepping scheme in addition to the Krylov approximation. As the approximate solution (4.4) of (4.2) is an Arnoldi approximation, the error representation (4.7) must coincide with that given in (2.11). To provide more insight on the restarted Arnoldi approximation for solving initial value problems, we proceed to show explicitly that the two error representations are the same. The key is the proper treatment of the parameter t. Denoting the error representation (2.11) with $f = f_t$ by

$$\widetilde{\boldsymbol{d}}_m(t) = \beta\gamma_m[\Delta_{w_m}f_t](A)\boldsymbol{v}_{m+1}, \tag{4.8}$$

we prove the following result.

THEOREM 4.1. *The error representation* (4.8) *for the Arnoldi approximation of* (4.3) *as a function of* t *solves the initial value problem* (4.5).

Proof. The initial condition $\widetilde{\boldsymbol{d}}_m(0) = \mathbf{0}$ follows from the fact that $f_0 \equiv 1$ and, since this function is interpolated without error, the associated divided difference is zero.

To verify that $\widetilde{\boldsymbol{d}}_m$ solves the differential equation, note first that differentiating the interpolant of f_t with respect to the parameter t results in

$$\partial_t[I_{w_m} f_t] = I_{w_m}(\partial_t f_t). \tag{4.9}$$

This can be seen by writing the interpolant as

$$[I_{w_m} f_t](\lambda) = \sum_{j=1}^{k} \sum_{\ell=0}^{n_j-1} f_t^{(\ell)}(\vartheta_j)\, q_{j,\ell}(\vartheta_j), \qquad \sum_{j=1}^{k} n_j = m,$$

in terms of the Hermite basis polynomials $q_{j,\ell} \in \mathcal{P}_{m-1}$, characterized by

$$q_{j,\ell}^{(p)}(\vartheta_q) = \delta_{j,q}\,\delta_{\ell,p}, \qquad j,q = 1,2,\ldots,k, \qquad \ell,p = 0,1,\ldots,n_j-1,$$

and exchanging the order of differentiation. As a consequence of (4.9) and the fact that $(\partial_t f_t)(\lambda) = \lambda f_t(\lambda)$, we also have

$$\partial_t[\Lambda_{w_m} f_t] = \Lambda_{w_m}(\partial_t f_t) = \Lambda_{w_m}(g f_t), \qquad \text{where } g(\lambda) = \lambda.$$

The product formula for divided differences (see, e.g., [25, Theorem 1.3.3]) now yields

$$\partial_t[\Lambda_{w_m} f_t](\lambda) = \lambda[\Lambda_{w_m} f_t](\lambda) + \pi_{m-1}(t), \tag{4.10}$$

where $\pi_{m-1}(t)$ is the leading coefficient of $I_{w_m} f_t$. Inserting A for λ in the scalar equation (4.10) and multiplying by $\boldsymbol{v}_{m+1}$ now gives us

$$\begin{aligned}(\partial_t - A)\widetilde{\boldsymbol{d}}_m(t) &= \beta\gamma_m \Big(A[\Lambda_{w_m} f_t](A) + \pi_{m-1}(t) I - A[\Lambda_{w_m} f_t](A) \Big) \boldsymbol{v}_{m+1} \\ &= \beta\gamma_m \pi_{m-1}(t) \boldsymbol{v}_{m+1}.\end{aligned}$$

A comparison with (4.6) reveals that what remains to be shown is that

$$\frac{\gamma_m}{\eta_{m+1,m}} \pi_{m-1}(t) = \boldsymbol{e}_m^\top e^{tH_m} \boldsymbol{e}_1.$$

The term on the right is the entry at the $(m,1)$ position of the matrix $e^{tH_m} = [I_{w_m} f_t](H_m)$. Due to the sparsity pattern of powers of an upper Hessenberg matrix, this entry is given by

$$\prod_{j=1}^{m-1} \eta_{j+1,j} \pi_{m-1}(t) = \frac{\gamma_m}{\eta_{m+1,m}} \pi_{m-1}(t)$$

and the proof is complete. ∎

The uniqueness of the solution of (4.5) together with Theorem 4.1 now imply once more that $\widetilde{\boldsymbol{d}}_m(t) = \boldsymbol{d}_m(t)$. We emphasize again that our restarted Arnoldi method approximates these error terms directly without recourse to a time-stepping scheme.

4.3. Convergence. The full Arnoldi approximation is known to converge superlinearly for the exponential function, as shown in, e.g., [17, 8]. For the case of solving linear systems of equations, i.e., the Arnoldi approximation for the function $f(\lambda) = 1/\lambda$, it is known that restarting the process can degrade or even destroy convergence. In this section we show that, for sufficiently smooth functions, restarting the Arnoldi approximation preserves its superlinear convergence. We state the following result for entire functions of order one (cf. [2, section 2.1]), a class which includes the exponential function, and note that the result generalizes to other orders with minor modifications.

THEOREM 4.2. *Given $A \in \mathbb{C}^{n\times n}$ and an entire function f of order one, let $\{\vartheta_j^{(m)}\}_{j=1}^m, m \geq 1$, denote an arbitrary node sequence contained in the field of values $W(A)$ of A with associated nodal polynomials $w_m \in \mathcal{P}_m$. Then there exist constants C and γ which are independent of m such that*

$$\|f(A)\boldsymbol{b} - [I_{w_m} f](A)\boldsymbol{b}\| \leq C\frac{\gamma^{m-1}}{(m-1)!}\|\boldsymbol{b}\| \qquad \text{for all } m. \tag{4.11}$$

Proof. We recall the well-known Hermite representation theorem for the interpolation error (cf. [6, Theorem 3.6.1]): Let $\Gamma \subset \mathbb{C}$ be a contour which contains $W(A)$, and hence also the interpolation nodes, in its interior, which we denote by Ω. Then for all $\lambda \in \Omega$ we have

$$f(\lambda) - [I_{w_m} f](\lambda) = \frac{1}{2\pi i}\int_\Gamma \frac{f(t)}{t-\lambda}\frac{w_m(\lambda)}{w_m(t)}\,dt. \tag{4.12}$$

By replacing f with $f - p$ in (4.12), we obtain for arbitrary polynomials $p \in \mathcal{P}_{m-1}$ the identity

$$f(\lambda) - [I_{w_m} f](\lambda) = \frac{1}{2\pi i}\int_\Gamma \frac{f(t)-p(t)}{t-\lambda}\frac{w_m(\lambda)}{w_m(t)}\,dt.$$

Inserting A for λ on both sides, multiplying with $\boldsymbol{b}$, and taking norms gives

$$\|f(A)\boldsymbol{b} - [I_{w_m} f](A)\boldsymbol{b}\| = \frac{1}{2\pi}\left\|\int_\Gamma [f(t)-p(t)](tI-A)^{-1}\frac{w_m(A)}{w_m(t)}\,dt\,\boldsymbol{b}\right\|. \tag{4.13}$$

We now bound each factor of the integrand. For any unit vector $\boldsymbol{u} \in \mathbb{C}^n$, we have $\boldsymbol{u}^H A\boldsymbol{u} \in W(A)$, and thus, for all $t \in \Gamma$,

$$\operatorname{dist}(\Gamma, W(A)) \leq |t - \boldsymbol{u}^H A\boldsymbol{u}| = |\boldsymbol{u}^H(tI-A)\boldsymbol{u}| \leq \|(tI-A)\boldsymbol{u}\|.$$

For arbitrary $\boldsymbol{v} \in \mathbb{C}^n$, it follows that $\|(tI-A)\boldsymbol{v}\| \geq \operatorname{dist}(\Gamma, W(A))\|\boldsymbol{v}\|$ and therefore

$$\|(tI-A)^{-1}\| \leq \frac{1}{\operatorname{dist}(\Gamma, W(A))}. \tag{4.14}$$

Similarly, since the nodes $\vartheta_j^{(m)}$ are contained in $W(A)$ by assumption, we have

$$|w_m(t)| = |(t-\vartheta_1^{(m)})(t-\vartheta_2^{(m)})\cdots(t-\vartheta_m^{(m)})| \geq \operatorname{dist}(\Gamma, W(A))^m, \quad t \in \Gamma. \tag{4.15}$$

Moreover, with $r(A) := \max\{|\lambda| : \lambda \in W(A)\}$ denoting the numerical radius of A, we may bound $\|w_m(A)\|$ by

$$\|w_m(A)\| \leq \prod_{j=1}^m \|A - \vartheta_j I\| \leq \prod_{j=1}^m (\|A\| + \vartheta_j) \leq [3r(A)]^m, \tag{4.16}$$

which follows from the well-known inequality $\|A\| \leq 2r(A)$ and since $\vartheta_j^{(m)} \in W(A)$.

Thus, from (4.13), (4.14), (4.15), (4.16), and the fact that $p \in \mathcal{P}_{m-1}$ was arbitrary, we obtain the bound

$$\|f(A)\boldsymbol{b} - [I_{w_m} f](A)\boldsymbol{b}\| \leq \frac{\ell(\Gamma)}{2\pi} \frac{\inf_{p\in\mathcal{P}_{m-1}} \|f-p\|_{\infty,\Omega}\, [3r(A)]^m}{\operatorname{dist}(\Gamma, W(A))^{m+1}} \|\boldsymbol{b}\|,$$

where $\ell(\Gamma)$ denotes the length of the contour Γ and $\|\cdot\|_{\infty,\Omega}$ denotes the supremum norm on Ω. The assertion now follows from the convergence rate of best uniform approximation of entire functions of order one by polynomials. In particular, it is known (see [11]) that there exist constants $\widetilde{C}$ and $\widetilde{\gamma}$ such that

$$\inf_{p\in\mathcal{P}_{m-1}} \|f-p\|_{\infty,\Omega} \leq \widetilde{C} \frac{\widetilde{\gamma}^{m-1}}{(m-1)!}. \qquad \square$$

COROLLARY 4.3. *The restarted Arnoldi approximation converges superlinearly for entire functions of order one.*

Proof. This follows from Theorem 4.2 by noting that, for the Arnoldi approximation, the set of interpolation nodes for each restart cycle are Ritz values of A and therefore contained in $W(A)$. $\square$

5. Numerical experiments. In this section we demonstrate the behavior of the restarted Arnoldi approximation for the exponential function using several examples from the literature. All computations were carried out in MATLAB version 7.0 (R14) on a 1.6 GHz Power Mac G5 computer with 1.5 GB of RAM.

5.1. Three-dimensional heat equation. Our first numerical experiment is based on one from [14]: Consider the initial boundary value problem

$$\dot{u} - \Delta u = 0 \quad \text{on } (0,1)^3 \times (0,T), \tag{5.1a}$$

$$u(x,t) = 0 \quad \text{on } \partial(0,1)^3 \text{ for all } t \in [0,T], \tag{5.1b}$$

$$u(x,0) = u_0(x), \quad x \in (0,1)^3. \tag{5.1c}$$

When the Laplacian is discretized by the usual seven-point stencil on a uniform grid involving n interior grid points in each Cartesian direction, problem (5.1) reduces to the initial value problem

$$\begin{aligned} \dot{\boldsymbol{u}}(t) &= A\boldsymbol{u}(t), \quad t \in (0,T), \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0, \end{aligned}$$

with an $N \times N$ matrix A ($N = n^3$) and an initial vector $\boldsymbol{u}_0$ consisting of the values $u_0(x)$ at the grid points x, the solution of which is given by

$$\boldsymbol{u}(t) = f_t(A)\boldsymbol{u}_0 = e^{tA}\boldsymbol{u}_0. \tag{5.2}$$

As in [14], we give the initial vector in terms of its expansion in eigenfunctions of the discrete Laplacian as

$$\boldsymbol{u}_0^{i,j,k} = \sum_{i',j',k'} \frac{1}{i'+j'+k'} \sin(ii'\pi h) \sin(jj'\pi h) \sin(kk'\pi h).$$

Here $h = 1/(n+1)$ is the mesh size and the triple indexing is relative to the lexicographic ordering of the mesh points in the unit cube.

TABLE 5.1

The full Arnoldi approximation applied to the three-dimensional heat equation with $h = 1/36$ and $t = 0.1 = n_{\text{steps}}\Delta t$. The dimension m of the Krylov spaces is chosen as the smallest to result in an error $\|\boldsymbol{e}\|_2$ less than 10^{-10} at $t = 0.1$.

Δt	n_{steps}	m	Time [s]	$\|\boldsymbol{e}\|_2$
1e-1	1	72	12.0	7.76e-11
5e-2	2	51	10.5	8.47e-11
2e-2	5	36	13.7	8.54e-11
1e-2	10	29	18.3	2.09e-11
5e-3	20	22	22.7	5.13e-11
1e-3	100	13	42.4	1.55e-11
5e-4	200	11	62.6	5.36e-12
1e-4	1000	8	172.2	1.20e-12
5e-5	2000	7	299.3	1.72e-12

TABLE 5.2

The restarted Arnoldi approximation applied to the three-dimensional heat equation with $h = 1/36$ and $t = 0.1$. The dimension m of the Krylov spaces is chosen to coincide with the runs in Table 5.1 and now the number of restarts k is chosen as the smallest to result in an error $\|\boldsymbol{e}\|_2$ less than 10^{-10} at $t = 0.1$.

Δt	k	m	Time [s]	$\|\boldsymbol{e}\|_2$
1e-1	2	51	10.2	2.22e-17
1e-1	2	36	5.2	3.61e-12
1e-1	3	29	5.0	7.78e-15
1e-1	4	22	4.1	9.54e-15
1e-1	6	13	2.2	4.37e-11
1e-1	7	11	1.8	1.29e-11
1e-1	10	8	1.7	7.01e-11
1e-1	12	7	1.6	3.27e-11

We first consider the case $n = 35$ and repeat a calculation in [14], where (5.2) is approximated at $t = 0.1$ using the unrestarted Arnoldi approximation. Writing the solution in the form $\boldsymbol{u}(t) = (e^{\Delta t A})^k \boldsymbol{u}_0$, where $k\Delta t = t$, one can compute the solution using k applications of the Arnoldi approximations involving the matrix $\Delta t A$, which has a smaller spectral interval than A and hence results in faster convergence. There is thus a tradeoff between using Krylov spaces of small dimension and having to take a small number of time steps of length Δt. The results in Table 5.1 show the execution times which result from fixing the time step Δt and using the smallest Krylov subspace dimension m which results in a Euclidean norm of less that 10^{-10} for the error vector $\boldsymbol{e}$ of the unrestarted Arnoldi approximation. We observe that using smaller time steps does allow one to use smaller Krylov spaces, but at a higher cost in terms of execution time.

We next consider the same problem, but instead of taking several time steps with the full Arnoldi approximation, we reduce the size of the Krylov spaces by restarting after every m steps. The results are given in Table 5.2. The dimension m of the Krylov spaces is chosen to coincide with the corresponding runs from Table 5.1, but now the number of restarts k is chosen as the smallest to result in an error less than 10^{-10}. Again there is a tradeoff between the size of the Krylov space and the number of restarts required until convergence. In contrast to Table 5.1, however, we note that the total execution times decrease rather than increase when smaller Krylov spaces are employed, in spite of the fact that this requires more restart cycles. Moreover, the longest execution time of the restarted variant is less than half of the shortest execution time of any of the full Arnoldi approximation runs.

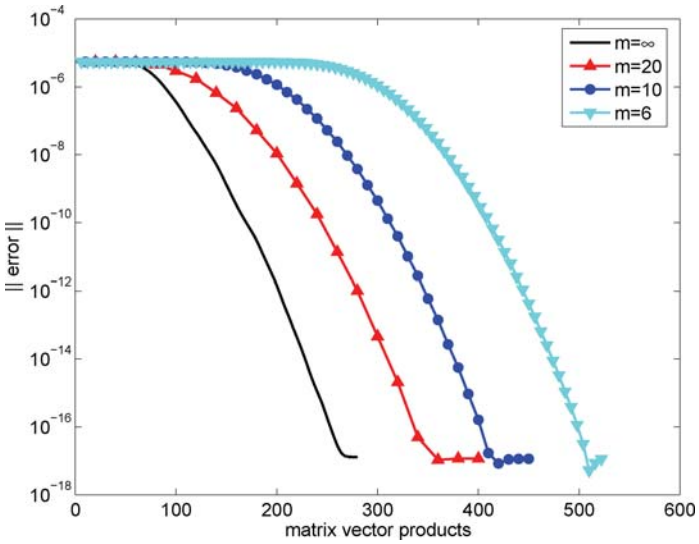


FIG. 5.1. *Error norm histories for the restarted Arnoldi approximation applied to the three-dimensional heat equation with $n = 50$, i.e., $N = 125\,000$, for several restart lengths m.*

TABLE 5.3
Execution times for the runs depicted in Figure 5.1: m denotes the restart length and k the number of restart cycles.

m	k	Time [s]
∞	1	1948
20	20	206
10	45	146
6	87	153

Finally, we consider the same problem for a finer discretization with $n = 50$, resulting in a matrix of dimension $N = 125\,000$. We apply the restarted Arnoldi approximation with restart lengths $m = 6, 10$, and 20 using the full Arnoldi approximation ($m = \infty$) as a reference. Each iteration is run until the accuracy no longer improves. The resulting error curves are shown in Figure 5.1, and the corresponding execution times in Table 5.3. We observe here that the method requires successively more restart cycles to converge as the restart length is decreased. Convergence, however, is merely delayed and is maintained down to the smallest restart length $m = 6$. In terms of execution time, there appears to be a point of diminishing returns using shorter and shorter restart lengths, as the shortest execution time was obtained for $m = 10$.

5.2. Skew-symmetric problem. Our next example is taken from [17]. We consider a matrix A with 1001 equidistant eigenvalues in $[-20i, 20i]$. In contrast to [17], we choose A to be block diagonal and real (and not diagonal and complex) in order to avoid complex arithmetic, as follows:

$$
\begin{aligned}
&A = \operatorname{blockdiag}(B_0, B_1, \ldots, B_{500}) \in \mathbb{R}^{1001\times 1001},\\
&B_0 = 0,\\
&B_j = \tfrac{j}{25}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad j = 1, 2, \ldots, 500.
\end{aligned}
\tag{5.3}
$$

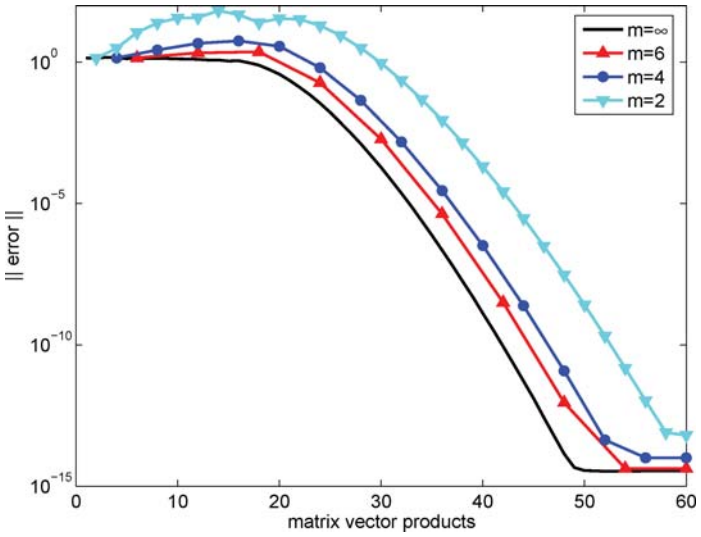


FIG. 5.2. *Error norm histories for the skew-symmetric problem of dimension* $n = 1001$.

The vector $\boldsymbol{b}$ is a random vector[5] of unit norm. The error curve of the full Arnoldi approximation ($m = \infty$) as well as those of the restarted Arnoldi approximation with restart lengths $m = 2, 4$, and 6 are shown in Figure 5.2.

We observe that the errors associated with the restarted Arnoldi approximations initially increase before tending to zero. We also observe that the final accuracy of the approximation deteriorates with decreasing restart length. This indicates that the restart length m is too small to "resolve" the spectral interval of A.

For an explanation, recall that the Arnoldi approximations $\boldsymbol{f}_m$ of $\exp(A)\boldsymbol{b}$ can be viewed as the result of an interpolation process: $\boldsymbol{f}_m = q_{m-1}(A)\boldsymbol{b}$, where q_{m-1} is an interpolating polynomial for the exponential function. For the unrestarted Arnoldi method, the interpolation nodes are the Ritz values of A with respect to $\mathcal{K}_m(A, \boldsymbol{b})$, which are approximately uniformly distributed over $[-20i, 20i]$ (cf. Figure 5.3, where the imaginary parts of the Ritz values are shown[6].) For the restarted Arnoldi method (with restart length m), however, the interpolation nodes are the collection of the Ritz values of A with respect to several Krylov spaces $\mathcal{K}_m\left(A, \boldsymbol{b}^{(j)}\right)$, $j = 0, 1, \ldots, k-1$ (after k restarts). These are far from uniformly distributed in $[-20i, 20i]$, but rather tend to accumulate at m discrete points (see Figure 5.3).

In the extreme case of restart length one, all interpolating nodes equal $\vartheta = 0$ (at least in exact arithmetic) and the interpolating polynomial $q_{k-1}(\lambda) = \sum_{j=0}^{k-1} \frac{1}{j!}\lambda^j$ is simply the truncated Taylor expansion of $\exp(\lambda)$. It is well known that, for $|\lambda| \gg 0$, intermediate partial sums are much larger than the final limit. An analogous statement holds for Hermite interpolating polynomials of the exponential function at too few nodes.

The phenomenon described above becomes more pronounced if we increase the spectral interval of A: Again, we consider the matrix A of (5.3), but now of dimension 10001 with equidistant eigenvalues in $[-200i, 200i]$. The resulting error curves for the restart lengths $m = 5, 10, 20$, and 40 are shown in Figure 5.4.

[5]Generated by the MATLAB syntax `randn('state',0); b = randn(1001,1)`

[6]Note that all Ritz values are purely imaginary because A is skew-symmetric and $\boldsymbol{b}$ is real.

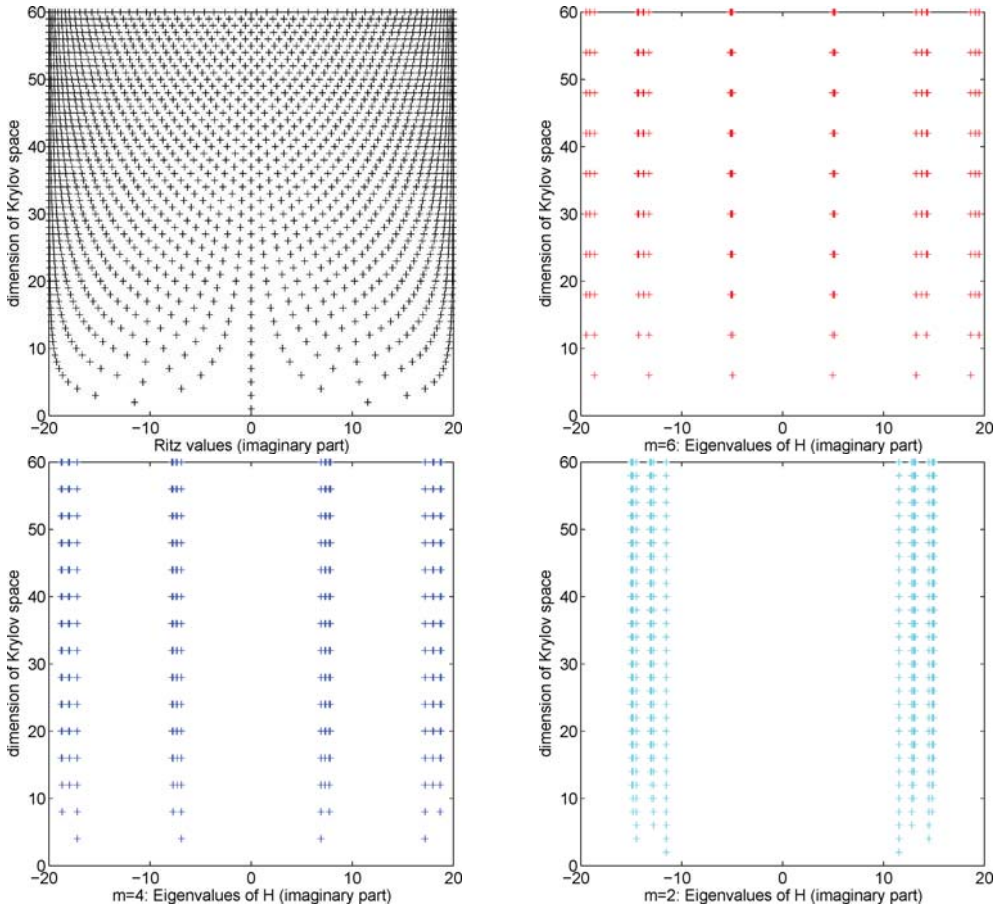


FIG. 5.3. *Interpolation nodes for the skew-symmetric problem of dimension* $n = 1001$.

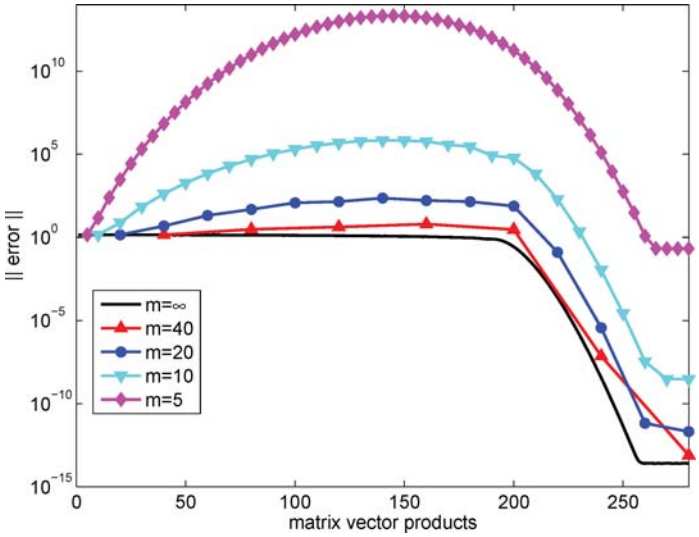


FIG. 5.4. *Error norm histories for the skew-symmetric problem of dimension* $n = 10001$.

TABLE 5.4
The restarted Arnoldi approximation applied to the skew-symmetric problem of dimension 10001*, for several restart lengths* m *(cf. Figure* 5.4*).*

m	Matrix vector products	Time[s]	Final accuracy	Largest error	Largest error / Final accuracy
∞	260	367	2.5e−14	1.4e+01 [1]	1.8e−14
40	280	48	7.8e−14	6.3e+01 [160]	1.2e−14
20	280	26	2.1e−12	2.3e+02 [140]	8.9e−15
10	270	16	2.9e−09	6.8e+05 [140]	4.3e−15
5	275	13	2.1e−01	2.2e+13 [145]	9.7e−15

Table 5.4 shows the number of matrix-vector products and the execution times which were required to reach a final accuracy for different restart length m. We also list the largest intermediate error (and after how many matrix-vector multiplications it is observed). Note that for every m the quotient of this largest error and the final accuracy approximately equals the machine precision of $2e-16$.

5.3. Convection-diffusion problem. Our final example is taken from [19, Example 6.1]: We consider the initial boundary value problem

$$\begin{aligned} \dot{u} - \Delta u + \tau_1 u_{x_1} + \tau_2 u_{x_2} &= 0 \quad \text{on } (0,1)^3 \times (0,T), \\ u(x,t) &= 0 \quad \text{on } \partial(0,1)^3 \text{ for all } t \in [0,T], \\ u(x,0) &= u_0(x), \quad x \in (0,1)^3. \end{aligned}$$

Discretizing the Laplacian by the usual seven-point stencil and the first-order derivatives, u_{x_1} and u_{x_2}, by central differences on a uniform grid with step size $h = 1/(n+1)$ leads—as in section 5.1—to an ordinary initial value problem

$$\begin{aligned} \dot{\boldsymbol{u}}(t) &= A\boldsymbol{u}(t), \quad t \in (0,T), \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0 \end{aligned}$$

with the matrix

$$A = I_n \otimes [I_n \otimes C_1] + [B \otimes I_n + I_n \otimes C_2] \otimes I_n$$

of dimension $N = n^3$. Here,

$$B = \tfrac{1}{h^2}\text{tridiag}(1,-2,1), \;\; C_j = \tfrac{1}{h^2}\text{tridiag}(1+\mu_j, -2, 1-\mu_j), \;\; j = 1,2,$$

where $\mu_j = \tau_j h/2$. The nonsymmetric matrix A is a popular test matrix because its eigenvalues are explicitly known: If $|\mu_j| > 1$ (for at least one j), they are complex; more precisely (cf. [19]),

$$\begin{aligned} \Lambda(A) \subset \tfrac{1}{h^2}\,[-6 - 2\cos(\pi h)\,\mathrm{Re}(\theta),\; -6 + 2\cos(\pi h)\,\mathrm{Re}(\theta)] \\ \times \tfrac{1}{h^2}\,[-2i\cos(\pi h)\,\mathrm{Im}(\theta),\; 2i\cos(\pi h)\,\mathrm{Im}(\theta)] \end{aligned}$$

with $\theta = 1 + \sqrt{1-\mu_1^2} + \sqrt{1-\mu_2^2}$. As in [19], we choose $h = 1/16$, $\tau_1 = 96$, $\tau_2 = 128$ ($\tau_1 = \tau_2 = 320$) which leads to $\mu_1 = 3$, $\mu_2 = 4$ ($\mu_1 = \mu_2 = 10$), and approximate $e^{tA}\boldsymbol{b}$, where $t = h^2$ and $\boldsymbol{b} = [1,1,\ldots,1]^\top$. The resulting error norm histories are shown in Figure 5.5. In the second example we again observe transient error growth. We attribute this, as in the skew-symmetric example, to the sufficiently large imaginary parts of the eigenvalues of $h^2 A$, which lie in

$$[-8.0, -4.0] \times i[-13.1, 13.1] \quad \text{for} \quad \mu_1 = 3, \mu_2 = 4$$

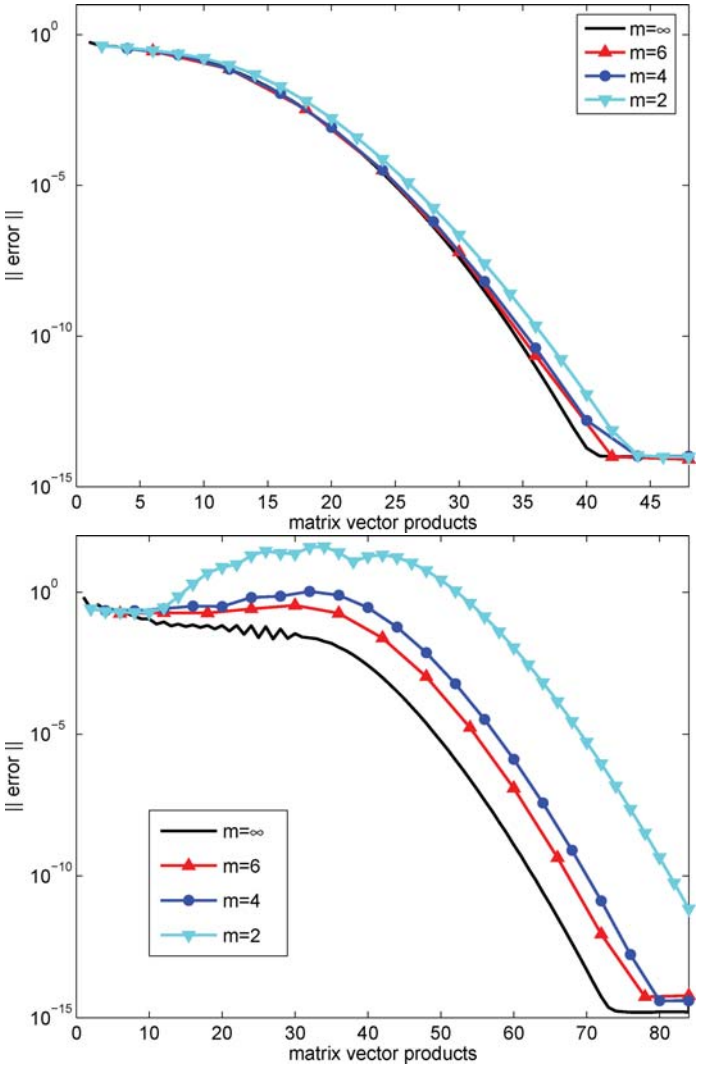


FIG. 5.5. *Error norm histories for the restarted Arnoldi approximation applied to the convection-diffusion problem with* $n = 15$*, i.e.,* $N = 3375$ *for several restart lengths* m*. As in* [19]*, we chose* $\mu_1 = 3$*,* $\mu_2 = 4$ *(top), and* $\mu_1 = \mu_2 = 10$ *(bottom).*

and

$$[-8.0, -4.0] \times i[-39.0, 39.0] \quad \text{for} \quad \mu_1 = \mu_2 = 10,$$

respectively.

6. Conclusions. We have shown how Krylov subspace methods for approximating $f(A)\boldsymbol{b}$ may be restarted. This permits the application of schemes like the Arnoldi approximation to very large matrices using a fixed amount of storage space. For functions f which are entire of order one, the restarted method retains the superlinear convergence property of the unrestarted method. In addition, we have identified the relationship of the restarted method to known algorithms in the cases $f(\lambda) = 1/\lambda$ and

$f_t(t\lambda) = e^{t\lambda}$. Moreover, we have demonstrated that the method performs well on several numerical examples from the literature. Related issues such as characterizing the convergence of the Arnoldi approximation using potential theoretic methods as well as yet more efficient implementations of the restarted algorithm will be the subject of future research.

REFERENCES

[1] E. J. Allen, J. Baglama, and S. K. Boyd, *Numerical approximation of the product of the square root of a matrix with a vector*, Linear Algebra Appl., 310 (2000), pp. 167–181.

[2] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[3] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.

[4] E. Celledoni and I. Moret, *A Krylov projection method for systems of ODEs*, Appl. Numer. Math., 24 (1997), pp. 365–378.

[5] P. I. Davies and N. J. Higham, *A Schur–Parlett algorithm for computing matrix functions*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 464–485.

[6] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.

[7] C. de Boor, *Divided differences*, Surv. Approximation Theory, 1 (2005), pp. 46–69.

[8] V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, Comput. Math. Math. Phys., 29 (1989), pp. 112–121.

[9] V. L. Druskin and L. A. Knizhnerman, *Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.

[10] M. Eiermann, O. G. Ernst, and O. Schneider, *Analysis of acceleration strategies for restarted minimal residual methods*, J. Comput. Appl. Math., 123 (2000), pp. 261–292.

[11] M. Freund and E. Görlich, *Polynomial approximation of entire functions and rate of growth of Taylor coefficients*, Proc. Edinb. Math. Soc., 28 (1985), pp. 341–348.

[12] R. W. Freund and M. Hochbruck, *Gauss-quadratures associated with the Arnoldi process and the Lanczos algorithm*, in Linear Algebra for Large-Scale and Real-Time Applications, M. S. Moonen, G. H. Golub, and B. L. R. de Moor, eds., Kluwer Academic Publishers, Dordrecht, 1993, pp. 377–380.

[13] R. W. Freund, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, J. Comput. Appl. Math., 43 (1992), pp. 135–158.

[14] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.

[15] F. R. Gantmacher, *The Theory of Matrices*, Vol. I, AMS Chelsea, Providence, RI, 1959.

[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[17] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[18] C. Moler and C. F. van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.

[19] I. Moret and P. Novati, *An interpolatory approximation of the matrix exponential based on Faber polynomials*, J. Comput. Appl. Math., 131 (2001), pp. 361–380.

[20] I. Moret and P. Novati, *Interpolating functions of matrices on zeros of quasi-kernel polynomials*, Numer. Linear Algebra Appl., 12 (2005), pp. 337–353.

[21] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.

[22] P. Novati, *A method based on Fejér points for the computation of functions of nonsymmetric matrices*, Appl. Numer. Math., 44 (2003), pp. 201–224.

[23] C. C. Paige, B. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.

[24] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1997.

[25] G. M. Phillips, *Interpolation and Approximation by Polynomials*, Springer-Verlag, New York, 2003.

[26] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[27] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[28] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.

[29] B. Singer and S. Spilerman, *The representation of social processes by Markov models*, Amer. J. Sociology, 8 (1976), pp. 1–54.
[30] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[31] D. E. Stewart and T. S. Leyk, *Error estimates for Krylov subspace approximations of matrix exponentials*, J. Comput. Appl. Math., 72 (1996), pp. 359–369.
[32] G. W. Stewart, *A Krylov–Schur algorithm for large eigenproblems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 601–614.
[33] H. Tal-Ezer, *High degree polynomial interpolation in Newton form*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 648–667.
[34] J. van den Eshof, A. Frommer, T. Lippert, and H. A. van der Vorst, *Numerical methods for the QCD overlap operator*. I. *Sign-function and error bounds*, Comput. Phys. Comm., 146 (2002), pp. 203–224.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2505–2538

DISCRETIZATION AND SIMULATION OF THE ZAKAI EQUATION*

EMMANUEL GOBET†, GILLES PAGÈS‡, HUYÊN PHAM‡, AND JACQUES PRINTEMS§

Abstract. This paper is concerned with numerical approximations for the stochastic partial differential Zakai equation of nonlinear filtering problems. The approximation scheme is based on the representation of the solutions as weighted conditional distributions. We first accurately analyze the error caused by an Euler-type scheme of time discretization. Sharp error bounds are calculated: we show that the rate of convergence is in general of order $\sqrt{\delta}$ (δ is the time step), but in the case when there is no correlation between the signal and the observation for the Zakai equation, the order of convergence becomes δ. This result is obtained by carefully employing techniques of Malliavin calculus. In a second step, we propose a simulation of the time discretization Euler scheme by a quantization approach. Formally, this consists in an approximation of the weighted conditional distribution by a conditional discrete distribution on finite supports. We provide error bounds and rate of convergence in terms of the number N of the grids of this support. These errors are minimal at some optimal grids which are computed by a recursive method based on Monte Carlo simulations. Finally, we illustrate our results with some numerical experiments arising from a correlated Kalman–Bucy filter.

Key words. stochastic partial differential equations, nonlinear filtering, Zakai equation, Euler scheme, quantization, Malliavin calculus

AMS subject classifications. 60H35, 60H15, 60G35, 60H07, 65C20

DOI. 10.1137/050623140

1. Introduction.

We are interested in numerical approximation for the measure-valued process V governed by the following stochastic partial differential equations (SPDE) written in weak form: for all test functions $f \in C_b^2(\mathbb{R}^d)$,

$$\begin{aligned}\langle V_t, f\rangle = \langle \mu_0, f\rangle &+ \int_0^t \langle V_s, Lf\rangle ds \\ &+ \int_0^t \langle V_s, hf + \gamma^\intercal \nabla f\rangle . dW_s,\end{aligned} \tag{1.1}$$

where μ_0 is an initial probability measure. We denote by $\mathcal{M}(\mathbb{R}^d)$ the set of finite signed measures on $\mathbb{R}^d$. Here L is the second-order differential operator,

$$Lf(x) = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(x)\partial^2_{x_i x_j} f(x) + \sum_{i=1}^{d} b_i(x)\partial_{x_i} f(x),$$

W is a q-dimensional Brownian motion, $a = (a_{ij})$ is a $d \times d$ matrix-valued function, $\gamma = (\gamma_{il})$ is a $d \times q$ matrix-valued function, $b = (b_i)$ is an $\mathbb{R}^d$-vector-valued function, and $h = (h_l)$ is an $\mathbb{R}^q$-vector-valued function defined on $\mathbb{R}^d$, in the form

$$\begin{aligned} a &= \sigma\sigma^\intercal + \gamma\gamma^\intercal, \\ b &= \beta + \gamma h, \end{aligned}$$

*Received by the editors January 21, 2005; accepted for publication (in revised form) June 23, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/sinum/44-6/62314.html
†LMC, ENSIMAG-INP Grenoble, Grenoble, France (emmanuel.gobet@imag.fr).
‡LPMA, Université Paris 6-Paris 7, Paris, France (gpa@ccr.jussieu.fr, pham@math.jussieu.fr).
§LAMA, Université Paris 12, Paris, France (printems@univ-paris12.fr).

for some $d\times d$ matrix-valued function $\sigma=(\sigma_{ij})$ and $\mathbb{R}^d$-vector-valued function $\beta=(\beta_i)$ on $\mathbb{R}^d$. The transpose and the scalar product are, respectively, denoted by $^\intercal$ and a dot. The Euclidean norm of a vector is denoted $|.|$, and one uses the norm $|\sigma|=\sqrt{\mathrm{Tr}(\sigma\sigma^\intercal)}$ for a matrix σ.

When the distribution V_t admits a density $v(t,x)$, one may usually rewrite (1.1) in the following form:

$$dv(t,x)=\left(\frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{x_ix_j}[a_{ij}(x)v(t,x)]-\sum_{i=1}^{d}\partial_{x_i}[b_i(x)v(t,x)]\right)dt + \left(h^\intercal(x)v(t,x)-\nabla[\gamma(x)v(t,x)]\right)dW_t. \tag{1.2}$$

Under appropriate conditions, it is proved in [21] that the solution V to (1.1) can be characterized through the following system of diffusions:

$$X_t = X_0+\int_0^t\beta(X_s)ds+\int_0^t\sigma(X_s)dB_s+\int_0^t\gamma(X_s)dW_s, \qquad X_0\rightsquigarrow\mu_0, \tag{1.3}$$

$$\xi_t=\exp(Z_t)=\exp\left(\int_0^t h(X_s).dW_s-\frac{1}{2}\int_0^t|h(X_s)|^2ds\right), \tag{1.4}$$

$$\langle V_t,f\rangle = E_W\left[f(X_t)\xi_t\right], \tag{1.5}$$

where B is an $\mathbb{R}^d$-Brownian motion independent of W, and E_W denotes the conditional expectation given W. We also denote by P_W the corresponding conditional probability.

Actually, (1.1) is the so-called Zakai equation arising from the nonlinear filtering problem: here, X given in (1.3) is a d-dimensional signal, and W is a q-dimensional observation process (with correlated noise when $\gamma\neq 0$) given by

$$W_t=\int_0^t h(X_s)ds+U_t,$$

on a probability space $(\Omega,\mathcal{F},P)$ equipped with filtration $(\mathcal{F}_t)$ under which B and U are independent Brownian motions. The nonlinear filtering problem consists in estimating the conditional distribution of X given W, i.e., we want to compute the measure-valued process π_t characterized by

$$\langle\pi_t,f\rangle=E^P[f(X_t)|\mathcal{F}_t^W],$$

where $\mathcal{F}_t^W$ is the filtration generated by the whole observation of W until t. Under suitable conditions, there exists a reference probability measure Q such that

$$\left.\frac{dP}{dQ}\right|_{\mathcal{F}_t}=\xi_t=\exp\left(\int_0^t h(X_s).dW_s-\frac{1}{2}\int_0^t|h(X_s)|^2ds\right),$$

and (B,W) are two independent Brownian motions under Q. By the Kallianpur–Striebel formula, we have

$$\langle\pi_t,f\rangle=\frac{\langle V_t,f\rangle}{\langle V_t,1\rangle},$$

where

$$\langle V_t, f\rangle = E_w^Q[f(X_t)\xi_t]$$

satisfies the Zakai equation (1.1). From now on, the symbol E will denote the expectation with respect to the probability Q.

1.1. A short discussion of related literature. Numerical approximations of the Zakai equation and more generally of SPDEs have been extensively studied in the literature. We cite the survey paper [17] and the references therein. Roughly summarizing, one may classify the following approaches:

– Approximations based on the analytic expression (1.2) vary from finite difference of finite elements methods, splitting up methods, or Galerkin's approximation. We cite, for instance, [33], [15], [16] for the finite difference method of the Zakai equation or SPDEs, and the recent paper [35] for the finite element method of SPDEs. For the splitting up method of the Zakai equation and SPDEs, see [4], [11], [23], [18]. See also [34] for a time discretization analysis of θ-schemes of parabolic-type SPDEs driven by a(n infinite-dimensional) Wiener process.

– A first algorithm based on some uniform quantization grids of the state process is mentioned in [20].

– Another point of view, developed and studied in [24] and [5], is based on the Wiener chaos decomposition of the solution to the Zakai equation. We mention also Wong–Zakai-type approximations considered in [19].

– The third approach is based on the probabilistic representation (1.5) of the solution as a weighted (or unnormalized) conditional distribution. For the Zakai equation of nonlinear filtering problem, papers [22] and [10] develop approximation methods by replacing the signal process by a finite state Markov chain on a uniform grid prescribed a priori. This method is somewhat equivalent to the finite difference method.

– The so-called particular Monte Carlo method is based on a particle approximation of the conditional distribution. It has recently given rise to extensive studies; see, for instance, [8], [6], [7] for the nonlinear filtering problem. We will compare some of our results to those obtained in [7] (in which the diffusion X does not depend on the observation process, i.e., $\gamma = 0$).

1.2. Contribution and organization of the paper. The first contribution of our work consists in accurately estimating the error due to time discretization on the conditional expectation (1.5). Without conditioning, classical results yield an error at most linear w.r.t. the time step δ (see, for instance, [3]). Here, the situation is unusual because of the conditional expectation, and our analysis makes clear the role of the correlation factor between the underlying process X and the observation process W. As concerns the proof, we use Malliavin calculus techniques, but the fact that we work conditionally to W induces some specific technicalities.

In a second part, we propose a simulation algorithm for the SPDE (1.1) based on an optimal quantization approach. Basically, this means a spatial discretization of the dynamics of the Euler time discretization (X_k, V_k) of (1.3)–(1.5) optimally fitted to its probabilistic features. To be more specific, we first recall some short background on optimal quantization of a random vector. Let $X : (\Omega, \mathcal{F}, P) \to \mathbb{R}^d$ be a random vector and let $\Gamma = \{x^1, \dots, x^N\}$ be a subset (or *grid*) of $\mathbb{R}^d$ having N elements. We approximate X by one of its Borel closest neighbor projections $\widehat{X}^\Gamma := \mathrm{Proj}_\Gamma(X)$ on Γ. Such a projection is canonically associated to a Voronoi tessellation $(C_i(\Gamma))_{1\le i\le N}$

that is a Borel partition of $\mathbb{R}^d$ satisfying for any $i = 1, \dots, N$

$$C_i(\Gamma) \subset \left\{ \xi \in \mathbb{R}^d : |\xi - x^i| = \min_j |\xi - x^j| \right\}.$$

Hence

$$\widehat{X}^\Gamma = \mathrm{Proj}_\Gamma(X) := \sum_{i=1}^N x^i \mathbf{1}_{\{X \in C_i(\Gamma)\}}.$$

As soon as $X \in L^p(\Omega, P, \mathbb{R}^d)$ the induced *L^p-quantization error* is given by

$$\|X - \widehat{X}^\Gamma\|_p = \left(E \min_{1 \le i \le N} |X - x^i|^p \right)^{\frac{1}{p}} < \infty.$$

The L^p-optimal N-quantization problem for X consists in finding a grid Γ^* which achieves the lowest L^p-quantization error among all grids of size at most N. Such an optimal grid does exist (see [14]), and its size is exactly N if the support of X is infinite; it is generally not unique (except in 1-dimension, where uniqueness holds when the distribution P_X of X has a log-concave density). The rate of convergence of the lowest L^p-quantization error as $N \to +\infty$ is ruled by the so-called Zador theorem (see [14]). For historical reasons, this theorem is usually stated with the pth power of the L^p-quantization error, known as the *L^p-distortion*.

THEOREM 1.1. *Assume that $X \in L^{p+\eta}(\Omega, P, \mathbb{R}^d)$ for some $\eta > 0$. Let f denote the probability density of the absolutely continuous part of its distribution P_X (f is possibly 0). Then,*

$$\lim_N \left(N^{\frac{p}{d}} \min_{|\Gamma| \le N} \|X - \widehat{X}^\Gamma\|_p^p \right) = J_{p,d} \|f\|_{\frac{d}{d+p}}.$$

The constant $J_{p,d}$ corresponds to the uniform distribution over $[0,1]^d$ and in that case the above $\lim_N$ also holds as an infimum.

The constant $J_{p,d}$ is unknown as soon as $d \ge 3$ although one knows that $J_{p,d} \sim (d/(2\pi e))^{\frac{p}{2}}$ as $d \to \infty$. This theorem says that the lowest L^p-quantization error goes to 0 at an $N^{-\frac{1}{d}}$-rate when $N \to \infty$. For more details about these results, we refer to [14] and the references therein.

From a computational viewpoint, no closed form is available for optimal quantization grids Γ^* except for some very specific 1-dimensional distributions like the uniform one. Several algorithms can be implemented to compute these optimal (or at least some efficient locally optimal) grids. Several of them rely on the differentiability of the L^p-distortion function as a function of the grid (viewed as an N-tuple of $(\mathbb{R}^d)^N$): if P_X is continuous, it is differentiable at any grid of size N and its gradient admits an integral representation with respect to the distribution of X. Consequently one may search for optimal grids by implementing a Newton–Raphson procedure (in 1-dimension) or a stochastic gradient descent (in d-dimension). These numerical aspects have been extensively investigated in [31] with special attention to the d-dim normal distribution. Efficient grids for these distributions are now available for many sizes in dimensions $d = 1$ up to 10 (which can be downloaded at www.quantification.finance-mathematique.com); the extension to the quantization of Markov chains, including its numerical aspects, has already been discussed in several papers for various fields

of applications, such as American option pricing, nonlinear filtering, or stochastic control (see, e.g., [1], [28], [30], or [29]).

We now briefly explain in this introduction how to apply the vector-quantization method to the Zakai SPDE (1.1). The process (X_k) is simply a time discretization of a diffusion independent of V. In particular, given an observation W, (X_k) can be easily simulated and the idea is to quantize optimally at each time step k the random vector X_k by a finite distribution $\hat{X}_k$. This provides in turn an approximation of (V_k) as the conditional distribution of $\hat{X}_k$, weighted by a Girsanov-like term.

Let us mention that this approach can be applied to a wider family of stochastic SPDEs, e.g., when the functions h and γ (and possibly β and σ in the diffusion process) depend upon V_t. This is the case of the stochastic McKean–Vlasov equation, where $h \equiv 0$ and $\gamma(x, V) = \int \bar{\gamma}(x, v)V(dv)$ (V positive measure). We refer to [13] for some theoretical and numerical developments on this equation.

Our main results concerning the rate of convergence can be summed up as follows. First we prove under some regularity assumptions that the error induced by a time discretization with step δ is in general of order $\sqrt{\delta}$, although in the case $\gamma = 0$ the order of convergence is improved to δ. As concerns spatial discretization error, we obtain $n^{\frac{3}{2}}/\bar{N}^{\frac{1}{d}}$ (where $\delta = T/n$ and $\bar{N} = N/n$ denotes the (average) size of the quantization grids used at every time step). Finally (when $\gamma \neq 0$), our global error term has the form

$$\frac{1}{\sqrt{n}} + \frac{n^{\frac{3}{2}}}{\bar{N}^{\frac{1}{d}}}.$$

Numerical experiments carried out in section 4 suggest that a significantly better space order holds true, such as (when $d = 1$) $\frac{c_1 + c_2 n + o(n)}{\bar{N}}$, where $c_2 \ll c_1$.

The finite element method applied to (1.2) would provide the same kind of rate (in [35] the Wiener process W is infinite-dimensional, which induces worst rates for time and space discretization). However, these methods require an implicit time integration in order to be stable. This requires us to invert an $N^d \times N^d$ linear system (even if it is sparse) at each time step, which becomes very expensive as the dimension d grows (say $d \geq 3$ or 4).

As concerns Monte Carlo methods based on interacting particles procedures like [8] or [6], the main difference of our approach in terms of complexity is that most parts of our computations (the quantization of the d-dimensional process X) can be made off-line. This compensates the dependency in d of its theoretical rate of convergence, at least in medium dimensions. Since the algorithm proposed here is similar to the quantized nonlinear filters developed in [28] from a computational point of view, we refer to the detailed discussion carried out in it.

The paper is organized as follows. Section 2 is devoted to the time discretization error of the SPDE (1.1). The above result is established using Malliavin calculus techniques. We describe precisely in section 3 the optimal quantization algorithm for the Zakai equation and we analyze the resulting error. Finally, we illustrate our results in section 4 with several simulations concerning the Zakai equation in the linear case.

2. Time discretization error. In this section, we study the error caused by a time discretization of the system (1.3)–(1.5) characterizing the solution to the SPDE (1.1) on a finite time interval $[0, T]$. We consider regular discretization times $t_k = k\delta$, $k = 0, \ldots, n$, where $\delta = T/n$ is the time step, and we denote $\phi(t) = \sup\{t_k : t_k \leq t\}$.

We then use an Euler scheme as follows:

$$\begin{aligned} X_t^\delta &= X_0 + \int_0^t \beta(X_{\phi(s)}^\delta)ds + \int_0^t \sigma(X_{\phi(s)}^\delta)dB_s + \int_0^t \gamma(X_{\phi(s)}^\delta)dW_s, \\ Z_t^\delta &= \int_0^t h(X_{\phi(s)}^\delta).dW_s - \frac{1}{2}\int_0^t |h(X_{\phi(s)}^\delta)|^2 ds, \\ \langle V_t^\delta, f\rangle &= E_W\left[f(X_t^\delta)\exp(Z_t^\delta)\right]. \end{aligned}$$

By denoting $\bar{X}_k = X_{t_k}^\delta$, $\bar{V}_k = V_{t_k}^\delta$, $\Delta\bar{B}_k = B_{t_k} - B_{t_{k-1}}$, $\Delta\bar{W}_k = W_{t_k} - W_{t_{k-1}}$, the Euler scheme reads at the discretization times t_k, $k = 0, \dots, n$,

$$\bar{X}_{k+1} = \bar{X}_k + \beta(\bar{X}_k)\delta + \sigma(\bar{X}_k)\Delta\bar{B}_{k+1} + \gamma(\bar{X}_k)\Delta\bar{W}_{k+1}, \tag{2.1}$$

$$\bar{X}_0 = X_0 \rightsquigarrow \mu_0, \tag{2.2}$$

$$\langle \bar{V}_k, f\rangle = E_W\left[f(\bar{X}_k)\exp\left(\sum_{j=0}^{k-1} g(\bar{X}_j, \Delta\bar{W}_{j+1})\right)\right], \tag{2.3}$$

where

$$g(x, \Delta\bar{W}) = h(x).\Delta\bar{W} - \frac{1}{2}|h(x)|^2\delta.$$

Denote by $\bar{P}_{k,W}(x, dx')$ the conditional probability of $\bar{X}_k$ given W and $\bar{X}_{k-1} = x$. From (2.1), we have

$$\bar{P}_{k,W}(x, dx') \rightsquigarrow \mathcal{N}\left(x + \beta(x)\delta + \gamma(x)\Delta\bar{W}_k, \delta\sigma(x)\sigma^\intercal(x)\right).$$

As usual, we set for any $f \in \mathcal{B}(\mathbb{R}^d)$ a set of bounded measurable functions on $\mathbb{R}^d$,

$$\bar{P}_{k,W} f(x) = E_W\left[f(\bar{X}_k)\middle|\, \bar{X}_{k-1} = x\right] = \int f(x')\bar{P}_{k,W}(x, dx'),$$

for any $x \in \mathbb{R}^d$. Hence, by the distribution of iterated conditional expectations, we have the following inductive formula for $\bar{V}_k$, $k = 0, \dots, n$:

$$\langle \bar{V}_{k+1}, f\rangle = \langle \bar{V}_k, \exp\left(g(., \Delta\bar{W}_{k+1})\right)\bar{P}_{k+1,W} f\rangle, \tag{2.4}$$

$$\bar{V}_0 = \mu_0. \tag{2.5}$$

We denote by $BL_1(\mathbb{R}^d)$ the unit ball of bounded Lipschitz functions on $\mathbb{R}^d$,

$$BL_1(\mathbb{R}^d) = \{f : \mathbb{R}^d \mapsto \mathbb{R} \text{ satisfying } |f(x)| \le 1 \text{ and } |f(x) - f(y)| \le |x - y| \ \forall\ x, y\},$$

and we consider the metric

$$\rho(V_1, V_2) = \sup\left\{|\langle V_1, f\rangle - \langle V_2, f\rangle|, \quad f \in BL_1(\mathbb{R}^d)\right\}$$

on $\mathcal{M}(\mathbb{R}^d)$ for any $V_1, V_2 \in \mathcal{M}(\mathbb{R}^d)$.

2.1. Main results. To simplify the following convergence analysis, we assume that the coefficients are very smooth and that they satisfy a uniform ellipticity condition.

(H1) (i) The functions β, σ, and γ are of class C^∞ with bounded derivatives.

(ii) The function h is of class C^∞ and is bounded, as are its derivatives.

(iii) For some $\epsilon_0 > 0$, one has $\sigma\sigma^\intercal(x) \geq \epsilon_0$ Id uniformly in x.

We recall some notation from [12]. We set $X_t^{\delta,\lambda} = X_t^\delta + \lambda(X_t - X_t^\delta)$ and $e^{\bar{Z}_T^\delta} = \int_0^1 e^{Z_T^\delta + \lambda(Z_T - Z_T^\delta)} d\lambda$. In addition, for any smooth function $a : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ we denote its derivative by a', which is $\mathbb{R}^{d'} \otimes \mathbb{R}^d$-valued. Finally, we repeatedly use the notation $a'(t) = \int_0^1 a'(X_t^{\delta,\lambda}) d\lambda$. Now, consider the unique solution of the linear equation $\mathcal{E}_t = \mathrm{Id} + \int_0^t \beta'(s)\mathcal{E}_s \, ds + \sum_{j=1}^d \int_0^t \sigma_j'(s)\mathcal{E}_s \, dB_s^j + \sum_{j=1}^q \int_0^t \gamma_j'(s)\mathcal{E}_s \, dW_s^j$ (as usual, σ_j and γ_j are the jth column of the matrix σ and γ). Then, Lemma 4.3 in [12] gives

$$
\begin{aligned}
X_t - X_t^\delta = \mathcal{E}_t \int_0^t \mathcal{E}_s^{-1} \Big\{ & [\beta(X_s^\delta) - \beta(X_{\phi(s)}^\delta)] \\
& - \sum_{j=1}^d \sigma_j'(s)[\sigma_j(X_s^\delta) - \sigma_j(X_{\phi(s)}^\delta)] \\
& - \sum_{j=1}^q \gamma_j'(s)[\gamma_j(X_s^\delta) - \gamma_j(X_{\phi(s)}^\delta)] \Big\} \, ds \\
& + \sum_{j=1}^d \mathcal{E}_t \int_0^t \mathcal{E}_s^{-1}[\sigma_j(X_s^\delta) - \sigma_j(X_{\phi(s)}^\delta)] \, dB_s^j \\
& + \sum_{j=1}^q \mathcal{E}_t \int_0^t \mathcal{E}_s^{-1}[\gamma_j(X_s^\delta) - \gamma_j(X_{\phi(s)}^\delta)] \, dW_s^j.
\end{aligned} \tag{2.6}
$$

For any $f \in BL_1(\mathbb{R}^d)$, we put $f_\delta(x) = E(f(x + \delta\tilde{B}_T))$, where $\tilde{B}$ is an extra d-dimensional Brownian motion independent on B and W. Clearly, f_δ is of class C_b^∞, $\|f_\delta\|_\infty + \sup_{x \neq y} \frac{|f_\delta(x) - f_\delta(y)|}{|x-y|} \leq C$, $\|f_\delta - f\|_\infty \leq C\delta$, both estimates being uniform in $BL_1(\mathbb{R}^d)$.

The main result of this section is the following.

THEOREM 2.1. *Assume* (H1). *For $f \in BL_1(\mathbb{R}^d)$, set*

$$
\begin{aligned}
A_1(f) &= -e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \left[\sum_{j=1}^q \int_0^T \left(\mathcal{E}_s^{-1} \int_{\phi(s)}^s \gamma_j'(X_r^\delta)\gamma(X_{\phi(r)}^\delta) dW_r \right) dW_s^j \right], \\
A_2(f) &= -e^{\bar{Z}_T^\delta} f(X_T) \left(\sum_{i=1}^q \int_0^T \left[\int_{\phi(s)}^s h_i'(X_r^\delta)\gamma(X_{\phi(r)}^\delta) dW_r \right] dW_s^i \right), \\
A_3(f) &= -\sum_{i,j=1}^q f(X_T) e^{\bar{Z}_T^\delta} \left(\int_0^T h_i'(s)\mathcal{E}_s \left(\int_0^s \mathcal{E}_r^{-1} \left[\int_{\phi(r)}^r \gamma_j'(X_u^\delta)\gamma(X_{\phi(u)}^\delta) dW_u \right] dW_r^j \right) dW_s^i \right), \\
A_4(f) &= \frac{1}{2} e^{\bar{Z}_T^\delta} f(X_T) \int_0^T \left[(\|h\|^2)'(s)\mathcal{E}_s \left(\sum_{j=1}^q \int_0^s \mathcal{E}_r^{-1} \left(\int_{\phi(r)}^r \gamma_j'(X_u^\delta)\gamma(X_{\phi(u)}^\delta) dW_u \right) dW_r^j \right) \right] ds.
\end{aligned}
$$

Then, one has

$$
\left\| \rho(V_T, V_T^\delta) \right\|_2 \leq C\delta + \sup_{f \in BL_1(\mathbb{R}^d)} \left\| E_W \left[A_1(f) + A_2(f) + A_3(f) + A_4(f) \right] \right\|_2,
$$

with

$$\sup_{f\in BL_1(\mathbb{R}^d)} \|E_W(A_1(f)+A_2(f)+A_3(f)+A_4(f))\|_2 \le C\sqrt{\delta}$$

for some constant C.

Remark 2.1. The fact that $\sqrt{\delta}$ is an upper bound for the error is clear, if we use classic L^p-estimates between X and X^δ. But we know that this argument involving pathwise errors is not optimal when errors on laws are considered [3]. The result above makes clear the role of the correlation in the error on conditional expectations.

1. When there is no correlation between signal and observation, i.e., $\gamma = 0$ (which is not really relevant in a filtering problem), the four terms $A_i(f)$, $i = 1, \dots, 4$, vanish and the rate of convergence for the approximation of V_T is of order δ, the time discretization step.

2. For constant function γ, the three contributions $A_1(f), A_3(f), A_4(f)$ vanish and there remains $A_2(f)$ of order $\sqrt{\delta}$ coming from the approximation of e^{Z_T}.

3. In the general case, the error will be inexorably of order $\sqrt{\delta}$. Indeed, main contributions in the error essentially behave like $\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}}(W_s - W_{t_i})dW_s = \frac{1}{2}\sum_{i=0}^{n-1}\big([W_{t_{i+1}} - W_{t_i}]^2 - [t_{i+1} - t_i]\big)$, where the L^2-norm equals $C\sqrt{\delta}$.

2.2. Proof of Theorem 2.1. The proof relies on Malliavin calculus techniques: we refer the reader to [26], from which we borrow our notation. For technical reasons, it will be useful to work with the extended Wiener process

$$\mathcal{W} = \begin{pmatrix} B \\ \tilde{B} \\ W \end{pmatrix};$$

all the further Malliavin calculus computations are made relative to $\mathcal{W}$. Set $H = L^2([0,T],\mathbb{R}^{2d+q})$ and denote $\tilde{X}_t^{\delta,\lambda} = X_t^{\delta,\lambda} + \frac{\delta}{\sqrt{2}}\tilde{B}_t$. For $F \in \mathbb{D}^{1,p}$, we write $\mathcal{D}F = (\mathcal{D}^B F, \mathcal{D}^{\tilde{B}} F, \mathcal{D}^W F)$ for the components relative to the three Brownian motions B, $\tilde{B}$, and W; the partial Malliavin covariance matrix of F is denoted by $\gamma^F = \int_0^T [\mathcal{D}_t^B F, \mathcal{D}_t^{\tilde{B}} F, 0][\mathcal{D}_t^B F, \mathcal{D}_t^{\tilde{B}} F, 0]^\intercal dt = \int_0^T \mathcal{D}_t^B F[\mathcal{D}_t^B F]^\intercal dt + \int_0^T \mathcal{D}_t^{\tilde{B}} F[\mathcal{D}_t^{\tilde{B}} F]^\intercal dt$ (see section 2.1 in [26]). Following section 1.3 in [26], the Skorokhod integral, i.e., the adjoint operator of $\mathcal{D}$, is denoted by $\boldsymbol{\delta}$ (with a boldface symbol to avoid confusion with the time step δ). For a process u in the domain of $\boldsymbol{\delta}$, for its Skorokhod integral we write $\boldsymbol{\delta}(u)$ and $\int_0^T u_t \boldsymbol{\delta}\mathcal{W}_t$ as well.

As in section 4.5.2 of [12], a localization factor $\psi_T^\delta \in [0,1]$ will be needed in the control of residual terms to justify integration by parts formulas. It satisfies the following properties:

(a) For any integers k and p, $\psi_T^\delta \in \mathbb{D}^{k,p}$ and $\sup_\delta \|\psi_T^\delta\|_{\mathbb{D}^{k,p}} \le \frac{C}{T^q}$ for some $C, q \ge 0$.

(b) For any $k \ge 1$, there are $C, q \ge 0$ such that $P(\psi_T^\delta \ne 1) \le \frac{C}{T^q}\delta^k$.

(c) $\{\psi_T^\delta \ne 0\} \subset \{\forall \lambda \in [0,1] : \det(\gamma^{\tilde{X}_T^{\delta,\lambda}}) \ge \frac{1}{2}\det(\gamma^{X_T})\}$.

We omit the details of its tedious construction and we simply refer to [12] (we mention that the nondegeneracy condition (H1) (iii) is used to get the above estimates with $1/T^q$, but it could also be replaced by a hypoellipticity-type assumption). To prepare the proof, we now state a series of technical results (justified later) which will help to derive a suitable stochastic analysis conditionally on W.

LEMMA 2.1. *In the following,* $\Phi(W)$ *stands for a functional measurable w.r.t.* W*, which belongs to* $\mathbb{D}^\infty$.

(i) *For any random variable* $Y \in L^2$, $E_W(Y)$ *is the unique random variable satisfying the equality* $E(Y\Phi(W)) = E(E_W(Y)\Phi(W))$ *for any functional* $\Phi(W) \in \mathbb{D}^\infty$.

(ii) *For any* $\Phi(W) \in \mathbb{D}^\infty$ *and* $F \in \mathbb{D}^{1,2}$, *one has* $\Phi(W)F \in \mathbb{D}^{1,1}$, *with* $\mathcal{D}^B(\Phi(W)F) = \Phi(W)\mathcal{D}^B F$ *and* $\mathcal{D}^{\tilde{B}}(\Phi(W)F) = \Phi(W)\mathcal{D}^{\tilde{B}}F$.

(iii) *For* $\Phi(W)$ *and* G *in* $\mathbb{D}^\infty$, $g \in C_b^\infty$, *and any multi-index* α, *one has*

$$\begin{cases} E\left(\Phi(W)\partial^\alpha g(X_T)G\right) = E\left(\Phi(W)g(X_T)H_\alpha(X_T,G)\right), \\ \|H_\alpha(X_T,G)\|_2 \le C\dfrac{\|G\|_{\mathbb{D}^{k,p}}}{T^q} \end{cases} \tag{2.7}$$

for some integers k, p, q. *Furthermore, if* $G = 0$ *on* $\{\psi_T^\delta = 0\}$, *then for any* $\lambda \in [0,1]$, *one has*

$$\begin{cases} E(\Phi(W)\partial^\alpha g(\tilde{X}_T^{\delta,\lambda})G) = E(\Phi(W)g(\tilde{X}_T^{\delta,\lambda})H_\alpha(\tilde{X}_T^{\delta,\lambda},G)), \\ \|H_\alpha(\tilde{X}_T^{\delta,\lambda},G)\|_2 \le C\dfrac{\|G\|_{\mathbb{D}^{k,p}}}{T^q} \end{cases} \tag{2.8}$$

with some constants C, k, p, q *uniform in* δ, *and* $\lambda \in [0,1]$.

The result below is one of the keys of our error analysis. The estimates of order δ are rather surprising. Indeed, at first glance, each stochastic integral (for fixed r) in the left-hand side of (2.9) is of order $\sqrt{\delta}$, but the mean over r helps in improving this estimate to get δ, provided that the processes g and h satisfy some suitable controls. Its proof is postponed until the end of this section.

PROPOSITION 2.1. *For* $g \in \mathbb{D}^\infty(H)$ *and* $h \in \mathbb{D}^\infty(H)$, *one has*

$$\begin{aligned} \int_0^T g_r \left(\int_{\phi(r)}^r h_u \boldsymbol{\delta}\mathcal{W}_u \right) dr &= \int_0^T \left(\int_0^T g_r h_u \mathbf{1}_{\phi(r)\le u\le r} dr \right) \boldsymbol{\delta}\mathcal{W}_u \\ &\quad + \int_0^T \left(\int_0^T \mathcal{D}_u g_r \cdot h_u \mathbf{1}_{\phi(r)\le u\le r} dr \right) du, \end{aligned} \tag{2.9}$$

and the above random variable belongs to $\mathbb{D}^\infty$. *Under extra assumptions, both terms in the right-hand side (r.h.s.) above are of order* δ.

(i) *Assume that* $N_{k,p}(g) = \sum_{j=0}^k \left[E\left(\int_0^T \|\mathcal{D}^j g_r\|^p_{L^p([0,T]^j)} dr\right)\right]^{1/p} < +\infty$ *and* $N_{k,p}(h) < +\infty$ *for any* k *and* p. *Then, the first term in the r.h.s. of* (2.9) *is of order* δ *in* $\mathbb{D}^{k,p}$, *for any* $k \in \mathbb{N}$ *and* $p > 1$:

$$\left\| \int_0^T \left(\int_0^T g_r h_u \mathbf{1}_{\phi(r)\le u\le r} dr \right) \boldsymbol{\delta}\mathcal{W}_u \right\|_{\mathbb{D}^{k,p}} \le C\ N_{k+1,q}(g) N_{k+1,q}(h)\ \delta \tag{2.10}$$

for some constants C *and* q *depending only on* k *and* p.

(ii) *Assume that* $M_{k,p}(g) = \sum_{j=1}^k \sup_{0\le r\le T} \left[E\|\mathcal{D}^j g_r\|^p_{L^p([0,T]^j)}\right]^{1/p} < +\infty$ *and* $N_{k,p}(h) < +\infty$ *for any* k *and* p. *Then, the second term in the r.h.s. of* (2.9) *is of order* δ *in* $\mathbb{D}^{k,p}$, *for any* $k \in \mathbb{N}$ *and* $p \ge 1$:

$$\left\| \int_0^T \left(\int_0^T \mathcal{D}_u g_r \cdot h_u \mathbf{1}_{\phi(r)\le u\le r} dr \right) du \right\|_{\mathbb{D}^{k,p}} \le C\ M_{k+1,q}(g)\ N_{k,q}(h)\ \delta \tag{2.11}$$

for some constants C *and* q *depending only on* k *and* p.

Let us turn to the proof of Theorem 2.1. It consists in proving

$$E(\Phi(W)[f(X_T^\delta)e^{Z_T^\delta} - f(X_T)e^{Z_T}]) \tag{2.12}$$
$$= E(\Phi(W)e^{Z_T^\delta}[(f-f_\delta)(X_T^\delta) - (f-f_\delta)(X_T)])$$
$$+ E(\Phi(W)e^{Z_T^\delta}[f_\delta(X_T^\delta) - f_\delta(X_T)]) \tag{2.13}$$
$$+ E(\Phi(W)f(X_T)[e^{Z_T^\delta} - e^{Z_T}]) \tag{2.14}$$
$$= E(\Phi(W)[A_1(f) + A_2(f) + A_3(f) + A_4(f) + R])$$

for any functional $\Phi(W) \in \mathbb{D}^\infty$, with $\|R\|_2 = O(\delta)$ uniformly w.r.t. $f \in BL_1(\mathbb{R}^d)$. Since $\|f - f_\delta\|_\infty \le C\delta$ for $f \in BL_1(\mathbb{R}^d)$, the term (2.12) can be neglected in our expansion.

In the following computations, we simply write Φ instead of $\Phi(W)$.

2.2.1. Contribution (2.13). A Taylor's formula combined with (2.6) and Ito's formula between $\phi(s)$ and s gives

$$E(\Phi e^{Z_T^\delta}[f_\delta(X_T^\delta) - f_\delta(X_T)]) \tag{2.15}$$
$$= E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha^{0,0}(u)du\right]ds\right) \tag{2.16}$$
$$+ E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha^{0,1}(u)dB_u\right]ds\right) \tag{2.17}$$
$$+ E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha^{0,2}(u)dW_u\right]ds\right) \tag{2.18}$$
$$+ E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha^{1,0}(u)du\right]dB_s\right) \tag{2.19}$$
$$+ \sum_{i=1}^d E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u\right]dB_s^i\right) \tag{2.20}$$
$$+ \sum_{i=1}^d E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha_i^{1,2}(u)dW_u\right]dB_s^i\right) \tag{2.21}$$
$$+ E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha^{2,0}(u)du\right]dW_s\right) \tag{2.22}$$
$$+ \sum_{i=1}^q \sum_{j=1}^d E\left(\Phi e^{Z_T^\delta} f_\delta'(T)\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\left[\int_{\phi(s)}^s \alpha_{i,j}^{2,1}(u)dB_u^j\right]dW_s^i\right) + E(\Phi A_1(f)), \tag{2.23}$$

where coefficients $\alpha_\cdot^\cdot \in \mathbb{D}^\infty(H)$ with $N_{k,p}(\alpha_\cdot^\cdot) + M_{k,p}(\alpha_\cdot^\cdot) < +\infty$ for any k, p, uniformly w.r.t. δ (actually, this is a consequence of the stronger estimate $\sup_{r\in[0,T]} \|\mathcal{D}^k_{s_1,\dots,s_k}\alpha_\cdot^\cdot(r)\|_p < \infty$; see, e.g., [12]). For instance, one can easily check that $\alpha_{i,j}^{2,1}(u) = -\gamma_i'(X^\delta_{\phi(u)})\sigma_j(X^\delta_{\phi(u)})$.

Terms in the factor of Φ in (2.15), (2.18), (2.21) clearly satisfy $\|R\|_2 = O(\delta)$ (recall that $\|f'\|_\infty \le C$ uniformly in $f \in BL_1(\mathbb{R}^d)$).

The contributions (2.16) and (2.17) give a contribution of order δ in L^p-norm by an application of estimates (2.10)–(2.11).

Terms in (2.19) contain most of the difficulties that we have to face in this error analysis; here, we give detailed arguments ((2.20) is handled in the same way). Note that $f_\delta(x) = E(f_{\delta/\sqrt{2}}(x + \frac{\delta}{\sqrt{2}}\tilde{B}_T))$ as well for the derivatives; thus, each term of the sum in (2.19) equals

$$\int_0^1 d\lambda E\left(\Phi\psi_T^\delta e^{Z_T^\delta} f'_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\Big[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u\Big]dB_s^i\right) \tag{2.24}$$

$$+\int_0^1 d\lambda E\left(\Phi(1-\psi_T^\delta)e^{Z_T^\delta} f'_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\Big[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u\Big]dB_s^i\right). \tag{2.25}$$

Since $P(\psi_T^\delta \neq 1) \leq C\frac{\delta^2}{T^q}$, (2.25) provides a negligible contribution. Besides, if we transform the Ito integral w.r.t. B^i into a Lebesgue integral, using the duality relationship (see section 1.3 in [26]) and property (ii) of Lemma 2.1, we obtain that (2.24) can be rewritten in the form

$$\int_0^1 d\lambda E\left(\Phi\int_0^T \mathcal{D}_s^{B^i}[\psi_T^\delta e^{Z_T^\delta} f'_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\mathcal{E}_T]\mathcal{E}_s^{-1}\Big[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u\Big]ds\right)$$
$$= \sum_{\kappa:|\kappa|=1,2}\int_0^1 d\lambda E\left(\Phi\ \partial_x^\kappa f_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\int_0^T \alpha_{\kappa,i}^{1,1}(s)\Big[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u\Big]ds\right),$$

where the summation holds on differentiation multi-indices κ with length equal to 1 and 2. In addition, the coefficients $\alpha_{\kappa,i}^{1,1}$ and $\alpha_{\kappa,i}^{1,1}$ satisfy $N_{k,p}(\alpha_{\kappa,i}^{1,1}) + M_{k,p}(\alpha_{\kappa,i}^{1,1}) < +\infty$ for any k and p. If we put $G = \int_0^T \alpha_{\kappa,i}^{1,1}(s)[\int_{\phi(s)}^s \alpha_i^{1,1}(u)dB_u]ds$, we remark that $G \in \mathbb{D}^\infty$, that $G = 0$ if $\psi_T^\delta = 0$ because of the local property of the derivative operator (Proposition 1.3.7 in [26]), and that $\|G\|_{\mathbb{D}^{k,p}} \leq C\delta$ by applying Proposition 2.1. Thus, Lemma 2.1 completes the estimate, and the factor of Φ in (2.24) is of order δ in L^2-norm, uniformly w.r.t. $f \in BL_1(\mathbb{R}^d)$.

We now consider (2.22). As for (2.19), we introduce ψ_T^δ; the term with $1-\psi_T^\delta$ can be neglected as before. Using analogous computations as above, it is straightforward to see that we have to control

$$\int_0^1 d\lambda E\left(\Phi\psi_T^\delta e^{Z_T^\delta} f'_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\mathcal{E}_T \int_0^T \mathcal{E}_s^{-1}\Big[\int_{\phi(s)}^s \alpha_{i,j}^{2,1}(u)dB_u^j\Big]dW_s^i\right)$$
$$= \int_0^1 d\lambda \int_0^T\int_0^T E\Big(\mathcal{D}_u^{B^j}[\mathcal{D}_s^{W^i}[\Phi\psi_T^\delta e^{Z_T^\delta} f'_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\mathcal{E}_T]\mathcal{E}_s^{-1}]\mathbf{1}_{\phi(s)\leq u\leq s}\alpha_{i,j}^{2,1}(u)\Big)du\ ds$$
$$= \sum_{\kappa:|\kappa|=1,2}\int_0^1 d\lambda E\left(\Phi\ \partial_x^\kappa f_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})\int_0^T\int_0^T \hat{\alpha}_{i,j}^{\kappa,2,1}(s)\mathbf{1}_{\phi(s)\leq u\leq s}\alpha_{i,j}^{2,1}(u)du\ ds\right)$$

(2.26)

$$+\sum_{\kappa:|\kappa|=1,2}\int_0^1 d\lambda E\left(\int_0^T \mathcal{D}_s^{W^i}[\Phi\ \partial_x^\kappa f_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda})]\Big(\int_0^T \alpha_{i,j}^{\kappa,2,1}(s)\mathbf{1}_{\phi(s)\leq u\leq s}\alpha_{i,j}^{2,1}(u)du\Big)ds\right).$$

(2.27)

For (2.26), it is enough to apply (2.8) with $G = \int_0^T\int_0^T \hat{\alpha}_{i,j}^{\kappa,2,1}(s)\mathbf{1}_{\phi(s)\leq u\leq s}\alpha_{i,j}^{2,1}(u)du\ ds$, which clearly satisfies $\|G\|_{\mathbb{D}^{k,p}} \leq C\delta$; this proves the expected estimate of order δ. The same conclusion holds for each term in (2.27): indeed, they can be transformed in

$\int_0^1 d\lambda E(\Phi\ \partial_x^\kappa f_{\delta/\sqrt{2}}(\tilde{X}_T^{\delta,\lambda}) \int_0^T (\int_0^T \alpha_{i,j}^{\kappa,2,1}(s)\mathbf{1}_{\phi(s)\le u\le s}\alpha_{i,j}^{2,1}(u)du)\boldsymbol{\delta}W_s^i)$ and we conclude with Lemma 2.1.

2.2.2. Contribution (2.14). It can be decomposed as $E(\Phi f(X_T)[e^{Z_T^\delta}-e^{Z_T}]) = E(\Phi f(X_T)e^{\bar{Z}_T^\delta}[Z_T^\delta - Z_T])$, that is,

$$E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\int_0^T [h(X_{\phi(s)}^\delta) - h(X_s^\delta)].dW_s\right)\right) \tag{2.28}$$

$$+ E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\int_0^T [h(X_s^\delta) - h(X_s)].dW_s\right)\right) \tag{2.29}$$

$$- \frac{1}{2}E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\int_0^T [\|h\|^2(X_{\phi(s)}^\delta) - \|h\|^2(X_s^\delta)]ds\right)\right) \tag{2.30}$$

$$- \frac{1}{2}E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\int_0^T [\|h\|^2(X_s^\delta) - \|h\|^2(X_s)]ds\right)\right). \tag{2.31}$$

In what follows, the main idea is to use Ito's formula and the stochastic expansion (2.6) to expand the differences $h(X_{\phi(s)}^\delta) - h(X_s^\delta)$, $h(X_s^\delta) - h(X_s)$, and so on. It will raise iterated stochastic integrals and, as before, the ones for which conditional expectation w.r.t. W is of order $\sqrt{\delta}$ are essentially of type $\int_0^T \cdots (\int_{\phi(s)}^s \cdots dW_u)dW_s$ (and not $\int_0^T \cdots (\int_{\phi(s)}^s \cdots dB_u)dW_s$ or $\int_0^T \cdots (\int_{\phi(s)}^s \cdots dW_u)dB_s$).

We now go into detail. Since (2.28) can be rewritten as $E(\Phi f(X_T)e^{\bar{Z}_T^\delta}(\sum_{i=1}^q \int_0^T [h_i(X_{\phi(s)}^\delta) - h_i(X_s^\delta)]dW_s^i))$, it equals

$$-E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\sum_{i=1}^q \int_0^T \left[\int_{\phi(s)}^s h_i'(X_r^\delta)\beta(X_{\phi(r)}^\delta)dr\right]dW_s^i\right)\right) \tag{2.32}$$

$$-E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\sum_{i=1}^q \int_0^T \left[\int_{\phi(s)}^s h_i'(X_r^\delta)\sigma(X_{\phi(r)}^\delta)dB_r\right]dW_s^i\right)\right) \tag{2.33}$$

$$-E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\sum_{i=1}^q \int_0^T \left[\int_{\phi(s)}^s h_i'(X_r^\delta)\gamma(X_{\phi(r)}^\delta)dW_r\right]dW_s^i\right)\right). \tag{2.34}$$

The factor of Φ in (2.32) clearly satisfies the required estimate and can be neglected. The term (2.33) can also be discarded from the main part of the error using the same arguments as for (2.22). Finally, the term (2.34) gives $A_2(f)$.

Term (2.29). Owing to (2.6), $\sum_{i=1}^q E(\Phi f(X_T)e^{\bar{Z}_T^\delta}(\int_0^T [h_i(X_s^\delta) - h_i(X_s)]dW_s^i))$ equals

$$-\sum_{i=1}^q \sum_{j=1}^d E\left(\Phi f(X_T)e^{\bar{Z}_T^\delta}\left(\int_0^T h_i'(s)\mathcal{E}_s \times\left(\int_0^s \mathcal{E}_r^{-1}\left[\int_{\phi(r)}^r \sigma_j'(X_u^\delta)\sigma(X_{\phi(u)}^\delta)dB_u\right]dB_r^j\right)dW_s^i\right)\right)$$

$$
\begin{aligned}
&-\sum_{i=1}^{q}\sum_{j=1}^{d} E\Bigg(\Phi f(X_T)e^{\bar{Z}_T^\delta}\Bigg(\int_0^T h_i'(s)\mathcal{E}_s \\
&\qquad\qquad \times\Bigg(\int_0^s \mathcal{E}_r^{-1}\Bigg[\int_{\phi(r)}^r \sigma_j'(X_u^\delta)\gamma(X_{\phi(u)}^\delta)dW_u\Bigg]dB_r^j\Bigg)dW_s^i\Bigg)\Bigg) \\
&-\sum_{i,j=1}^{q} E\Bigg(\Phi f(X_T)e^{\bar{Z}_T^\delta}\Bigg(\int_0^T h_i'(s)\mathcal{E}_s \\
&\qquad\qquad \times\Bigg(\int_0^s \mathcal{E}_r^{-1}\Bigg[\int_{\phi(r)}^r \gamma_j'(X_u^\delta)\sigma(X_{\phi(u)}^\delta)dB_u\Bigg]dW_r^j\Bigg)dW_s^i\Bigg)\Bigg) \qquad (2.35) \\
&-\sum_{i,j=1}^{q} E\Bigg(\Phi f(X_T)e^{\bar{Z}_T^\delta}\Bigg(\int_0^T h_i'(s)\mathcal{E}_s \\
&\qquad\qquad \times\Bigg(\int_0^s \mathcal{E}_r^{-1}\Bigg[\int_{\phi(r)}^r \gamma_j'(X_u^\delta)\gamma(X_{\phi(u)}^\delta)dW_u\Bigg]dW_r^j\Bigg)dW_s^i\Bigg)\Bigg) + E(\Phi R) \qquad (2.36)
\end{aligned}
$$

with $\|R\|_2 = O(\delta)$ by estimates (2.10)–(2.11). The term (2.36) gives $A_3(f)$, while the other contributions can be neglected. To justify this assertion, let us consider, for instance, (2.35), with techniques being the same for the other ones. First, we can replace f by f_δ since $\|f - f_\delta\|_\infty \le C\delta$. Then, three applications of the duality relationship yield

$$
\begin{aligned}
&E\Bigg(\Phi f_\delta(X_T)e^{\bar{Z}_T^\delta}\Bigg(\int_0^T h_i'(s)\mathcal{E}_s\Bigg(\int_0^s \mathcal{E}_r^{-1}\Bigg[\int_{\phi(r)}^r \gamma_j'(X_u^\delta)\sigma(X_{\phi(u)}^\delta)dB_u\Bigg]dW_r^j\Bigg)dW_s^i\Bigg)\Bigg) \\
&\quad = \int_0^T\int_0^T\int_0^T E(\mathcal{D}_u^B[\mathcal{D}_r^{W^j}[\mathcal{D}_s^{W^i}[\Phi f_\delta(X_T)e^{\bar{Z}_T^\delta}]h_i'(s)\mathcal{E}_s]\mathcal{E}_r^{-1}] \\
&\qquad \cdot\gamma_j'(X_u^\delta)\sigma(X_{\phi(u)}^\delta)\mathbf{1}_{\phi(r)\le u\le r})du\ dr\ ds.
\end{aligned}
$$

The term inside the expectation can be split into a sum involving the derivative of Φ and of f. Presumably, the more difficult term to estimate is of the form

$$
\int_0^T\int_0^T\int_0^T E(\mathcal{D}_r^{W^j}[\mathcal{D}_s^{W^i}[\Phi\ \partial_x^\kappa f_\delta(X_T)]]\alpha(u,r,s)\mathbf{1}_{\phi(r)\le u\le r})du\ dr\ ds.
$$

We omit the details for the other ones, which are easier to handle. Two integrations by parts with fixed W (see (iii) in Lemma 2.1) show that it equals

$$
E\Bigg(\Phi\ \partial_x^\kappa f_\delta(X_T)\int_0^T\Bigg(\int_0^T\Bigg(\int_0^T \alpha(u,r,s)\mathbf{1}_{\phi(r)\le u\le r}du\Bigg)\boldsymbol{\delta}W_r^j\Bigg)\boldsymbol{\delta}W_s^i\Bigg).
$$

Then, we conclude using (2.7) with $\|\int_0^T(\int_0^T(\int_0^T \alpha(u,r,s)\mathbf{1}_{\phi(r)\le u\le r}du)\boldsymbol{\delta}W_r^j)\boldsymbol{\delta}W_s^i\|_{\mathbb{D}^{k,p}} \le C\delta$.

Term (2.30). It yields a contribution of order δ, by an application of Ito's formula and inequalities (2.10)–(2.11). At last, the term (2.31) is equal to $-\frac{1}{2}\int_0^T E(\Phi f(X_T)e^{\bar{Z}_T^\delta}[\|h\|^2(X_s^\delta) - \|h\|^2(X_s)])ds$; in this form, the analysis is analogous to that of (2.13) and we omit the details. It gives the contribution $A_4(f)$ and some residual terms of order δ.

2.2.3. Proof of Lemma 2.1. The two first statements are straightforward. Statement (i) immediately follows from the fact that any $\Phi(W) \in L^2$ can be approximated in L^2 by a sequence of $\mathbb{D}^\infty$-r.v. using the chaos expansion (see Theorem 1.1.1 in [26]). Statement (ii) is clear from the definition of $\mathbb{D}^{1,p}$, $\mathcal{D}^B$, and $\mathcal{D}^{\tilde{B}}$.

Statement (iii) is an integration by parts formula that puts the differentiation/integration only on B and $\tilde{B}$, but not on W. Its proof is an easy adaptation of Proposition 3.2.1 in [27]. The estimate (2.7) is standard using in particular $\|[\gamma^{X_T}]^{-1}\|_p \le \frac{C}{T^q}$ under the nondegeneracy condition (H1) (iii) (see Theorem 3.3.1 in [27]). We only prove (2.8), which is less usual because of the localization factor G. Using (ii), one obtains the following equalities:

$$[\mathcal{D}^B(\Phi(W)g(\tilde{X}_T^{\delta,\lambda})), \mathcal{D}^{\tilde{B}}(\Phi(W)g(\tilde{X}_T^{\delta,\lambda}))] = \Phi(W)g'(\tilde{X}_T^{\delta,\lambda})[\mathcal{D}^B\tilde{X}_T^{\delta,\lambda}, \mathcal{D}^{\tilde{B}}\tilde{X}_T^{\delta,\lambda}],$$

$$\int_0^T \mathcal{D}_t(\Phi(W)g(\tilde{X}_T^{\delta,\lambda}))[\mathcal{D}_t^B\tilde{X}_T^{\delta,\lambda}, \mathcal{D}_t^{\tilde{B}}\tilde{X}_T^{\delta,\lambda}, 0]^\intercal dt = \Phi(W)g'(\tilde{X}_T^{\delta,\lambda})\gamma^{\tilde{X}_T^{\delta,\lambda}}.$$

Note that $\gamma^{\tilde{X}_T^{\delta,\lambda}} \ge \frac{\delta^2}{2}\mathrm{Id}$ and thus $\gamma^{\tilde{X}_T^{\delta,\lambda}}$ is invertible (it is the purpose of the small perturbation of $X^{\delta,\lambda}$ with $\delta\tilde{B}/\sqrt{2}$). Then, the duality relationship leads to

$$\begin{aligned}
&E(\Phi(W)\partial_{x_i}g(\tilde{X}_T^{\delta,\lambda})G)\\
&= E\left(\int_0^T \mathcal{D}_t(\Phi(W)g(\tilde{X}_T^{\delta,\lambda}))[Ge^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^B\tilde{X}_T^{\delta,\lambda}, Ge^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^{\tilde{B}}\tilde{X}_T^{\delta,\lambda}, 0]^\intercal dt\right)\\
&= E\left(\Phi(W)g(\tilde{X}_T^{\delta,\lambda})\int_0^T [Ge^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^B\tilde{X}_T^{\delta,\lambda}, Ge^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^{\tilde{B}}\tilde{X}_T^{\delta,\lambda}, 0]\boldsymbol{\delta}\mathcal{W}_t\right).
\end{aligned}$$

For longer multi-index α, we iterate the procedure and construct $H_\alpha(\tilde{X}_T^{\delta,\lambda}, G)$ by the recurrence formula $H_{\alpha'+[e^i]^\intercal}(\tilde{X}_T^{\delta,\lambda}, G) = \int_0^T [H_{\alpha'}(\tilde{X}_T^{\delta,\lambda}, G)e^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^B\tilde{X}_T^{\delta,\lambda}, H_{\alpha'}(\tilde{X}_T^{\delta,\lambda}, G)e^i\cdot[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathcal{D}_t^{\tilde{B}}\tilde{X}_T^{\delta,\lambda}, 0]\boldsymbol{\delta}\mathcal{W}_t$. Concerning the estimation on $\|H_\alpha(\tilde{X}_T^{\delta,\lambda}, G)\|_2$, note first that since the derivative operator and the Skorokhod integral are local (see Propositions 1.3.6 and 1.3.7 in [26]), one has $H_\alpha(\tilde{X}_T^{\delta,\lambda}, G) = H_\alpha(\tilde{X}_T^{\delta,\lambda}, G)\mathbf{1}_{\psi_T^\delta>0}$ owing to the property on G. Using the standard inequality $\|H_\alpha(\tilde{X}_T^{\delta,\lambda}, G)\mathbf{1}_A\|_p \le C\|[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathbf{1}_A\|_{q_1}^{p_1}\|\tilde{X}_T^{\delta,\lambda}\|_{k_2,q_2}^{p_2}\|G\|_{\mathbb{D}^{k_3,q_3}}$ (Proposition 2.4 in [3]) combined with $\|[\gamma^{\tilde{X}_T^{\delta,\lambda}}]^{-1}\mathbf{1}_{\psi_T^\delta>0}\|_p \le \frac{C}{T^q}$ (take into account property (c) of ψ_T^δ; see section 2.2), we easily complete the expected estimation.

2.2.4. Proof of Proposition 2.1. To prove (2.9), take $\Psi \in \mathbb{D}^\infty$ and write using Fubini's theorem twice and the duality relationship alternatively as follows:

$$\begin{aligned}
E\left(\Psi\int_0^T g_r\left(\int_{\phi(r)}^r h_u\boldsymbol{\delta}\mathcal{W}_u\right)dr\right) &= \int_0^T E\left(\Psi g_r\left(\int_{\phi(r)}^r h_u\boldsymbol{\delta}\mathcal{W}_u\right)\right)dr\\
&= \int_0^T\int_0^T E\big(\mathcal{D}_u[\Psi g_r]\mathbf{1}_{\phi(r)\le u\le r}\cdot h_u\big)du\,dr\\
&= \int_0^T E\left(\mathcal{D}_u\Psi\cdot\int_0^T g_r h_u\mathbf{1}_{\phi(r)\le u\le r}dr\right)du
\end{aligned}$$

$$+\int_0^T E\left(\Psi\int_0^T \mathcal{D}_u g_r\cdot h_u \mathbf{1}_{\phi(r)\le u\le r}dr\right)du$$

$$= E\left(\Psi\int_0^T\left(\int_0^T g_r h_u \mathbf{1}_{\phi(r)\le u\le r}dr\right)\boldsymbol{\delta}\mathcal{W}_u\right)$$

$$+E\left(\Psi\int_0^T\left(\int_0^T \mathcal{D}_u g_r\cdot h_u \mathbf{1}_{\phi(r)\le u\le r}dr\right)du\right).$$

It is standard to check that $\int_0^T g_r(\int_{\phi(r)}^r h_u\boldsymbol{\delta}\mathcal{W}_u)dr$ belongs to $\mathbb{D}^\infty$ (see Lemma 1.3.4 in [26]). The original feature of our result is specifically related to (2.10) and (2.11). For this, we use the following general estimates, which we prove at the end.

LEMMA 2.2. *For appropriately defined random variables* $(g_{r,s}, h_{u,s}, g_{r,s,u})_{r,s,u}$, *we have*

(2.37)

$$\left[E\left(\int_{[0,T]^j}ds\int_0^T du\left|\int_0^T g_{r,s}h_{u,s}\mathbf{1}_{\phi(r)\le u\le r}dr\right|^2\right)^{p/2}\right]^{1/p}$$

$$\le C_{p,q}(T)\ \delta\left[E\left(\int_{[0,T]^{j+1}}|h_{u,s}|^q\ duds\right)\right]^{1/q}\left[E\left(\int_{[0,T]^{j+1}}|g_{r,s}|^q\ drds\right)\right]^{1/q},$$

$$\left[E\left(\int_{[0,T]^j}ds\int_0^T du\left|\int_0^T g_{r,s,u}h_{u,s}\mathbf{1}_{\phi(r)\le u\le r}dr\right|^2\right)^{p/2}\right]^{1/p}$$

(2.38)

$$\le C_{p,q}(T)\delta\left[E\left(\int_{[0,T]^{j+1}}|h_{u,s}|^q\ duds\right)\right]^{1/q}\sup_{0\le r\le T}\left[E\left(\int_{[0,T]^{j+1}}|g_{r,s,u}|^q\ dsdu\right)\right]^{1/q}$$

for q large enough.

We are now in a position to derive (2.10). Consider first $k=0$. To control the L^p-norms of the first term in the r.h.s. of (2.9), we invoke the continuity of the Skorokhod integral (Proposition 2.4.3 in [27]) to get

(2.39)

$$\left\|\int_0^T\left(\int_0^T g_r h_u\mathbf{1}_{\phi(r)\le u\le r}dr\right)\boldsymbol{\delta}\mathcal{W}_u\right\|_p \le C\left(\left\|\int_0^T g_r h_\cdot \mathbf{1}_{\phi(r)\le\cdot\le r}dr\right\|_{L^p(\Omega,H)}\right.$$

$$\left.+\left\|\int_0^T \mathcal{D}(g_r h_\cdot)\mathbf{1}_{\phi(r)\le\cdot\le r}dr\right\|_{L^p(\Omega,H^{\otimes 2})}\right).$$

From (2.37), we easily get that the first term above is bounded by $N_{0,q}(h)N_{0,q}(h)\delta$ for q large enough. With analogous computations, the second term in the r.h.s. of (2.39) is bounded by $CN_{1,q}(h)N_{1,q}(h)\delta$. Estimates (2.10) have been proved when $k=0$. For $k\ge 1$, the successive derivatives of the r.h.s. of (2.9) are standard to compute and can be expressed in a similar form as before. Then, analogous computations can be performed and this proves (2.10) for any k. The derivation of (2.11) is analogous, using in addition (2.38).

Proof of Lemma 2.2. The Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\int_0^T du \left| \int_0^T g_{r,s,u} h_{u,s} \mathbf{1}_{\phi(r)\le u\le r} dr \right|^2 \\
&\quad \le \int_0^T du |h_{u,s}|^2 \left(\int_u^{\phi(u)+\delta} |g_{r,s,u}|\, dr \right)^2 \\
&\quad \le \left[\int_0^T du |h_{u,s}|^4 \right]^{1/2} \left[\int_0^T du \left(\int_u^{\phi(u)+\delta} |g_{r,s,u}|\, dr \right)^4 \right]^{1/2} \\
&\quad \le \delta^{3/2} \left[\int_0^T du |h_{u,s}|^4 \right]^{1/2} \left[\int_0^T du \int_u^{\phi(u)+\delta} |g_{r,s,u}|^4\, dr \right]^{1/2}.
\end{aligned}$$

If g does not depend on u, the last term above is bounded by $\delta^{1/2}[\int_0^T |g_{r,s}|^4\, dr]^{1/2}$. Then, the derivation of (2.37) is easy, using Hölder's inequalities. To obtain (2.38), i.e., when g depends on u, the previous computation to get the missing factor $\delta^{1/2}$ does not work directly; first, one has to integrate over s and ω, the other arguments remaining unchanged. ∎

3. Simulation of the Zakai equation and quantization error.

3.1. The quantization algorithm. In this section, we propose a quantization approach for the numerical implementation of formulas in (2.1), (2.3), and (2.5). Here, those formulas are written as

$$\begin{aligned}
\bar{X}_{k+1} &= \bar{X}_k + \beta(\bar{X}_k)\delta + \sigma(\bar{X}_k)\Delta\bar{B}_{k+1} + \gamma(\bar{X}_k)\Delta\bar{W}_{k+1} \\
&=: F_\delta(\bar{X}_k, \Delta\bar{B}_{k+1}, \Delta\bar{W}_{k+1}), \qquad (3.1)\\
\langle \bar{V}_{k+1}, f\rangle &= \langle \bar{V}_k, \exp\left(g(.,\Delta\bar{W}_{k+1})\right) \bar{P}_{k+1,W} f\rangle \qquad (3.2)
\end{aligned}$$

for $k = 0, \dots, n-1$, with

$$g(x,\Delta W) = h(x).\Delta W - \frac{1}{2}|h(x)|^2\delta, \qquad (3.3)$$

and $\bar{P}_{k+1,W}(x, dx')$ is a normal distribution with mean $x + \beta(x)\delta + \gamma(x)\Delta\bar{W}_{k+1}$ and variance $\sigma(x)\sigma^\intercal(x)\delta$.

We construct an approximation of $\bar{V}_k$ as follows. At each time t_k, $k = 0, \dots, n$, we are given the following grid $\Gamma_k = \{x_k^1, \dots, x_k^{N_k}\}$ of N_k points in $\mathbb{R}^d$, associated to Voronoi tessellations $C_i(\Gamma_k)$, $i = 1, \dots, N_k$:

$$C_i(\Gamma_k) = \left\{ u \in \mathbb{R}^d : |u - x_k^i| = \min_j |u - x_k^j| \right\}.$$

We then approximate the process $(\bar{X}_k)$ by the marginal quantized process $(\hat{X}_k)$ defined as

$$\hat{X}_k = \mathrm{Proj}_{\Gamma_k}(\bar{X}_k) := \sum_{i=1}^{N_k} x_k^i 1_{\{\bar{X}_k \in C_i(\Gamma_k)\}}.$$

We thus define the conditional probability $\hat{P}_{k,W}$ of $\hat{X}_k$ given $\hat{X}_{k-1}$ and W. In other words, $\hat{P}_{k,W}$ is a (random) probability transition matrix $\{\hat{p}_{k,W}^{ij},\ i = 1, \dots, N_{k-1}, j =$

$1, \dots, N_k\}$ characterized by

$$\hat{p}^{ij}_{k,W} = P_W\left[\hat{X}_k = x^j_k \middle| \hat{X}_{k-1} = x^i_{k-1}\right].$$

Finally, the random measure-valued process $(\bar{V}_k)$ is approximated by the discrete random measure process $(\hat{V}_k)$ defined by

$$\begin{aligned} \hat{V}_0 &= \text{law of} \hat{X}_0, \\ \langle \hat{V}_{k+1}, f\rangle &= \langle \hat{V}_k, \exp\left(g(.,\Delta\bar{W}_{k+1})\right)\hat{P}_{k+1,W} f\rangle. \end{aligned} \tag{3.4}$$

From an algorithmic viewpoint, this reads as

$$\hat{V}_k = \sum_{i=1}^{N_k} \hat{v}^i_k \delta_{x^i_k} \quad (\delta_x \text{ is the Dirac mass at } x)$$

for $k = 0, \dots, n$, where the weights $\hat{v}^i_k$ are computed in a forward induction as follows:

$$\begin{aligned} \hat{v}^i_0 = \hat{p}^i_0 &:= P[\hat{X}_0 = x^i_0] = P[\bar{X}_0 \in C_i(\Gamma_0)], \quad i = 1, \dots, N_0, \\ \hat{v}^j_{k+1} &= \sum_{i=1}^{N_k} \hat{v}^i_k \hat{p}^{ij}_{k+1,W} \exp\left(g(x^i_k, \Delta\bar{W}_{k+1})\right), \quad j = 1, \dots, N_{k+1}. \end{aligned}$$

The implementation of the above method requires optimally for each $k = 0, \dots, n$

- a grid Γ_k which minimizes the L^p-quantization error

$$\|\Delta_k\|_p = \|\bar{X}_k - \hat{X}_k\|_p$$

as well as an estimation of this error, and

- the weights of the joint distribution $(\hat{X}_{k-1}, \hat{X}_k)$ and marginal distribution $\hat{X}_{k-1}$,

$$\begin{aligned} \hat{r}^{ij}_{k,W} &= P_W\left[\hat{X}_k = x^j_k,\ \hat{X}_{k-1} = x^i_{k-1}\right] = P_W\left[\bar{X}_k \in C_j(\Gamma_k),\ \bar{X}_{k-1} \in C_i(\Gamma_{k-1})\right], \\ \hat{q}^i_{k-1,W} &= P_W\left[\hat{X}_{k-1} = x^i_k\right] = P_W\left[\bar{X}_{k-1} \in C_i(\Gamma_{k-1})\right] \end{aligned}$$

for $i = 1, \dots, N_{k-1}$, $j = 1, \dots, N_k$, so that

$$\hat{p}^{ij}_{k,W} = \frac{\hat{r}^{ij}_{k,W}}{\hat{q}^i_{k-1,W}}.$$

This program is achieved as follows:

– Monte Carlo simulation of M independent copies $(\bar{X}^{(m)}_0, \dots, \bar{X}^{(m)}_n)$, $m = 1, 2, \dots, M$, distributed according to $(\bar{X}_0, \dots, \bar{X}_n)$.

– Recursive optimization of the grids $\Gamma_0, \dots, \Gamma_n$ by a *competitive learning vector quantization* procedure and computation of the probability weights $\hat{r}^{ij}_{k,W}$ and $\hat{q}^i_{k-1,W}$, $k = 1, \dots, n$. As a byproduct, we also have an estimation of the L^2-quantization errors $\|\Delta_k\|_2$, $k = 0, \dots, n$.

3.2. Analysis of quantization error. The next theorem states an error estimation for the approximation of $\bar{V}_n$ under the following condition on the coefficients of the SDE X:

(H2) (i) The functions β, σ, and γ are Lipschitz.

(ii) The function h is bounded and Lipschitz.

THEOREM 3.1. *Under* (H2)*, for all* $p \in [1, +\infty)$ *and* $p' > p$*, there exists a positive real constant* $C_{p,p'}$ *such that*

$$\left\| \rho(\bar{V}_n, \hat{V}_n) \right\|_p \leq C_{p,p'} \frac{1}{\sqrt{\delta}} \sum_{k=0}^{n} \|\Delta_k\|_{p'} \qquad (\textit{with } \delta = T/n).$$

We first need the following classic result about the L^p-Lipschitz property of Euler schemes.

LEMMA 3.1. *Let* G_δ *be a functional in the form*

$$G_\delta(x, \varepsilon) = x + \delta B(x) + \sqrt{\delta}\, \Sigma(x)\varepsilon,$$

where B *and* Σ *are Lipschitz functions on* $\mathbb{R}^d$*, and* ε *is a Gaussian white noise. Then, for all* $p \in [1, \infty)$*, there exists a constant* C_p *such that for all* $x, x' \in \mathbb{R}^d$*,*

$$\|G_\delta(x, \varepsilon) - G_\delta(x', \varepsilon)\|_p \leq C_p(1 + \delta)|x - x'|.$$

We refer, e.g., to [30] for a detailed proof in a slightly more general setting where ε is only symmetric and lies in L^p.

One defines for every $k = 1, \ldots, n$ the operator $\bar{H}_{k,W}$ by

$$\bar{H}_{k,W}(f)(x) = \exp g(x, \Delta \bar{W}_k) \bar{P}_{k,W}(f)(x) \quad \forall f \in BL_1(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d,$$

where g is defined by (3.3). One defines

$$\bar{H}_{0,W}(f) = \langle \mu_0, f \rangle.$$

One easily checks that (with the former notations)

$$\langle \bar{V}_k, f \rangle = E_W(\bar{H}_{k,W}(f)(\bar{X}_{k-1})) = \langle \bar{V}_{k-1}, \bar{H}_{k,W}(f) \rangle$$

so that, for every $k = 0, \ldots, n$,

$$\langle \bar{V}_k, f \rangle \ = \ (\bar{H}_{0,W} \circ \bar{H}_{1,W} \circ \cdots \circ \bar{H}_{k,W})(f).$$

This equality can be written either in forward or backward recursive form. The backward form will be an important tool for proofs:

$$\begin{aligned} \bar{U}_{n,W} f &:= f, \\ \bar{U}_{k-1,W} f &:= \bar{H}_{k,W}(\bar{U}_{k,W} f), \qquad k = 1, \ldots n. \end{aligned} \tag{3.5}$$

Then, one checks *using the Markov property and the iterated conditional expectation rule* that

$$\bar{U}_{0,W} f = \langle \bar{V}_n, f \rangle.$$

For every $k = 1, \dots, n$, one approximates the operator $\bar{H}_{k,W}$ by its natural quantized counterpart $\widehat{H}_{k,W}$ defined on the grid $\Gamma_{k-1} = \{x_{k-1}^1, \dots, x_{k-1}^i, \dots, x_{k-1}^{N_{k-1}}\}$ by

$$\widehat{H}_{k,W}(f)(x_{k-1}^i) := \exp g(x_{k-1}^i, \Delta\bar{W}_k) \sum_j f(x_k^j) P_W(\widehat{X}_k = x_k^j \mid \widehat{X}_{k-1} = x_{k-1}^i)$$

so that

$$\widehat{H}_{k,W}(f)(\widehat{X}_{k-1}) = \exp g(\widehat{X}_{k-1}, \Delta\bar{W}_k)\, E_W(f(\widehat{X}_k) \mid \widehat{X}_{k-1}).$$

Then, one sets

$$\widehat{H}_{0,W}(f) := \sum_j f(x_0^j) P_W(\widehat{X}_0 = x_0^j).$$

We then notice that the approximation of $\bar{V}_k$ defined in (3.4) satisfies the following:

$$\langle \widehat{V}_k, f\rangle = (\widehat{H}_{0,W} \circ \widehat{H}_{1,W} \circ \cdots \circ \widehat{H}_{k,W})(f), \quad k = 1, \dots, n. \tag{3.6}$$

Once again, this equality can be read in backward form as follows:

$$\begin{aligned} &\widehat{U}_{n,W} f(x_n^i) := f(x_n^i), \quad i = 1, \dots, N_n, \\ &\widehat{U}_{k-1,W} f(x_{k-1}^i) := \widehat{H}_{k,W}(\widehat{U}_{k,W} f)(x_{k-1}^i), \quad i = 1, \dots, N_{k-1}, \ \ k = 1, \dots n, \end{aligned} \tag{3.7}$$

so that

$$\langle \widehat{V}_n, f\rangle = \widehat{U}_{0,W} f. \tag{3.8}$$

The proof is designed as follows: we wish to establish a backward induction between the error terms $\|\bar{U}_{k,W} f(\bar{X}_k) - \widehat{U}_{k,W} f(\widehat{X}_k)\|_p$ at successive times k and $k+1$ involving the quantization error $\|\bar{X}_{k+1} - \hat{X}_{k+1}\|_p$ of the Euler scheme. Unfortunately a naive approach makes the final error explode because of successive use of the Hölder inequality. So we are led to introduce a process $\bar{Y}_k$ starting at $\bar{X}_0$ but produced by a *biased* dynamics $G_{\delta,p}$ (instead of F_δ) which corresponds to a step-by-step discrete Girsanov (implicit) change of probability. Thus we can simultaneously take advantage of the martingale property of the Doléans exponential and of the independence property of the increments $\Delta\bar{W}_k$; it makes it possible not to use the Hölder inequality at a crucial step (see (3.15) below), which would cause an explosion of the constants. Finally, we use a revert Girsanov change of probability to come back to the quantization error of the original dynamics $(\bar{X}_k)$.

Proof of Theorem 3.1. We will assume for convenience that $\delta = T/n \in (0, 1]$ throughout the proof.

Step 1 (backward induction on the error $\|\bar{U}_{k,W} f(\bar{Y}_k) - \widehat{U}_{k,W} f(\widehat{Y}_k)\|_p$). Set temporarily

$$\begin{aligned} G_{\delta,p}(y, v, w) &:= F_\delta(y, v, w + p\delta h(y)) \\ &= y + \delta(\beta(y) + p\gamma(y)h(y)) + \sigma(y)v + \gamma(y)w, \\ \bar{Y}_k &:= G_{\delta,p}(\bar{Y}_{k-1}, \Delta\bar{B}_k, \Delta\bar{W}_k), \quad k \geq 1, \\ \bar{Y}_0 &= X_0, \\ \widetilde{Y}_k &:= F_\delta(\bar{Y}_{k-1}, \Delta\bar{B}_k, \Delta\bar{W}_k), \quad k \geq 1. \end{aligned}$$

Let $\bar{\mathcal{F}}_k$ denote the σ-field $\sigma(\Delta\bar{B}_\ell, \Delta\bar{W}_\ell, \ell = 1, \dots, k)$. Set, for every $k = 0, \dots, n$,

$$\hat{Y}_k := \mathrm{Proj}_{\Gamma_k}(\bar{Y}_k) \quad \text{and} \quad \widehat{\widetilde{Y}}_k := \mathrm{Proj}_{\Gamma_k}(\widetilde{Y}_k).$$

With these notations, one checks that for every $f \in BL_1(\mathbb{R}^d)$,

$$\bar{H}_{k,W}(f)(\bar{Y}_{k-1}) = \exp g(\bar{Y}_{k-1}, \Delta\bar{W}_k) E_W(f(\widetilde{Y}_k) \,|\, \bar{Y}_{k-1}) \tag{3.9}$$

and

$$\widehat{H}_{k,W}(f)(\widehat{Y}_{k-1}) = \exp g(\bar{Y}_{k-1}, \Delta\bar{W}_k) E_W(f(\widehat{\widetilde{Y}}_k) \,|\, \hat{Y}_{k-1}). \tag{3.10}$$

Consequently

$$\begin{aligned}
&\bar{U}_{k-1,W} f(\bar{Y}_{k-1}) - \widehat{U}_{k-1,W} f(\widehat{Y}_{k-1}) \\
&\quad = \bar{H}_{k,W}(\bar{U}_{k,W} f)(\bar{Y}_{k-1}) - \widehat{H}_{k,W}(\widehat{U}_{k,W} f)(\widehat{Y}_{k-1}) \\
&\quad = (\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) - E_w\big((\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) \,|\, \widehat{Y}_{k-1}\big) \\
&\qquad + E_w\big(\bar{H}_{k,W}(\bar{U}_{k,W} f)(\bar{Y}_{k-1}) - \widehat{H}_{k,W}(\widehat{U}_{k,W} f)(\widehat{Y}_{k-1}) \,|\, \widehat{Y}_{k-1}\big).
\end{aligned}$$

Let us deal with the above two terms successively. The random vector $\widehat{Y}_{k-1}$ being a function of $\bar{Y}_{k-1}$ and conditional expectation $E(\,.\,|\,W, \widehat{Y}_{k-1})$ being an L^p-contraction, one gets

$$\begin{aligned}
&\big\|\bar{U}_{k-1,W} f(\bar{Y}_{k-1}) - E_w\big((\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) | \widehat{Y}_{k-1}\big)\big\|_p \\
&\quad \le \big\|(\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) - (\bar{U}_{k-1,W} f)(\widehat{Y}_{k-1})\big\|_p \\
&\qquad + \big\|E_w\big((\bar{U}_{k-1,W} f)(\widehat{Y}_{k-1}) - (\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) \,|\, \widehat{Y}_{k-1}\big)\big\|_p \\
&\quad \le 2\big\|(\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) - (\bar{U}_{k-1,W} f)(\widehat{Y}_{k-1})\big\|_p.
\end{aligned}$$

Consequently, using the expressions (3.9) and (3.10) and once again the contraction property and the $\sigma(\bar{Y}_{k-1})$-measurability of $\hat{Y}_{k-1}$ yields

$$\begin{aligned}
&\big\|\bar{U}_{k-1,W} f(\bar{Y}_{k-1}) - \widehat{U}_{k-1,W} f(\widehat{Y}_{k-1})\big\|_p \\
&\quad \le 2\big\|(\bar{U}_{k-1,W} f)(\bar{Y}_{k-1}) - (\bar{U}_{k-1,W} f)(\widehat{Y}_{k-1})\ \big\|_p \\
&\qquad + \big\|e^{g(\bar{Y}_{k-1}, \Delta\bar{W}_k)}(\bar{U}_{k,W} f)(\widetilde{Y}_k) - e^{g(\widehat{Y}_{k-1}, \Delta\bar{W}_k)}(\widehat{U}_{k,W} f)(\widehat{\widetilde{Y}}_k)\big\|_p
\end{aligned} \tag{3.11}$$

(when $p = 2$, the 2 factor can be deleted). Let us deal now with the second term of the sum in the r.h.s. First note that

$$\begin{aligned}
&\big\|e^{g(\bar{Y}_{k-1}, \Delta\bar{W}_k)}(\bar{U}_{k,W} f)(\widetilde{Y}_k) - e^{g(\widehat{Y}_{k-1}, \Delta\bar{W}_k)}(\widehat{U}_{k,W} f)(\widehat{\widetilde{Y}}_k)\big\|_p \\
&\quad = \big\| \exp g(\bar{Y}_{k-1}, \Delta\bar{W}_k)\big(\bar{U}_{k,W} f(\widetilde{Y}_k) - \exp\big(g(\widehat{Y}_{k-1}, \Delta\bar{W}_k) \\
&\qquad - g(\bar{Y}_{k-1}, \Delta\bar{W}_k)\big)\widehat{U}_{k,W} f(\widehat{\widetilde{Y}}_k)\big)\big\|_p.
\end{aligned}$$

Set $L_p(\delta) := \exp\left((p-1)\|h\|_\infty^2\, \delta/2\right)$. A change of variable "à la Girsanov" yields for every nonnegative Borel function Θ and every $p \in (1, +\infty)$

$$\begin{aligned}
&\big\|\exp\left(g(\bar{Y}_{k-1}, \Delta\bar{W}_k)\right)\Theta(\bar{Y}_{k-1}, \Delta\bar{B}_k, \Delta\bar{W}_k)\big\|_p^p \\
&\quad \le (L_p(\delta))^p E\left(\exp\left(ph(\bar{Y}_{k-1}).\Delta\bar{W}_k - p^2|h(\bar{Y}_{k-1})|^2\delta/2\right)\Theta^p(\bar{Y}_{k-1}, \Delta\bar{B}_k, \Delta\bar{W}_k)\right) \\
&\quad \le (L_p(\delta))^p E\left(\Theta^p(\bar{Y}_{k-1}, \Delta\bar{B}_k, \Delta\bar{W}_k + p\delta h(\bar{Y}_{k-1}))\right)
\end{aligned}$$

so that

$$\begin{aligned}(3.12)\qquad &\left\|\exp\left(g(\bar{Y}_{k-1},\Delta\bar{W}_k)\right)\Theta(\bar{Y}_{k-1},\Delta\bar{B}_k,\Delta\bar{W}_k)\right\|_p\\ &\quad\le L_p(\delta)\|\Theta(\bar{Y}_{k-1},\Delta\bar{B}_k,\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\|_p.\end{aligned}$$

Applying the above inequality with $\Theta(y,v,w)=(\bar{U}_{k,W}f)(G_{\delta,p}(y,v,w))$ leads to

$$\begin{aligned}&\left\|e^{g(\bar{Y}_{k-1},\Delta\bar{W}_k)}(\bar{U}_{k,W}f)(\widetilde{Y}_k)-e^{g(\widehat{Y}_{k-1},\Delta\bar{W}_k)}(\widehat{U}_{k,W}f)(\widehat{\widetilde{Y}}_k)\right\|_p\\ &\le L_p(\delta)\Big\|\big(\bar{U}_{k,W}f(\bar{Y}_k)-\exp\big(g(\widehat{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\\ &\qquad -g(\bar{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\big)\widehat{U}_{k,W}f(\widehat{Y}_k)\big)\Big\|_p\\ &\le L_p(\delta)\left\|\bar{U}_{k,W}f(\bar{Y}_k)-\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_p\\ &\quad+L_p(\delta)\Big\|\big(1-\exp\big(g(\widehat{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\\ &\qquad -g(\bar{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\big)\big)\widehat{U}_{k,W}(f)(\widehat{Y}_k)\Big\|_p\\ &\le L_p(\delta)\left\|\bar{U}_{k,W}f(\bar{Y}_k)-\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_p\\ &\quad+L_p(\delta)\Big\|1-\exp\big(g(\widehat{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\\ &\qquad -g(\bar{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\big)\Big\|_{rp}\left\|\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_{sp},\end{aligned}\qquad(3.13)$$

where $r>1$ and $s=\frac{r}{r-1}$ are conjugate Hölder exponents. Now

$$\left\|\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_{sp}=\left\|\exp g(\widehat{Y}_k,\Delta\bar{W}_k)\,\widehat{U}_{k+1,W}f(\widehat{Y}_k)\right\|_{sp}.$$

Applying (3.12) (with sp) yields

$$\left\|\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_{sp}\le L_{sp}(\delta)\left\|\widehat{U}_{k+1,W}f(\widehat{Y}_{k+1}^{(sp)})\right\|_{sp}$$

for some $\bar{\mathcal{F}}_{k+1}$-measurable random vector $\widehat{Y}_{k+1}^{(sp)}$ which we have no need to specify (since f is bounded). One derives by induction that

$$\begin{aligned}(3.14)\qquad \left\|\widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_{sp}&\le(L_{sp}(\delta))^{n-k}\left\|\widehat{U}_{n,W}f(\widehat{Y}_n^{(sp)})\right\|_{sp}\\ &\le(L_{sp}(\delta))^{n-k}\|f\|_\infty\le C_{p,r,\|h\|_\infty,T}\|f\|_\infty\end{aligned}$$

with $K_{p,r,\|h\|_\infty,T}=\exp\left((sp-1)\|h\|_\infty^2T/2\right)$.

Let us deal now with the L^{rp}-norm of the exponential term. First, temporarily set $\widehat{\Delta}_k(h):=h(\widehat{Y}_k)-h(\bar{Y}_k)$. Then, standard computations show that

$$\begin{aligned}&\left\|1-\exp\Big(g(\widehat{Y}_{k-1},\Delta\bar{W}_k+p\,\delta\,h(\bar{Y}_{k-1}))-g(\bar{Y}_{k-1},\Delta\bar{W}_k+p\delta h(\bar{Y}_{k-1}))\Big)\right\|_{rp}\\ &\quad=\left\|1-\exp\Big((p-1)\delta h(\bar{Y}_{k-1}).\widehat{\Delta}_{k-1}(h)+\widehat{\Delta}_{k-1}(h)\Delta\bar{W}_k-|\widehat{\Delta}_{k-1}(h)|^2\delta/2\Big)\right\|_{rp}.\end{aligned}$$

Now using the elementary inequality $|e^x - 1| \le |x| e^{x_+}$, where $x_+ := \max(x, 0)$, and the fact that $x \mapsto x_+$ is nondecreasing yields

$$\begin{aligned}
&\big\|1 - \exp\big(g(\widehat{Y}_{k-1}, \Delta\bar{W}_k + p\,\delta\,h(\bar{Y}_{k-1})) - g(\bar{Y}_{k-1}, \Delta\bar{W}_k + p\delta h(\bar{Y}_{k-1}))\big)\big\|_{rp} \\
&\quad \le \big\| |\widehat{\Delta}_{k-1}(h)|(p-1)\delta h(\bar{Y}_{k-1}) + \Delta\bar{W}_k \\
&\qquad - (\widehat{\Delta}_{k-1}(h))\delta/2| \exp(2(p-1)\delta\|h\|_\infty^2 + 2\|h\|_\infty |\Delta\bar{W}_k|)\big\|_{rp} \\
&\quad \le L_{4p-3}(\delta)\sqrt{\delta}[h]_{\mathrm{Lip}} \big\| |\bar{Y}_{k-1} - \widehat{Y}_{k-1}|((p-1)\sqrt{\delta}\|h\|_\infty + |Z_k| + \|h\|_\infty\sqrt{\delta}) \\
&\qquad \times \exp(2\|h\|_\infty\sqrt{\delta}|Z_k|)\big\|_{rp},
\end{aligned}$$

where $Z_k := \frac{\Delta\bar{W}_k}{\sqrt{\delta}}$ is an $\mathcal{N}(0; I_d)$ random vector *independent* of $\bar{\mathcal{F}}_{k-1}$. Finally,

$$\begin{aligned}
&\big\|1 - \exp\big(g(\widehat{Y}_{k-1}, \Delta\bar{W}_k + p\delta h(\bar{Y}_{k-1})) - g(\bar{Y}_{k-1}, \Delta\bar{W}_k + p\delta h(\bar{Y}_{k-1}))\big)\big\|_{rp} \\
&\qquad \le C_{p,r,\delta,\|h\|_\infty,T}\sqrt{\delta}[h]_{\mathrm{Lip}}\big\|\widehat{Y}_{k-1} - \bar{Y}_{k-1}\big\|_{rp} \qquad (3.15)
\end{aligned}$$

with

$$C_{p,r,\delta,\|h\|_\infty,T} = L_{4p-3}(\delta)\big\|\big((p-1)\sqrt{\delta}\|h\|_\infty + |Z| + \sqrt{\delta}\|h\|_\infty\big)\exp\big(2\|h\|_\infty\sqrt{\delta}|Z|\big)\big\|_{rp}.$$

(Note that this real constant is increasing as a function of δ.) Plugging the estimates in (3.15) and (3.14) into (3.13) yields for every $k = 1, \dots, n$

$$\begin{aligned}
&\big\|e^{g(\bar{Y}_{k-1},\Delta\bar{W}_k)}(\bar{U}_{k,W}f)(\widetilde{Y}_k) - e^{g(\widehat{Y}_{k-1},\Delta\bar{W}_k)}(\widehat{U}_{k,W}f)(\widehat{\widetilde{Y}}_k)\big\|_p \\
&\qquad \le L_p(\delta)\big\|\bar{U}_{k,W}f(\bar{Y}_k) - \widehat{U}_{k,W}f(\widehat{Y}_k)\big\|_p + B(\delta)\big\|\bar{Y}_{k-1} - \widehat{Y}_{k-1}\big\|_{rp} \qquad (3.16)
\end{aligned}$$

with $B(\delta) := C_{p,r,\|h\|_\infty,T}\sqrt{\delta}[h]_{Lip}\|f\|_\infty$ (with $C_{p,r,\|h\|_\infty,T} = C_{p,r,1,\|h\|_\infty,T}K_{p,r,\|h\|_\infty,T}$ $L_p(1)$).

Now let us pass to the first term in the r.h.s. of (3.11). Let $(\bar{Y}_\ell^{k,y})_{\ell=k,\dots,n}$ be the sequence obtained by iterating $G_{p,\delta}(., \Delta\bar{B}_\ell, \Delta\bar{W}_\ell)$ from y at time $\ell = k$, i.e.,

$$\forall\, \ell \in \{k+1, \dots, n\}, \qquad \bar{Y}_\ell^{k,y} = G_{p,\delta}(\bar{Y}_{\ell-1}^{k,y}, \Delta\bar{B}_\ell, \Delta\bar{W}_\ell), \quad \bar{Y}_k^{k,y} := y.$$

The same proof as above shows that, for any couple (Z_{k-1}, Z'_{k-1}) of $\bar{\mathcal{F}}_{k-1}$-measurable L^p-integrable random variables,

$$\begin{aligned}
&\big\|(\bar{U}_{k-1,W}f)(Z_{k-1}) - (\bar{U}_{k-1,W}f)(Z'_{k-1})\ \big\|_p \\
&\quad \le L_p(\delta)\big\|\bar{U}_{k,W}(\bar{Y}_k^{k-1,Z_{k-1}}) - \bar{U}_{k,W}(\bar{Y}_k^{k-1,Z'_{k-1}})\big\|_p \\
&\qquad + B(\delta)\big\|\bar{Y}_{k-1}^{k-1,Z_{k-1}} - \bar{Y}_{k-1}^{k-1,Z'_{k-1}}\big\|_{rp},
\end{aligned}$$

so that by induction

$$\begin{aligned}
&\Big\|(\bar{U}_{k-1,W}f)(\bar{Y}_{k-1}) - (\bar{U}_{k-1,W}f)(\widehat{Y}_{k-1})\ \Big\|_p \\
&\quad \le B(\delta)\sum_{\ell=k}^{n}(L_p(\delta))^{\ell-k}\Big\|\bar{Y}_{\ell-1}^{k-1,\bar{Y}_{k-1}} - \bar{Y}_{\ell-1}^{k-1,\widehat{Y}_{k-1}}\Big\|_{rp} \\
&\qquad + (L_p(\delta))^{n+1-k}[f]_{\mathrm{Lip}}\Big\|\bar{Y}_n^{k-1,\bar{Y}_{k-1}} - \bar{Y}_n^{k-1,\widehat{Y}_{k-1}}\Big\|_{rp}.
\end{aligned}$$

Now, Lemma 3.1 (applied to $G_{\delta,p}$) implies the existence of a real constant $C_{rp} > 0$ such that

$$\left\|\bar{Y}_\ell^{k-1,\bar{Y}_{k-1}} - \bar{Y}_\ell^{k-1,\widehat{Y}_{k-1}}\right\|_{rp} \le (1 + C_{rp}\delta)^{\ell+1-k}\left\|\bar{Y}_{k-1} - \widehat{Y}_{k-1}\right\|_{rp}.$$

Setting $L'_{p,r}(\delta) = L_p(\delta)(1 + C_{rp}\delta)$ finally yields for every $k = 1, \dots, n$

$$\left\|(\bar{U}_{k-1,W}f)(\bar{Y}_{k-1}) - (\bar{U}_{k-1,W}f)(\widehat{Y}_{k-1})\ \right\|_p \le C(\delta)\left\|\bar{Y}_{k-1} - \widehat{Y}_{k-1}\right\|_{2p}$$

with

$$C(\delta) = L_p(T)e^{C_{rp}}\left(C_{p,r,\|h\|_\infty,T}\frac{[h]_{\mathrm{Lip}}\|f\|_\infty\sqrt{\delta}}{L'_{p,r}(\delta) - 1} + [f]_{\mathrm{Lip}}\right) \tag{3.17}$$

$$\le L_p(T)e^{C_{rp}}\left(C'_{p,r,\|h\|_\infty,T}\frac{[h]_{\mathrm{Lip}}\|f\|_\infty}{\sqrt{\delta}} + [f]_{\mathrm{Lip}}\right). \tag{3.18}$$

Plugging (3.16) and (3.17) into (3.11) finally yields the induction

$$\begin{aligned}\left\|\bar{U}_{k-1,W}f(\bar{Y}_{k-1}) - \widehat{U}_{k-1,W}f(\widehat{Y}_{k-1})\right\|_p &\le L_p(\delta)\left\|\bar{U}_{k,W}f(\bar{Y}_k) - \widehat{U}_{k,W}f(\widehat{Y}_k)\right\|_p \\ &\quad + A(\delta)\left\|\bar{Y}_{k-1} - \widehat{Y}_{k-1}\right\|_{rp}\end{aligned}$$

with

$$\begin{aligned}A(\delta) = B(\delta) + 2\,C(\delta) &\le C''_{p,r,\|h\|_\infty,T}\left([h]_{\mathrm{Lip}}\|f\|_\infty\left(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}\right) + [f]_{\mathrm{Lip}}\right) \\ &\le \frac{C_{p,r,\|h\|_\infty,[h]_{\mathrm{Lip}},\|f\|_\infty,[f]_{\mathrm{Lip}},T}}{\sqrt{\delta}}\end{aligned}$$

since $\delta \in (0, 1]$. A new induction leads to

$$\begin{aligned}\left\|\langle\bar{V}_n, f\rangle - \langle\hat{V}_n, f\rangle\right\|_p &= \left\|\bar{U}_{0,W}f(\bar{X}_0) - \widehat{U}_{0,W}f(\widehat{X}_0)\right\|_p \\ &= \left\|\bar{U}_{0,W}f(\bar{Y}_0) - \widehat{U}_{0,W}f(\widehat{Y}_0)\right\|_p \\ &\le A(\delta)\sum_{k=0}^{n}(L_p(\delta))^k\|\bar{Y}_k - (\widehat{U}_{n,W}f)(\widehat{Y}_n)\|_{rp} \\ &\quad + (L_p(\delta))^n\|(\bar{U}_{n,W}f)(\bar{Y}_n) - \widehat{Y}_k\|_p \\ &\le \frac{C_{p,r,\|h\|_\infty,[h]_{\mathrm{Lip}},\|f\|_\infty,[f]_{\mathrm{Lip}},T}}{\sqrt{\delta}}\sum_{k=0}^{n}\|\bar{Y}_k - \widehat{Y}_k\|_{rp} \\ &\quad + L_p(T)[f]_{\mathrm{Lip}}\|\bar{Y}_n - \widehat{Y}_n\|_{rp}.\end{aligned} \tag{3.19}$$

Step 2 (global revert Girsanov transform). Now, we aim to come back to $\bar{X}_k$ by introducing a revert Girsanov transform:

$$\|\bar{Y}_k - \widehat{Y}_k\|_{rp}^{rp} = E(Z_k(Z_k)^{-1}|\bar{Y}_k - \widehat{Y}_k|^{rp}),$$

where

$$Z_k = \exp\left(-\sum_{\ell=1}^{k} ph(\bar{Y}_{\ell-1}).\Delta\bar{W}_\ell - p^2|h(\bar{Y}_{\ell-1})|^2\frac{\delta}{2}\right).$$

It follows that

$$E(Z_k(Z_k)^{-1}|\bar{Y}_k - \widehat{Y}_k|^{rp}) = E\left(\exp\left(\sum_{\ell=1}^{k} ph(\bar{X}_{\ell-1}).\Delta\bar{W}_\ell - p^2|h(\bar{X}_{\ell-1})|^2\frac{\delta}{2}\right)|\bar{X}_k - \widehat{X}_k|^{rp}\right)$$

so that by the Hölder inequality applied with two conjugate exponents r', $s' > 1$,

$$\begin{aligned}\|\bar{Y}_k - \widehat{Y}_k\|_{rp}^{rp} &\le \left(E\exp\left(\sum_{\ell=1}^{k} s'ph(\bar{X}_{\ell-1}).\Delta\bar{W}_\ell - s'p^2|h(\bar{X}_{\ell-1})|^2\delta/2\right)\right)^{1/s'} \\ &\quad \cdot \left(E|\bar{X}_k - \widehat{X}_k|^{rr'p}\right)^{1/r'} \\ &\le \exp\left(k(s'-1)p^2\|h\|_\infty^2\delta/2\right)\|\bar{X}_k - \widehat{X}_k\|_{rr'p}^{rp}.\end{aligned}$$

Finally,

$$\|\bar{Y}_k - \widehat{Y}_k\|_{rp} \le \exp\left(kp\|h\|_\infty^2\delta/4\right)\|\bar{X}_k - \widehat{X}_k\|_{4p} \le C_{p,r,r',\|h\|_\infty,T}\|\bar{X}_k - \widehat{X}_k\|_{rr'p}.$$

One completes the proof by setting $r = r' = \sqrt{p'/p} > 1$ and plugging this last inequality into (3.19). □

3.3. Global error. Combining the results established in the former sections, we obtain the following result.

THEOREM 3.2. *Assume* (H1)–(H2). *Let* $p' > 2$ *and let* $N \ge n \ge 1$. *Assume that for every* $k \in \{0, \dots, n\}$, Γ_k *is an* $L^{p'}$*-optimal grid of size* $[N/(n+1)]$ *for* X_k. *There exists a real constant* C (*depending on* p' *but not* n) *such that*

$$\|\rho(V_T, \widehat{V}_n)\|_2 \le C\left(\frac{1}{n^\theta} + \frac{n^{\frac{3}{2}}}{\bar{N}^{\frac{1}{d}}}\right) \tag{3.20}$$

with $\theta = 0$ *if* $\gamma \equiv 0$ *and* $\theta = 1/2$ *otherwise, and* $\bar{N} = N/n$.

Proof. Combining results obtained in Theorems 2.1 and 3.1 yields the following:

$$\|\rho(V_T, \widehat{V}_n)\|_2 \le C\left(\frac{1}{n^\theta} + \sqrt{\delta}\sum_{k=0}^{n}\|\Delta_k\|_{p'}\right),$$

where $\Delta_k = X_k - \widehat{X}_k = X_k - \mathrm{Proj}_{\Gamma_k}(X_k)$. It follows from the nonparametric version of Zador's theorem, recently established in [25], that for every $p, \delta > 0$ there exists a universal real constant $C_{p,\delta}$ such that for every $N \ge 1$ and every $\mathbb{R}^d$-valued random vector Y,

$$\min_{\Gamma\subset\mathbb{R}^d, |\Gamma|\le N}\|Y - \widehat{Y}^\Gamma\|_p \le C_{p,\delta}\|Y\|_{p+\delta}N^{-\frac{1}{d}}.$$

Applying this result to our framework yields (with $\delta = 1$)

$$\begin{aligned}\sum_{k=0}^{n}\|\Delta\|_{p'} &\le C_{p'}\sup_n\max_{0\le k\le n}\|\bar{X}_k\|_{p'+1}(n+1)(N/(n+1))^{\frac{1}{d}} \\ &\le Cn^{\frac{3}{2}}(N/n)^{-\frac{1}{d}},\end{aligned}$$

where C is a finite real constant since we know that (b and σ, γ having at most linear growth) the family of Euler schemes $((\bar{X}_k)_{0\le k\le n})_{n\ge 1}$ satisfies $\sup_n \max_{0\le k\le n} \|\bar{X}_k\|_r < +\infty$ for any $r > 0$. □

Remark 3.1. • The $n^{\frac{3}{2}}$ in the spatial error term of (3.20) is most likely not optimal (see section 4). It probably comes from the specific technicalities induced by quantization. It corresponds, e.g., to the rate obtained for the "quenched error" in [7]. As shown by our numerical experiments, the spatial error term most likely behaves as $O(n \times (N/n)^{-\frac{1}{d}})$ or $O((N/n)^{-\frac{1}{d}})$, depending on some stability conditions between n and $\bar{N}$ (see section 4 for a detailed explanation).

• As an example, one can compare our error rate with that obtained in [7] (in the $\gamma \equiv 0$ setting) where an error bound is of the form

$$\frac{1}{n} + \sqrt{\frac{n}{M}},$$

where M denotes the number of Monte Carlo trials obtained under some regularity assumptions on the diffusion coefficients h and f (regardless of the dimension). In this case, M can be compared with our N/n, i.e., the mean value of points per time layers in our algorithm.

4. Numerical simulations and estimation of the rates of convergence. Since the expression of the global error given by (3.20) does not separate clearly the time and space parameters, we will try in this section to investigate separately the rate of convergence in time and in space in the following (linear) case:

$$\begin{aligned} \beta(x) = (A - \Gamma H)x, \quad h(x) = Hx, \\ \gamma(x) = \Gamma, \quad \sigma(x) = \Sigma, \end{aligned}$$

where A, Γ, Σ, and H are constant matrices of appropriate dimensions. We also suppose that μ_0 is a Gaussian law with mean m_0 and covariance matrix R_0. Then it is well known that the solution to the Zakai equation (1.1) is explicitly given by

$$\langle V_t, f\rangle = \left[\int f(\hat{m}_t + R(t)^{\frac{1}{2}}x)\frac{\exp\left(-\frac{1}{2}|x|^2\right)}{(2\pi)^{\frac{d}{2}}}dx\right]\langle V_t, 1\rangle, \tag{4.1}$$

where $R(t)$ is the solution to the Riccati equation

$$\begin{aligned} \frac{dR}{dt} &= AR + RA^\intercal + \Sigma\Sigma^\intercal + \Gamma\Gamma^\intercal - (RH^\intercal + \Gamma)(HR + \Gamma^\intercal), \\ R(0) &= R_0; \end{aligned} \tag{4.2}$$

$\hat{m}_t$ is the solution of

$$\begin{aligned} d\hat{m}_t &= A\hat{m}_t dt + (RH^\intercal + \Gamma)(dW_t - H\hat{m}_t dt), \\ \hat{m}_0 &= m_0; \end{aligned} \tag{4.3}$$

and

$$\langle V_t, 1\rangle = \exp\left(\int_0^t H\hat{m}_s.dW_s - \frac{1}{2}\int_0^t |H\hat{m}_s|^2 ds\right). \tag{4.4}$$

In other words, the normalized measure π_t defined by

$$\langle \pi_t, f \rangle = \frac{\langle V_t, f \rangle}{\langle V_t, 1 \rangle}$$

is a Gaussian distribution with mean $\hat{m}_t$ and variance $R(t)$.

We now introduce the quantized normalized filter for a given function $f \in BL_1(\mathbb{R})$ as

$$\langle \hat{\pi}_k^\delta, f \rangle := \frac{\langle \hat{V}_k, f \rangle}{\langle \hat{V}_k, 1 \rangle}, \quad k = 0, \dots, n,$$

where we have emphasized the dependence of the filter in $\delta = T/n$ by a superscript. The unnormalized filters $\hat{V}_k$ are computed according to algorithm (3.4).

The exact normalized filter is approximated owing to (4.1) using the following method. Since R is an explicitly known function (solution of (4.2)), it is sufficient to approximate $\hat{m}_t$, the solution of the SDE (4.3) with a refined Euler scheme of step, as

$$\delta_{ref} = \frac{T}{1024} \ll \delta.$$

Indeed, for each path of the observation W, (4.3) and (4.4) are discretized as

$$\bar{m}_{l+1} = \bar{m}_l + \delta_{ref} A \bar{m}_l + (R(l\delta_{ref})H^\intercal + \Gamma)(W_{(l+1)\delta_{ref}} - W_{l\delta_{ref}} - H\bar{m}_l \delta_{ref}), \tag{4.5}$$

$$\bar{Z}_{l+1} = \bar{Z}_l + H\bar{m}_l.(W_{(l+1)\delta_{ref}} - W_{l\delta_{ref}}) - \frac{1}{2}|H\bar{m}_l|^2\delta_{ref}, \quad \bar{\xi}_l = \exp(\bar{Z}_l), \tag{4.6}$$

and thus a very close approximation of the exact normalized filter, in the sense that it can be considered as the exact solution as long as δ remains considerably larger than δ_{ref}, is

$$\langle \pi_{l\delta_{ref}}^{\delta_{ref}}, f \rangle := \int f(\bar{m}_l + R(l\delta_{ref})^{\frac{1}{2}} x) \frac{\exp\left(-\frac{1}{2}|x|^2\right)}{(2\pi)^{\frac{d}{2}}} dx,$$

where $R(t)$ is computed owing to an exact quadrature formula.

We now estimate the rate of convergence of the scheme with respect to the spatial and time discretization. In order to smooth undesirable time oscillations of the error, we focus on the following temporal mean of the quadratic quantization error for the normalized filter, namely

$$\mathrm{Err}(\delta, \bar{N}) = \frac{1}{n} E \sum_{k=0}^{n} \left| \langle \hat{\pi}_k^\delta, f \rangle - \langle \pi_{t_k}^{\delta_{ref}}, f \rangle \right|^2, \tag{4.7}$$

where $t_k = k\delta = l(k)\delta_{ref}$ and $\bar{N} = N/n$ denotes the mean number of points per time layers. Then $\mathrm{Err}(\delta, \bar{N})$ is simply an approximation of the squared $L^2([0,T], dt)$-norm of the error.

We test the error for the following test functions:

$$f_0(x) = x, \quad f_1(x) = \exp(-x^2), \quad f_2(x) = \exp(-x). \tag{4.8}$$

The expectation in (4.7) is computed by a Monte Carlo method with $M = 100$ trajectories of the observations W.

The parameters of our simulations are

$$\Sigma = 1, \quad B = -0.5, \quad H = 1, \quad T = 1.$$

Such a choice of parameters is motivated by the fact that it provides values for $R(t)$ that are not too small. Otherwise, there would not be enough points around $m_0 = 0$ to be able to "capture" the behavior of the signal around its mean 0.

We will also change the model a bit and consider the following equations:

$$\begin{cases} dX_t = BX_t\,dt + \Sigma\,dB_t + \Gamma dW_t, \\ \\ dW_t = HX_t dt + \varepsilon dU_t. \end{cases} \tag{4.9}$$

The formulas above need to be changed as follows: $\Gamma \rightsquigarrow \varepsilon\Gamma$ and $H \rightsquigarrow H/\varepsilon$. The reason for introducing this new degree of freedom on the noise level may look paradoxical since small ε will provide large errors. But precisely, these large errors make it possible to display the rate of convergence more efficiently than with $\varepsilon = 1$, which produces smaller errors. Let us take the example of the spatial order. Indeed, we will see that as the discretization parameters $\bar{N}$ get larger and larger the error $\mathrm{Err}(\delta, \bar{N})$ is decreasing as a function of $\bar{N}$ until some threshold, depending a priori on δ and on the number M of observations (i.e., paths of W). Beyond this threshold, the error becomes more or less constant because the difference with the exact solution will be of the same order of the temporal discretization. Subsequently the sum of the two errors will become indistinguishable from the temporal one. Therefore, a small ε will provide bigger errors and so we will have more relevant points before reaching this threshold.

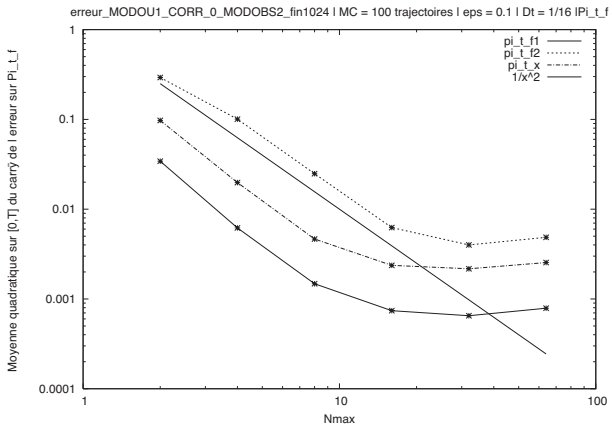
• *Estimation of the spatial discretization rate.* We first estimate the spatial rate of convergence in the case $\Gamma = 0$ (no correlation between the signal process X and the observation process W). For four values of $n = 1/\delta \in \{16, 32, 64, 256\}$, we estimate $\bar{N} \mapsto Err(\delta, \bar{N})$ with $\bar{N} = 2^{-\ell}$, $\ell = 1, \ldots, 7$. As a first step, for each value of n and of $\bar{N}$, we compute an optimal quantization $(\hat{X}_k)_k$ of the Euler scheme $(\bar{X}_k)_k$ of (4.9) (which is a version of (3.1)), according to the algorithm described in subsection 3.1. Then, for each test function f in (4.8) and each observation path of W, we compute recursively $\langle \hat{V}_k^\delta, f\rangle$ and $\langle \hat{V}_k^\delta, 1\rangle$ using (3.4) and then $\langle \hat{\pi}_k^\delta, f\rangle$. On the other hand, we compute the exact solutions using (4.5), and finally we compute $Err(\delta, \bar{N})$ as defined by (4.7) by summing up over the M trajectories sampled from the observation process W.

Note that since $\Gamma = 0$, the quantization optimization procedure of $(X_k)_k$ is a one-shot process which does not depend on the observations W.

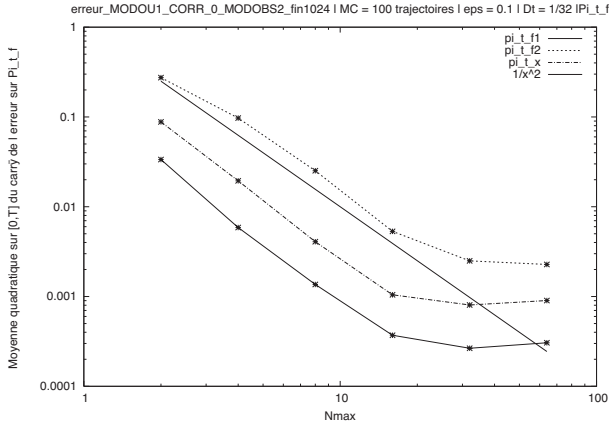
The results are summarized in Figures 1 and 2. It seems to have two regimes of convergence when $\bar{N}$ becomes larger. On the one hand, Figure 1 displays the error (4.7) for low values of n. It seems that its square root behaves like $O(1/\bar{N})$ for the three values of n before a threshold depending (linearly) on n; after that the error remains unchanged.

On the other hand, for high values of n (but still below $n_{ref} = 1024$), Figure 2 suggests a slower rate of convergence in $O((\bar{N})^{-\mathbf{1/2}})$.

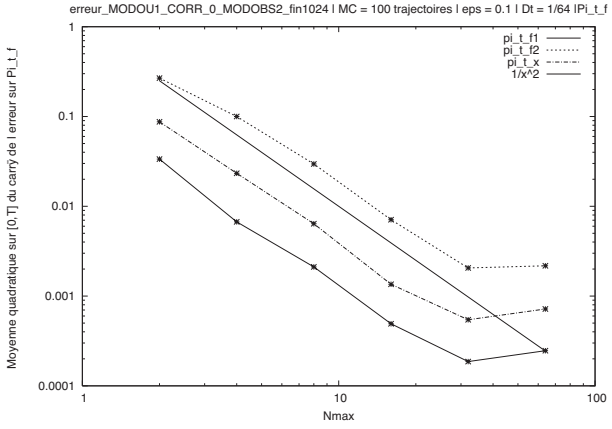
This suggests, keeping in mind (3.20) and Remark 3.1, a decomposition of the



(a) $n = 16$



(b) $n = 32$



(c) $n = 64$

FIG. 1. *Error* $Err(\delta, \bar{N})$ *as a function of* $\bar{N}$ *for several time discretizations* n. *The straight line depicts* $\bar{N} \mapsto 1/\bar{N}^2$, *and the dashed lines denote the errors computed with the different functions* (4.8). *Here* $\varepsilon = 0.1$.

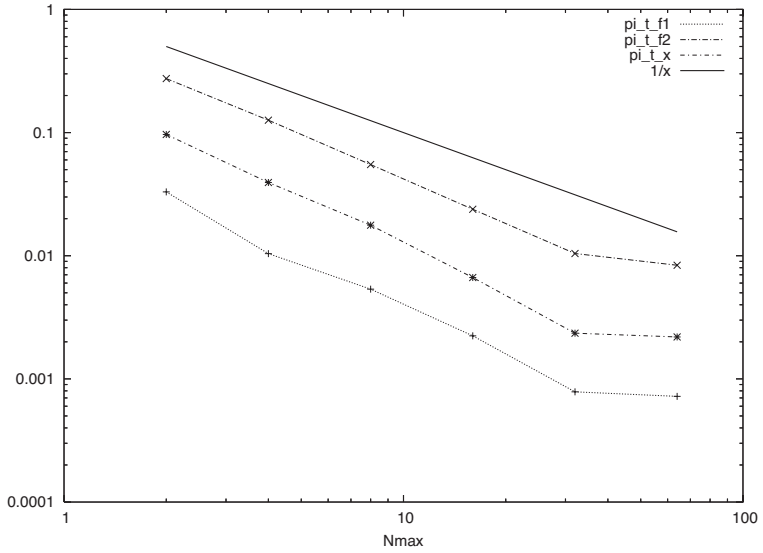


FIG. 2. *Rate of convergence of* (4.7) *with* $n = 256$. *Here again* $\varepsilon = 0.1$.

global error of the form

$$\frac{C_1}{n} + \frac{C_2(n)}{\bar{N}},$$

where $C_1 > 0$ and $C_2(n) = C_2 + c_2 n + o(n)$ with $C_2 > 0$, $c_2 > 0$, and $c_2 \ll C_2$.

For low values of n, C_2 remains constant and hence we get, obviously,

$$\frac{C_1}{n} \leq \frac{C_2}{\bar{N}} \Longleftrightarrow \bar{N} \leq Cn = \bar{N}_1^*(n),$$

and thus we get an order $O(1/\bar{N})$.

For high values of n, the linear part of C_2 becomes larger and hence we get, obviously, in the same manner

$$\frac{C_1}{n} \leq \frac{c_2 n}{\bar{N}} \Longleftrightarrow \bar{N} \leq C' n^2 = \bar{N}_2^*(n),$$

and hence we have the order $O((\bar{N})^{-1/2})$.

In fact, this emphasizes that the scheme needs some stability criterion involving n and $\bar{N}$ in order to converge at the true rate $O(1/\bar{N})$.

The quantization step of the algorithm can also be the cause of this rate. Indeed, during the quantization optimization of the signal X, we need to simulate at each time step an Euler increment of X in (4.9). This simulation is used to compute the weights of the "quantization tree" of X (weight of the Voronoi cells and the transition probabilities) and to process the optimization. Here the Euler increment of X, namely $\Sigma\sqrt{\delta}\,\chi$, where χ denotes a real valued normal random variable, becomes very small as n grows; and so it is when $n = 256$. This implies that the Euler increment will mainly "hit" the closest cell in the upper time layer (not to mention the ability of a random number generator to simulate the tail of distributions). Consequently, the transition

probabilities are not computed accurately enough, given the size of the simulation, and can explain the downgrading of the rate of convergence in time. One can conclude this experiment by saying that there is a CFL involving the mean spatial unit length and the time step parameter and a second CFL involving the time discretization parameter and the size of the simulation (this one has been precisely analyzed in [2]).

These results clarify Remark 3.1 concerning the improvement of Theorem 3.1.

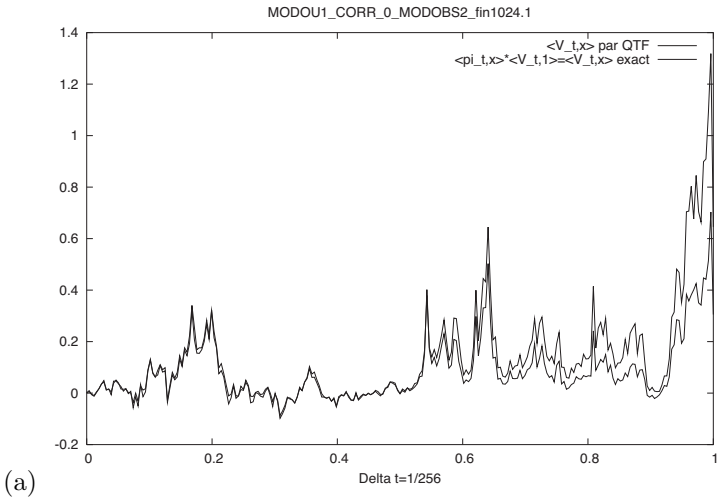
• *Estimation of the time discretization rate of convergence.* Now we look for the rate of convergence with respect to δ. For that purpose, we use $\bar{N} = 100$ quantization points in each time layer. The rate of convergence in time will be estimated with

$$\Gamma \in \{0, 0.5\}, \quad \varepsilon \in \{0.1, 0.5, 1.0\}, \quad \delta = 2^{-m}, \quad m = 1, \dots, 8.$$

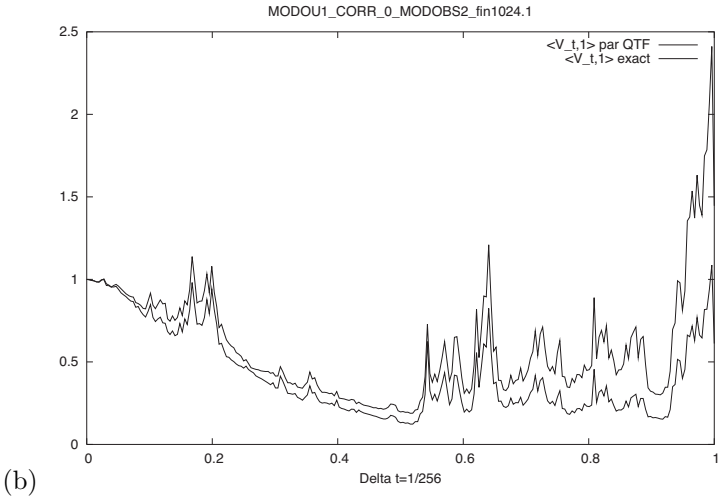
Let us see now why we used the normalized filter instead of the unnormalized one. In Figure 3 are displayed typical examples of graphs $k \mapsto \langle \hat{V}_k^\delta, f \rangle$, $t \mapsto \langle V_t, f \rangle$, $k \mapsto \langle \hat{\pi}_k^\delta, x \rangle$, and $t \mapsto \langle \pi_t, x \rangle$ for $\Gamma = 0$, $\varepsilon = 0.1$, $\delta = 1/256$, and $\bar{N} = N_n = 100$. The exact filters are still computed using (4.5) and (4.6). We verify on that example that the normalized filter seems to be better computed than the unnormalized one. It explains why we did not use the unnormalized version of the error. Indeed, for such a level of noise for the observations ($\varepsilon = 0.1$), the unnormalized filter $\langle \hat{V}_k^\delta, f \rangle$ has very large values. This is true for all tested functions f and all time discretizations $\delta = 1/n$. Furthermore, it is also true on all sampled trajectories of W (not all depicted). Therefore, it is difficult for numerical reasons to compute errors based on $\langle \hat{V}_k^\delta, f \rangle$ for $\varepsilon = 0.1$.

Let us consider first the uncorrelated case ($\Gamma = 0$). Figure 4 shows the error plotted against the time step in a log-log scale for f given by (4.8). We can see again that for a given fixed ε, the time error decreases until a threshold and then remains flat. We also see that this threshold grows as the inverse of the noise level ε. Before reaching this threshold, for every ε and every function f, the rate seems to be of order $\delta = 1/n$ as established in Theorem 2.1.

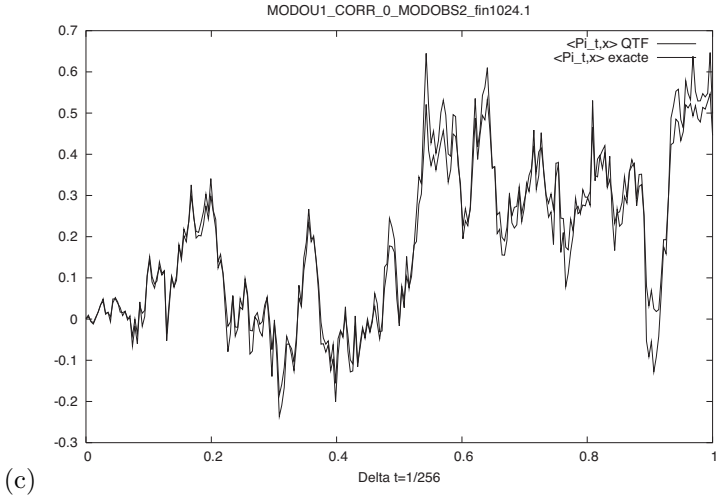
Let us emphasize that, once again in this case, the quantization procedure does not depend on the observations. Therefore, it can be carried out *off-line*. This is no longer true in the correlated case. Then (e.g., if $\Gamma = 0.5$), we will have to compute $M = 100$ quantizations (one per observation path) of the signal $(X_k)_k$ for every $n \in \{2, 4, 8, 16, 32, 64, 128, 256\}$, i.e., 800 optimal grids. The previous study in the uncorrelated case seems to indicate that we need a small level of noise on the observations in order to display a rate with a significant number of time steps. This is why we have chosen $\varepsilon = 0.1$ for the simulations. Figure 5 shows the errors obtained as a function of n in a log-log scale for the functions (4.8). The rates of convergence are the same in each case. A linear regression seems to indicate a rate of $O(n^{-3/4})$ which is better than the $O(n^{-1/2})$ stated in Theorem 2.1. An explanation of this unexpected behavior could be the following one. The constant in the factor of the term $n^{-1/2}$ is presumably very small compared to the one associated to n^{-1}; thus, small values of n make an intermediate rate of convergence appear, while the rate $n^{-1/2}$ would be observed for larger n (in the asymptotic regime).



(a)

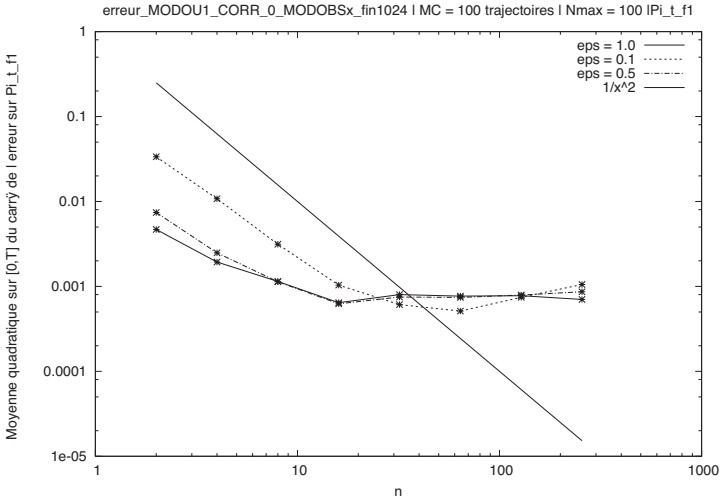


(b)

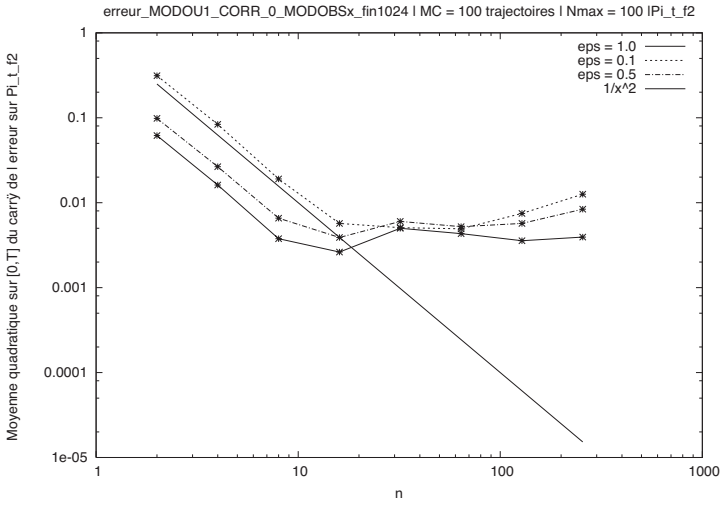


(c)

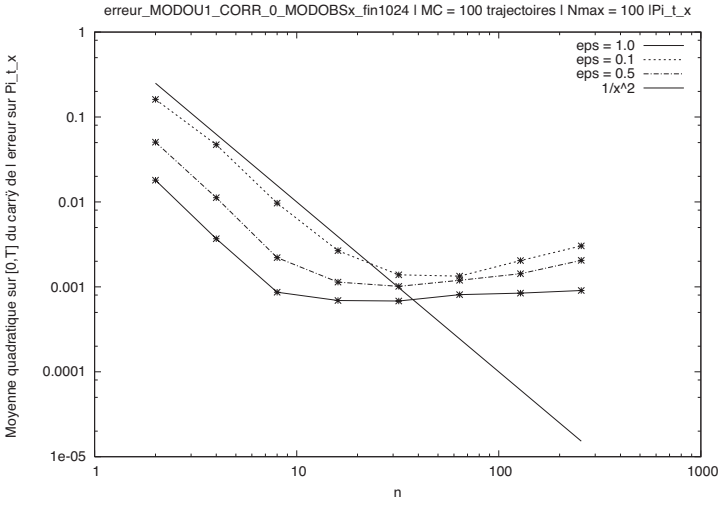
FIG. 3. *Examples of curves* (a) $k \mapsto \langle \hat{V}_k^\delta, x \rangle$, (b) $k \mapsto \langle \hat{V}_k^\delta, 1 \rangle$, (c) $k \mapsto \langle \hat{\pi}_k^\delta, x \rangle$ *with* $\delta = 1/256$ *and* $N_n = 100$ *computed with the same trajectory of observation. Here* $\varepsilon = 0.1$ *and* $\Gamma = 0$. *The thick line depicts the exact filter computed according to a time step* $\delta_{ref} = 1/1024$, *and the thin line depicts the quantized filter.*



(a)



(b)



(c)

FIG. 4. *Square of the error* (4.7) *where* (a) $f(x) = \exp(-x^2)$, (b) $f(x) = \exp(-x)$, *and* (c) $f(x) = x$ *as a function of the time step* n *in a log-log scale. Uncorrelated case.*

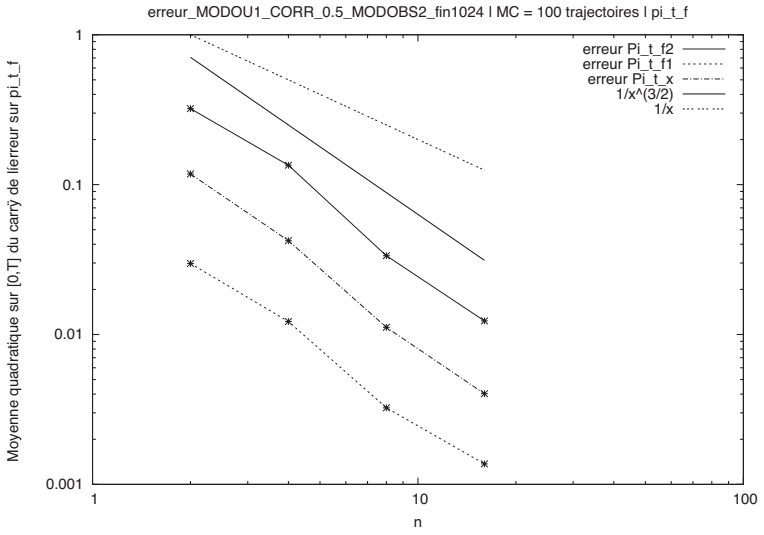


FIG. 5. *Error* (4.7) *as a function of the time step* n *in a log-log scale. Correlated case. The three functions* I_d*,* $f_1(x) = \exp(-x)$*, and* $f_2(x) = \exp(-x^2)$ *are depicted.*

REFERENCES

[1] V. BALLY AND G. PAGÈS, *A quantization algorithm for solving discrete time multi-dimensional optimal stopping problems*, Bernoulli, 9 (2003), pp. 1003–1049.

[2] V. BALLY AND G. PAGÈS, *Error analysis of the optimal quantization algorithm for obstacle problems*, Stoch. Process. Appl., 106 (2003), pp. 1–40.

[3] V. BALLY AND D. TALAY, *The distribution of the Euler scheme for stochastic differential equations:* I. *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[4] A. BENSOUSSAN, R. GLOWINSKI, AND R. RASCANU, *Approximation of Zakai equation by the splitting-up method*, in Stochastic Systems and Optimization, Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York, 1989, pp. 257–265.

[5] A. BUDHIRAJA AND G. KALLIANPUR, *Approximations to the solution of the Zakai equation using multiple Wiener and Stratonovitch integral expansions*, Stoch. Stoch. Rep., 56 (1996), pp. 271–315.

[6] D. CRISAN AND T. LYONS, *A particle approximation of the solution of the Kushner-Stratonovitch equation*, Probab. Theory Related Fields, 115 (1999), pp. 549–578.

[7] D. CRISAN, P. DEL MORAL, AND T. LYONS, *Interacting particle systems approximations of the Kushner-Stratonovitch equation*, Adv. Appl. Probab., 31 (1999), pp. 819–838.

[8] P. DEL MORAL, *Nonlinear filtering using random particles*, Theory Probab. Appl., 40 (1995), pp. 690–701.

[9] L. DEVROYE, *Non-uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

[10] G. DI MASI, M. PRATELLI, AND W. RUNGGALDIER, *An Approximation for the nonlinear filtering problem with error bounds*, Stoch. Stoch. Rep., 14 (1985), pp. 247–271.

[11] P. FLORCHINGER AND F. LE GLAND, *Time-discretization of the Zakai equation for diffusion processes observed in correlated noise*, Stoch. Stoch. Rep., 35 (1991), pp. 233–256.

[12] E. GOBET AND R. MUNOS, *Sensitivity analysis using Itô–Malliavin calculus and martingales, and application to stochastic optimal control*, SIAM J. Control Optim., 43 (2005), pp. 1676–1713.

[13] E. GOBET, G. PAGÈS, H. PHAM, AND J. PRINTEMS, *Discretization and Simulation for a Class of SPDEs with Applications to Zakai and McKean Vlasov Equations*, Preprint LPMA 958, Université Paris 6-Paris 7, Paris, 2005.

[14] S. GRAF AND H. LUSCHGY, *Foundations of Quantization for Random Vectors*, Lecture Notes in Math. 1730, Springer-Verlag, New York, 2000.

[15] I. GYÖNGY, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*, I. Potential Anal., 9 (1998), pp. 1–25.

[16] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*, II. Potential Anal., 11 (1999), pp. 1–37.
[17] I. Gyöngy, *Approximations of stochastic partial differential equations*, in Stochastic Partial Differential Equations, Lecture Notes in Pure and Appl. Math. 227, Dekker, New York, 2002, pp. 287–307.
[18] I. Gyöngy and N. Krylov, *On the splitting-up method and stochastic partial differential equations*, Ann. Probab., 31 (2003), pp. 564–591.
[19] Y. Hu, G. Kallianpur, and J. Xiong, *An approximation for Zakai equation*, Appl. Math. Optim., 45 (2002), pp. 23–44.
[20] H. Korezlioglu and W. J. Runggaldier, *Filtering for non-linear systems driven by non-white noises: An approximation scheme*, Stoch. Stoch. Rep., 44 (1993), pp. 65–102.
[21] T. Kurtz and J. Xiong, *Particle representations for a class of nonlinear SPDEs*, Stochastic Process. Appl., 83 (1999), pp. 103–126.
[22] H. Kushner, *A robust discrete state approximation of the optimal nonlinear filter for a diffusion*, Stoch. Stoch. Rep., 3 (1979), pp. 75–83.
[23] F. Le Gland, *Splitting-up approximation for SPDEs and SDEs with application to nonlinear filtering*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci., 176, Springer-Verlag, New York, 1992, pp. 177–187.
[24] S. Lototsky, R. Mikulevicius, and B. Rozovskii, *Nonlinear filtering revisited: A spectral approach*, SIAM J. Control Optim., 33 (1997), pp. 1716–1730.
[25] H. Luschgy and G. Pagès, *Functional Quantization Rate and Mean Pathwise Regularity of Processes with an Application to Lévy Processes*, Preprint LPMA 1048, Université Paris 6-Paris 7, Paris, 2006.
[26] D. Nualart, *Malliavin Calculus and Related Topics*, Springer-Verlag, New York, 1995.
[27] D. Nualart, *Analysis on Wiener space and anticipating stochastic calculus*, in Lectures on Probability Theory and Statistics, Springer-Verlag, Berlin, 1998, pp. 123–167.
[28] G. Pagès and H. Pham, *Optimal quantization methods for non-linear filtering with discrete-time observations*, Bernoulli, 11 (2005), pp. 893–932.
[29] G. Pagès, H. Pham, and J. Printems, *Optimal quantization methods and applications to numerical problems in finance*, in Handbook of Numerical Methods in Finance, S. Rachev, ed., Birkhäuser Boston, Boston, 2004, pp. 253–297.
[30] G. Pagès, H. Pham, and J. Printems, *An optimal Markovian quantization algorithm for multidimensional stochastic control problems*, Stoch. Dyn., 4 (2004), pp. 501–545.
[31] G. Pagès and J. Printems, *Optimal quadratic quantization for numerics: The Gaussian case*, Monte Carlo Methods Appl., 9 (2003) pp. 135–166.
[32] J. Picard, *Approximations of nonlinear filtering problems and order of convergence*, in Filtering and Control of Random Processes, Lect. Notes in Control and Inform. Sci. 61, Springer-Verlag, New York, 1984, pp. 219–236.
[33] M. Piccioni, *Convergence of implicit discretization schemes for linear differential equations with application to filtering*, in Stochastic Partial Differential Equations and Applications, Lecture Notes in Math. 1236, Springer-Verlag, New York, 1987, pp. 208–229.
[34] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1055–1078.
[35] J. B. Walsh, *Finite element methods for parabolic stochastic PDE's*, Potential Anal., 23 (2005), pp. 1–43.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2539–2558

ON A PARALLEL ROBIN-TYPE NONOVERLAPPING DOMAIN DECOMPOSITION METHOD*

LIZHEN QIN† AND XUEJUN XU†

Abstract. In recent years, a nonoverlapping Robin-type domain decomposition method (DDM) for the finite element discretization systems of the second order elliptic equations, which is based on using Robin-type boundary conditions as information transmission conditions on the subdomain interfaces, has been developed and analyzed since it was first proposed by P. L. Lions in [*On the Schwarz alternating method* III: *A variant for nonoverlapping subdomains*, in Proceedings of the 3rd International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, PA, 1990, pp. 202–223]. However, the convergence rate of this DDM with many subdomains remains open when the lower term of equations vanishes. This open problem will be considered in this paper. The convergence rate is almost $1 - O(h^{1/2}H^{-1/2})$ in certain cases—for example, the case of a small number of subdomains, where h is the mesh size and H is the size of subdomain. In order to get the desirous convergence results, two mathematics skills are introduced in this paper; one is complexification of real linear space and the other is the spectral radius formula.

Key words. nonconforming finite elements, nonoverlapping domain decomposition, convergence rate, geometric convergence

AMS subject classifications. 65F10, 65N30

DOI. 10.1137/05063790X

1. Introduction.

The Robin-type nonoverlapping domain decomposition method (DDM), which is based on using Robin-type boundary conditions as information transmission conditions on the subdomain interfaces, has become an increasingly important tool for solving the following second order elliptic equations:

$$\begin{cases} -\sum\limits_{i,j} \frac{\partial}{\partial x_i}(a_{ij}(x)\frac{\partial u}{\partial x_j}) + b(x)u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{cases}$$

The idea of employing Robin-type boundary conditions as interface conditions was first proposed by P. L. Lions in [21]. Recently there have been several theoretical analyses and applications of this DDM; cf. [1], [2], [8], [9], [13], [14], [15], [19], [23], [10], [16], [17], [18], [20], and [11] for details. In [17], Gander, Halpern, and Nataf considered the second elliptic problems in the case of the two subdomains. It is first pointed out by them that the optimal choice of relaxation parameter was $O(h^{-1/2})$ and the convergence rate $1-O(h^{1/2})$ can be achieved in this special case. In this paper, we will discuss the second order elliptic problems in the case of many subdomains.

Compared with other DDMs, this method has several advantages. First, the iterative procedure is very simple. Second, in contrast to other procedures, it need not to solve global problems. So the iterative procedure is much more highly parallel than others. Last, by the results in [22] and in this paper, we know that the convergence

*Received by the editors August 10, 2005; accepted for publication (in revised form) June 17, 2006; published electronically December 5, 2006. This work was supported by the special funds for major state basic research projects (973) under 2005CB321701 and the National Science Foundation (NSF) of China (10471144).

http://www.siam.org/journals/sinum/44-6/63790.html

†LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, People's Republic of China (qinlz@lsec.cc.ac.cn, xxj@lsec.cc.ac.cn).

rate of this method is $1 - O(h^{1/2}H^{-1/2})$ in certain cases, where h is the size of the mesh and H is the size of the subdomain, which is much better than that of the Schwarz DDMs with small overlapping. Based on the analysis in [12] and [4], we know that the realistic convergence rate is only $1 - O(hH^{-1})$ for the additive and multiplicative Schwarz DDMs.

In the case of $b(x) \geq b > 0$, the lower term is positive strictly. The analysis of this method has been highly developed. In [6], [7], Deng proved that this method was convergent. The result in [7] claims that the convergence of this method is geometric and the convergence rate is $1 - O(h)$. This result had been improved in [22], which tells us this convergence rate could be $1-O(h^{1/2}H^{-1/2})$. Meanwhile counterexamples constructed in [22] show that the convergence rate cannot be improved in general.

However, the analysis for convergence is much harder when the lower term vanishes, i.e., $b(x) = 0$. In [6], [7], Deng proved that this method was also convergent in this case, but he was not able to tell us the convergence rate. In [22], we showed that the convergence rate was $1 - O(h^{1/2}H^{-1/2})$ if at least one face of each subdomain belongs to the boundary of domain. But for the general domain decomposition with interior subdomains, the analysis of the convergence rate remains open.

In this paper, we get the convergence rate in general in the case when the lower term vanishes. We will show that the convergence rate can be $1 - O(C^N h^{1/2}H^{-1/2})$, where $C \in (0, 1)$ and N is a geometric parameter which will be introduced in section 4. The convergence rate will be $1-O(h^{1/2}H^{-1/2})$ when N is not large. The numerical experiments in this paper support our theoretical results. Our proof is based on the energy method. The energy method was first proposed by P. L. Lions in [21] and has been developed in [6] and [7]. We will introduce two other techniques for obtaining the convergence rate in this paper; one is the complexification of real linear space and the other is the spectral radius formula.

The outline of this paper is as follows. In section 2, we will introduce the model problem and some basic results of this method. The analysis of this paper is based on two skills; one is the complexification of real linear space and the other is the spectral radius formula. We will introduce them in section 3. In section 4, we will introduce some geometric aspects of domain decomposition. In section 5, we prove our main result, Theorem 5.1. Finally, in section 6, we will give some numerical results to support our theory.

2. Model problem and preliminaries. We consider the following model problem:

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & \text{in } \Omega, \\ u & = & 0 & \text{on } \partial\Omega, \end{array} \right. \tag{2.1}$$

where Ω is a bounded polyhedral domain in $R^d (d = 2, 3)$, $f \in L^2(\Omega)$.

Partition Ω into nonoverlapping subdomains $\Omega_i (i = 1, \ldots, K)$ quasi-uniformly and regularly. Then the domain decomposition iterative procedure can be written as follows (cf. [7] for details):

$$\left\{ \begin{array}{rcll} -\Delta u_i^n & = & f & \text{in } \Omega_i, \\ \frac{\partial u_i^n}{\partial \nu_i} + \lambda_{ij} u_i^n & = & g_{ij}^n & \text{on } \gamma_{ij}, \quad j \in N(i), \\ u_i^n & = & 0 & \text{on } \Gamma_i, \\ g_{ij}^{n+1} & = & 2\lambda_{ij} u_j^n - g_{ji}^n, & \end{array} \right. \tag{2.2}$$

where $\Gamma_i = \partial\Omega_i \bigcap \partial\Omega$, $\gamma_{ij} = \partial\Omega_i \bigcap \partial\Omega_j$, $\lambda_{ij} = \lambda_{ji} > 0$ are parameters,

$$N(i) = \{j \neq i \mid |\gamma_{ij}| > 0\}, \tag{2.3}$$

and $|\gamma_{ij}|$ is the measure of γ_{ij}.

This algorithm was first proposed by P. L. Lions in [21] and then improved by Deng in [6]. It can be accelerated by using Krylov methods via a substructuring of this algorithm.

Let $\mathcal{T}_h$ be a quasi-uniform and regular finite element triangulation of Ω. The mesh size is h. Let X be a nonconforming Crouzeix–Raviart finite element space over $\mathcal{T}_h$ (cf. [5]). The function of X vanishes at every freedom on $\partial\Omega$.

Denote N_h as the set of all interior nodal points, i.e., N_h contains all midpoints of interior edges ($d = 2$) or barycenters of interior faces ($d = 3$) of elements.

Let $X_i = X|_{\Omega_i}$. Define $X_i^0 \subseteq X_i$ whose functions vanish at every freedom on $\partial\Omega_i$.

Next we define two spaces Y_i and Y_{ij} on $\partial\Omega_i$ and γ_{ij}, respectively. Define Y_i to be a piecewise constant space on triangulation $\mathcal{T}_{\partial\Omega_i}$, where $\mathcal{T}_{\partial\Omega_i}$ is the triangulation of $\partial\Omega_i \setminus \partial\Omega$ inherited from $\mathcal{T}_h$, i.e., $\mathcal{T}_{\partial\Omega_i} = \mathcal{T}_h|_{\partial\Omega_i \setminus \partial\Omega}$. Furthermore let $Y_{ij} = Y_i|_{\gamma_{ij}}$.

All above spaces are equipped with the L^2 norm denoted by $\|\cdot\|_0$. Denote the L^2 inner product by $\langle\cdot,\cdot\rangle$. X can also be equipped with norms $|\cdot|_{1,h}$ and $\|\cdot\|_{1,h}$, where

$$|v|_{1,h}^2 = \sum_{T\in\mathcal{T}_h} \int_T |\nabla v|^2,$$

$$\|v\|_{1,h,\Omega_i}^2 = |v|_{1,h,\Omega_i}^2 + H^{-2}\|v\|_{0,\Omega_i}^2,$$

$$\|v\|_{1,h}^2 = \sum_{i=1}^{K} \|v\|_{1,h,\Omega_i}^2.$$

Define π_i and π_{ij} to be linear operators from X_i to Y_i and Y_{ij}, respectively. $\forall v_i \in X_i$, $\pi_i v_i \in Y_i$ and

$$\pi_i v_i|_\tau \equiv v_i(p) \quad \forall \tau \in \mathcal{T}_{\partial\Omega_i},$$

where p is the midpoint of $\tau(d = 2)$ or the barycenter of $\tau(d = 3)$, and $\pi_{ij} v_i \in Y_{ij}$ is defined to be $\pi_i v_i|_{\gamma_{ij}}$.

Conversely, we define S_i and S_{ij} to be linear operators from Y_i and Y_{ij} to X_i, respectively. $\forall w_i \in Y_i$, let $S_i w_i \in X_i$ and

$$S_i w_i = \begin{cases} w_i & \text{freedom on } \partial\Omega_i, \\ 0 & \text{other freedom.} \end{cases}$$

$\forall w_{ij} \in Y_{ij}$, we also define $S_{ij} w_{ij} \in X_i$ and

$$S_{ij} w_{ij} = \begin{cases} w_{ij} & \text{freedom on } \gamma_{ij}, \\ 0 & \text{other freedom.} \end{cases}$$

Note that $\pi_i v_i \neq v_i|_{\partial\Omega_i}$, $S_i w_i|_{\partial\Omega_i} \neq w_i$ in general. However,

$$v_i - S_i \pi_i v_i \in X_i^0, \tag{2.4}$$

$$\pi_i S_i = Id_i, \qquad \pi_{ij} S_{ij} = Id_{ij}, \tag{2.5}$$

where Id_i and Id_{ij} are identity operators on Y_i and Y_{ij}, respectively. By (2.5), we know that both π_i and π_{ij} are surjective. Furthermore, we have the following lemma.

LEMMA 2.1.

$$\|\pi_{ij}v_i\|_{0,\gamma_{ij}} \le C\|v_i|_{\gamma_{ij}}\|_{0,\gamma_{ij}} \quad \forall v_i \in X_i,$$

$$\|S_{ij}w_{ij}\|_{0,\Omega_i} \le Ch^{1/2}\|w_{ij}\|_0,$$

$$|S_{ij}w_{ij}|_{1,h,\Omega_i} \le Ch^{-1/2}\|w_{ij}\|_0 \quad \forall w_{ij} \in Y_{ij},$$

where C is a constant independent of h.

Proof. The first inequality can be verified by direct computation.

By scaling arguments, we have

$$\begin{aligned}\|S_{ij}w_{ij}\|_{0,\Omega_i} &\le Ch^{d/2}\Big(\sum_{p\in\Omega_i\bigcap N_h}(S_{ij}w_{ij})(p)^2\Big)^{1/2}\\ &= Ch^{d/2}\Big(\sum_{p\in\gamma_{ij}\bigcap N_h} w_{ij}(p)^2\Big)^{1/2}\\ &\le Ch^{1/2}\|w_{ij}\|_{0,\gamma_{ij}},\end{aligned}$$

where $d = 2, 3$ is the dimension of domain Ω. Using the inverse inequality,

$$|S_{ij}w_{ij}|_{1,h,\Omega_i} \le Ch^{-1}\|S_{ij}w_{ij}\|_{0,\Omega_i},$$

we get the last two inequalities. ∎

The discrete finite element approximation of (2.1) is to find $\hat{u} \in X$ such that

$$a(\hat{u}, v) = f(v) \quad \forall v \in X, \tag{2.6}$$

where

$$\begin{aligned}a(\hat{u}, v) &= \sum_{T\in\mathcal{T}_h}\int_T \nabla u\cdot\nabla v,\\ f(v) &= \int_\Omega fv.\end{aligned}$$

Let

$$a_i(\cdot,\cdot) = a(\cdot,\cdot)|_{X_i}, \qquad f_i(\cdot) = f(\cdot)|_{X_i}.$$

The discrete finite element version of (2.2) can be written as follows:

$$\begin{cases} a_i(u_i^n, v) + \sum\limits_{j\in N(i)}\lambda_{ij}\int_{\gamma_{ij}}\pi_{ij}u_i^n\cdot\pi_{ij}v &=& \int_{\Omega_i} fv + \sum\limits_{j\in N(i)}\int_{\gamma_{ij}} g_{ij}^n\cdot\pi_{ij}v,\\ g_{ij}^{n+1} &=& 2\lambda_{ij}\pi_{ji}u_j^n - g_{ji}^n. \end{cases} \tag{2.7}$$

Remark 2.1. $\forall u, v \in X_i$,

$$\int_{\gamma_{ij}}\pi_{ij}u\cdot\pi_{ij}v = \sum_{p\in\gamma_{ij}\bigcap N_h} u(p)v(p)|s_p|,$$

where s_p is the element face with p as its barycenter and $|s_p|$ is the measure of s_p. In [7], $\int_{\gamma_{ij}}\pi_{ij}u\cdot\pi_{ij}v$ is replaced by $\sum_{p\in\gamma_{ij}\bigcap N_h} u(p)v(p)|s_p|$ in (2.7). In fact, these two iterative procedures are equivalent.

DEFINITION 2.2. *Let $\hat{u}$ be the finite element solution and u^n be the solution of the nth iterative step, respectively. If*

$$\|u^n - \hat{u}\| \le CL^n \|u^0 - \hat{u}\|, \tag{2.8}$$

where $\|\cdot\|$ is the certain norm, $L \in [0,1)$, and C is independent of n, then this is geometric convergence. L is the convergence rate.

It is well known that $\hat{u}$ can be approximated by the procedure (2.7). The following theorem will tell us that $\hat{u}$ is the fixed point of (2.7). Its proof can be found in [22].

THEOREM 2.3. *If $\hat{u}$ is the solution of* (2.6) *and $\lambda_{ij} = \lambda_{ji} > 0$, then there exists $\hat{g}_{ij} \in Y_{ij}$ such that*

$$a_i(\hat{u}, v) + \sum_{j \in N(i)} \lambda_{ij} \int_{\gamma_{ij}} \pi_{ij}\hat{u}_i \cdot \pi_{ij} v_i = f_i(v) + \sum_{j \in N(i)} \int_{\gamma_{ij}} \hat{g}_{ij} \cdot \pi_{ij} v_i, \tag{2.9}$$

where $\hat{u}_i = \hat{u}|_{\Omega_i}$, $v = (v_1, \ldots, v_k) \in \prod_{i=1}^K X_i$,

$$\hat{g}_{ij} = 2\lambda_{ij}\pi_{ji}\hat{u}_j - \hat{g}_{ji}. \tag{2.10}$$

Define W to be a subspace of $\prod_{i=1}^K X_i \times \prod_{i=1}^K Y_i$ such that $(e_1, \ldots, e_K, \varepsilon_1, \ldots, \varepsilon_k) \in W$ if and only if $\forall v_i \in X_i$

$$a_i(e_i, v_i) + \sum_{j \in N(i)} \lambda_{ij} \int_{\gamma_{ij}} \pi_{ij} e_i \cdot \pi_{ij} v_i = \sum_{j \in N(i)} \int_{\gamma_{ij}} \varepsilon_{ij} \cdot \pi_{ij} v_i, \tag{2.11}$$

where $\varepsilon_{ij} = \varepsilon_i|_{\gamma_{ij}} \in Y_{ij}$.

Define A to be a linear operator of W. $\forall (e_1, \ldots, e_K, \varepsilon_1, \ldots, \varepsilon_k)$,

$$A(e_1, \ldots, e_K, \varepsilon_1, \ldots, \varepsilon_k) = (\tilde{e}_1, \ldots, \tilde{e}_K, \tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_k), \tag{2.12}$$

where

$$\tilde{\varepsilon}_{ij} = 2\lambda_{ij}\pi_{ji} e_j - \varepsilon_{ji}.$$

$\tilde{e}_i$ is determined by $(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_k)$ and (2.11).

Define

$$e_i^n = u_i^n - \hat{u}_i, \varepsilon_i^n = g_i^n - \hat{g}_i,$$

and

$$\varepsilon_{ij}^n = \varepsilon_i^n|_{\gamma_{ij}} = g_{ij}^n - \hat{g}_{ij}.$$

By Theorem 2.3 and (2.12), we get a corollary immediately.

COROLLARY 2.4. *$(e^n, \varepsilon^n) \in W$; i.e., (e^n, ε^n) satisfies* (2.11). *Moreover,*

$$(e^{n+1}, \varepsilon^{n+1}) = A(e^n, \varepsilon^n). \tag{2.13}$$

Proof. Subtracting (2.7) by (2.9), we get

$$a_i(e_i^n, v_i) + \sum_{j \in N(i)} \lambda_{ij} \int_{\gamma_{ij}} \pi_{ij} e_i^n \cdot \pi_{ij} v_i = \sum_{j \in N(i)} \int_{\gamma_{ij}} \varepsilon_{ij}^n \cdot \pi_{ij} v_i,$$

so $(e^n, \varepsilon^n) \in W$.

By Theorem 2.3, we have

$$\varepsilon_{ij}^{n+1} = g_{ij}^{n+1} - \hat{g}_{ij} = 2\lambda_{ij}\pi_{ji}e_j^n - \varepsilon_{ji}^n.$$

Thus we deduce (2.13) from the definition of A. ∎

We need the trace theorem and the Poincaré inequality for nonconforming finite element space in order to get the convergence rate. First, we introduce the Sarkis' isomorphism between the Crouzeix–Raviart nonconforming element space and the Courant conforming element space (see [24, Isomorphism 1, Lemmas 3, 4, and 5, and Isomorphism 2]).

LEMMA 2.5 (Sarkis' isomorphism). *Suppose γ_{ij} is a face (when $d = 2$, a face is an edge) of subdomain Ω_i; then $\forall u \in X_i$, there is a piecewise linear function $\tilde{u} \in H^1(\Omega_i)$ such that*

$$c|u|_{1,h,\Omega_i} \le |\tilde{u}|_{1,\Omega_i} \le C|u|_{1,h,\Omega_i},$$

$$c\|u\|_{1,h,\Omega_i} \le \|\tilde{u}\|_{1,\Omega_i} \le C\|u\|_{1,h,\Omega_i},$$

$$\int_{\gamma_{ij}} \tilde{u} = \int_{\gamma_{ij}} u,$$

and if γ_{ik} is an arbitrary face of Ω_i, then

$$c\|u|_{\gamma_{ik}}\| \le \|\tilde{u}|_{\gamma_{ik}}\| \le C\|u|_{\gamma_{ik}}\|,$$

where c and C are constants independent of h and the diameter of Ω_i.

Proof. Proofs of the first two inequalities and equality can be found in [24, Lemmas 3, 4, and 5]. The last inequality can be verified by direct computation thanks to the construction of $\tilde{u}$ and the equivalence between the L^2 norm and the L^2 discrete norm. ∎

The following theorem is a Poincaré–Friedrichs inequality (cf. [3, (1.1)] and [24, Lemma 4]. It can be proved by using equivalence norm of quotient spaces (see [5, the proof of Theorem 3.1.1] and the scaling argument.

THEOREM 2.6 (Poincaré–Friedrichs inequality). *If the diameter of each subdomain Ω_i $(i = 1, \ldots, K)$ is $O(H)$, and γ_{ij} is a face of Ω_i, then $\forall \tilde{u} \in H^1(\Omega_i)$, we have*

$$\|\tilde{u}\|_{0,\Omega_i}^2 \le CH^2|\tilde{u}|_{1,\Omega_i}^2 + CH^{2-d}\left(\int_{\gamma_{ij}} \tilde{u}\right)^2,$$

where $d = 2, 3$ is the dimension of Ω_i and C is a constant independent of Ω_i.

Furthermore, we can get a special trace theorem of Crouzeix–Raviart element space.

LEMMA 2.7 (special trace theorem). *If the diameter of each subdomain Ω_i $(i = 1, \ldots, K)$ is $O(H)$, γ_{ij}, and γ_{ik} are two faces of Ω_i, then $\forall v_i \in X_i$ we have*

$$\|\pi_{ik}v_i\|_{0,\gamma_{ik}}^2 \le CH|v_i|_{1,h,\Omega_i}^2 + C\|\pi_{ij}v_i\|_{0,\gamma_{ij}}^2,$$

where C is a constant independent of Ω_i.

Proof. By Lemma 2.5, there is a $\tilde{v}_i \in H^1(\Omega_i)$ such that

$$c|v_i|_{1,h,\Omega_i} \le |\tilde{v}_i|_{1,\Omega_i} \le C|v_i|_{1,h,\Omega_i},$$

$$\int_{\gamma_{ij}} \tilde{v}_i = \int_{\gamma_{ij}} v_i,$$

and

$$c\|v_i|_{\gamma_{ik}}\| \le \|\tilde{v}_i|_{\gamma_{ik}}\| \le C\|v_i|_{\gamma_{ik}}\|.$$

As a result, we have

$$\begin{aligned}
\|\pi_{ik}v_i\|^2_{0,\gamma_{ik}} &\le C\|v_i|_{\gamma_{ik}}\|^2_{0,\gamma_{ik}} \\
&\le C\|\tilde{v}_i|_{\gamma_{ik}}\|^2_{0,\gamma_{ik}} \\
&\le CH|\tilde{v}_i|^2_{1,\Omega_i} + CH^{-1}\|\tilde{v}_i\|^2_{0,\Omega_i} \\
&\le CH|\tilde{v}_i|^2_{1,\Omega_i} + CH^{1-d}\left(\int_{\gamma_{ij}} \tilde{v}_i\right)^2 \\
&\le CH|v_i|^2_{1,h,\Omega_i} + CH^{1-d}\left(\int_{\gamma_{ij}} v_i\right)^2,
\end{aligned}$$

where we have used Lemma 2.1, the trace theorem for $H^1(\Omega_i)$, and Theorem 2.6 in the first, third, and fourth inequalities, respectively. By the definition of π_{ij}, we know that

$$\int_{\gamma_{ij}} v_i = \int_{\gamma_{ij}} \pi_{ij}v_i;$$

then

$$\begin{aligned}
\|\pi_{ik}v_i\|^2_{0,\gamma_{ik}} &\le CH|v_i|^2_{1,h,\Omega_i} + CH^{1-d}\left(\int_{\gamma_{ij}} \pi_{ij}v_i\right)^2 \\
&\le CH|v_i|^2_{1,h,\Omega_i} + C\|\pi_{ij}v_i\|^2_{0,\gamma_{ij}}. \qquad \square
\end{aligned}$$

3. Some results in linear space. In this section we will introduce some notions and skills in linear space, such as complexification and the spectral radius formula which will play important roles in the following analysis.

We will consider $\|A^n\|$ in order to get the convergence rate, where A is the operator of linear space W defined in (2.12). If V is a complex linear space and T is a complex linear operator of V, then we have the following well-known spectral radius formula:

$$\lim_{n\to\infty} \|T^n\|^{1/n} = \rho(T),$$

where $\rho(T)$ is the spectral radius of T. So we can estimate $\|T^n\|$ in terms of $\rho(T)$. Unfortunately, W is a real linear space and A is a real linear operator while the spectral radius formula does not hold in the real case in general. So the complexification of a real linear space and a real linear operator is necessary.

Constructing the complexification of a real linear space is just like constructing a complex number field by a real number field.

DEFINITION 3.1. *Suppose V is a real n dimensional linear space; we call the tensor product space $C \otimes V$ the complexification of V, where C is the complex number field or one dimensional complex linear space. In other words, $C \otimes V$ is a complex n dimensional space such that*

$$C \otimes V = \{x + \sqrt{-1}y | x, y \in V\}.$$

Then $C \otimes V$ has the following addition and scalar multiplication properties:

$$(x_1 + \sqrt{-1}y_1) + (x_2 + \sqrt{-1}y_2) = (x_1 + x_2) + \sqrt{-1}(y_1 + y_2),$$

$$(a + \sqrt{-1}b)(x + \sqrt{-1}y) = (ax - by) + \sqrt{-1}(bx + ay), \quad a + \sqrt{-1}b \in C.$$

LEMMA 3.2. *Suppose V is a real linear space equipped with inner product $\langle \cdot, \cdot \rangle$; then we can define an inner product of $C \otimes V$ such that*

$$\langle x_1 + \sqrt{-1}y_1, x_2 + \sqrt{-1}y_2 \rangle = \langle x_1, x_2 \rangle + \langle y_1, y_2 \rangle - \sqrt{-1}\langle x_1, y_2 \rangle + \sqrt{-1}\langle y_1, x_2 \rangle.$$

If $\| \cdot \|$ is the norm induced by the inner product, then

$$\|x + \sqrt{-1}y\|^2 = \|x\|^2 + \|y\|^2.$$

The proof of this lemma will be presented in the appendix.

Remark 3.1. As for the semi-innerproduct, i.e., $\langle x, x \rangle \geq 0$ and may be equal to 0 even when $x \neq 0$, the counterpart conclusion of Lemma 3.2 also holds. In this case the semi-innerproduct will induce a seminorm. The proof of this conclusion is similar to that of Lemma 3.2.

Remark 3.2. From now on, all norms and seminorms in this paper can be induced by innerproducts and semi-innerproducts.

DEFINITION 3.3. *If V is a real linear space and T is a real linear operator of V, we define a complex linear operator $1 \otimes T$ of $C \otimes V$ such that*

$$1 \otimes T(x + \sqrt{-1}y) = Tx + \sqrt{-1}Ty.$$

We call $1 \otimes T$ the complexification of T. For convenience, we also denote $1 \otimes T$ by $\bar{T}$.

LEMMA 3.4. *If V is a real linear space and T_1, T_2 are real linear operators of V, then*

$$(1 \otimes T_1)(1 \otimes T_2) = 1 \otimes (T_1 T_2).$$

In particular,

$$1 \otimes (T^n) = (1 \otimes T)^n;$$

we denote $1 \otimes (T^n)$ or $(1 \otimes T)^n$ by $\bar{T}^n$.

We will prove this lemma in the appendix.

LEMMA 3.5. *If V is a finite dimensional real linear space equipped with an innerproduct, and T is a real linear operator of V, then*

$$\|\bar{T}\| = \|T\|.$$

This lemma will also be proved in the appendix.

Now let us discuss the linear space W and the linear operator A defined in (2.12) again. Our main skill to deal with $\|A^n\|$ is that $\|A^n\|$ is dominated by $\rho(\bar{A})$, where $\rho(\bar{A})$ is the spectral radius of $\bar{A}$, the complexification of A. In other words, we use the following lemma.

LEMMA 3.6. *If W is equipped with an innerproduct and*

$$\rho(\bar{A}) \leq 1 - R, \quad R \in (0, 1),$$

then for all positive integer number n, there is a constant C independent of n such that

$$\|A^n\| \leq C(1 - R/2)^n.$$

Proof. By Lemmas 3.4 and 3.5, we know that

$$\|A^n\| = \|\bar{A}^n\|.$$

Since $\bar{A}$ is a complex linear operator of the complex linear space $C \otimes W$, then by the spectral radius formula,

$$\lim_{n\to\infty} \|\bar{A}^n\|^{1/n} = \rho(\bar{A}).$$

So $\forall \varepsilon > 0$, there is an integer N such that when $n > N$, we have

$$\|\bar{A}^n\|^{1/n} \leq \rho(\bar{A}) + \varepsilon,$$

or

$$\|\bar{A}^n\| \leq (\rho(\bar{A}) + \varepsilon)^n.$$

Choose a constant $C > 1$ such that

$$\|\bar{A}^n\| \leq C(\rho(\bar{A}) + \varepsilon)^n$$

for $n = 1, \ldots, N$. Then $\forall\ n$,

$$\|A^n\| = \|\bar{A}^n\| \leq C(\rho(\bar{A}) + \varepsilon)^n.$$

Letting $\varepsilon = R/2$, we get the conclusion, where C is independent of n, although it may depend on R. ∎

Besides the complexification of W and A, we will also use those of other real linear spaces such as X_i, and Y_i, and other real linear operators such as π_{ij}.

4. Some geometric aspects of domain decomposition. Our analysis depends on the geometric aspects of subdomain decomposition. Now we introduce some notions and an assumption.

We define a sequence of sets D_i whose elements are subdomains by induction:

$$D_1 = \{\Omega_i | \text{at least one face of } \Omega_i \text{ belongs to } \partial\Omega\},$$

$$D_{r+1} = \{\Omega_i | \Omega_i \notin D_r, \Omega_i \text{ share one face with some } \Omega_j \in D_r \text{ at least}\}.$$

Now we define a geometric parameter of subdomain decomposition.

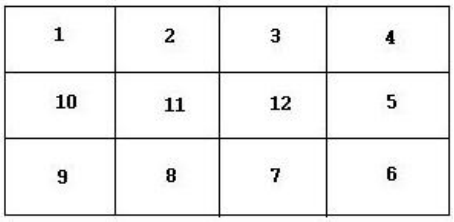


FIG. 1.

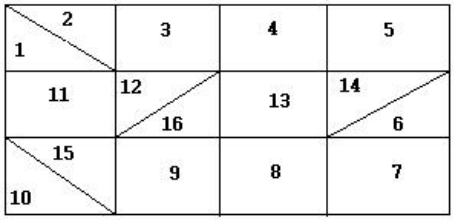


FIG. 2.

DEFINITION 4.1. *There is an integer N such that $\bigcup_{i=1}^{N} D_i$ contains all subdomains of Ω and we call N the winding number of domain decomposition.*

For example (see Figure 1), the integer i in each subdomain means that this subdomain is Ω_i. So

$$D_1 = \{\Omega_i | i = 1, \ldots, 10\},$$

$$D_2 = \{\Omega_{11}, \Omega_{12}\},$$

and winding number $N = 2$.

See also Figure 2:

$$D_1 = \{\Omega_i | i = 1, \ldots, 11\},$$

$$D_2 = \{\Omega_i | i = 12, \ldots, 16\},$$

and winding number N is also 2.

For convenience, we denote a subdomain belongs to D_r by Ω_{i^r}.

DEFINITION 4.2. *$\forall \Omega_{i^r} \in D_r$, we call a set*

$$P = \{\Omega_{i^1}, \ldots, \Omega_{i^r}\} \subseteq \bigcup_{i=1}^{r} D_r$$

a path connecting Ω_{i^r} with the boundary provided that $\Omega_{i^r} \in P$ and $P \cap D_k$ $(k = 1, \ldots, r)$ has exact one element.

Of course, the path connecting Ω_{i^r} with the boundary may not be unique. For example (see Figure 1), Ω_{11} has three paths, $\{\Omega_2, \Omega_{11}\}$, $\{\Omega_8, \Omega_{11}\}$, and $\{\Omega_{10}, \Omega_{11}\}$.

Now we make an assumption on the domain decomposition.

ASSUMPTION 4.1. *The domain decomposition satisfies the following two conditions.*

(1) *For each Ω_i, there is a path connecting Ω_i with the boundary. Then we assign this path to Ω_i.*

(2) *If* $\Omega_i, \Omega_j \in D_r$ *and* $\Omega_i \neq \Omega_j$, *then the path assigned to* Ω_i *has no intersection with that assigned to* Ω_j.

See Figure 2. We assign $\{\Omega_{11}, \Omega_{12}\}$, $\{\Omega_4, \Omega_{13}\}$, $\{\Omega_6, \Omega_{14}\}$, $\{\Omega_{10}, \Omega_{15}\}$, and $\{\Omega_9, \Omega_{16}\}$, to $\Omega_{11}, \ldots, \Omega_{15}$, respectively. Each pair has no intersection.

Remark 4.1. It is necessary to point out that Assumption 4.1 is not strict at all. It is almost not an assumption because most domain decompositions are satisfied by it.

Remark 4.2. The relations among N, H, and number of subdomains are delicate. N must be small provided that H is not small or the number of subdomains is small. However, N will not be large necessarily when H is small or the number of subdomains is large. A case in point is that the domain Ω is a strip. For example, decompose a narrow rectangle domain into 4×25 small rectangles uniformly. Then the number of subdomains is 100 while N is 2.

5. Convergence rate. Now we state and prove the main theorem of this paper. We define

$$\|(e^n, \varepsilon^n)\|^2 = \sum_{i=1}^{K} \|e_i^n\|_{1,h,\Omega_i}^2 + \sum_{i=1}^{K} \|\varepsilon_i^n\|_{0,\partial\Omega_i}^2. \tag{5.1}$$

THEOREM 5.1. *If domain decomposition satisfies Assumption* 4.1, *choose* $\lambda_{ij} = \lambda = O(h^{-1/2}H^{-1/2})$ $\forall i,\ j$; *then*

$$\|(e^n, \varepsilon^n)\| \leq C_2(1 - C_1(C_0)^N h^{1/2} H^{-1/2})^n \|(e^0, \varepsilon^0)\|, \tag{5.2}$$

where N *is the winding number of the domain decomposition,* $C_0 \in (0,1)$ *and* $C_1 > 0$ *are constants independent of* h, H, N, *and* n, *and* C_2 *is a positive constant independent of* n.

Remark 5.1. Counterexamples have been constructed in [22] to show that the convergence rate of this method can never be better than $1 - O(h^{1/2}H^{-1/2})$. The convergence rate, $1 - O(h^{1/2}H^{-1/2})$, can only be attained if we choose $\lambda_{ij} = \lambda = O(h^{-1/2}H^{-1/2})$ $\forall i,\ j$ when at least one face of each subdomain belongs to $\partial\Omega$. (See [22, sections 4 and 5] for details.) So we choose $\lambda = O(h^{-1/2}H^{-1/2})$ here.

Remark 5.2. If the winding number N is small, it can be seen from (5.2) that the convergence rate is $1 - O(h^{1/2}H^{-1/2})$. But when the winding number N goes up, the convergence rate of the DDM will deteriorate.

From now on, let $\lambda = O(h^{-1/2}H^{-1/2})$. We define a new norm over the space $\prod_{i=1}^{k} Y_i$, that is, $\forall \varepsilon \in \prod_{i=1}^{k} Y_i$,

$$\|\varepsilon\|_*^2 = \lambda^{-1}\|\varepsilon\|_0^2, \quad \|\varepsilon_{ij}\|_*^2 = \lambda^{-1}\|\varepsilon_{ij}\|_0^2. \tag{5.3}$$

Some lemmas are needed to prove Theorem 5.1.

LEMMA 5.2. *If* $(\bar{e}, \bar{\varepsilon}) \in C \otimes W$, $j, k \in N(i)$, *then*

$$\|\bar{\varepsilon}_{ik}\|_*^2 \leq C(h^{-1/2}H^{1/2}|\bar{e}_i|_{1,h,\Omega_i}^2 + h^{-1/2}H^{-1/2}\|\bar{\pi}_{ij}\bar{e}_i\|_{0,\gamma_{ij}}^2),$$

where $\bar{\pi}_{ij}$ *is the complexification of* π_{ij} *just as in Definition* 3.1 *and* C *is independent of* h *and* H.

Proof. Suppose

$$(\bar{e}, \bar{\varepsilon}) = (\tilde{e}, \tilde{\varepsilon}) + \sqrt{-1}(\hat{e}, \hat{\varepsilon}),$$

where $(\tilde{e},\tilde{\varepsilon}),(\hat{e},\hat{\varepsilon}) \in W$; then by Lemma 3.2 and Remark 3.1,

$$\|\bar{\varepsilon}_{ik}\|_*^2 = \|\tilde{\varepsilon}_{ik}\|_*^2 + \|\hat{\varepsilon}_{ik}\|_*^2,$$

$$|\bar{e}_i|_{1,h,\Omega_i}^2 = |\tilde{e}_i|_{1,h,\Omega_i}^2 + |\hat{e}_i|_{1,h,\Omega_i}^2,$$

and

$$\|\bar{\pi}_{ij}\bar{e}_i\|_{0,\gamma_{ij}}^2 = \|\pi_{ij}\tilde{e}_i\|_{0,\gamma_{ij}}^2 + \|\pi_{ij}\hat{e}_i\|_{0,\gamma_{ij}}^2.$$

So we need only to verify that $\forall (e,\varepsilon) \in W$, we have

$$\|\varepsilon_{ik}\|_*^2 \le C(h^{-1/2}H^{1/2}|e_i|_{1,h,\Omega_i}^2 + h^{-1/2}H^{-1/2}\|\pi_{ij}e_i\|_{0,\gamma_{ij}}^2).$$

By Lemma 2.7, we know that

$$\|\pi_{ik}e_i\|_{0,\gamma_{ik}}^2 \le CH|e_i|_{1,h,\Omega_i}^2 + C\|\pi_{ij}e_i\|_{0,\gamma_{ij}}^2. \tag{5.4}$$

Since $(e,\varepsilon) \in W$, replace V_i in (2.11) by $S_{ij}\varepsilon_{ij}$, and we have

$$\begin{aligned}
\|\epsilon_{ik}\|_0^2 &= \int_{\gamma_{ik}} \varepsilon_{ik}\cdot\pi_{ik}S_{ik}\varepsilon_{ik} \\
&= a_i(e_i, S_{ik}\varepsilon_{ik}) + \lambda\int_{\gamma_{ik}} \pi_{ik}e_i\cdot\epsilon_{ik} \\
&\le |e_i|_{1,h,\Omega_i}|S_{ik}\epsilon_{ik}|_{1,h,\Omega_i} + \lambda\|\pi_{ik}e_i\|_{0,\gamma_{ik}}\|\varepsilon_{ik}\|_0 \\
&\le Ch^{-1/2}|e_i|_{1,h,\Omega_i}\|\varepsilon_{ik}\|_0 + C\lambda(H^{1/2}|e_i|_{1,h,\Omega_i} + \|\pi_{ij}e_i\|_{0,\gamma_{ij}})\|\varepsilon_{ik}\|_0 \\
&\le C(h^{-1/2} + \lambda H^{1/2})|e_i|_{1,h,\Omega_i}\|\varepsilon_{ik}\|_0 + C\lambda\|\pi_{ij}e_i\|_{0,\gamma_{ij}}\|\varepsilon_{ik}\|_0.
\end{aligned}$$

Here we have used (2.5) in the first and second equalities and Lemma 2.1 and (5.4) in the second inequality, respectively. Then

$$\|\epsilon_{ik}\|_0^2 \le C(h^{-1/2} + \lambda H^{1/2})^2|e_i|_{1,h,\Omega_i}^2 + C\lambda^2\|\pi_{ij}e_i\|_{0,\gamma_{ij}}^2,$$

or

$$\|\epsilon_{ik}\|_*^2 \le C(\lambda^{-1/2}h^{-1/2} + \lambda^{1/2}H^{1/2})^2|e_i|_{1,h,\Omega_i}^2 + C\lambda\|\pi_{ij}e_i\|_{0,\gamma_{ij}}^2.$$

Since $\lambda = O(h^{-1/2}H^{-1/2})$, we finish the proof. ∎

LEMMA 5.3. *If $(\bar{e},\bar{\varepsilon}) \in C \otimes W$ is an eigenvector of $\bar{A}$ such that $\bar{A}(\bar{e},\bar{\varepsilon}) = \sigma(\bar{e},\bar{\varepsilon})$, then*

$$\bar{\varepsilon} \neq 0, \quad |\sigma| \le 1, \tag{5.5}$$

$$\sigma\bar{\varepsilon}_{ij} = 2\lambda\bar{\pi}_{ji}\bar{e}_j - \bar{\varepsilon}_{ji}, \tag{5.6}$$

$$|\sigma|^2\|\bar{\varepsilon}\|_*^2 = \|\bar{\varepsilon}\|_*^2 - 4a(\bar{e},\bar{e}), \tag{5.7}$$

where $|\sigma|$ is modulus of σ.

Proof. By the definition of A, (2.12), and Definition 3.3, we get (5.6) immediately. We can conclude that $\bar{\varepsilon} \neq 0$. Otherwise, if $\bar{\varepsilon} = 0$, by (2.11), we know $\bar{e} = 0$, so $(\bar{e},\bar{\varepsilon}) = 0$. However, since $(\bar{e},\bar{\varepsilon})$ is an eigenvector, then $(\bar{e},\bar{\varepsilon}) \neq 0$. This is a contradiction.

Suppose

$$(\bar{e},\bar{\varepsilon}) = (\tilde{e},\tilde{\varepsilon}) + \sqrt{-1}(\hat{e},\hat{\varepsilon}),$$

where $(\tilde{e},\tilde{\varepsilon}),(\hat{e},\hat{\varepsilon}) \in W$; then

$$\|\sigma\bar{\varepsilon}_{ij}\|_0^2 = \|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon}_{ji}\|_0^2 + \|2\lambda\pi_{ji}\hat{e}_j - \hat{\varepsilon}_{ji}\|_0^2,$$

or

$$|\sigma|^2\|\bar{\varepsilon}_{ij}\|_*^2 = \|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon}_{ji}\|_*^2 + \|2\lambda\pi_{ji}\hat{e}_j - \hat{\varepsilon}_{ji}\|_*^2, \tag{5.8}$$

$$\|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon}_{ji}\|_*^2 = \|\tilde{\varepsilon}_{ji}\|_*^2 - 4\langle\pi_{ji}\tilde{e}_j, \tilde{\varepsilon}_{ji} - \lambda\pi_{ji}\tilde{e}_j\rangle,$$

$$\begin{aligned}
&\sum_{1\le i<j\le K}(\|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon}_{ji}\|_*^2 + \|2\lambda\pi_{ij}\tilde{e}_i - \tilde{\varepsilon}_{ij}\|_*^2)\\
&= \sum_{\substack{1\le i<j\le K\\ j\in N(i)}}(\|\tilde{\epsilon}_{ij}\|_*^2 + \|\tilde{\epsilon}_{ji}\|_*^2 - 4\langle\pi_{ji}\tilde{e}_j, \tilde{\epsilon}_{ji} - \lambda\pi_{ji}\tilde{e}_j\rangle|_{\gamma_{ji}} - 4\langle\pi_{ij}\tilde{e}_i, \tilde{\epsilon}_{ij} - \lambda\pi_{ij}\tilde{e}_i\rangle|_{\gamma_{ij}})\\
&= \sum_{i=1}^K\sum_{j\in N(i)}\|\tilde{\epsilon}_{ij}\|_*^2 - 4\sum_{i=1}^K\sum_{j\in N(i)}\langle\pi_{ij}\tilde{e}_i, \tilde{\epsilon}_{ij} - \lambda\pi_{ij}\tilde{e}_i\rangle|_{\gamma_{ij}}.
\end{aligned}$$

Since $(\tilde{e},\tilde{\varepsilon}) \in W$, by (2.11),

$$\sum_{j\in N(i)}\langle\pi_{ij}\tilde{e}_i, \tilde{\epsilon}_{ij} - \lambda\pi_{ij}\tilde{e}_i\rangle|_{\gamma_{ij}} = a_i(\tilde{e}_i,\tilde{e}_i).$$

Then

$$\sum_{1\le i<j\le K}(\|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon_{ji}}\|_*^2 + \|2\lambda\pi_{ij}\tilde{e}_i - \tilde{\varepsilon_{ij}}\|_*^2) = \|\tilde{\varepsilon}\|_*^2 - 4a(\tilde{e},\tilde{e}).$$

Similarly, we also have

$$\sum_{1\le i<j\le K}(\|2\lambda\pi_{ji}\hat{e}_j - \hat{\varepsilon_{ji}}\|_*^2 + \|2\lambda\pi_{ij}\hat{e}_i - \hat{\varepsilon_{ij}}\|_*^2) = \|\hat{\varepsilon}\|_*^2 - 4a(\hat{e},\hat{e}).$$

So

$$\begin{aligned}
&|\sigma|^2\|\bar{\varepsilon}\|_*^2\\
&= \sum_{1\le i<j\le K}(|\sigma|^2\|\bar{\varepsilon}_{ij}\|_*^2 + |\sigma|^2\|\bar{\varepsilon}_{ji}\|_*^2)\\
&= \sum_{1\le i<j\le K}(\|2\lambda\pi_{ji}\tilde{e}_j - \tilde{\varepsilon_{ji}}\|_*^2 + \|2\lambda\pi_{ij}\tilde{e}_i - \tilde{\varepsilon_{ij}}\|_*^2)\\
&\quad + \sum_{1\le i<j\le K}(\|2\lambda\pi_{ji}\hat{e}_j - \hat{\varepsilon_{ji}}\|_*^2 + \|2\lambda\pi_{ij}\hat{e}_i - \hat{\varepsilon_{ij}}\|_*^2)\\
&= \|\tilde{\varepsilon}\|_*^2 - 4a(\tilde{e},\tilde{e}) + \|\hat{\varepsilon}\|_*^2 - 4a(\hat{e},\hat{e})\\
&= \|\bar{\varepsilon}\|_*^2 - 4a(\bar{e},\bar{e}).
\end{aligned}$$

Thus we get (5.7).

We know $|\sigma| \le 1$ since $a(\bar{e},\bar{e}) \ge 0$ and $\bar{\varepsilon} \ne 0$. ∎

Remark 5.3. By a similar argument in the proof of Lemma 5.3, we can prove the following important equality,

$$\|\epsilon^{n+1}\|_*^2 = \|\epsilon^n\|_*^2 - 4a(e^n, e^n). \tag{5.9}$$

This equality discloses the convergence behavior of this method explicitly.

LEMMA 5.4. *If $(\bar{e},\bar{\varepsilon})$ is an eigenvector of $\bar{A}$ such that $\bar{A}(\bar{e},\bar{\varepsilon}) = \sigma(\bar{e},\bar{\varepsilon})$, $\sigma \ne 0$, $j \in N(i)$, then*

$$\|\bar{\pi}_{ij}\bar{e}_i\|_0^2 \le C|\sigma|^{-2}h^{1/2}H^{1/2}\|\bar{\varepsilon}_{ji}\|_*^2 + C|\sigma|^{-2}\|\bar{\pi}_{ji}\bar{e}_j\|_0^2,$$

where C is a constant independent of h and H.

Proof. By (5.6), we have

$$|\sigma|^2\|\bar{\varepsilon}_{ij}\|_0^2 \le C\lambda^2\|\bar{\pi}_{ji}\bar{e}_j\|_0^2 + C\|\bar{\varepsilon}_{ji}\|_0^2$$

or

$$\|\bar{\varepsilon}_{ij}\|_*^2 \le C|\sigma|^{-2}\lambda\|\bar{\pi}_{ji}\bar{e}_j\|_0^2 + C|\sigma|^{-2}\|\bar{\varepsilon}_{ji}\|_*^2. \tag{5.10}$$

On the other hand, also by (5.6), we know

$$2\lambda\bar{\pi}_{ij}\bar{e}_i = \sigma\bar{\varepsilon}_{ji} + \bar{\varepsilon}_{ij};$$

then

$$\lambda^2\|\bar{\pi}_{ij}\bar{e}_i\|_0^2 \le C|\sigma|^2\|\bar{\varepsilon}_{ji}\|_0^2 + C\|\bar{\varepsilon}_{ij}\|_0^2,$$

and

$$\begin{aligned}\lambda\|\bar{\pi}_{ij}\bar{e}_i\|_0^2 &\le C|\sigma|^2\|\bar{\varepsilon}_{ji}\|_*^2 + C\|\bar{\varepsilon}_{ij}\|_*^2\\ &\le C|\sigma|^2\|\bar{\varepsilon}_{ji}\|_*^2 + C|\sigma|^{-2}\lambda\|\bar{\pi}_{ji}\bar{e}_j\|_0^2 + C|\sigma|^{-2}\|\bar{\varepsilon}_{ji}\|_*^2\\ &\le C2|\sigma|^{-2}\|\bar{\varepsilon}_{ji}\|_*^2 + C|\sigma|^{-2}\lambda\|\bar{\pi}_{ji}\bar{e}_j\|_0^2,\end{aligned}$$

where we have used (5.10) and the fact that $|\sigma| \le 1$ in the second and third inequalities, respectively. As a result,

$$\|\bar{\pi}_{ij}\bar{e}_i\|_0^2 \le C|\sigma|^{-2}\lambda^{-1}\|\bar{\varepsilon}_{ji}\|_*^2 + C|\sigma|^{-2}\|\bar{\pi}_{ji}\bar{e}_j\|_0^2.$$

Recall that $\lambda = O(h^{-1/2}H^{-1/2})$, and we get the conclusion. ∎

LEMMA 5.5. *If $\Omega_{i^r} \in D_r$ and there is a path $\{\Omega_{i^1},\dots,\Omega_{i^r}\}$ connecting Ω_{i^r} with boundary, then for all eigenvectors $(\bar{e},\bar{\varepsilon})$ such that $\bar{A}(\bar{e},\bar{\varepsilon}) = \sigma(\bar{e},\bar{\varepsilon})$, $\sigma \ne 0$,we have*

$$\begin{aligned}&\|\bar{\varepsilon}_{i^rk}\|_*^2\\ &\le C_3h^{-1/2}H^{1/2}|\bar{e}_{i^r}|_{1,h,\Omega_{i^r}}^2 + (C_3)^3 2|\sigma|^{-2}h^{-1/2}H^{1/2}|\bar{e}_{i^{r-1}}|_{1,h,\Omega_{i^{r-1}}}^2\\ &\quad + \cdots + (C_3)^{2r-1}(2|\sigma|^{-2})^{r-1}h^{-1/2}H^{1/2}|\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2,\end{aligned}$$

where $C_3 > 1$ is independent of h, H, and the choice of eigenvector $(\bar{e},\bar{\varepsilon})$.

Proof. We verify it by induction on r.

By Lemma 5.2, Lemma 5.4, and (5.4), we know that there is a constant $C_3 > 1$ such that $\forall j,k \in N(i)$, the following three inequalities hold:

$$\|\bar{\pi}_{ik}\bar{e}_i\|_{0,\gamma_{ik}}^2 \le C_3H|\bar{e}_i|_{1,h,\Omega_i}^2 + C_3\|\bar{\pi}_{ij}\bar{e}_i\|_{0,\gamma_{ij}}^2, \tag{5.11}$$

$$\|\bar{\varepsilon}_{ik}\|_*^2 \le C_3 h^{-1/2} H^{1/2} |\bar{e}_i|_{1,h,\Omega_i}^2 + C_3 h^{-1/2} H^{-1/2} \|\bar{\pi}_{ij}\bar{e}_i\|_{0,\gamma_{ij}}^2, \tag{5.12}$$

$$\|\bar{\pi}_{ij}\bar{e}_i\|_0^2 \le C_3 |\sigma|^{-2} h^{1/2} H^{1/2} \|\bar{\varepsilon}_{ji}\|_*^2 + C_3 |\sigma|^{-2} \|\bar{\pi}_{ji}\bar{e}_j\|_0^2. \tag{5.13}$$

It is necessary to point out that C_3 is independent of h, H, and the choice of eigenvector $(\bar{e}, \bar{\varepsilon})$.

When $r = 1$, i.e., $\Omega_{i^1} \in D_1$, there is at least one face of Ω_{i^1} belonging to $\partial\Omega$. Let this face play the role of γ_{ij} in (5.11) and (5.12); since $\bar{\pi}\bar{e}_{i^1}$ vanishes on this face, we have

$$\|\bar{\pi}_{i^1 k}\bar{e}_{i^1}\|_0^2 \le C_3 H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2, \tag{5.14}$$

$$\|\bar{\varepsilon}_{i^1 k}\|_*^2 \le C_3 h^{-1/2} H^{1/2} |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2. \tag{5.15}$$

So the conclusion holds in the case of $r = 1$.

When $r = 2$, i.e., $\Omega_{i^2} \in D_2$, there is a path $\{\Omega_{i^1}, \Omega_{i^2}\}$. So on one hand we can apply (5.11), (5.12), and (5.13) to Ω_{i^2}; on the other hand we can also use the hypothesis of induction (5.14) and (5.15) since Ω_{i^1} must belong to D_1.

By (5.13), (5.15), and (5.14) we have

$$\begin{aligned} \|\bar{\pi}_{i^2 i^1}\bar{e}_{i^2}\|_0^2 &\le C_3 |\sigma|^{-2} h^{1/2} H^{1/2} \|\bar{\varepsilon}_{i^1 i^2}\|_*^2 + C_3 |\sigma|^{-2} \|\bar{\pi}_{i^1 i^2}\bar{e}_{i^1}\|_0^2 \\ &\le (C_3)^2 |\sigma|^{-2} H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2 + (C_3)^2 |\sigma|^{-2} H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2 \\ &= (C_3)^2 2 |\sigma|^{-2} H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2. \end{aligned} \tag{5.16}$$

By (5.11) and (5.16), we have $\forall k \in N(i^2)$

$$\begin{aligned} &\|\bar{\pi}_{i^2 k}\bar{e}_{i^2}\|_0^2 \\ &\le C_3 H |\bar{e}_{i^2}|_{1,h,\Omega_{i^2}}^2 + C_3 \|\bar{\pi}_{i^2 i^1}\bar{e}_{i^2}\|_0^2 \\ &\le C_3 H |\bar{e}_{i^2}|_{1,h,\Omega_{i^2}}^2 + (C_3)^3 2 |\sigma|^{-2} H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2. \end{aligned} \tag{5.17}$$

By (5.12) and (5.16), we also get

$$\begin{aligned} &\|\bar{\varepsilon}_{i^2 k}\|_*^2 \\ &\le C_3 h^{-1/2} H^{1/2} |\bar{e}_{i^2}|_{1,h,\Omega_{i^2}}^2 + C_3 h^{-1/2} H^{-1/2} \|\bar{\pi}_{i^2 i^1}\bar{e}_{i^2}\|_0^2 \\ &\le C_3 h^{-1/2} H^{1/2} |\bar{e}_{i^2}|_{1,h,\Omega_{i^2}}^2 + (C_3)^3 2 |\sigma|^{-2} h^{-1/2} H^{1/2} |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2. \end{aligned} \tag{5.18}$$

So the conclusion also holds for $r = 2$.

In general, if the conclusion holds for $r - 1$, we consider the case of r. We know there is a path $\{\Omega_{i^1}, \ldots, \Omega_{i^{r-1}}, \Omega_{i^r}\}$. We can get the estimation for Ω_{i^r} by using (5.11), (5.12), (5.13), and the conclusion for $r - 1$ since $\{\Omega_{i^1}, \ldots, \Omega_{i^{r-1}}\}$ is a path connecting $\Omega_{i^{r-1}}$ with the boundary. In other words, we can get $\forall k \in N(i^r)$

$$\begin{aligned} &\|\bar{\pi}_{i^r k}\bar{e}_{i^r}\|_0^2 \\ &\le C_3 H |\bar{e}_{i^r}|_{1,h,\Omega_{i^r}}^2 + (C_3)^3 2 |\sigma|^{-2} H |\bar{e}_{i^{r-1}}|_{1,h,\Omega_{i^{r-1}}}^2 \\ &\quad + \cdots + (C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} H |\bar{e}_{i^1}|_{1,h,\Omega_{i^1}}^2, \end{aligned}$$

$$\begin{aligned}&\|\bar{\varepsilon}_{i^r k}\|_*^2\\ &\le C_3 h^{-1/2} H^{1/2} |\bar{e}_{i^r}|^2_{1,h,\Omega_{i^r}} + (C_3)^3 2|\sigma|^{-2} h^{-1/2} H^{1/2} |\bar{e}_{i^{r-1}}|^2_{1,h,\Omega_{i^{r-1}}}\\ &\quad + \cdots + (C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} h^{-1/2} H^{1/2} |\bar{e}_{i^1}|^2_{1,h,\Omega_{i^1}}.\end{aligned}$$

Thus we get the conclusion in general. □

Now we are in a position to prove the main theorem.

Proof of Theorem 5.1. By Corollary 2.4, we know

$$\|(e^n, \varepsilon^n)\| \le \|A^n\| \|(e^0, \varepsilon^0)\|. \tag{5.19}$$

In addition, by Lemma 3.6, we need only to estimate $\rho(\bar{A})$. This is our main task.

Suppose $(\bar{e}, \bar{\varepsilon})$ is an eigenvector of $\bar{A}$ and $\sigma \neq 0$ is certain eigenvalue. By the condition (1) of Assumption 4.1, $\forall \Omega_{i^r} \in D_r$, Ω_{i^r} has a path $\{\Omega_{i^1}, \Omega_{i^2}, \ldots, \Omega_{i^r}\}$. By Lemma 5.5, $\forall k \in N(i^r)$,

$$\begin{aligned}&\|\bar{\varepsilon}_{i^r k}\|_*^2\\ &\le C_3 h^{-1/2} H^{1/2} |\bar{e}_{i^r}|^2_{1,h,\Omega_{i^r}} + (C_3)^3 2|\sigma|^{-2} h^{-1/2} H^{1/2} |\bar{e}_{i^{r-1}}|^2_{1,h,\Omega_{i^{r-1}}}\\ &\quad + \cdots + (C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} h^{-1/2} H^{1/2} |\bar{e}_{i^1}|^2_{1,h,\Omega_{i^1}}.\end{aligned}$$

We assume each Ω_i has at most M faces. So

$$\begin{aligned}&\sum_{k\in N(i^r)} \|\bar{\varepsilon}_{i^r k}\|_*^2\\ &\le MC_3 h^{-1/2} H^{1/2} |\bar{e}_{i^r}|^2_{1,h,\Omega_{i^r}} + M(C_3)^3 2|\sigma|^{-2} h^{-1/2} H^{1/2} |\bar{e}_{i^{r-1}}|^2_{1,h,\Omega_{i^{r-1}}}\\ &\quad + \cdots + M(C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} h^{-1/2} H^{1/2} |\bar{e}_{i^1}|^2_{1,h,\Omega_{i^1}}.\end{aligned}$$

By the condition (2) of Assumption 4.1, we have

$$\begin{aligned}&\sum_{\Omega_{i^r}\in D_r} \sum_{j\in N(i^r)} \|\bar{\varepsilon}_{i^r j}\|_*^2\\ &\le MC_3 h^{-1/2} H^{1/2} \sum_{\Omega_{i^r}\in D_r} |\bar{e}_{i^r}|^2_{1,h,\Omega_{i^r}}\\ &\quad + M(C_3)^3 2|\sigma|^{-2} h^{-1/2} H^{1/2} \sum_{\Omega_{i^{r-1}}\in D_{r-1}} |\bar{e}_{i^{r-1}}|^2_{1,h,\Omega_{i^{r-1}}}\\ &\quad + \cdots + M(C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} h^{-1/2} H^{1/2} \sum_{\Omega_{i^1}\in D_1} |\bar{e}_{i^1}|^2_{1,h,\Omega_{i^1}}.\end{aligned} \tag{5.20}$$

Summing up all subdomains by (5.20), we get

$$\begin{aligned}\|\bar{\varepsilon}\|_*^2 &= \sum_{r=1}^{N} \sum_{\Omega_{i^r}\in D_r} \sum_{j\in N(i^r)} \|\bar{\varepsilon}_{i^r j}\|_*^2\\ &\le Mh^{-1/2} H^{1/2} \sum_{r=1}^{N} (C_3)^{2r-1} (2|\sigma|^{-2})^{r-1} a(\bar{e}, \bar{e})\\ &\le (C_1)^{-1} |\sigma|^{-2N} (2(C_3)^2)^N h^{-1/2} H^{1/2} a(\bar{e}, \bar{e}),\end{aligned}$$

where C_1 is independent of h, H, N, and the choice of eigenvector $(\bar{e}, \bar{\varepsilon})$. By (5.7), we have

$$\begin{aligned}|\sigma|^2\|\bar{\varepsilon}\|_*^2 &= \|\bar{\varepsilon}\|_*^2 - 4a(\bar{e},\bar{e})\\ &\le \|\bar{\varepsilon}\|_*^2 - 4C_1|\sigma|^{2N}(2(C_3)^2)^{-N}h^{1/2}H^{-1/2}\|\bar{\varepsilon}\|_*^2.\end{aligned}$$

Moreover, by (5.5), $\bar{\varepsilon} \neq 0$, so

$$|\sigma|^2 \le 1 - 4C_1|\sigma|^{2N}(2(C_3)^2)^{-N}h^{1/2}H^{-1/2}.$$

If $|\sigma|^2 > 1/2$, then

$$|\sigma|^2 \le 1 - 4C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2}.$$

As a result,

$$|\sigma|^2 \le \max\{1/2, 1 - 4C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2}\}.$$

Since

$$1 - 4C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2} \ge 1/2,$$

when h tends to 0, we get

$$|\sigma|^2 \le 1 - 4C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2}.$$

Thus

$$|\sigma| \le (1 - 4C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2})^{1/2} \le 1 - 2C_1(4(C_3)^2)^{-N}h^{1/2}H^{-1/2}.$$

Let $C_0 = (4(C_3)^2)^{-1}$; then

$$\rho(\bar{A}) \le 1 - 2C_1(C_0)^N h^{1/2}H^{-1/2}, \tag{5.21}$$

where $C_0 \in (0,1)$ and C_1 are independent of h, H, and N.

By Lemma 3.6, (5.19), and (5.21) we finish this proof. ∎

6. Numerical experiments. Now we carry out some numerical experiments to check the convergence behavior of this method. We find that this method converges quickly when the winding number N is not very large. But the convergence rate will deteriorate when the winding number N goes up.

The domain in which our problems are defined is $[0, L \times H] \times [0, M \times H]$. We decompose the domain into $L \times M$ subdomains. Each subdomain is a square whose edge length is H. The problem is (2.1) and the exact solution u is $x(x - L \times H)y(y - M \times H)$. We triangulate the domain uniformly and the mesh size is h.

We will consider four cases. In these cases, $L \times M$ is 10×2, 5×3, 5×5, and 7×7, respectively. The winding numbers of these four cases are 1, 2, 3, and 4, respectively. In addition we will modify H in each case.

We choose the initial guess $g_{ij}^0 = 0$ in each case (see (2.7)). The stop criterion is $\|u^n - u\|_\infty \le 10^{-4}$. The iteration number is n. Let the initial error be $e^0 = \|u^0 - u\|_\infty$ and the final one be $e^n = \|u^n - u\|_\infty$. Since the iteration number n is dependent on e^0, we compare the convergence speed to the convergence rate $(e^n/e^0)^{1/n}$ instead of n. The smaller $(e^n/e^0)^{1/n}$ is, the quicker the convergence will be.

In this paper, we choose the relaxation parameter $\lambda = O(h^{-1/2}H^{-1/2})$ in theory. Here we denote $h^{-1/2}H^{-1/2}$ by λ_0. We search the optimal parameter λ_{opt} in each experiment. We find that $\lambda_{opt}/\lambda_0 \approx 0.2$. Table 1 displays the numerical results.

It is seen from the table that the convergence rate of the DDM converges quickly when the winding number is not very large, but it will deteriorate when the winding number N goes up. This numerical result confirms our theoretical investigation.

TABLE 1

$L \times M$	N	h	H	λ_{opt}/λ_0	n	e^n	$(e^n/e^0)^{1/n}$
10×2	1	0.01	0.3	0.24	8	$6.6749e-5$	0.3805
		0.01	0.2	0.25	6	$6.4250e-5$	0.3623
5×3	2	0.01	0.3	0.20	10	$7.3636e-5$	0.4965
		0.01	0.2	0.25	7	$4.0014e-5$	0.4250
5×5	3	0.01	0.3	0.16	14	$8.2826e-5$	0.5647
		0.01	0.2	0.20	10	$6.6438e-5$	0.5166
7×7	4	0.01	0.3	0.13	24	$6.7312e-5$	0.6688
		0.01	0.2	0.16	18	$3.9613e-5$	0.6214

7. Conclusion. The Robin-type nonoverlapping DDM is a convenient, applicable, and highly parallel tool for solving the second order elliptic partial differential equations. The convergence rate is $1-O(h^{1/2}H^{-1/2})$ in certain cases when the lower term of the equation is strictly positive (see [22]). In this paper, we point out the convergence rate is $1-O((C_0)^N h^{1/2}H^{-1/2})$ when the lower term of the equation vanishes. Here h is size of mesh, H is the size of subdomain, and N is the winding number of domain decomposition (see Definition 4.1). $C_0 \in (0,1)$ is independent of h, H, and N. The numerical results in this paper support our theory. Given the fact claimed in [22] that the convergence rate of this method cannot be better than $1-O(h^{1/2}H^{-1/2})$ whether or not the lower term is positive definite strictly, the result in this paper is also sharp.

Appendix. Now we give the proofs of Lemmas 3.2, 3.4, and 3.5.

Proof of Lemma 3.2. It is easy to check that

$$\langle x+\sqrt{-1}y, x+\sqrt{-1}y\rangle \geq 0,$$

where the equality holds if and only if $x+\sqrt{-1}y=0$, and

$$\langle x_1+\sqrt{-1}y_1, x_2+\sqrt{-1}y_2\rangle = \overline{\langle x_2+\sqrt{-1}y_2, x_1+\sqrt{-1}y_1\rangle},$$

where $\bar{z}$ means the conjugate complex number of z. Moreover

$$\begin{aligned}&\langle (x_1+\sqrt{-1}y_1)+(x_3+\sqrt{-1}y_3), x_2+\sqrt{-1}y_2\rangle\\ &= \langle x_1+\sqrt{-1}y_1, x_2+\sqrt{-1}y_2\rangle + \langle x_3+\sqrt{-1}y_3, x_2+\sqrt{-1}y_2\rangle,\end{aligned}$$

$$\langle (a+\sqrt{-1}b)(x_1+\sqrt{-1}y_1), x_2+\sqrt{-1}y_2\rangle = (a+\sqrt{-1}b)\langle x_1+\sqrt{-1}y_1, x_2+\sqrt{-1}y_2\rangle.$$

So we have defined an inner product of complex linear space $C\otimes V$. Meanwhile

$$\|x+\sqrt{-1}y\|^2 = \langle x+\sqrt{-1}y, x+\sqrt{-1}y\rangle = \langle x,x\rangle + \langle y,y\rangle = \|x\|^2+\|y\|^2. \qquad \square$$

Proof of Lemma 3.4.

$$\begin{aligned}(1\otimes T_1)(1\otimes T_2)(x+\sqrt{-1}y) &= (1\otimes T_1)(T_2x+\sqrt{-1}T_2y)\\ &= T_1T_2x+\sqrt{-1}T_1T_2y\\ &= 1\otimes(T_1T_2)(x+\sqrt{-1}y),\end{aligned}$$

so we get the conclusion. $\square$

Proof of Lemma 3.5. By Lemma 3.2,

$$\|\bar{T}\|^2 = \sup_{x+\sqrt{-1}y\neq 0} \frac{\|\bar{T}(x+\sqrt{-1}y)\|^2}{\|x+\sqrt{-1}y\|^2} = \sup_{x+\sqrt{-1}y\neq 0} \frac{\|Tx\|^2+\|Ty\|^2}{\|x\|^2+\|y\|^2}$$
$$\leq \sup_{x+\sqrt{-1}y\neq 0} \frac{\|T\|^2\|x\|^2+\|T\|^2\|y\|^2}{\|x\|^2+\|y\|^2} = \|T\|^2,$$

so

$$\|\bar{T}\| \leq \|T\|.$$

On the other hand, we know there is a vector $x_0 \in V$ such that $\|x_0\| = 1$ and $\|Tx_0\| = \|T\|$, and also by Lemma 3.2,

$$\|\bar{T}\|^2 \geq \frac{\|\bar{T}(x_0+\sqrt{-1}x_0)\|^2}{\|x_0+\sqrt{-1}x_0\|^2} = \frac{2\|Tx_0\|^2}{2\|x_0\|^2} = \|T\|^2.$$

So $\|\bar{T}\| = \|T\|$. □

Acknowledgment. We thank the anonymous referees for many constructive comments and suggestions which led to an improved presentation of this paper.

REFERENCES

[1] T. Achdou, C. Japhet, P. Le Tallec, F. Nataf, F. Rogier, and M. Vidrascu, *Domain decomposition methods for nonsymmetric problems*, in Proceedings of the Eleventh International Conference on Domain Decopmsition Methods, DDM.org, Augsburg, Germany, 1999, pp. 3–17.

[2] L. S. Bennethum and X. Feng, *A domain decomposition method for solving a Helmholtz-like problem in elasticity based on the Wilson nonconforming element*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 1–25.

[3] S. C. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[4] S. C. Brenner, *Lower bounds for two-level additive Schwarz preconditioners with small overlap*, SIAM J. Sci. Comput., 21 (2000), pp. 1657–1669.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] Q. Deng, *An analysis for a nonoverlapping domain decomposition iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1525.

[7] Q. Deng, *A nonoverlapping domain decomposition method for nonconforming finite element problems*, Commun. Pure Appl. Anal., 2 (2003), pp. 295–306.

[8] B. Despres, *Domain decomposition method and Helmholtz problem*, in Mathematical and Numerical Aspects of Wave Propagation Phenomena, G. Cohen, L. Halpern, and P. Joly, eds., SIAM, Philadelphia, 1991, pp. 44–52.

[9] B. Despres, P. Joly, and J. E. Roberts, *A domain decomposition method for harmonic Maxwell equations*, in Iterative Methods in Linear Algebra, North–Holland, Amsterdam, 1992, pp. 475–484.

[10] S. Deparis, M. Discacciati, and A. Quarteroni, *A Domain Decomposition Framework for Fluid-structure Interaction Problems*, Tech. Report MOX 45, MOX - Modeling and Scientific Computing, Milan, Italy, 2004.

[11] M. Discacciati, *An Operator-splitting Approach to Nonoverlapping Domain Decomposition Methods*, Tech. Report 14, Section de Mathmatiques, EPFL, Lausanne, Switzerland, 2004.

[12] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[13] J. Douglas and C. S. Huang, *Accelerated domain decomposition iterative procedures for mixed methods based on Robin transmission conditions*, Calcolo, 35 (1998), pp. 131–147.

[14] J. Douglas and C. S. Huang, *An accelerated domain decomposition procedure based on Robin transmission conditions*, BIT, 37 (1997), pp. 678–686.

[15] X. Feng, *Analysis of a domain decomposition method for the nearly elastic wave equations based on mixed finite element methods*, IMA J. Numer. Anal., 18 (1998), pp. 229–250.

[16] M. J. Gander and G. H. Golub, *A nonoverlapping optimized Schwarz method which converges with an arbitrarily weak dependence on h*, in Proceedings of the Fourteenth International Conference on Domain Decomposition Methods, UNAM, Mexico City, 2002.

[17] M. J. Gander, L. Halpern, and F. Nataf, *Optimized Schwarz methods*, in Proceedings of the Twelfth International Conference on Domain Decomposition Methods, Chiba, Japan, 2001, pp. 15–28.

[18] M. J. Gander, F. Magoulès, and F. Nataf, *Optimized Schwarz methods without overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 24 (2002), pp. 38–60.

[19] W. Guo and L. S. Hou, *Generalizations and accelerations of Lions' nonoverlapping domain decomposition method for linear elliptic PDE*, SIAM J. Numer. Anal., 41 (2003), pp. 2056–2080.

[20] C. Japhet, F. Nataf, and F. Rogier, *The optimized order* 2 *method, application to convection-diffusion problems*, Future Generation Computer Systems FUTURE, 18 (2001).

[21] P. L. Lions, *On the Schwarz alternating method* III: *A variant for nonoverlapping subdomains*, in Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Perianx, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[22] L. Qin and X. Xu, *On the convergence rate of a parallel nonoverlapping domain decomoposition method*, SIAM J. Numer. Anal., submitted.

[23] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Oxford, UK, 1999.

[24] M. Sarkis, *Nonstandard coarse spaces and Schwarz methods for elliptic problems with discontinuous coefficients using nonconforming elements*, Numer. Math., 77 (1997), pp. 383–406.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2559–2583

STRONG TRACTABILITY OF QUASI-MONTE CARLO QUADRATURE USING NETS FOR CERTAIN BANACH SPACES*

RONG-XIAN YUE† AND FRED J. HICKERNELL‡

Abstract. We consider multivariate integration in the weighted spaces of functions with mixed first derivatives bounded in L_p norms and the weighted coefficients introduced via ℓ_q norms, where $p, q \in [1, \infty]$. The integration domain may be bounded or unbounded. The worst-case error and randomized error are investigated for quasi-Monte Carlo quadrature rules. For the worst-case setting the quadrature rule uses deterministic $((T_u), s)$-sequences in base b, and for the randomized setting the quadrature rule uses randomly scrambled digital $((T_u), m, s)$-nets in base b. Sufficient conditions are found under which multivariate integration is strongly tractable in the worst-case and randomized settings, respectively. Similar results hold for the Banach spaces of finite-order weights. Results presented in this article extend and improve upon those found previously.

Key words. weighted integration, quasi-Monte Carlo quadrature, strong tractability

AMS subject classifications. 65D32, 65C05

DOI. 10.1137/040621776

1. Introduction.

In many practical problems arising in statistics [FW94, Gen92], finance [PT96], and physics [Kei96], one often needs to approximate the integral of a function $f(\mathbf{x})$ over the s-dimensional unit cube,

$$I(f) = \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x}, \tag{1}$$

or over a bounded or unbounded s-dimensional box D with a nonnegative weight function $\rho(\mathbf{x})$,

$$I_\rho(f) = \int_D f(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x}. \tag{2}$$

It is assumed in this article that D is an s-dimensional box of the form

$$D = \overline{(\mathbf{a}, \mathbf{b})} := \overline{(a_1, b_1)} \times \cdots \times \overline{(a_s, b_s)} \subseteq R^s, \tag{3}$$

where each of the $\overline{(a_k, b_k)}$ may possibly be a finite, semi-infinite, or infinite interval. In addition, the weight function $\rho(\mathbf{x})$ is assumed to have product form

$$\rho(\mathbf{x}) = \prod_{k=1}^{s} \rho_k(x_k) \tag{4}$$

*Received by the editors December 30, 2004; accepted for publication (in revised form) May 26, 2006; published electronically December 11, 2006. This work was partially supported by Hong Kong Research Grants Council grant RGC/HKBU/2020/02P, Shanghai Municipal Education Commission (05DZ03), E-Institutes of Shanghai Municipal Education Commission (E03004), Shanghai Leading Academic Discipline Project (T0401), the Special Funds for Major Specialties of Shanghai Municipal Education Committee, and the NSFC grant 10671129.

http://www.siam.org/journals/sinum/44-6/62177.html

†Division of Scientific Computation, E-Institute of Shanghai Universities, and Department of Applied Mathematics, Shanghai Normal University, 100 Guilin Road, Shanghai 200234, People's Republic of China (yue2@shnu.edu.cn).

‡Department of Applied Mathematics, Illinois Institute of Technology, 10 West 32nd Street, Chicago, IL 60616-3793 (hickernell@iit.edu).

for nonnegative functions ρ_k, which are assumed to be probability density functions on (a_k, b_k) for simplicity. After a suitable transformation, the integration problem in (2) can be written as the form in (1). We call integration in (1) the classical problem. In this article we concentrate mainly on the classical problem, and then extend the results to the general problem.

For the classical problem with large s the integral is often approximated by algorithms of the following form:

$$Q_n(f) = \frac{1}{n} \sum_{j=0}^{n-1} f(\mathbf{x}^j), \tag{5}$$

where the point set $P = \{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{n-1}\}$ is carefully chosen from the unit cube $[0,1]^s$. In this article P is chosen by quasi-Monte Carlo and randomized quasi-Monte Carlo methods known as (T, m, s)-nets and (T, s)-sequences [Nie92, Chapter 4] and their scrambled versions [Owe95]. For fixed dimension s, quasi-Monte Carlo methods are usually considered to be more accurate than Monte Carlo methods, but for Monte Carlo methods it can be much easier to estimate accuracy. Randomized Monte Carlo methods combine the best of Monte Carlo and quasi-Monte Carlo methods [Owe95, Owe97a, Owe97b, Owe98, YM99, HY00, HHY04].

It is interesting to investigate the performance of quasi-Monte Carlo and randomized quasi-Monte Carlo methods for high or very high dimensional integration. This problem is related to the concepts of tractability and strong tractability [SW98]. Tractability means that one can reduce the initial error by a factor $\varepsilon \in (0,1)$ by using a number of function values which are polynomial in s and ε^{-1}. Strong tractability means that the number of samples is independent of s and depends polynomially on ε^{-1}. The smallest (or the infimum of) power of ε^{-1} is called the strong exponent of tractability. See [SW98] for a more precise description of tractability and strong tractability.

Strong tractability for integration is related to the class of integrands. There have been a few investigations in recent years concentrating on the strong tractability problem of quasi-Monte Carlo and randomized quasi-Monte Carlo methods based on (T, m, s)-nets and (T, s)-sequences for reproducing kernel Hilbert spaces [DP05, Wan03, YH01, YH05].

However, as pointed out in [Slo02], it is necessary to consider integration problems for Banach spaces of functions since the reproducing kernel Hilbert space methods are too restrictive for some classes of problems. For example, integrals from mathematical finance are typically with respect to probability measures over unbounded domains. After mapping to the unit cube most problems of this kind yield integrands whose derivatives are integrable, *but not square integrable*.

Consider the European put option as a simple one-dimensional example. The payoff is given by

$$g(z) = \max(K - S_0 \exp\{(r - \sigma^2/2)\tau + \sigma\sqrt{\tau} z\}, 0),$$

where S_0 is the initial asset price, τ is the time to maturity, r is the interest rate, σ is the volatility, K is the strike price, and z is a random variable with the standard Gaussian distribution, $\Phi(z)$. The fair price of the option is the mean or expected value of the payoff, i.e., $\int_{-\infty}^{\infty} g(z)\Phi'(z)\,dz = \int_0^1 f(x)\,dx$, where $f(x) = g(\Phi^{-1}(x))$. The integral in terms of x after a variable transformation is the classical integration

problem. The integral of the pth power of the first derivative of f is

$$\int_0^1 |f'(x)|^p \, dx = \int_{-\infty}^{\infty} |g'(z)|^p [\Phi'(z)]^{1-p} \, dz$$
$$= (2\pi)^{(p-1)/2} \int_{-\infty}^{\infty} |g'(z)|^p e^{(p-1)z^2/2} \, dz.$$

For the example in question $g'(z)$ decays exponentially to zero as $z \to -\infty$, is zero for z large enough, and has a jump discontinuity where $g(z)$ becomes positive. Thus, $g'(z)$ is p-integrable for all $p \geq 1$; however, due to the term $[\Phi'(z)]^{1-p}$, the function f' is integrable, but not square integrable. In fact, it is not p-integrable for any $p > 1$. The same conclusion holds if other fatter tailed distributions, such as the variance-gamma, are substituted for the Gaussian distribution.

Therefore, a fundamental difficulty arises in applying the Hilbert space results to such problems. Recently, there have been some studies of the tractability problem for weighted integration based on general quadrature rules and lattice rules for weighted Banach spaces of functions whose mixed partial derivatives are bounded in L_p norms for $p \in [1, \infty]$; see [HSW04a, HSW04b, HSW04c]. We think that it is also interesting to consider the quadrature rules that use (T, m, s)-nets and (T, s)-sequences for these spaces.

This article studies the tractability problems for weighted Banach spaces of integrands, in which two quasi-Monte Carlo rules are considered. One uses *deterministic* Niederreiter (T, s)-sequences, and another uses *randomly scrambled* Niederreiter digital (T, m, s)-nets. We refer to [Nie88] for Niederreiter sequences and [Nie92, Chapter 4; NX01, Chapter 8] for constructions of digital nets. For *deterministic* Niederreiter sequence rules we assume that the integrands f lie in a weighted Banach space, $\mathcal{F}^{(1)}_{p,q,\gamma,s}$, of functions whose mixed *anchored* first derivatives are bounded in L_p norms, and the weighted coefficients, $\gamma = \{\gamma_k\}_k$, are introduced via ℓ_q norms over the index u, where $p, q \in [1, \infty]$. This space is the same as that in [HSW04b]. For the *randomly scrambled* Niederreiter net rules, the class of integrands is a weighted Banach space, $\mathcal{F}^{(2)}_{p,q,\gamma,s}$, of functions whose *unanchored* mixed first derivatives are bounded in L_p norms and the weighted coefficients, $\gamma = \{\gamma_k\}_k$, are introduced via ℓ_q norms, where $p, q \in [1, \infty]$.

Note that the spaces $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ and $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ have the same smoothness. But the space $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ works technically better for the worst-case setting, and the space $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ works technically better for the randomized setting.

The reason for using Niederreiter sequences is that Niederreiter sequences are telescoping; i.e., to obtain a sequence in dimension $s + 1$, it suffices to add the last component x^j_{s+1} to the term of the s-dimensional sequence $(x^j_1, \ldots, x^j_s)$ for $j = 1, 2, \ldots$. Moreover, for any nonempty subset u of $\{1, \ldots, s\}$, the projection of the s-dimensional Niederreiter (T, s)-sequence onto the axes in u forms a $|u|$-dimensional Niederreiter sequence with quality parameter [Nie88]

$$T_u = \sum_{k \in u} [\deg(g_k) - 1], \tag{6}$$

where $g_1, \ldots, g_s$ denote the first s monic irreducible polynomials over the finite field F_b. This allows us to write the Niederreiter sequence and net as $((T_u), s)$-sequence and $((T_u), m, s)$-net, respectively, where (T_u) denotes a $(2^s - 1)$-dimensional vector of the quality parameters corresponding to all nonempty subsets u of $\{1, \ldots, s\}$. From

[Nie92, Wan03] one has the following upper bound on $\deg(g_k)$:

$$\deg(g_k) \le \log_b k + \log_b \log_b(k+b) + 2, \quad k = 1, 2, \ldots. \tag{7}$$

These properties make Niederreiter sequences suitable for tractability studies.

Note that very similar properties hold for some other sequences, such as Sobol sequences [Sob67, Sob69], Halton sequences [Hal60, HW02], generalized Niederreiter sequences [Nie92, Tez95], and certain constructions by Niederreiter and Xing [NX01]. Hence results very similar to the one presented in this article hold for those sequences. For instance, the Sobol sequence (in base 2) also has the telescopic property. And for any nonempty subset u of $\{1, \ldots, s\}$, the projection of the s-dimensional Sobol (T, s)-sequence onto the axes in u forms a $|u|$-dimensional Sobol sequence with quality parameter [Sob67]

$$T_u = \sum_{k \in u} [\deg(P_k) - 1],$$

where $P_1, \ldots, P_s$ denote the first s primitive polynomials. From [Sob69, Wan03] the degree $\deg(P_k)$ can be bounded by

$$\deg(P_k) \le \log_2 k + \log_2 \log_2(k+1) + \log_2 \log_2 \log_2(k+3) + C, \quad k = 1, 2, \ldots,$$

where C is a constant independent of k and s. Hence very similar strong tractability results to the one presented for Niederreiter sequences hold for Sobol sequences but the sufficient conditions are slightly stronger.

The main results of this article are Theorems 1 to 4, which provide sufficient conditions on strong tractability for $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ in the worst-case setting and $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ in the randomized setting. These results are summarized in Table 1, where p^* and q^* denote the conjugates of p and q, i.e.,

$$\frac{1}{p} + \frac{1}{p^*} = 1, \quad \frac{1}{q} + \frac{1}{q^*} = 1.$$

The asymptotic orders of the quadrature errors are given under the assumption that the sufficient condition for strong tractability holds. The parameter ϵ is an arbitrary positive number. For comparison, related results in [HSW04b, Wan03, YH05] are also listed in the table.

The following points are worth noting about these results:

(i) The spaces considered in [Wan03] are the weighted reproducing kernel Hilbert spaces, and the original weights γ_k in [Wan03] are the square of our weights. Therefore, the space in [Wan03] in which the weights are replaced with γ_k^2 becomes $\mathcal{F}^{(1)}_{2,2,\gamma,s}$. Our results substantially extend and improve upon the results of [Wan03] by considering more general spaces of integrands and deriving weaker sufficient conditions for strong tractability.

(ii) The setting in [YH05] is the *randomized worst case*, and the quadrature rule is based on a randomly scrambled Niederreiter sequence. Moreover, the weighted Sobolev–Hilbert spaces, $\mathcal{H}^{\mathrm{SH}}_{s,\gamma}$, are nearly the same as $\mathcal{F}^{(1)}_{2,2,\gamma,s}$. The sufficient condition there for strong tractability is somewhat weaker than the one in the present article. We do not know yet whether the condition for the worst-case setting in this article can be weakened or not.

TABLE 1
Summary results in the present article and [HSW04b, Wan03, YH05].

Setting	Article	Space	Rule	Sufficient condition for strong tractability	Error
Worst case	Present	$\mathcal{F}^{(1)}_{p,q,\gamma,s}$	sequence	$\sum_{k=1}^{\infty} \gamma_k^a k \ln k < \infty$ $(1 \le a \le q^*)$	$\mathcal{O}(n^{-1/a+\epsilon})$
Worst case	[HSW04b]	$\mathcal{F}^{(1)}_{p,q,\gamma,s}$	lattice	$\sum_{k=1}^{\infty} \gamma_k^a < \infty$ $(1 \le a \le q^*)$	$\mathcal{O}(n^{-1/a+\epsilon})$
Worst case	[Wan03]	$\mathcal{F}^{(1)}_{2,2,\gamma,s}$	sequence	$\sum_{k=1}^{\infty} \gamma_k k \ln k < \infty$	$\mathcal{O}(n^{-1+\epsilon})$
Randomized worst case	[YH05]	$\mathcal{H}^{\mathrm{SH}}_{s,\gamma}$	sequence	$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^2 < \infty$	$\mathcal{O}(n^{-1+\epsilon})$
Randomized	Present	$\mathcal{F}^{(2)}_{p,q,\gamma,s}$	net	$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^2 < \infty$ $(1 \le p \le \infty)$	$\mathcal{O}(n^{-1+\epsilon})$
				$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^3 < \infty$ $(2 \le p \le \infty)$	$\mathcal{O}(n^{-3/2+\epsilon})$
Randomized	[YH05]	$\mathcal{H}^{\mathrm{SH}}_{s,\gamma}$	net	$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^3 < \infty$	$\mathcal{O}(n^{-3/2+\epsilon})$

(iii) Compared with the lattice rules in [HSW04b], our sufficient condition for digital sequences in the worst-case setting is somewhat more stringent. In fact, we can conclude that our condition $\sum_{k=1}^{\infty} \gamma_k^a k \ln k < \infty$ $(1 \le a \le q^*)$ is roughly equivalent to $\sum_{k=1}^{\infty} \gamma_k^{a/2} < \infty$ $(1 \le a \le q^*)$ from the following Lemma 1, provided that $\gamma_1 \ge \gamma_2 \ge \cdots \ge 0$.

(iv) As far as we are aware, there have been few studies on randomized settings in literature. Sloan and Woźniakowski [SW01] studied the randomized error of the classical Monte Carlo algorithm for weighted Korobov spaces. Yue and Hickernell [YH05] studied the randomized error of the quasi-Monte Carlo algorithm based on scrambled Niederreiter nets and sequences for weighted Sobolev–Hilbert spaces. The result of the randomized error for $p = q = 2$ in this article is the same as that of [YH05]. Therefore, the results of this article extend the result of [YH05] by considering more general spaces of integrands.

(v) Compared with the condition with $a = 1$ in the worst-case setting for $\mathcal{F}^{(1)}_{p,q,\gamma,s}$, the condition in the randomized setting for $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ is weaker for $p \ge 1$. However, the results in the randomized setting are just for Niederreiter digital nets, unlike Niederreiter sequences in the worst-case setting.

LEMMA 1. *Let* $\{\gamma_k\}$ *be a nonnegative nonincreasing sequence, and let* λ_k *be a sequence satisfying*

$$\tilde{c}_{r,\delta} k^{r-\delta} \le \lambda_k \le c_{r,\delta} k^{r+\delta} \quad \forall \delta \in \left(0, \frac{1}{r}\right), \quad k = 1, 2, \ldots,$$

where $\tilde{c}_{r,\delta}$ and $c_{r,\delta}$ are two nonnegative constants depending only on r and δ. Then

$$\tilde{L}_{r,\delta}\left[\sum_{k=1}^{\infty}\gamma_k^{\frac{a(1+\delta)}{1+r}}\right]^{\frac{1+r}{1+\delta}} \le \sum_{k=1}^{\infty}\gamma_k^a\lambda_k \le L_{r,\delta}\left[\sum_{k=1}^{\infty}\gamma_k^{\frac{a(1-r\delta)}{1+r}}\right]^{\frac{1+r}{1-r\delta}},$$

where $\tilde{L}_{r,\delta}$ and $L_{r,\delta}$ are two nonnegative constants depending only on r and δ.

The proof of Lemma 1 is given in the appendix.

The article proceeds as follows: Section 2 considers the strong tractability of quasi-Monte Carlo rules that use deterministic Niederreiter sequences for the classical problem. Section 3 considers strong tractability of quasi-Monte Carlo rules that use the randomly scrambled Niederreiter digital nets for the classical problem. Section 4 extends the strong tractability results for the classical problem to the weighted integration over a general domain. Some concluding remarks are given in section 5, where it is shown that similar results to the one presented in previous sections hold for the Banach spaces of finite-order weights.

2. Tractability in worst-case settings for the classical problem. In this section, we deal with the strong tractability problem of multivariate integration using the deterministic Niederreiter sequence for the *classical problem*. We first introduce the spaces of our integrands, which are defined as in [HSW04a]. We briefly recall the definition as follows.

For a given $p \in [1,\infty]$, let $\mathcal{H}_{p,k}$ be the space of absolutely continuous functions $h : [0,1] \to R$ with p-integrable first derivatives, i.e., $h'(x) \in L_p([0,1])$. Let $\mathcal{H}_p^s = \otimes_{k=1}^s \mathcal{H}_{p,k}$ be the space consisting of linear combinations of functions of the following tensor product form:

$$f : [0,1]^s \to R \quad \text{and} \quad f(\mathbf{x}) = \prod_{k=1}^{s} h_k(x_k) \text{ with } h_k \in \mathcal{H}_{p,k}.$$

Take a sequence $\gamma = \{\gamma_k\}_k$ of positive numbers, and let $\gamma_\emptyset = 1$ and $\gamma_u = \prod_{k\in u}\gamma_k$ for nonempty subset $u \subseteq \{1,\dots,s\}$. Let $\mathbf{c}$ be a fixed point in $[0,1]^s$ and define the mixed *anchored* first derivative $f'_{u,\mathbf{c}}$ of $f(\mathbf{x})$ for any nonempty subset u by

$$f'_{u,\mathbf{c}}(\mathbf{x}_u) := \left(\prod_{k\in u}\frac{\partial}{\partial x_k}\right) f(\mathbf{x}_u, \mathbf{c}_{\bar{u}}),$$

where $(\mathbf{x}_u, \mathbf{c}_{\bar{u}})$ denotes the s-dimensional vector whose kth component is x_k if $k \in u$, and is c_k if $k \in \bar{u}$. Given an additional parameter $q \in [1,\infty]$, define the space $\mathcal{F}_{p,q,\gamma,s}^{(1)}$ to be the completion of $\mathcal{H}_p^s$ with respect to the norm

$$\|f\|_{p,q,\gamma,s} := \begin{cases} \left(\displaystyle\sum_{u\subseteq\{1,\dots,s\}} \gamma_u^{-q}\|f'_{u,\mathbf{c}}\|_{L_p}^q\right)^{1/q} & \text{for } q < \infty, \\ \displaystyle\max_{u\subseteq\{1,\dots,s\}}\{\gamma_u^{-1}\|f'_{u,\mathbf{c}}\|_{L_p}\} & \text{for } q = \infty, \end{cases} \tag{8}$$

where the sum is over all 2^s subsets of coordinates of $[0,1]^s$, and $f'_{\emptyset,\mathbf{c}} = f(\mathbf{c})$.

For the significance of the weights γ_k we refer to, e.g., [SW98, NW01] and note that in some articles, including [SW98, NW01, Wan03], $p = q = 2$ and the definition of $\|\cdot\|_{2,2,\gamma,s}$ uses γ_u^{-1} instead of γ_u^{-2}. Hence, our weights γ_k in (8) for $p = q = 2$ are the square roots of those in [SW98, NW01, Wan03].

For the weighted Banach space $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ defined above, the worst-case error of the quasi-Monte Carlo quadrature Q_n is defined as

$$e^{\mathrm{wo}}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s}) := \sup_{\|f\|_{p,q,\gamma,s}\le 1} |I(f) - Q_n(f)|. \tag{9}$$

An expression for $e^{\mathrm{wo}}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s})$ is given in [HSW04a], in which the functions $M_u(\mathbf{x}, \mathbf{t})$ are important. For each $k \in \{1, \dots, s\}$ let

$$M_k(x,t) := \begin{cases} 1 & \text{if } c_k \le t < x, \\ -1 & \text{if } x \le t < c_k, \\ 0 & \text{otherwise,} \end{cases}$$

and for each subset u of $\{1, \dots, s\}$ let

$$M_u(\mathbf{x}_u, \mathbf{t}_u) := \prod_{k\in u} M_k(x_k, t_k)$$

with the convention that $M_\emptyset \equiv 1$. It is shown in [HSW04c] that

$$f(\mathbf{x}) = \sum_{u\subseteq\{1,\dots,s\}} \int_{[0,1]^u} f'_{u,\mathbf{c}}(\mathbf{t}_u) M_u(\mathbf{x}_u, \mathbf{t}_u) d\mathbf{t}_u.$$

Define

$$h_u(\mathbf{t}_u) := I(M_u(\cdot, \mathbf{t}_u)) - Q_n(M_u(\cdot, \mathbf{t}_u)). \tag{10}$$

When $\mathbf{c} = \mathbf{1}$, then $h_u(\mathbf{t}_u)$ has the following expression:

$$h_u(\mathbf{t}_u) = \mathrm{vol}([\mathbf{0}, \mathbf{t}_u)) - \frac{1}{n}\sum_{j=0}^{n-1} 1_{[\mathbf{0},\mathbf{t}_u)}(\mathbf{x}_u^j). \tag{11}$$

For the case where $\mathbf{c}$ is in the interior of the unit cube $[0,1]^s$, $h_u(\mathbf{t}_u)$ has a similar expression replacing the cube $[\mathbf{0}, \mathbf{t}_u)$ by a certain box, which is described below. Note that the anchor $\mathbf{c} \in (0,1)^s$ partitions the unit cube $[0,1]^s$ into 2^s quadrants. Given a $\mathbf{t} = (t_1, \dots, t_s)^T$ in one of these quadrants, let $B(\mathbf{t}; \mathbf{c})$ denote the box with one corner at $\mathbf{t}$ and the opposite corner given by the unique vertex of $[0,1]^s$ that lies in the same quadrant as $\mathbf{t}$. Then $h_u(\mathbf{t}_u)$ can be expressed as

$$h_u(\mathbf{t}_u) = \mathrm{vol}(B_u(\mathbf{t}_u; \mathbf{c}_u)) - \frac{1}{n}\sum_{j=0}^{n-1} 1_{B_u(\mathbf{t}_u;\mathbf{c}_u)}(\mathbf{x}_u^j), \tag{12}$$

where $B_u(\mathbf{t}_u; \mathbf{c}_u)$ is the projection of $B(\mathbf{t}; \mathbf{c})$ onto the axes in u.

In terms of the functions h_u the worst-case error $e^{\rm wo}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s})$ in (9) is given by [HSW04a]

$$e^{\rm wo}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s}) = \begin{cases} \left(\sum_{u \neq \emptyset} \gamma_u^{q^*} \|h_u\|_{L_{p^*}}^{q^*} \right)^{1/q^*} & \text{for } q > 1, \\ \max_{u \neq \emptyset} \{\gamma_u \|h_u\|_{L_{p^*}}\} & \text{for } q = 1, \end{cases} \tag{13}$$

where p^* and q^* are the conjugates of p and q, respectively.

The result in the following lemma will be used several times in the proofs of the main theorems in this article.

LEMMA 2. *Let $\alpha, \beta, \theta > 0$ and $\tau \geq 1$. For the first n points of a Niederreiter $((T_u), s)$-sequence in base b, define*

$$\Phi(\alpha, \beta, \theta, \tau) := \sum_{\emptyset \neq u \subseteq 1,\ldots,s} \gamma_u^{\alpha} b^{\beta T_u} (\theta \ln(\tau n))^{|u|}. \tag{14}$$

If the γ_k satisfy

$$\sum_{k=1}^{\infty} \gamma_k^{\alpha} (k \ln k)^{\beta} < \infty, \tag{15}$$

then for any fixed $\epsilon > 0$ there exists a constant C_ϵ independent of s and n such that

$$\Phi(\alpha, \beta, \theta, \tau) \leq C_\epsilon n^{\epsilon}.$$

Proof. This lemma can be proved by an argument similar to that used in [Wan03, Theorem 4]. □

Now we can proceed to prove the strong tractability result of multivariate integration using Niederreiter sequences in the worst-case setting.

THEOREM 1. *Assume $p, q \in [1, \infty]$. Let $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ be the Banach space of functions f with norm* (8)*. Assume that the quasi-Monte Carlo quadrature Q_n uses a Niederreiter $((T_u), s)$-sequence in a prime power base b. If*

$$\sum_{k=1}^{\infty} \gamma_k^{a} k \ln k < \infty \tag{16}$$

for any $a \in [1, q^*]$, then the corresponding integration is strongly tractable in the worst-case setting, and for any fixed $\epsilon > 0$ there exists a constant C independent of s and n such that

$$e^{wo}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s}) \leq C n^{-1/a+\epsilon}. \tag{17}$$

Proof. It follows from the expression in (13) for the worst-case error and the fact $\|h_u\|_{L_{p^*}} \leq \|h_u\|_{L_\infty}$ for any $p^* \in [1, \infty]$ that it is sufficient to consider the case with $p^* = \infty$. For the case of $\mathbf{c} = \mathbf{1}$,

$$\|h_u\|_{L_\infty} = \sup_{\mathbf{t}_u \in [0,1]^u} \left| \mathrm{vol}([\mathbf{0}, \mathbf{t}_u)) - \frac{1}{n} \sum_{j=0}^{n-1} 1_{[\mathbf{0}, \mathbf{t}_u)}(\mathbf{x}_u^j) \right| = D^*_{u,\infty}(P)$$

due to the expression (11), where $D^*_{u,\infty}(P)$ is the local *star discrepancy* of P corresponding to the subset u. Note that if $\{\mathbf{x}^j\}_{j\geq 0}$ is a Niederreiter $((T_u),s)$-sequence in base b, then its projection onto the axes in u is a Niederreiter $(T_u,|u|)$-sequence in base b, where T_u is as given by (6). It follows from [Wan03, Lemma 1] that

$$D^*_{u,\infty}(P) \leq n^{-1} b^{T_u} (\theta \ln(bn))^{|u|},$$

where $\theta = b/\ln b$. Noting that $\|h_u\|_{L_\infty} = D^*_{u,\infty}(P) \leq 1$, we then have the following for $q \in [1,\infty]$ and any $\tilde{a} \in [1, q^*]$:

$$\begin{aligned}\left(\sum_u \gamma_u^{q^*} \|h_u\|^{q^*}_{L_\infty}\right)^{1/q^*} &\leq \left(\sum_u \gamma_u^{q^*} \|h_u\|^{\tilde{a}}_{L_\infty}\right)^{1/q^*} \leq \left(\sum_u \gamma_u^{q^*/\tilde{a}} \|h_u\|_{L_\infty}\right)^{\tilde{a}/q^*} \\ &\leq n^{-\tilde{a}/q^*} \left(\sum_u \gamma_u^{q^*/\tilde{a}} b^{T_u} (\theta \ln(bn))^{|u|}\right)^{\tilde{a}/q^*} \\ &= n^{-\tilde{a}/q^*} [\Phi(q^*/\tilde{a}, 1, \theta, b)]^{\tilde{a}/q^*}.\end{aligned}$$

Set $a = q^*/\tilde{a}$, and then $a \in [1, q^*]$ since $1 \leq \tilde{a} \leq q^*$. It follows from (13) that

$$e^{\mathrm{wo}}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s}) \leq n^{-1/a} [\Phi(a, 1, \theta, b)]^{1/a}.$$

Applying Lemma 2 to $\Phi(a,1,\theta,b)$ gives the upper bound (17) for $e^{\mathrm{wo}}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s})$ if condition (16) holds.

For the case with $\mathbf{c}$ in the interior of the unit cube $[0,1]^s$, from expression (12) we have

$$\begin{aligned}\|h_u\|_{L_\infty} &= \sup_{\mathbf{t}_u \in [0,1]^u} \left| \mathrm{vol}(B_u(\mathbf{t}_u; \mathbf{c}_u)) - \frac{1}{n} \sum_{j=0}^{n-1} 1_{B_u(\mathbf{t}_u;\mathbf{c}_u)}(\mathbf{x}^j_u) \right| \\ &\leq \sup_{J_u} \left| \mathrm{vol}(J_u) - \frac{1}{n} \sum_{j=0}^{n-1} 1_{J_u}(\mathbf{x}^j_u) \right| = D_{u,\infty}(P),\end{aligned}$$

where J_u denotes subintervals of $[0,1]^s$ of the form $\prod_{k\in u}[\alpha_k, \beta_k)$, and $D_{u,\infty}(P)$ is the local *extreme discrepancy* (or *unanchored discrepancy*) of P corresponding to the subset u. From Proposition 2.4 in [Nie92] we have

$$D_{u,\infty}(P) \leq 2^{|u|} D^*_{u,\infty}(P).$$

Then the upper bound for $e^{\mathrm{wo}}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s})$ follows from the previous argument for $\mathbf{c} = \mathbf{1}$.

Note that the initial error in multivariate integration in the space $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ is

$$e^{\mathrm{wo}}(0, \mathcal{F}^{(1)}_{p,q,\gamma,s}) = \left(\sum_{u \subseteq \{1,\ldots,s\}} \gamma_u^{q^*} \|\mathrm{vol}(B_u(\cdot; \mathbf{c}_u))\|^{q^*}_{L_{p^*}} \right)^{1/q^*}.$$

By the definition of $B_u(\mathbf{t}_u; \mathbf{c}_u)$ it can be verified that

$$\begin{aligned}\|\mathrm{vol}(B_u(\cdot;\mathbf{c}_u))\|_{L_{p^*}}^{p^*} &= \sum_{v\subseteq u}\left(\prod_{k\in v}\int_0^{c_k} t_k^{p^*}\,dt_k\right)\left(\prod_{k\in u-v}\int_{c_k}^1 (1-t_k)^{p^*}\,dt_k\right)\\ &= (p^*+1)^{-|u|}\sum_{v\subseteq u}\left(\prod_{k\in v} c_k^{p^*+1}\right)\left(\prod_{k\in u-v}(1-c_k)^{p^*+1}\right)\\ &= (p^*+1)^{-|u|}\prod_{k\in u}\left[c_k^{p^*+1}+(1-c_k)^{p^*+1}\right].\end{aligned}$$

It follows that

$$\begin{aligned}e^{\mathrm{wo}}(0,\mathcal{F}^{(1)}_{p,q,\gamma,s}) &= \left[\sum_{u\subseteq\{1,\dots,s\}}\gamma_u^{q^*}(p^*+1)^{-|u|q^*/p^*}\prod_{k\in u}\left(c_k^{p^*+1}+(1-c_k)^{p^*+1}\right)^{q^*/p^*}\right]^{1/q^*}\\ &= \prod_{k=1}^{s}\left[1+\gamma_k^{q^*}(p^*+1)^{-q^*/p^*}\left(c_k^{p^*+1}+(1-c_k)^{p^*+1}\right)^{q^*/p^*}\right]^{1/q^*},\end{aligned}$$

which is uniformly bounded in s under condition (16). From this factor and the upper bound for the worst-case error above, we assert that the multivariate integration is strongly tractable in the worst-case setting. This concludes the proof. □

Remark 1. From Theorem 1 the following facts are observed:

(i) If condition (16) holds for $a = 1$, then the worst-case error is $O(n^{-1+\epsilon})$, and in the case with $p = q = 2$, this result is the same as that obtained in [Wan03] for reproducing kernel Hilbert spaces. Therefore, our result in Theorem 1 is an extension of that in [Wan03] to arbitrary p and q.

(ii) Although the convergence rate for $a > 1$ is smaller than that for $a = 1$, by introducing the number $a \in (1, q^*]$, the sequence of weights $\{\gamma_k\}_k$ is transformed at the same time and the strong tractability condition is weaker. Therefore, when considering the question of strong tractability, this theorem allows one to accept a lower convergence rate in turn for less restrictive conditions on the weights. This was done in [HSW04b].

3. Tractability in randomized settings for the classical problem. In this section we consider the strong tractability problem of integration using randomly scrambled Niederreiter digital net rules in randomized settings for the classical problem. The function spaces $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ defined below have slightly different norms than $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ considered in the previous section although the smoothness assumptions are the same. The reason for this is that the arguments for the randomized setting are different from that used in the previous section. In analyzing the randomized error, the mean square error of the quadrature rule is expressed in terms of Fourier coefficients of the integrand under the orthogonormal system of multivariate Haar wavelets. By making use of integration by parts, each of the Fourier coefficients is expressed by an integral of the product of two functions: one is the unanchored mixed first derivative of the integral, and another is defined via the Haar wavelet.

For $p \in [1,\infty]$ let $\mathcal{H}_p^s$ be defined as in the previous section. Given an additional parameter $q \in [1,\infty]$ and a sequence $\gamma = \{\gamma_k\}_k$ of positive numbers, $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ is defined

to be the completion of $\mathcal{H}_p^s$ with respect to the norm

$$\|f\|_{p,q,\gamma,s} := \begin{cases} \left(\sum_{u \subseteq \{1,\ldots,s\}} \gamma_u^{-q} \|f'_u\|_{L_p}^q \right)^{1/q} & \text{for } q < \infty, \\ \max_{u \subseteq \{1,\ldots,s\}} \{\gamma_u^{-1} \|f'_u\|_{L_p}\} & \text{for } q = \infty, \end{cases} \tag{18}$$

where $\gamma_\emptyset = 1$, $\gamma_u = \prod_{k \in u} \gamma_k$, the derivative $f'_u(\mathbf{x})$ for $u \neq \emptyset$ is defined by

$$f'_u(\mathbf{x}) := \left(\prod_{k \in u} \frac{\partial}{\partial x_k} \right) f(\mathbf{x}),$$

and $f'_\emptyset(\mathbf{x})$ denotes $f(\mathbf{x})$. For $p = q = 2$ the space becomes the Sobolev space considered in section 3.1 of [SW01].

For the weighted Banach spaces defined above, we define the randomized error of the quadrature rule Q_n as

$$e^{\mathrm{ra}}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s}) := \sup_{\|f\|_{p,q,\gamma,s} \leq 1} \sqrt{E|I(f) - Q_n(f)|^2}, \tag{19}$$

where the expectation E is taken with respect to the random samples.

In what follows, we first give some background about the randomly scrambled digital sequences and Haar wavelets. Then we deal with the randomized error of the quadrature rules.

3.1. Randomly scrambled digital sequences. The sequence $\{\mathbf{x}^j\}_{j \geq 0}$ of points in the unit cube $[0,1]^s$ is generated in the following way, which is called scrambling [Owe95, Owe00]. Let $b \geq 2$ be a prime power base. The kth component of the jth point $\mathbf{x}^j = (x_1^j, \ldots, x_s^j)^T$ is determined by the b-ary expression

$$x_k^j = \frac{x_{jk1}}{b} + \frac{x_{jk2}}{b^2} + \cdots.$$

Here, the digits x_{jkl} are generated by

$$\begin{pmatrix} x_{jk1} \\ x_{jk2} \\ \vdots \end{pmatrix} = \boldsymbol{L}_k \boldsymbol{C}_k \begin{pmatrix} j_1 \\ j_2 \\ \vdots \end{pmatrix} + \mathbf{e}_k \bmod b, \quad k = 1, \ldots, s,$$

where $(j_1, j_2, \ldots)^T$ is the vector of b-ary digits of $j \in Z_+$, i.e., $j = j_1 + j_2 b + j_3 b^2 + \cdots$, the $\boldsymbol{C}_k$ are the prescribed $\infty \times \infty$ generator matrices, the $\boldsymbol{L}_k$ are lower triangular $\infty \times \infty$ scrambling matrices, and the $\mathbf{e}_k$ are $\infty \times 1$ digital shifts. The scrambling matrices and shifts are chosen randomly.

Choosing the first $n = b^m$, $m \in Z_+$, points of a digital sequence in base b gives a digital net in base b. The quality of the digital net is defined below.

For a vector $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots)^T \in Z_+^\infty$ with $\|\boldsymbol{\ell}\|_1 := \sum_\alpha \ell_\alpha < \infty$, let $\boldsymbol{C}(\boldsymbol{\ell})$ be the $\|\boldsymbol{\ell}\|_1 \times \infty$ matrix formed by the first ℓ_1 rows of $\boldsymbol{C}_1$ followed by the first ℓ_2 rows of $\boldsymbol{C}_2$, etc. For any integer $m \in Z_+$ let $\boldsymbol{C}(\boldsymbol{\ell}, m)$ denote the $\|\boldsymbol{\ell}\|_1 \times m$ matrix formed by the first m columns of $\boldsymbol{C}(\boldsymbol{\ell})$. Define

$$T_u := \min\{T \geq 0 : \ \mathrm{rank}(\boldsymbol{C}(\boldsymbol{\ell}, m)) = \|\boldsymbol{\ell}\|_1 \ \forall m, \forall \boldsymbol{\ell} \text{ with } U(\boldsymbol{\ell}) \subseteq u, \|\boldsymbol{\ell}\|_1 = m - T\}.$$

Smaller values of T_u correspond to better nets. Note that the quality measure introduced here is the same as the T_u in (6). See [Owe95, Owe00] for a more detailed explanation of the scrambled scheme, and [Nie92, Chapter 4; NX01, Chapter 8] for constructions of digital nets and sequences.

3.2. Haar wavelets. To deal with the randomized error of the randomly scrambled digital net rules, we use Haar wavelets. Multidimensional Haar wavelets are tensor products of one-dimensional wavelets. We refer to [Wal02, Chapter 5] for one-dimensional Haar wavelets and to [Ent97, Ent98] for earlier work on Haar wavelets and nets.

We first introduce some notation. For any $\nu \in Z_+$ define the function

$$\lg(\nu) := \begin{cases} \lfloor \log_b(\nu) \rfloor + 1 & \text{if } \nu > 0, \\ 0 & \text{if } \nu = 0, \end{cases}$$

which means that the base b representation of ν has $\lg(\nu)$ digits if one ignores leading zeros. Let $\tilde{\nu}$ denote the leading digit of ν when written in base b, and define

$$z_\nu := \nu b^{1-\lg(\nu)} - \tilde{\nu}.$$

For any $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_s)^T \in Z_+^s$ define

$$\lg(\boldsymbol{\nu}) := (\lg(\nu_1), \ldots, \lg(\nu_s))^T, \quad \mathbf{z}_{\boldsymbol{\nu}} := (z_{\nu_1}, \ldots, z_{\nu_s})^T.$$

Also, let $U(\boldsymbol{\nu})$ denote the set of all k for which $\nu_k > 0$ and let $|U(\boldsymbol{\nu})|$ denote the cardinality of $U(\boldsymbol{\nu})$. Moreover, for $u \subseteq \{1, \ldots, s\}$ let $\mathbf{1}_u$ denote the s-dimensional vector whose kth component is 1 for $k \in u$ and 0 otherwise. For any $\mathbf{x} = (x_1, \ldots, x_s)^T, \mathbf{y} = (y_1, \ldots, y_s)^T \in [0,1]^s$, and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_s)^T \in Z_+^s$, let $\delta(\mathbf{x}, \mathbf{y}, \boldsymbol{\ell}) = 1$ if the first ℓ_k digits of x_k and y_k are the same for all $k = 1, \ldots, s$, and let $\delta(\mathbf{x}, \mathbf{y}, \boldsymbol{\ell}) = 0$ otherwise.

Multidimensional Haar wavelets $\psi_{\boldsymbol{\nu}}(\mathbf{x})$ for $\boldsymbol{\nu} \in Z_+^s$ are piecewise constant functions, which are defined as

$$\psi_{\boldsymbol{\nu}}(\mathbf{x}) := b^{(\|\lg(\boldsymbol{\nu})\|_1 - |U(\boldsymbol{\nu})|)/2} \exp\left(\frac{2\pi i}{b} \sum_{k=1}^{s} \tilde{\nu}_k x_{\bullet k \lg(\nu_k)}\right) \delta(\mathbf{x}, \mathbf{z}_{\boldsymbol{\nu}}, \lg(\boldsymbol{\nu}) - \mathbf{1}_{U(\boldsymbol{\nu})}), \tag{20}$$

where $x_{\bullet k\ell}$ denotes the ℓth b-ary digit of the kth component of $\mathbf{x}$, and $i := \sqrt{-1}$. Note that the support of $\psi_{\boldsymbol{\nu}}$ is a box, $S(\boldsymbol{\nu})$, of volume $b^{|U(\boldsymbol{\nu})| - \|\lg(\boldsymbol{\nu})\|_1}$. In fact,

$$S(\boldsymbol{\nu}) = \prod_{k \in U(\boldsymbol{\nu})} \left[z_{\nu_k}, z_{\nu_k} + b^{1-\lg(\nu_k)} \right) \times [0,1)^{\{1,\ldots,s\} \setminus U(\boldsymbol{\nu})}. \tag{21}$$

It is known that $\{\psi_{\boldsymbol{\nu}}(\mathbf{x})\}_{\boldsymbol{\nu}}$ is a sequence of complex-valued, integrable, orthogonormal basis functions. Any $f \in L_2([0,1]^s)$ can be represented as an infinite series

$$f(\mathbf{x}) = \sum_{\boldsymbol{\nu}} F(\boldsymbol{\nu}) \psi_{\boldsymbol{\nu}}(\mathbf{x}), \quad \mathbf{x} \in [0,1]^s, \tag{22}$$

where the $F(\boldsymbol{\nu})$ are Fourier coefficients given by

$$F(\boldsymbol{\nu}) := \int_{[0,1]^s} f(\mathbf{x}) \overline{\psi_{\boldsymbol{\nu}}(\mathbf{x})} d\mathbf{x}; \tag{23}$$

here $\overline{\psi_{\boldsymbol{\nu}}(\mathbf{x})}$ denotes the complex conjugate of $\psi_{\boldsymbol{\nu}}(\mathbf{x})$. The following facts will play an important role in our randomized error analysis. For each $\nu \in Z_+$ and any $x \in [0,1]$ let

$$\xi_\nu(x) := \int_0^x \overline{\psi_\nu(t)}dt.$$

For each $\boldsymbol{\nu} \in Z_+^s$ and any $\mathbf{x} \in [0,1]^s$ let

$$\xi_{\boldsymbol{\nu}}(\mathbf{x}) := \prod_{k \in U(\boldsymbol{\nu})} \xi_{\nu_k}(x_k). \tag{24}$$

Note that the support of $\xi_{\boldsymbol{\nu}}$ is the same as that of $\psi_{\boldsymbol{\nu}}$.

The following upper bounds on the different norms of $\xi_{\boldsymbol{\nu}}$ can be verified by immediate calculations, which are omitted here.

LEMMA 3. *For each* $\boldsymbol{\nu} \in Z_+^s$,

$$\begin{aligned}
&\|\xi_{\boldsymbol{\nu}}\|_{L_1} \le 2^{-|U(\boldsymbol{\nu})|} b^{-3(\|lg(\boldsymbol{\nu})\|_1 - |U(\boldsymbol{\nu})|)/2},\\
&\|\xi_{\boldsymbol{\nu}}\|_{L_2} \le 3^{-|U(\boldsymbol{\nu})|/2} b^{-\|lg(\boldsymbol{\nu})\|_1 + |U(\boldsymbol{\nu})|},\\
&\|\xi_{\boldsymbol{\nu}}\|_{L_\infty} \le b^{-(\|lg(\boldsymbol{\nu})\|_1 - |U(\boldsymbol{\nu})|)/2}.
\end{aligned}$$

3.3. Upper bounds for the randomized error. This subsection will give upper bounds for the randomized error defined in (19) for different values of $p \in [1,\infty]$, and find sufficient conditions under which multivariate integration using the randomly scrambled Niederreiter digital nets is strongly tractable in the randomized setting.

LEMMA 4. *Let* $\psi_{\boldsymbol{\nu}}(\mathbf{x})$, $\boldsymbol{\nu} \in Z_+^s$, *be the Haar wavelets defined by* (20), *and let* Q_n *be the quasi-Monte Carlo quadrature that uses randomly scrambled digital nets with* $n = b^m$, $m \in Z_+$. *By* $\mathrm{MSE}(Q_n, f)$ *denote the mean square error of the approximation* Q_n, *i.e.,*

$$\mathrm{MSE}(Q_n, f) = E|I(f) - Q_n(f)|^2.$$

Then for any $f \in L_2([0,1]^s)$

$$\begin{aligned}
\mathrm{MSE}(Q_n, f) &= \sum_{\boldsymbol{\nu} \ne \mathbf{0}} |F(\boldsymbol{\nu})|^2 E[Q_n(\psi_{\boldsymbol{\nu}}) Q_n(\overline{\psi_{\boldsymbol{\nu}}})]\\
&\le \sum_{\boldsymbol{\nu}: \|lg(\boldsymbol{\nu})\|_1 + T_U(\boldsymbol{\nu}) > m} |F(\boldsymbol{\nu})|^2 3^{|U(\boldsymbol{\nu})|} b^{T_{U(\nu)} - m},
\end{aligned} \tag{25}$$

where the $F(\boldsymbol{\nu})$ *are the coefficients of* f *under the Haar wavelets.*

Proof. Making use of the expansion in (22) and noting that $F(\mathbf{0}) = I(f)$ yields

$$I(f) - Q_n(f) = -\frac{1}{n} \sum_{j=0}^{n-1} \sum_{\boldsymbol{\nu} \ne \mathbf{0}} F(\boldsymbol{\nu}) \psi_{\boldsymbol{\nu}}(\mathbf{x}^j) = -\sum_{\boldsymbol{\nu} \ne \mathbf{0}} F(\boldsymbol{\nu}) Q_n(\psi_{\boldsymbol{\nu}}),$$

and then

$$|I(f) - Q_n(f)|^2 = \sum_{\boldsymbol{\nu} \ne \mathbf{0}} \sum_{\boldsymbol{\omega} \ne \mathbf{0}} F(\boldsymbol{\nu}) \overline{F(\boldsymbol{\omega})} Q_n(\psi_{\boldsymbol{\nu}}) Q_n(\overline{\psi_{\boldsymbol{\omega}}}).$$

It is proved in [HD04, Lemma 10] that for scrambled digital nets

$$\begin{aligned} E[Q_n(\psi_{\boldsymbol{\nu}})Q_n(\overline{\psi_{\boldsymbol{\omega}}})] &= 0 \quad \text{for } \boldsymbol{\nu} \neq \boldsymbol{\omega}, \\ E[Q_n(\psi_{\boldsymbol{\nu}})Q_n(\overline{\psi_{\boldsymbol{\nu}}})] &= 0 \quad \text{for } \|\mathrm{lg}(\boldsymbol{\nu})\|_1 + T_{U(\boldsymbol{\nu})} \le m, \\ E[Q_n(\psi_{\boldsymbol{\nu}})Q_n(\overline{\psi_{\boldsymbol{\nu}}})] &\le 3^{|U(\boldsymbol{\nu})|} b^{T_{U(\boldsymbol{\nu})}-m} \quad \text{for } \|\mathrm{lg}(\boldsymbol{\nu})\|_1 + T_{U(\boldsymbol{\nu})} > m. \end{aligned}$$

The results in (25) then follow immediately from the facts mentioned above. ∎

For $f \in \mathcal{F}^{(2)}_{p,q,\gamma,s}$, by making use of integration by parts we can express the Fourier coefficients $F(\boldsymbol{\nu})$ in terms of the functions $\xi_{\boldsymbol{\nu}}$ defined in (24) as follows:

$$F(\boldsymbol{\nu}) = (-1)^{|U(\boldsymbol{\nu})|} \int_{S(\boldsymbol{\nu})} \xi_{\boldsymbol{\nu}}(\mathbf{x}) f'_{U(\boldsymbol{\nu})}(\mathbf{x}) d\mathbf{x}, \tag{26}$$

where $S(\boldsymbol{\nu})$ is the support of $\xi_{\boldsymbol{\nu}}$ given in (21). By Hölder's inequality applied to (26),

$$|F(\boldsymbol{\nu})| \le \|f'_{U(\boldsymbol{\nu})}\|_{L_p,S(\boldsymbol{\nu})} \|\xi_{\boldsymbol{\nu}}\|_{L_{p^*}}, \tag{27}$$

where

$$\|f'_{U(\boldsymbol{\nu})}\|_{L_p,S(\boldsymbol{\nu})} := \left(\int_{S(\boldsymbol{\nu})} |f'_{U(\boldsymbol{\nu})}(\mathbf{x})|^p d\mathbf{x}\right)^{1/p}, \quad \|\xi_{\boldsymbol{\nu}}\|_{L_{p^*}} := \left(\int_{[0,1]^s} |\xi_{\boldsymbol{\nu}}(\mathbf{x})|^{p^*} d\mathbf{x}\right)^{1/p^*}.$$

It follows from (25) that

$$\begin{aligned} \mathrm{MSE}(Q_n, f) &\le n^{-1} \sum_{\boldsymbol{\nu}:\|\mathrm{lg}(\boldsymbol{\nu})\|_1 + T_{U(\boldsymbol{\nu})} > m} \|f'_{U(\boldsymbol{\nu})}\|^2_{L_p,S(\boldsymbol{\nu})} \|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}} 3^{|U(\boldsymbol{\nu})|} b^{T_{U(\boldsymbol{\nu})}} \\ &= n^{-1} \sum_{u \neq \emptyset} \sum_{\substack{\boldsymbol{\nu}:\, U(\boldsymbol{\nu})=u, \\ \|\mathrm{lg}(\boldsymbol{\nu})\|_1 + T_u > m}} 3^{|u|} b^{T_u} \|f'_u\|^2_{L_p,S(\boldsymbol{\nu})} \|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}}. \end{aligned} \tag{28}$$

For simplicity in notation, we define $\mathcal{L}_{u,m}$ as the set of s-dimensional vector $\boldsymbol{\ell}$ with integer components $\ell_k > 0$ for $k \in u$, $\ell_k = 0$ for $k \in \bar{u}$, and $\|\boldsymbol{\ell}\|_1 > m - T_u$, and for each $\boldsymbol{\ell}$ define $\mathcal{N}_{u,\boldsymbol{\ell}}$ as the set of s-dimensional integer vector $\boldsymbol{\nu}$ with $U(\boldsymbol{\nu}) = u$ and $\mathrm{lg}(\boldsymbol{\nu}) = \boldsymbol{\ell}$, i.e.,

$$\mathcal{L}_{u,m} := \{\boldsymbol{\ell} = (\ell_1, \ldots, \ell_s)^T : \|\mathrm{lg}(\boldsymbol{\ell})\|_1 > m - T_u,\, \ell_k > 0\, \forall k \in u,\, \ell_k = 0\, \forall k \in \bar{u}\}, \tag{29}$$

$$\mathcal{N}_{u,\boldsymbol{\ell}} := \{\boldsymbol{\nu} = (\nu_1, \ldots, \nu_s)^T : U(\boldsymbol{\nu}) = u,\, \mathrm{lg}(\boldsymbol{\nu}) = \boldsymbol{\ell}\}. \tag{30}$$

The inner sum in the last expression in (28) can be written as follows:

$$\Psi_{u,m,p} := \sum_{\boldsymbol{\ell} \in \mathcal{L}_{u,m}} \sum_{\boldsymbol{\nu} \in \mathcal{N}_{u,\boldsymbol{\ell}}} 3^{|u|} b^{T_u} \|f'_u\|^2_{L_p,S(\boldsymbol{\nu})} \|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}}. \tag{31}$$

Then (28) becomes

$$\mathrm{MSE}(Q_n, f) \le n^{-1} \sum_{u \neq \emptyset} \Psi_{u,m,p}. \tag{32}$$

Before stating the main results for the randomized strong tractability, we give the following lemma that will be used in the proof of Theorem 2. This lemma can be proved by the binomial theory.

LEMMA 5. *Let m, t, r, and b be integers with $m \geq t \geq 0$, $r \geq 1$, and $b \geq 2$. Then for $\eta \in (0, \infty)$*

$$\Omega(\eta, m, t, r) := \sum_{l=m-t+1}^{\infty} \binom{l-1}{r-1} b^{-\eta l} < \left(\frac{b^{\eta} m}{b^{\eta}-1}\right)^{r} b^{-\eta(m-t)}. \tag{33}$$

THEOREM 2. *Let $p, q \in [1, \infty]$. Let $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ be the Banach space of functions with norm* (18). *Let Q_n be the quasi-Monte Carlo quadrature that uses randomly scrambled Niederreiter digital $((T_u), m, s)$-nets in prime power base b.*

(i) *For $p \in [1, \infty]$, if*

$$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^2 < \infty, \tag{34}$$

then the corresponding integration is strongly tractable in the randomized setting, and for any fixed $\epsilon > 0$, there exists a constant C independent of s and n such that

$$e^{ra}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s}) \leq C n^{-1+\epsilon}. \tag{35}$$

(ii) *In particular, for $p \in [2, \infty]$, if*

$$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^3 < \infty, \tag{36}$$

then the corresponding integration is strongly tractable in the randomized setting, and for any fixed $\epsilon > 0$, there exists a constant C independent of s and n such that

$$e^{ra}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s}) \leq C n^{-\frac{3}{2}+\epsilon}. \tag{37}$$

Proof. For item (i), we first note that $\mathcal{F}^{(2)}_{p,q,\gamma,s} \subseteq \mathcal{F}^{(2)}_{2,q,\gamma,s}$ for any $p > 2$, and then

$$e^{\mathrm{ra}}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s}) \leq e^{\mathrm{ra}}(Q_n, \mathcal{F}^{(2)}_{2,q,\gamma,s}) \quad \forall p > 2. \tag{38}$$

Hence it is sufficient to consider the case $p \in [1, 2]$. In this case we have $p^* \in (1, \infty]$ and

$$\|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}} \leq \|\xi_{\boldsymbol{\nu}}\|^2_{L_\infty} \leq b^{-\|\mathrm{lg}(\boldsymbol{\nu})\|_1 + |U(\boldsymbol{\nu})|}$$

by Lemma 3. Then by the definition (31) of $\Psi_{u,m,p}$,

$$\begin{aligned} \Psi_{u,m,p} &\leq \sum_{\boldsymbol{\ell} \in \mathcal{L}_{u,m}} \sum_{\boldsymbol{\nu} \in \mathcal{N}_{u,\boldsymbol{\ell}}} (3b)^{|u|} b^{T_u - \|\boldsymbol{\ell}\|_1} \|f'_u\|^2_{L_p, S(\boldsymbol{\nu})} \|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}} \\ &\leq (3b)^{|u|} b^{T_u} \sum_{\boldsymbol{\ell} \in \mathcal{L}_{u,m}} b^{-\|\boldsymbol{\ell}\|_1} \left(\sum_{\boldsymbol{\nu} \in \mathcal{N}_{u,\boldsymbol{\ell}}} \int_{S(\boldsymbol{\nu})} |f'_u(\mathbf{x})|^p d\mathbf{x} \right)^{\frac{2}{p}}, \end{aligned} \tag{39}$$

where the last inequality holds due to $\frac{2}{p} \geq 1$. By the definition of the support $S(\boldsymbol{\nu})$ in (21), for a fixed $|s|$-dimensional vector $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_s)^T$ with $\ell_k > 0$ for $k \in u$ and $\ell_k = 0$ for $k \in \bar{u}$ we have

$$\sum_{\boldsymbol{\nu} \in \mathcal{N}_{u,\boldsymbol{\ell}}} \int_{S(\boldsymbol{\nu})} |f'_u(\mathbf{x})|^p d\mathbf{x} = (b-1)^{|u|} \int_{[0,1]^s} |f'_u(\mathbf{x})|^p d\mathbf{x} = (b-1)^{|u|} \|f'_u\|^p_{L_p}. \tag{40}$$

Moreover, note that for a given positive integer l with $l > m - T_u$ there is a total number $\binom{l-1}{|u|-1}$ of the vectors $\boldsymbol{\ell}$ in $\mathcal{L}_{u,\boldsymbol{\ell}}$ such that $\|\boldsymbol{\ell}\|_1 = l$. Therefore, (39) becomes

$$\begin{aligned}\Psi_{u,m,p} &\le \|f'_u\|^2_{L_p}[3b(b-1)]^{|u|}b^{T_u}\sum_{l=m-T_u+1}^{\infty}\binom{l-1}{|u|-1}b^{-\frac{p}{2}l} \\ &= \|f'_u\|^2_{L_p}[3b(b-1)]^{|u|}b^{T_u}\Omega(1,m,T_u,|u|),\end{aligned} \tag{41}$$

where Ω is as defined in Lemma 5. Making use of the inequality in Lemma 5 and the fact $m = \log_b n = \ln n/\ln b$ in (41) yields

$$\Psi_{u,m,p} \le n^{-1}\|f'_u\|^2_{L_p}b^{2T_u}(\theta_1 \ln n)^{|u|}, \tag{42}$$

where $\theta_1 = 3b^2/\ln b$. Applying this inequality to (32) we have

$$\mathrm{MSE}(Q_n, f) \le n^{-2}\sum_{u\ne\emptyset}\|f'_u\|^2_{L_p}b^{2T_u}(\theta_1 \ln n)^{|u|}.$$

Therefore, for $q \in [1,\infty]$, by Hölder's inequality and the definition of norm in (18) we have the following:

$$\begin{aligned}\mathrm{MSE}(Q_n, f) &\le n^{-2}\left[\sum_{u\ne\emptyset}\left(\gamma_u^{-2}\|f'_u\|^2_{L_p}\right)^q\right]^{\frac{1}{q}}\left[\sum_{u\ne\emptyset}\left(\gamma_u^2 b^{2T_u}(\theta_1\ln n)^{|u|}\right)^{q^*}\right]^{\frac{1}{q^*}} \\ &\le n^{-2}\|f\|^2_{p,q,\gamma,s}\left[\sum_{u\ne\emptyset}\left(\gamma_u^2 b^{2T_u}(\theta_1\ln n)^{|u|}\right)^{q^*}\right]^{\frac{1}{q^*}} \\ &\le n^{-2}\|f\|^2_{p,q,\gamma,s}\sum_{u\ne\emptyset}\gamma_u^2 b^{2T_u}(\theta_1\ln n)^{|u|} \\ &= n^{-2}\|f\|^2_{p,q,\gamma,s}\Phi(2,2,\theta_1,1),\end{aligned}$$

where Φ is defined by (14). It follows from Lemma 2 that for any fixed $\epsilon > 0$, there exists a constant C independent of s and n such that

$$e^{\mathrm{ra}}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s}) = \sup_{\|f\|_{p,q,\gamma,s}\le 1}\sqrt{\mathrm{MSE}(Q_n,f)} \le Cn^{-1+\epsilon}$$

under condition (34).

We now consider the initial error in the space $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ with $p \in [1,2]$. Note from (27) and Lemma 3 that

$$\begin{aligned}|F(0)|^2 &\le \sum_{\boldsymbol{\nu}}|F(\boldsymbol{\nu})|^2 \le \sum_u\sum_{\boldsymbol{\nu}:U(\boldsymbol{\nu})=u}\|f'_u\|^2_{L_p,S(\boldsymbol{\nu})}\|\xi_{\boldsymbol{\nu}}\|^2_{L_{p^*}} \\ &\le \sum_u\sum_{\boldsymbol{\ell}_u}\sum_{\boldsymbol{\nu}:\lg(\boldsymbol{\nu})=\boldsymbol{\ell}_u} b^{|u|-\|\boldsymbol{\ell}_u\|_1}\|f'_u\|^2_{L_p,S(\boldsymbol{\nu})} \\ &\le \sum_u\sum_{\boldsymbol{\ell}_u} b^{|u|-\|\boldsymbol{\ell}_u\|_1}\left(\sum_{\boldsymbol{\nu}:\lg(\boldsymbol{\nu})=\boldsymbol{\ell}_u}\int_{S(\boldsymbol{\nu})}|f'_u(\mathbf{x})|^p d\mathbf{x}\right)^{\frac{2}{p}} \\ &= \sum_u\sum_{\boldsymbol{\ell}_u}[b(b-1)]^{|u|}b^{-\|\boldsymbol{\ell}_u\|_1}\|f'_u\|^2_{L_p},\end{aligned}$$

where $\boldsymbol{\ell}_u$ denotes the s-dimensional integer vector with components $\ell_k > 0$ for $k \in u$ and $\ell_k = 0$ for $k \in \bar{u}$. Because

$$\sum_{\boldsymbol{\ell}_u} b^{-\|\boldsymbol{\ell}_u\|_1} = \sum_{l=|u|}^{\infty} \binom{l-1}{|u|-1} b^{-l} = (b-1)^{-|u|},$$

we then have

$$\begin{aligned}
|F(0)|^2 &\le \sum_u \|f_u'\|_{L_p}^2 b^{|u|} \\
&\le \left[\sum_u \left(\gamma_u^{-2}\|f_u'\|_{L_p}^2\right)^q\right]^{\frac{1}{q}} \left[\sum_u \left(\gamma_u^2 b^{2|u|}\right)^{q^*}\right]^{\frac{1}{q^*}} \\
&\le \|f\|_{p,q,\gamma,s}^2 \prod_{k=1}^{s} \left(1 + (b\gamma_k)^{2q^*}\right)^{\frac{1}{q^*}}.
\end{aligned}$$

Therefore,

$$e^{\mathrm{wo}}(0, \mathcal{F}_{p,q,\gamma,s}^{(1)}) = \sup_{\|f\|_{p,q,\gamma,s}\le 1} |F(0)| \le \prod_{k=1}^{s} \left(1 + (b\gamma_k)^{2q^*}\right)^{\frac{1}{q^*}},$$

which is uniformly bounded in s under condition (34). It follows that the multivariate integration is strongly tractable in the randomized setting.

As for item (ii), it is also sufficient to consider the case $p = 2$ due to (38). For $p = 2$ ($p^* = 2$), from (31), (40), and Lemma 3 we have

$$\begin{aligned}
\Psi_{u,m,2} &\le b^{2|u|+T_u} \sum_{\boldsymbol{\ell}\in\mathcal{L}_{u,m}} \sum_{\boldsymbol{\nu}\in\mathcal{N}_{u,\boldsymbol{\ell}}} b^{-2\|\boldsymbol{\ell}\|_1} \|f_u'\|_{L_2,S(\boldsymbol{\nu})}^2 \\
&= b^{2|u|+T_u} \sum_{\boldsymbol{\ell}\in\mathcal{L}_{u,m}} b^{-2\|\boldsymbol{\ell}\|_1} (b-1)^{|u|} \|f_u'\|_{L_2}^2 \\
&= \|f_u'\|_{L_2}^2 [b^2(b-1)]^{|u|} b^{T_u} \sum_{l=m-T_u+1}^{\infty} \binom{l-1}{|u|-1} b^{-2l} \\
&= \|f_u'\|_{L_2}^2 [b^2(b-1)]^{|u|} b^{T_u} \Omega(2, m, T_u, |u|).
\end{aligned}$$

From Lemma 5 we then have

$$\Psi_{u,m,2} \le n^{-2} \|f_u'\|_{L_2}^2 (\theta_2 \ln n)^{|u|} b^{3T_u},$$

where $\theta_2 = b^4/[(b+1)\ln b]$. Therefore, from (32),

$$\mathrm{MSE}(Q_n, f) \le n^{-3} \sum_{u\ne\emptyset} \|f_u'\|_{L_2}^2 b^{3T_u} (\theta_2 \ln n)^{|u|}.$$

For $q \in [1, \infty]$, by Hölder's inequality

$$\begin{aligned}
\mathrm{MSE}(Q_n, f) &\le n^{-3} \|f\|_{2,q,\gamma,s}^2 \left[\sum_{u\ne\emptyset} \left(\gamma_u^2 b^{3T_u} (\theta_2 \ln n)^{|u|}\right)^{q^*}\right]^{\frac{1}{q^*}} \\
&\le n^{-3} \|f\|_{2,q,\gamma,s}^2 \sum_{u\ne\emptyset} \gamma_u^2 b^{3T_u} (\theta_2 \ln n)^{|u|} \\
&= n^{-3} \|f\|_{2,q,\gamma,s}^2 \Phi(2, 3, \theta_2, 1).
\end{aligned}$$

Applying Lemma 2 to $\Phi(2,3,\theta_2,1)$ gives the desired result for the case $p=2$ and then for $p\in(2,\infty]$. This completes the proof for the theorem. □

Remark 2. For different values of p in $[2,\infty]$, we have two kinds of sufficient conditions for strong tractability in the randomized setting. The convergence rates are different under these two conditions. The condition under which the rate is higher is a little bit stronger.

Remark 3. The results in Theorem 2 are just for randomly scrambled Niederreiter digital nets but not for sequences. The difficulty for randomly scrambled sequences is the calculation of the expectation $E[Q_n(\psi_{\boldsymbol{\nu}})Q_n(\overline{\psi_{\boldsymbol{\nu}}})]$.

4. Weighted integration over a general domain. In this section, we extend the results presented in the previous two sections to the weighted integration problem over a general domain in (3), i.e., $D=\overline{(a_1,b_1)}\times\cdots\times\overline{(a_s,b_s)}$ and the general weight function ρ defined in (4).

4.1. Worst-case error analysis. The general problem for the worst case of integration can be reduced to the classical problem with domain $[0,1]^s$ and uniform weight $\rho\equiv 1$. Specifically, we define the following transformations:

$$W_k(x):=\int_{a_k}^{x}\rho_k(z)dz,\quad k=1,\dots,s. \tag{43}$$

Then $W_k:\overline{(a_k,b_k)}\to[0,1]$ is onto and increasing. Define

$$\boldsymbol{W}(\mathbf{x})=(W_1(x_1),\dots,W_s(x_s))^T$$

and $\mathbf{d}=\boldsymbol{W}(\mathbf{c})$ for a given anchor $\mathbf{c}\in D$. By $\boldsymbol{W}^{-1}$ denote the inverse transformation of $\boldsymbol{W}$.

By these transformations the integral $I_\rho(f)=\int_D f(\mathbf{x})\rho(\mathbf{x})d\mathbf{x}$ may be written as

$$\int_{[0,1]^s} g(\mathbf{y})d\mathbf{y},\quad \text{where}\quad g(\mathbf{y})=f(\boldsymbol{W}^{-1}(\mathbf{y})). \tag{44}$$

Let $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ be the weighted Banach space of functions defined on D with the corresponding norm of the form (8). The worst-case error

$$e^{\mathrm{wo}}(Q_n,\mathcal{F}^{(1)}_{p,q,\gamma,s})=\sup_{\|f\|_{p,q,\gamma,s}\le 1}|I_\rho(f)-Q_n(f)|$$

is still of the form (13) [HSW04a], where

$$h_u(\mathbf{t}_u)=I_\rho(M_u(\cdot,\mathbf{t}_u))-Q_n(M_u(\cdot,\mathbf{t}_u)),\quad \mathbf{t}_u\in D_u:=\prod_{k\in u}\overline{(a_k,b_k)}.$$

Denote the $|u|$-dimensional vectors $\mathbf{d}_u=(W_k(c_k))_{k\in u}$, $\mathbf{y}_u=(W_k(t_k))_{k\in u}$, $\mathbf{z}_u^j=(W_k(x_k^j))_{k\in u}$. By u_- denote the subset of u containing those indices k for which $t_k<c_k$. Then $h_u(\mathbf{t}_u)$ can be expressed as

$$h_u(\mathbf{t}_u)=(-1)^{u_-}\left[\mathrm{vol}(B_u(\mathbf{y}_u,\mathbf{d}_u))-\frac{1}{n}\sum_{j=0}^{n-1}1_{B_u(\mathbf{y}_u,\mathbf{d}_u)}(\mathbf{z}_u^j)\right],$$

where B_u is defined as in section 2. It follows that for $p \in [1,\infty]$ with conjugate p^*

$$\begin{aligned}\|h_u\|_{L_{p^*}} &= \left(\int_{D_u} |h_u(\mathbf{t}_u)|^{p^*} \rho_u(\mathbf{t}_u) d\mathbf{t}_u\right)^{\frac{1}{p^*}} \\ &= \left(\int_{[0,1]^u} \left|\mathrm{vol}(B_u(\mathbf{y}_u,\mathbf{d}_u)) - \frac{1}{n}\sum_{j=0}^{n-1} 1_{B_u(\mathbf{y}_u,\mathbf{d}_u)}(\mathbf{z}_u^j)\right|^{p^*} d\mathbf{y}_u\right)^{\frac{1}{p^*}} \\ &= D_{u,p^*}(\mathbf{z}^0,\ldots,\mathbf{z}^{n-1}),\end{aligned}$$

which is the *local L_{p^*} anchored discrepancy* of the point set $\{\mathbf{z}^0,\ldots,\mathbf{z}^{n-1}\}$. It is known that [Nie92]

$$D_{u,p^*}(\mathbf{z}^0,\ldots,\mathbf{z}^{n-1}) \le \kappa D^*_{u,\infty}(\mathbf{z}^0,\ldots,\mathbf{z}^{n-1}),$$

where $\kappa = 1$ if $\mathbf{d} = \boldsymbol{W}(\mathbf{c}) = \mathbf{1}$ and $\kappa = 2^{|u|}$ if $\mathbf{d} = \boldsymbol{W}(\mathbf{c})$, is in the interior of the unit cube $[0,1]^s$. Therefore, we have the following strong tractability result immediately following from Theorem 1.

THEOREM 3. *Let $p, q \in [1,\infty]$ and $\mathcal{F}^{(1)}_{p,q,\gamma,s}$ be defined as in section 2 for functions on the domain $D = \overline{(\mathbf{a},\mathbf{b})}$. Assume that $\{\mathbf{z}^j\}_{j\ge 0}$ is a Niederreiter $((T_u),s)$-sequence in base b, and $\{\mathbf{x}^j\}_{j\ge 0}$ is the transformed sequence according to the transformations in* (43)*. If*

$$\sum_{k=1}^{\infty} \gamma_k^a k \ln k < \infty$$

for any $a \in [1,q^*]$, then the corresponding integration is strongly tractable in the worst-case setting, and for any fixed $\epsilon > 0$, where C is some constant independent of s and n such that

$$e^{wo}(Q_n, \mathcal{F}^{(1)}_{p,q,\gamma,s}) \le C n^{-\frac{1}{a}+\epsilon}.$$

4.2. Randomized error analysis. Here we note that the assumption, $\|f\|_{p,q,\gamma,s} < \infty$, in the space $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ defined in section 3 might be too restrictive when D is unbounded. To alleviate this problem, we consider a modification following the approach from [HSW04b].

Consider a transformation of variables

$$\mathbf{y} = \boldsymbol{W}(\mathbf{x}) = (W_1(x_1),\ldots,W_s(x_s))^T,$$

where each W_k is a cumulative distribution function on interval (a_k,b_k) with density $w_k(x_k) = W_k'(x_k)$. Then integral $I_\rho(f) = \int_D f(\mathbf{x})\rho(\mathbf{x})d\mathbf{x}$ may be written as

$$\int_{[0,1]^s} g(\mathbf{y})d\mathbf{y}, \quad \text{where} \quad g(\mathbf{y}) = f(\boldsymbol{W}^{-1}(\mathbf{y}))\phi(\boldsymbol{W}^{-1}(\mathbf{y})), \tag{45}$$

$$\phi(\mathbf{x}) = \frac{\rho(\mathbf{x})}{w(\mathbf{x})}, \quad \text{and} \quad w(\mathbf{x}) = \prod_{k=1}^{s} w_k(x_k).$$

See [HSW04b] for a discussion about the necessity of introducing $w(\mathbf{x})$.

We now suppose that $\{\mathbf{z}^0, \dots, \mathbf{z}^{n-1}\}$ is a randomly scrambled digital net, and the integral of g in (45) is approximated by $n^{-1}\sum_{j=0}^{n-1} g(\mathbf{z}^j)$. This is equivalent to the rule

$$Q_n(f) = \frac{1}{n}\sum_{j=0}^{n-1} f(\boldsymbol{W}^{-1}(\mathbf{z}^j))\phi(\boldsymbol{W}^{-1}(\mathbf{z}^j)) \tag{46}$$

for $I_\rho(f)$. By Lemma 4 the error of this approximation has the following upper bound:

$$\sum_{\boldsymbol{\nu}:\|\lg(\boldsymbol{\nu})\|_1+T_{U(\boldsymbol{\nu})}>m} |G(\boldsymbol{\nu})|^2 3^{|U(\boldsymbol{\nu})|} b^{T_{U(\boldsymbol{\nu})}-m},$$

where the $G(\boldsymbol{\nu})$ are the Fourier coefficients of $g(\mathbf{y}) = f(\boldsymbol{W}^{-1}(\mathbf{y}))\phi(\boldsymbol{W}^{-1}(\mathbf{y}))$ under the Haar wavelets $\psi_{\boldsymbol{\nu}}$. In terms of $\xi_{\boldsymbol{\nu}}$ defined in (24) $G(\boldsymbol{\nu})$ can be expressed as

$$G(\boldsymbol{\nu}) = (-1)^{|U(\boldsymbol{\nu})|}\int_{S(\boldsymbol{\nu})} \xi_{\boldsymbol{\nu}}(\mathbf{y}) g'_{U(\boldsymbol{\nu})}(\mathbf{y})\,d\mathbf{y},$$

where $S(\boldsymbol{\nu})$ is the support of the wavelet $\psi_{\boldsymbol{\nu}}$, which is given in (21). Now for each nonempty subset $u \subseteq \{1, \dots, s\}$,

$$g'_u(\mathbf{y}) = \frac{1}{w_u(\boldsymbol{W}^{-1}(\mathbf{y}))}\left.\frac{\partial^{|u|}(f\phi)(\mathbf{x})}{\partial \mathbf{x}_u}\right|_{\mathbf{x}=\boldsymbol{W}^{-1}(\mathbf{y})},$$

where $f\phi$ just denotes the multiplication of the two functions, and $w_u(\mathbf{x}) = \prod_{k\in u} w_k(x_k)$. Applying Hölder's inequality yields the following:

$$\begin{aligned}|G(\boldsymbol{\nu})| \le{}& \left(\int_{S(\boldsymbol{\nu})} \left|\left.\frac{\partial^{|u|}(f\phi)(\mathbf{x})}{\partial \mathbf{x}_u}\right|_{\mathbf{x}=\boldsymbol{W}^{-1}(\mathbf{y})} \frac{1}{w_u(\boldsymbol{W}^{-1}(\mathbf{y}))}\right|^p d\mathbf{y}\right)^{\frac{1}{p}} \\ &\times \left(\int_{[0,1]^s} |\xi_{\boldsymbol{\nu}}(\mathbf{y})|^{p^*} d\mathbf{y}\right)^{\frac{1}{p^*}}.\end{aligned}$$

Define the norm

$$\|f\|_{p,q,\gamma,s,\rho,w} := \begin{cases} \left(\displaystyle\sum_{u\subseteq\{1,\dots,s\}} \gamma_u^{-q}\|(f\rho/w)'_u w_u^{-1}\|_{L_p}^q\right)^{\frac{1}{q}} & \text{for } q<\infty, \\ \displaystyle\max_{u\subseteq\{1,\dots,s\}} \left\{\gamma_u^{-1}\|(f\rho/w)'_u w_u^{-1}\|_{L_p}\right\} & \text{for } q=\infty. \end{cases} \tag{47}$$

Then we modify the space $\mathcal{F}^{(2)}_{p,q,\gamma,s}$ to $\mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}$, i.e., we let $\mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}$ be the weighted Banach space of all absolutely continuous functions f defined on D with $\|f\|_{p,q,\gamma,s,\rho,w} < \infty$. The randomized error is defined as

$$e^{\mathrm{ra}}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}) := \sup_{\|f\|_{p,q,\gamma,s,\rho,w}\le 1} \sqrt{E|I_\rho(f) - Q_n(f)|^2}.$$

From Theorem 2 we have the strong tractability results of integration over the general domain in the randomized setting.

THEOREM 4. *Consider the integration problem of approximating integral* (2) *by rule* (46) *with randomly scrambled Niederreiter digital nets in base b. Let* $\mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}$ *be defined as above.*

(i) *For* $p \in [1,\infty]$, *if*

$$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^2 < \infty,$$

then the corresponding integration is strongly tractable in the randomized setting, and for any fixed $\epsilon > 0$, *where* C *is some constant independent of* s *and* n *such that*

$$e^{ra}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}) \le C n^{-1+\epsilon}.$$

(ii) *In particular, for* $p \in [2,\infty]$, *if*

$$\sum_{k=1}^{\infty} \gamma_k^2 (k \ln k)^3 < \infty,$$

then the corresponding integration is strongly tractable in the randomized setting, and for any fixed $\epsilon > 0$, *where* C *is some constant independent of* s *and* n *such that*

$$e^{ra}(Q_n, \mathcal{F}^{(2)}_{p,q,\gamma,s,\rho,w}) \le C n^{-\frac{3}{2}+\epsilon}.$$

5. Concluding remarks. We have considered the strong tractability problems for multivariate integration using deterministic and randomly scrambled Niederreiter sequences. The spaces of integrands are weighted Banach spaces with parameters p, q, γ, s. The definitions of these spaces are slightly different in the worst-case and randomized settings. The main results of this article are summarized below.

Each of the conditions we found for strong tractability in worst-case and randomized settings is of the form $\sum_{k=1}^{\infty} \gamma_k^{\alpha} (k \ln k)^{\beta} < \infty$ for some positive numbers α and β. The values of α and β are determined by p or q. The larger α and smaller β imply weaker conditions of strong tractability. In the worst-case setting, the parameter p has little influence on the convergence rate of the worst-case error; however, the parameter q plays a significant role in determining strong tractability. In the randomized setting, only the parameter p plays a significant role in determining both strong tractability and convergence rate of the randomized error.

The factors $(k \ln k)^{\beta}$ in each term of the summation in the strong tractability conditions come from the quality parameter vector (T_u) of a Niederreiter sequence due to its telescopic property. These factors make the strong tractability conditions in this article slightly stronger than the conditions on integration using lattice rules. It is shown in [HSW04b, Theorem 3] that integration using lattice rules for the weighted Banach spaces is strongly tractable under the condition that $\sum_{k=1}^{\infty} \gamma_k^a < \infty$ for $a \in [1, q^*]$, and the convergence rate of the worst-case error is $O(n^{-1/a+\epsilon})$. We do not know yet whether integration using Niederreiter sequences is strongly tractable under the same condition as using lattice rules, but this is an interesting question worth pursuing.

The functions considered in this article depend on successive variables. We associate the first variable x_1 to the weight γ_1, the second variable x_2 to the weight γ_2, the variables $\mathbf{x}_u$ to the weights $\gamma_u = \prod_{k \in u} \gamma_k$, and so on. Although dimension-dependent weights are sometimes used in practice, we have some philosophical difficulties with

them. What this means is that a function that just depends on, say, the first five variables, could have its norm decrease as the nominal dimension, s, changes just because the weights change. On the other hand, weights not in product form are important; see, e.g., the work of [SWW04] and others on finite-order weights and tractability. We observe that each of the results in Theorems 1 to 4 holds for the Banach spaces of finite-order weights after a modification to the sufficient condition. This modification can be made according to the following lemma, compared with the proofs of Theorems 1 to 4.

LEMMA 6. *Let the finite-order weights $\{\gamma_{s,u}\}$ be of order d^*, i.e.,*

$$d^* = \min\{d : \gamma_{s,u} = 0 \quad \forall s \ \text{ and } \ \forall u \text{ with } |u| > d\}.$$

Let $\alpha, \beta, \theta > 0$ and $\tau \geq 1$. For the first n points of a Niederreiter $((T_u), s)$-sequence in base b, define

$$\Phi(\alpha, \beta, \theta, \tau) := \sum_{\emptyset \neq u \subseteq \{1,\dots,s\}, |u| \leq d^*} \gamma_{s,u}^{\alpha} b^{\beta T_u} (\theta \ln(\tau n))^{|u|}. \tag{48}$$

If the weights $\{\gamma_{s,u}\}$ satisfy

$$M_{\alpha,\beta} := \sup_{s=1,2,\cdots} \sum_{\emptyset \neq u \subseteq \{1,\dots,s\}, |u| \leq d^*} \gamma_{s,u}^{\alpha} \prod_{k \in u} [k \ln(k+b)]^{\beta} < \infty, \tag{49}$$

then for any fixed $\epsilon > 0$ there exists a constant C_ϵ independent of s and n such that

$$\Phi(\alpha, \beta, \theta, \tau) \leq C_\epsilon n^{\epsilon}.$$

Proof. From (6), (7), and (49) we have

$$\begin{aligned}
\Phi(\alpha, \beta, \theta, \tau) &\leq \sum_{\emptyset \neq u \subseteq \{1,\dots,s\}, |u| \leq d^*} \left(\tilde{\theta} \ln(\tau n)\right)^{|u|} \gamma_{s,u}^{\alpha} \prod_{k \in u} [k \ln(k+b)]^{\beta} \\
&= \sum_{r=1}^{d^*} \left(\tilde{\theta} \ln(\tau n)\right)^{r} \sum_{|u|=r} \gamma_{s,u}^{\alpha} \prod_{k \in u} [k \ln(k+b)]^{\beta} \leq M_{\alpha,\beta} \sum_{r=1}^{d^*} \left(\tilde{\theta} \ln(\tau n)\right)^{r},
\end{aligned}$$

where $\tilde{\theta} = \theta(b/\ln b)^{\beta}$. For any fixed $\epsilon > 0$ define

$$B_\epsilon = \max_{r=1,2,\dots,d^*} \left[r! (\tilde{\theta}/\epsilon)^r \right],$$

and we then have

$$\Phi(\alpha, \beta, \theta, \tau) \leq M_{\alpha,\beta} B_\epsilon \sum_{r=1}^{d^*} \frac{(\epsilon \ln(\tau n))^r}{r!} \leq M_{\alpha,\beta} B_\epsilon e^{\epsilon \ln(\tau n)} = C_\epsilon n^{\epsilon},$$

where $C_\epsilon = M B_\epsilon$. This completes the proof for the lemma. ∎

Therefore, we have the following modification:

1. For the Banach space, $\mathcal{F}_{p,q,\gamma,s}^{(1)}$, of finite-order weights $\gamma = \{\gamma_{s,u}\}$ with order d^*, condition (16) is modified by

$$M_{a,1} = \sup_{s=1,2,\cdots} \sum_{\emptyset \neq u \subseteq \{1,\dots,s\}, |u| \leq d^*} \gamma_{s,u}^{a} \prod_{k \in u} [k \ln(k+b)] < \infty.$$

2. For the Banach space, $\mathcal{F}^{(2)}_{p,q,\gamma,s}$, of finite-order weights $\gamma = \{\gamma_{s,u}\}$ with order d^*, conditions (34) and (36) are modified by

$$M_{2,2} = \sup_{s=1,2,\cdots} \sum_{\emptyset \neq u \subseteq \{1,\ldots,s\}, |u| \leq d^*} \gamma_{s,u}^2 \prod_{k \in u} [k \ln(k+b)]^2 < \infty$$

and

$$M_{2,3} = \sup_{s=1,2,\cdots} \sum_{\emptyset \neq u \subseteq \{1,\ldots,s\}, |u| \leq d^*} \gamma_{s,u}^2 \prod_{k \in u} [k \ln(k+b)]^3 < \infty,$$

respectively.

Appendix: Proof of Lemma 1. First we have

$$\sup_k \{\gamma_k^\beta k\} \leq \sum_{k=1}^\infty \gamma_k^\beta \quad \forall \beta > 0,$$

since for any integer $K > 0$

$$\sum_{k=1}^\infty \gamma_k^\beta \geq \sum_{k=1}^K \gamma_k^\beta \geq K \gamma_K^\beta.$$

Now for any $\delta, \delta' > 0$ applying the assumption on λ_k we have

$$\begin{aligned}\sum_{k=1}^\infty \gamma_k^a \lambda_k &\leq \sup_k \left\{ \gamma_k^{\frac{ar(1+\delta)}{1+r}} c_{r,\delta} k^{r+\delta'} \right\} \sum_{k=1}^\infty \gamma_k^{\frac{a(1-r\delta)}{1+r}} \\ &\leq \sum_{k=1}^\infty \gamma_k^a c_{r,\delta'} k^{r+\delta'} = \sum_{k=1}^\infty c_{r,\delta'} k^{r+\delta'} \gamma_k^{\frac{ar(1+\delta)}{1+r}} \gamma_k^{\frac{a(1-r\delta)}{1+r}}.\end{aligned}$$

We choose δ' such that $r + \delta' = r(1+\delta)/(1-r\delta)$, and then we have

$$\sup_k \left\{ \gamma_k^{\frac{ar(1+\delta)}{1+r}} k^{r+\delta'} \right\} = \left[\sup_k \gamma_k^{\frac{a(1-r\delta)}{1+r}} k \right]^{\frac{r(1+\delta)}{1-r\delta}} \leq \left[\sum_{k=1}^\infty \gamma_k^{\frac{a(1-r\delta)}{1+r}} \right]^{\frac{r(1+\delta)}{1-r\delta}}.$$

It follows that

$$\sum_{k=1}^\infty \gamma_k^a \lambda_k \leq c_{r,\delta'} \left[\sum_{k=1}^\infty \gamma_k^{\frac{a(1-r\delta)}{1+r}} \right]^{\frac{r(1+\delta)}{1-r\delta}+1},$$

which gives the right-hand side of the desired result.

As for the left-hand side of the desired result, we write for any $\delta > 0$

$$\sum_{k=1}^\infty \gamma_k^{\frac{a(1+\delta)}{1+r}} = \sum_{k=1}^\infty \gamma_k^{\frac{a(1+\delta)}{1+r}} k^{\frac{(r-\delta)(1+\delta)}{1+r}} k^{\frac{(\delta-r)(1+\delta)}{1+r}}.$$

Applying Hölder's inequality yields

$$\begin{aligned}\sum_{k=1}^\infty \gamma_k^{\frac{a(1+\delta)}{1+r}} &\leq \left[\sum_{k=1}^\infty \gamma_k^a k^{r-\delta} \right]^{\frac{1+\delta}{1+r}} \left[\sum_{k=1}^\infty k^{\frac{(1+\delta)(\delta-r)}{r-\delta}} \right]^{\frac{r-\delta}{1+r}} \\ &= \left[\sum_{k=1}^\infty \gamma_k^a k^{r-\delta} \right]^{\frac{1+\delta}{1+r}} \left[\sum_{k=1}^\infty k^{-(1+\delta)} \right]^{\frac{r-\delta}{1+r}}.\end{aligned}$$

Noting that $\sum_{k=1}^{\infty} k^{-(1+\delta)} < \infty$ and applying the assumption on λ_k again yields

$$\sum_{k=1}^{\infty} \gamma_k^{\frac{a(1+\delta)}{1+r}} \le \left[\sum_{k=1}^{\infty} \gamma_k^a \lambda_k\right]^{\frac{1+\delta}{1+r}} \hat{c}_{r,\delta},$$

which gives the left-hand side of the desired result. The proof of Lemma 1 is complete. □

Acknowledgments. The authors thank the two anonymous referees for a number of detailed, constructive suggestions that contributed greatly to the manuscript.

REFERENCES

[DP05] J. Dick and F. Pillichshammer, *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity, 21 (2005), pp. 149–195.

[Ent97] K. Entacher, *Quasi-Monte Carlo methods for numerical integration of multivariate Haar series* I, BIT, 37 (1997), pp. 846–861.

[Ent98] K. Entacher, *Quasi-Monte Carlo methods for numerical integration of multivariate Haar series* II, BIT, 38 (1998), pp. 283–292.

[FW94] K. T. Fang and Y. Wang, *Number-theoretic Methods in Statistics*, Chapman and Hall, London, 1994.

[Gen92] A. Genz, *Statistics applications of subregion adaptive multiple numerical integration*, in Numerical Integration—Recent Developments, Software and Applications, T. O. Espelid and A. Genz, eds., Kluwer Academic Publishers, Dordrecht, 1992, pp. 267–280.

[Hal60] J. H. Halton, *On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals*, Numer. Math., 2 (1960), pp. 84–90.

[HD04] F. J. Hickernell and J. Dick, *An algorithm-driven approach to error analysis for multidimensional integration*, submitted, 2004.

[HHY04] S. Heinrich, F. J. Hickernell, and R. X. Yue, *Optimal quadrature for Haar wavelet spaces*, Math. Comp., 73 (2004), pp. 259–277.

[HSW04a] F. J. Hickernell, I. H. Sloan, and G. W. Wasilkowski, *On tractability of weighted integration for certain Banach spaces of functions*, in Monte Carlo and Quasi-Monte Carlo Methods 2002, H. Niederreiter, ed., Springer-Verlag, Berlin, 2004, pp. 51–71.

[HSW04b] F. J. Hickernell, I. H. Sloan, and G. W. Wasilkowski, *The strong tractability of multivariate integration using lattice rules*, in Monte Carlo and Quasi-Monte Carlo Methods 2002, H. Niederreiter, ed., Springer-Verlag, Berlin, 2004, pp. 259–273.

[HSW04c] F. J. Hickernell, I. H. Sloan, and G. W. Wasilkowski, *On tractability of weighted integration over bounded and unbounded regions in R^s*, Math. Comp., 73 (2004), pp. 1885–1901.

[HW02] F. J. Hickernell and X. Wang, *The error bounds and tractability of quasi-Monte Carlo algorithms in infinite dimension*, Math. Comp., 71 (2002), pp. 1641–1661.

[HY00] F. J. Hickernell and R. X. Yue, *The mean square discrepancy of scrambled (t, s)-sequences*, SIAM J. Numer. Anal., 38 (2000), pp. 1089–1112.

[Kei96] B. D. Keister, *Multidimensional quadrature algorithms*, Comput. Phys., 10 (1996), pp. 119–122.

[Nie88] H. Niederreiter, *Low-discrepancy and low-dispersion sequences*, J. Number Theory, 30 (1988), pp. 51–70.

[Nie92] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Con. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.

[NX01] H. Niederreiter and C. P. Xing, *Rational Points on Curves over Finite Fields: Theory and Applications*, London Math. Soc. Lecture Note Ser. 285, Cambridge University Press, Cambridge, UK, 2001.

[NW01] E. Novak and H. Woźniakowski, *When are integration and discrepancy tractable?*, in Foundation of Computational Mathematics, R. A. DeVore, A. Iserles, and E. Süli, eds., Cambridge University Press, Cambridge, UK, 2001.

[Owe95] A. B. Owen, *Randomly permuted (t, m, s)-nets and (t, s)-sequences*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Lecture Notes in Statist.

106, H. Niederreiter and P. J. S. Shiue, eds., Springer-Verlag, New York, 1995, pp. 299–317.

[Owe97a] A. B. Owen, *Monte Carlo variance of scrambled net quadrature*, SIAM J. Numer. Anal., 34 (1997), pp. 1884–1910.

[Owe97b] A. B. Owen, *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25 (1997), pp. 1541–1562.

[Owe98] A. B. Owen, *Scrambled Sobol and Niederreiter-Xing points*, J. Complexity, 14 (1998), pp. 466–489.

[Owe00] A. B. Owen, *Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo*, in Monte Carlo and Quasi-Monte Carlo Methods 1998, H. Niederreiter and P. J. S. Shiue, eds., Springer-Verlag, Berlin, 2000.

[Owe05] A. B. Owen, *Multidimensional variation for quasi-Monte Carlo*, in Contemporary Multivariate Analysis and Experimental Design (Singapore), J. Fan and G. Li, eds., Series in Biostatistics, Vol. 2, World Scientific, River Edge, NJ, 2005, pp. 49–74.

[PT96] A. Papageorgiou and J. F. Traub, *Beating Monte Carlo*, Risk, 9 (1996), pp. 63–65.

[Slo02] I. H. Sloan, *QMC integration—beating intractability by weighting the coordinate directions*, in Monte Carlo and Quasi-Monte Carlo Methods 2000, K. T. Fang, F. J. Hickernell, and H. Niederreiter, eds., Springer-Verlag, Berlin, 2002, pp. 103–123.

[Sob67] I. M. Sobol', *The distribution of points in a cube and the accurate evaluation of integrals*, Zh. Vychisl. Mat. Mat. Fiz., 7 (1967), pp. 784–802 (in Russian).

[Sob69] I. M. Sobol', *Multidimensional Quadrature Formulas and Haar Functions*, Izdat. Nauka, Moscow, 1969 (in Russian).

[SWW04] I. H. Sloan, X. Wang, and H. Woźniakowski, *Finite-order weights imply tractability of multivariate integration*, J. Complexity, 20 (2004), pp. 46–74.

[SW98] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.

[SW01] I. H. Sloan and H. Woźniakowski, *Tractability of multivariate integration for weighted Korobov classes*, J. Complexity, 17 (2001), pp. 697–721.

[Tez95] S. Tezuka, *Uniform Random Numbers: Theory and Practice*, Kluwer Academic Publishers, Dordrecht, 1995.

[Wal02] D. F. Walnut, *An Introduction to Wavelet Analysis*, Birkhäuser Boston, Cambridge, MA, 2002.

[Wan03] X. Wang, *Strong tractability of multivariate integration using quasi-Monte Carlo algorithms*, Math. Comp., 72 (2003), pp. 823–838.

[YH01] R. X. Yue and F. J. Hickernell, *Integration and approximation based on scramble sampling in arbitrary dimensions*, J. Complexity, 17 (2001), pp. 881–897.

[YH05] R. X. Yue and F. J. Hickernell, *Strong tractability of integration using scrambled Niederreiter points*, Math. Comp., 74 (2005), pp. 1871–1893.

[YM99] R. X. Yue and S. S. Mao, *On the variance of quadrature over scrambled nets and sequences*, Statist. Probab. Lett., 44 (1999), pp. 267–280.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2584–2609

LOCAL LINEARIZATION METHOD FOR NUMERICAL INTEGRATION OF DELAY DIFFERENTIAL EQUATIONS*

J. C. JIMENEZ†, L. M. PEDROSO†, F. CARBONELL†, AND V. HERNANDEZ†

Abstract. In this paper, a new approach for the numerical computation of delay differential equations (DDEs) is introduced. The essential idea consists of obtaining numerical integrators that use a code expressly developed for linear DDEs, in contrast with the conventional approach of using a code for ordinary differential equations. Specifically, two numerical schemes of this new class of integrators are proposed and their numerical viability analyzed. It includes the estimation of the convergence rate, the evaluation of the computational cost of the schemes, and a simulation study. It is proved that these one-step explicit integrators converge uniformly with order two to the solution of nonlinear DDEs and are able to integrate stiff equations in a satisfactory way with low computational cost.

Key words. delay differential equations, local linearization methods, numerical integrators

AMS subject classifications. 34K28, 65L05, 65L20

DOI. 10.1137/040607356

1. Introduction.

In recent years interest has increased in the numerical solution of delay differential equations (DDEs) with constant delay. Such interest was motivated by their applicability in the mathematical modeling of several physical, chemical, and biological processes, where they provide the best and sometimes the only realistic simulation of the observable phenomena [7, 17, 30].

There exists a variety of such numerical integrators, which essentially have two main ingredients [2]: (1) the emulation of the method of the steps in order to obtain piecewise ordinary differential equations (ODEs), and (2) the application of a variable step-size ODE code with a suitable approximation of the retarded solutions. Examples include the schemes proposed in [8, 21, 23, 25, 27, 33, 34, 41], which use several ODE codes (e.g., Euler, Runge–Kutta, multisteps, and local linearization (LL)) and several ways to approximate the retarded solutions (e.g., polynomial functions, θ-methods, continuous extensions of Runge–Kutta, and LL methods). Although the convergence and linear stability of these methods have been well studied, it is not the case of the preserving qualitative features of such methods [2]. It is well known [13, 39] that, in general, conventional numerical integrators for ODEs do not preserve the dynamical properties of the original ODEs. Therefore, it is expected that the numerical integrators for DDEs derived from the above-mentioned ODE codes do not preserve the dynamic properties of the original DDEs either.

It is also well known that for the stability analysis of ODEs, as well as for DDEs, there are two main techniques [2]: (1) the Lyapunov theory, and (2) the stability theory in first approximation. The latter, simpler one is based on the local linearization of the differential equations. This kind of linearization is also the main component of the so-called LL integrators for ODEs. In recent papers [26, 15], it has been shown that this type of scheme preserves the dynamical properties of the original equations

*Received by the editors April 26, 2004; accepted for publication (in revised form) May 30, 2006; published electronically December 11, 2006. This work was partially supported by research grant 03-059 RG/MATHS/LA from the Third World Academy of Science.

http://www.siam.org/journals/sinum/44-6/60735.html

†Instituto de Cibernética Matemática y Físisca, Calle 15 No. 551 entre C y D, Vedado, La Habana 10400, Cuba (jcarlos@icmf.inf.cu, liuva@icmf.inf.cu, felix@icmf.inf.cu, vivky@icmf.inf.cu).

much better than the conventional numerical integrators. For instance, under quite general conditions, they do not have spurious equilibrium points and preserve the local stability of the exact solution at hyperbolic equilibrium points and periodic orbits. On the other hand, this linearization approach has been the key for the construction of efficient and stable numerical schemes for the integration and estimation of various classes of random dynamical systems (see [28, 29, 36, 37] and the references therein). Specifically, in the framework of stochastic and random differential equations, simulations studies have shown that the LL integrators have similar stability properties to the conventional implicit integrators with the computational efficiency of the explicit ones (see, for instance, [5, 10, 9]). In addition, in the framework of the nonlinear filtering problems the LL filters have similar features (see, for instance, [35]). In all cases, the piecewise linearization of the vector fields that define the differential equations to be integrated is the keystone in the construction of the LL integrators and, at the same time, the main difference with the conventional numerical integrators (which are typically derived from a primary expansion of the unknown solution in power series). Thus, the application of this LL approach for the integration of DDEs is also attractive.

The goal of this paper is to study the numerical viability of the LL approach for defining a new type of numerical integrators for DDEs, leaving the qualitative analysis of them for a future paper. The essential ideas of this approach are (1) approximate linearly the vector field of the DDE in order to obtain a piecewise linear DDE, and (2) compute the solution of such linear equations by the variation-of-constants formula with a suitable approximation of the retarded solutions.

The paper is organized as follow. In section 2, the LL method is introduced and two numerical schemes are proposed. In section 3, the convergence of the method is studied, while in the last section a simulation study is carried out in order to illustrate the performance of the method.

2. LL method. Let $\mathbf{f} : \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$ be a differentiable function and let $\mathbf{x}(t)$ be the solution of the m-dimensional DDE

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}\left(t, \mathbf{x}(t), \mathbf{x}_t(-\tau)\right), \quad t \in [t_0, T], \tag{2.1}$$

$$\mathbf{x}_{t_0}(s) = \varphi(s), \quad s \in [-\tau, 0], \tag{2.2}$$

at the point $t \in [t_0 - \tau, T]$, where $\tau > 0$ is a constant delay, $\varphi : [-\tau, 0] \longrightarrow \mathbb{R}^m$ is a given initial function, and $\mathbf{x}_t : [-\tau, 0] \longrightarrow \mathbb{R}^m$ is the segment function defined as

$$\mathbf{x}_t(s) := \mathbf{x}(t+s), \quad s \in [-\tau, 0],$$

for all $t \in [t_0, T]$. Lipschitz and smoothness conditions on the function $\mathbf{f}$ are also assumed in order to ensure a unique solution for (2.1)–(2.2).

Let $(t)_h = \{t_0 < t_1 < \cdots < t_n < \cdots \leq T\}$ be a partition of the time interval $[t_0, T]$ such that

$$\sup_n (t_{n+1} - t_n) \leq h < 1, \tag{2.3}$$

and define

$$n_t := \max\{n = 0, 1, 2, \ldots, : t_n \leq t \text{ and } t_n \in (t)_h\}$$

for $t \in [t_0, T]$. Throughout this paper it will be assumed that condition $h < \tau$ holds.

Suppose that, for all $t_n \in (t)_h$, $\mathbf{y}_n \in \mathbb{R}^m$ is a point close to $\mathbf{x}(t_n)$. For all $t \in [t_0, T]$, let $\widetilde{\mathbf{y}}_t : [-\tau, 0] \longrightarrow \mathbb{R}^m$ be a segment function that approximates to $\mathbf{x}_t$, such that $\widetilde{\mathbf{y}}_{t_n}(0) = \mathbf{y}_n$ for all $t_n \in (t)_h$.

In addition, let us consider the first order Taylor expansion of the function $\mathbf{f}$ around the point $(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))$,

$$\begin{aligned}\mathbf{f}(s, \mathbf{u}, \mathbf{v}) \approx \mathbf{f}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau)) + \mathbf{f}_x(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))(\mathbf{u} - \mathbf{y}_n) \\ + \mathbf{f}_{x_t}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))(\mathbf{v} - \widetilde{\mathbf{y}}_{t_n}(-\tau)) + \mathbf{f}_t(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))(s - t_n)\end{aligned}$$

for $s \in \mathbb{R}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$, where $\mathbf{f}_x$, $\mathbf{f}_{x_t}$, and $\mathbf{f}_t$ denote the partial derivatives of $\mathbf{f}$ with respect to the variables $\mathbf{x}$, $\mathbf{x}_t$, and t, respectively. Taking into account that $\mathbf{f}$ can be linearly approximated by its first order Taylor expansion, the solution of (2.1)–(2.2) can be locally approximated on each interval $[t_n, t_{n+1})$ by the solution of the linear DDE

$$\begin{aligned}\frac{d\mathbf{z}(t)}{dt} &= \mathbf{A}_n\mathbf{z}(t) + \mathbf{B}_n\mathbf{z}_t(-\tau) + \mathbf{c}_n t - \mathbf{c}_n t_n + \mathbf{d}_n \\ &\quad -\mathbf{A}_n\mathbf{y}_n - \mathbf{B}_n\widetilde{\mathbf{y}}_{t_n}(-\tau),\ t \in [t_n, t_{n+1}), \\ \mathbf{z}_{t_n}(s) &= \widetilde{\mathbf{y}}_{t_n}(s), \quad s \in [-\tau, 0],\end{aligned} \tag{2.4}$$

which is given by [20]

$$\begin{aligned}\mathbf{z}(t) = e^{\mathbf{A}_n(t-t_n)}\Big\{\mathbf{y}_n + \int_0^{t-t_n} e^{-\mathbf{A}_n u}(\mathbf{B}_n(\widetilde{\mathbf{y}}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(-\tau)) + \mathbf{c}_n u + \mathbf{d}_n \\ - \mathbf{A}_n\mathbf{y}_n)du\Big\},\end{aligned} \tag{2.5}$$

where $\mathbf{A}_n = \mathbf{f}_x(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))$, $\mathbf{B}_n = \mathbf{f}_{x_t}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))$ are constant matrices, $\mathbf{c}_n = \mathbf{f}_t(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))$, $\mathbf{d}_n = \mathbf{f}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}_{t_n}(-\tau))$ are constant vectors, and $t_n, t_{n+1} \in (t)_h$. Further, by using the identity

$$\int_0^{\Delta} e^{-\mathbf{A}_n u}du\ \mathbf{A}_n = -(e^{-\mathbf{A}_n\Delta} - \mathbf{I}), \qquad \Delta \geq 0, \tag{2.6}$$

and simple rules from the integral calculus, the above expression can be conveniently rewritten as

$$\mathbf{z}(t) = \mathbf{y}_n + \Phi(t_n, \mathbf{y}_n, t - t_n; \widetilde{\mathbf{y}}_{t_n}), \tag{2.7}$$

where

$$\begin{aligned}\Phi(t_n, \mathbf{y}_n, t - t_n; \widetilde{\mathbf{y}}_{t_n}) = \int_0^{t-t_n} e^{\mathbf{A}_n(t-t_n-u)}(\mathbf{B}_n(\widetilde{\mathbf{y}}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(-\tau)) + \mathbf{d}_n)du \\ + \int_0^{t-t_n}\int_0^{u} e^{\mathbf{A}_n(t-t_n-u)}\mathbf{c}_n dr du.\end{aligned} \tag{2.8}$$

In this way, by setting $\mathbf{y}_0 = \mathbf{x}(t_0)$ and iteratively evaluating the expression (2.7) at t_{n+1} (for $n = 0, 1, \dots$) a sequence of points

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Phi(t_n, \mathbf{y}_n, t_{n+1} - t_n; \widetilde{\mathbf{y}}_{t_n})$$

can be obtained as an approximation to the solution $\mathbf{x}$ of (2.1)–(2.2) at each point $t_{n+1} \in (t)_h$. This just defines the local linear discretization of a DDE. More precisely, we have the following definition.

DEFINITION 2.1. *For a time discretization $(t)_h$, the local linear discretization of the solution of* (2.1)–(2.2) *at each point $t_{n+1} \in (t)_h$ is defined by the recursive expression*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Phi(t_n, \mathbf{y}_n, t_{n+1} - t_n; \widetilde{\mathbf{y}}_{t_n}), \tag{2.9}$$

where $\widetilde{\mathbf{y}}_{t_n} : [-\tau, 0] \longrightarrow \mathbb{R}^m$ is a segment function that approximate to $\mathbf{x}_{t_n}$, such that $\widetilde{\mathbf{y}}_{t_n}(0) = \mathbf{y}_n$.

Moreover, an approximation for $\mathbf{x}$ in the whole interval $[t_0 - \tau, T]$ is stated in the definition below.

DEFINITION 2.2. *For a time discretization $(t)_h$, the local linear approximation of the solution of* (2.1)–(2.2) *is defined by the function*

$$\mathbf{y}(t) = \mathbf{y}_{n_t} + \Phi(t_{n_t}, \mathbf{y}_{n_t}, t - t_{n_t}; \widetilde{\mathbf{y}}_{t_{n_t}}) \tag{2.10}$$

for all $t \in [t_0, T]$ and by

$$\mathbf{y}(t) = \varphi(t)$$

for $t \in [t_0 - \tau, t_0]$. Here, $\mathbf{y}_{n_t}$ denotes the LL discretization (2.9) *at n_t, and $\widetilde{\mathbf{y}}_{t_{n_t}}$ is the segment function of Definition* 2.1.

In addition, for $t \in [t_0, T]$, let $\mathbf{y}_t : [-\tau, 0] \longrightarrow \mathbb{R}^m$ be the segment function defined as

$$\mathbf{y}_t(s) := \mathbf{y}(t + s),\ s \in [-\tau, 0],$$

where $\mathbf{y}(t + s)$ is the LL approximation (2.10) evaluated at the point $t + s$.

It is clear that the LL approximation is a continuous function that coincides with the LL discretization at each point of the time discretization $(t)_h$.

As can be noted from the definition, to compute the LL discretization at the time t_{n+1}, a suitable approximation $\widetilde{\mathbf{y}}_{t_n}$ to $\mathbf{x}_{t_n}$ is assumed to be given. Based on the choice of such an approximation, different kinds of LL schemes could be defined. In the following subsections, two LL schemes will be introduced.

2.1. Natural LL scheme. Let us consider the numerical scheme that is defined in a natural way by taking $\widetilde{\mathbf{y}}_t(s)$ as the LL approximation $\mathbf{y}_t(s)$ for all $s \in [-\tau, 0]$ and $t + s \in [t_0, T]$. Specifically, the scheme shall be defined through the expression (2.9) with $\widetilde{\mathbf{y}}_{t_n} \equiv \mathbf{y}_{t_n}$ for all $t_n \in (t)_h \setminus t_0$, and $\widetilde{\mathbf{y}}_{t_0}(s) \equiv \widetilde{\varphi}(s)$ for all $s \in [-\tau, 0]$, where $\widetilde{\varphi}$ is an approximation to φ that shall be defined below.

In this subsection, and only here, it is assumed that the points in the time discretization $(t)_h$ are equidistant, i.e., $t_{n+1} - t_n = h = \frac{\tau}{N_0}$, for a fixed $N_0 \in \mathbb{N}_+$.

Consider the times t_n, $n = 0, \ldots, N_0$, for which $t_0 \le t_n \le t_0 + \tau$, and denote $s_n = t_n - \tau$. Suppose that in the interval $[s_n, s_{n+1}]$, the initial function φ in (2.2) can be exponentially approximated by the function

$$\widetilde{\varphi}(s_n + u) = \varphi(s_n) + \mathbf{L} e^{\mathbf{T}_n u} \mathbf{R}, \quad u \in [0, h], \tag{2.11}$$

where $\mathbf{T}_n$, $\mathbf{L}$, and $\mathbf{R}$ are certain constant matrices such that $\mathbf{L} e^{\mathbf{T}_n u} \mathbf{R} \in \mathbb{R}^m$. For instance, when $\varphi(s_n + u)$ is approximated by the interpolating polynomial $\sum_{i=0}^{p} \alpha_{i,n} u^i$

these matrices can be chosen as

$$\mathbf{T}_n = \begin{pmatrix} \mathbf{0}_{m\times m} & p!\alpha_{p,n} & (p-1)!\alpha_{p-1,n} & \cdots & 2\alpha_{2,n} & \alpha_{1,n} \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(m+p)\times(m+p)},$$

$\mathbf{L} = (\mathbf{I}_m, \mathbf{0}_{m\times p})$, and $\mathbf{R}^\intercal = (\mathbf{0}_{1\times m+p-1}, 1)$, where the coefficients $\alpha_{i,n}$ are obtained from interpolating conditions. The statement above is straightforwardly derived from Lemma 6.1 in the appendix and the expression

$$\sum_{i=0}^{p} \alpha_{i,n} u^i = \alpha_{0,n} + \int_0^u \alpha_{1,n} ds + \sum_{i=2}^{p} i!\alpha_{i,n} \int_0^u \int_0^{s_1} \cdots \int_0^{s_{i-1}} ds_i ds_{i-1} \ldots ds_1, \tag{2.12}$$

with $\alpha_{0,n} = \varphi(s_n)$.

Taking into account that $\widetilde{\mathbf{y}}_{t_n}(u-\tau) = \widetilde{\varphi}(s_n + u)$ it follows that

$$\Phi(t_n, \mathbf{y}_n, h; \widetilde{\mathbf{y}}_{t_n}) = \int_0^h e^{\mathbf{A}_n(h-u)}(\mathbf{B}_n\mathbf{L}e^{\mathbf{T}_n u}\mathbf{R} + \mathbf{d}_n)du + \int_0^h \int_0^u e^{\mathbf{A}_n(h-u)}\mathbf{c}_n dr du,$$

which, by (2.6), can be rewritten as

$$\begin{aligned}\Phi(t_n, \mathbf{y}_n, h; \widetilde{\mathbf{y}}_{t_n}) = &\int_0^h \int_0^u e^{\mathbf{A}_n(h-u)}\mathbf{B}_n\mathbf{L}e^{\mathbf{T}_n r}\mathbf{T}_n\mathbf{R}dr du + \int_0^h e^{\mathbf{A}_n(h-u)}\mathbf{B}_n\mathbf{L}\mathbf{R}du \\ &+ \int_0^h e^{\mathbf{A}_n(h-u)}\mathbf{d}_n du + \int_0^h \int_0^u e^{\mathbf{A}_n(h-u)}\mathbf{c}_n dr du.\end{aligned}$$

Now, since $\mathbf{L}\mathbf{R} = \mathbf{0}_{m\times 1}$, by Lemma 6.1 it is obtained that

$$\Phi(t_n, \mathbf{y}_n, h; \widetilde{\mathbf{y}}_{t_n}) = \mathbf{L}_0 e^{\mathbf{T}_{0,n}h}\mathbf{R}_0,$$

where

$$\mathbf{T}_{0,n} = \begin{pmatrix} \mathbf{A}_n & \mathbf{B}_n\mathbf{L} & \mathbf{c}_n & \mathbf{d}_n \\ 0 & \mathbf{T}_n & 0 & \mathbf{T}_n\mathbf{R} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad t_n \in [t_0, t_0+\tau],$$

$\mathbf{L}_0 = (\mathbf{I}_m, \mathbf{0}_{m\times m+p+2})$, and $\mathbf{R}_0^\intercal = (\mathbf{0}_{1\times 2m+p+1}, 1)$. Therefore, for $t_0 \le t_n \le t_0+\tau$, the expression

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{L}_0 e^{\mathbf{T}_{0,n}h}\mathbf{R}_0, \quad n = 0, \ldots, N_0 - 1,$$

defines an LL discretization, while the expression

$$\mathbf{y}(t) = \mathbf{y}_{n_t} + \mathbf{L}_0 e^{\mathbf{T}_{0,n}(t-t_{n_t})}\mathbf{R}_0, \quad n = 0, \ldots, N_0 - 1, \tag{2.13}$$

defines an LL approximation for all $t \in [t_0, t_0+\tau]$.

Taking into account the analogy between the expressions (2.13) and (2.11), the procedure above can be used to extend the LL approximation to $t_0+\tau \le t_n \le t_0+2\tau$,

and so on. In this way, for $t_0 + k\tau \le t_n \le t_0 + (k+1)\,\tau$ with $k = 1, 2, \ldots,$ we obtain the expression

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{L}_k e^{\mathbf{T}_{k,n} h} \mathbf{R}_k, \quad n = kN_0, \ldots, (k+1)N_0 - 1,$$

which defines the natural LL scheme. Here, the matrices $\mathbf{T}_{k,n}$, $\mathbf{L}_k$, and $\mathbf{R}_k$ are recursively defined by

$$\mathbf{T}_{k,n} = \begin{pmatrix} \mathbf{A}_n & \mathbf{B}_n\mathbf{L}_{k-1} & \mathbf{c}_n & \mathbf{d}_n \\ 0 & \mathbf{T}_{k-1,n} & 0 & \mathbf{T}_{k-1,n}\mathbf{R}_{k-1} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad t_n \in [t_0 + k\tau, t_0 + (k+1)\tau],$$

$\mathbf{L}_k = (\mathbf{L}_{k-1}, \mathbf{0}_{m+2})$, and $\mathbf{R}_k^\intercal = (\mathbf{0}_{1\times m+2}, \mathbf{R}_{k-1}^\intercal)$.

Note that the natural LL scheme produces the exact solution of linear DDEs with polynomial or exponential initial condition. Thus, obviously, the numerical solution provided by it preserves the stability properties of linear DDEs. However, observe also that the dimension of the matrices $\mathbf{T}_{k,n}$ increases with k, which increases the computational cost of the scheme when $\tau \ll T$. In this case, the use of Krylov subspace methods [22] to compute these high dimensional exponential matrices are highly recommended in order to reduce the computational cost of the natural LL scheme.

2.2. Polynomial LL scheme. Let us consider a piecewise polynomial approximation $\widetilde{\mathbf{y}}_{t_n}(s)$ to $\mathbf{y}_{t_n}(s)$ for all $s \in [-\tau, 0]$ defined in such a way that $\widetilde{\mathbf{y}}_{t_n}(0) = \mathbf{y}_{t_n}(0)$ and

$$\widetilde{\mathbf{y}}_{t_n}(u - \tau) = \alpha_{0,n} + \sum_{i=1}^{p} \alpha_{i,n} u^i, \quad u \in [0, h_n], \tag{2.14}$$

where the coefficients $\alpha_{i,n}$ are obtained from either interpolating or smoothness conditions, $h_n = t_{n+1} - t_n$ and $t_{n+1}, t_n \in (t)_h$.[1] By taking into account the integral representation (2.12) for polynomials and that $\widetilde{\mathbf{y}}_{t_n}(-\tau) = \alpha_{0,n}$, the function Φ can be rewritten as

$$\begin{aligned}\Phi(t_n, \mathbf{y}_n, h_n; \widetilde{\mathbf{y}}_{t_n}) = &\int_0^{h_n} e^{\mathbf{A}_n(h_n - u)} \mathbf{d}_n du + \int_0^{h_n}\int_0^{u} e^{\mathbf{A}_n(h_n-u)} \left(\mathbf{c}_n + \mathbf{B}_n\alpha_{1,n}\right) ds du \\ &+ \sum_{i=2}^{p} (i!\alpha_{i,n}) \int_0^{h_n}\int_0^{u}\int_0^{s_1} \cdots \int_0^{s_{i-1}} e^{\mathbf{A}_n(h_n-u)} \mathbf{B}_n ds_i ds_{i-1} \ldots ds_1 du.\end{aligned}$$

Then, from Lemma 6.1 it follows that

$$\Phi(t_n, \mathbf{y}_{t_n}, h_n; \widetilde{\mathbf{y}}_{t_n}) = \mathbf{L} e^{\mathbf{T}_n h_n} \mathbf{R},$$

where $\mathbf{T}_n \in \mathbb{R}^{(m+p+1)\times(m+p+1)}$ is given by

$$\mathbf{T}_n = \begin{pmatrix} \mathbf{A}_n & p!\mathbf{B}_n\alpha_{p,n} & (p-1)!\mathbf{B}_n\alpha_{p-1,n} & \cdots & 2\mathbf{B}_n\alpha_{2,n} & \mathbf{c}_n + \mathbf{B}_n\alpha_{1,n} & \mathbf{d}_n \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

[1] Note that no specification for the piecewise polynomial $\widetilde{\mathbf{y}}_{t_n}$ is given for $s \in (h_{n+1} - \tau, 0)$.

$\mathbf{L} = (\mathbf{I}_m, \mathbf{0}_{m\times p+1})$, and $\mathbf{R}^\intercal = (\mathbf{0}_{1\times m+p}, 1)$. Therefore, for all $t_n \in (t)_h$, the expression

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{L}e^{\mathbf{T}_n h_n}\mathbf{R} \tag{2.15}$$

defines the polynomial LL scheme.

Note that, in contrast to the natural LL scheme, the polynomial LL scheme is defined in terms of a matrix exponential of fixed dimension for all t_n. Therefore, this scheme is computationally feasible and its numerical implementation is reduced to using a convenient algorithm to compute matrix exponentials, e.g., those based on rational Padé approximations [18], the Schur decomposition [18], or Krylov subspace methods [22] (for a recent review see [38]). The selection of one of them will mainly depend on the size and structure of the matrices $\mathbf{T}_n$. For instance, for many low-dimensional systems of equations it is enough to use the algorithm developed in [40], which takes advantage of the special structure of the matrices $\mathbf{T}_n$, whereas for large systems of equations the Krylov subspace methods are strongly recommended.

2.3. LL schemes for equations with multiple delays. Let $\mathbf{f} : \mathbb{R} \times \prod_{i=1}^{d+1} \mathbb{R}^m \longrightarrow \mathbb{R}^m$ be a differentiable function. Consider the m-dimensional DDE with d constant delays defined by

$$\begin{aligned} &\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}\left(t, \mathbf{x}(t), \mathbf{x}_t(-\tau_1), \ldots, \mathbf{x}_t(-\tau_d)\right), \quad t \in [t_0, T], \\ &\mathbf{x}_{t_o}(s) = \varphi(s), \quad s \in [-\tau, 0], \end{aligned}$$

where $\varphi : [-\tau, 0] \longrightarrow \mathbb{R}^m$ is a given initial function with $\tau = \max_{i=1,\ldots,d}\{\tau_i\}$, and $\tau_i > 0$, $i = 1, \ldots, d$, are constant delays. $\mathbf{x}_t$ is a segment function defined as at the beginning of section 2.

By following the same ideas of the previous subsections, the definitions of the LL discretization and LL approximation are easily extended to equations with multiple delays. In this case, the expressions (2.9) and (2.10) are also obtained but with Φ defined by the form

$$\begin{aligned} \Phi(t_n, \mathbf{y}_n, h_n; \widetilde{\mathbf{y}}^1_{t_n}, \ldots, \widetilde{\mathbf{y}}^d_{t_n}) = &\int_0^{h_n} e^{\mathbf{A}_n(h_n-u)}\left(\sum_{i=1}^{d}\mathbf{B}^i_n(\widetilde{\mathbf{y}}^i_{t_n}(u-\tau_i) - \widetilde{\mathbf{y}}^i_{t_n}(-\tau_i)) + \mathbf{d}_n\right)du \\ &+ \int_0^{h_n}\int_0^{u} e^{\mathbf{A}_n(h_n-u)}\mathbf{c}_n dr du, \end{aligned} \tag{2.16}$$

where $\widetilde{\mathbf{y}}^i_{t_n} : [-\tau_i, 0] \longrightarrow \mathbb{R}^m$ is the segment function defined by

$$\widetilde{\mathbf{y}}^i_{t_n}(s) := \widetilde{\mathbf{y}}^i(t_n + s), \quad s \in [-\tau_i, 0],$$

and $\widetilde{\mathbf{y}}^i : [t_n - \tau_i, t_n] \longrightarrow \mathbb{R}^m$ is a suitable approximation to $\mathbf{x}(t)$ for all $t \in [t_n - \tau_i, t_n]$ such that $\widetilde{\mathbf{y}}^i(t_n) = \mathbf{y}_n$. In expression (2.16), $t_n, t_{n+1} \in (t)_h$, $h_n = t_{n+1} - t_n$,

$$\mathbf{A}_n = \mathbf{f}_x(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}^1_{t_n}(-\tau_1), \ldots, \widetilde{\mathbf{y}}^d_{t_n}(-\tau_d)),\ \mathbf{B}^i_n = \mathbf{f}_{x_t(-\tau_i)}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}^1_{t_n}(-\tau_1), \ldots, \widetilde{\mathbf{y}}^d_{t_n}(-\tau_d))$$

are constant matrices and

$$\mathbf{c}_n = \mathbf{f}_t(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}^1_{t_n}(-\tau_1), \ldots, \widetilde{\mathbf{y}}^d_{t_n}(-\tau_d)), \quad \mathbf{d}_n = \mathbf{f}(t_n, \mathbf{y}_n, \widetilde{\mathbf{y}}^1_{t_n}(-\tau_1), \ldots, \widetilde{\mathbf{y}}^d_{t_n}(-\tau_d))$$

are constant vectors. $\mathbf{f}_t$, $\mathbf{f}_x$, and $\mathbf{f}_{x_t(-\tau_i)}$ denote, respectively, the partial derivatives of $\mathbf{f}$ with respect to the variables t, $\mathbf{x}$, and $\mathbf{x}_t(-\tau_i)$.

In this way, the LL schemes proposed in the subsections above are easily extended to DDEs with multiple delays.

3. Convergence analysis. For the sake of simplicity, in this section it is assumed that there is a single delay. Suppose also that the initial function φ in (2.2) satisfies the boundedness and Lipschitz conditions

$$\sup_{-\tau\le s\le 0}\|\varphi(s)\|\le M_0 \tag{3.1}$$

and

$$\|\varphi(s_2)-\varphi(s_1)\|\le M_1(s_2-s_1) \tag{3.2}$$

for $-\tau\le s_1\le s_2\le 0$; and the function $\mathbf{f}$ in (2.1) and its first partial derivatives satisfy the Lipschitz conditions

$$\|\mathbf{f}(t,\mathbf{u}_1,\mathbf{v}_1)-\mathbf{f}(t,\mathbf{u}_2,\mathbf{v}_2)\|\le\lambda_0(\|\mathbf{u}_1-\mathbf{u}_2\|+\|\mathbf{v}_1-\mathbf{v}_2\|), \tag{3.3}$$

$$\|\mathbf{f}_{\mathbf{x}}(t,\mathbf{u}_1,\mathbf{v}_1)-\mathbf{f}_{\mathbf{x}}(t,\mathbf{u}_2,\mathbf{v}_2)\|\le\lambda_1(\|\mathbf{u}_1-\mathbf{u}_2\|+\|\mathbf{v}_1-\mathbf{v}_2\|), \tag{3.4}$$

$$\|\mathbf{f}_{\mathbf{x}_t}(t,\mathbf{u}_1,\mathbf{v}_1)-\mathbf{f}_{\mathbf{x}_t}(t,\mathbf{u}_2,\mathbf{v}_2)\|\le\lambda_2(\|\mathbf{u}_1-\mathbf{u}_2\|+\|\mathbf{v}_1-\mathbf{v}_2\|), \tag{3.5}$$

$$\|\mathbf{f}_t(t,\mathbf{u}_1,\mathbf{v}_1)-\mathbf{f}_t(t,\mathbf{u}_2,\mathbf{v}_2)\|\le\lambda_3(\|\mathbf{u}_1-\mathbf{u}_2\|+\|\mathbf{v}_1-\mathbf{v}_2\|) \tag{3.6}$$

for all $\mathbf{u}_1,\mathbf{u}_2,\mathbf{v}_1,\mathbf{v}_2\in\mathbb{R}^m$, $t\in\mathbb{R}$. Further, suppose that $\mathbf{f}$ and its first and second partial derivatives satisfy the linear growth and boundedness conditions

$$\|\mathbf{f}(t,\mathbf{u},\mathbf{v})\|+\|\mathbf{f}_t(t,\mathbf{u},\mathbf{v})\|\le K_0(1+\|\mathbf{u}\|+\|\mathbf{v}\|), \tag{3.7}$$

$$\|\mathbf{f}_{\mathbf{x}}(t,\mathbf{u},\mathbf{v})\|+\|\mathbf{f}_{\mathbf{x}_t}(t,\mathbf{u},\mathbf{v})\|\le K_1, \tag{3.8}$$

and

$$\begin{aligned}&\|f_{\mathbf{xx}}(t,\mathbf{u},\mathbf{v})\|+\|f_{\mathbf{xx}_t}(t,\mathbf{u},\mathbf{v})\|+\|f_{\mathbf{x}_t\mathbf{x}_t}(t,\mathbf{u},\mathbf{v})\|\\&+\|f_{tt}(t,\mathbf{u},\mathbf{v})\|+\|f_{t\mathbf{x}}(t,\mathbf{u},\mathbf{v})\|+\|f_{t\mathbf{x}_t}(t,\mathbf{u},\mathbf{v})\|\le K_2\end{aligned} \tag{3.9}$$

for all $\mathbf{u},\mathbf{v}\in\mathbb{R}^m$, $t\in\mathbb{R}$.

3.1. Local truncation error. In this subsection, the local truncation error of the LL discretization shall be derived. With that proposal the next two lemmas shall be used. The first one establishes a uniform bound and Lipschitz condition for the solution of the DDE, whereas the second one states a Lipschitz-type condition for the function Φ with respect to its second and fourth arguments.

LEMMA 3.1. *Assuming that conditions* (3.1), (3.2), *and* (3.7) *hold, there exist positive constants* C_0 *and* C_1 *such that*

$$\sup_{t_0-\tau\le t\le T}\|\mathbf{x}(t)\|\le C_0 \tag{3.10}$$

and

$$\|\mathbf{x}(s_2)-\mathbf{x}(s_1)\|\le C_1(s_2-s_1) \tag{3.11}$$

hold for $t_0-\tau\le s_1\le s_2\le T$, *where* $\mathbf{x}$ *is the solution of* (2.1)–(2.2).

Proof. From the integral form of (2.1)–(2.2) it follows that

$$\|\mathbf{x}(t)\|\le\|\varphi(0)\|+\int_{t_0}^{t}\|\mathbf{f}\left(s,\mathbf{x}\left(s\right),\mathbf{x}_s(-\tau)\right)\|\,ds,\quad t\ge t_0,$$

which by conditions (3.1) and (3.7) leads to

$$\|\mathbf{x}(t)\| \leq M_0 + \int_{t_0}^{t} K_0(1 + \|\mathbf{x}(s)\| + \|\mathbf{x}_s(-\tau)\|)ds.$$

Hence

$$\sup_{t_0-\tau\leq u\leq t} \|\mathbf{x}(u)\| \leq M_0 + \int_{t_0}^{t} K_0\left(1 + 2\sup_{t_0-\tau\leq u\leq s} \|\mathbf{x}(u)\|\right)ds,$$

and (3.10) follows from the Gronwall inequality.

On the other hand, for $t_0 - \tau \leq s_1 \leq s_2 \leq t_0$ inequality (3.11) follows from condition (3.2), whereas for $t_0 \leq s_1 \leq s_2 \leq T$ we obtain

$$\begin{aligned}\|\mathbf{x}(s_2) - \mathbf{x}(s_1)\| &\leq \int_{s_1}^{s_2} \|\mathbf{f}(s, \mathbf{x}(s), \mathbf{x}_s(-\tau))\| ds \\ &\leq (s_2 - s_1)K_0\left(1 + \sup_{s_1\leq s\leq s_2} (\|\mathbf{x}(s)\| + \|\mathbf{x}_s(-\tau)\|)\right),\end{aligned}$$

which by (3.10) gives

$$\|\mathbf{x}(s_2) - \mathbf{x}(s_1)\| \leq K_0(1 + 2C_0)(s_2 - s_1),$$

and so we conclude the proof. □

LEMMA 3.2. *Let $t_n, t_{n+1} \in (t)_h$ and $h_n = t_{n+1} - t_n$. Under conditions* (3.1)–(3.8)*, there exists a positive constant P such that*

$$\begin{aligned}&\|\Phi(t_n, \mathbf{v}, h_n; \mathbf{z}_{t_n}) - \Phi(t_n, \mathbf{x}(t_n), h_n; \mathbf{x}_{t_n})\| \\ &\quad\leq h_n P\left(\|\mathbf{v} - \mathbf{x}(t_n)\| + \sup_{s\in[0,h_n]} \|\mathbf{z}_{t_n}(s-\tau) - \mathbf{x}_{t_n}(s-\tau)\|\right),\end{aligned}$$

where $\mathbf{x}$ is the solution of (2.1)–(2.2)*, Φ is defined as in* (2.8)*, $\mathbf{v} \in \mathbb{R}^m$,$\mathbf{z}_{t_n} : [-\tau, 0] \longrightarrow \mathbb{R}^m$ is a segment function defined as*

$$\mathbf{z}_{t_n}(s) := \mathbf{z}(t_n + s),\ \ s \in [-\tau, 0],$$

and $\mathbf{z} : [t_n - \tau, t_n] \longrightarrow \mathbb{R}^m$ is a function.

Proof. Define

$$R_n := \Phi(t_n, \mathbf{v}, h_n; \mathbf{z}_{t_n}) - \Phi(t_n, \mathbf{x}(t_n), h_n; \mathbf{x}_{t_n}),$$

which in turn can be written as

$$\begin{aligned}R_n = &\int_0^{h_n} e^{\mathbf{A}_n(h_n-u)}(\mathbf{B}_n(\mathbf{z}_{t_n}(u-\tau) - \mathbf{z}_{t_n}(-\tau)) + u\mathbf{c}_n + \mathbf{d}_n)du \\ &- \int_0^{h_n} e^{\overline{\mathbf{A}}_n(h_n-u)}(\overline{\mathbf{B}}_n(\mathbf{x}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(-\tau)) + u\overline{\mathbf{c}}_n + \overline{\mathbf{d}}_n)du,\end{aligned}$$

where $\mathbf{A}_n = \mathbf{f}_{\mathbf{x}}(t_n, \mathbf{v}, \mathbf{z}_{t_n}(-\tau))$, $\mathbf{B}_n = \mathbf{f}_{\mathbf{x}_t}(t_n, \mathbf{v}, \mathbf{z}_{t_n}(-\tau))$, $\overline{\mathbf{A}}_n = \mathbf{f}_{\mathbf{x}}(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$, $\overline{\mathbf{B}}_n = \mathbf{f}_{\mathbf{x}_t}(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$ are constant matrices and $\mathbf{c}_n = \mathbf{f}_t(t_n, \mathbf{v}, \mathbf{z}_{t_n}(-\tau))$, $\mathbf{d}_n = \mathbf{f}(t_n, \mathbf{v}, \mathbf{z}_{t_n}(-\tau))$, $\overline{\mathbf{c}}_n = \mathbf{f}_t(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$, $\overline{\mathbf{d}}_n = \mathbf{f}(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$ are constant vectors.

By using Lemma 6.3 in the appendix we obtain

$$\begin{aligned}\|R_n\| \leq & \int_0^{h_n} \left\| e^{\mathbf{A}_n(h_n-u)} - e^{\overline{\mathbf{A}}_n(h_n-u)} \right\| \left(\left\| \overline{\mathbf{B}}_n \right\| \|\mathbf{x}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(-\tau)\| + u \|\overline{\mathbf{c}}_n\| + \left\| \overline{\mathbf{d}}_n \right\| \right) du \\ & + \int_0^{h_n} \left\| e^{\mathbf{A}_n(h_n-u)} \right\| \left(\left\| \mathbf{B}_n \mathbf{z}_{t_n}(-\tau) - \overline{\mathbf{B}}_n \mathbf{x}_{t_n}(-\tau) \right\| + u \|\mathbf{c}_n - \overline{\mathbf{c}}_n\| + \left\| \mathbf{d}_n - \overline{\mathbf{d}}_n \right\| \right. \\ & \left. + \left\| \mathbf{B}_n \mathbf{z}_{t_n}(u-\tau) - \overline{\mathbf{B}}_n \mathbf{x}_{t_n}(u-\tau) \right\| \right) du.\end{aligned}$$

From the finite increments inequality, conditions (3.8), (3.4), and constraint (2.3) it is follows that

$$\begin{aligned}\left\| e^{\mathbf{A}_n(h_n-u)} - e^{\overline{\mathbf{A}}_n(h_n-u)} \right\| &\leq e^{K_1 h_n} h_n \left\| \mathbf{A}_n - \overline{\mathbf{A}}_n \right\| \\ &\leq e^{K_1} \lambda_1 h_n \Gamma_n,\end{aligned}$$

where

$$\Gamma_n = \|\mathbf{v} - \mathbf{x}(t_n)\| + \sup_{u \in [0,h_n]} \|\mathbf{z}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(u-\tau)\|.$$

In addition, from Lemmas 6.3 and 3.1 and conditions (3.5) and (3.8) it is follows that

$$\begin{aligned}\left\| \mathbf{B}_n \mathbf{z}_{t_n}(u-\tau) - \overline{\mathbf{B}}_n \mathbf{x}_{t_n}(u-\tau) \right\| &\leq \|\mathbf{B}_n\| \|\mathbf{z}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(u-\tau)\| \\ &\quad + \left\| \mathbf{B}_n - \overline{\mathbf{B}}_n \right\| \|\mathbf{x}_{t_n}(u-\tau)\| \\ &\leq (K_1 + \lambda_2 C_0)\, \Gamma_n\end{aligned}$$

for all $u \in [0, h_n]$. By using the two previous inequalities, Lemma 3.1, and conditions (3.3), (3.6), (3.7), and (3.8) we obtain

$$\begin{aligned}\|R_n\| &\leq \Gamma_n \int_0^{h_n} e^{K_1} \lambda_1 h_n (2K_1 C_0 + K_0(1+2C_0)u + K_0(1+2C_0)) du \\ &\quad + \Gamma_n \int_0^{h_n} e^{K_1} (2(K_1 + \lambda_2 C_0) + \lambda_0 + \lambda_3 u) du \\ &\leq h_n P\, \Gamma_n,\end{aligned}$$

where $P = 2e^{K_1}\lambda_1 (K_1 C_0 + K_0(1+2C_0)) + e^{K_1}(2(K_1 + \lambda_2 C_0) + \lambda_0 + \lambda_3)$. Constraint (2.3) has also been used to obtain P. □

Let us denote by L_{n+1} the local truncation error of the LL discretization at t_{n+1}, i.e.,

$$L_{n+1} = \|\mathbf{x}(t_{n+1}) - \mathbf{x}(t_n) - \Phi(t_n, \mathbf{x}(t_n), h_n; \widetilde{\mathbf{y}}_{t_n})\|, \tag{3.12}$$

where $t_n, t_{n+1} \in (t)_h$, $h_n = t_{n+1} - t_n$, $\mathbf{x}$ is the solution of (2.1)–(2.2), and $\Phi, \widetilde{\mathbf{y}}_{t_n}$ are defined as in Definition 2.1.

THEOREM 3.3. *Suppose that conditions* (3.1)–(3.9) *hold. Then*

$$L_{n+1} \leq L h_n^3 + P h_n \sup_{s \in [0,h_n]} \|\widetilde{\mathbf{y}}_{t_n}(s-\tau) - \mathbf{x}_{t_n}(s-\tau)\|,$$

where L and P are positive constants.

Proof. Let $\mathbf{u}$ and $\mathbf{v}$ be the solutions of the nonautonomous ODEs

$$\begin{aligned} \frac{d\mathbf{u}(t)}{dt} &= \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}_t(-\tau)), \quad t \in [t_n, t_{n+1}], \\ \mathbf{u}(t_n) &= \mathbf{x}(t_n), \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} \frac{d\mathbf{v}(t)}{dt} &= \mathbf{g}(t, \mathbf{v}(t), \mathbf{x}_t(-\tau)), \quad t \in [t_n, t_{n+1}], \\ \mathbf{v}(t_n) &= \mathbf{x}(t_n), \end{aligned}$$

respectively, where function $\mathbf{g}$ is the first order Taylor expansion of the function $\mathbf{f}$ around the point $(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$. That is,

$$\mathbf{g}(t, \mathbf{v}(t), \mathbf{x}_t(-\tau)) = \mathbf{A}_n \left(\mathbf{v}\left(t\right) - \mathbf{x}\left(t_n\right)\right) + \mathbf{B}_n \left(\mathbf{x}_t\left(-\tau\right) - \mathbf{x}_{t_n}\left(-\tau\right)\right) + \mathbf{c}_n(t - t_n) + \mathbf{d}_n,$$

where $\mathbf{A}_n = \mathbf{f}_{\mathbf{x}}(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$, $\mathbf{B}_n = \mathbf{f}_{\mathbf{x}_t}(t_n, \mathbf{x}\left(t_n\right), \mathbf{x}_{t_n}\left(-\tau\right))$ are constant matrices and $\mathbf{c}_n = \mathbf{f}_t(t_n, \mathbf{x}\left(t_n\right), \mathbf{x}_{t_n}\left(-\tau\right))$, $\mathbf{d}_n = \mathbf{f}(t_n, \mathbf{x}(t_n), \mathbf{x}_{t_n}(-\tau))$ are constant vectors. In addition, let

$$\begin{aligned} \varepsilon &:= \sup_{t \in [t_n, t_{n+1}]} \left\| \frac{d\mathbf{v}(t)}{dt} - \mathbf{f}(t, \mathbf{v}(t), \mathbf{x}_t(-\tau)) \right\| \\ &= \sup_{t \in [t_n, t_{n+1}]} \left\| \mathbf{g}(t, \mathbf{v}(t), \mathbf{x}_t(-\tau)) - \mathbf{f}(t, \mathbf{v}(t), \mathbf{x}_t(-\tau)) \right\|. \end{aligned} \tag{3.14}$$

By applying the Taylor formulae with Lagrange remainder for functions defined on a Banach space [12] and condition (3.9) we obtain

$$\varepsilon \leq \frac{3}{2} K_2 \sup_{t \in [t_n, t_{n+1}]} (\|t - t_n\|^2 + \|\mathbf{v}(t) - \mathbf{v}\left(t_n\right)\|^2 + \|\mathbf{x}_t(-\tau) - \mathbf{x}_{t_n}(-\tau)\|^2).$$

Moreover, by using Lemma 3.1, conditions (3.7)–(3.8), and constraint (2.3) we obtain

$$\begin{aligned} \|\mathbf{v}(t) - \mathbf{v}\left(t_n\right)\| &= \|\Phi(t_n, \mathbf{v}\left(t_n\right), t - t_n; \mathbf{x}_{t_n})\| \\ &\leq \int_0^{t-t_n} \left\| e^{\mathbf{A}_n(t-t_n-u)} \right\| (\|\mathbf{d}_n\| + \|\mathbf{c}_n\| u) du \\ &\quad + \int_0^{t-t_n} \left\| e^{\mathbf{A}_n(t-t_n-u)} \right\| \|\mathbf{B}_n\| \, \|\mathbf{x}_{t_n}\left(u - \tau\right) - \mathbf{x}_{t_n}\left(-\tau\right)\| \, du \\ &\leq C_2 h_n, \end{aligned}$$

where $C_2 = 2e^{K_1}(K_0(1 + 2C_0) + K_1 C_0)$. However, Lemma 3.1 also implies that

$$\|\mathbf{x}_t(-\tau) - \mathbf{x}_{t_n}(-\tau)\| \leq C_1 h_n.$$

Therefore,

$$\varepsilon \leq \frac{3}{2} K_2 (1 + C_2^2 + C_1^2) h_n^2.$$

Now, by applying Lemma 6.2 to the functions $\mathbf{u}$ and $\mathbf{v}$ (i.e., by using (3.13) and (3.14) for the first and second differential inequality in that lemma, respectively) it is obtained that

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq \frac{\varepsilon}{\lambda_0} (e^{\lambda_0 (t - t_n)} - 1)$$

for $t \in [t_n, t_{n+1}]$. Moreover, from the mean value theorem it follows that

$$(e^{\lambda_0(t-t_n)} - 1) \leq \lambda_0 e^{\lambda_0(t-t_n)}(t - t_n),$$

which implies that

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq \varepsilon e^{\lambda_0(t-t_n)}(t - t_n).$$

Taking into account that $\mathbf{u} \equiv \mathbf{x}$ in $[t_n, t_{n+1}]$ and that

$$\mathbf{v}(t) \equiv \mathbf{x}(t_n) + \Phi\left(t_n, \mathbf{x}\left(t_n\right), t - t_n; \mathbf{x}_{t_n}\right)$$

we obtain

$$\begin{aligned} l_{n+1} &= \|\mathbf{x}\left(t_{n+1}\right) - \mathbf{x}\left(t_n\right) - \Phi(t_n, \mathbf{x}\left(t_n\right), h_n; \mathbf{x}_{t_n})\| \\ &\leq L h_n^3, \end{aligned}$$

with $L = \frac{3}{2}K_2(1 + C_2^2 + C_1^2)e^{\lambda_0}$. Constraint (2.3) has been again used to obtain L.

The proof is completed by applying Lemma 3.2 to the second term of the inequality

$$L_{n+1} \leq l_{n+1} + \|\Phi(t_n, \mathbf{x}\left(t_n\right), h_n; \widetilde{\mathbf{y}}_{t_n}) - \Phi(t_n, \mathbf{x}\left(t_n\right), h_n; \mathbf{x}_{t_n})\|. \qquad \square$$

3.2. Uniform convergence. The main result of this subsection is stated in the following theorem.

THEOREM 3.4. *For $t \in [t_0, T]$, let $\mathbf{y}_t : [-\tau, 0] \longrightarrow \mathbb{R}^m$ be the segment function defined as*

$$\mathbf{y}_t(s) := \mathbf{y}(t + s), \quad s \in [-\tau, 0],$$

where $\mathbf{y}(t + s)$ is the LL approximation of the DDE (2.1)–(2.2) at the point $t + s$. Further, let $\widetilde{\mathbf{y}}_{t_n} : [-\tau, 0] \longrightarrow \mathbb{R}^m$ be a segment function defined as

$$\widetilde{\mathbf{y}}_{t_n}(s) := \widetilde{\mathbf{y}}(t_n + s), \quad s \in [-\tau, 0],$$

where $\widetilde{\mathbf{y}}(t_n + s)$ is an approximation of $\mathbf{y}(t_n + s)$ such that

$$\sup_{u \in [0, h_n]} \|\mathbf{y}_{t_n}(u - \tau) - \widetilde{\mathbf{y}}_{t_n}(u - \tau)\| \leq C_r h_n^r \tag{3.15}$$

and $\widetilde{\mathbf{y}}_{t_n}(0) = \mathbf{y}(t_n)$ for all $t_n \in (t)_h$, with $C_r > 0$ and $r \in \mathbb{N}_+$. Then, under conditions (3.1)–(3.9), there exists a positive constant M such that

$$\|\mathbf{x}\left(t\right) - \mathbf{y}\left(t\right)\| \leq M h^{\min\{2,r\}}$$

for every $t \in [t_0 - \tau, T]$, where $\mathbf{x}$ is the solution of (2.1)–(2.2).

Proof. Suppose that the numerical integration has reached t_n and let E_n be a uniform bound on $\|\mathbf{x}\left(t\right) - \mathbf{y}\left(t\right)\|$ for every $t \in [t_0 - \tau, t_n]$.

By definition of LL approximation

$$\begin{aligned} \mathbf{x}\left(t\right) - \mathbf{y}\left(t\right) = \mathbf{x}(t) - \mathbf{y}\left(t_n\right) - \Phi(t_n, \mathbf{y}\left(t_n\right), t - t_n; \widetilde{\mathbf{y}}_{t_n}) + \mathbf{x}(t_n) - \mathbf{x}\left(t_n\right) \\ + \Phi(t_n, \mathbf{x}\left(t_n\right), t - t_n; \widetilde{\mathbf{y}}_{t_n}) - \Phi(t_n, \mathbf{x}\left(t_n\right), t - t_n; \widetilde{\mathbf{y}}_{t_n}) \end{aligned}$$

for $t \in [t_n, t_{n+1}]$, and thus

$$\|\mathbf{x}(t) - \mathbf{y}(t)\| \leq \|\mathbf{x}(t_n) - \mathbf{y}(t_n)\| + L_{n+1} + R_{n+1}, \tag{3.16}$$

where L_{n+1} denotes the local truncation error (3.12) and

$$R_{n+1} = \|\Phi(t_n, \mathbf{x}(t_n), t - t_n; \widetilde{\mathbf{y}}_{t_n}) - \Phi(t_n, \mathbf{y}(t_n), t - t_n; \widetilde{\mathbf{y}}_{t_n})\|.$$

Taking into account that

$$\begin{aligned} \sup_{u\in[0,h_n]} \|\widetilde{\mathbf{y}}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(u-\tau)\| &\leq \sup_{u\in[0,h_n]} \|\mathbf{x}_{t_n}(u-\tau) - \mathbf{y}_{t_n}(u-\tau)\| \\ &\quad + \sup_{u\in[0,h_n]} \|\mathbf{y}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(u-\tau)\| \\ &\leq E_n + C_r h_n^r \end{aligned} \tag{3.17}$$

we obtain

$$L_{n+1} \leq L h_n^3 + h_n P\,(E_n + C_r h_n^r) \tag{3.18}$$

and

$$\begin{aligned} R_{n+1} &\leq \|\Phi(t_n, \mathbf{x}(t_n), t - t_n; \widetilde{\mathbf{y}}_{t_n}) - \Phi(t_n, \mathbf{x}(t_n), t - t_n; \mathbf{x}_{t_n})\| \\ &\quad + \|\Phi(t_n, \mathbf{x}(t_n), t - t_n; \mathbf{x}_{t_n}) - \Phi(t_n, \mathbf{y}(t_n), t - t_n; \widetilde{\mathbf{y}}_{t_n})\| \\ &\leq h_n P \left\{ \|\mathbf{x}(t_n) - \mathbf{y}(t_n)\| + 2 \sup_{u\in[0,h_n]} \|\widetilde{\mathbf{y}}_{t_n}(u-\tau) - \mathbf{x}_{t_n}(u-\tau)\| \right\} \\ &\leq h_n P(3E_n + 2C_r h_n^r) \end{aligned} \tag{3.19}$$

for $t \in [t_n, t_{n+1}]$. Here, (3.15), Theorem 3.3, and Lemma 3.2 were used to derive (3.17), (3.18), and (3.19).

Thus, from (3.16), (3.18), and (3.19) and constraint (2.3) we obtain

$$\|\mathbf{x}(t) - \mathbf{y}(t)\| \leq (1 + hP_1)\, E_n + L_1 h^{\min\{3, r+1\}},$$

where $P_1 = 4P$ and $L_1 = L + 3PC_r$. Note that this expression gives an error bound for all $t \in [t_n, t_{n+1}]$. Therefore, by definition of E_n, that bound also holds for $t \in [t_0, t_n]$. Thus,

$$E_{n+1} \leq (1 + hP_1)\, E_n + L_1 h^{\min\{3, r+1\}}.$$

Finally, by induction, the last inequality implies that

$$\begin{aligned} E_{n+1} &\leq \frac{((1 + hP_1)^{n+1} - 1)}{hP_1}\; L_1 h^{\min\{3, r+1\}} \\ &\leq M\; h^{\min\{2, r\}}, \end{aligned}$$

where $M = L_1(e^{P_1(T-t_0)} - 1)/P_1$. This completes the proof. □

Thus, we are ready to analyze the convergence rate of the two LL schemes introduced in the previous section. Denote by $\mathcal{C}^r[-\tau, 0]$ the class of functions on $[-\tau, 0]$ with continuous derivatives up to order r, and by $\overline{\mathcal{C}}^r[-\tau, 0]$ the class of functions with continuous r-derivatives everywhere in $[-\tau, 0]$ except at a finite number of points.

For the natural LL scheme it is obvious that $\max_{u\in[0,h_n]} \|\mathbf{y}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(u-\tau)\| = 0$ for all $t_n \in (t)_h$ whenever the initial condition φ is either a piecewise polynomial or an exponential function. However, if $\varphi \in \mathcal{C}^r[-\tau, 0]$, the function φ can be approximated by a piecewise interpolating polynomial of order r and so condition (3.15) holds for all $t_n \in (t)_h \cap [t_0, t_0+\tau]$, and $\max_{u\in[0,h_n]} \|\mathbf{y}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(u-\tau)\| = 0$ for all $t_n \in (t)_h \cap [t_0+\tau, T]$. Thus, in the first case the order of convergence of the natural LL scheme is two, while in the second case it is $\min\{2, r\}$.

However, for the polynomial LL schemes, the convergence analysis is not so simple. Note that the first derivative of the LL approximation $\mathbf{y}$ satisfies (2.4) for each $t \in [t_n, t_{n+1})$, and thus its rth derivative is not continuous at all of the points $t_n \in (t)_h$ for all $r \in \mathbb{N}_+$. That is, $\mathbf{y}_t \in \overline{\mathcal{C}}^r[-\tau, 0]$. Therefore, the conventional results from the approximation theory are not straightforward applicable and additional results are needed. For example, the next theorem deals with the case of interpolating polynomials for $\mathbf{y}$.

THEOREM 3.5. *Let $\widetilde{\mathbf{y}}$ be the order r polynomial that interpolate $\mathbf{y}$ in r points s_i on $[t_n-\tau, t_{n+1}-\tau]$ for each $t_n \in (t)_h$. Then*

$$\sup_{u\in[0,h_n]} \|\mathbf{y}_{t_n}(u-\tau) - \widetilde{\mathbf{y}}_{t_n}(u-\tau)\| \le C_{r,n} h_n^r,$$

where

$$C_{r,n} = D_r \left(1 + \sup_{u\in[0,h_n]} \|\lambda_r(u+t_n-\tau)\|\right) \sup_{u\in[0,h_n]} \left\| \frac{d^r \mathbf{y}_{t_n}}{ds^r}(u-\tau) \right\|.$$

Here, $\lambda_r(t) = \sum_{i=1}^r |\prod_{j\neq i}(t-s_j)/\prod_{j\neq i}(s_i-s_j)|$ is the well-known Lebesgue function [14], *$\frac{d^r \mathbf{y}_{t_n}}{ds^r}(u-\tau)$ denotes the rth derivative of $\mathbf{y}_{t_n}$ evaluated at $(u-\tau)$, and D_r is a positive constant depending only on r.*

Proof. Let $\mathcal{P}_r$ be the space of polynomials of order r on $[t_n-\tau, t_{n+1}-\tau]$. Then, by Theorem XII.5 in [14] for the polynomial approximation of functions with continuous derivatives on a bounded interval except at a finite number of points, there exists $\mathbf{p} \in \mathcal{P}_r$ such that

$$\|\mathbf{y}-\mathbf{p}\|_\infty \le D_r \left\| \frac{d^r \mathbf{y}}{dt^r} \right\|_\infty (t_{n+1}-t_n)^r, \tag{3.20}$$

where $\|.\|_\infty$ denotes the uniform norm on $[t_n-\tau, t_{n+1}-\tau]$ and D_r is a positive constant depending only on r.

Let $I_r\mathbf{y}(t)$ be an order r polynomial that interpolates $\mathbf{y}$ in r points s_i on $[t_n-\tau, t_{n+1}-\tau]$. Taking into account the Lagrange form

$$I_r\mathbf{y}(t) = \sum_{i=1}^r \mathbf{y}(s_i) l_i(t), \text{ with } l_i(t) = \prod_{j\neq i}(t-s_j)/\prod_{j\neq i}(s_i-s_j)$$

of that polynomial, it is follows that

$$\|I_r\mathbf{y}\|_\infty \le \|\mathbf{y}\|_\infty \|\lambda_r\|_\infty, \tag{3.21}$$

where $\lambda_r(t) = \sum_{i=1}^r |l_i(t)|$ is the Lebesgue function.

Now, by taking into account that $I_r\mathbf{p} = \mathbf{p}$ for all $\mathbf{p} \in \mathcal{P}_r$, and using the inequalities (3.20) and (3.21) we obtain

$$\begin{aligned} \|\mathbf{y} - I_r\mathbf{y}\|_\infty &\le \|\mathbf{y}-\mathbf{p}\|_\infty + \|I_r(\mathbf{p}-\mathbf{y})\|_\infty \\ &\le (1+\|\lambda_r\|_\infty)\|\mathbf{p}-\mathbf{y}\|_\infty \\ &\le C_{r,n} h_n^r. \end{aligned}$$

The proof is completed by noting that $\widetilde{\mathbf{y}} \equiv I_r \mathbf{y}$. ∎

In this way, if $C_{r,n}$ were bounded for all n, Theorems 3.4 and 3.5 would imply the order of convergence $\min\{2, r\}$ for the interpolating polynomial LL schemes. Accordingly, the polynomial LL scheme (2.15) with linear interpolation would provide the best performance with respect to the trade-off between convergence rate and computational cost. For this kind of LL approximation, the next theorem states an upper bound for its second derivative in such a way that condition (3.15) in Theorem 3.4 holds for $r = 2$.

THEOREM 3.6. *For all $t_n \in (t)_h$, let*

$$\widetilde{\mathbf{y}}_{t_n}(u - \tau) = \alpha_{0,n} + \alpha_{1,n} u, \quad \text{with } u \in [0, h_n],$$

be a piecewise linear interpolant of $\mathbf{y}$, where $\alpha_{0,n} = \mathbf{y}_{t_n}(-\tau)$ and $\alpha_{1,n} = (\mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau))/h_n$. Then

$$\sup_{u \in [0, h_n]} \|\lambda_2(u + t_n - \tau)\| = 1,$$

and, under conditions (3.7)–(3.8)*, we obtain*

$$\sup_{u \in [0, h_n]} \left\| \frac{d^2 \mathbf{y}_{t_n}}{ds^2}(u - \tau) \right\| \leq M,$$

where M is a constant independent of n.

Proof. By definition

$$\begin{aligned} \lambda_2(u + t_n - \tau) &= \left| \frac{u + t_n - t_{n+1}}{t_n - t_{n+1}} \right| + \left| \frac{u}{t_{n+1} - t_n} \right| \\ &= \frac{t_{n+1} - t_n - u}{t_{n+1} - t_n} + \frac{u}{t_{n+1} - t_n} \\ &= 1 \end{aligned}$$

for $u \in [0, h_n]$, which implies the first assertion of the theorem.

To prove the second one, the boundedness and Lipschitz condition for the LL approximation $\mathbf{y}$ shall be derived first and afterward bounds for the first and second derivatives of $\mathbf{y}$.

From the definition of LL approximation

$$\mathbf{y}(t) = \mathbf{y}(t_0) + \sum_{n=0}^{n_t - 1} \Phi(t_n, \mathbf{y}_n, h_n; \widetilde{\mathbf{y}}_{t_n}) + \Phi(t_{n_t}, \mathbf{y}_{n_t}, t - t_{n_t}; \widetilde{\mathbf{y}}_{t_{n_t}}),$$

where

$$\Phi(t_n, \mathbf{y}_n, t - t_n; \widetilde{\mathbf{y}}_{t_n}) = \int_0^{t - t_n} e^{\mathbf{A}_n (t - t_n - u)} \left(\mathbf{B}_n (\mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau)) \frac{u}{h_n} + \mathbf{d}_n + \mathbf{c}_n u \right) du$$

with $\mathbf{A}_n = \mathbf{f}_x(t_n, \mathbf{y}_n, \mathbf{y}_{t_n}(-\tau))$, $\mathbf{B}_n = \mathbf{f}_{x_t}(t_n, \mathbf{y}_n, \mathbf{y}_{t_n}(-\tau))$, $\mathbf{c}_n = \mathbf{f}_t(t_n, \mathbf{y}_n, \mathbf{y}_{t_n}(-\tau))$, and $\mathbf{d}_n = \mathbf{f}(t_n, \mathbf{y}_n, \mathbf{y}_{t_n}(-\tau))$ for all $t_n \in (t)_h$.

Now, by using conditions (3.7)–(3.8) and constraint (2.3) it follows that

$$\begin{aligned}
R_n(t) &= \|\Phi(t_n, \mathbf{y}_n, t-t_n; \widetilde{\mathbf{y}}_{t_n})\| \\
&\leq \int_0^{t-t_n} \left\| e^{\mathbf{A}_n(t-t_n-u)} \right\| \left(\|\mathbf{B}_n\| \left\| \mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau) \right\| \left\| \frac{u}{h_n} \right\| + \|\mathbf{d}_n\| + \|\mathbf{c}_n\| u \right) du \\
&\leq e^{K_1} K_1 \int_0^{t-t_n} \left\| \mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau) \right\| du \\
&\quad + e^{K_1} K_0 \int_0^{t-t_n} (1 + \|\mathbf{y}_n\| + \|\mathbf{y}_{t_n}(-\tau)\|)(1+u) du \\
&\leq 2e^{K_1} K_1 \int_{t_n}^{t} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du \\
&\quad + (1+h_n) e^{K_1} K_0 \int_{t_n}^{t} \left(1 + 2 \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \right) du. \qquad (3.22)
\end{aligned}$$

In this way,

$$\begin{aligned}
\|\mathbf{y}(t)\| &\leq \|\mathbf{y}(t_0)\| + \sum_{n=0}^{n_t-1} R_n(h_n) + R_{n_t}(t-t_{n_t}) \\
&\leq \|\mathbf{y}(t_0)\| + 2e^{K_1} \sum_{n=0}^{n_t-1} \left(K_1 \int_{t_n}^{t_{n+1}} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du \right. \\
&\quad \left. + K_0 \int_{t_n}^{t_{n+1}} (1 + 2 \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\|) du \right) + 2e^{K_1} \left(K_1 \int_{t_{n_t}}^{t} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du \right. \\
&\quad \left. + K_0 \int_{t_{n_t}}^{t} (1 + 2 \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\|) du \right) \\
&\leq \|\mathbf{y}(t_0)\| + 2e^{K_1} \left(K_1 \int_{t_0}^{t} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du + K_0 \int_{t_0}^{t} (1 + 2 \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\|) du \right) \\
&\leq M_1 + M_2 \int_{t_0}^{t} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du,
\end{aligned}$$

where $M_1 = \|\mathbf{y}(t_0)\| + 2e^{K_1} K_0 (T - t_0)$ and $M_2 = 2e^{K_1}(K_1 + 2K_0)$. Therefore

$$\sup_{t_0-\tau \leq s \leq t} \|\mathbf{y}(s)\| \leq M_1 + M_2 \int_{t_0}^{t} \sup_{t_0-\tau \leq s \leq u} \|\mathbf{y}(s)\| \, du.$$

Now, by applying the Gronwall inequality we obtain

$$\sup_{t_0-\tau \leq t \leq T} \|\mathbf{y}(t)\| \leq M_3, \qquad (3.23)$$

where $M_3 = M_1 M_2 e^{M_2(T-t_0)}$ is a positive constant.

Let $s_2 = t_{n_{s_2}} + \Delta_2$ and $s_1 = t_{n_{s_1}} + \Delta_1$ be two points in $[t_0, T]$ such that $s_2 \geq s_1$. Thus,

$$\begin{aligned}\mathbf{y}(s_2) - \mathbf{y}(s_1) = {} & \Phi(t_{n_{s_1}}, \mathbf{y}_{n_{s_1}}, h_{n_{s_1}}; \widetilde{\mathbf{y}}_{t_{n_{s_1}}}) - \Phi(t_{n_{s_1}}, \mathbf{y}_{n_{s_1}}, \Delta_1; \widetilde{\mathbf{y}}_{t_{n_{s_1}}}) \\ & + \sum_{n=n_{s_1}+1}^{n_{s_2}-1} \Phi(t_n, \mathbf{y}_n, h_n; \widetilde{\mathbf{y}}_{t_n}) + \Phi(t_{n_{s_2}}, \mathbf{y}_{n_{s_2}}, \Delta_2; \widetilde{\mathbf{y}}_{t_{n_{s_2}}}).\end{aligned}$$

From (3.22) it is straightforward to obtain

$$\|\Phi(t_n, \mathbf{y}_n, h_n; \widetilde{\mathbf{y}}_{t_n})\| = M_4 h_n,$$

where $M_4 = 2e^{K_1}(K_1 M_3 + K_0(1 + 2M_3))$, while the inequality

$$\left\|\Phi(t_{n_{s_1}}, \mathbf{y}_{n_{s_1}}, h_{n_{s_1}}; \widetilde{\mathbf{y}}_{t_{n_{s_1}}}) - \Phi(t_{n_{s_1}}, \mathbf{y}_{n_{s_1}}, \Delta_1; \widetilde{\mathbf{y}}_{t_{n_{s_1}}})\right\| \leq M_4(h_{n_{s_1}} - \Delta_1)$$

can be derived by following the same steps used to obtain (3.22). Then

$$\begin{aligned}\|\mathbf{y}(s_2) - \mathbf{y}(s_1)\| &\leq M_4 \left(h_{n_{s_1}} - \Delta_1 + \sum_{n=n_{s_1}+1}^{n_{s_2}-1} h_n + \Delta_2 \right) \\ &\leq M_4 \left(t_{n_{s_1}+1} - t_{n_{s_1}} - \Delta_1 + \sum_{n=n_{s_1}+1}^{n_{s_2}-1} (t_{n+1} - t_n) + s_2 - t_{n_{s_2}} \right) \\ &\leq M_4(s_2 - s_1). \qquad (3.24)\end{aligned}$$

By definition, for each $t_n \in (t)_h$, the LL approximation $\mathbf{y}$ satisfies the linear DDE (2.4), which can be written in terms of the nonautonomous ODE

$$\begin{aligned}\frac{d\mathbf{z}\,(t)}{dt} &= \mathbf{g}^n(t, \mathbf{z}(t)), \quad t \in [t_n, t_{n+1}), \qquad (3.25) \\ \mathbf{z}(t_n) &= \widetilde{\mathbf{y}}_{t_n}(0),\end{aligned}$$

where

$$\begin{aligned}\mathbf{g}^n(t, \mathbf{z}(t)) = {} & \mathbf{A}_n \mathbf{z}(t) + \mathbf{B}_n(\mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau)) \frac{(t - t_n)}{h_n} \\ & + \mathbf{c}_n(t - t_n) + \mathbf{d}_n - \mathbf{A}_n \mathbf{y}_n.\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\frac{d\mathbf{z}\,(t)}{dt}\right\| &= \|\mathbf{g}^n(t, \mathbf{z}(t))\| \\ &\leq 2(\|\mathbf{A}_n\| + \|\mathbf{B}_n\|) \sup_{t_0 - \tau \leq t \leq T} \|\mathbf{y}\,(t)\| + \|\mathbf{c}_n h_n + \mathbf{d}_n\|\end{aligned}$$

for all $t \in [t_n, t_{n+1})$. However, from conditions (3.7)–(3.8) and constraint (2.3) it is

obtained that

$$\begin{aligned}\|\mathbf{c}_n h_n + \mathbf{d}_n\| &\leq (1+h_n)K_0\Big(1+2\sup_{t_0-\tau\leq t\leq T}\|\mathbf{y}\,(t)\|\Big)\\ &\leq 2K_0(1+2M_3)\end{aligned}$$

and $\|\mathbf{A}_n\|+\|\mathbf{B}_n\| \leq K_1$. Thus, by taking into account that $\mathbf{z}(t) \equiv \mathbf{y}(t)$ for $t \in [t_n, t_{n+1})$ and $t_n, t_{n+1} \in (t)_h$, it is obtained that

$$\sup_{t_n\leq t\leq t_{n+1}}\left\|\frac{d\mathbf{y}\,(t)}{dt}\right\| \leq M_5, \tag{3.26}$$

where $M_5 = 2K_1M_3 + 2K_0(1+2M_3)$ is a positive constant independent of n.

Now, by taking the derivative in (3.25) we obtain

$$\frac{d^2\mathbf{z}\,(t)}{dt^2} = \mathbf{g}_t^n(t,\mathbf{z}(t)) + \mathbf{g}_{\mathbf{z}}^n(t,\mathbf{z}(t))\frac{d\mathbf{z}\,(t)}{dt},$$

where $\mathbf{g}_t^n$ and $\mathbf{g}_{\mathbf{z}}^n$ denote the partial derivatives of $\mathbf{g}^n$ with respect to t and $\mathbf{z}$, respectively. Thus,

$$\left\|\frac{d^2\mathbf{z}\,(t)}{dt^2}\right\| \leq \|\mathbf{g}_t^n(t,\mathbf{z}(t))\| + \|\mathbf{g}_{\mathbf{z}}^n(t,\mathbf{z}(t))\|\left\|\frac{d\mathbf{z}\,(t)}{dt}\right\| \tag{3.27}$$

for $t \in [t_n, t_{n+1})$. From conditions (3.7)–(3.8) and (3.23)–(3.24) it follows that

$$\begin{aligned}\|\mathbf{g}_t^n(u,\mathbf{z}(u))\| &\leq \frac{1}{h_n}\|\mathbf{B}_n\|\left\|\mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau)\right\| + \|\mathbf{c}_n\|\\ &\leq \frac{1}{h_n}K_1\left\|\mathbf{y}_{t_{n+1}}(-\tau) - \mathbf{y}_{t_n}(-\tau)\right\| + K_0\Big(1+2\sup_{t_0-\tau\leq t\leq T}\|\mathbf{y}\,(t)\|\Big)\\ &\leq K_1M_4 + K_0(1+2M_3) \qquad (3.28)\end{aligned}$$

and

$$\|\mathbf{g}_{\mathbf{z}}^n(u,\mathbf{z}(u))\| \leq K_1 \tag{3.29}$$

for $u \in [t_n, t_{n+1}]$. Finally, by using (3.26), (3.28), and (3.29) in (3.27), and by taking into account that $\mathbf{z}(t) \equiv \mathbf{y}(t)$ for $t \in [t_n, t_{n+1})$ and $t_n, t_{n+1} \in (t)_h$, we obtain

$$\sup_{t_n\leq t\leq t_{n+1}}\left\|\frac{d^2\mathbf{y}\,(t)}{dt^2}\right\| \leq M_6,$$

where $M_6 = K_1(M_4+M_5) + K_0(1+2M_3)$ is a positive constant independent of n. Thus, the proof is complete. □

4. Simulation results. In this section the performance of the LL method is illustrated by means of numerical simulations. To do so, we shall use the polynomial LL scheme (2.15) with linear interpolation and fixed step-size h. Specifically, in each interval $[t_n - \tau_i, t_{n+1} - \tau_i]$, the LL approximation $\mathbf{y}$ is approximated by the function

$$\widetilde{\mathbf{y}}\left(t_n + s - \tau_i\right) = \mathbf{y}\left(t_n - \tau_i\right) + \alpha_{1,n}^i s$$

with $s \in [0, h]$ and $\alpha_{1,n}^i = \left(\mathbf{y}\left(t_{n+1} - \tau_i\right) - \mathbf{y}\left(t_n - \tau_i\right)\right)/h$ for each delay τ_i; and the LL scheme is defined by the iteration

$$\mathbf{y}_{t_{n+1}} = \mathbf{y}_{t_n} + \mathbf{h}(\mathbf{y}_{t_n}) \tag{4.1}$$

for all $t_n \in (t)_h$, where the vector $\mathbf{h}(\mathbf{y}_{t_n})$ is obtained from the expression

$$\begin{bmatrix} \mathbf{F} & \mathbf{g}_1 & \mathbf{h}(\mathbf{y}_{t_n}) \\ 0 & 1 & \mathbf{g}_2 \\ 0 & 0 & 1 \end{bmatrix} = e^{h\mathbf{T}_n}$$

with

$$\mathbf{T}_n = \begin{bmatrix} \mathbf{A}_n & \mathbf{c}_n + \sum_{i=1}^{d} \mathbf{B}_n^i \alpha_{1,n}^i & \mathbf{d}_n \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(m+2)\times(m+2)}.$$

Here, the matrix $e^{h\mathbf{T}_n}$ is computed by the rational Padé approximation with the "scaling and squaring" procedure (see Algorithm 11.3.1 in [18] for details).

Two DDEs with a variety of complexity were selected. This includes nonlinear equations with single and multiple delays, with low order discontinuities, and stiff equations.

Example 1. The first example is an epidemic model due to Cooke [31]. It describes the fraction of a population that is infected by a virus at time t through the equation

$$\frac{dx\left(t\right)}{dt} = bx\left(t-7\right)\left(1 - x\left(t\right)\right) - cx\left(t\right), \quad t \in [0, 70]. \tag{4.2}$$

Here b and c are positive constants. If $b > c$, then the solution $x\left(t\right) = 1 - c/b$ is an equilibrium point. The equation is integrated for $b = 2$, $c = 1$ and initial condition function $x\left(t\right) = 0.8$ for $t \in [-7, 0]$.

Figure 4.1 shows the numerical solution converging to the equilibrium point. Figure 4.2 shows the maximum errors of the LL scheme (4.1) in the computation of the solution of (4.2) in $(t)_h$ for $h = 0.1$, 0.01, 0.001 and the straight line that fits these points in the minimum square sense. The slope of that line is 2, which agrees with the theoretical estimate obtained in the previous section. For a "true" solution we used the trajectory obtained by the LL scheme with $h = 0.00001$. For a comparison, Figure 4.2 also shows the results obtained by the polynomial LL scheme with first and third order interpolating polynomial $\widetilde{\mathbf{y}}$, i.e., schemes of the form (2.15) with $p = 0$ and $p = 2$, respectively. In each case, the slope of their respective lines are 1 and 2, which also agrees with the theoretical estimate.

Example 2. The second example is the stiff DDE proposed in [6] to describe the dynamic of an antiviral immune response. The disease dynamic is governed by the

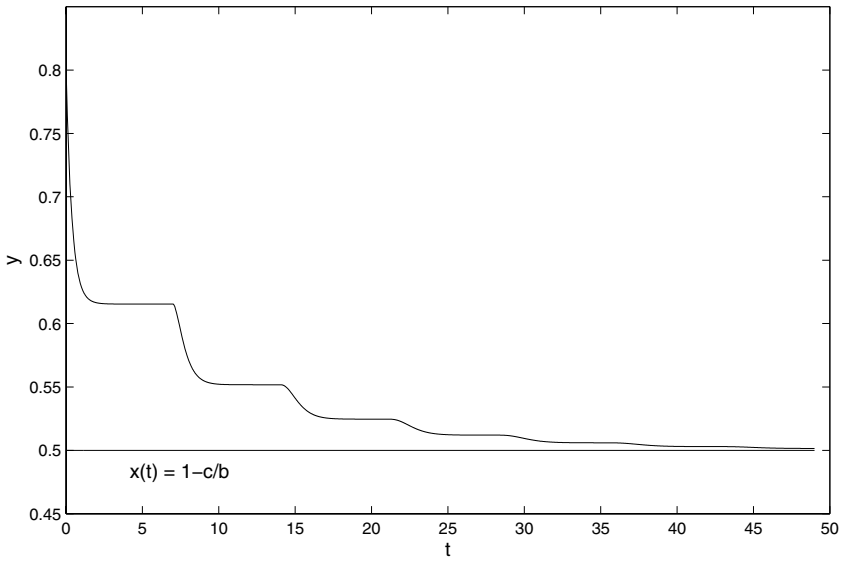


FIG. 4.1. *Numerical solution for Example* 1 *with* $b = 2$, $c = 1$ *obtained by the LL scheme* (4.1) *with* $h = 0.01$. *The straight line represents the equilibrium point* $x(t) = 1 - c/b$.

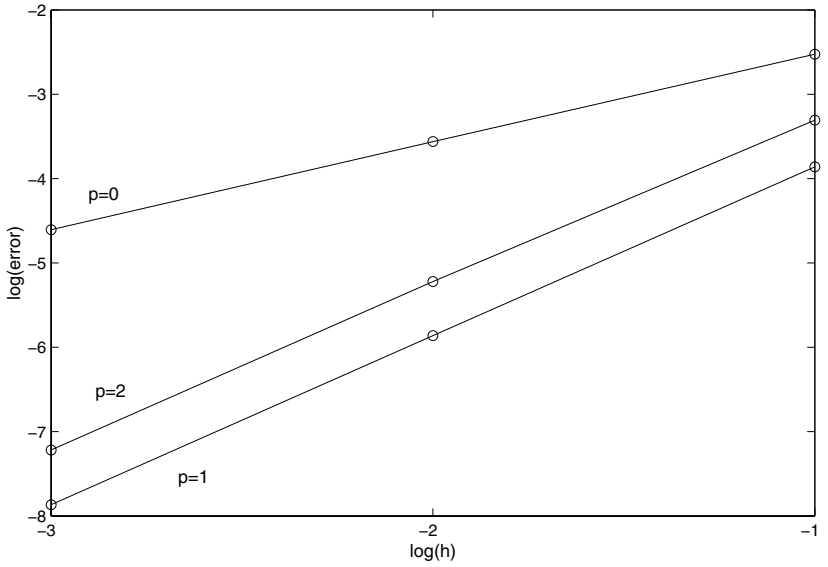


FIG. 4.2. *Maximum errors of interpolating polynomial LL schemes produced in the integration of Example* 1 *at the points* $(t)_h$, *with* $h = 10^{-1}$, 10^{-2}, 10^{-3}. *The slopes of the straight line that fits these points are* 1, 2, *and* 2 *for the schemes with interpolating polynomials of grade* $p = 0, 1, 2$, *respectively.*

TABLE 4.1
Values of the parameters in the DDE of Example 2.

j/a	a_j	a_{1j}	a_{2j}	a_{3j}
0	–	0.15	2	16
1	83	9.4×10^{9}	8×10^{28}	0.1
2	5	10^{-15}	1	10^{-18}
3	6.6×10^{14}	1.2	10^{-19}	1.7×10^{30}
4	3×10^{11}	2.7×10^{16}	5.3×10^{33}	3
5	0.4	2	16	0.4
6	2.5×10^{7}	5.3×10^{27}	1.6×10^{14}	4.3×10^{-22}
7	5×10^{-13}	1	0.4	8.5×10^{6}
8	2.3×10^{9}	10^{-18}	10^{-18}	8.6×10^{11}
9	0.052	2.7×10^{16}	8×10^{32}	0.043

system of ten-dimensional DDEs,

$$
\begin{aligned}
\frac{dx_1}{dt} &= a_1 x_2 + a_2 a_3 x_2 x_7 - a_4 x_1 x_{10} - a_5 x_1 - a_6 x_1 (a_7 - x_2 - x_3),\\
\frac{dx_2}{dt} &= a_8 x_1 (a_7 - x_2 - x_3) - a_3 x_2 x_7 - a_9 x_2,\\
\frac{dx_3}{dt} &= a_3 x_2 x_7 + a_9 x_2 - a_{10} x_3,\\
\frac{dx_4}{dt} &= a_{11} a_{12} x_1 - a_{13} x_4,\\
\frac{dx_5}{dt} &= a_{14}((1 - x_3/a_7) a_{15} x_4(t - \tau_1) x_5(t - \tau_1) - x_4 x_5) - a_{16} x_4 x_5 x_7 + a_{17}(a_{18} - x_5),\\
\frac{dx_6}{dt} &= a_{19}((1 - x_3/a_7) a_{20} x_4(t - \tau_2) x_6(t - \tau_2) - x_4 x_6) - a_{21} x_4 x_6 x_8 + a_{22}(a_{23} - x_6),\\
\frac{dx_7}{dt} &= a_{24}((1 - x_3/a_7) a_{25} x_4(t - \tau_3) x_5(t - \tau_3) x_7(t - \tau_3) - x_4 x_5 x_7) - a_{26} x_2 x_7\\
&\quad + a_{27}(a_{28} - x_7),\\
\frac{dx_8}{dt} &= a_{29}((1 - x_3/a_7) a_{30} x_4(t - \tau_4) x_6(t - \tau_4) x_8(t - \tau_4) - x_4 x_6 x_8) + a_{31}(a_{32} - x_8),\\
\frac{dx_9}{dt} &= a_{33}(1 - x_3/a_7) a_{34} x_4(t - \tau_5) x_6(t - \tau_5) x_8(t - \tau_5) + a_{35}(a_{36} - x_9),\\
\frac{dx_{10}}{dt} &= a_{37} x_9 - a_{38} x_{10} x_1 - a_{39} x_{10},
\end{aligned}
$$

with five time delays $\tau_1 = \tau_2 = 0.6$, $\tau_3 = \tau_4 = 2$, and $\tau_5 = 3$, and initial conditions $x_1(t) = 2.9 \times 10^{-16}$, $x_2(t) = x_3(t) = x_4(t) = 0$, $x_5(t) = a_{18}$, $x_6(t) = a_{23}$, $x_7(t) = a_{28}$, $x_8(t) = a_{32}$, $x_9(t) = a_{36}$, and $x_{10}(t) = a_{36} a_{37} a_{39}$ for all $t \in [-\tau_5, 0]$. The parameters a_i are given in Table 4.1.

That equation has been used as a test example to compare the performance of numerical integrators in the case of multidimensional stiff equations with multiple delays. It has been reported in [6] that a number of them fail to produce a numerical solution after $t = 110$ because the appearance of sharp picks in the solution. Figure 4.3 illustrates this problem. In this case, the explicit continuous extension Runge–Kutta $(2, 3)$ scheme with step-size $h = 0.01$ was used.

On the contrary, Figure 4.4 shows the numerical solution until $t = 150$ obtained

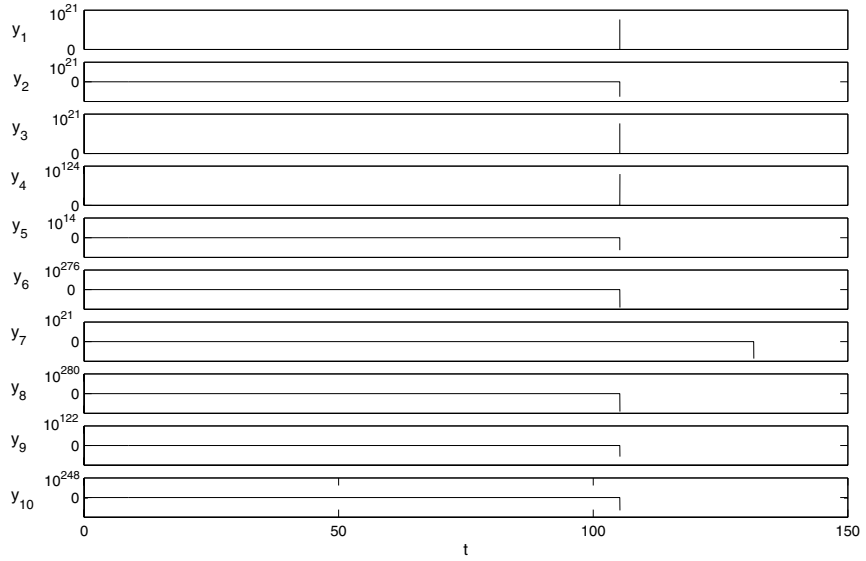


FIG. 4.3. *Approximate solution of Example* 2 *computed by the continuous extension Runge–Kutta* $(2,3)$ *scheme with step-size* $h = 0.01$.

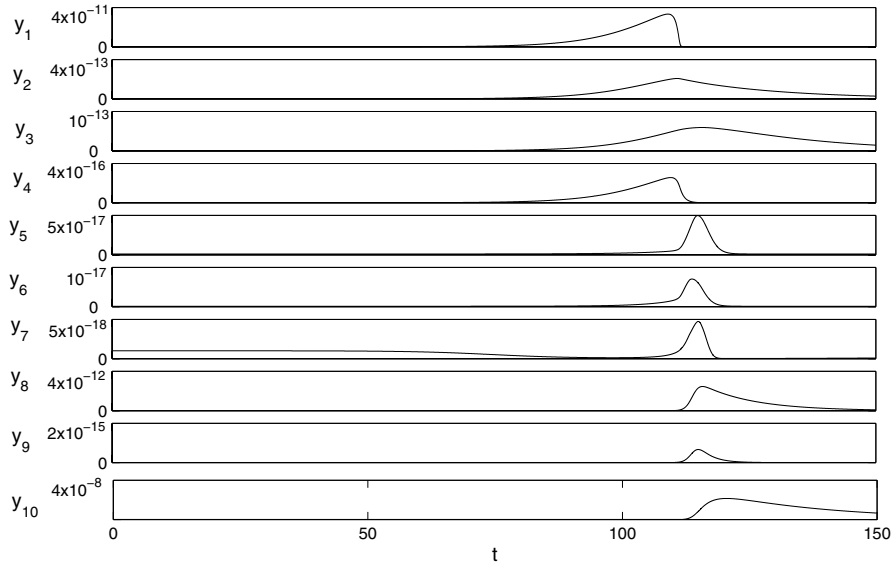


FIG. 4.4. *Approximate solution of Example* 2 *computed by the LL scheme* (4.1) *with step-size* $h = 0.01$.

TABLE 4.2
Maximum of the relative errors (%) produced by the LL scheme (4.1) in the integration of Example 2 on the interval $[0, 150]$ with different step-size h.

h/x	1	2	3	4	5	6	7	8	9	10
0.1	17.86	0.0377	0.0478	0.7256	0.3140	0.5269	1.3753	0.6410	0.6258	1.7621
0.01	0.13	0.0004	0.0035	0.0780	0.0031	0.0054	0.0120	0.0730	0.0071	0.0132

by the LL scheme (4.1) with the same step-size. Table 4.2 presents the maximum of the relative errors of the LL scheme in $(t)_h$ versus h for each variable. For a "true" solution we used the trajectory obtained by the LL scheme with $h = 0.00001$. The time for computing such a solution (until $t = 100$) was 1.7 times longer than the time used by the Runge–Kutta scheme mentioned above. Thus, no small step-size is necessary to integrate that equation with an adequate precision and computational cost, which reveals the potential of the LL method to integrate stiff DDEs.

5. Conclusions. The local linearization approach for the numerical integration of DDEs was introduced and two numerical schemes were considered. The first one, called the natural LL scheme, preserves the stability of multidimensional linear DDEs with multiple delays, but its computational cost is high in the case that the smallest delay of the DDE is much lower than the final integration time. On the contrary, according to the simulation study carried out, the computational cost of the polynomial LL scheme is comparable to the cost of the conventional explicit integrators but with the advantage of integrating stiff systems. This last result agrees with similar performances of the LL integrators of other classes of differential equations (ordinary, random, and stochastic). The two LL schemes proposed in this paper are explicit and have second order of convergence. Nevertheless, high order schemes of this family can also be derived by just following the same ideas that have been used to construct high order LL integrators for ODEs and SDEs [15, 16].

6. Appendix. The following result is a generalization of Theorem 1 in [40].

LEMMA 6.1 (Theorem 1 in [11]). *Let n, d_1, $d_2, \ldots, d_n$ be positive integers and $\mathbf{A}$ an $n \times n$ block triangular matrix defined by*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \ldots & \mathbf{A}_{1n} \\ \mathbf{0} & \mathbf{A}_{22} & \ldots & \mathbf{A}_{2n} \\ \mathbf{0} & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{nn} \end{pmatrix},$$

where $(\mathbf{A}_{lj})$ are $d_l \times d_j$ matrices, with $l, j = 1, \ldots, n$. Then for $t \geqslant 0$

$$e^{\mathbf{A}t} = \begin{pmatrix} \mathbf{B}_{11}(t) & \mathbf{B}_{12}(t) & \ldots & \mathbf{B}_{1n}(t) \\ \mathbf{0} & \mathbf{B}_{22}(t) & \ldots & \mathbf{B}_{2n}(t) \\ \mathbf{0} & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{nn}(t) \end{pmatrix},$$

with

$$\mathbf{B}_{ll}(t) = e^{\mathbf{A}_{ll}t}, \quad l = 1, \ldots, n,$$
$$\mathbf{B}_{lj}(t) = \int_0^t \mathbf{M}^{(l,j)}(t, s_1) ds_1, \quad l = 1, \ldots, n-1, j = l+1,$$

$$\mathbf{B}_{lj}(t) = \int_0^t \mathbf{M}^{(l,j)}(t,s_1)ds_1 + \sum_{k=1}^{j-l-1} \int_0^t\int_0^{s_1} \cdots \int_0^{s_k} \sum_{l<i_1<\cdots<i_k<j} \mathbf{M}^{(l,i_1,\ldots,i_k,j)}(t,s_1,\ldots,s_{k+1})ds_{k+1}\ldots ds_1,$$
$$l = 1,\ldots,n-2,\ j = l+2,\ldots,n,$$

where, for any multi-index $(i_1,\ldots,i_k) \in \mathbb{N}^k$ *and vector* $(s_1,\ldots,s_k) \in \mathbb{R}^k$*, the matrices* $\mathbf{M}^{(i_1,\ldots,i_k)}(s_1,\ldots,s_k)$ *are defined by*

$$\mathbf{M}^{(i_1,\ldots,i_k)}(s_1,\ldots,s_k) = \left(\prod_{r=1}^{k-1} e^{\mathbf{A}_{i_r i_r}(s_r - s_{r+1})}\mathbf{A}_{i_r i_{r+1}}\right) e^{\mathbf{A}_{i_k i_k} s_k}.$$

The thesis of the next lemma is known as the fundamental inequality in the framework of ODEs.

LEMMA 6.2 (fundamental inequality; Theorem 2 in [24, p. 6]). *Let* $\mathbf{f}(t,\mathbf{x}) : [t_0,t_1] \times D \longrightarrow \mathbb{R}^m$, $D \subset \mathbb{R}^m$*, be a continuous function that satisfies*

$$\|\mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{y})\| \le \lambda_0 \|\mathbf{x}-\mathbf{y}\|, \qquad \lambda_0 \ge 0,$$

for all $t \in [t_0,t_1]$ *and* $\mathbf{x},\mathbf{y} \in D$*. Let* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *be functions such that*

$$\left\| \frac{d\mathbf{u}(t)}{dt} - \mathbf{f}(t,\mathbf{u}(t)) \right\| \le \varepsilon_1,$$
$$\left\| \frac{d\mathbf{v}(t)}{dt} - \mathbf{f}(t,\mathbf{v}(t)) \right\| \le \varepsilon_2$$

for all $t \in [t_0,t_1]$*. Set*

$$\mathbf{p}(t) = \mathbf{u}(t) - \mathbf{v}(t) \qquad \text{and} \qquad \varepsilon = \varepsilon_1 + \varepsilon_2.$$

Then

$$\|\mathbf{p}(t)\| \le e^{\lambda_0(t-t_0)} \|\mathbf{p}(t_0)\| + \frac{\varepsilon}{\lambda_0}(e^{\lambda_0(t-t_0)} - 1)$$

for all $t \in [t_0,t_1]$.

LEMMA 6.3. *Let* $\mathbf{A}$, $\mathbf{C}$ *be* $n \times m$ *matrices, and let* $\mathbf{B}$, $\mathbf{D}$ *be* $m \times r$ *matrices. Then*

$$\|\mathbf{AB} - \mathbf{CD}\| \le \|\mathbf{A}-\mathbf{C}\| \|\mathbf{D}\| + \|\mathbf{A}\| \|\mathbf{B}-\mathbf{D}\|.$$

Proof. The lemma is obtained by applying the triangular inequality to the identity $\mathbf{AB} - \mathbf{CD} = (\mathbf{A}-\mathbf{C})\mathbf{D} + \mathbf{A}(\mathbf{B}-\mathbf{D})$. ∎

Acknowledgment. The authors are very grateful to a referee, whose comments and suggestions contributed to significantly improve the paper.

REFERENCES

[1] C. T. H. BAKER, *Retarded differential equations*, J. Comput. Appl. Math., 125 (2000), pp. 309–335.

[2] C. T. H. BAKER, C. A. H. PAUL, AND D. R. WILLE, *Issues in the numerical solution of evolutionary delay differential equations*, Adv. Comput. Math., 3 (1995), pp. 171–196.

[3] T. H. Baker and C. A. H. Paul, *Computing stability regions: Runge–Kutta methods for delay differential equations*, IMA J. Numer. Anal., 14 (1994), pp. 347–362.
[4] A. Bellen and S. Maset, *Numerical solution of constant coefficient linear delay differential equations as abstract Cauchy problems*, Numer. Math., 84 (2000), pp. 351–374.
[5] R. Biscay, J. C. Jimenez, J. Riera, and P. Valdes, *Local linearization method for the numerical solution of stochastic differential equations*, Ann. Inst. Statist. Math., 48 (1996), pp. 631–644.
[6] G. A. Bocharov, G. I. Marchuk, and A. A. Romanyukha, *Numerical solution by LMMs of stiff delay differential systems modelling an immune response*, Numer. Math., 73 (1996), pp. 131–148.
[7] G. A. Bocharov and F. A. Rihan, *Numerical modelling in biosciences using delay differential equations*, J. Comput. Appl. Math., 125 (2000), pp. 183–199.
[8] M. Calvo and T. Grande, *On the asymptotic stability of θ-methods for delay differential equations*, Numer. Math., 54 (1988), pp. 257–269.
[9] F. Carbonell, J. C. Jimenez, and R. J. Biscay, *Weak local linear discretizations for stochastic differential equations: Convergence and numerical schemes*, J. Comput. Appl. Math., 197 (2006), pp. 578–596.
[10] F. Carbonell, J. C. Jimenez, R. Biscay, and H. de la Cruz, *The local linearization method for numerical integration of random differential equations*, BIT, 45 (2005), pp. 1–14.
[11] F. Carbonell, J. C. Jimenez, and L. Pedroso, *Computing Multiple Integrals Involving Matrix Exponentials*, Technical report 2005-328, Instituto de Cibernética Matemática y Física, La Habana, Cuba, 2005.
[12] H. Cartan, *Calcul Differentiel*, Hermann, Paris, 1967.
[13] J. H. E. Cartwright and O. Piro, *The dynamics of Runge–Kutta methods*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 427–449.
[14] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
[15] H. de la Cruz, R. J. Biscay, F. Carbonell, T. Ozaki, and J. C. Jimenez, *A Higher Order Local Linearization Method for Solving Ordinary Differential Equations*, Technical report 2005-337, Instituto de Cibernética Matemática y Física, La Habana, Cuba, 2005.
[16] H. de la Cruz, R. J. Biscay, F. Carbonell, J. C. Jimenez, and T. Ozaki, *Local linearization–Runge–Kutta (LLRK) methods for solving ordinary differential equations*, in Computational Science—ICCS 2006, Lecture Notes in Comput. Sci. 3991, Springer-Verlag, New York, 2006, pp. 132–139.
[17] R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, New York, 1977.
[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[19] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems*, 2nd ed., Springer-Verlag, Berlin, 1993.
[20] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
[21] D. J. Higham and I. T. Famelis, *Equilibrium states of adaptive algorithms for delay differential equations*, J. Comput. Appl. Math., 58 (1995), pp. 151–169.
[22] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM Numer. Anal., 34 (1997), pp. 1911–1925.
[23] T. Hongjiong and K. Jiaoxun, *The stability of the θ-methods in the numerical solution of delay differential equations with several delay terms*, J. Comput. Appl. Math., 58 (1995), pp. 171–181.
[24] W. Hurewicz, *Lectures on Ordinary Differential Equations*, Edicion Revolucionaria, La Habana, Cuba, 1966.
[25] K. J. In't Hout, *A new interpolation procedure for adapting Runge–Kutta methods to delay differential equations*, BIT, 32 (1992), pp. 634–649.
[26] J. C. Jimenez, R. Biscay, C. Mora, and L. M. Rodriguez, *Dynamic properties of the local linearization method for initial-value problems*, Appl. Math. Comput., 131 (2002), pp. 21–37.
[27] J. C. Jimenez and T. Ozaki, *Identification of Continuous–Discrete State Space Models with Delays*, Research memo 658, The Institute of Statistical Mathematics, Tokyo, Japan, 1997.
[28] J. C. Jimenez and T. Ozaki, *Local linearization filters for nonlinear continuous-discrete state space models with multiplicative noise*, Internat. J. Control, 76 (2003), pp. 1159–1170.
[29] J. C. Jimenez and T. Ozaki, *An approximate innovation method for the estimation of diffusion processes from discrete data*, J. Time Ser. Anal., 27 (2006), pp. 77–97.
[30] V. R. Kolmanovskii and A. Myshkis, *Applied Theory of Functional Differential Equations*, Kluwer, Dordrecht, The Netherlands, 1992.
[31] N. MacDonald, *Time Lags in Biological Models*, Springer-Verlag, Berlin, 1978.

[32] S. E. A. Mohammed, *Stochastic Functional Differential Equations*, Pitman, Boston, MA, 1984.
[33] K. W. Neves, *Automatic integration of functional differential equations*, ACM Trans. Math. Software, 1 (1975), pp. 357–368.
[34] H. J. Orbele and H. J. Pesch, *Numerical treatment of delay differential equations by Hermite interpolation*, Numer. Math., 37 (1981), pp. 235–255.
[35] T. Ozaki, *A local linearization approach to nonlinear filtering*, Internat. J. Control, 57 (1993), pp. 75–96.
[36] B. L. S. Prakasa-Rao, *Statistical Inference for Diffusion Type Processes*, Oxford University Press, Oxford, UK, 1999.
[37] H. Schurz, *Numerical analysis de stochastic differential equations without tears*, in Handbook of Stochastic Analysis and Applications, D. Kannan and V. Lakahmikamtham, eds., Marcel Dekker, New York, 2002, pp. 237–358.
[38] R. B. Sidje, *EXPOKIT: Software package for computing matrix exponentials*, AMC Trans. Math. Software, 24 (1998), pp. 130–156.
[39] I. Stewart, *Numerical methods: Warning—handle with care!*, Nature, 355 (1992), pp. 16–17.
[40] C. F. Van Loan, *Computing integrals involving the matrix exponential*, IEEE Trans. Automat. Control, 23 (1978), pp. 395–404.
[41] M. Zennaro, *Natural continuous extensions of Runge–Kutta methods*, Math. Comput., 46 (1986), pp. 119–133.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2610–2632

IMPROVED SIMULATION FOR THE KILLED BROWNIAN MOTION IN A CONE*

STÉPHANE MENOZZI†

Abstract. In this paper, we first give an error expansion of the weak error associated to a discretely killed Brownian motion in a cone that writes as an intersection of half spaces. We exploit this result to derive an original correction method to improve the initial convergence rate. This method is based on the sensitivity of the underlying Dirichlet problem w.r.t. the domain and turns out to be a numerically cheaper and sharper alternative to standard extrapolation techniques.

Key words. discretely killed Brownian motion, nonsmooth domains, overshoot above the boundary, domain correction

AMS subject classifications. 65C30, 60H35, 60J65

DOI. 10.1137/050636966

1. Introduction: Statement of the problem. Let $(X_t)_{t\in[0,T]}$ be a d-dimensional diffusion process. For a real valued functional ψ of the process, the numerical estimation of $Q_T := \mathbb{E}[\psi\big((X_t)_{t\in[0,T]}\big)]$ arises in a large class of problems. In financial mathematics, Q_T corresponds to the price of an option with possibly path-dependent pay-off ψ if one assumes the dynamics of the underlying asset is given by X; see, e.g., [KS98]. For some ψ, Q_T is also the probabilistic representation of the solution of a linear PDE. This is the well-known Feynman–Kac formula; see [Fre85]. In this context, the probabilistic approximation is particularly adapted in large dimensions.

Concerning the numerical estimation of Q_T, the easiest case is $\psi\big((X_t)_{t\in[0,T]}\big) = f(X_T)$, i.e. when Q_T corresponds to the price of a European option with pay-off f or to the solution of a Cauchy problem in a PDE setting. This problem has been analyzed in detail by Talay and Tubaro [TT90] and then by Bally and Talay [BT96a], [BT96b]. These works provide an expansion of the weak error when the process is discretized with the Euler scheme.

We are going to deal with the trickier case $\psi\big((X_t)_{t\in[0,T]}\big) = f(X_T)\mathbf{1}_{\forall t\in[0,T]:\ X_t\in D}$, D being a given domain (i.e., an open connected subset) of $\mathbb{R}^d$. This is an irregular functional of the path. In this framework, Q_T corresponds to the price of a barrier option (see Andersen and Brotherton-Ratcliffe [ABR96] for references on the subject) or to the solution of a Cauchy–Dirichlet problem. Assuming the domain is smooth, Gobet and the author [Gob00], [GM04] obtained bounds for the weak error associated to the discretely killed Euler scheme but no error expansion. In this work, we derive an error expansion for the special case of the Brownian motion (Black–Scholes setting) and an intersection of half spaces (nonsmooth domains). Even though this context can seem quite restrictive, it already induces a major difficulty w.r.t. the previous works. Namely, for the proofs one has to handle the singularities of the heat kernel in a cone.

*Received by the editors July 27, 2005; accepted for publication (in revised form) June 23, 2006; published electronically December 11, 2006.
http://www.siam.org/journals/sinum/44-6/63696.html
†Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII, Denis Diderot, 175 Rue du Chevaleret, 75013 Paris, France (menozzi@math.jussieu.fr).

Let $(X_t)_{t\in[0,T]}$ be a d-dimensional Brownian motion (BM) with dynamics

$$X_t = x + \mu t + \sigma W_t \tag{1.1}$$

with fixed initial data x and terminal time T. Here W is a standard d-dimensional BM defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ with the usual assumptions on $(\mathcal{F}_t)_{t\in[0,T]}$. We assume $\sigma\sigma^*$ to be positive definite.

Let D be a domain of $\mathbb{R}^d$. Define $\tau := \inf\{t \geq 0 : X_t \notin D\}$. Consider a regular time mesh of the interval $[0,T]$ with N time steps, $(t_i = ih)_{i\in[\![0,N]\!]}$, $h = T/N$ being the step size. Introduce $\tau^N := \inf\{t_i \geq 0 : X_{t_i} \notin D\}$. For a measurable nonnegative function f and an initial point $x \in D$, denoting $\mathbb{E}_x[.] = \mathbb{E}[.|X_0 = x]$, we refer to the quantity

$$\mathrm{Err}(T, h, f, x) = \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau^N > T}] - \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau > T}]$$

as the weak error associated to the discrete time killing of X w.r.t. the domain D. Note that since $\tau^N > \tau$ a.s., $\mathrm{Err}(T, h, f, x) \geq 0$.

The discrete approximation of the exit time allows us to define a simple Monte Carlo procedure to estimate the previous quantity. In this context, $\mathrm{Err}(T, h, f, x)$ can thus be seen as the error associated to the discretization of the exit time.

Let us first recall some controls on $\mathrm{Err}(T, h, f, x)$ given in the literature. In [GM04] and Chapter I of [Men04], we proved, in the wider framework of killed diffusion processes approximated by their corresponding Euler schemes, that for smooth domains and functions f satisfying either some support condition w.r.t. D or some smoothness properties and compatibility conditions, one had that $\mathrm{Err}(T, h, f, x)$ was upper and lower bounded at order $1/2$ w.r.t. h. We also showed in [GM05] that, for a large class of Itô processes, the upper bound holds true for an intersection of smooth domains.

Still in [GM04], we stated an expansion and correction result for $\mathrm{Err}(T, h, f, x)$ in the special case of the half space in a Brownian framework. In this work, we extend these results to the case of an intersection of half spaces which is of particular interest in mathematical finance since the domain of a multiasset barrier option is often defined as a product domain.

Let $D \subset \mathbb{R}^d$ be a domain of the form $D = \cap_{i=1}^m D^i$, $m \in [\![1, d]\!]$, where the $(D^i)_{i\in[\![1,m]\!]}$ are d-dimensional half spaces with nonempty intersection. Under suitable smoothness properties up to the boundary for $v(t, x) := \mathbb{E}_x[f(X_{T-t})\mathbf{1}_{\tau > T-t}]$, we obtain an error expansion at order $\frac{1}{2}$ w.r.t. h for $\mathrm{Err}(T, h, f, x)$.

As emphasized in [GM04], the leading term in the weak error is still the one associated to the overshoot of the killed process above the boundary (the overshoot being defined as the distance to the boundary of the process when it exits the domain).

In the special case of Brownian motion, for half spaces or intersections of half spaces forming a cone, we are able to obtain the asymptotic distribution of the overshoot, extending previous results obtained by Siegmund [Sie79]. To derive the error expansion we then use usual techniques based on Taylor's expansions. The smoothness of v is needed for this last step. From a theoretical point of view, the main difficulty is analytical and consists in having good smoothness properties of v up to the boundary of a nonsmooth domain.

From a numerical point of view, the error expansion is the preliminary step for a procedure that aims to improve the convergence rate. A standard one in this framework is the Romberg extrapolation; see Talay and Tubaro [TT90] and section 3 for details.

In this paper, we propose an alternative correction method based on the recent work of Costantini, El Karoui, and Gobet [CKG03] concerning the sensitivity of the Dirichlet problem w.r.t. the domain. Unlike the Romberg extrapolation, we do not need to refine the time step and thus the procedure is computationally cheaper. Note also that the empirical variance associated to the Monte Carlo estimator is by construction smaller. We simply proceed to the simulation w.r.t. a more constrained domain. Namely, instead of killing the process when it exits from D at one of the discretization times, we kill it when it leaves $D_h := \cap_{i=1}^m D_h^i,\ D_h^i := \{y \in \mathbb{R}^d : y - C_i\sqrt{h}n_i \in D^i\}$, where n_i denotes the inner unit normal associated to the half space D^i. We will see that, for suitable positive constants C_i, this new choice of discrete time killing allows us to remove the leading term in the error development.

We mention that in a one-dimensional setting, both the expansion result and the correction procedure could be derived by direct computations from the work of Broadie, Glasserman, and Kou [BGK99].

For the sake of completeness, let us also mention that concerning the weak approximation of killed or reflected diffusion processes, another approach to improve the initial convergence rate can be found in Gobet [Gob01]. The techniques introduced therein strongly rely on some explicit transition probabilities for the Brownian motion in a half space and cannot be easily adapted to the orthant case.

Outline of the paper. We state our main results in section 2. Numerical results are presented in section 3. They confirm that the correction procedure is rather accurate and also numerically extends to a wider context for the underlying processes and domains. The proofs of the main results are developed in section 4. In section 5 we give some smoothness properties of v in nonsmooth domains. We conclude in section 6 evoking possible extensions and open problems. Appendices A and B are respectively devoted to the proofs of technical points concerning the asymptotic behavior of the overshoot and the killed heat kernel in a nonsmooth domain.

2. Main results.

2.1. Current working assumptions. We suppose our domain satisfies the following assumption:

(D) $D = \cap_{j=1}^m D^j\ \forall j \in [\![1,m]\!],\ D^j := \{y \in \mathbb{R}^d : y_j > b_0^j\}$, where $m \in [\![1,d]\!]$.

We introduce the following:

(BM) The d-dimensional process $(X_s)_{s\geq 0}$ has the form $X_s := x + \sigma_0 W_s$, where W is a standard d-dimensional BM and $\sigma_0\sigma_0^* = \begin{pmatrix} \Sigma & 0 \\ 0 & \mathbf{I}_{d-m} \end{pmatrix}$ is assumed to be positive definite and Σ is a correlation matrix with coefficients $(\rho_{ij})_{(i,j)\in[\![1,m]\!]^2}$. The integer $m \in [\![1,d]\!]$ is the same as in assumption **(D)**.

Suppose **(BM)**, **(D)** are in force. For a given positive measurable function f we define $\forall (t,y) \in [0,T]\times\mathbb{R}^d,\ v(t,y) := \mathbb{E}_y[f(X_{T-t})\mathbf{1}_{\tau>T-t}]$.

In the following, for an open set $U \subset \mathbb{R}^d$ we denote by $C_b^{k+\alpha}(\bar{U}), k \in \mathbf{N}^*,\ \alpha \in (0,1)$, the space of functions possessing k bounded and uniformly α-Hölder continuous spatial derivatives in $\bar{U}$. We also introduce $C_b^{k/2+\alpha/2,k+\alpha}([0,T]\times\bar{U}), k \in \mathbf{N}^*,\ \alpha \in (0,1)$, the space of functions possessing k bounded and uniformly α-Hölder continuous spatial derivatives in $[0,T]\times\bar{U}$ and $\lfloor k/2 \rfloor$ bounded and uniformly $(\mathbf{1}_{k=1}+\alpha)/2$-Hölder continuous time derivatives in $[0,T]\times\bar{U}$. Continuity is intended w.r.t. the parabolic metric; i.e., $\forall (P,Q) = ((t,x),(t',x')) \in ([0,T]\times\bar{U})^2,\ d(P,Q) = (|t-t'|+|x-x'|^2)^{1/2}$.

Now, under **(BM)**, **(D)**, we assume the following:

(S) The function f vanishes on the boundary. The associated function v belongs

to $C_b^{1/2+\alpha/2,1+\alpha}([0,T]\times\bar{D})\cap C^{1,2}([0,T)\times D)$, $\alpha>0$; i.e., there exists a constant $C>0$, s.t.

$$\sup_{\substack{(x,\,y)\,\in\,\bar{D}^2,\\(s,\,t)\,\in\,[0,\,T]^2}}\frac{|\nabla v(s,x)-\nabla v(t,y)|}{|s-t|^{\alpha/2}+|x-y|^{\alpha}}+\sup_{(s,\,t)\,\in\,[0,\,T]^2,\ x\,\in\,\bar{D}}\frac{|v(s,x)-v(t,x)|}{|s-t|^{\frac{1+\alpha}{2}}}\leq C.$$

$$|v|_{\infty,[0,T]\times\bar{D}}+|\nabla v|_{\infty,[0,T]\times\bar{D}}+$$

In particular, **(S)** means that the function f is at least $C_b^{1+\alpha}(\bar{D})$. We specify in section 5 sufficient conditions on f to obtain **(S)** in special cases.

We mention that the previous assumptions on f, i.e., that it is a smooth function vanishing on the boundary, are essentially technical. Numerically speaking they seem to have little influence; see section 3 for details.

2.2. Statement of the main theorems.

THEOREM 2.1 (error expansion for the correlated Brownian motion in an orthant). *Assume* **(BM)**, **(D)**, *and* **(S)**. *For h small enough the error writes*

$$Err(T,h,f,x)=C_1\sqrt{h}+o(\sqrt{h})$$

with $C_1=C_0\sum_{i=1}^m\mathbb{E}_x[\mathbf{1}_{\tau^i\leq T,\wedge_{j\in[\![1,m]\!]\setminus\{i\}}\tau^j>\tau^i}(\partial_{y_i}v(\tau^i,X_{\tau^i}))]$, $\tau^i:=\inf\{s\geq 0: X_s^i=b_0^i\}$, $C_0=\frac{\mathbb{E}_0[s_{\tau^+}^2]}{2\mathbb{E}_0[s_{\tau^+}]}$, *where* $s_0:=0,\forall n\geq 1, s_n:=\sum_{i=1}^n G^i$, *the G^i being i.i.d. standard centered normal variables and* $\tau^+:=\inf\{n\geq 0: s_n>0\}$.

One knows from [Sie79] and [AGP95] that $\frac{\mathbb{E}_0[s_{\tau^+}^2]}{2\mathbb{E}_0[s_{\tau^+}]}=-\frac{\zeta(1/2)}{\sqrt{2\pi}}=0.5823\ldots$, where ζ denotes Riemann's Zeta function.

Under our current assumptions **(BM)**, **(D)**, **(S)**, the next theorem improves the accuracy of the numerical procedure by removing the term of order $\frac{1}{2}$ in the error. For this, the simulation of $(X_{t_i})_{0\leq i\leq N}$ is performed in a modified domain, namely, $D^h:=\{y\in\mathbb{R}^d:\forall i\in[\![1,m]\!],\ y_i>b_0^i+C_0\sqrt{h}\}$ instead of $D:=\{y\in\mathbb{R}^d:\forall i\in[\![1,m]\!],\ y_i>b_0^i\}$.

We denote $\tau_{D^h}^N$ (resp., τ_{D^h}) the discrete (resp., continuous) exit time from this domain D^h.

THEOREM 2.2. *Assume* **(BM)**, **(D)**, **(S)**. *For h small enough we have*

$$Err'(T,h,f,x):=\mathbb{E}_x[f(X_T)\mathbf{1}_{\tau_{D^h}^N>T}]-\mathbb{E}_x[f(X_T)\mathbf{1}_{\tau>T}]=o(\sqrt{h}).$$

Remark 2.1. Consider now the more general case $X_s=x+\mu s+\sigma W_s, D:=\{x\in\mathbb{R}^d:(Ax)_i>b_i\ \forall i\in[\![1,m]\!]\}$, where $A=(a_1\ a_2\ \ldots\ a_m)^*$ is of rank m. Using Girsanov's theorem and a rotation of coordinates using a matrix Λ (with the ith row equal to $\frac{\sigma^*a_i}{\|\sigma^*a_i\|}$ for $i\in[\![1,m]\!]$ and the remaining rows forming an orthonormal basis of $\{\mathrm{Span}((\sigma^*a_j)_{j\in[\![1,m]\!]})\}^{\perp}$) preserving the Wiener measure, one obtains that for a Borelian bounded function f

$$\mathrm{Err}(T,h,f,x)=\mathbb{E}_0[f_0^x(\check{W}_T)(\mathbf{1}_{\tau_{D_0^x}^N>T}-\mathbf{1}_{\tau_{D_0^x}>T})],$$

where $\check{W}$ is a centered d-dimensional Brownian motion with covariance matrix $\sigma_0\sigma_0^*=\begin{pmatrix}\Sigma & 0\\ 0 & \mathbf{I}_{d-m}\end{pmatrix}$, and $\forall(i,j)\in[\![1,m]\!]^2$, $\Sigma_{ij}=\langle\sigma^*a_i,\sigma^*a_j\rangle/(\|\sigma^*a_i\|\|\sigma^*a_j\|)$. The domain D_0^x writes $D_0^x:=\cap_{j=1}^m D_0^{x,j}$, where $\forall\ j\in[\![1,m]\!]$, $D_0^{x,j}:=\{y\in\mathbb{R}^d:\ y_j>b_0^{x,j}\}$, $b_0^{x,j}=$

$\frac{b_j - a_j \cdot x}{\|\sigma^* a_j\|}$. Denoting $\tau_{D_0^{x,j}} := \inf\{s \geq 0 : \check{W}_s \notin D_0^{x,j}\}$, $\tau^N_{D_0^{x,j}} := \inf\{s_i \geq 0 : \check{W}_{s_i} \notin D_0^{x,j}\}$, we have $\tau_{D_0^x} := \wedge_{j=1}^m \tau_{D_0^{x,j}}, \tau^N_{D_0^x} := \wedge_{j=1}^m \tau^N_{D_0^{x,j}}$. The function f_0^x writes $f_0^x(y) = \exp(\sigma^{-1}\mu \cdot \Lambda^{-1} y - \frac{\|\sigma^{-1}\mu\|^2}{2} T) f(x + \sigma\Lambda^{-1} y)$.

Introduce, $\forall (t,y) \in [0,T] \times D_0^x$, $v_0^x(t,y) = \mathbb{E}_y[f_0^x(\check{W}_{T-t}) \mathbf{1}_{\tau_{D_0^x} > T-t}]$. If assumption **(S)** is fulfilled by v_0^x (one could weaken the boundedness condition in **(S)** to an exponential growth condition), then the former error expansion remains valid. We can as in the previous theorem remove the leading term of the error by simulating w.r.t. $D^h := \{x \in \mathbb{R}^d : (Ax)_i > (b + C_0 e\sqrt{h})_i,\ i \in [\![1,m]\!]\}$, $e = (\|\sigma^* a_1\|, \ldots, \|\sigma^* a_m\|)^*$.

Remark 2.2. In the half space case, i.e., for $m = 1$, the above transformation illustrates that the problem is essentially one-dimensional. This last aspect still holds true for a domain delimited by parallel hyperplanes. This is the reason why we did not take this case into consideration in **(D)**.

3. Numerical results. In this section we provide some numerical tests and compare the method from Theorem 2.2 with the usual Romberg correction that we briefly recall.

From Theorem 2.1, we derive $\frac{1}{\sqrt{2}-1}\mathbb{E}[f(X_T)(\sqrt{2}\mathbf{1}_{\tau^{2N} > T} - \mathbf{1}_{\tau^N > T})] - \mathbb{E}[f(X_T)\mathbf{1}_{\tau > T}] = o(\sqrt{h})$. The Romberg extrapolation technique consists in approximating by a Monte Carlo method the first term in the left-hand side of the previous equation.

We point out that our correction is numerically less expensive than the Romberg procedure that requires refining the time step. The Monte Carlo estimator deriving from Theorem 2.2 also has by construction a smaller empirical variance.

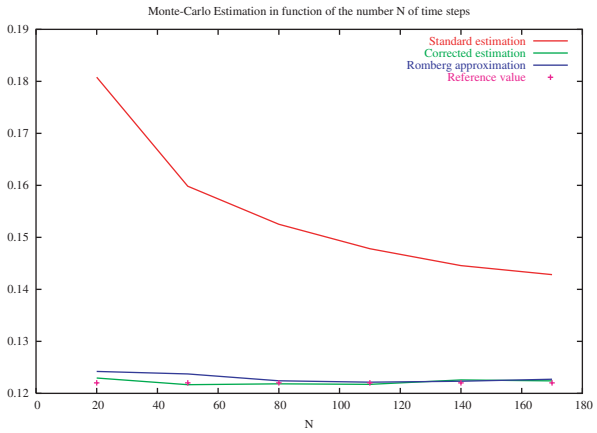
Bidimensional cone. We consider a two-dimensional risky asset following the Black–Scholes–Merton dynamics, $S_t^1 = S_0^1 \exp(\sigma_1 W_t^1 + (r - \frac{\sigma_1^2}{2})t)$, $S_t^2 = S_0^2 \exp(\sigma_2(\rho W_t^1 + (1-\rho^2)^{1/2} W_t^2) + (r - \frac{\sigma_2^2}{2})t)$, where $W = (W^1, W^2)$ is a standard bidimensional BM. For a fixed final time T, a given strike K, and threshold B, put $D := \{(s_1, s_2) \in \mathbb{R}^2 : s_1 > B, s_2 > B\}$. We are interested in computing $\mathbb{E}[e^{-rT}\mathbf{1}_{\tau > T}\varphi(S_T)]$, where φ is a smooth approximation of the indicator function that one expects in the case of a digital barrier option. We take $\varphi(s) := \mathbf{1}^\star_{K,\varepsilon}(s_1)\mathbf{1}^\star_{K,\varepsilon}(s_2)$ with $\mathbf{1}^\star_{K,\varepsilon}(s_1) = 0$ if $s_1 \leq K - \varepsilon$, $\mathbf{1}^\star_{K,\varepsilon}(s_1) = 1$ if $s_1 \geq K$, and in between we use the smooth interpolating function $\mathbf{1}^\star_{K,\varepsilon}(s_1) = 10\varepsilon^{-3}(s_1 - (K-\varepsilon))^3 - 15\varepsilon^{-4}(s_1 - (K-\varepsilon))^4 + 6\varepsilon^{-5}(s_1 - (K-\varepsilon))^5$. As soon as $K > B + \varepsilon$, the previous function φ satisfies conditions that guarantee **(S)** is fulfilled up to an exponential growth condition; see assumption **(F)**, section 5.2, Proposition 5.3, and Remark 2.1.

For $r = .04$, $\sigma_1 = \sigma_2 = .3$, $\rho = .5$, $S_0^1 = S_0^2 = 100 = K$, $B = 90$, $T = 1$, $\varepsilon = 5$ we compute the standard Monte Carlo approximation, the Romberg approximation, and the correction proposed in Theorem 2.2 for 10^6 paths (see Figure 1). The reference value has been computed with 10^6 paths and 15000 times steps with the Monte Carlo procedure.

The width of the 95% confidence interval is essentially equal to 1.5 10^{-3}. Furthermore we also observe that the associated empirical variance is lower than the one of the Romberg extrapolation.

From a numerical point of view a natural question concerns the behavior of the $o(\sqrt{h})$ appearing in Theorem 2.2. We have not experimentally emphasized a constant exponent; anyhow it turns out that the numerical rest is smaller than $O(h^{3/2})$.

Experiments in higher dimensions. Set $d \in \mathbf{N}^*, d > 2$. Let $X_s = x + \sigma_0 W_s$, where W is a d-dimensional standard BM and $\forall (i,j) \in [\![1,d]\!]^2$, $(\sigma_0 \sigma_0^*)_{ij} = \mathbf{1}_{i=j} +$



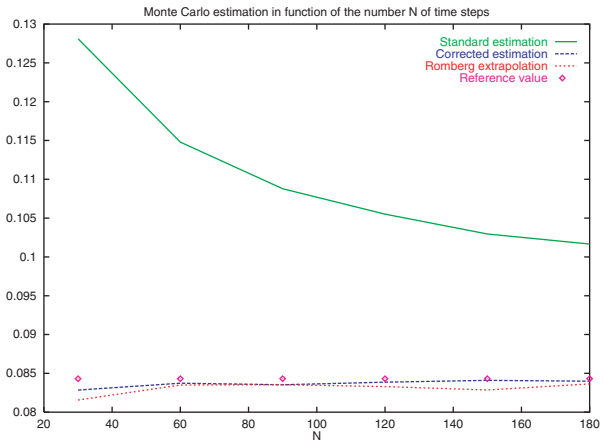
Empirical mean

	MC	MC Shift	Romberg	Reference
$N = 20$	.180807	0.12295	0.124209	0.122017
$N = 50$	.159831	0.121661	0.123736	0.122017
$N = 80$	.15251	0.12184	0.122422	0.122017
$N = 110$	.147821	0.121735	0.122145	0.122017
$N = 140$	.144559	0.122581	0.122338	0.122017
$N = 170$	.142827	0.122393	0.122717	0.122017

Unbiased variance estimator

	MC	MC Shift	Romberg	Reference
$N = 20$	.146365	0.106928	.298985	.10413
$N = 50$	.132417	0.105984	.230244	.10413
$N = 80$	.127643	0.106117	.205762	.10413
$N = 110$	.12445	0.106048	.191531	.10413
$N = 140$	.122845	0.106683	.182411	.10413
$N = 170$	.121412	0.106549	.175806	.10413

FIG. 1.



Empirical mean

	MC	MC Shift	Romberg	Reference
$N = 30$	.128099	.0828428	.0815582	.0843068
$N = 60$	.114797	.0837263	.0834953	.0843068
$N = 90$	.108789	.0835279	.0835393	.0843068
$N = 120$	.105514	.0838578	.083296	.0843068
$N = 150$	.102956	.0840968	.082859	.0843068
$N = 180$	.101663	.0839845	.0836351	.0843068

Unbiased variance estimator

	MC	MC Shift	Romberg	Reference
$N = 30$	.111558	.0759777	.187642	.0771784
$N = 60$	.101524	.0767086	.151835	.0771784
$N = 90$	.096877	.0765401	.137229	.0771784
$N = 120$	.0943122	.0768126	.128758	.0771784
$N = 150$	.0922929	.0770106	.124373	.0771784
$N = 180$	.0912674	.0769148	.119289	.0771784

FIG. 2.

$\frac{\alpha}{d-1}\mathbf{1}_{i\neq j}, \alpha \in [0,1)$, so that $\sigma_0\sigma_0^*$ has dominant diagonal and is thus positive definite. We take $D := \{x \in \mathbb{R}^d : x_i > 0 \; \forall i \in [\![1,d]\!]\}$ and we are interested in approximating the quantity $\mathbb{E}_x[f(X_T)\mathbf{1}_{\tau>T}]$, where $\forall x \in D,\ f(x) = \prod_{i=1}^d \mathbf{1}_{K<S_0^i \exp(sx_i)}$. Here $S_0 \in \mathbb{R}^d$ is a fixed vector and s is a fixed scale factor. Note that f is not as smooth as required in **(S)**.

The results in Figure 2 have been obtained with $d = 5$, $T = 1$, $K = 100$, $\alpha = s =$

$.05$, $S_0^i = 95\exp(sx_i)$, $x_i = .85\ \forall i \in [\![1,d]\!]$. The reference value has been computed with the standard Monte Carlo procedure for 2×10^6 paths and 2×10^4 steps. We used 10^6 paths for the other Monte Carlo computations.

The size of the 95% confidence interval is essentially equal to $1.5\ 10^{-3}$.

Even though we are not under the assumptions of the main theorems, the correction technique gives a good result. It is still sharper and provides smaller empirical variance than the Romberg extrapolation.

Numerical correction in a non-Brownian setting. We introduce in this section an algorithm that aims to extend the correction method of Theorem 2.2 to the diffusion case and to a wider class of domains. Namely, we assume in the following that D is a Lipschitz domain. In particular, the inner normal unit is defined a.e. on ∂D. Let $(X_s)_{s\geq 0}$ be a diffusion process with dynamics

$$X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s, \tag{3.1}$$

where b,σ are bounded and Lipschitz continuous. We approximate it by its Euler scheme $X_t^N = x + \int_0^t b(X_{\phi(s)}^N)ds + \int_0^t \sigma(X_{\phi(s)}^N)dW_s$, $\phi(s) := \inf\{t_i \geq 0 : t_i \leq s < t_{i+1}\}$ and define the discrete exit time $\tau^N := \inf\{t_i \geq 0 : X_{t_i}^N \notin D\}$. Note that the Euler scheme is locally in time nothing else but a Brownian motion with constant drift and diffusion coefficients. Thus, on the set $\tau^N > t_i$, $i \in [\![0, N-1]\!]$, and for $X_{t_i}^N$ in a neighborhood of the boundary s.t. $\Pi_{\partial D}$ is a.e. well defined, mimicking the procedure introduced in Theorem 2.2, we heuristically extend the correction by killing the Euler scheme in t_{i+1} whenever it is outside $D_h(X_{t_i}^N) := D\backslash V_{\partial D}(C_0\sqrt{h}\|\sigma^*(X_{t_i}^N)n(\Pi_{\partial D}(X_{t_i}^N))\|)$, where $\forall a>0, V_{\partial D}(a) := \{y\in\mathbb{R}^d : d(y,\partial D)\leq a\}$.

From an algorithmic point of view, the computation of $n(\Pi_{\partial D}(X_{t_i}^N))$ can be very demanding. Anyhow, on $\{\tau^N > t_i\}$ for a given $\eta > 0$ s.t. $X_{t_i}^N \in D\backslash V_{\partial D}(h^{1/2-\eta})$ we derive from Bernstein's inequality (cf. Lemma 4.1) and standard computations that $\mathbb{P}[\tau^N = t_{i+1}|\mathcal{F}_{t_i}] \leq C\exp(-ch^{-\eta})$. It is therefore useless to refine the simulation procedure for those events.

We sum up this heuristic correction in the following algorithm.

Algorithm 3.1 (empirical correction in a diffusion framework).

- *Assume X follows the dynamics of* (3.1).
- *Fix $\eta > 0$ and set*

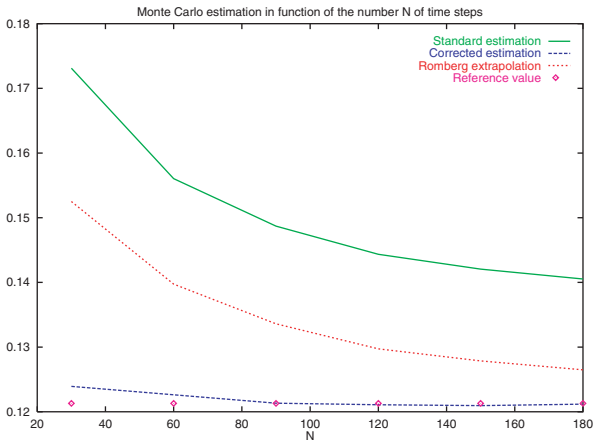
(i) $X_0^N = x$.

(ii) $\forall i \in [\![0, N-1]\!]$ *s.t.* $\tau^N > t_i$, $X_{t_{i+1}}^N := X_{t_i}^N + b(X_{t_i}^N)h + \sigma(X_{t_i}^N)(W_{t_{i+1}} - W_{t_i})$.

- If $X_{t_i} \in V_{\partial D}(h^{1/2-\eta})$ *and* $X_{t_{i+1}} \notin D_h(X_{t_i}^N)$*:* Kill the path.
- If $X_{t_i} \notin V_{\partial D}(h^{1/2-\eta})$ *and* $X_{t_{i+1}} \notin D$*:* Kill the path (rare event).
- If $i+1 \neq N$ *and no killing, iterate step* (ii).

We first give some results that illustrate that the proposed extension has a good numerical behavior. Namely, we take $b(x) = (b^i(x))_{i\in[\![1,d]\!]} = (\sin(x_i))_{i\in[\![1,d]\!]}, \sigma(x) = (\sigma^{ij}(x))_{(i,j)\in[\![1,d]\!]^2} = (\mathbf{1}_{i=j} + \sin(x_i)/(2(d-1))\mathbf{1}_{j\neq i})_{(i,j)\in[\![1,d]\!]^2}$ in (3.1) and a domain corresponding to a hypercube. For $f(x) = 1$ (corresponding to the estimation of the complementary distribution function of the exit time), $d = 3$, $D = (-1, 1.5)\times(-1.5, 1)\times(-2, 2.5)$, $T = 1$, $X_0 = 0$, we compute a reference value using the Romberg technique with $N = 12000$ and 2×10^6 paths. We obtain for 10^6 paths the results in Figure 3.

The size of the 95% confidence interval is still essentially equal to $1.5\ 10^{-3}$. For



Empirical mean

	MC	MC Shift	Romberg	Reference
$N = 30$	.173109	.123927	.145253	.121302
$N = 60$	.15604	.122632	.135507	.121302
$N = 90$	.148699	.121327	.132945	.121302
$N = 120$	.14434	.121088	.130965	.121302
$N = 150$	.142049	.120945	.129252	.121302
$N = 180$	.14053	.121177	.128577	.121302

Unbiased variance estimator

	MC	MC Shift	Romberg	Reference
$N = 30$	.143142	.108569	.152509	.108385
$N = 60$	.131692	.107594	.13973	.108385
$N = 90$	.126588	.106607	.133594	.108385
$N = 120$	.123506	.106426	.129721	.108385
$N = 150$	.121871	.106317	.127861	.108385
$N = 180$	.120781	.106493	.126486	.108385

FIG. 3.

this example, our correction is significantly more accurate.

We conclude this subsection giving a less favorable case. Take X a standard bidimensional BM, $D = B(0,1) \subset \mathbb{R}^2$, and $f(y) = (\frac{1}{2} - \|y\|^2)^+$. The 95% confidence interval associated to the reference value computed for $N_{MCR} = 2 \times 10^6, N_R = 14400$ by a standard Monte Carlo method is $I_C(N_{MCR}, N_R) = [.0203672, .0204871]$.

With $N_{MC} = 10^6$ paths, the previous correction algorithm gives the following:

	$N = 15$	$N = 30$	$N = 60$
$I_C(N_{MC}, N)$	[.019911, .0202323]	[.01999915, .020314]	[.0200514, .0203743]

	$N = 120$	$N = 240$	$N = 480$
$I_C(N_{MC}, N)$	[.0200527, .0203756]	[.0199975, .0203202]	[.0199531, .0202749]

Even though most of the intervals $I_C(N_{MC}, N)$ do not intersect $I_C(N_{MCR}, N_R)$, they are quite close to it. This is promising because the computational time employed to get the above estimates is significantly reduced w.r.t. to the one needed to obtain $I_C(N_{MCR}, N_R)$. Furthermore, the fact that the quantity to estimate is small brings additional numerical difficulty.

4. Proof of the main results.

4.1. Additional notation and usual controls. For smooth functions $g(t,x)$, the notation $H_g(t,x)$ stands for the Hessian matrix of g w.r.t. x. Time derivatives are denoted by $\partial_t^\beta g(t,x),\ \beta \in \mathbf{N}^*$.

We will keep the same notation C (or C') for all finite, nonnegative constants which will appear in our computations: they may depend on D, T, σ_0, or f, but they will not depend on the number of time steps N and the initial value x. We reserve the notation c and c' for constants that are also independent of T and f. Other possible dependences for the constants are explicitly indicated.

In the following $O_{pol}(h)$ (resp., $O(h)$) stands for every quantity $R(h)$ such that

$\forall n \in \mathbf{N}$, for some $C > 0$, one has $|R(h)| \leq Ch^n$ (resp., $|R(h)| \leq Ch$) (uniformly in x).

LEMMA 4.1 (Bernstein's inequality). *Assume* **(BM)**. *Consider two stopping times S, S' upper bounded by T with $0 \leq S' - S \leq \Delta \leq T$. Then for any $p \geq 1$, there are some constants $c > 0$ and C, such that for any $\eta \geq 0$, one has a.s.*

$$\mathbb{P}[\sup_{t\in[S,S']} \|X_t - X_S\| \geq \eta \mid \mathcal{F}_S] \leq C \exp\left(-c\frac{\eta^2}{\Delta}\right),$$
$$\mathbb{E}[\sup_{t\in[S,S']} \|X_t - X_S\|^p \mid \mathcal{F}_S] \leq C\Delta^{p/2}.$$

For a proof of the first inequality we refer to Chapter 3, section 3 in [RY99]. The other inequality easily follows from the first one or from the BDG inequalities.

4.2. Proof of Theorem 2.1 (expansion result). Let us briefly outline the scheme of the proof. First, we write the error as a sum of increments of the function v. Using Taylor expansions, we then introduce the overshoot terms of the process in the previous development (the overshoot being defined as the distance of the process to the domain when it exits the domain). We finally conclude using the asymptotic independence of the rescaled overshoot, as well as its integrability properties, and the discrete exit time.

Step 1: Decomposition of the error. Recalling that the function v vanishes on D^c, we write

$$\begin{aligned}\mathrm{Err}(T,h,f,x) &= \mathbb{E}_x[v(T\wedge\tau^N, \Pi_{\bar{D}}(X_{T\wedge\tau^N}))] - v(0,x)\\ &= \sum_{i=0}^{N-1} \mathbb{E}_x[\big(v(t_{i+1}\wedge\tau^N, \Pi_{\bar{D}}(X_{t_{i+1}\wedge\tau^N})) - v(t_i\wedge\tau^N, \Pi_{\bar{D}}(X_{t_i\wedge\tau^N}))\big)]\\ &= \sum_{i=0}^{N-1} \mathbb{E}_x[\mathbf{1}_{\tau^N > t_i}\big(v(t_{i+1}, \Pi_{\bar{D}}(X_{t_{i+1}})) - v(t_i, X_{t_i})\big)].\end{aligned}$$

Introduce $\forall t \in [0,T), \tau_t := \inf\{s \geq t : X_s \notin D\}$. Recall also that the function v satisfies the PDE

$$\begin{cases} (\partial_t v + \frac{1}{2}\mathrm{tr}(H_v \sigma_0 \sigma_0^*))(t,y) = 0,\ (t,y) \in [0,T)\times D,\\ v(t,.)|_{\partial D} = 0, t \in [0,T],\ v(T,y) = f(y),\ y \in \bar{D}.\end{cases} \tag{4.1}$$

Hence, $(v(s\wedge\tau_t, X_{s\wedge\tau_t}))_{s\in[t,T]}$ is a martingale. We have

$$\mathrm{Err}(T,h,f,x) = \sum_{i=0}^{N-1} \mathbb{E}_x[\mathbf{1}_{\tau^N > t_i}\mathbf{1}_{\tau_{t_i} < t_{i+1}}(v(t_{i+1}, \Pi_{\bar{D}}(X_{t_{i+1}})) - v(\tau_{t_i}, X_{\tau_{t_i}}))].$$

Define now $\forall j \in [\![1,m]\!]$, $\tau_t^j := \inf\{s \geq t : X_s^j = b_0^j\}$. Since $\mathbb{P}[\tau_{t_i}^j = \tau_{t_i}^k | \mathcal{F}_{t_i}] = 0$, $j \neq k$, one gets $\mathbb{P}[X_{t_{i+1}} \notin D|\mathcal{F}_{t_i}] = \sum_{j=1}^m \mathbb{E}[\mathbf{1}_{\tau_{t_i} < t_{i+1}, \tau_{t_i} = \tau_{t_i}^j} \mathbb{P}[X_{t_{i+1}} \notin D | \mathcal{F}_{\tau_{t_i}^j}]|\mathcal{F}_{t_i}]$ $\geq \frac{1}{2}\mathbb{P}[\tau_{t_i} < t_{i+1}|\mathcal{F}_{t_i}]$. Hence,

$$\sum_{i=0}^{N-1} \mathbb{P}_x[\tau^N > t_i, \tau_{t_i} \leq t_{i+1}] \leq 2\sum_{i=0}^{N-1} \mathbb{P}_x[\tau^N = t_{i+1}] \leq 2. \tag{4.2}$$

This last identity will be frequently used from now on to isolate the remainders; see, e.g., the last equality below.

Using **(S)** and Lemma 4.1 we obtain

$$\mathrm{Err}(T,h,f,x)=\sum_{i=0}^{N-1}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i,\tau_{t_i}<t_{i+1}}\{\nabla v(\tau_{t_i},X_{\tau_{t_i}})\cdot(\Pi_{\bar D}(X_{t_{i+1}})-X_{\tau_{t_i}})$$
$$+o(h^{1/2})\}]=\sum_{i=0}^{N-1}\sum_{j=1}^{m}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i,\tau^j_{t_i}<t_{i+1},\tau_{t_i}=\tau^j_{t_i}}\partial_{x_j}v(\tau^j_{t_i},X_{\tau^j_{t_i}})(X^j_{t_{i+1}}-b^j_0)^+]+o(h^{1/2}).$$

For the last equality we used the explicit expression of the projection on $\bar D$, namely, $\Pi_{\bar D}(y)=(b^1_0+(y_1-b^1_0)^+,\dots,b^m_0+(y_m-b^m_0)^+,y_{m+1},\dots,y_d)$ and also that $\forall(j,k)\in[\![1,m]\!]^2,\ j\neq k,\ \forall s\in[0,T],\ \partial_{x_k}v(s,y)|_{y\in\mathbb{R}^d:y_j=b^j_0}=0$. This is a simple consequence of the fact that v vanishes on D^c. By symmetry, assumption **(S)**, and the previous arguments we derive

$$\mathrm{Err}(T,h,f,x)=\sum_{j=1}^{m}\mathbb{E}_x[\mathbf{1}_{\tau^N\le T}\partial_{x_j}v(\tau^N,\Pi_{\bar D}(X_{\tau^N}))(X^j_{\tau^N}-b^j_0)^-]+R+o(h^{1/2}),\tag{4.3}$$

where

$$|R|:=\left|\sum_{i=0}^{N-1}\sum_{j=1}^{m}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i}\mathbf{1}_{\tau^j_{t_i}<t_{i+1},\tau_{t_i}\neq\tau^j_{t_i}}\partial_{x_j}v(\tau^j_{t_i},\Pi_{\bar D}(X_{\tau^j_{t_i}}))(X^j_{t_{i+1}}-b^j_0)^-]\right|$$
$$\le C\sqrt{h}\sum_{i=0}^{N-1}\sum_{j=1}^{m}\sum_{k\in[\![1,m]\!],\ k\neq j}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i}\mathbf{1}_{\tau^j_{t_i}<t_{i+1},\tau_{t_i}=\tau^k_{t_i}}|\partial_{x_j}v(\tau^j_{t_i},\Pi_{\bar D}(X_{\tau^j_{t_i}}))|].$$

Define $\forall\eta>0$, $V^{jk}(h^\eta):=\{y\in\mathbb{R}^d:|y_l-b^l_0|\le h^\eta,\ l\in\{j,k\}\}$. Put also $\mathcal{CO}^{jk}:=\{y\in\mathbb{R}^d:y_l=b^l_0,\ l\in\{j,k\}\}$. In short, $V^{jk}(h^\eta)$ is a neighborhood of width h^η of the corner $\mathcal{CO}^{jk}$. Note that $\forall(t,y)\in[0,T]\times\mathcal{CO}^{jk}\cap\bar D,\ \nabla v(t,y)=0$. Thus, from Lemma 4.1, assumption **(S)**, and (4.2) we derive that for h small enough

$$|R|\le C\sqrt{h}\sum_{i=0}^{N-1}\sum_{j=1}^{m}\sum_{k\in[\![1,m]\!],\ k\neq j}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i}\mathbf{1}_{X_{t_i}\in V^{jk}(h^{1/4})}\mathbf{1}_{\tau^j_{t_i}<t_{i+1},\tau_{t_i}=\tau^k_{t_i}}$$
$$\times|\partial_{x_j}v(\tau^j_{t_i},\Pi_{\bar D}(X_{\tau^j_{t_i}}))-\partial_{x_j}v(\tau^j_{t_i},\Pi_{\bar D}(\Pi_{\mathcal{CO}^{jk}}(X_{\tau^j_{t_i}})))|]+O_{pol}(h)$$
$$\le C\sqrt{h}\sum_{i=0}^{N-1}\sum_{j=1}^{m}\sum_{k\in[\![1,m]\!],\ k\neq j}\mathbb{E}_x[\mathbf{1}_{\tau^N>t_i}\mathbf{1}_{X_{\tau^j_{t_i}}\in V^{jk}(h^{1/8})}\mathbf{1}_{\tau^j_{t_i}<t_{i+1},\tau_{t_i}=\tau^k_{t_i}}$$
$$\times|X_{\tau^j_{t_i}}-\Pi_{\mathcal{CO}^{jk}}(X_{\tau^j_{t_i}})|^\alpha]+O_{pol}(h)=O(h^{\frac12+\frac{\alpha}{8}}):=o(h^{1/2}).$$

Plugging this last estimate into (4.3), we have

$$\mathrm{Err}(T,h,f,x)=\sum_{j=1}^{m}\mathbb{E}_x[\mathbf{1}_{\tau^N\le T}\partial_{x_j}v(\tau^N,\Pi_{\bar D}(X_{\tau^N}))(X^j_{\tau^N}-b^j_0)^-]+o(h^{1/2})$$
$$:=\sum_{j=1}^{m}E_j+o(h^{1/2}).\tag{4.4}$$

Remark 4.1. We emphasize that, up to now, we have not used the specific Brownian dynamics of the process X. The expansion (4.4) is valid for the error associated to the discretization of a diffusion process approximated by its discretely killed Euler scheme, provided that the process is nonadherent to the boundary and that **(S)** is fulfilled for the associated function v.

Step 2: Use of the asymptotic independence of the hitting time and the overshoot in the Brownian case. We now detail the asymptotic behavior of E_1. The other terms could be handled in exactly the same way. The following lemma, whose proof is postponed to Appendix A and strongly relies on the Brownian setting, is the main tool needed.

LEMMA 4.2. *Assume* **(BM)**, **(D)**. *Put* $\forall i \in [\![1,m]\!]$, $\tau^i := \inf\{t \geq 0 : X_t^i = b_0^i\}$, $\tau^{N,i} := \inf\{t_j := jh \geq 0 : X_{t_j}^i \leq b_0^i\}$. *One has* $\forall y \in \mathbb{R}^{+,*}$,

$$\mathbb{P}_x[\sqrt{h}^{-1}(X_{\tau^N}^1 - b_0^1)^- \geq y, \tau^N \leq t, \tau^{N,1} \leq \wedge_{i=2}^m \tau^{N,i}] \underset{N}{\longrightarrow} (1-H(y))\mathbb{P}_x[\tau^1 \leq t, \tau^1 < \wedge_{i=2}^m \tau^i],$$

where, using the notation of Theorem 2.1, $H(y) := (\mathbb{E}_0[s_{\tau^+}])^{-1}\int_0^y dz \mathbb{P}_0[s_{\tau^+} > z]$. *The limit is uniform on* $[0,T]$.

In order to isolate the rescaled overshoot $Z_N := \sqrt{h}^{-1}(X_{\tau^N}^1 - b_0^1)^-$ in E_1, we rewrite the components $X^2, \ldots, X^m$, of the correlated part of X in terms of X^1 and an additional correlated $(m-1)$-dimensional BM $\tilde{X}$ independent of X^1 and $(X^i)_{i\in[\![m+1,d]\!]}$. Namely, $\forall i \in [\![2,m]\!]$, $X_s^i = \rho_{1i}X_s^1 + (1-\rho_{1i}^2)^{1/2}\tilde{X}_s^{i-1}$, $\tilde{X}_0^{i-1} = (x_0^i - \rho_{1i}x_0^1)/(1-\rho_{1i}^2)^{1/2}$. Set also $(\rho_{1.}X_s^1 + (1-\rho_{1.}^2)^{1/2}\tilde{X}_s^{\cdot,-1})^{2,m} := (\rho_{12}X_s^1 + (1-\rho_{12}^2)^{1/2}\tilde{X}_s^1, \ldots, \rho_{1m}X_s^1 + (1-\rho_{1m}^2)^{1/2}\tilde{X}_s^{m-1}) = (X_s^2, \ldots, X_s^m) := X_s^{2,m}$, $X_s^{m+1,d} := (X_s^{m+1}, \ldots, X_s^d)$.

For notational convenience we introduce $\forall y \in \mathbb{R}^{d-1}$, $\Pi_{\bar{D}^{2,d}}(y) := (b_0^2 + (y_1 - b_0^2)^+, \ldots, b_0^m + (y_{m-1} - b_0^m)^+, y_m, \ldots, y_{d-1})$. The term E_1, defined in (4.4), writes

$$\begin{aligned}
&E_1 = \sqrt{h}\mathbb{E}_x[Z_N \mathbf{1}_{\tau^N \leq T}\partial_{x_1} v(\tau^N, b_0^1, \Pi_{\bar{D}^{2,d}}((\rho_{1.}(b_0^1 - \sqrt{h}Z_N)\\
&+(1-\rho_{1.}^2)^{1/2}\tilde{X}_{\tau^N}^{\cdot,-1})^{2,m}, X_{\tau^N}^{m+1,d}))]\\
&= \sqrt{h}\mathbb{E}_x[Z_N \mathbf{1}_{\tau^N \leq T}\partial_{x_1} v(\tau^N, b_0^1, \Pi_{\bar{D}^{2,d}}((\rho_{1.}b_0^1 + (1-\rho_{1.}^2)^{1/2}\tilde{X}_{\tau^N}^{\cdot,-1})^{2,m}, X_{\tau^N}^{m+1,d}))] + R_1,
\end{aligned}$$

where

$$\begin{aligned}
R_1 := \sqrt{h}\mathbb{E}_x[Z_N \mathbf{1}_{\tau^N \leq T}\big(&\partial_{x_1} v(\tau^N, b_0^1, \Pi_{\bar{D}^{2,d}}((\rho_{1.}(b_0^1 - \sqrt{h}Z_N)\\
&+(1-\rho_{1.}^2)^{1/2}\tilde{X}_{\tau^N}^{\cdot,-1})^{2,m}, X_{\tau^N}^{m+1,d}))\\
&-\partial_{x_1} v(\tau^N, b_0^1, \Pi_{\bar{D}^{2,d}}((\rho_{1.}b_0^1 + (1-\rho_{1.}^2)^{1/2}\tilde{X}_{\tau^N}^{\cdot,-1})^{2,m}, X_{\tau^N}^{m+1,d})))].
\end{aligned}$$

Under **(S)** the function $\partial_{x_1} v$ is continuous and bounded. Proposition 6 from [GM04] gives the uniform integrability of Z_N on the event $\tau^{N,1} \leq T$. We thus derive from Lemma 4.2 by convergence in law that for h small enough

$$E_1 = \sqrt{h}\mathbb{E}[Z]\mathbb{E}_x[\mathbf{1}_{\tau^1 \leq T, \wedge_{j=2}^m \tau^j > \tau^1}\partial_{x_1} v(\tau^1, X_{\tau^1})] + o(\sqrt{h}) + R_1,$$

where the distribution function of Z is given by H defined in Lemma 4.2. Recalling that $\mathbb{E}[Z] = \frac{\mathbb{E}[s_{\tau^+}^2]}{2\mathbb{E}[s_{\tau^+}]} = C_0$, we obtain $E_1 = \sqrt{h}C_0\mathbb{E}[\mathbf{1}_{\tau^1 \leq T, \wedge_{j=2}^m \tau^j > \tau^1}\partial_{x_1} v(\tau^1, X_{\tau^1})] + o(\sqrt{h}) + R_1$. Now, from assumption **(S)** we have $|R_1| \leq Ch^{\frac{1+\alpha}{2}}\mathbb{E}_x[Z_N^{1+\alpha}\mathbf{1}_{\tau^{N,1} \leq T}]$. Thus, by Proposition 6 in [GM04] we have $R_1 = O(h^{\frac{1+\alpha}{2}})$, which completes the proof.

Remark 4.2. The controls in the previous proof as well as the residual terms appearing in the computations are locally uniform w.r.t. the domain D.

Remark 4.3. To conclude this section, we would like to emphasize that the main difficulty in applying the previous theorem consists in finding conditions on f that guarantee that **(S)** is fulfilled. We provide some sufficient conditions in section 5 but in all generality this is far from easy.

4.3. Proof of Theorem 2.2 (correction result). In this subsection we detail how the arguments from Costantini, El Karoui, and Gobet (see [CKG03]) can be employed to prove our correction result. We write

$$\begin{aligned}\mathrm{Err}'(T,h,f,x) &= \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau^N_{D^h}>T}] - \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau_{D^h}>T}]\\ &+\mathbb{E}_x[f(X_T)\mathbf{1}_{\tau_{D^h}>T}] - \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau>T}] := E_1+E_2.\end{aligned}$$

From Remark 4.2 we derive that one could show just like in Theorem 2.1 that even though the domain depends on h we have $E_1 = C_1\sqrt{h} + o(\sqrt{h})$, where C_1 denotes the constant introduced in the quoted theorem. For E_2 we adapt some ideas from [CKG03] concerning the sensitivity of the Dirichlet problem w.r.t. the domain.

For a given $c \in \mathbb{R}^d$, let us denote $\forall \eta > 0,\ D_\eta := \{y \in \mathbb{R}^d : y - \eta c \in D\}$. We define $\tau_{D_\eta} := \inf\{s > 0 : X_s \notin D_\eta\}$ and we introduce $\forall\ x \in D$ the mapping $\mathcal{J}_c^x : \eta \longrightarrow \mathbb{E}_x[f(X_T)\mathbf{1}_{\tau_{D_\eta}>T}]$. We show below that under the assumptions of Theorem 2.2, the mapping $\mathcal{J}_c^x$ is differentiable in $\eta = 0$ and for $c = (\underbrace{1,\dots,1}_{m},\underbrace{0,\dots,0}_{d-m})$, one has

$$\begin{aligned}\partial_\eta \mathcal{J}_c^x(\eta)|_{\eta=0} &= -\mathbb{E}_x[\nabla v(\tau, X_\tau).c\mathbf{1}_{\tau<T}]\\ &= -\sum_{i=1}^{m}\mathbb{E}_x[\partial_{x_i} v(\tau^i, X_{\tau^i})\mathbf{1}_{\tau^i\le T,\wedge_{j\neq i}\tau^j>\tau^i}].\end{aligned} \tag{4.5}$$

From (4.5) we then derive that $E_2 = \mathcal{J}_c^x(C_0\sqrt{h}) - \mathcal{J}_c^x(0) = \partial_\eta \mathcal{J}_c^x(0)C_0\sqrt{h} + o(\sqrt{h}) = -C_1\sqrt{h} + o(\sqrt{h})$ which proves the theorem.

Proof of (4.5). Let us define $X_s^\eta := X_s - \eta c,\ \tau^{D,\eta} = \inf\{s > 0 : X_s^\eta \notin D\}$. Note that $\tau_{D_\eta} = \tau^{D,\eta}$. Denoting $\Delta_\eta := \mathbb{E}_x[f(X_T^\eta + \eta c)\mathbf{1}_{\tau^{D,\eta}>T}] - v(0,x)$, we have to identify the limit of Δ_η/η as $\eta \longrightarrow 0$. We have

$$\begin{aligned}\Delta_\eta &= \mathbb{E}_x[f(X_T^\eta + \eta c)\mathbf{1}_{\tau^{D,\eta}>T}] - \mathbb{E}_x[f(X_T^\eta)\mathbf{1}_{\tau^{D,\eta}>T}]\\ &+\mathbb{E}_x[f(X_T^\eta)\mathbf{1}_{\tau^{D,\eta}>T} - v(T\wedge\tau^{D,\eta}, X_{T\wedge\tau^{D,\eta}})]\\ &+\mathbb{E}_x[v(T\wedge\tau^{D,\eta}, X_{T\wedge\tau^{D,\eta}})] - v(0,x) := \Delta_{\eta,1}+\Delta_{\eta,2}+\Delta_{\eta,3}.\end{aligned}$$

Since $(M_t)_{t\in[0,T]} := (v(t\wedge\tau, X^{0,x}_{t\wedge\tau}))_{t\in[0,T]}$ is a martingale and $\tau^{D,\eta} < \tau$, we readily get $\Delta_{\eta,3} = 0$.

Note that $f(X_T^\eta)\mathbf{1}_{\tau^{D,\eta}>T} = v(T\wedge\tau^{D,\eta}, X^\eta_{T\wedge\tau^{D,\eta}})$. One also has $\tau_{D_\eta} \underset{\eta\longrightarrow 0,\ \text{a.s}}{\longrightarrow} \tau$. From assumption **(S)**, v is continuously differentiable. Thus, one gets $\lim_{\eta\longrightarrow 0}\Delta_{\eta,2}/\eta = -\mathbb{E}_x[\nabla v(T\wedge\tau, X_{T\wedge\tau}).c]$.

On the other hand, since we assumed f to be continuously differentiable we obtain $\lim_{\eta\longrightarrow 0}\Delta_{\eta,1}/\eta = \mathbb{E}_x[\nabla f(X_T).c\ \mathbf{1}_{\tau>T}]$. Recalling $\forall x \in \bar{D},\ v(T,x) = f(x)$ we write

$$\partial_\eta \mathcal{J}_c^x(\eta)|_{\eta=0} = -\mathbb{E}_x\big[\big(\nabla v(T\wedge\tau, X_{T\wedge\tau}) - \nabla f(X_T)\mathbf{1}_{\tau>T}\big).c\big] = -\mathbb{E}_x[\mathbf{1}_{\tau\le T}\nabla v(\tau, X_\tau).c]$$

which for $c = (\underbrace{1,\dots,1}_{m},\underbrace{0,\dots,0}_{d-m})$ proves (4.5). ∎

Remark 4.4. We mention that in the special case of a half space, an alternative proof based on the explicit expression of the hitting time densities is possible; see Chapter III of [Men04].

5. Some smoothness properties of v under (D): Sufficient conditions to fulfill (S). In the whole section we assume **(BM)**, **(D)**. We deal only with the case $d \geq 2, m \geq 2$. Indeed, for $m = 1$ the domain D is smooth and standard results, based on the explicit expression of the density of the killed BM, can be used to derive the required smoothness in **(S)**.

This section is divided into two parts. In subsection 5.1 we recall the explicit expression of the killed heat kernel under **(D)** and also state a control of its derivatives in dimension 2. Using these results and some standard PDE techniques we then derive in subsection 5.2 some sufficient conditions to get **(S)** when $d = m = 2$.

5.1. Explicit expression of the heat kernel under (D) and associated controls. Note first that

$$\mathbb{P}_x[\tau > t, X_t \in [y, y + dy)]/dy = |\det(\sigma_0^{-1})|\mathbb{P}_{\tilde{x}}[\tilde{\tau} > t, \tilde{X}_t \in [z, z + dz)]/dz, \tag{5.1}$$

where $\tilde{x} = \sigma_0^{-1}x,\ z = \sigma_0^{-1}y,\ \tilde{X}_t = \tilde{x} + W_t$ with W standard BM, $\tilde{D} := \{z \in \mathbb{R}^d : (\sigma_0 z)_i > b_0^i\}$, and $\tilde{\tau} := \inf\{s \geq 0 : \tilde{X}_s \notin \tilde{D}\}$.

Equation (5.1) gives the expression of the killed density of X under **(D)** in function of the density of the killed standard BM in a convex cone $\tilde{D}$ that writes as an intersection of half spaces. The domain $\tilde{D}$ is piecewise C^∞ and hence the trace of the cone on the unit sphere $\mathbb{S}^{d-1}$ with center at the vertex of $\tilde{D}$ is a normal domain in the sense of Chavel [Cha84] (see definition on page 16 of this reference). Denoting this trace by $\Gamma := \tilde{D} \cap \mathbb{S}^{d-1}$, we derive from page 169 of the above reference that we have a Sturm–Liouville spectral decomposition of the Laplace–Beltrami operator for the elliptic Dirichlet problem on Γ; i.e., the normalized eigenfunctions $(m_j)_{j\in\mathbf{N}^*}$ of $\Delta_{\mathbb{S}^{d-1}}$ form an orthonormal basis of $L^2(\Gamma)$ and the eigenvalues $(\lambda_j)_{j\in\mathbf{N}^*}$ are s.t. $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \uparrow +\infty$.

As a direct consequence of (2.2) in Bañuelos and Smits [BS97], we derive the following proposition.

PROPOSITION 5.1. *Let $\tilde{D}$ be a cone with origin 0 that writes as a nonempty intersection of half spaces. One has*

$$\begin{aligned}
&\forall (t, x, y) \in \mathbb{R}^{+*} \times \tilde{D}^2,\ x = r\theta,\ y = \rho\eta,\ (\theta, \eta) \in (\mathbb{S}^{d-1})^2,\ (\rho, r) \in (\mathbb{R}^{+*})^2,\\
&\mathbb{P}_x[\tilde{X}_t \in dy, \tilde{\tau} > t] = \frac{e^{-\frac{\rho^2+r^2}{2t}}}{t(\rho r)^{\frac{d}{2}-1}} \sum_{j=1}^{+\infty} I_{\nu_j}\left(\frac{\rho r}{t}\right) m_j(\theta) m_j(\eta) \rho^{d-1} d\rho d\sigma(\eta)\\
&:= \mathbf{q_t}(\mathbf{x}, \mathbf{y})\mathbf{dy},\ \nu_j = (\lambda_j + (\tfrac{d}{2} - 1)^2)^{1/2},
\end{aligned}$$

where λ_j (resp., m_j) are the eigenvalues (resp., the normalized eigenfunctions) of $\Delta_{\mathbb{S}^{d-1}}$ on Γ for the elliptic Dirichlet problem and I_ν denotes the modified Bessel function of order ν.

Remark 5.1. The result of Proposition 5.1 is standard in the bidimensional case. It is in that case a simple extension of the well-known method of images that consists, for special angles, in writing the killed heat kernel as a suitable sum of standard Gaussian kernels alternating heat sources and sinks in order to satisfy the boundary conditions. We refer to Carslaw and Jaeger [CJ59] or to Iyengar [Iye85] for details.

Remark 5.2. In the special case $d = m = 2$ the eigenvalues (resp., the normalized eigenfunctions) write $\lambda_j = (\pi j/\omega)^2$ (resp., $m_j(\theta) = \sqrt{\frac{2}{\omega}} \sin(\frac{\pi j}{\omega} \arg(\theta))$), where $\omega \in (0, 2\pi)$ is the angle of the cone. For $m > 2$, we do not have such an explicit expression, but analysis techniques (see Weyl's lemma [Cha84, p. 172]) give some controls on the behavior of these eigenvalues; see also Remark B.2.

LEMMA 5.2 (radial control of the derivatives when $\mathbf{d = m = 2}$). *For a given* $\omega \in (0, \pi)$, *let* $\tilde{D} := \{x = (r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : r > 0,\ \theta \in (0, \omega)\}$. *Note that* $\tilde{D}$ *is a convex cone.* $\forall\ R > 0, T > 0$, *there exist positive constants* $C := C(R, T), c, \xi$ *s.t.* $\forall (t, x, y) \in (0, T] \times (\tilde{D} \cap B(0, R)) \times \tilde{D}$, $x = (r\cos\theta, r\sin\theta)$, $y = (\rho\cos\eta, \rho\sin\eta)$, $(\theta, \eta) \in (0, \omega)^2$, $(\rho, r) \in (\mathbb{R}^{+*})^2$, *one has*

$$q_t(x, y) + |\partial_t q_t(x, y)| + |\nabla_x q_t(x, y)| \le \frac{C}{t^\xi} \exp\left(-c\frac{|r - \rho|^2}{t}\right)$$

and $\exists \alpha_0 := \alpha_0(\omega) > 0$,

$$\sup_{(x,\, x') \,\in\, (\tilde{D} \,\cap\, B(0,\, R))^2} \frac{|\nabla q_t(x, y) - \nabla q_t(x', y)|}{|x - x'|^{\alpha_0}} \le \frac{C}{t^\xi} \exp\left(-c\frac{|r - \rho|^2 \wedge |r' - \rho|^2}{t}\right).$$

The proof of the above lemma is postponed to Appendix B.

5.2. Derivation of (S) when $d = m = 2$. In Remark 5.1 we mentioned that for special angles of the cone, one could express the killed heat kernel q in terms of a sum of standard Gaussian kernels. To be precise, this can be done when the angle of the cone writes $\omega = \pi/m_0, m_0 \in \mathbf{N}^*$. For our original problem (4.1), one can establish a connection between the killed heat kernel q and the density of the killed BM in the orthant thanks to (5.1). One therefore deduces that for some particular correlation coefficients, corresponding to angles that have the previous form, under suitable assumptions on the final condition f one has the "usual" smoothness properties for the solution v of problem (4.1), and hence **(S)** is satisfied. We now give a smoothness result for the solution v of (4.1) for general correlation coefficients. Using the notation of section 2.1, we introduce the following assumption:

(F) The function $f \in C_b^{2+\alpha}(\bar{D})$, $\alpha > 0$, $f|_{\partial D} = \mathrm{Tr}(H_f \sigma_0 \sigma_0^*)|_{\partial D} = 0$, and $d(\mathrm{supp}(f), b_0) \ge 2\varepsilon > 0$.

PROPOSITION 5.3. *Assume* **(D)**, **(BM)**, **(F)**. *For* $D := \{x \in \mathbb{R}^2 : x_1 > b_0^1, x_2 > b_0^2\}$, $\sigma_0\sigma_0^* = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $\rho \in (-1, 1)$, *there exists* $\alpha' > 0$ *s.t. the unique solution* v *of* (4.1) *belongs to* $C_b^{1/2+\alpha'/2, 1+\alpha'}([0, T] \times \bar{D})$. *In particular,* **(S)** *is satisfied.*

Proof. From Proposition 5.1 we derive that problem (4.1) has a unique solution $v \in C^{1,2}([0, T) \times D) \cap C_b^0([0, T] \times \bar{D})$, where $C_b^0([0, T] \times \bar{D})$ denotes the space of bounded continuous functions on $[0, T] \times \bar{D}$.

Let us now note that as a consequence of the support condition in **(F)** and the radial control of Lemma 5.2 there exists $\alpha_0 > 0$ s.t. $v \in C_b^{1/2+\alpha_0/2, 1+\alpha_0}([0, T] \times B(b_0, \varepsilon) \cap \bar{D})$. Choose now D_1 to be a C^3 domain s.t. $d(\bar{D}_1, b_0) \ge \varepsilon/3 > 0$ and $\{x \in \mathbb{R}^2 : |x - b_0| \ge \varepsilon, x \in \partial D\} = \{x \in \mathbb{R}^2 : |x - b_0| \ge \varepsilon, x \in \partial D_1\}$. From the techniques used in Chapter IV of Friedman [Fri64] to prove the boundary Schauder estimates and Theorem 5.2 in Chapter 4 in [LSU68], we derive that $v \in C_b^{1+\alpha/2, 2+\alpha}([0, T] \times \bar{D}_1)$. Put $\alpha' := \alpha \wedge \alpha_0$. The proof is complete. □

6. Conclusion. In this paper we obtained an expansion result for the weak error in the special case of a discretely killed Brownian motion in an orthant provided

we had smoothness properties of the solution of the underlying Cauchy–Dirichlet problem. We exploited the explicit asymptotic distribution of the overshoot above the boundary that had previously been characterized as the leading term of the weak error; see [GM04]. Finally, the correction method we introduced has given promising results. A natural question concerns its possible extension to a wider framework than the Brownian one. The theoretical analysis of Algorithm 3.1 introduced to this end will concern further research.

The main motivation that led us to deal with conical cases comes from mathematical finance. Indeed, with multiassets, one often defines the domain of a barrier option as a product domain. For the moment we are only able to treat in whole generality the case of bidimensional domains in a Black–Scholes framework.

Concerning further extensions in bigger dimensions, the key point concerns the smoothness properties of the underlying function $v(t,x) = \mathbb{E}_x[f(X_{T-t})\mathbf{1}_{\tau>T-t}]$. Anyhow, for some special angles, or equivalently for special correlation coefficients, we can extend the method of images to express the transition density as a sum of standard Gaussian kernels. In that case, under suitable assumptions on f, we have the usual smoothness properties on v, and both the expansion and correction results hold true.

Let us mention that the proof of the main results would work if v had a uniform Hölder continuous first spatial derivative with exponential growth only in a neighborhood of the boundary. This could allow us to relax the boundedness assumption on f.

Appendix A. Asymptotic behavior of the overshoot. This section is dedicated to the proof of Lemma 4.2 introduced in section 4.2 concerning the asymptotic behavior of the overshoot. In the following, we freely use the notation introduced in Theorem 2.1 and Lemma 4.2.

A.1. Asymptotic independence of the overshoot and the exit time. We first state a one-dimensional result due to Siegmund [Sie79].

LEMMA A.1 (asymptotic independence of the overshoot and the discrete exit time). *Let W be a standard linear BM. Put $x > 0$ and consider the domain $D :=]-\infty, x[$. We have for any $y \geq 0$*

$$\lim_{h \longrightarrow 0} \mathbb{P}_0[\tau^N \leq t, (W_{\tau^N} - x) \leq y\sqrt{h}] = \mathbb{P}_0[\tau \leq t] H(y). \tag{A.1}$$

The limit is uniform in $t \in [0,T]$.

Proof. Equation (A.1) is a direct consequence of Lemma 3 in [Sie79] for a fixed t. We derive the uniformity on $[0,T]$ using Dini-like arguments noting that the left-hand side of (A.1) defines a sequence of (discontinuous) increasing functions and that the simple limit is continuous (see, e.g., problem 7.2.3 in [Die71] or subsection A.2). ∎

From now on, we assume $m \geq 2$ and proceed to the proof.

Proof of Lemma 4.2. Let us first show that $\forall (t,y) \in [0,T] \times \mathbb{R}^{+,*}$, $\zeta_N(t) := \mathbb{P}_x[\sqrt{h}^{-1}(X^1_{\tau^{N,1}} - b^1_0)^- \geq y, \tau^N \leq t, \tau^{N,1} \leq \wedge_{i=2}^m \tau^{N,i}] \underset{N}{\longrightarrow} (1 - H(y))\mathbb{P}_x[\tau^1 \leq t, \wedge_{i=2}^m \tau^i > \tau^1] := \zeta(t)$. We write

$$\begin{aligned} \zeta_N(t) &= \mathbb{P}_x[\sqrt{h}^{-1}(X^1_{\tau^{N,1}} - b^1_0)^- \geq y, \tau^{N,1} \leq t] \\ &\quad -\mathbb{P}_x[\sqrt{h}^{-1}(X^1_{\tau^{N,1}} - b^1_0)^- \geq y, \tau^{N,1} \leq t, \tau^{N,1} > \wedge_{i=2}^m \tau^{N,i}] \\ &:= (\zeta^1_N - \zeta^2_N)(t). \end{aligned} \tag{A.2}$$

From Lemma A.1 one gets

$$\zeta_N^1(t) \underset{N}{\longrightarrow} \zeta^1(t) := (1-H(y))\mathbb{P}_x[\tau^1 \le t] \tag{A.3}$$

uniformly on $[0,T]$. Let us turn to the control of ζ_N^2. As a consequence of the strong Markov property of X, we have

$$\begin{aligned}\zeta_N^2(t) &= \mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m \tau^{N,i} \le t}\mathbf{1}_{\tau^{N,1}>\wedge_{i=2}^m \tau^{N,i}}\mathbb{P}[\sqrt{h}^{-1}(X^1_{\tau^{N,1}} - b_0^1)^- \ge y, \tau^{N,1} \le t | \mathcal{F}_{\wedge_{i=2}^m \tau^{N,i}}]] \\ &:= \mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m \tau^{N,i} \le t}\mathbf{1}_{\tau^{N,1}>\wedge_{i=2}^m \tau^{N,i}}\xi_N(X^1_{\wedge_{i=2}^m \tau^{N,i}}, \wedge_{i=2}^m \tau^{N,i}, t)]\end{aligned}$$

with $\xi_N(X^1_{\wedge_{i=2}^m \tau^{N,i}}, \wedge_{i=2}^m \tau^{N,i}, t) = \mathbb{P}_{X^1_{\wedge_{i=2}^m \tau^{N,i}}}[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}} - b_0^1)^- \ge y, \tilde{\tau}^{N,1} \le t - \wedge_{i=2}^m \tau^{N,i}]$, where $(\tilde{X}^1_t)_{t\ge 0}$ is a standard BM with starting point $X^1_{\wedge_{i=2}^m \tau^{N,i}}$ and $\tilde{\tau}^{N,1} := \inf\{t_i := ih \ge 0 : \tilde{X}^1_{t_i} \le b_0^1\}$.

For a given arbitrary compact interval $\mathcal{K} := [\underline{\mathcal{K}}, \overline{\mathcal{K}}] \subset (b_0^1, +\infty)$ we split $\zeta_N^2(t)$ into two parts.

$$\begin{aligned}\zeta_N^2(t) &= \mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m \tau^{N,i} \le t}\mathbf{1}_{\tau^{N,1}>\wedge_{i=2}^m \tau^{N,i}}\mathbf{1}_{X^1_{\wedge_{i=2}^m \tau^{N,i}} \in \mathcal{K}}\xi_N(X^1_{\wedge_{i=2}^m \tau^{N,i}}, \wedge_{i=2}^m \tau^{N,i}, t)] \\ &+ \mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m \tau^{N,i} \le t}\mathbf{1}_{\tau^{N,1}>\wedge_{i=2}^m \tau^{N,i}}\mathbf{1}_{X^1_{\wedge_{i=2}^m \tau^{N,i}} \notin \mathcal{K}}\xi_N(X^1_{\wedge_{i=2}^m \tau^{N,i}}, \wedge_{i=2}^m \tau^{N,i}, t)] \\ &:= \zeta_N^{21}(t) + \zeta_N^{22}(t).\end{aligned}$$

Fix $\varepsilon > 0$. We now show that one can choose $\mathcal{K}(\varepsilon),\ N_0 := N_0(\varepsilon, \mathcal{K}(\varepsilon))$ s.t. for $N \ge N_0$,

$$\zeta_N^2(t) = (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t] + O(\varepsilon). \tag{A.4}$$

Control of $\zeta_N^{21}(t)$. Write first

$$\begin{aligned}\zeta_N^{21}(t) &= \Big(\zeta_N^{21}(t) - (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t, X^1_{\wedge_{i=2}^m \tau^i} \in \mathcal{K}]\Big) \\ &\quad + (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t] - R(t,\mathcal{K}),\end{aligned}$$

where $R(t,\mathcal{K}) = (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t, X^1_{\wedge_{i=2}^m \tau^i} \notin \mathcal{K}]$. Note that

$$\begin{aligned}0 \le R(t,\mathcal{K}) &\le \mathbb{P}_x\big[\wedge_{i=2}^m \tau^i \le T, X^1_{\wedge_{i=2}^m \tau^i} \ge \overline{\mathcal{K}}\big] \\ &\quad + \mathbb{P}_x\big[\wedge_{i=2}^m \tau^i \le T, X^1_{\wedge_{i=2}^m \tau^i} \in (b_0^1, \underline{\mathcal{K}}]\big] := R_1(\overline{\mathcal{K}}) + R_2(\underline{\mathcal{K}}).\end{aligned}$$

Lemma 4.1 readily gives $R_1(\overline{\mathcal{K}}) \le C\exp(-c\frac{(\overline{\mathcal{K}}-x_1)^2}{T})$. On the other hand, $R_2(\underline{\mathcal{K}}) \underset{\underline{\mathcal{K}} \longrightarrow b_0^1}{\longrightarrow} 0$. Hence, for $\varepsilon > 0$ we can choose $\mathcal{K} = \mathcal{K}(\varepsilon)$ s.t.

$$\begin{aligned}\zeta_N^{21}(t) &= \Big(\zeta_N^{21}(t) - (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t, X^1_{\wedge_{i=2}^m \tau^i} \in \mathcal{K}]\Big) \\ &\quad + (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t] + O(\varepsilon) \\ &:= \delta_N(t) + (1-H(y))\mathbb{P}_x[\wedge_{i=2}^m \tau^i < \tau^1, \tau^1 \le t] + O(\varepsilon).\end{aligned}$$

For the term $\delta_N(t)$ we introduce the following lemma whose proof is postponed to the end of the section.

LEMMA A.2. *Let $\tilde{X}^1$ be a standard BM with starting point $\tilde{x}$ in a given compact interval $\mathcal{K} = [\underline{\mathcal{K}}, \overline{\mathcal{K}}] \subset (b_0^1, +\infty)$. Then*

$$\mathbb{P}_{\tilde{x}}[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}} - b_0^1)^- \ge y, \tilde{\tau}^{N,1} \le u] \underset{N}{\longrightarrow} (1-H(y))\mathbb{P}_{\tilde{x}}[\tilde{\tau}^1 \le u]$$

uniformly on $(\tilde{x},u)\in\mathcal{K}\times[0,T]$.

From Lemma A.2, $\forall\varepsilon>0,\ \exists\tilde{N}_0:=\tilde{N}_0(\mathcal{K}(\varepsilon),\varepsilon)$, s.t. $N\geq\tilde{N}_0$

$$\begin{aligned}
&\delta_N(t)=(1-H(y))\Big\{\mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m\tau^{N,i}\leq t}\mathbf{1}_{\wedge_{i=2}^m\tau^{N,i}<\tau^{N,1}}\mathbf{1}_{X^1_{\wedge_{i=2}^m\tau^{N,i}}\in\mathcal{K}}\\
&\times\mathbb{P}_{X^1_{\wedge_{i=2}^m\tau^{N,i}}}[\tilde{\tau}^1\leq t-\wedge_{i=2}^m\tau^{N,i}]]-\mathbb{P}_x[\wedge_{i=2}^m\tau^i<\tau^1,\tau^1\leq t,X^1_{\wedge_{i=2}^m\tau^i}\in\mathcal{K}]\Big\}+O(\varepsilon)\\
&:=(1-H(y))\Big(\mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m\tau^{N,i}\leq t}\mathbf{1}_{\wedge_{i=2}^m\tau^{N,i}<\tau^{N,1}}\mathbf{1}_{X^1_{\wedge_{i=2}^m\tau^{N,i}}\in\mathcal{K}}\xi_t(X^1_{\wedge_{i=2}^m\tau^{N,i}},\wedge_{i=2}^m\tau^{N,i})]\\
&-\mathbb{E}_x[\mathbf{1}_{\wedge_{i=2}^m\tau^i\leq t}\mathbf{1}_{\wedge_{i=2}^m\tau^i<\tau^1}\mathbf{1}_{X^1_{\wedge_{i=2}^m\tau^i}\in\mathcal{K}}\xi_t(X^1_{\wedge_{i=2}^m\tau^i},\wedge_{i=2}^m\tau^i)]\Big)+O(\varepsilon).
\end{aligned}$$

Note that $\xi_t(x,u):=\mathbb{P}_x[\tilde{\tau}\leq t-u]$ is continuous in $(x,u)\in(b_0^1,+\infty)\times[0,t]$. Recall that $\tau^{N,i}\xrightarrow[N]{\text{a.s}}\tau^i,\ i\in[\![1,m]\!]$, and by continuity $X^1_{\wedge_{i=2}^m\tau^{N,i}}\xrightarrow[N]{\text{a.s}}X^1_{\wedge_{i=2}^m\tau^i}$. One can check that the law of $(\tau^1,\wedge_{i=2}^m\tau^i,X^1_{\wedge_{i=2}^m\tau^i})$ is absolutely continuous w.r.t. the Lebesgue measure. We thus derive by convergence in law that for N large enough

$$\delta_N(t)\ =\ O(\varepsilon),\ \zeta_N^{21}(t)=(1-H(y))\mathbb{P}_x[\wedge_{i=2}^m\tau^i<\tau^1,\tau^1\leq t]+O(\varepsilon). \tag{A.5}$$

Control of $\zeta_N^{22}(t)$. The arguments we use to control this term are quite similar to those introduced to treat the terms $R_1(\overline{\mathcal{K}}),R_2(\underline{\mathcal{K}})$ above.

Indeed, since $\xi_N\in[0,1]$ one gets

$$\begin{aligned}
\zeta_N^{22}(t)\ &\leq\ \mathbb{P}_x[\wedge_{i=2}^m\tau^{N,i}\leq T,X^1_{\wedge_{i=2}^m\tau^{N,i}}\geq\overline{\mathcal{K}}]+\mathbb{P}_x[\wedge_{i=2}^m\tau^{N,i}\leq T,X^1_{\wedge_{i=2}^m\tau^{N,i}}\in(b_0^1,\underline{\mathcal{K}}]]\\
&:=\ R_1^N(\overline{\mathcal{K}})+R_2^N(\underline{\mathcal{K}}).
\end{aligned}$$

From Lemma 4.1 we get $R_1^N(\overline{\mathcal{K}})\leq C\exp(-c\frac{(\overline{\mathcal{K}}-x_1)^2}{T})$. The previous choice of $\overline{\mathcal{K}}$ gives $R_1^N(\overline{\mathcal{K}})=O(\varepsilon)$. Write now $R_2^N(\underline{\mathcal{K}}):=(R_2^N(\underline{\mathcal{K}})-R_2(\underline{\mathcal{K}}))+R_2(\underline{\mathcal{K}})$. On the one hand, the former choice of $\underline{\mathcal{K}}$ yields $R_2(\underline{\mathcal{K}})=O(\varepsilon)$. On the other hand, for the difference $(R_2^N-R_2)(\underline{\mathcal{K}})$, since $\tau^{N,i}\xrightarrow[N]{\text{a.s}}\tau^i,\ i\in[\![2,m]\!],\ X^1_{\wedge_{i=2}^m\tau^{N,i}}\xrightarrow[N]{\text{a.s}}X^1_{\wedge_{i=2}^m\tau^i}$, with the same arguments we employed to control $\delta_N(t)$, we derive by convergence in law $\exists N_0:=N_0(\underline{\mathcal{K}},\varepsilon),\ N\geq N_0,\ |(R_2^N-R_2)(\underline{\mathcal{K}})|\leq\varepsilon$. Hence, for $N:=N(\mathcal{K},\varepsilon)$ large enough, we write $\zeta_N^{22}(t)=O(\varepsilon)$ which together with (A.5) gives (A.4). From (A.4), (A.3), and (A.2) we derive the simple convergence of ζ_N to ζ for a fixed $t\in[0,T]$.

The uniformity in $t\in[0,T]$ derives from the fact that $\zeta_N(t)$ is a cumulative distribution function with continuous limit; see also the arguments at the beginning of the proof of Lemma A.1. □

A.2. Proof of Lemma A.2. Let us define $\forall(x,u)\in(\mathcal{K}:=[\underline{\mathcal{K}},\overline{\mathcal{K}}])\times[0,T],\ \underline{\mathcal{K}}>b_0^1,\ \Psi_N(x,u)=\mathbb{P}_x[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}}-b_0^1)^-\geq y,\tilde{\tau}^{N,1}\leq u]$. For a fixed $x\in\mathcal{K}$, Lemma A.1 yields that $\Psi_N(x,u)\xrightarrow[N]{}(1-H(y))\mathbb{P}_x[\tau^1\leq u]:=\Psi(x,u)$ uniformly on $u\in[0,T]$. Let us now show that for a fixed $u\in[0,T]$ we have the uniform convergence w.r.t. $x\in\mathcal{K}$. Write

$$\begin{aligned}
&\Psi_N(x,u)=\mathbb{P}_x[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}}-b_0^1)^-\geq y]-\mathbb{P}_x[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}}-b_0^1)^-\geq y,\tilde{\tau}^{N,1}>u]\\
&:=\Psi_N^1(x)-\Psi_N^2(x,u).
\end{aligned}$$

With the notation of Theorem 2.1, introducing $\forall a \geq 0, \bar{\tau}_a := \inf\{n \in \mathbf{N} : s_n > a\}$, we write $\Psi_N^1(x) = \mathbb{P}_x[\sqrt{h}^{-1}(\tilde{X}^1_{\tilde{\tau}^{N,1}} - b_0^1)^- \geq y] = \mathbb{P}_0[(s_{\bar{\tau}_{(x-b_0^1)/\sqrt{h}}} - (x-b_0^1)/\sqrt{h}) \geq y]$. Equation (19) from [Sie79] gives $\lim_{b\to\infty} \mathbb{P}_0[s_{\bar{\tau}_b} - b \geq y] = 1 - H(y)$. Hence, $\Psi_N^1(x) \underset{N}{\longrightarrow} (1 - H(y))$ uniformly on $x \in \mathcal{K}$. We develop Ψ_N^2 like in the proof of Lemma 3 from the same reference, controlling that we can isolate uniform rests. Recalling $\phi(u) := \inf\{t_i \geq 0 : t_i \leq u < t_{i+1}\}$, we get

$$\begin{aligned}\Psi_N^2(x,u) &= \mathbb{P}[(s_{\bar{\tau}_{(x-b_0^1)/\sqrt{h}}} - (x-b_0^1)/\sqrt{h}) \geq y, \bar{\tau}_{(x-b_0^1)/\sqrt{h}} > \phi(u)/h] \\ &= \int_0^\infty \mathbb{P}[\bar{\tau}_{(x-b_0^1)/\sqrt{h}} > \phi(u)/h, (x-b_0^1)/\sqrt{h} - s_{\phi(u)/h} \in [z, z+dz)] \\ &\quad \times \mathbb{P}[s_{\bar{\tau}_z} - z \geq y].\end{aligned}$$

We split the above integral into three terms $\Psi_N^{21}, \Psi_N^{22}, \Psi_N^{23}$ respectively associated to the intervals $(0, \varepsilon(x-b_0^1)/\sqrt{h})$, $(\varepsilon(x-b_0^1)/\sqrt{h}, (x-b_0^1)/(\varepsilon\sqrt{h}))$, $((x-b_0^1)/(\varepsilon\sqrt{h}), \infty)$ for an arbitrary $\varepsilon \in (0,1)$. One has

$$\begin{aligned}\Psi_N^{21}(x,u) &\leq \mathbb{P}\left[\frac{(1-\varepsilon)(x-b_0^1)}{\sqrt{h}\sqrt{\phi(u)/h}} \leq \mathcal{N}(0,1) \leq \frac{x-b_0^1}{\sqrt{h}\sqrt{\phi(u)/h}}\right] \\ &\leq \mathbb{P}\left[\frac{(1-\varepsilon)(x-b_0^1)}{T^{1/2}} \leq \mathcal{N}(0,1) \leq \frac{x-b_0^1}{T^{1/2}}\right] \leq \tfrac{C\varepsilon(\overline{\mathcal{K}}-b_0^1)}{T^{1/2}}\end{aligned}$$

uniformly for $x \in \mathcal{K}$. We also have

$$\begin{aligned}\Psi_N^{23}(x,u) &\leq \mathbb{P}\left[\mathcal{N}(0,1) \leq (1-\varepsilon^{-1})\frac{x-b_0^1}{\phi(u)^{1/2}}\right] \leq \mathbb{P}\left[\mathcal{N}(0,1) \leq (1-\varepsilon^{-1})\frac{\underline{\mathcal{K}}-b_0^1}{T^{1/2}}\right] \\ &\leq \frac{CT^{1/2}}{\underline{\mathcal{K}}-b_0^1}\frac{\varepsilon}{1-\varepsilon},\end{aligned}$$

which is still uniform w.r.t. $x \in \mathcal{K}$. From these computations we derive that for N large enough, $\Psi_N^{22}(x,u) = (1-H(y))\mathbb{P}_x[\tilde{\tau}^{N,1} > u] + O(\varepsilon)$, where the rest is uniform w.r.t. $\mathcal{K}$. It therefore remains to show $\mathbb{P}_x[\tilde{\tau}^{N,1} > u] := \gamma_N(u,x) \underset{N}{\longrightarrow} \gamma(u,x) := \mathbb{P}_x[\tilde{\tau}^1 > u]$ uniformly on $\mathcal{K}$. We note that $1 - \gamma_N(u,x) = \mathbb{P}_0[\sup_{i\in[\![0,\phi(u)/h]\!]} \tilde{X}^1_{t_i} \geq (x-b_0^1)]$ is decreasing in x, so that $\gamma_N(u,.)$ is increasing. Since the simple limit is continuous, we derive the uniformity using the same arguments as in the proof of Lemma A.1.

Now, we have shown that for a fixed parameter $x \in \mathcal{K}$, $u \in [0,T]$, we have the uniform convergence w.r.t. the other. Let us now show the joint uniform convergence. The limit Ψ is uniformly continuous on $\mathcal{K} \times [0,T]$. This reads

(A.6)
$$\forall \varepsilon > 0, \exists \eta := \eta(\varepsilon), \forall (x,x') \times (t,t') \in \mathcal{K}^2 \times [0,T]^2,\ |t-t'| + |x-x'| \leq \eta, \\ |\Psi(x,t) - \Psi(x',t')| \leq \varepsilon.$$

In particular, $|t-t'| \leq \eta \Rightarrow \sup_{x\in\mathcal{K}} |\Psi(x,t) - \Psi(x,t')| \leq \varepsilon$. Let us now consider a regular grid $\Lambda := \{s_i\}_{i\in[\![1,a]\!]}$ of $[0,T]$ with step $s = s_{i+1} - s_i \leq \eta$. Since for a fixed $t \in [0,T]$ we have uniform convergence in space,

(A.7) $$\forall \varepsilon > 0,\ \exists \tilde{N}_0 = \max_{i\in[\![1,a]\!]} \tilde{N}_0(s_i),\ N \geq \tilde{N}_0,\ \sup_{i\in[\![1,a]\!]} \sup_{x\in\mathcal{K}} |\Psi_N(x,s_i) - \Psi(x,s_i)| \leq \varepsilon.$$

Noting that both $\Psi_N(x,.), \Psi(x,.)$ are increasing functions we derive from (A.6), (A.7)

$$\begin{array}{l}\forall t \in [s_i, s_{i+1}],\ \Psi(x,s_i) - \Psi(x,s_{i+1}) + \Psi(x,s_{i+1}) - \Psi_N(x,s_{i+1}) \\ \leq \Psi(x,t) - \Psi_N(x,t) \leq \Psi(x,s_{i+1}) - \Psi(x,s_i) + \Psi(x,s_i) - \Psi_N(x,s_i), \\ \forall \varepsilon > 0, \exists N_0,\ N \geq N_0,\ \sup_{t\in[0,T]} \sup_{x\in\mathcal{K}} |\Psi(x,t) - \Psi_N(x,t)| \leq \varepsilon,\end{array}$$

which shows the joint uniformity and completes the proof.

Appendix B. Results about the killed heat kernel: Proof of Lemma 5.2. One of the key tools in the proof of the lemma is the following identity:

$$\forall x > 0,\ \forall \mu > \nu \geq 0,\ I_\mu(x) < I_\nu(x). \tag{B.1}$$

Relation (B.1) was proved by Jones in [Jon68]. From identity 9.6.34 in Abramowitz and Stegun we also get that $\exp(z) = I_0(z) + 2\sum_{n=1}^{\infty} I_n(z)$. Hence, from the explicit expression of the killed heat kernel (see Proposition 5.1 and Remark 5.2) and recalling that $\nu_n := n\pi/\omega > n$, (B.1) yields

$$q_t(x,y) \leq \frac{2\exp(-\frac{r^2+\rho^2}{2t})}{t\omega} \sum_{n=1}^{\infty} I_{\nu_n}\left(\frac{r\rho}{t}\right) \leq \frac{2\exp(-\frac{r^2+\rho^2}{2t})}{t\omega} \sum_{n=1}^{\infty} I_n\left(\frac{r\rho}{t}\right) \leq \frac{\exp(-\frac{|r-\rho|^2}{2t})}{t\omega}.$$

Put $A_t := \sum_{n=1}^{\infty} \partial_t(\sin(\nu_n\theta)\sin(\nu_n\eta) I_{\nu_n}\left(\frac{r\rho}{t}\right))$. From the recurrence relations on modified Bessel functions (see formula 9.6.26 in [AS72]) one gets $|A_t| \leq \frac{r\rho}{2t^2}\sum_{n=1}^{\infty}(I_{\nu_n+1} + I_{\nu_n-1})\left(\frac{r\rho}{t}\right) \leq C\frac{r\rho}{t^2}\exp\left(\frac{r\rho}{t}\right)$. Thus,

$$\begin{array}{rcl}\exp\left(-\frac{r^2+\rho^2}{2t}\right)|A_t| & \leq & C\frac{r\rho}{t^2}\exp\left(-\frac{|r-\rho|^2}{2t}\right) \leq C(\frac{R^2}{t^2} + \frac{R|r-\rho|}{t^2})\exp\left(-c\frac{|r-\rho|^2}{2t}\right) \\ & \leq & C(\frac{R^2}{t^2} + \frac{R}{t^{3/2}})\exp\left(-c\frac{|r-\rho|^2}{t}\right) \leq \frac{C}{t^\xi}\exp\left(-c\frac{|r-\rho|^2}{t}\right),\end{array}$$

which gives the result for the time derivative.

The boundedness and Hölder continuity of the gradient is somehow trickier to obtain. Let us show these properties for the partial derivative of the heat kernel w.r.t. the first parameter. They could be obtained for the other one exactly in the same way. Bare hand calculations yield

$$\begin{array}{rcl}\partial_{x_1} q_t(x,y) & = & -\frac{x_1}{t} q_t(x,y) + \frac{\rho}{t^2\omega}\exp\left(-\frac{r^2+\rho^2}{2t}\right)\sum_{n=1}^{\infty}\sin(\nu_n\eta) \\ & & \times\left\{\sin((\nu_n-1)\theta) I_{\nu_n-1}\left(\frac{r\rho}{t}\right) + \sin((\nu_n+1)\theta) I_{\nu_n+1}\left(\frac{r\rho}{t}\right)\right\}.\end{array}$$

The previous arguments give the stated control for $|\partial_{x_1} q_t(x,y)|$.

Now, the most "singular" term in the expression of $\partial_{x_1} q_t(x,y)$ is the one involving the modified Bessel functions of lowest order. Thus, we have to prove the Hölder continuity of

$$\begin{array}{rcl} g_t(x,y) & := & \frac{\rho}{t^2\omega}\exp\left(-\frac{r^2+\rho^2}{2t}\right)\sum_{n=1}^{\infty}\sin(\nu_n\eta)\sin((\nu_n-1)\theta) I_{\nu_n-1}\left(\frac{r\rho}{t}\right) \\ & := & \frac{\rho}{t^2\omega}\exp\left(-\frac{r^2+\rho^2}{2t}\right) B_t(x,y).\end{array}$$

Still by direct computation we get that

$$|\nabla_x B_t(x,y)| \leq \frac{C\rho}{t}\left\{\exp\left(\frac{r\rho}{t}\right) + \sum_{n=1}^{\infty} I_{\nu_n-2}\left(\frac{r\rho}{t}\right)\right\}.$$

Now, since $z \in \mathbb{R}^{+*}$, from equation 9.6.20 in [AS72]

$$I_\nu(z) = \frac{1}{\pi}\int_0^\pi \exp(z\cos(\gamma))\cos(\nu\gamma)d\gamma - \frac{\sin(\nu\pi)}{\pi}\int_0^\infty \exp(-z\cosh(t)-\nu t)dt.$$

From this expression it is easily seen that $\forall n \in \mathbf{N},\ z>0,\ I_n(z) = I_{-n}(z)$, and $\forall \nu > 0,\ \forall \varepsilon > 0$

$$\begin{aligned}&|I_\nu - I_{-\nu}|(z) \le C\int_0^\infty \exp(-z\cosh(t))\exp(\nu t)dt\\ &= C\int_0^1 \exp\left(-\frac{z}{2}(u^{-1}+u)\right)u^{-(1+\nu)}du \le Cz^{-(\nu+\varepsilon)}\int_0^1 u^{-(1-\varepsilon)}du \le \frac{C}{\varepsilon}z^{-(\nu+\varepsilon)}.\end{aligned}$$

Thus, for all $\varepsilon > 0$

$$|\nabla_x B_t(x,y)| \le \frac{C\rho}{t}\left\{\exp\left(\frac{r\rho}{t}\right) + \varepsilon^{-1}\left(\frac{r\rho}{t}\right)^{-(2-\nu_1+\varepsilon)}\mathbf{1}_{\nu_1<2}\right\}. \tag{B.2}$$

Take $(x,x') \in (B(0,R)\cap\tilde{D})^2$, s.t. $r<r'$. For $\alpha_0 \in (0,1]$ to be specified later on,

$$\begin{aligned}&\tfrac{|g_t(x,y)-g_t(x',y)|}{|x-x'|^{\alpha_0}} \le \tfrac{\rho}{t^2\omega}\exp\left(-\tfrac{\rho^2}{2t}\right)\left[\left|\exp\left(-\tfrac{r'^2}{2t}\right)-\exp\left(-\tfrac{r^2}{2t}\right)\right||B_t(x,y)|\right.\\ &\left.+\exp\left(-\tfrac{r'^2}{2t}\right)|B_t(x',y)-B_t(x,y)|\right]\times|x-x'|^{-\alpha_0} := (A_t^1+A_t^2)(x,x').\end{aligned}$$

Recalling that $|x-x'| \ge |r-r'|$ and $|B_t(x,y)| \le C\exp\left(\frac{r\rho}{t}\right)$ we derive

$$\begin{aligned}A_t^1(x,x') &\le \tfrac{C\rho}{t^3}\exp\left(-\tfrac{\rho^2}{2t}\right)\sup_{s\in[r,r']}\exp\left(-\tfrac{s^2}{2t}\right)|r-r'|^{1-\alpha_0}\times\exp\left(\tfrac{r\rho}{t}\right)\\ &\le \tfrac{C}{t^\xi}\exp\left(-c\tfrac{|r-\rho|^2}{t}\right).\end{aligned} \tag{B.3}$$

We also have $A_t^2(x,x') \le \frac{C\rho}{t^2}\exp(-\frac{r'^2+\rho^2}{2t})\sup_{u\in[0,1]}|\nabla_x B_t(ux+(1-u)x',y)||x-x'|^{1-\alpha_0}$. Hence, from (B.2) we get

$$\begin{aligned}A_t^2(x,x') \le \tfrac{C\rho^2}{t^3}\exp\left(-\tfrac{r'^2+\rho^2}{2t}\right)&\left\{\exp\left(\tfrac{r'\rho}{t}\right)+\varepsilon^{-1}\left(\tfrac{r\rho}{t}\right)^{-(2-\nu_1+\varepsilon)}\mathbf{1}_{\nu_1<2}\right\}\\ &\times|x-x'|^{1-\alpha_0}.\end{aligned}$$

If $\nu_1 \ge 2$, the above control together with (B.3) give the statement of the proposition with $\alpha_0 = 1$; i.e., the gradient is Lipschitz continuous. So from now on, we consider the case $\nu_1 < 2$.

If $|x-x'| \le r$, we derive from the previous expression that for $\nu_1 > \varepsilon$

$$\begin{aligned}A_t^2(x,x') &\le \tfrac{C}{(1\wedge\varepsilon)t^\xi}\exp\left(-c\tfrac{|r'-\rho|^2}{t}\right)(1+r^{1-\alpha_0-(2-\nu_1+\varepsilon)})\\ &\le \tfrac{C}{(1\wedge\varepsilon)t^\xi}\exp\left(-c\tfrac{|r'-\rho|^2}{t}\right)(1+r^{\nu_1-1-\varepsilon-\alpha_0}).\end{aligned} \tag{B.4}$$

On the other hand, if $|x-x'|>r$, we write

$$
\begin{aligned}
A_t^2(x,x') &\leq \tfrac{C\rho}{t^2}\exp\left(-\tfrac{r'^2+\rho^2}{2t}\right)|B_t(x,y)-B_t(x',y)|\times|x-x'|^{-\alpha_0}\\
&\leq \tfrac{C\rho}{t^2}\exp\left(-\tfrac{r'^2+\rho^2}{2t}\right)\Big(\textstyle\sum_{n=1}^{\infty}(\nu_n-1)I_{\nu_n-1}\left(\tfrac{r\rho}{t}\right)\\
&\quad+\textstyle\sum_{n=1}^{\infty}\left|I_{\nu_n-1}\left(\tfrac{r\rho}{t}\right)-I_{\nu_n-1}\left(\tfrac{r'\rho}{t}\right)\right|\Big)\times|x-x'|^{-\alpha_0}\\
&:= \tfrac{C\rho}{t^2}\exp\left(-\tfrac{r'^2+\rho^2}{2t}\right)(D_t^1+D_t^2).
\end{aligned}
\tag{B.5}
$$

From the recurrence relations on modified Bessel functions, see equation 9.6.26 in [AS72], and since $|x-x'|>r$ one gets

$$
\begin{aligned}
D_t^1 &\leq \tfrac{Cr\rho}{t}\textstyle\sum_{n\geq 1}(I_{\nu_n-2}+I_{\nu_n})\left(\tfrac{r\rho}{t}\right)r^{-\alpha_0}\leq \tfrac{Cr^{1-\alpha_0}\rho}{(1\wedge\varepsilon)t}\exp\left(\tfrac{\rho r}{t}\right)\left\{\left(\tfrac{r\rho}{t}\right)^{-(2-\nu_1+\varepsilon)}+1\right\}\\
&\leq \tfrac{C}{1\wedge\varepsilon}\exp\left(\tfrac{r\rho}{t}\right)\left\{\tfrac{\rho}{t}+\left(\tfrac{\rho}{t}\right)^{\nu_1-1-\varepsilon}r^{\nu_1-1-\varepsilon-\alpha_0}\right\}.
\end{aligned}
\tag{B.6}
$$

Recall now formula 9.6.18 from [AS72]; i.e.,

$$
\forall\nu>-1/2,\ I_\nu(z)=\frac{(\frac{1}{2}z)^\nu}{\pi^{1/2}\Gamma(\nu+\frac{1}{2})}\int_{-1}^{1}(1-u^2)^{\nu-1/2}\cosh(zu)du.
$$

Hence,

$$
\begin{aligned}
D_t^2 &\leq C\left\{\left|I_{\nu_1-1}\left(\tfrac{r\rho}{t}\right)-I_{\nu_1-1}\left(\tfrac{r'\rho}{t}\right)\right|+\tfrac{\rho|r-r'|}{t}\exp\left(\tfrac{r'\rho}{t}\right)\right\}|x-x'|^{-\alpha_0}\\
&\leq C\left\{\left(\tfrac{\rho}{t}\right)^{\nu_1-1}|r^{\nu_1-1}-r'^{\nu_1-1}|+\left(\tfrac{\rho}{t}\right)^{\nu_1}|r-r'|+\tfrac{\rho|r-r'|}{t}\right\}\exp\left(\tfrac{r'\rho}{t}\right)|r-r'|^{-\alpha_0}\\
&\leq C\left\{\left(\tfrac{\rho}{t}\right)^{\nu_1-1}|r-r'|^{\nu_1-1-\alpha_0}+\left(\tfrac{\rho}{t}\right)^{\nu_1}+\tfrac{\rho}{t}\right\}\exp\left(\tfrac{r'\rho}{t}\right).
\end{aligned}
\tag{B.7}
$$

Plugging (B.6) and (B.7) into (B.5) we derive that for $|x-x'|>r$

$$
A_t^2(x,x')\leq\frac{C}{(1\wedge\varepsilon)t^\xi}\exp\left(-c\frac{|r'-\rho|^2}{t}\right)\left\{1+|r-r'|^{\nu_1-1-\alpha_0}+r^{\nu_1-1-\varepsilon-\alpha_0}\right\}.
\tag{B.8}
$$

Take now $\varepsilon>0$ s.t. $\nu_1-1-\varepsilon>0$. Set $\alpha_0=\nu_1-1-\varepsilon$. From (B.4) and (B.8) the proof is complete.

Remark B.1. The spectral theory suggests our previous Hölder constant for the gradient is somehow optimal. Indeed, if ϕ_1 denotes the first eigenfunction of the elliptic Dirichlet problem for the Laplacian in a bidimensional truncated cone of angle ω and vertex 0, we have from Example 4.6.5 in Davies [Dav89] that $\phi_1(x)=O(r^{\nu_1})$, $\nu_1=\pi/\omega$ when $x\longrightarrow 0$ nontangentially, and the heat kernel also writes $q_t(x,y)=\sum_{i=1}^{\infty}\exp(-E_it)\phi_i(x)\phi_i(y)$, where the $(E_i)_{i\in\mathbf{N}^*}$ are the eigenvalues of the Laplacian in the truncated cone ($0<E_1\leq E_2\leq\cdots\uparrow\infty$) and the $(\phi_i)_{i\in\mathbf{N}^*}$ the orthonormal eigenfunctions.

The spectral decomposition of the heat kernel also suggests that we cannot expect more spatial smoothness than that of the elliptic problem. A general study of this kind

of problem is far from easy. A Sobolev approach can be found in Dauge [Dau88] and Kozlov, Maz'ya, and Rossmann [KMR97]. The possible application of their arguments to the parabolic case will concern further research.

Remark B.2. The control of the time derivative stated in Lemma 5.2 holds true up to $d = 4$ without major changes in the proof. The main tool needed is Weyl's asymptotic lemma that gives some controls on the behavior of the eigenfunctions that appear in Proposition 5.1; see [Cha84, p. 172]. The remaining computations are rather similar to the previous ones.

REFERENCES

[ABR96] L. Andersen and R. Brotherton-Ratcliffe, *Exact exotics*, Risk, 9 (1996), pp. 85–89.

[AGP95] S. Asmussen, P. Glynn, and J. Pitman, *Discretization error in simulation of one-dimensional reflecting Brownian motion*, Ann. Appl. Probab., 5 (1995), pp. 875–896.

[AS72] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[BGK99] M. Broadie, P. Glasserman, and S. Kou, *Connecting discrete and continuous path-dependent options*, Finance Stoch., 3 (1999), pp. 55–82.

[BS97] R. Bañuelos and R. G. Smits, *Brownian motion in cones*, Probab. Theory Related Fields, 108 (1997), pp. 299–319.

[BT96a] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations:* I. *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[BT96b] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations,* II. *Convergence rate of the density*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

[Cha84] I. Chavel, *Eigenvalues in Riemannian Geometry*, Wiley, New York, 1984.

[CJ59] H. Carslaw and J. Jaeger, *Conduction of Heat in Solids*, Oxford University Press, Oxford, UK, 1959.

[CKG03] C. Costantini, N. El Karoui, and E. Gobet, *Représentation de Feynman-Kac dans des domaines temps-espace et sensibilité par rapport au domaine*, C.R. Math. Acad. Sci. Paris, 337 (2003), pp. 337–342.

[Dau88] M. Dauge, *Elliptic Boundary Value Problems in Corner Domains. Smoothness and Asymptotics of Solution*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[Dav89] E. B. Davies, *Heat Kernels and Spectral Theory*, Cambridge University Press, Cambridge, UK, 1989.

[Die71] J. Dieudonné, *Eléments d'Analyse, Vol.* 1, Gauthiers-Villars, Paris, 1971.

[Fre85] M. Freidlin, *Functional Integration and Partial Differential Equations*, Ann. Math. Stud., Princeton University Press, Princeton, NJ, 1985.

[Fri64] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[GM04] E. Gobet and S. Menozzi, *Exact approximation rate of killed hypoelliptic diffusions using the discrete Euler scheme*, Stochastic Process Appl., 112 (2004), pp. 210–223.

[GM05] E. Gobet and S. Menozzi, *Discrete sampling of functionals of Itô processes*, Tech. report $n°559$ CMAP, in Séminaire de Probabilités Volume XL, Lecture Notes in Math., Springer-Verlag, New York, to appear.

[Gob00] E. Gobet, *Euler schemes for the weak approximation of killed diffusion*, Stochastic Process Appl., 87 (2000), pp. 167–197.

[Gob01] E. Gobet, *Euler schemes and half-space approximation for the simulation of diffusions in a domain*, ESAIM Probab. Statist., 5 (2001), pp. 261–297.

[Iye85] S. Iyengar, *Hitting lines with two-dimensional Brownian motion*, SIAM J. Appl. Math., 45 (1985), pp. 983–989.

[Jon68] A. L. Jones, *An extension of an inequality involving modified Bessel functions*, J. Math. Phys., 48 (1968), pp. 220–221.

[KMR97] V. A. Kozlov, V. G. Maz'ya, and J. Rossmann, *Elliptic Boundary Values Problem in Domains with Singularities*, Math. Surveys Monogr. 52, AMS, Providence, RI, 1997.

[KS98] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Applications of Mathematics (New York) 39, Springer-Verlag, New York, 1998.

[LSU68] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[Men04] S. Menozzi, *Discrétisations associées à un processus dans un domaine et schémas numériques probabilistes pour les EDP paraboliques quasi-linaires*, Ph.D. Thesis, University Paris VI, Paris, France, 2004.

[RY99] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Grundlehren Math. Wiss. 293, Springer-Verlag, New York, 1999.

[Sie79] D. Siegmund, *Corrected diffusion approximations in certain random walk problems*, Adv. in Appl. Probab., 11 (1979), pp. 701–719.

[TT90] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Anal. Appl., 8 (1990), pp. 94–120.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2633–2649

A UNIFIED APPROACH FOR UZAWA ALGORITHMS*

CONSTANTIN BACUTA†

Abstract. We present a unified approach in analyzing Uzawa iterative algorithms for saddle point problems. We study the classical Uzawa method, the augmented Lagrangian method, and two versions of inexact Uzawa algorithms. The target application is the Stokes system, but other saddle point systems, e.g., arising from mortar methods or Lagrange multipliers methods, can benefit from our study. We prove convergence of Uzawa algorithms and find optimal rates of convergence in an abstract setting on finite- or infinite-dimensional Hilbert spaces. The results can be used to design multilevel or adaptive algorithms for solving saddle point problems. The discrete spaces do not have to satisfy the LBB stability condition.

Key words. Uzawa algorithms, saddle point system, multilevel methods, augmented Lagrangian method, Stokes problem

AMS subject classifications. 74S05, 74B05, 65N22, 65N55

DOI. 10.1137/050630714

1. Introduction.

In this paper, we provide a unified approach for Uzawa methods for linear saddle point systems. Such systems arise in solving various partial differential equations (PDEs) or systems of PDEs at the continuous level or at the discrete level. Typical examples of such PDEs are second-order elliptic problems, Stokes equations, and elasticity problems. We analyze the classical Uzawa Method (UM) [1], the augmented Lagrangian Uzawa method (ALUM) [14], the inexact Uzawa method (IUM) [7, 13], and a modified (or multilevel) inexact Uzawa method (MIUM) under a general approach on abstract Hilbert spaces. The motivation for considering abstract versions of Uzawa algorithms on infinite-dimensional Hilbert spaces is that the analysis at the continuous level of an algorithm for solving a PDE gives the right strategy for discretizing the PDE. In addition, the convergence factors of certain multilevel or adaptive algorithms for solving saddle point systems depend on the stability parameters of the continuous problem, and in many cases the discrete LBB stability condition is not required to be satisfied (see [4, 12] or section 6). Next, we formulate the general framework of the saddle point problem to be studied in this paper and indicate the way the paper is organized.

We let $\mathbf{V}$ and P be two Hilbert spaces with inner products $a(\cdot,\cdot)$ and $(\cdot,\cdot)$, with the corresponding induced norms $|\cdot|_{\mathbf{V}} = |\cdot| = a(\cdot,\cdot)^{1/2}$ and $\|\cdot\|_P = \|\cdot\| = (\cdot,\cdot)^{1/2}$. The dual parings on $\mathbf{V}^* \times \mathbf{V}$ and $P^* \times P$ are denoted by $\langle\cdot,\cdot\rangle$ and $(\cdot,\cdot)$, respectively. Here, $\mathbf{V}^*$ and P^* denote the dual of $\mathbf{V}$ and P, respectively. We identify P^* and P as Hilbert spaces so that $(\cdot,\cdot)$ represents both the inner product on P and the duality between P^* and P. In applications to Stokes systems, $\mathbf{V} = (H_0^1)^d (d = 2, 3, \dots)$, P is a subspace of L^2 of codimension one and $(\cdot,\cdot)$ is the standard inner product on L^2. Next, we consider that $b(\cdot,\cdot)$ is a continuous bilinear form on $\mathbf{V} \times P$, satisfying the

*Received by the editors May 5, 2005; accepted for publication (in revised form) July 10, 2006; published electronically December 11, 2006. This work was partially supported by the University of Delaware Research Foundation.

http://www.siam.org/journals/sinum/44-6/63071.html

†Department of Mathematical Sciences, University of Delaware, 501 Ewing Hall, Newark, DE 19716 (bacuta@math.udel.edu).

inf-sup condition. More precisely, we assume that

$$\inf_{p\in P}\sup_{\mathbf{v}\in\mathbf{V}}\frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|}=m>0 \tag{1.1}$$

and

$$\sup_{p\in P}\sup_{\mathbf{v}\in\mathbf{V}}\frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|}=M<\infty. \tag{1.2}$$

For $f\in\mathbf{V}^*$, $g\in P^*$, we consider the following variational problem:
Find $(\mathbf{u},p)\in\mathbf{V}\times P$ such that

$$\begin{array}{lcll} a(\mathbf{u},\mathbf{v}) & + & b(\mathbf{v},p) & =\langle\mathbf{f},\mathbf{v}\rangle \quad \text{for all } \mathbf{v}\in\mathbf{V},\\ b(\mathbf{u},q) & & & =(g,q) \quad \text{for all } q\in P. \end{array} \tag{1.3}$$

It is known that the above variational problem has a unique solution for any $f\in\mathbf{V}^*$, $g\in P^*$ (see [9, 10, 15] or Lemma 2.1). With the forms a and b, we associate two linear operators $A:V\to V^*$ and $B:V\to P$ defined by

$$\langle A\mathbf{u},\mathbf{v}\rangle=a(\mathbf{u},\mathbf{v}) \quad \text{for all } \mathbf{u},\mathbf{v}\in\mathbf{V}$$

and

$$(B\mathbf{u},q)=b(\mathbf{u},q) \quad \text{for all } \mathbf{u}\in\mathbf{V},\ q\in P.$$

Let $B^*:P\to V^*$ be the dual operator of B defined by

$$\langle B^*q,\mathbf{v}\rangle=(q,B\mathbf{v})=(B\mathbf{v},q)=b(\mathbf{v},q) \quad \text{for all } \mathbf{v}\in\mathbf{V},\ q\in P.$$

The problem (1.3) is equivalent to the following problem:
Find $(\mathbf{u},p)\in\mathbf{V}\times P$ such that

$$\begin{array}{lcll} A\mathbf{u} & + & B^*p & =\mathbf{f},\\ B\mathbf{u} & & & =g. \end{array} \tag{1.4}$$

In this framework, we analyze Uzawa algorithms for solving the system (1.3) or (1.4). We consider that the form a gives the inner product and the norm on $\mathbf{V}$. A more general case of (1.3) is considered in [9, 10, 15]. Our particular assumptions for the form a give rise to a simplified analysis. For the general case, we obtain sharp convergence estimates only in terms of the two constants m and M.

The rest of the paper is organized as follows. In section 2, we analyze the convergence of the classical Uzawa algorithm. The augmented Lagrangian Uzawa method is analyzed in section 3 (Fortin and Glowinski [14]). In section 4, we shall investigate the convergence of the inexact Uzawa algorithm (Bramble, Pasciak, and Vassilev [7] and Elman and Golub [13]) in the above abstract framework. Applications to discretizations on stable pairs are presented in section 5. A modified inexact Uzawa algorithm with applications in constructing multilevel methods and adaptive methods for solving (1.3) is illustrated in section 6. In section 7, we present applications of our abstract results to the Stokes system.

2. The abstract Uzawa algorithm. We begin this section with two lemmas which provide basic properties of norms and operators introduced in section 1. The proofs are based on the Riesz representation theorem (see, e.g., [20]). For completeness, we include the proofs.

LEMMA 2.1. *The operator $A : V \to V^*$ is invertible and the Schur complement operator $BA^{-1}B^* : P \to P$ is symmetric and a positive definite operator satisfying*

$$(BA^{-1}B^*p, p) = \sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v}, p)^2}{|\mathbf{v}|^2}, \tag{2.1}$$

$$m^2\|p\|^2 \le (BA^{-1}B^*p, p) \le M^2\|p\|^2, \quad p \in P. \tag{2.2}$$

Consequently, the problem (1.3) *(or* (1.4)*) has a unique solution.*

Proof. From the definition of A, we get that A is a bounded injective operator. Using the Riesz representation theorem, it follows that A is also a surjective operator. Let us further note that A satisfies

$$\langle A\mathbf{u}, \mathbf{v}\rangle = a(\mathbf{u}, \mathbf{v}) = a(\mathbf{v}, \mathbf{u}) = \langle A\mathbf{v}, \mathbf{u}\rangle,$$

and the changes of variable $A\mathbf{u} = \mathbf{u}^*$ and $A\mathbf{v} = \mathbf{v}^*$ lead to

$$\langle \mathbf{u}^*, A^{-1}\mathbf{v}^*\rangle = \langle \mathbf{v}^*, A^{-1}\mathbf{u}^*\rangle, \quad \mathbf{u}^*, \mathbf{v}^* \in \mathbf{V}^*. \tag{2.3}$$

Using (2.3), we obtain

$$\begin{aligned}(BA^{-1}B^*p, q) &= \langle B^*q, A^{-1}B^*p\rangle = \langle B^*p, A^{-1}B^*q\rangle \\ &= (BA^{-1}B^*q, p) = (p, BA^{-1}B^*q), \quad p, q \in P.\end{aligned}$$

To prove (2.1), we let $p \in P$ be fixed and consider the following problem:

Find $\mathbf{u} \in \mathbf{V}$ such that

$$a(\mathbf{u}, \mathbf{v}) = b(\mathbf{v}, p) \qquad \text{for all } \mathbf{v} \in \mathbf{V}. \tag{2.4}$$

Since the functional $\mathbf{v} \to b(\mathbf{v}, p)$ is continuous on $\mathbf{V}$, by the Riesz representation theorem we have that the unique solution $\mathbf{u}$ of (2.4) satisfies

$$a(\mathbf{u}, \mathbf{u}) = \|\mathbf{v} \to b(\mathbf{v}, p)\|^2_{\mathbf{V}^*} = \sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v}, p)^2}{|\mathbf{v}|^2}. \tag{2.5}$$

On the other hand, from (2.4) we have

$$A\mathbf{u} = B^*p \ \text{ or } \ \mathbf{u} = A^{-1}B^*p$$

and

$$a(\mathbf{u}, \mathbf{u}) = \langle A\mathbf{u}, \mathbf{u}\rangle = \langle B^*p, A^{-1}B^*p\rangle = (p, BA^{-1}B^*p). \tag{2.6}$$

Thus, (2.1) follows from (2.5) and (2.6). The estimate (2.2) follows imediately from (2.1), (1.1), and (1.2).

To prove the existence and uniqueness of (1.3) (or (1.4)), we substitute $\mathbf{u}$ from the first equation of (1.4) into the second equation of (1.4). The resulting equation in p,

$$BA^{-1}B^*p = BA^{-1}\mathbf{f} - g,$$

has a unique solution due to the fact that $BA^{-1}B^* : P \to P$ is symmetric and a positive definite operator. ∎

Remark 2.2. From the general theory of symmetric operators and Lemma 2.1, we have that $\sigma(BA^{-1}B^*) \subset [m^2, M^2]$ and $m^2, M^2 \in \sigma(BA^{-1}B^*)$. In the finite-dimensional case, m^2 and M^2 are the extreme eigenvalues of the Schur complement $BA^{-1}B^*$.

LEMMA 2.3. *The following norm estimates are valid:*

$$\|\phi\|^2_{\mathbf{V}^*} = a(A^{-1}\phi, A^{-1}\phi) = |A^{-1}\phi|^2, \quad \phi \in \mathbf{V}^*, \tag{2.7}$$

$$\|A\mathbf{u}\|_{\mathbf{V}^*} = |\mathbf{u}|, \quad \mathbf{u} \in \mathbf{V}, \tag{2.8}$$

$$\|B^*q\|_{\mathbf{V}^*} = |A^{-1}B^*q| = (BA^{-1}B^*q, q)^{1/2} \le M\|q\|, \quad q \in P, \tag{2.9}$$

$$\|B\| = M, \quad \textit{hence} \quad \|B\mathbf{u}\| \le M\ |\mathbf{u}|, \quad \mathbf{u} \in \mathbf{V}. \tag{2.10}$$

Proof. By the Riesz representation theorem, we have that for any $\phi \in \mathbf{V}^*$ the problem

Find $\mathbf{u} \in \mathbf{V}$ such that

$$\langle A\mathbf{u}, \mathbf{v}\rangle = a(\mathbf{u}, \mathbf{v}) = \langle \phi, \mathbf{v}\rangle \qquad \text{for all } \ \mathbf{v} \in \mathbf{V} \tag{2.11}$$

has a unique solution, and the solution $\mathbf{u}$ satisfies

$$a(\mathbf{u}, \mathbf{u}) = \sup_{\mathbf{v}\in\mathbf{V}} \frac{\langle \phi, \mathbf{v}\rangle^2}{a(\mathbf{v}, \mathbf{v})} = \|\phi\|^2_{\mathbf{V}^*}. \tag{2.12}$$

From (2.11), we have that $\mathbf{u} = A^{-1}\phi$, which combined with (2.12) gives (2.7). The equality (2.8) is a consequence of (2.7), and (2.9) follows from (2.8) and (2.2). The last estimate follows from the definition of B and the assumption in (1.2). ∎

Next, we present the Uzawa algorithm [1] for solving the solution of the abstract problem (1.3). Given a parameter $\alpha > 0$, called a relaxation parameter, the Uzawa algorithm for approximating the solution $(\mathbf{u}, p)$ of (1.3) can be described as follows.

ALGORITHM 2.4 (Uzawa method (UM)). *Let p_0 be any approximation for p, and for $k = 1, 2, \ldots,$ construct $(\mathbf{u}_k, p_k)$ by*

$$\begin{aligned} a(\mathbf{u}_k, \mathbf{v}) &= (\mathbf{f}, \mathbf{v}) - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}, \\ p_k &= p_{k-1} + \alpha(B\mathbf{u}_k - g). \end{aligned} \tag{2.13}$$

The convergence of the UM is discussed for particular cases in, e.g., [10, 14, 15, 17]. It shows that the UM is convergent for small enough α and that the convergence rate is the same as the convergence rate of the Richardson iterative methods for the Schur complement $BA^{-1}B^*$. For completeness, we include the proof.

THEOREM 2.5. *Let $(\mathbf{u}, p)$ be the solution of (1.3) and let $(\mathbf{u}_k, p_k)$ be the sequence of approximations built by the UM (2.13). Then, the following holds.*

(i) *The sequences $\mathbf{u} - \mathbf{u}_k$ and $p - p_k$ satisfy*

$$\begin{aligned} a(\mathbf{u} - \mathbf{u}_k, \mathbf{u} - \mathbf{u}_k)^{1/2} &\le M\ \|p - p_{k-1}\|, \\ \|p - p_k\| &\le \|I - \alpha BA^{-1}B^*\|\ \|p - p_{k-1}\|. \end{aligned}$$

(ii) *For $\alpha < \frac{2}{M^2}$, the UM is convergent and*

$$\|I - \alpha BA^{-1}B^*\| = \max\{|1 - \alpha m^2|, |1 - \alpha M^2|\} < 1.$$

(iii) *For* $\alpha = \frac{1}{M^2}$, *the convergence factor is* $\|I - \alpha BA^{-1}B^*\| = 1 - \frac{m^2}{M^2}$.
(iv) *The optimal convergence factor is achieved for*

$$\alpha_{opt} = \frac{2}{M^2 + m^2} \quad \textit{and} \quad \|I - \alpha_{opt} BA^{-1}B^*\| = \frac{M^2 - m^2}{M^2 + m^2}.$$

Proof. From the first equation of (1.3) and the first equation of (2.13), we have that

$$a(\mathbf{u} - \mathbf{u}_k, \mathbf{v}) = b(\mathbf{v}, p_{k-1} - p) \qquad \text{for all } \mathbf{v} \in \mathbf{V}. \tag{2.14}$$

The above relation implies

$$\begin{aligned} |\mathbf{u} - \mathbf{u}_k|^2 &= a(\mathbf{u} - \mathbf{u}_k, \mathbf{u} - \mathbf{u}_k) = (BA^{-1}B^*(p_{k-1} - p), (p_{k-1} - p)) \\ &\leq M^2 \|p_{k-1} - p\|^2, \end{aligned}$$

which proves the first part of (i). From the second equation of (1.4) and the second equation of (2.13), we have that

$$p - p_k = p - p_{k-1} + \alpha B(\mathbf{u} - \mathbf{u}_k).$$

Combining with (2.14), we get

$$p - p_k = (I - \alpha BA^{-1}B^*)(p - p_{k-1}), \tag{2.15}$$

which gives the second part of (i). From Lemma 2.1, we have that $(I - \alpha BA^{-1}B^*)$ is a symmetric operator, and for any $p \in$P, $p \neq 0$,

$$1 - \alpha M^2 \leq \frac{((I - \alpha BA^{-1}B^*)p, p)}{\|p\|^2} \leq 1 - \alpha m^2,$$

which justifies part (ii). The rest of the proof follows from (ii). ∎

3. Augmented Lagrangian Uzawa algorithm. The main idea of the augmented Lagrangian method, introduced by Fortin and Glowinski [14], is to use the constraint condition for the variable p and another tuning parameter $\rho > 0$ in order to improve the convergence factor of the Uzawa algorithm. We will consider the approach for abstract Hilbert spaces $\mathbf{V}$ and P and prove sharp convergence estimates for the corresponding Uzawa algorithm.

Let $(\mathbf{u}, p)$ be the solution of the variational problem (1.3). Then, from the second equation of (1.4), we have that

$$(B\mathbf{u}, B\mathbf{v}) = (g, B\mathbf{v}), \quad \mathbf{v} \in \mathbf{V}.$$

Thus, for any $\rho > 0$, $(\mathbf{u}, p)$ is also a solution of

$$\begin{array}{ll} a(\mathbf{u}, \mathbf{v}) + \rho(B\mathbf{u}, B\mathbf{v}) + b(\mathbf{v}, p) &= \langle \mathbf{f}, \mathbf{v} \rangle + \rho(g, B\mathbf{v}), \\ b(\mathbf{u}, q) &= (g, q). \end{array} \tag{3.1}$$

Using the notation

$$a_\rho(\mathbf{u}, \mathbf{v}) := a(\mathbf{u}, \mathbf{v}) + \rho(B\mathbf{u}, B\mathbf{v}) \text{ and } \mathbf{f}_\rho := \mathbf{f} + \rho B^* g,$$

we have that

$$\begin{array}{ll} a_\rho(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= \langle \mathbf{f}_\rho, \mathbf{v} \rangle, \quad \mathbf{v} \in \mathbf{V}, \\ b(\mathbf{u}, q) &= (g, q), \quad q \in P. \end{array} \tag{3.2}$$

With the form a_ρ, we associate the linear operator $A_\rho : V \to V^*$,

$$\langle A_\rho \mathbf{u}, \mathbf{v}\rangle = a_\rho(\mathbf{u}, \mathbf{v}) \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbf{V}.$$

Thus, an equivalent form of (3.2) is

$$\begin{aligned} A_\rho \mathbf{u} + B^* p &= \mathbf{f}_\rho, \\ B\mathbf{u} &= g. \end{aligned} \tag{3.3}$$

Since $a_\rho(\cdot,\cdot)$ and $a(\cdot,\cdot)$ give rise to equivalent norms on $\mathbf{V}$, we have that (3.2) (or (3.3)) has a unique solution. Consequently, problems (1.3) and (3.2) are equivalent. In what follows, the Uzawa algorithm applied to (3.2) will be called the augmented Lagrangian Uzawa method (ALUM).

Given a relaxation parameter $\alpha > 0$, the augmented Lagrangian Uzawa algorithm for approximating the solution $(\mathbf{u}, p)$ of (1.3) is as follows.

ALGORITHM 3.1 (ALUM). *Let p_0 be any approximation for p, and for $k = 1, 2, \ldots,$ construct $(\mathbf{u}_k, p_k)$ by*

$$\begin{aligned} a_\rho(\mathbf{u}_k, \mathbf{v}) &= (\mathbf{f}_\rho, \mathbf{v}) - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}, \\ p_k &= p_{k-1} + \alpha(B\mathbf{u}_k - g). \end{aligned}$$

To study the convergence of (3.1), we shall calculate first

$$M_\rho := \sup_{p\in P} \sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v}, p)}{\|p\|(a_\rho(\mathbf{v}, \mathbf{v}))^{1/2}} \tag{3.4}$$

and

$$m_\rho := \inf_{p\in P} \sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v}, p)}{\|p\|(a_\rho(\mathbf{v}, \mathbf{v}))^{1/2}}. \tag{3.5}$$

THEOREM 3.2. *For any $\rho > 0$, we have*

$$BA_\rho^{-1}B^* = \left(\rho I + (BA^{-1}B^*)^{-1}\right)^{-1}, \tag{3.6}$$

$$M_\rho^2 = \frac{1}{\rho + \frac{1}{M^2}} \quad \textit{and} \quad m_\rho^2 = \frac{1}{\rho + \frac{1}{m^2}}. \tag{3.7}$$

Proof. To prove (3.6), we need two identities. First, we note that for any invertible linear operator $C : P \to P$ such that $I + \rho C$ is also invertible, we have

$$\left(\rho I + C^{-1}\right)^{-1} = C - \rho C(I + \rho C)^{-1}C. \tag{3.8}$$

This can be proved by checking that the proposed inverse verifies the algebraic definition of the inverse. The second identity is based on the Sherman–Morrison–Woodbury formula and can be proved again just by algebraic manipulations:

$$(A + \rho B^* B)^{-1} = A^{-1} - \rho A^{-1}B^*(I + \rho BA^{-1}B^*)^{-1}BA^{-1}. \tag{3.9}$$

From (3.9), we get

$$\begin{aligned} B(A + \rho B^* B)^{-1}B^* = {} & BA^{-1}B^* \\ & - \rho BA^{-1}B^*(I + \rho BA^{-1}B^*)^{-1}BA^{-1}B^*. \end{aligned} \tag{3.10}$$

If we take $C = BA^{-1}B^*$ in (3.8) and combine it with (3.10), we obtain (3.6). To verify (3.7), we notice that by applying Lemma 2.1 with a_ρ instead of a we have

$$\sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v},p)^2}{\|p\|^2\ a_\rho(\mathbf{v},\mathbf{v})} = (BA_\rho^{-1}B^*p,p), \quad p\in P. \tag{3.11}$$

Thus, we get

$$\begin{aligned} M_\rho^2 &= \sup_{p\in P} \frac{(BA_\rho^{-1}B^*p,p)}{(p,p)} = \sup_{p\in P} \frac{\left((\rho I + (BA^{-1}B^*)^{-1}\ )^{-1}p,p\right)}{(p,p)} \\ &= \frac{1}{\inf\limits_{q\in P} \frac{((\rho I+(BA^{-1}B^*)^{-1})q,q)}{(q,q)}} = \frac{1}{\rho + \inf\limits_{q\in P} \frac{((BA^{-1}B^*)^{-1}q,q)}{(q,q)}} \\ &= \left(\rho + \frac{1}{\sup\limits_{r\in P} \frac{(BA^{-1}B^*r,r)}{(r,r)}}\right)^{-1} = \left(\rho + \frac{1}{M^2}\right)^{-1}. \end{aligned}$$

Here, we have used the changes of variable $(\rho I + (BA^{-1}B^*)^{-1}\ )^{-1/2}p = q$ and $(BA^{-1}B^*)^{-1/2}q = r$. The proof for m_ρ is similar. ∎

The above result gives formulas for the inf-sup and sup-sup constants for the ALUM in terms of m, M, and ρ. In applications, the constant m is more difficult to obtain. The following theorem gives the convergence rate of the ALUM.

THEOREM 3.3. *Let* $(\mathbf{u},p)$ *be the solution of* (1.3) *and let* $(\mathbf{u}_k,p_k)$ *be the sequence of approximations built by Algorithm* 3.1. *Then, the following holds true:*

(i) *The sequences* $\mathbf{u}-\mathbf{u}_k$ *and* $p-p_k$ *satisfy*

$$\begin{aligned} &a_\rho(\mathbf{u}-\mathbf{u}_k,\mathbf{u}-\mathbf{u}_k)^{1/2} \le M_\rho\ \|p-p_{k-1}\|, \\ &\|p-p_k\| \le \|I-\alpha BA_\rho^{-1}B^*\|\ \|p-p_{k-1}\|. \end{aligned}$$

(ii) *For* $\alpha < \frac{2}{M_\rho^2}$, *the ALUM is convergent and*

$$\|I-\alpha BA_\rho^{-1}B^*\| = \max\{|1-\alpha m_\rho^2|, |1-\alpha M_\rho^2|\} < 1.$$

(iii) *For* $\alpha = \frac{1}{M_\rho^2}$, *the convergence factor is*

$$\|I-\alpha BA_\rho^{-1}B^*\| = \left(1-\frac{m^2}{M^2}\right)\frac{1}{m^2\rho+1}.$$

(iv) *The optimal convergence factor is achieved for* $\alpha_{opt} = \frac{2}{M_\rho^2+m_\rho^2}$ *and*

$$\|I-\alpha_{opt}BA_\rho^{-1}B^*\| = \frac{M_\rho^2-m_\rho^2}{M_\rho^2+m_\rho^2} = \left(1-\frac{m^2}{M^2}\right)\frac{1}{2m^2\rho+1+m^2/M^2}.$$

Proof. The result is a direct consequence of Theorem 2.5 and (3.7). ∎

A similar result for the discrete version of the Stokes system can be found in [19]. As it was pointed out in [14] and [19], the choice of a very large ρ improves on the rate of convergence of the ALUM, but at the same time, the operator A_ρ becomes more difficult to invert. For the continuous and discrete Stokes system, estimates for the convergence factor of the ALUM were recently obtained by Nochetto and Pyo in [16]. The question raised in [16] on how much we can improve the rate of convergence of the ALUM if information about the spectral value m is available can be easily answered now by comparing part (iii) and part (iv) of Theorem 3.3 or Theorem 7.1.

4. Inexact Uzawa method. Throughout the rest of the paper we will keep the notation and assumptions of section 2. In this section, following the ideas in [7, 13], we shall introduce and investigate the convergence of an abstract inexact Uzawa algorithm where the exact solve of the elliptic problem (the action of A^{-1}) is replaced by an approximation process, which might not be a linear operator. We describe the approximate inverse of A as a map $C : \mathbf{V}^* \to \mathbf{V}$ which, for $\phi \in \mathbf{V}^*$, returns an approximation of $\xi = A^{-1}\phi$ such that

$$|C\phi - A^{-1}\phi|_{\mathbf{V}} \leq \delta \|\phi\|_{\mathbf{V}^*} \qquad \text{for all } \phi \in \mathbf{V}^* \tag{4.1}$$

for some $\delta \in (0,1)$. We notice here that (4.1) is a strong condition for the infinite-dimensional case. The condition can be weakened by requiring to be satisfied only for certain values $\phi \in \mathbf{V}^*$. If $\mathbf{V}$ and P are finite-dimensional spaces, then C can be considered as a linear or nonlinear process for inverting A and (4.1) is a reasonable assumption (see [7]). One example of nonlinear process C is the approximate inverse associated with the preconditioned conjugate gradient algorithm. A practical case would be to consider $C\phi = \xi_{num}$, where ξ_{num} is the numerical approximation of ξ defined by

$$a(\xi, \mathbf{v}) = \langle \phi, \mathbf{v} \rangle \qquad \text{for all } v \in \mathbf{V}.$$

In any case, if $A\xi = \phi$ and $C\phi$ is defined by $C\phi = \xi_{ap}$, an approximation of ξ, then, according to (2.7), the assumption (4.1) is equivalent to

$$|\xi_{ap} - \xi|_{\mathbf{V}} \leq \delta |\xi|_{\mathbf{V}} \qquad \text{for all } \xi \in \mathbf{V}. \tag{4.2}$$

The inexact Uzawa algorithm for approximating the solution $(\mathbf{u}, p)$ of (1.3) is as follows.

ALGORITHM 4.1 (inexact Uzawa method (IUM)). *Let* $(\mathbf{u}_0, p_0)$ *be any approximation for* $(\mathbf{u}, p)$*, and for* $k = 1, 2, \ldots,$ *construct* $(\mathbf{u}_k, p_k)$ *by*

$$\begin{aligned} \mathbf{u}_k &= \mathbf{u}_{k-1} + C(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1}), \\ p_k &= p_{k-1} + \alpha(B\mathbf{u}_k - g). \end{aligned}$$

Before we study the stability and convergence rate of Algorithm 4.1 we shall introduce the following notation. For $k = 0, 1, \ldots,$ let $e_k^{\mathbf{u}} = \mathbf{u} - \mathbf{u}_k$, $e_k^p = p - p_k$, and

$$E_k = \begin{pmatrix} |e_k^{\mathbf{u}}| \\ \|e_k^p\| \end{pmatrix}.$$

Let

$$\mathbf{M} := \begin{pmatrix} \delta & M(1+\delta) \\ \alpha M \delta & \gamma + \alpha M^2 \delta \end{pmatrix},$$

where $\gamma := \|I - \alpha B A^{-1} B^*\| = \max\{|1 - \alpha m^2|, |1 - \alpha M^2|\}$. On $\mathbf{R}^2$ we introduce the inner product $[\cdot, \cdot]_w$ defined by

$$\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right]_w = w_1 x_1 y_1 + w_2 x_2 y_2,$$

where w_1, w_2 are any two positive numbers such that

$$\frac{w_1}{w_2} = \frac{\alpha\delta}{1+\delta},$$

and δ is a positive number such that (4.1) is satisfied. We note that $\mathbf{M}$ is symmetric with respect to the $[\cdot,\cdot]_w$ inner product. We will denote the norm induced by $[\cdot,\cdot]_w$ with $\|\cdot\|_w$.

THEOREM 4.2. *Let $0 < \alpha < 2/M^2$ and assume that C satisfies* (4.1) *with*

$$\delta < \frac{1-\gamma}{1-\gamma+2\alpha M^2}. \tag{4.3}$$

Then, the IUM converges. If r is the spectral radius of the matrix $\mathbf{M}$, then $0 < r < 1$ and

$$\|E_k\|_w \leq \ r^k \ \|E_0\|_w, \quad k = 1, 2, \ldots . \tag{4.4}$$

Proof. We follow the proof of a similar result in [7] for the finite-dimensional case. From the first equation of (1.4) and the first equation of Algorithm 4.1, we have

$$\begin{aligned} e_k^{\mathbf{u}} &= e_{k-1}^{\mathbf{u}} - C(Ae_{k-1}^{\mathbf{u}} + B^* e_{k-1}^p) \\ &= (A^{-1} - C)(Ae_{k-1}^{\mathbf{u}} + B^* e_{k-1}^p) - A^{-1}B^* e_{k-1}^p. \end{aligned} \tag{4.5}$$

From the second equation of (1.4) and the second equation of Algorithm 4.1, we get

$$e_k^p = e_{k-1}^p + \alpha B e_k^{\mathbf{u}}. \tag{4.6}$$

If we substitute $e_k^{\mathbf{u}}$ from (4.5) into (4.6), then

$$e_k^p = (I - \alpha B A^{-1} B^*) e_{k-1}^p + \alpha B (A^{-1} - C)(Ae_{k-1}^{\mathbf{u}} + B^* e_{k-1}^p). \tag{4.7}$$

From (4.5) and (4.7), by the triangle inequality, and from the estimates (2.9) and (2.10) and the assumption (4.1), we obtain

$$|e_k^{\mathbf{u}}| \leq \delta \ |e_{k-1}^{\mathbf{u}}| + M(1+\delta) \ \|e_{k-1}^p\|$$

and

$$\|e_k^p\| \leq \alpha M \delta \ |e_{k-1}^{\mathbf{u}}| + (\gamma + \alpha M^2 \delta) \|e_{k-1}^p\|.$$

Using the notation introduced above, we have

$$E_k \leq \mathbf{M} \ E_{k-1}, \tag{4.8}$$

where

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

means $x_1 \leq y_1$ and $x_2 \leq y_2$. From (4.8), we deduce

$$E_k \leq \mathbf{M}^k \ E_0. \tag{4.9}$$

Since $\mathbf{M}$ is symmetric with respect to $[\cdot,\cdot]_w$-inner product, we have

$$\|E_k\|_w^2 = [E_k, E_k]_w \leq [\mathbf{M}^k \ E_0, \mathbf{M}^k \ E_0]_w = [\mathbf{M}^{2k} \ E_0, E_0]_w \leq r^{2k} \ \|E_0\|_w^2,$$

which proves (4.2). To complete the proof, we have to show that $r \in (0,1)$, provided that $0 < \alpha < 2/M^2$ and (4.3) holds. The characteristic equation of the matrix $\mathbf{M}$ is

$$\lambda^2 - \lambda(\delta + \gamma + \alpha M^2 \delta) + \delta(\gamma - \alpha M^2) = 0.$$

Since $\mathbf{M}$ has positive entries, the characteristic equation has real roots and the largest (positive) root agrees with the spectral radius of $\mathbf{M}$. Consequently,

$$r = \frac{1}{2}\left(\delta + \gamma + \alpha M^2\delta + \sqrt{(\delta+\gamma+\alpha M^2\delta)^2 - 4\delta(\gamma - \alpha M^2)}\right).$$

Using that $\gamma = \max\{|1-\alpha m^2|, |1-\alpha M^2|\}$ and $\alpha \in (0, 2/M^2)$, it is easy to verify that the function $\delta \to r = r(\delta)$ is an increasing function on $(0,1)$ and that $r = 1$ for

$$\delta = \delta_0 := \frac{1-\gamma}{1-\gamma+2\alpha M^2}. \tag{4.10}$$

This completes the proof of the theorem. □

Remark 4.3. For $0 < \alpha \leq \frac{2}{M^2+m^2}$ we have that $\gamma = 1 - \alpha m^2$ and the threshold δ_0 becomes

$$\delta_0 = \frac{m^2}{m^2 + 2M^2},$$

which is independent of α. For $\frac{2}{M^2+m^2} \leq \alpha < \frac{2}{M^2}$ we have that $\gamma = \alpha M^2 - 1$ and the threshold δ_0 becomes

$$\delta_0 = \frac{2 - \alpha M^2}{2 + \alpha M^2}.$$

Nevertheless, the optimal (maximal) value of δ_0 as the function of $\alpha \in [\frac{2}{M^2+m^2}, \frac{2}{M^2})$ is $\delta_0 = \frac{m^2}{m^2+2M^2}$ and is achieved for $\alpha = \frac{2}{M^2+m^2}$. Thus, a good choice for α (independent of m) is $\alpha = 1/M^2$. In this case we still have $\delta_0 = \frac{m^2}{m^2+2M^2}$.

Remark 4.4. We can apply the IUM for the augmented Lagrangian formulation. The only changes in Algorithm 4.1 is that A is replaced by A_ρ and $\mathbf{f}$ is replaced by $\mathbf{f}_\rho$. The convergence analysis follows from Theorem 4.2. Let us further notice that in this case $|e_k^{\mathbf{u}}|^2 = a_\rho(e_k^{\mathbf{u}}, e_k^{\mathbf{u}})$ and for $\alpha = 1/M_\rho^2$ the threshold δ_0 which assures convergence for the IUM is

$$\delta_0(\rho) = \frac{m_\rho^2}{m_\rho^2 + 2M_\rho^2} = \frac{m^2 + \rho m^2 M^2}{m^2 + 2M^2 + 3\rho m^2 M^2} \to \frac{1}{3} \quad \text{as} \quad \rho \to \infty.$$

Thus, if the IUM for the augmented Lagrangian formulation is applied with sufficiently large ρ, $\alpha = 1/M_\rho^2$ and with the approximation operator C satisfying

$$\|C - A_\rho^{-1}\| \leq \delta_0(\rho) < 1/3,$$

then the method converges.

Remark 4.5. A different approach in analyzing the IUM in the finite-dimensional case is presented by Cheng in [11]. From his analysis for $\alpha = 1$ and $M = 1$, it follows that the IUM converges (with a different estimate for the convergence factor), under the weaker assumption that $\delta < \delta_0 = 1/3$. Cheng's result for the infinite-dimensional case seems not to have been investigated. A positive answer for this problem would be an interesting result, since in practice it is difficult to estimate the spectral value m.

5. Discretization with the inf-sup condition. In this section we assume that the variational form of a PDE (or system of PDEs) leads to (1.3) and let $\mathbf{V}_h$ and P_h be two finite-dimensional spaces, $\mathbf{V}_h \subset \mathbf{V}$, $P_h \subset P$, with good approximation properties. We further assume that

$$\inf_{p\in P_h} \sup_{\mathbf{v}\in\mathbf{V}_h} \frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|} = m(h) > 0 \text{ and } \sup_{p\in P_h} \sup_{\mathbf{v}\in\mathbf{V}_h} \frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|} = M(h). \tag{5.1}$$

For an overview of numerical methods for solving saddle point systems, we refer the reader to the recently published review paper [5] by Benzi, Golub, and Liesen.

From Lemma 2.1 and Remark 2.2 we see that $m(h), M(h)$ are the lowest and the largest eigenvalues of the Schur complement $B_h A_h^{-1} B_h^*$ associated with the discrete spaces $\mathbf{V}_h$ and P_h. Then (see, e.g., [9, 15]), the discrete variational problem

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times P_h$ such that

$$\begin{array}{lcll} a(\mathbf{u}_h,\mathbf{v}) & + & b(\mathbf{v},p_h) & = \langle \mathbf{f},\mathbf{v}\rangle, \quad \mathbf{v}\in\mathbf{V}_h, \\ b(\mathbf{u}_h,q) & & & = (g,q), \quad q\in P_h, \end{array} \tag{5.2}$$

has a unique solution and

$$|\mathbf{u}-\mathbf{u}_h| + \|p-p_h\| \le c\left(\inf_{\mathbf{v}\in\mathbf{V}_h} |\mathbf{u}-\mathbf{v}| + \inf_{q\in P_h} \|p-q\|\right),$$

where c is a constant depending only on $m(h)$ and $M(h)$. In this case, the exact or inexact Uzawa algorithms can be applied for the discrete variational problem (5.2) on $\mathbf{V}_h \times P_h$; see, e.g., [7]. The convergence factors depend on $m(h)$ and $M(h)$ and could deteriorate as $h \to 0$ if the pair $(\mathbf{V}_h, P_h)$ is not stable. We recall here that a pair $(\mathbf{V}_h, P_h)$, or more precisely a family of pairs $\{(\mathbf{V}_h, P_h)\}_h$, is called stable if $m(h)$ defined in (5.1) satisfies

$$m(h) \ge m > 0,$$

with m independent of h. In the next section we use the inexact Uzawa algorithm at the continuous level to construct algorithms which avoid building stable pairs $(\mathbf{V}_h, P_h)$.

6. Modified inexact Uzawa method. Eliminating the discrete inf-sup condition. We shall apply the IUM to construct discrete approximations $(\mathbf{u}_k, p_k) \in (\mathbf{V}_k, P_k)$, where $\mathbf{V}_k \subset \mathbf{V}$ and $P_k \subset P$ are finite-dimensional spaces such that the pairs $(\mathbf{V}_k, P_k)$ do not have to be stable pairs.

The algorithm proposed in this section can be used for building multilevel or adaptive methods for solving the system (1.3). Adaptive methods for saddle point problems have been the subject for recent research in numerical analysis (see, e.g., [12, 4]). Our new approach, combined with standard techniques of a posteriori error estimate theory, could lead to new and efficient adaptive algorithms for solving saddle point systems. To describe our new algorithm, we assume that a sequence of nested subspaces,

$$\mathbf{V}_0 \subset \mathbf{V}_1 \subset \mathbf{V}_2 \subset \cdots \subset \mathbf{V},$$

was determined and for $k = 1, 2, \ldots$, a linear or nonlinear process $C_k : \mathbf{V}^* \to \mathbf{V}_k$ approximating A^{-1} is available such that for a fixed $\phi \in \mathbf{V}^*$, $C_k\phi \in \mathbf{V}_k$ is an approximation of $\xi = A^{-1}\phi$. To construct a good approximate inverse $C_k : \mathbf{V}^* \to \mathbf{V}_k$

one might need to increase the space $\mathbf{V}_{k-1}$ to a space with better approximation properties using an adaptive method. Thus, the embedding assumption $\mathbf{V}_{k-1} \subset \mathbf{V}_k$ is needed. On the other hand, in the proposed algorithms, the variable p is updated at the continuous level and no inversion is used. Thus, the P_k's are just subsets of the space P and do not have to be nested.

The modified inexact Uzawa algorithm for approximating the solution $(\mathbf{u}, p)$ of (1.3) can be stated now as follows.

ALGORITHM 6.1 (modified inexact Uzawa method (MIUM)). *Let $\mathbf{u}_0 \in \mathbf{V}_0$ be any approximation for $\mathbf{u}$ and let $p_0 \in P$ be any approximation for p. For $k = 1, 2, \ldots$, construct $(\mathbf{u}_k, p_k)$, with $\mathbf{u}_k \in \mathbf{V}_k$, by*

$$\begin{aligned} \mathbf{u}_k &= \mathbf{u}_{k-1} + C_k(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1}), \\ p_k &= p_{k-1} + \alpha(B\mathbf{u}_k - g). \end{aligned} \tag{6.1}$$

THEOREM 6.2. *Let $0 < \alpha < 2/M^2$, $\gamma = \max\{|1 - \alpha m^2|, |1 - \alpha M^2|\} = \|I - \alpha B A^{-1} B^*\|$, and assume that for $k = 1, 2, \ldots, C_k$ satisfies*

$$\|(C_k - A^{-1})(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1})\|_{\mathbf{V}} \leq \delta \|(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1})\|_{\mathbf{V}^*}, \tag{6.2}$$

with

$$\delta < \frac{1 - \gamma}{1 - \gamma + 2\alpha M^2}. \tag{6.3}$$

Then, the MIUM converges and the convergence rate is given by (4.4).

Proof. It is similar to the proof of Theorem 4.2. ∎

We notice here that, for a fixed α, the threshold δ which assures the convergence of the MIUM depends only on the constants m and M. In the case $g = 0$, we have $p_k \in P_k := B\mathbf{V}_k$. Nevertheless, no matter the choice of the spaces $\mathbf{V}_k, P_k$, the pairs $(\mathbf{V}_k, P_k)$ do not have to be stable pairs.

For the rest of this section, the first equation in (6.1) will be considered in a variational form as follows. Let $\mathbf{d}_k \in \mathbf{V}_k$ be the solution of

$$a(\mathbf{d}_k, \mathbf{v}) = \langle f, \mathbf{v} \rangle - a(\mathbf{u}_{k-1}, v) - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}_k. \tag{6.4}$$

Take $\tilde{\mathbf{d}}_k := C_k(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1})$ to be an approximation of $\mathbf{d}_k$. For example, $\tilde{\mathbf{d}}_k$ could be a numerical approximation of $\mathbf{d}_k$. Let us assume that $\mathbf{D}_{k-1} \in \mathbf{V}$ is the solution of the continuous problem

$$a(\mathbf{D}_{k-1}, \mathbf{v}) = \langle f, \mathbf{v} \rangle - a(\mathbf{u}_{k-1}, v) - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}. \tag{6.5}$$

From the Riesz representation theorem

$$\|(\mathbf{f} - A\mathbf{u}_{k-1} - B^* p_{k-1})\|_{\mathbf{V}^*} = |\mathbf{D}_{k-1}|_{\mathbf{V}}.$$

Thus, the assumption (6.2) can be rewritten as

$$|\tilde{\mathbf{d}}_k - \mathbf{D}_{k-1}|_{\mathbf{V}} \leq \delta \, |\mathbf{D}_{k-1}|_{\mathbf{V}}. \tag{6.6}$$

Since $\mathbf{d}_k \in \mathbf{V}_k$ is the Galerkin approximation of $\mathbf{D}_{k-1} \in \mathbf{V}$, we have that $|\mathbf{d}_k|_{\mathbf{V}} \leq |\mathbf{D}_{k-1}|_{\mathbf{V}}$. A sufficient condition for the assumption (6.2) is

$$|\tilde{\mathbf{d}}_k - \mathbf{D}_{k-1}|_{\mathbf{V}} \leq \delta \, |\mathbf{d}_k|_{\mathbf{V}}. \tag{6.7}$$

6.1. Multilevel exact Uzawa. In this subsection we assume that the problem (6.4) can be solved exactly on $\mathbf{V}_k$, i.e., $\tilde{\mathbf{d}}_k = \mathbf{d}_k$. Then, $\mathbf{u}_k = \mathbf{u}_{k-1} + \mathbf{d}_k$ and consequently,

$$a(\mathbf{u}_k, \mathbf{v}) = \langle f, \mathbf{v}\rangle - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}_k.$$

If $\mathbf{U}_{k-1} \in \mathbf{V}$ satisfies

$$a(\mathbf{U}_{k-1}, \mathbf{v}) = \langle f, \mathbf{v}\rangle - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V},$$

then $\mathbf{D}_{k-1} = \mathbf{U}_{k-1} - \mathbf{u}_{k-1}$ and (6.6) is equivalent to

$$|\mathbf{u}_k - \mathbf{U}_{k-1}|_{\mathbf{V}} \le \delta\ |\mathbf{u}_{k-1} - \mathbf{U}_{k-1}|_{\mathbf{V}}. \tag{6.8}$$

If $\eta_k > 0$ is a computable estimator for $|\mathbf{u}_k - \mathbf{U}_{k-1}|_{\mathbf{V}}$, i.e.,

$$|\mathbf{u}_k - \mathbf{U}_{k-1}|_{\mathbf{V}} \le \eta_k, \tag{6.9}$$

then, using (6.7), we get that a sufficient condition for (6.8) is

$$\eta_k \le \delta\ |\mathbf{u}_k - \mathbf{u}_{k-1}|_{\mathbf{V}}. \tag{6.10}$$

ALGORITHM 6.3 (multilevel exact Uzawa). *Let $p_0 \in P$ be any approximation for p. For $k = 1, 2, \ldots$, construct $(\mathbf{u}_k, p_k)$, with $\mathbf{u}_k \in \mathbf{V}_k$, by*

$$\begin{aligned} a(\mathbf{u}_k, \mathbf{v}) &= \langle f, \mathbf{v}\rangle - b(\mathbf{v}, p_{k-1}), \quad \mathbf{v} \in \mathbf{V}_k, \\ p_k &= p_{k-1} + \alpha(B\mathbf{u}_k - g). \end{aligned}$$

As a consequence of Theorem 6.2 we have the following.

COROLLARY 6.4. *Let $0 < \alpha < 2/M^2$, $\gamma = \|I - \alpha BA^{-1}B^*\|$, and assume that (6.8) or (6.9)–(6.10) are satisfied with $\delta < \frac{1-\gamma}{1-\gamma+2\alpha M^2}$. Then, the multilevel exact Uzawa algorithm converges and the convergence rate is given by (4.4).*

6.2. Multilevel inexact Uzawa. In this subsection, we assume that the problem (6.4) can be solved on each $\mathbf{V}_k$ with an absolute error $\epsilon_k \in [0, \delta)$, i.e.,

$$|\mathbf{d}_k - \tilde{\mathbf{d}}_k|_{\mathbf{V}} \le \epsilon_k |\mathbf{d}_k|_{\mathbf{V}}. \tag{6.11}$$

If $\eta_k > 0$ is a computable estimator for $|\mathbf{d}_k - \mathbf{D}_{k-1}|_{\mathbf{V}}$, i.e.,

$$|\mathbf{d}_k - \mathbf{D}_{k-1}|_{\mathbf{V}} \le \eta_k, \tag{6.12}$$

then a computable sufficient condition for (6.6) is

$$\eta_k \le \frac{\delta - \epsilon_k}{1 + \epsilon_k}\ |\tilde{\mathbf{d}}_k|_{\mathbf{V}}. \tag{6.13}$$

Indeed, from (6.7) and (6.11)–(6.13) and the triangle inequality we have

$$\begin{aligned} |\tilde{\mathbf{d}}_k - \mathbf{D}_{k-1}|_{\mathbf{V}} &\le |\mathbf{d}_k - \mathbf{D}_{k-1}|_{\mathbf{V}} + |\mathbf{d}_k - \tilde{\mathbf{d}}_k|_{\mathbf{V}} \\ &\le \eta_k + \epsilon_k |\mathbf{d}_k|_{\mathbf{V}} \le \frac{\delta_k - \epsilon_k}{1 + \epsilon_k}\ |\tilde{\mathbf{d}}_k|_{\mathbf{V}} + \epsilon_k |\mathbf{d}_k|_{\mathbf{V}} \\ &\le \frac{\delta_k - \epsilon_k}{1 + \epsilon_k}(1 + \epsilon_k)|\mathbf{d}_k|_{\mathbf{V}} + \epsilon_k |\mathbf{d}_k|_{\mathbf{V}} = \delta_k |\mathbf{d}_k|_{\mathbf{V}} \le \delta_k\ |\mathbf{D}_{k-1}|_{\mathbf{V}}. \end{aligned}$$

We conclude this subsection with a corollary and some remarks.

COROLLARY 6.5. *Let $0 < \alpha < 2/M^2$, $\gamma = \|I - \alpha BA^{-1}B^*\|$, and let $\mathbf{d}_k, \tilde{\mathbf{d}}_k$ satisfy (6.11)–(6.13) with $\delta < \frac{1-\gamma}{1-\gamma+2\alpha M^2}$. Then, the MIUM converges and the convergence rate is given by (4.4).*

6.3. Multilevel and adaptive interpretation of the inexact Uzawa algorithm. We note that the modified inexact Uzawa algorithm can be interpreted as a multilevel algorithm. We consider that a sequence $\{\mathbf{M}_k\}$ of approximating subspaces of $\mathbf{V}$ is constructed such that $\mathbf{M}_k$ is strictly larger than $\mathbf{M}_{k-1}$ and that $\mathbf{M}_k$ is built from $\mathbf{M}_{k-1}$ by a uniform refinement strategy (see, e.g., [3, 6, 8, 19]). Based on this existing sequence of nested subspaces of $\mathbf{V}$, we can now build a sequence $\{\mathbf{V}_k\}$ so that (6.6) holds as follows.

Take $\mathbf{V}_0 = \mathbf{M}_0$, and for any positive integer k, assuming that $\mathbf{V}_{k-1} = \mathbf{M}_j$ is known, define $\mathbf{V}_{k+i} := \mathbf{M}_j$ for $i = 0, 1, \ldots$ as long as (6.13) is satisfied for k replaced by $k+i$. In other words, we update $\mathbf{u}_{k-1}$ without enlarging the space $\mathbf{V}_{k-1}$ as long as (6.13) is satisfied. When (6.13) fails to hold, we solve for the $\mathbf{u}_k$ on the next discrete level space.

The modified inexact Uzawa algorithm can be also interpreted as an adaptive method. We construct the sequence $\{\mathbf{V}_k\}$ (so that (6.6) holds) by starting with a subspace $\mathbf{V}_0$ of $\mathbf{V}$ with good approximation properties and by building the sequence $\{\mathbf{V}_k\}_{k\geq 1}$ in a similar manner. If (6.13) fails to hold for $\mathbf{V}_k = \mathbf{V}_{k-1}$, then the new discrete space $\mathbf{V}_k$ is constructed by using an adaptive strategy which assures that (6.13) and consequently (6.6) hold.

7. Applications to the Stokes system. We consider the stationary Stokes equations

$$\begin{array}{lll} -\Delta \mathbf{u} & -\nabla p & = \mathbf{f} \quad \text{in } \Omega, \\ \operatorname{div} \mathbf{u} & & = g \quad \text{in } \Omega, \end{array} \tag{7.1}$$

with vanishing Dirichlet boundary condition $\mathbf{u} = 0$ on $\partial\Omega$ and g satisfying the constraint

$$\int_\Omega g \, dx = 0.$$

In this section we apply the abstract Uzawa results presented in the previous sections to solve (7.1).

Let $\mathbf{V} := (H_0^1(\Omega))^d$, $d = 2$ or $= 3$, and

$$P = L_0^2(\Omega) := \left\{ h \in L^2(\Omega) | \int_\Omega h \, dx = 0 \right\}.$$

We assume that $\mathbf{f} \in (L^2(\Omega))^d$ and $g \in L^2(\Omega)$. The variational formulation of (7.1) becomes

Find $\mathbf{u} \in \mathbf{V}, p \in P$ such that

$$\begin{array}{lll} (\nabla \mathbf{u}, \nabla \mathbf{v}) & +(\operatorname{div} \mathbf{v}, p) & = (\mathbf{f}, \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}. \\ (\operatorname{div} \mathbf{u}, q) & & = (g, q), \quad q \in P, \end{array} \tag{7.2}$$

where $(\cdot, \cdot)$ represents the standard L^2-inner product. We will denote by $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ the bilinear forms

$$a(\mathbf{u}, \mathbf{v}) := (\nabla \mathbf{u}, \nabla \mathbf{v}) = \sum_{i=1}^{d} (\nabla \mathbf{u}_i, \nabla \mathbf{v}_i)$$

and

$$b(\mathbf{v}, p) := (\operatorname{div} \mathbf{v}, p), \quad \mathbf{v} \in \mathbf{V}, p \in P.$$

We note that, for Ω smooth enough, we have

$$a(\mathbf{u},\mathbf{v}) := (\nabla\mathbf{u},\nabla\mathbf{v}) = (\operatorname{curl}\mathbf{u},\operatorname{curl}\mathbf{v}) + (\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v}), \quad \mathbf{u},\mathbf{v}\in\mathbf{V}. \tag{7.3}$$

We denote the norm induced by a with $|\cdot|_{\mathbf{V}}$ or $|\cdot|$. The norm on P is the L^2-standard norm and is simply denoted by $\|\cdot\|$. With the above notation, the variational formulation of (7.1) becomes (1.3).

It is known that for Ω smooth enough, the following LBB condition holds. More precisely, we have

$$\inf_{p\in P}\sup_{\mathbf{v}\in\mathbf{V}}\frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|} = c_0 > 0. \tag{7.4}$$

On the other hand, from (7.3) we get that

$$\sup_{p\in P}\sup_{\mathbf{v}\in\mathbf{V}}\frac{b(\mathbf{v},p)}{\|p\|\ |\mathbf{v}|} = 1. \tag{7.5}$$

We notice that for the Stokes problem the operator $A:\mathbf{V}\to\mathbf{V}^*$ consists of d copies of $-\Delta : H_0^1(\Omega)\to H^{-1}(\Omega)$, $B\mathbf{v} = \operatorname{div}\mathbf{v}$, $B^*p = -\nabla p$, and for $\rho>0$,

$$a_\rho(\mathbf{u},\mathbf{v}) := a(\mathbf{u},\mathbf{v}) + \rho(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v}), \quad \mathbf{u},\mathbf{v}\in\mathbf{V}.$$

The next two theorems are direct consequences of Theorems 2.5 and 3.3, respectively.

THEOREM 7.1. *Let* $(\mathbf{u},p)$ *be the solution of* (7.2) *and let* $(\mathbf{u}_k,p_k)$ *be the sequence of approximations built by the UM* (2.13). *Then the statements* (i)–(iv) *of Theorem* 2.5 *hold with* $m = c_0$ *and* $M = 1$.

THEOREM 7.2. *Let* $(\mathbf{u},p)$ *be the solution of* (7.2) *and let* $(\mathbf{u}_k,p_k)$ *be the sequence of approximations built by the ALUM* (3.1). *Then the statements* (i)–(iv) *of Theorem* 3.3 *hold with* $m = c_0$ *and* $M = 1$.

According to section 5, both the UM and ALUM can be applied to any discretization of (7.2), provided that $(\mathbf{V_h},P_h)$, with $\mathbf{V_h}\subset\mathbf{V}$ and $P_h\subset P$, is a stable pair. Let us assume that a fixed pair $(\mathbf{V}_h,P_h)$ satisfies the discrete inf-sup and sup-sup conditions with constants $m(h) = c_d > 0$ and $M(h) = 1$. If $Q_h : P\to P_h$ is the L^2-orthogonal projection, then, with the new spaces, the operators associated with the forms a and b are A_h and B_h, respectively, where $A_h : \mathbf{V}_h\to\mathbf{V}_h^*$ consists of d copies of the discrete Laplacian and $B_h\mathbf{v} = Q_h\operatorname{div}\mathbf{v}$. Thus, the update for the pressure becomes

$$p_k = p_{k-1} + \alpha\ Q_h(\operatorname{div}\mathbf{u}_k - g).$$

The analysis of the discrete versions of the UM and the ALUM can be carried on similarly. The only difference in describing the convergence of the two algorithms for the discrete case is that c_0 in Theorems 7.1 and 7.2 is replaced by c_d.

The inexact Uzawa algorithm can be also applied for the discretization of (7.2) on $(\mathbf{V_h},\mathbf{P_h})$ (see, e.g., [7]). Taking for example $C_h : \mathbf{V}_h^*\to\mathbf{V}_h$ to be a preconditioner for A_h such that (4.1) is satisfied with $\delta < \frac{c_d^2}{2+c_d^2}$, we have that the IUM converges for any $\alpha\in(0,2)$. We can also apply the inexact Uzawa algorithm for the augmented Lagrangian Uzawa formulation on $(\mathbf{V_h},\mathbf{P_h})$ (see Remark 4.4).

According to Corollary 6.5, the MIUM for solving (7.2) can be also applied for any $\delta < c_0^2/(2+c_0^2)$. The main difficulty in doing so is to find the sequence of spaces $\{\mathbf{V}_k\}$ such that (6.6) or (6.13) is satisfied. Residual-type a posteriori estimators η_k (see, e.g., [2], [18]) could be involved in finding the right sequence $\{\mathbf{V}_k\}$. Constructing and testing multilevel or adaptive algorithms for solving the Stokes system based on the MIUM remains a challenging new problem and is a subject for future work.

8. Conclusion. The paper gives a unified analysis approach of various Uzawa-like algorithms for solving continuous or discrete saddle point problems. The convergence condition and the convergence factors depend upon the extreme spectral bounds of the Schur complement $BA^{-1}B^*$ only. To the best of our knowledge, the result concerning the optimal convergence factor of the ALUM for the infinite-dimensional case is new. The analysis of the modified inexact Uzawa algorithm at the continuous level, which was introduced in section 6, gives a general strategy for solving saddle point systems. Our inexact Uzawa algorithm is similar to the algorithm for solving the Stokes system presented in [4]. The differences are in the way the error bounds are imposed (see (6.6), (6.13)) and the way the pressure is updated. Our analysis, combined with standard techniques of a posteriori error estimates, could lead to new and efficient adaptive algorithms for solving saddle point systems. The main difficulty in implementing concrete algorithms based on the MIUM is finding error estimators η_k such that the conditions (6.6) or (6.13) are satisfied. Finding spaces $\{\mathbf{V}_k\}$ such that conditions similar to (6.6) are satisfied will be the focus of our future work.

Acknowledgment. The author would like to thank the two reviewers for the valuable suggestions towards improving this paper.

REFERENCES

[1] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.

[2] I. Babuška and A. Miller, *A feedback finite element method with a posteriori error estimations: Part* I. *The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg., 61 (1987), pp. 1–40.

[3] C. Bacuta, J. H. Bramble, and J. Pasciak, *New interpolation results and applications to finite element methods for elliptic boundary value problems*, Numer. Linear Algebra Appl., 10 (2003), pp. 33–64.

[4] E. Bänsch, P. Morin, and R. H. Nocheto, *An adaptive Uzawa FEM for the Stokes problem: Convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229.

[5] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solutions of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[6] J. H Bramble, J. E. Pasciak, and P. S. Vassilevski, *Computational scales of Sobolev norms with application to preconditioning*, Math. Comp., 69 (2000), pp. 463–480.

[7] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.

[8] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook for Numerical Analysis, P. G. Ciarlet and J .L. Lions, eds., Vol. VII, North–Holland, Amsterdam, 2000, pp. 173–415.

[9] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[10] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[11] X.-L. Cheng, *On the nonlinear inexact Uzawa algorithm for saddle-point problems*, SIAM J. Numer. Anal., 37 (2000), pp. 1930–1934.

[12] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems–optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[13] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[14] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solutions of Boundary Value Problems*, Stud. Math. Appl. 15, North–Holland, Amsterdam, 1983.

[15] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[16] R. H. Nochetto and J. Pyo, *Optimal relaxation parameter for the Uzawa method*, Numer.

Math., 98 (2004), pp. 695–702.
[17] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1984.
[18] R. Verfurth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, 1996.
[19] J. Xu, *Multilevel Finite Element Theory*, manuscript.
[20] K. Yosida *Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, 1978.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2650–2670

DISCONTINUOUS GALERKIN FINITE ELEMENT APPROXIMATION OF NONLINEAR NON-FICKIAN DIFFUSION IN VISCOELASTIC POLYMERS*

BÉATRICE RIVIÈRE† AND SIMON SHAW‡

Abstract. We consider discrete schemes for a nonlinear model of non-Fickian diffusion in viscoelastic polymers. The model is motivated by, but not the same as, that proposed by Cohen, White, and Witelski in *SIAM J. Appl. Math.*, 55 (1995), pp. 348–368. The spatial discretization is effected with both the symmetric and nonsymmetric interior penalty discontinuous Galerkin finite element method, and the time discretization is of Crank–Nicolson type. We also discuss two means of handling the nonlinearity: either implicitly, which requires the solution of nonlinear equations at each time level, or through a linearization based on extrapolating from previous time levels. The same optimal orders of convergence are proven in both cases and, to verify this, some numerical results are also given for the linearized scheme.

Key words. non-Fickian diffusion, viscoelasticity, finite element method, error estimates

AMS subject classifications. 35K55, 65M60, 74D99

DOI. 10.1137/05064480X

1. Introduction.

In [24] Thomas and Windle demonstrated by experiment that the diffusion of organic penetrants into glassy polymers does not obey the classical Fick's law. At moderate temperatures the profile of diffusing penetrant (methanol in their case) forms a steep front which travels at a constant speed into the polymer. In [25] they developed a model for this "anomalous" diffusion in terms of an ordinary differential equation for the fractional swelling of the polymer.

However, in order to have more predictive value, a mathematical model for this behavior in the form of a partial differential equation is more desirable. Such a model has been proposed by Cohen, White, and Witelski in [6] (see also the references therein). Recognizing that viscoelastic stress relaxation effects are significant in polymers, they add such a term to Fick's law and drive this stress through a nonlinear relaxation equation which is adjoined to the diffusion equation. Solving the system then results in a heat equation with a nonlinear viscoelastic memory term in the form of a Volterra integral—typical of continuum models of polymers (see, e.g., [8] for polymer theory and [14] for a similar model of heat conduction).

In terms of the underlying physics, it seems that high levels of penetrant concentration can cause a *rubber-glass phase change*. The polymer's viscoelastic properties change dramatically across this transition layer, and this can cause sharp fronts to develop in the diffusing penetrant.

The model proposed in [6] seems to be difficult to handle in terms of obtaining estimates and so, as a stepping stone to that model, we deal here with a simpler version which involves a vector of stresses in the diffusion equation, rather than (as

*Received by the editors November 10, 2005; accepted for publication (in revised form) June 27, 2006; published electronically December 21, 2006.

http://www.siam.org/journals/sinum/44-6/64480.html

†Computational Mathematics Research Group, Department of Mathematics, 301 Thackeray, University of Pittsburgh, Pittsburgh, PA 15260 (riviere@math.pitt.edu).

‡BICOM (Brunel Institute of Computational Mathematics), Brunel University, Uxbridge UB8 3PH, England (simon.shaw@brunel.ac.uk). This author would like to acknowledge the support of the Engineering and Physical Sciences Research Council, GR/R10844/01, and also the US Army Research Office, DAAD19-00-1-0421.

in [6]) the gradient of a scalar stress. The reason this is a simplification can be better explained once we have seen the equations.

Our model is as follows. For an open bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) and a time interval $I := (0, T)$, for some $T > 0$, we want to find the "concentration," $u\colon \Omega \times I \to \mathbb{R}$, and viscoelastic stress, $\boldsymbol{\sigma}\colon \Omega \times I \to \mathbb{R}^d$, such that in $\Omega \times I$,

$$u_t(t) - \nabla \cdot D\nabla u(t) = f(t) + \nabla \cdot K\boldsymbol{\sigma}(t), \tag{1}$$

$$\boldsymbol{\sigma}_t(t) + \gamma(u)\boldsymbol{\sigma}(t) = \mu\nabla u(t), \tag{2}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)^T$. These are subject to the initial conditions,

$$u(\boldsymbol{x}, 0) = \breve{u}(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{\sigma}(\boldsymbol{x}, 0) = \breve{\boldsymbol{\sigma}}(\boldsymbol{x}), \tag{3}$$

and the boundary conditions,

$$\begin{aligned} &u(\boldsymbol{x}, t) = 0 \text{ on } \Gamma_D \times I \qquad \text{and} \\ &(D\nabla u(\boldsymbol{x}, t) + K\boldsymbol{\sigma}(\boldsymbol{x}, t)) \cdot \boldsymbol{n}(\boldsymbol{x}) = g(\boldsymbol{x}, t) \text{ on } \Gamma_N \times I, \end{aligned} \tag{4}$$

where $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\Gamma_D \cap \Gamma_N = \varnothing$, Γ_N has outward normal $\boldsymbol{n}$, and Γ_D is closed with positive surface measure. Note that in (1) and (2), and usually below, we drop the $\boldsymbol{x}$ dependence.

In [6] the vector of stresses, $\boldsymbol{\sigma}$, is replaced by the gradient of a scalar stress, $\nabla\sigma$. Our model is a simplification because, in weak form, we can generate the term $(\boldsymbol{\sigma}, \nabla u)$ in both equations and therefore easily merge them as a starting point for estimates. This is clearly illustrated below in Theorem 1.1.

In our equations D, K, and μ are positive constants. Also, the nonlinear function

$$\gamma(u) = \frac{1}{2}(\gamma_R + \gamma_G) + \frac{1}{2}(\gamma_R - \gamma_G)\tanh\left(\frac{u - u_{RG}}{\Delta}\right), \tag{5}$$

with constants $\gamma_R \gg \gamma_G > 0$, models the sharp change in material properties across the *rubber-glass transition*. The sharpness of the change is controlled by the positive constant Δ, and the location of the change is controlled by the constant transition concentration u_{RG}. Regions where $u \ll u_{RG}$ correspond to the "glassy" phase while regions where $u \gg u_{RG}$ are "rubbery." When u is in or near a Δ-neighborhood of u_{RG} the polymer is in a nebulous phase transition state.

Since this simplified model is motivated by a diffusion problem we have continued to refer to u as a concentration. However, because the underlying physics may have been lost in the simplification, it may not actually have the correct physical properties.

We note that

$$0 < \gamma_G \leqslant \gamma(y) \leqslant \gamma_R \qquad \forall y \in \mathbb{R} \tag{6}$$

and

$$\gamma'(y) = \frac{\gamma_R - \gamma_G}{2\Delta}\operatorname{sech}^2\left(\frac{y - u_{RG}}{\Delta}\right), \tag{7}$$

so that

$$0 \leqslant \gamma'(y) \leqslant C'_\gamma := \frac{\gamma_R - \gamma_G}{2\Delta} \qquad \forall y \in \mathbb{R}. \tag{8}$$

Also,

$$\gamma''(y) = -\left(\frac{\gamma_R - \gamma_G}{\Delta^2}\right) \tanh\left(\frac{y - u_{RG}}{\Delta}\right) \operatorname{sech}^2\left(\frac{y - u_{RG}}{\Delta}\right),$$

which gives

$$|\gamma''(y)| \leqslant C''_\gamma := \frac{\gamma_R - \gamma_G}{\Delta^2} \qquad \forall y \in \mathbb{R}. \tag{9}$$

We also note that we can solve (2) to get

$$\boldsymbol{\sigma}(t) = \breve{\boldsymbol{\sigma}} e^{-\int_0^t \gamma(u(\xi))\, d\xi} + \mu \int_0^t e^{-\int_s^t \gamma(u(\xi))\, d\xi} \nabla u(s)\, ds \tag{10}$$

and use this in (1) to arrive at (assuming $\breve{\boldsymbol{\sigma}} = \mathbf{0}$)

$$u_t(t) - \nabla \cdot D\nabla u(t) = f(t) + \nabla \cdot \mu K \int_0^t e^{-\int_s^t \gamma(u(\xi))\, d\xi} \nabla u(s)\, ds. \tag{11}$$

We recognize this as a parabolic partial differential equation with a nonlinear Volterra-type memory term typical of that arising in viscoelasticity theory. We could work directly with this formulation in constructing our numerical approximation, but we prefer to work with the system, (1) with (2), since we then need not be concerned with the discretization of the Volterra integral. Also, representing viscoelasticity through evolution equations for internal variables is often preferred to the use of Volterra integrals. See, for example, [11, 10, 3]. It is important to realize that introducing internal variables does not introduce more unknowns and lead to a more complex scheme than would result from using the Volterra formulation directly. In the latter case the "history" in the Volterra integral needs to be stored and updated at each time step. This is exactly analogous to storing the previous value of the internal variable and then updating it through a time-stepping scheme.

This is the third in a series of papers extending the (spatially) *discontinuous Galerkin (DG) finite element method (FEM)* to viscoelasticity problems. In [16] we considered an elliptic stress analysis problem with memory and in [17] we extended this to a second-order hyperbolic problem with memory. Both of these deal only with linear problems, but below we "complete the set" by considering a parabolic problem and including a physically relevant nonlinearity.

DG methods offer several advantages. The lack of continuity constraints between the local approximations allows for an easy implementation of mesh adaptivity. Unlike the classical continuous FEM, the DG method can handle unstructured nonconforming meshes with several hanging nodes per edge (or face). In addition, increasing the polynomial degree does not require any major modification of the software. It is relatively easy as well to have polynomial degrees that vary from one mesh element to the next. Finally, one inherent property of DG methods is the local mass conservation. While this property is essential in many flow and transport problems, it remains to be seen that local mass conservation is important for non-Fickian polymer diffusion problems. A deeper numerical investigation is needed. This will be the object of future work.

The layout of this article is as follows. In section 2 the equations are spatially discretized using an interior penalty DG FEM, and we consider both the symmetric and nonsymmetric variants. The time discretization is a standard Crank–Nicolson

method with a choice of treatments for the nonlinear term. Either this term is approximated in an implicit way, which involves a nonlinear equation set at each time level, or it is handled by extrapolating the current approximation of u to the current time level from the two previous time levels (similarly to [5]). Special care is needed at the first time step, but we can show optimal second-order convergence in each case. The error estimates are contained in section 3 and some numerical experiments are given in section 4. We finish with some comments regarding our model and approach in section 5, as well as discuss the potential for extending this work to Cohen, White, and Witelski's model.

For background on the DG FEM we refer the reader to [18, 21, 9, 20], and for the numerical analysis of generic parabolic problems with memory we refer the reader to [5, 12, 23, 26, 13] (but there are many others).

However, apart from [17], we are not aware of any error analysis for numerical approximations to viscoelasticity problems where the Volterra integral is replaced with internal variable evolution equations, such as (2).

Our notation is standard. For $\omega \subseteq \bar{\Omega}$ we use $(\cdot,\cdot)_\omega$ to denote the $L_2(\omega)$ inner product and simply write $(\cdot,\cdot)$ when $\omega = \Omega$. Also, we use $\|\cdot\|_{p,\omega}$ to denote the $H^p(\omega)$ norm and write $\|\cdot\|_m$ as an abbreviation for $\|\cdot\|_{m,\Omega}$.

We set $\boldsymbol{H}^p(\Omega) := (H^p(\Omega))^d$, but in the notation just described we do not distinguish between inner products and norms on $H^p(\Omega)$ (as used for u) and inner products and norms on $\boldsymbol{H}^p(\Omega)$ (as used for $\boldsymbol{\sigma}$).

Since our functions are time dependent we take the usual approach of thinking of them as maps from I to some underlying Banach space, X. For $1 \leqslant p < \infty$ the $L_p(0,t;X)$ norms are given by

$$\|v\|_{L_p(0,t;X)} := \left(\int_0^t \|v(t)\|_X^p \, dt\right)^{1/p},$$

with the usual "ess sup" modification when $p = \infty$.

To finish this introduction we derive a stability estimate for this problem; the proof is a model of how to proceed with the estimates for the discrete scheme that follows. To begin we note that if

$$V := \{v \in H^1(\Omega) \colon v = 0 \text{ on } \Gamma_D\},$$

then a variational formulation of (1)–(2) is as follows: find maps $u \colon I \to V$ and $\boldsymbol{\sigma} \colon I \to \boldsymbol{L}_2(\Omega)$ such that

$$(12)\quad (u_t(t), v) + (D\nabla u(t), \nabla v) + (K\boldsymbol{\sigma}(t), \nabla v) = L(t; v) \qquad \forall v \in V,$$

$$(13)\quad (\boldsymbol{\sigma}_t(t) + \gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{w}) = (\mu\nabla u(t), \boldsymbol{w}) \qquad \forall \boldsymbol{w} \in \boldsymbol{L}_2(\Omega),$$

where

$$(14)\qquad L(t; v) := (f(t), v) + (g(t), v)_{\Gamma_N}.$$

We can now state a basic stability estimate which does not require Gronwall's lemma.

THEOREM 1.1 (basic stability). *There exists a constant $C > 0$, independent of T, such that, if $(u, \boldsymbol{\sigma})$ is a solution of (12), (13), then*

$$\|u(t)\|_0^2 + \|\boldsymbol{\sigma}(t)\|_0^2 + \int_0^t \left(\|D^{1/2}\nabla u(s)\|_0^2 + \|\boldsymbol{\sigma}(s)\|_0^2\right) ds \leqslant C\left(\|\breve{u}\|_0^2 + \|\breve{\boldsymbol{\sigma}}\|_0^2 + \|f\|_{L_2(0,t;L_2(\Omega))}^2 + \|g\|_{L_2(0,t;L_2(\Gamma_N))}^2\right)$$

for all $t > 0$.

Proof. Choose $v = u$ in (12) and $\boldsymbol{w} = (K/\mu)\boldsymbol{\sigma}(t)$ in (13) and add the resulting equations to get

$$\begin{aligned}(u_t(t), u(t)) + (D\nabla u(t), \nabla u(t)) + (K\boldsymbol{\sigma}(t), \nabla u(t))\\ + \frac{K}{\mu}(\boldsymbol{\sigma}_t(t), \boldsymbol{\sigma}(t)) + \frac{K}{\mu}(\gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{\sigma}(t)) - (K\boldsymbol{\sigma}(t), \nabla u(t))\\ = (f(t), u(t)) + (g(t), u(t))_{\Gamma_N}.\end{aligned}$$

Hence, using Poincaré's inequality

$$\begin{aligned}\frac{d}{dt}\|u(t)\|_0^2 + \frac{K}{\mu}\frac{d}{dt}\|\boldsymbol{\sigma}(t)\|_0^2 + 2\|D^{1/2}\nabla u(t)\|_0^2 + \frac{2K}{\mu}(\gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{\sigma}(t))\\ \leqslant 2C\|f(t)\|_0\|D^{1/2}\nabla u(t)\|_0 + 2C\|g(t)\|_{0,\Gamma_N}\|D^{1/2}\nabla u(t)\|_0\\ \leqslant 2C^2\|f(t)\|_0^2 + 2C^2\|g(t)\|_{0,\Gamma_N}^2 + \|D^{1/2}\nabla u(t)\|_0^2.\end{aligned}$$

Integrating then gives

$$\begin{aligned}\|u(t)\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}(t)\|_0^2 + \int_0^t \left(\|D^{1/2}\nabla u(s)\|_0^2 + \frac{2K\gamma_G}{\mu}\|\boldsymbol{\sigma}(s)\|_0^2\right) ds\\ \leqslant \|\breve{u}\|_0^2 + \frac{K}{\mu}\|\breve{\boldsymbol{\sigma}}\|_0^2 + 2C^2\left(\|f\|_{L_2(0,t;L_2(\Omega))}^2 + \|g\|_{L_2(0,t;L_2(\Gamma_N))}^2\right).\end{aligned}$$

This concludes the proof. ∎

Last in this section, we recall Young's inequality in the form

$$ab \leqslant \frac{a^p}{p\epsilon^p} + \frac{\epsilon^q b^q}{q} \tag{15}$$

for all $a, b \geqslant 0$, $\epsilon > 0$, and $p, q \in (1, \infty)$ such that $1/p + 1/q = 1$.

2. The numerical scheme. The first step is to establish notation for the spatial discretization. Let $\mathcal{E}_h = \{E\}$ be a nondegenerate quasiuniform subdivision of Ω, where E is a triangle if $d = 2$, or a tetrahedron if $d = 3$. The nondegeneracy requirement is that there exists $\rho > 0$ such that if $h_E = \operatorname{diam}(E)$, then E contains a ball of radius ρh_E in its interior. Let $h = \max\{h_E : E \in \mathcal{E}_h\}$; the quasiuniformity requirement is that there exists $\tau > 0$ such that $h/h_E \leqslant \tau$ for all $E \in \mathcal{E}_h$. We denote by Γ_h the set of interior edges (faces for $d = 3$) of $\mathcal{E}_h$. With each edge (or face) e, we associate a unit normal vector $\boldsymbol{n}_e$. For a boundary edge e, $\boldsymbol{n}_e$ is taken to be the unit outward vector normal to $\partial\Omega$.

We now define the average and the jump operators. For each of the interior edges, suppose that e is shared by E_1^e and E_2^e such that $\boldsymbol{n}_e$ points from E_1^e to E_2^e and for a boundary edge, suppose that e belongs to E_1^e. We define the averaging operator $\{\cdot\}$ by

$$\{w\} := \begin{cases} \frac{1}{2}(w|_{E_1^e})|_e + \frac{1}{2}(w|_{E_2^e})|_e & \text{if } e \subset \Omega,\\ (w|_{E_1^e})|_e & \text{if } e \subset \partial\Omega \end{cases}$$

and the jump operator $[\cdot]$ by

$$[w] := \begin{cases} (w|_{E_1^e})|_e - (w|_{E_2^e})|_e & \text{if } e \subset \Omega,\\ (w|_{E_1^e})|_e & \text{if } e \subset \partial\Omega. \end{cases}$$

The distinction between $[\cdot]$ and $-[\cdot]$ can be made because each edge e_a has a unit normal associated with it. The "direction" in which the jump takes place is unimportant.

These operators are well defined if $w|_{E_a^i} \in H^{\frac{1}{2}+\epsilon}(E_a^i)$ for $i = 1, 2$ and $\epsilon > 0$. Below, we use $|e|$ to denote the $(d-1)$-dimensional surface measure of the edge/face e. We also frequently use the estimate, $|e| \leqslant Ch^{d-1}$ which arises as a consequence of our assumptions.

Define the broken spaces for any integer $r > 0$ as

$$\begin{aligned}
\mathcal{D}_r(\mathcal{E}_h) &= \{v \in L_2(\Omega) : v|_E \in \mathbb{P}_r(E) \quad \forall E \in \mathcal{E}_h\},\\
\boldsymbol{\mathcal{D}}_r(\mathcal{E}_h) &= \mathcal{D}_r(\mathcal{E}_h)^d,\\
H^n(\mathcal{E}_h) &= \{v \in L_2(\Omega) : v|_E \in H^n(E) \quad \forall E \in \mathcal{E}_h\}.
\end{aligned}$$

For these finite element spaces we have the following interpolation-error estimates. If $v \in H^n(\mathcal{E}_h) \cap C(\bar{\Omega})$ and $\mu = \min\{r+1, n\}$ then there is an interpolant $\hat{v} \in \mathcal{D}_r(\mathcal{E}_h) \cap C(\bar{\Omega})$ such that for each $E \in \mathcal{E}_h$,

$$\|v - \hat{v}\|_{m,E} \leqslant Ch_E^{\mu-m}\|v\|_{n,E} \qquad \text{for } n \geqslant m \geqslant 0, \tag{16}$$

$$\|v - \hat{v}\|_{m,\gamma} \leqslant Ch_E^{\mu-m-1/2}\|v\|_{n,E} \qquad \text{for } m = 0, 1 \text{ and } n \geqslant m, \tag{17}$$

where $\gamma \subseteq \partial E$.

Define the bilinear forms

$$J_0^{\delta,\beta}(w,v) = \sum_{e\in\Gamma_h\cup\Gamma_D} \frac{\delta}{|e|^\beta}\int_e [w][v] \qquad \text{for } \beta \geqslant (d-1)^{-1},$$

$$\begin{aligned}
A(w,v) = \sum_E \int_E D\nabla w\cdot\nabla v + J_0^{\delta,\beta}(w,v) - \sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{D\nabla w\cdot \boldsymbol{n}_e\}[v]\\
+ \kappa\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{D\nabla v\cdot \boldsymbol{n}_e\}[w].
\end{aligned}$$

Here κ is a switch: we set $\kappa = 1$ to obtain the nonsymmetric DG scheme, and $\kappa = -1$ to obtain the symmetric scheme. Following from these definitions are the norm and seminorm

$$\|v\|_{\mathcal{A}} := \left(|v|_{\mathcal{E}}^2 + J_0^{\delta,\beta}(v,v)\right)^{\frac{1}{2}} \qquad \text{and} \qquad |v|_{\mathcal{E}} := \left(\sum_{E\in\mathcal{E}_h}\int_E D\nabla v\cdot\nabla v\, dE\right)^{\frac{1}{2}}.$$

We will need the following estimates.

LEMMA 2.1. *We have*

$$\|v\|_0 \leqslant C_f\|v\|_{\mathcal{A}} \qquad \forall v \in H^1(\mathcal{E}_h) \tag{18}$$

and

$$\|v\|_{0,\Gamma_N} \leqslant C_g h^{-1/2}\|v\|_{\mathcal{A}} \qquad \forall v \in \mathcal{D}_r(\mathcal{E}_h),$$

for constants C_f and C_g, independent of h.

Proof. For the first inequality we refer to [9, Lemma 6.2], and for the second inequality we use the first one with Sobolev interpolation to get

$$\|v\|_{0,\Gamma_N}^2 = \sum_{e\in\Gamma_N}\|v\|_{0,e}^2 \leqslant C\sum_E h^{-1}\|v\|_{0,E}\|\nabla v\|_{0,E} \leqslant Ch^{-1}(C_f^2 + D^{-1})\|v\|_{\mathcal{A}}^2.$$

This completes the proof. ∎

We note that if $u(t) \in C(\bar{\Omega})$ for each t, then

$$(u_t(t), v) + A(u(t), v) = L(t; v) - \sum_E (K\boldsymbol{\sigma}(t), \nabla v)_E + \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K\boldsymbol{\sigma}(t) \cdot \boldsymbol{n}_e\}[v] \qquad \forall v \in \mathcal{D}_r(\mathcal{E}_h), \tag{19}$$

and

$$(\boldsymbol{\sigma}_t(t) + \gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{w}) = \sum_E (\mu \nabla u(t), \boldsymbol{w})_E \qquad \forall \boldsymbol{w} \in \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h). \tag{20}$$

The first of these arises by elementwise partial integration and "adding zero" (see, e.g., [21]).

To construct a fully discrete approximation we set $k = T/N$ for some $N \in \mathbb{N}$, and write $t_i = ik$. To ease notation we define

$$\partial_t w_i := \frac{w(t_i) - w(t_{i-1})}{k} \qquad \text{and} \qquad \bar{w}_i := \frac{w(t_i) + w(t_{i-1})}{2}.$$

The fully discrete approximations, u^h and $\boldsymbol{\sigma}^h$, to u and $\boldsymbol{\sigma}$ are continuous and piecewise linear in time, and discontinuous in space. We set $u_i^h := u^h(t_i)$ and $\boldsymbol{\sigma}_i^h := \boldsymbol{\sigma}^h(t_i)$.

An issue is how to handle the nonlinearity, $\gamma(u)$, in the numerical scheme. We offer two possibilities by approximating $\gamma(u)|_{(t_{i-1},t_i)}$ by $\gamma(\mathcal{B}_{i,n}u^h)$ for $n = 1$ or 2, where

$$\mathcal{B}_{i,1}u^h := \bar{u}_i^h \qquad \text{and} \qquad \mathcal{B}_{i,2}u^h := \mathcal{E}_i u^h,$$

with $\mathcal{E}_i$ an extrapolation operator defined by

$$\mathcal{E}_i u^h := \begin{cases} u_0^h & \text{for } i = 1; \\ \frac{3}{2}u_{i-1}^h - \frac{1}{2}u_{i-2}^h & \text{for } i = 2, \ldots, N. \end{cases}$$

In the first case we approximate $\gamma(u)|_{(t_{i-1},t_i)}$ by taking the true average, $\bar{u}_i^h$, of the discrete solution. This will result in a nonlinear system to be solved at each time level. To linearize this system, the second method linearly extrapolates to the average based on the two previous solutions. This is not possible at the first time step and so this first step will require special treatment in the error estimation. This extrapolation technique is widely used in coupled flow and transport problems such as miscible displacement. See, for example, [7] and [19].

The fully discrete approximations (i.e., for $n = 1$ or 2) are based on sampling (19) and (20) at the temporal midpoints, $t_{i-1/2}$. They are defined as follows: for each $i = 1, 2, \ldots, N$, find a pair $\{u_i^h, \boldsymbol{\sigma}_i^h\} \in \mathcal{D}_r(\mathcal{E}_h) \times \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h)$ such that

$$(\partial_t u_i^h, v) + A(\bar{u}_i^h, v) = L_i(v) - \sum_E (K\bar{\boldsymbol{\sigma}}_i^h, \nabla v)_E + \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K\bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\}[v] \qquad \forall v \in \mathcal{D}_r(\mathcal{E}_h), \tag{21}$$

and

$$(\partial_t \boldsymbol{\sigma}_i^h + \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^h, \boldsymbol{w}) = \sum_E (\mu \nabla \bar{u}_i^h, \boldsymbol{w})_E \qquad \forall \boldsymbol{w} \in \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h), \tag{22}$$

where

$$L_i(v) := \frac{1}{2}\Big(L(t_i; v) + L(t_{i-1}; v)\Big),$$

and the discrete initial data are given by

$$\begin{aligned} (u_0^h, v) &= (\breve{u}, v) && \forall v \in \mathcal{D}_r(\mathcal{E}_h), \\ (\boldsymbol{\sigma}_0^h, \boldsymbol{w}) &= (\breve{\boldsymbol{\sigma}}, \boldsymbol{w}) && \forall \boldsymbol{w} \in \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h). \end{aligned}$$

We now give a stability estimate for this discrete approximation and note that Gronwall's lemma is not used. We also note that the "h^{-1}" factor appearing in front of the boundary term is a weakness in the proof and is not observed in computations. It appears that the removal of this factor is an open problem (although, see Remark 3.6 later).

THEOREM 2.2 (discrete basic stability). *If $\beta \geqslant (d-1)^{-1}$ and $h \leqslant \hat{h}$ we have for $m = 1, 2, \ldots, N$ that*

$$\begin{aligned} \|u_m^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_m^h\|_0^2 + C^\star k \sum_{i=1}^m \left(\|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2K\|\bar{\boldsymbol{\sigma}}_i^h\|_0^2\right) \\ \leqslant \|\breve{u}\|_0^2 + \frac{K}{\mu}\|\breve{\boldsymbol{\sigma}}\|_0^2 + 6k\sum_{i=1}^m \left(C_f^2\|\bar{f}_i\|_0^2 + C_g^2 h^{-1}\|\bar{g}_i\|_{0,\Gamma_N}^2\right), \end{aligned}$$

provided that

$$\delta \geqslant 3C\hat{h}^{(d-1)\beta-1}\max\left\{\frac{4D}{1-C^\star}, \frac{\mu K}{2\gamma_G - 2\mu C^\star}\right\},$$

where $C^\star < \min\{1, \gamma_G/\mu\}$ is some chosen positive constant, C is independent of h, and C_f and C_g are those in Lemma 2.1.

Proof. Choose $v = \bar{u}_i^h$ in (21) and $\boldsymbol{w} = (K/\mu)\bar{\boldsymbol{\sigma}}_i^h$ in (22) and note that

$$\begin{aligned} (\partial_t u_i^h, \bar{u}_i^h) &= \frac{1}{2k}\|u_i^h\|_0^2 - \frac{1}{2k}\|u_{i-1}^h\|_0^2, \\ (\partial_t \boldsymbol{\sigma}_i^h, \bar{\boldsymbol{\sigma}}_i^h) &= \frac{1}{2k}\|\boldsymbol{\sigma}_i^h\|_0^2 - \frac{1}{2k}\|\boldsymbol{\sigma}_{i-1}^h\|_0^2, \\ A(\bar{u}_i^h, \bar{u}_i^h) &= \sum_E (D\nabla\bar{u}_i^h, \nabla\bar{u}_i^h)_E + J_0^{\delta,\beta}(\bar{u}_i^h, \bar{u}_i^h) \\ &\qquad + (\kappa - 1)\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{D\nabla\bar{u}_i^h \cdot \boldsymbol{n}_e\}[\bar{u}_i^h]. \end{aligned}$$

Adding the two resulting equations then gives

$$\begin{aligned} &\frac{1}{2k}\|u_i^h\|_0^2 - \frac{1}{2k}\|u_{i-1}^h\|_0^2 + \frac{K}{2k\mu}\|\boldsymbol{\sigma}_i^h\|_0^2 - \frac{K}{2k\mu}\|\boldsymbol{\sigma}_{i-1}^h\|_0^2 + \|\bar{u}_i^h\|_{\mathcal{A}}^2 + \frac{K}{\mu}(\gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^h, \bar{\boldsymbol{\sigma}}_i^h) \\ &\quad = L_i(\bar{u}_i^h) - (\kappa-1)\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{D\nabla\bar{u}_i^h \cdot \boldsymbol{n}_e\}[\bar{u}_i^h] + \sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{K\bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\}[\bar{u}_i^h], \end{aligned}$$

and summing over $i = 1, \ldots, m$ and multiplying by $2k$ yields

$$\begin{aligned} &\|u_m^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_m^h\|_0^2 + 2k\sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k\sum_{i=1}^m \frac{K}{\mu}(\gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^h, \bar{\boldsymbol{\sigma}}_i^h) \\ &\quad = \|u_0^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_0^h\|_0^2 + 2k\sum_{i=1}^m L_i(\bar{u}_i^h) + I + II, \end{aligned}$$

where

$$I = 2k \sum_{i=1}^{m} \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\} [\bar{u}_i^h],$$
$$II = 2k \sum_{i=1}^{m} (1-\kappa) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D \nabla \bar{u}_i^h \cdot \boldsymbol{n}_e\} [\bar{u}_i^h].$$

Now, using $\|\bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\|_{0,\partial E} \leqslant C h^{-1/2} \|\bar{\boldsymbol{\sigma}}_i^h\|_{0,E}$ and recalling that $|e| \leqslant C h^{d-1}$, for I we have

$$\begin{aligned}
|I| &\leqslant 2k \sum_{i=1}^{m} \sum_{e \in \Gamma_h \cup \Gamma_D} K \|\{\bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\}\|_{0,e} \|[\bar{u}_i^h]\|_{0,e}, \\
&\leqslant 2\epsilon_1 k \sum_{i=1}^{m} \sum_{e \in \Gamma_h \cup \Gamma_D} K^2 \left(\frac{|e|^\beta}{\delta} \right) \|\{\bar{\boldsymbol{\sigma}}_i^h \cdot \boldsymbol{n}_e\}\|_{0,e}^2 + \frac{k}{2\epsilon_1} \sum_{i=1}^{m} \sum_{e \in \Gamma_h \cup \Gamma_D} \left(\frac{\delta}{|e|^\beta} \right) \|[\bar{u}_i^h]\|_{0,e}^2, \\
&\leqslant 2\epsilon_1 k \sum_{i=1}^{m} \frac{K^2 C h^{(d-1)\beta - 1}}{\delta} \|\bar{\boldsymbol{\sigma}}_i^h\|_0^2 + \frac{k}{2\epsilon_1} \sum_{i=1}^{m} J_0^{\delta,\beta}(\bar{u}_i^h, \bar{u}_i^h).
\end{aligned}$$

Similarly, since $|\kappa - 1| \leqslant 2$,

$$\begin{aligned}
|II| &\leqslant 4k \sum_{i=1}^{m} \sum_{e \in \Gamma_h \cup \Gamma_D} \left(\frac{|e|^\beta}{\delta} \right)^{1/2} \|\{D \nabla \bar{u}_i^h \cdot \boldsymbol{n}_e\}\|_{0,e} \left(\frac{\delta}{|e|^\beta} \right)^{1/2} \|[\bar{u}_i^h]\|_{0,e}, \\
&\leqslant 2\epsilon_2 k \sum_{i=1}^{m} \sum_{E} \frac{D C h^{(d-1)\beta - 1}}{\delta} \|D^{1/2} \nabla \bar{u}_i^h\|_{0,E}^2 + \frac{2k}{\epsilon_2} \sum_{i=1}^{m} J_0^{\delta,\beta}(\bar{u}_i^h, \bar{u}_i^h).
\end{aligned}$$

With these we arrive at

$$\begin{aligned}
&\|u_m^h\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\sigma}_m^h\|_0^2 + \left(2 - \frac{1}{2\epsilon_1} - \frac{2}{\epsilon_2} \right) k \sum_{i=1}^{m} \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^{m} \frac{K}{\mu} (\gamma(\mathcal{B}_{i,n} u^h) \bar{\boldsymbol{\sigma}}_i^h, \bar{\boldsymbol{\sigma}}_i^h) \\
&\quad \leqslant \|u_0^h\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\sigma}_0^h\|_0^2 + 2k \left| \sum_{i=1}^{m} L_i(\bar{u}_i^h) \right| \\
&\qquad + 2k \sum_{i=1}^{m} \frac{C h^{(d-1)\beta - 1}}{\delta} \left(K^2 \epsilon_1 \|\bar{\boldsymbol{\sigma}}_i^h\|_0^2 + D \epsilon_2 \|\bar{u}_i^h\|_{\mathcal{A}}^2 \right).
\end{aligned}$$

Now, using Lemma 2.1,

$$\begin{aligned}
2k \left| \sum_{i=1}^{m} L_i(\bar{u}_i^h) \right| &= k \left| \sum_{i=1}^{m} \left(L(t_i; \bar{u}_i^h) + L(t_{i-1}; \bar{u}_i^h) \right) \right| \\
&= k \left| \sum_{i=1}^{m} \left((f(t_i), \bar{u}_i^h) + (f(t_{i-1}), \bar{u}_i^h) + (g(t_i), \bar{u}_i^h)_{\Gamma_N} + (g(t_{i-1}), \bar{u}_i^h)_{\Gamma_N} \right) \right| \\
&= 2k \left| \sum_{i=1}^{m} \left((\bar{f}_i, \bar{u}_i^h) + (\bar{g}_i, \bar{u}_i^h)_{\Gamma_N} \right) \right| \\
&\leqslant 2k \sum_{i=1}^{m} C_f \|\bar{f}_i\|_0 \|\bar{u}_i^h\|_{\mathcal{A}} + 2k \sum_{i=1}^{m} C_g h^{-1/2} \|\bar{g}_i\|_{0,\Gamma_N} \|\bar{u}_i^h\|_{\mathcal{A}}
\end{aligned}$$

$$\leqslant \epsilon_3 k \sum_{i=1}^{m} C_f^2 \|\bar{f}_i\|_0^2 + \epsilon_3 k \sum_{i=1}^{m} C_g^2 h^{-1} \|\bar{g}_i\|_{0,\Gamma_N}^2 + \frac{2k}{\epsilon_3} \sum_{i=1}^{m} \|\bar{u}_i^h\|_{\mathcal{A}}^2.$$

With this and (6), we now have

$$\begin{aligned}
\|u_m^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_m^h\|_0^2 + \left(2 - \frac{1}{2\epsilon_1} - \frac{2}{\epsilon_2} - \frac{2}{\epsilon_3}\right) k \sum_{i=1}^{m} \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^{m} \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\sigma}}_i^h\|_0^2 \\
\leqslant \|u_0^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_0^h\|_0^2 + \epsilon_3 k \sum_{i=1}^{m} \Big(C_f^2 \|\bar{f}_i\|_0^2 + C_g^2 h^{-1} \|\bar{g}_i\|_{0,\Gamma_N}^2 \Big) \\
+ 2k \sum_{i=1}^{m} \frac{C h^{(d-1)\beta - 1}}{\delta} \left(K^2 \epsilon_1 \|\bar{\boldsymbol{\sigma}}_i^h\|_0^2 + D\epsilon_2 \|\bar{u}_i^h\|_{\mathcal{A}}^2 \right).
\end{aligned}$$

Setting $\epsilon_2 = \epsilon_3 = 6$ and $\epsilon_1 = 3/2$ means that we can write this as

$$\begin{aligned}
\|u_m^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_m^h\|_0^2 + \left(1 - \frac{12DC\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \sum_{i=1}^{m} \|\bar{u}_i^h\|_{\mathcal{A}}^2 \\
+ \left(\frac{\gamma_G}{\mu} - \frac{3KC\hat{h}^{(d-1)\beta-1}}{2\delta}\right) 2Kk \sum_{i=1}^{m} \|\bar{\boldsymbol{\sigma}}_i^h\|_0^2 \\
\leqslant \|u_0^h\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\sigma}_0^h\|_0^2 + 6k \sum_{i=1}^{m} \Big(C_f^2 \|\bar{f}_i\|_0^2 + C_g^2 h^{-1} \|\bar{g}_i\|_{0,\Gamma_N}^2 \Big),
\end{aligned}$$

and choosing some positive constant $C^\star < \min\{1, \gamma_G/\mu\}$ and requiring that

$$\delta \geqslant 3C\hat{h}^{(d-1)\beta-1} \max\left\{ \frac{4D}{1 - C^\star}, \frac{\mu K}{2\gamma_G - 2\mu C^\star} \right\},$$

we arrive at the theorem. ∎

Since this is a finite dimensional problem, we can infer existence from uniqueness in the linear case where $n = 2$. Since this is the more practical of the two algorithms we are content with this. Also, at least for the original model of Cohen et al., [6], it seems from [2] that such analysis for the nonlinear problem is highly nontrivial.

THEOREM 2.3 (discrete existence and uniqueness). *Under the conditions of Theorem* 2.2*, the discrete solution exists for* $n = 2$ *and is unique.*

REMARK 2.4. *The condition that* δ *"be large enough" in Theorem* 2.2 *can be removed in the nonsymmetric case,* $\kappa = 1$*, by requiring a small enough time step,* k*. To see this note that the term* II *in the proof vanishes and that the second term in the bound for* I *can be moved to the left with an appropriate choice of* ϵ_1*. After applying the triangle inequality to* $\|\bar{\boldsymbol{\sigma}}_i^h\|_0^2$*, the term* $\|\boldsymbol{\sigma}_m^h\|_0^2$ *can also be moved to the left if* k *is small enough, and the remaining terms are bounded by a discrete Gronwall inequality.*

3. Error estimate. In this section we derive error estimates for our schemes encompassing the cases $\kappa = \pm 1$ and $n = 1$ or 2. First we need some standard Taylor series estimates, and it is convenient to define

$$\Delta_i v := \frac{v_t(t_i) + v_t(t_{i-1})}{2} - \frac{v(t_i) - v(t_{i-1})}{k},$$

which we recognize as (the negative of) the error in the trapezium rule.

LEMMA 3.1 (Taylor estimates). *Whenever v has the indicated regularity we have positive constants, C, independent of h and k such that*

$$\|v(t_{i-1/2}) - \bar{v}_i\|_0 \leqslant Ck^{3/2}\|v_{tt}\|_{L_2(t_{i-1},t_i;L_2(\Omega))}, \tag{23}$$

$$\|v(k/2) - v(0)\|_0 \leqslant Ck\|v_t\|_{L_\infty(0,k/2;L_2(\Omega))}, \tag{24}$$

$$\left\|v(t_{i-1/2}) - \frac{3v(t_{i-1}) - v(t_{i-2})}{2}\right\|_0 \leqslant Ck^{3/2}\|v_{tt}\|_{L_2(t_{i-2},t_{i-1/2};L_2(\Omega))}, \tag{25}$$

$$\|v_t(t_{i-1/2}) - \partial_t v_i\|_0 \leqslant Ck^{3/2}\|v_{ttt}\|_{L_2(t_{i-1},t_i;L_2(\Omega))}, \tag{26}$$

and

$$\|\Delta_i v\|_0 \leqslant Ck^{3/2}\|v_{ttt}\|_{L_2(t_{i-1},t_i;L_2(\Omega))}, \tag{27}$$

from the Peano kernel theorem applied to the trapezoidal rule for numerical integration.

We define

$$\chi_i := u_i^h - u^\perp(t_i), \qquad \boldsymbol{\eta}_i := \boldsymbol{\sigma}_i^h - \boldsymbol{\sigma}^*(t_i),$$
$$\xi(t_i) := u(t_i) - u^\perp(t_i), \qquad \boldsymbol{\theta}(t_i) := \boldsymbol{\sigma}(t_i) - \boldsymbol{\sigma}^*(t_i),$$

where $\boldsymbol{\sigma}^* \in \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h)$ is the nodal interpolant to $\boldsymbol{\sigma}$, and $u^\perp \in \mathcal{D}_r(\mathcal{E}_h)$ is the elliptic projection of u defined by

$$A(u^\perp, v) = A(u, v) \qquad \forall v \in \mathcal{D}_r(\mathcal{E}_h). \tag{28}$$

PROPOSITION 3.2 (estimates for the elliptic projection). *If $u \in C(\bar{\Omega})$ and $u^\perp \in \mathcal{D}_r(\mathcal{E}_h)$ is defined through* (28) *for $\kappa = \pm 1$, we have for $m = 0, 1, 2, \ldots$ and $t \geq 0$ that*

$$\left\|\frac{\partial^m}{\partial t^m}\Big(u(t) - u^\perp(t)\Big)\right\|_{\mathcal{A}} \leqslant Ch^s \left\|\frac{\partial^m u}{\partial t^m}(t)\right\|_{s+1}, \tag{29}$$

$$\left\|\frac{\partial^m}{\partial t^m}\Big(u(t) - u^\perp(t)\Big)\right\|_0 \leqslant Ch^s \left\|\frac{\partial^m u}{\partial t^m}(t)\right\|_{s+1}, \tag{30}$$

$$\left\|\frac{\partial^m u^\perp}{\partial t^m}(t)\right\|_{\mathcal{A}} \leqslant C \left\|\frac{\partial^m u}{\partial t^m}(t)\right\|_2, \tag{31}$$

whenever $\partial^m u(t)/\partial t^m \in H^{s+1}(\Omega)$ and $1 \leqslant s \leqslant r$.

When $m = 0$ the proof of (29) is given in [21] (the "NIPG" scheme) for the nonsymmetric case, $\kappa = 1$, and can be readily established for $\kappa = -1$ by similar arguments. The nonoptimal (30) then follows from (29) and (18) (an optimal L_2 estimate is also given in [21], but we do not need it here). The stability estimate, (31), follows from

$$\|u^\perp(t)\|_{\mathcal{A}} \leqslant \|u(t) - u^\perp(t)\|_{\mathcal{A}} + \|u(t)\|_{\mathcal{A}},$$

along with (29) (with $s = 1$) and the fact that $[u(t)] = 0$. The estimates then follow for $m \geqslant 1$ by differentiating (28).

For use later, we note also that

$$\|\boldsymbol{\sigma}^*(t)\|_{\boldsymbol{L}_\infty(\Omega)} \leqslant C\|\boldsymbol{\sigma}(t)\|_{\boldsymbol{L}_\infty(\Omega)}. \tag{32}$$

The next result is a lemma that deals with the error generated by the nonlinear term.

LEMMA 3.3 (nonlinearity error). *For $n = 1$ or 2 we have*

$$\begin{aligned}
&\left|\left((\overline{(\gamma(u)\boldsymbol{\sigma})_i}) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i\right)\right| \\
&\quad\leqslant \frac{Ch^{2r}}{\epsilon}\left(\|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{H}^r(\Omega))} + \|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\|u\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}\right) \\
&\qquad + \frac{Ck^3}{\epsilon}\left(\|(\gamma(u)\boldsymbol{\sigma})_{tt}\|^2_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))} + \|\boldsymbol{\sigma}_{tt}\|^2_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}\right) \\
&\qquad\quad + \frac{C}{\epsilon}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\|\chi_{i-1}\|_0^2 \\
&+ \begin{cases}
\dfrac{C}{\epsilon}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\left(k^3\|u_{tt}\|^2_{L_2(t_{i-1},t_i;H^2(\Omega))} + \|\chi_i\|_0^2\right) + \dfrac{\gamma_G\epsilon}{2}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \\
\qquad \textit{for } n = 1,\ i = 1,\dots,N, \\[2ex]
\dfrac{Ck^2}{\epsilon}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\|u_t\|^2_{L_\infty(0,k/2;H^2(\Omega))} + \dfrac{\gamma_G\epsilon}{2}\|\boldsymbol{\eta}_1\|_0^2 + \dfrac{\gamma_G\epsilon}{2}\|\boldsymbol{\eta}_0\|_0^2 \\
\qquad \textit{for } n = 2,\ i = 1, \\[2ex]
\dfrac{C}{\epsilon}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\left(k^3\|u_{tt}\|^2_{L_2(t_{i-2},t_{i-1/2};H^2(\Omega))} + \|\chi_{i-2}\|_0^2\right) + \dfrac{\gamma_G\epsilon}{2}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \\
\qquad \textit{for } n = 2,\ i = 2,\dots,N,
\end{cases}
\end{aligned}$$

for a constant C independent of h, k, and ϵ and for all $\epsilon > 0$.

Proof. We have, from (23) in Lemma 3.1,

$$\begin{aligned}
\left|\left((\overline{\gamma(u)\boldsymbol{\sigma}(t_i)}) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i\right)\right| &\leqslant \|\bar{\boldsymbol{\eta}}_i\|_0\Big(\|\overline{\gamma(u)\boldsymbol{\sigma}(t_i)} - \gamma(u(t_{i-1/2}))\boldsymbol{\sigma}(t_{i-1/2})\|_0 \\
&\quad + \|\gamma(u(t_{i-1/2}))\boldsymbol{\sigma}(t_{i-1/2}) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*\|_0\Big), \\
&\leqslant Ck^{3/2}\|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}\|\bar{\boldsymbol{\eta}}_i\|_0 \\
&\quad + \|\bar{\boldsymbol{\eta}}_i\|_0\Big(\|\gamma(u(t_{i-1/2}))(\boldsymbol{\sigma}(t_{i-1/2}) - \bar{\boldsymbol{\sigma}}_i^*)\|_0 \\
&\quad + \|\gamma(u(t_{i-1/2})) - \gamma(\mathcal{B}_{i,n}u^h)\|_0\|\bar{\boldsymbol{\sigma}}_i^*\|_{\boldsymbol{L}_\infty(\Omega)}\Big), \\
&\leqslant Ck^{3/2}\|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}\|\bar{\boldsymbol{\eta}}_i\|_0 \\
&\quad + \gamma_R\|\bar{\boldsymbol{\eta}}_i\|_0\Big(\|\boldsymbol{\sigma}(t_{i-1/2}) - \bar{\boldsymbol{\sigma}}_i\|_0 + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^*\|_0\Big) \\
&\quad + C'_\gamma\|\bar{\boldsymbol{\eta}}_i\|_0\|\bar{\boldsymbol{\sigma}}_i^*\|_{\boldsymbol{L}_\infty(\Omega)}\|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0,
\end{aligned}$$

where we observed, using (8), that

$$\begin{aligned}
\|\gamma(u(t_{i-1/2})) - \gamma(\mathcal{B}_{i,n}u^h)\|_0 &= \left\|\int_0^1 \gamma'(su(t_{i-1/2}) + (1-s)\mathcal{B}_{i,n}u^h)\,ds\ (u(t_{i-1/2}) \right. \\
&\qquad \left. - \mathcal{B}_{i,n}u^h)\right\|_0 \\
&\leqslant C'_\gamma\|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0.
\end{aligned}$$

Using (23), (24), and (25) from Lemma 3.1, along with (16) and (32), we therefore

arrive at

$$\begin{aligned}\left|\left(\overline{(\gamma(u)\boldsymbol{\sigma}(t_i))}-\gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*,\bar{\boldsymbol{\eta}}_i\right)\right| &\leqslant Ck^{3/2}\|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}\|\bar{\boldsymbol{\eta}}_i\|_0\\ &\quad+\gamma_R\|\bar{\boldsymbol{\eta}}_i\|_0\left(Ck^{3/2}\|\boldsymbol{\sigma}_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}+Ch^r\|\boldsymbol{\sigma}\|_{L_\infty(0,T;\boldsymbol{H}^r(\Omega))}\right)\\ &\quad+C\|\bar{\boldsymbol{\eta}}_i\|_0\|\boldsymbol{\sigma}\|_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\|u(t_{i-1/2})-\mathcal{B}_{i,n}u^h\|_0.\end{aligned}$$

Now, using Proposition 3.2,

$$\begin{aligned}\|u(t_{i-1/2})-\mathcal{B}_{i,n}u^h\|_0 &\leqslant \|u(t_{i-1/2})-u^\perp(t_{i-1/2})\|_0+\|u^\perp(t_{i-1/2})-\mathcal{B}_{i,n}u^\perp\|_0\\ &\quad+\|\mathcal{B}_{i,n}u^\perp-\mathcal{B}_{i,n}u^h\|_0,\\ &\leqslant Ch^r\|u\|_{L_\infty(0,T;H^{r+1}(\Omega))}+\|u^\perp(t_{i-1/2})-\mathcal{B}_{i,n}u^\perp\|_0+\|\mathcal{B}_{i,n}\chi\|_0,\end{aligned}$$

and this is as far as we can get without distinguishing between $n=1$ and $n=2$.

So, first, for $n=1$ we have

$$\begin{aligned}\|u(t_{i-1/2})-\mathcal{B}_{i,1}u^h\|_0 &\leqslant Ch^r\|u\|_{L_\infty(0,T;H^{r+1}(\Omega))}+Ck^{3/2}\|u_{tt}\|_{L_2(t_{i-1},t_i;H^2(\Omega))}\\ &\quad+\frac{1}{2}\|\chi_i\|_0+\frac{1}{2}\|\chi_{i-1}\|_0,\end{aligned}$$

where we used (23) from Lemma 3.1 and (30) with $m=0$.

Second, using (24) from Lemma 3.1, in the case $n=2$ we have when $i=1$ that

$$\|u(t_{i-1/2})-\mathcal{B}_{i,2}u^h\|_0\leqslant Ch^r\|u\|_{L_\infty(0,T;H^{r+1}(\Omega))}+Ck\|u_t\|_{L_\infty(0,k/2;H^2(\Omega))}+\|\chi_0\|_0,$$

while if $i>1$, with (25) from Lemma 3.1,

$$\begin{aligned}\|u(t_{i-1/2})-\mathcal{B}_{i,2}u^h\|_0 &\leqslant Ch^r\|u\|_{L_\infty(0,T;H^{r+1}(\Omega))}+Ck^{3/2}\|u_{tt}\|_{L_2(t_{i-2},t_{i-1/2};H^2(\Omega))}\\ &\quad+\frac{3}{2}\|\chi_{i-1}\|_0+\frac{1}{2}\|\chi_{i-2}\|_0.\end{aligned}$$

Assembling these estimates then gives

$$\begin{aligned}&\left|\left(\overline{((\gamma(u)\boldsymbol{\sigma})_i)}-\gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*,\bar{\boldsymbol{\eta}}_i\right)\right|\\ &\quad\leqslant Ck^{3/2}\Big(\|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}+\gamma_R\|\boldsymbol{\sigma}_{tt}\|_{L_2(t_{i-1},t_i;\boldsymbol{L}_2(\Omega))}\Big)\|\bar{\boldsymbol{\eta}}_i\|_0\\ &\qquad+Ch^r\Big(\|\boldsymbol{\sigma}\|_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\|u\|_{L_\infty(0,T;H^{r+1}(\Omega))}+\gamma_R\|\boldsymbol{\sigma}\|_{L_\infty(0,T;\boldsymbol{H}^r(\Omega))}\Big)\|\bar{\boldsymbol{\eta}}_i\|_0\\ &\qquad+\|\bar{\boldsymbol{\eta}}_i\|_0\|\boldsymbol{\sigma}\|_{L_\infty(0,T;\boldsymbol{L}_\infty(\Omega))}\times\begin{cases}Ck^{3/2}\|u_{tt}\|_{L_2(t_{i-1},t_i;H^2(\Omega))}+\frac{1}{2}\|\chi_i\|_0+\frac{1}{2}\|\chi_{i-1}\|_0\\ \qquad\text{for } n=1,\ i\geqslant 1;\\ Ck\|u_t\|_{L_\infty(0,k/2;H^2(\Omega))}+\|\chi_0\|_0\\ \qquad\text{for } n=2,\ i=1;\\ Ck^{3/2}\|u_{tt}\|_{L_2(t_{i-2},t_{i-1/2};H^2(\Omega))}\\ \qquad+\frac{3}{2}\|\chi_{i-1}\|_0+\frac{1}{2}\|\chi_{i-2}\|_0\\ \qquad\text{for } n=2,\ i>1.\end{cases}\end{aligned}$$

Several applications of Young's inequality then complete the proof. ∎

Before giving the error estimate we recall from, e.g., [1, Theorem 4.12] that if $\Omega \subset \mathbb{R}^d$ for $d = 2$ or 3 satisfies a cone condition, then $\|v\|_{L_\infty(\Omega)} \leqslant C\|v\|_m$ for $m > d/2$. Moreover,

$$H^1(\Omega) \hookrightarrow L_q(\Omega) \qquad \text{for} \qquad \begin{cases} 2 \leqslant q < \infty & \text{if } d = 2, \\ 2 \leqslant q \leqslant 6 & \text{if } d = 3, \end{cases} \tag{33}$$

and then $\|v\|_{L_q(\Omega)} \leqslant C\|v\|_1$ for all $v \in H^1(\Omega)$. Also, if $(X, \|\cdot\|_X)$ is a Banach space then, for $v\colon I \to X$, we have

$$\|v\|_{L_\infty(0,\tau;X)} \leqslant C(\tau)\Big(\|v(0)\|_X + \|v_t\|_{L_p(0,\tau;X)}\Big) \qquad \forall \tau \in \bar{I} \tag{34}$$

and for $1 \leqslant p \leqslant \infty$.

Now we can state the error estimate. The regularity requirements stated in this are given simply as they appear in the proof and in Lemma 3.3. We return to this point later.

THEOREM 3.4 (error estimate). *Let $\hat{h} \leqslant diam(\Omega)$ and $\hat{k} \leqslant T$ be positive constants and for $r \geqslant 1$ assume that $\breve{u} \in H^{r+1}(\Omega)$, $\breve{\boldsymbol{\sigma}} \in \boldsymbol{H}^r(\Omega)$,*

- *$u \in W^1_\infty(I; H^{r+1}(\Omega)) \cap H^2(I; H^2(\Omega)) \cap H^3(I; L_2(\Omega))$,*
- *$\boldsymbol{\sigma} \in L_\infty(I; \boldsymbol{L}_\infty(\Omega)) \cap W^1_\infty(I; \boldsymbol{H}^r(\Omega)) \cap H^3(I; \boldsymbol{L}_2(\Omega))$,*
- *$(\gamma(u)\boldsymbol{\sigma})_{tt} \in L_2(I; \boldsymbol{L}_2(\Omega))$;*

then for $\beta \geqslant (d-1)^{-1}$, $h \leqslant \hat{h}$, $\hat{h}^{(d-1)\beta-1}/\delta$ small enough (for $n = 1$ and 2), and $k \leqslant \hat{k}$, where $\hat{k}$ is small enough (for $n = 1$ only), we have a positive constant, C, independent of h and k such that

$$\|u(t_m) - u^h_m\|_0 + \|\boldsymbol{\sigma}(t_m) - \boldsymbol{\sigma}^h_m\|_0 + \left(k\sum_{i=1}^m \|\bar{u}_i - \bar{u}^h_i\|^2_{\mathcal{A}} + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}^h_i\|^2_0\right)^{1/2} \leqslant C(h^r + k^2)$$

for each $m = 1, \ldots, N$.

Proof. We average (19) between t_i and t_{i-1} and subtract it from (21), and do the same with (20) and (22). Adding the results of these then gives an error equation,

$$\begin{aligned}(\partial_t\chi_i, v) + (\partial_t\boldsymbol{\eta}_i, \boldsymbol{w}) + A(\bar{\chi}_i, v) &= (\Delta_i u, v) + (\Delta_i\boldsymbol{\sigma}, \boldsymbol{w}) + (\partial_t\xi_i, v) + (\partial_t\boldsymbol{\theta}_i, \boldsymbol{w}) + A(\bar{\xi}_i, v) \\ &\quad - \sum_E (K\bar{\boldsymbol{\eta}}_i, \nabla v)_E + \sum_E (K\bar{\boldsymbol{\theta}}_i, \nabla v)_E \\ &\quad + \sum_E (\mu\nabla\bar{\chi}_i, \boldsymbol{w})_E - \sum_E (\mu\nabla\bar{\xi}_i, \boldsymbol{w})_E \\ &\quad + \sum_{e\in\Gamma_h\cup\Gamma_D} \int_e \{K\bar{\boldsymbol{\eta}}_i \cdot \boldsymbol{n}_e\}[v] - \sum_{e\in\Gamma_h\cup\Gamma_D} \int_e \{K\bar{\boldsymbol{\theta}}_i \cdot \boldsymbol{n}_e\}[v] \\ &\quad + (\overline{\gamma(u)\boldsymbol{\sigma}(t_i)} - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}^h_i, \boldsymbol{w}) \qquad \forall v \in \mathcal{D}_r(\mathcal{E}_h) \text{ and } \forall \boldsymbol{w} \in \boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h).\end{aligned}$$

We now choose $v = \bar{\chi}_i$ and $\boldsymbol{w} = (K/\mu)\bar{\boldsymbol{\eta}}_i$, multiply by $2k$, and sum over $i = 1, \ldots, m \leqslant N$ to get

$$\begin{aligned}\|\chi_m\|^2_0 + \frac{K}{\mu}\|\boldsymbol{\eta}_m\|^2_0 + 2k\sum_{i=1}^m \|\bar{\chi}_i\|^2_{\mathcal{A}} + 2k\sum_{i=1}^m \frac{K}{\mu}(\gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\eta}}_i, \bar{\boldsymbol{\eta}}_i) \\ = \|\chi_0\|^2_0 + \frac{K}{\mu}\|\boldsymbol{\eta}_0\|^2_0 + 2k\sum_{i=1}^m (\Delta_i u, \bar{\chi}_i) + 2k\sum_{i=1}^m \frac{K}{\mu}(\Delta_i\boldsymbol{\sigma}, \bar{\boldsymbol{\eta}}_i)\end{aligned}$$

$$
\begin{aligned}
&+ 2k\sum_{i=1}^m(\partial_t\xi_i,\bar\chi_i) + 2k\sum_{i=1}^m A(\bar\xi_i,\bar\chi_i) + \frac{2Kk}{\mu}\sum_{i=1}^m(\partial_t\boldsymbol{\theta}_i,\bar{\boldsymbol{\eta}}_i)\\
&+ 2k\sum_{i=1}^m\sum_E(K\bar{\boldsymbol{\theta}}_i,\nabla\bar\chi_i)_E - 2k\sum_{i=1}^m\sum_E(K\nabla\bar\xi_i,\bar{\boldsymbol{\eta}}_i)_E\\
&+ 2k\sum_{i=1}^m(1-\kappa)\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e\{D\nabla\bar\chi_i\cdot\boldsymbol{n}_e\}[\bar\chi_i]\\
&+ 2k\sum_{i=1}^m\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e\{K\bar{\boldsymbol{\eta}}_i\cdot\boldsymbol{n}_e\}[\bar\chi_i] - 2k\sum_{i=1}^m\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e\{K\bar{\boldsymbol{\theta}}_i\cdot\boldsymbol{n}_e\}[\bar\chi_i]\\
&+ \frac{2Kk}{\mu}\sum_{i=1}^m(\overline{\gamma(u)\boldsymbol{\sigma}(t_i)} - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*,\bar{\boldsymbol{\eta}}_i),\\
&= T_1+\cdots+T_{13}.
\end{aligned}
$$

We now take each term in turn. By the $L_2(\Omega)$ projection we have $(\chi_0,v)=(\xi(0),v)$ for all $v\in\mathcal{D}_r(\mathcal{E}_h)$, which, from (30), results in

$$|T_1| = \|\chi_0\|_0^2 \leqslant \|\xi(0)\|_0^2 \leqslant Ch^{2r}\|\breve{u}\|_{r+1}^2.$$

Similarly, we have $(\boldsymbol{\eta}_0,\boldsymbol{w})=(\boldsymbol{\theta}(0),\boldsymbol{w})$ for all $\boldsymbol{w}\in\boldsymbol{\mathcal{D}}_{r-1}(\mathcal{E}_h)$ and, from (16), this gives

$$|T_2| = \|\boldsymbol{\eta}_0\|_0^2 \leqslant \|\boldsymbol{\theta}(0)\|_0^2 \leqslant Ch^{2r}\|\breve{\boldsymbol{\sigma}}\|_r^2.$$

For T_3 and T_4 we appeal to (27) from Lemma 3.1 and (18) to get

$$|T_3| \leqslant \frac{Ck}{\epsilon_3}\sum_{i=1}^m\|\Delta_i u\|_0^2 + \epsilon_3 k\sum_{i=1}^m\|\bar\chi_i\|_{\mathcal{A}}^2 \leqslant \frac{Ck^4}{\epsilon_3}\|u_{ttt}\|_{L_2(0,t_m;L_2(\Omega))}^2 + \epsilon_3 k\sum_{i=1}^m\|\bar\chi_i\|_{\mathcal{A}}^2,$$

and

$$
\begin{aligned}
|T_4| &\leqslant \frac{Kk}{\mu\gamma_G\epsilon_4}\sum_{i=1}^m\|\Delta_i\boldsymbol{\sigma}\|_0^2 + \epsilon_4 k\sum_{i=1}^m\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2,\\
&\leqslant \frac{Ck^4}{\epsilon_4}\|\boldsymbol{\sigma}_{ttt}\|_{L_2(0,t_m;\boldsymbol{L}_2(\Omega))}^2 + \epsilon_4 k\sum_{i=1}^m\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2.
\end{aligned}
$$

Using (18), (30), and (26) from Lemma 3.1, we have for T_5 that

$$
\begin{aligned}
|T_5| &\leqslant \frac{Ck}{\epsilon_5}\sum_{i=1}^m\left(\|\partial_t\xi_i - \xi_t(t_{i-1/2})\|_0^2 + \|\xi_t(t_{i-1/2})\|_0^2\right) + \epsilon_5 k\sum_{i=1}^m\|\bar\chi_i\|_{\mathcal{A}}^2,\\
&\leqslant \frac{Ck^4}{\epsilon_5}\|u_{ttt}\|_{L_2(0,t_m;H^2(\Omega))} + \frac{Ct_mh^{2r}}{\epsilon_5}\|u_t\|_{L_\infty(0,t_m;H^{r+1}(\Omega))}^2 + \epsilon_5 k\sum_{i=1}^m\|\bar\chi_i\|_{\mathcal{A}}^2,
\end{aligned}
$$

where we used (18) and (31) to get $\|\xi_{ttt}\|_0\leqslant C\|u_{ttt}\|_2$.

Now, $T_6=0$ from (28) and for T_7 we argue similarly as for T_5 and obtain

$$
\begin{aligned}
|T_7| &\leqslant \frac{Kk}{\mu\gamma_G\epsilon_7}\sum_{i=1}^m\|\partial_t\boldsymbol{\theta}_i\|_0^2 + \epsilon_7 k\sum_{i=1}^m\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2,\\
&\leqslant \frac{Ck^4}{\epsilon_7}\|\boldsymbol{\sigma}_{ttt}\|_{L_2(0,t_m;\boldsymbol{L}_2(\Omega))} + \frac{Ct_mh^{2r}}{\epsilon_7}\|\boldsymbol{\sigma}_t\|_{L_\infty(0,t_m;\boldsymbol{H}^r(\Omega))}^2 + \epsilon_7 k\sum_{i=1}^m\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2,
\end{aligned}
$$

where we used the estimate $\|\boldsymbol{\theta}_{ttt}\|_0 \leqslant C\|\boldsymbol{\sigma}_{ttt}\|_0$. For T_8,

$$|T_8| \leqslant \frac{Ck}{\epsilon_8 D^2}\sum_{i=1}^m \|\bar{\boldsymbol{\theta}}_i\|_0^2 + \epsilon_8 k\sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \leqslant \frac{Ct_m h^{2r}}{\epsilon_8}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,t_m;\boldsymbol{H}^r(\Omega))} + \epsilon_8 k\sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,$$

and for T_9,

$$\begin{aligned}|T_9| &\leqslant 2k\sum_{i=1}^m \frac{K}{D}\|\bar{\xi}_i\|_{\mathcal{A}}\|\bar{\boldsymbol{\eta}}_i\|_0 \leqslant \frac{k}{\epsilon_9}\sum_{i=1}^m \frac{\mu K}{\gamma_G D^2}\|\bar{\xi}_i\|_{\mathcal{A}}^2 + \epsilon_9 k\sum_{i=1}^m \frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2,\\ &\leqslant \frac{Ct_m h^{2r}}{\epsilon_9}\|u\|^2_{L_\infty(0,t_m;H^{r+1}(\Omega))} + \epsilon_9 k\sum_{i=1}^m \frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2.\end{aligned}$$

We now note that $T_{10} = 0$ if $\kappa = 1$ (the nonsymmetric scheme) and in general we have

$$\begin{aligned}|T_{10}| &\leqslant 2(1-\kappa)k\sum_{i=1}^m \sum_{e\in\Gamma_h\cup\Gamma_D}\left(\frac{|e|^\beta}{\delta}\right)^{1/2}\|\{D\nabla\bar{\chi}_i\cdot\boldsymbol{n}_e\}\|_{0,e}\left(\frac{\delta}{|e|^\beta}\right)^{1/2}\|[\bar{\chi}_i]\|_{0,e},\\ &\leqslant 2(1-\kappa)k\sum_{i=1}^m \frac{Ch^{(d-1)\beta/2-1/2}}{\delta^{1/2}}\|D^{1/2}\nabla\bar{\chi}_i\|_0\ J_0^{\delta,\beta}(\bar{\chi}_i,\bar{\chi}_i)^{1/2},\\ &\leqslant (1-\kappa)\epsilon_{10}k\sum_{i=1}^m \frac{Ch^{(d-1)\beta-1}}{\delta}\|\bar{\chi}_i\|_{\mathcal{A}}^2 + \frac{(1-\kappa)k}{\epsilon_{10}}\sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2.\end{aligned}$$

For T_{11} a similar argument produces

$$|T_{11}| \leqslant \epsilon_{11}k\sum_{i=1}^m \frac{Ch^{(d-1)\beta-1}}{\delta}\|\bar{\boldsymbol{\eta}}_i\|_0^2 + \frac{k}{\epsilon_{11}}\sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,$$

and, for T_{12},

$$\begin{aligned}|T_{12}| &\leqslant \epsilon_{12}k\sum_{i=1}^m \frac{Ch^{(d-1)\beta}}{\delta}\left(\sum_E \|\bar{\boldsymbol{\theta}}_i\|_{L_2(\partial E)}\right)^2 + \frac{k}{\epsilon_{12}}\sum_{i=1}^m J_0^{\delta,\beta}(\bar{\chi}_i,\bar{\chi}_i),\\ &\leqslant \frac{Ct_m\epsilon_{12}h^{2r-1+(d-1)\beta}}{\delta}\|\boldsymbol{\sigma}\|^2_{L_\infty(0,t_m;\boldsymbol{H}^r(\Omega))} + \frac{k}{\epsilon_{12}}\sum_{i=1}^m J_0^{\delta,\beta}(\bar{\chi}_i,\bar{\chi}_i).\end{aligned}$$

Setting $\epsilon_{10} = 2$ and choosing

$$\epsilon_3+\epsilon_5+\epsilon_8+\frac{1}{\epsilon_{12}} = \frac{1}{4}, \qquad \epsilon_4+\epsilon_7+\epsilon_9 = 1, \qquad \text{and} \qquad \epsilon_{11} = 4,$$

we then assemble these estimates and obtain

$$\begin{aligned}\|\chi_m\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\eta}_m\|_0^2 &+ \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right)k\sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2\\ &+ \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K\gamma_G}\right)k\sum_{i=1}^m \frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2\\ &\leqslant C(h^{2r}+k^4) + \frac{2kK}{\mu}\sum_{i=1}^m \left|\left(\overline{\gamma(u)\boldsymbol{\sigma}(t_i)} - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i\right)\right|,\end{aligned}$$

where we recalled that $\beta \geqslant (d-1)^{-1}$. Now we make several appeals to Lemma 3.3. First, when $n = 1$ we have, for $k \leq \hat{k}$, that

$$\left(1 - \frac{C\hat{k}}{\epsilon}\right)\|\chi_m\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\eta}_m\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right)k\sum_{i=1}^{m}\|\bar{\chi}_i\|_{\mathcal{A}}^2 + \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K\gamma_G} - \epsilon\right)k\sum_{i=1}^{m}\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \leqslant C(h^{2r}+k^4) + \frac{Ck}{\epsilon}\sum_{i=0}^{m-1}\|\chi_i\|_0^2.$$

Choosing $\epsilon = 1/2$, $\hat{k}$, and $\hat{h}^{(d-1)\beta-1}/\delta$ small enough, an application of Gronwall's lemma then results in

$$\|\chi_m\|_0^2 + \|\boldsymbol{\eta}_m\|_0^2 + k\sum_{i=1}^{m}\|\bar{\chi}_i\|_{\mathcal{A}}^2 + k\sum_{i=1}^{m}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \leqslant C(h^{2r}+k^4).$$

Second, for the linearized scheme where $n = 2$, we have by Lemma 3.3 for $m = 1$ and with $\epsilon = (2\gamma_G k)^{-1}$ that

$$\|\chi_1\|_0^2 + \frac{K}{2\mu}\|\boldsymbol{\eta}_1\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right)k\|\bar{\chi}_1\|_{\mathcal{A}}^2 + \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K\gamma_G}\right)k\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_1\|_0^2 \leqslant C(h^{2r}+k^4) + Ck^2\|\chi_0\|_0^2 + C\|\boldsymbol{\eta}_0\|_0^2.$$

Now use the estimates given above for T_1 and T_2 and again select $\hat{h}^{(d-1)\beta-1}/\delta$ small enough to get

$$\|\chi_1\|_0^2 + \|\boldsymbol{\eta}_1\|_0^2 + k\|\bar{\chi}_1\|_{\mathcal{A}}^2 + k\|\bar{\boldsymbol{\eta}}_1\|_0^2 \leqslant C(h^{2r}+k^4).$$

On the other hand, for $m > 1$ we estimate the first term in the sum (corresponding to $i = 1$) in T_{13} by choosing $\epsilon = 1/k$ in Lemma 3.3 and then use the estimates just obtained. For the remaining terms we choose $\epsilon = 1/2$. With empty sums set to zero, we then have for $m = 2, 3, 4, \ldots$ that

$$\|\chi_m\|_0^2 + \frac{K}{\mu}\|\boldsymbol{\eta}_m\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right)k\sum_{i=1}^{m}\|\bar{\chi}_i\|_{\mathcal{A}}^2 + \left(\frac{1}{2} - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K\gamma_G}\right)k\sum_{i=1}^{m}\frac{K\gamma_G}{\mu}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \leqslant C(h^{2r}+k^4) + Ck\sum_{i=2}^{m-1}\|\chi_i\|_0^2,$$

by the same estimates for the initial conditions as used previously. Once again, we choose $\hat{h}^{(d-1)\beta-1}/\delta$ small enough and use Gronwall's lemma to arrive at

$$\|\chi_m\|_0^2 + \|\boldsymbol{\eta}_m\|_0^2 + k\sum_{i=1}^{m}\|\bar{\chi}_i\|_{\mathcal{A}}^2 + k\sum_{i=1}^{m}\|\bar{\boldsymbol{\eta}}_i\|_0^2 \leqslant C(h^{2r}+k^4).$$

We now see that this inequality holds for all $m \in \{1, \ldots, N\}$ in both of the cases

$n = 1$ and $n = 2$. By the triangle inequality we then have

$$\begin{aligned}\|u(t_m) - u_m^h\|_0 + \|\boldsymbol{\sigma}(t_m) - \boldsymbol{\sigma}_i^h\|_0 + \left(k\sum_{i=1}^m \|\bar{u}(t_i) - \bar{u}_i^h\|_{\mathcal{A}}^2\right)^{1/2} \\ + \left(k\sum_{i=1}^m \|\bar{\boldsymbol{\sigma}}(t_i) - \bar{\boldsymbol{\sigma}}_i^h\|_0^2\right)^{1/2} \\ \leqslant \|\xi(t_m)\|_0 + \|\boldsymbol{\theta}(t_m)\|_0 + \left(k\sum_{i=1}^m \|\bar{\xi}(t_i)\|_{\mathcal{A}}^2\right)^{1/2} + \left(k\sum_{i=1}^m \|\bar{\boldsymbol{\theta}}(t_i)\|_0^2\right)^{1/2} \\ + \|\chi(t_m)\|_0 + \|\boldsymbol{\eta}(t_m)\|_0 + \left(k\sum_{i=1}^m \|\bar{\chi}(t_i)\|_{\mathcal{A}}^2\right)^{1/2} + \left(k\sum_{i=1}^m \|\bar{\boldsymbol{\eta}}(t_i)\|_0^2\right)^{1/2},\end{aligned}$$

and our estimates, along with (16) and (30) and the fact that $(a^2 + b^2)^{1/2} \leqslant a + b$ for $a, b \geqslant 0$, then complete the proof. □

Note that due to the much larger number of terms involved this proof used Gronwall's inequality, unlike the proof of Theorem 2.2. It is possible that more careful estimation could remove the need for an exponentially large "Gronwall constant" in the error estimate, but we leave this as a problem for another time.

If we replace the $\mathcal{D}_r(\mathcal{E}_h)$-approximation of u by a standard conforming piecewise polynomial finite element space containing the essential boundary condition, then the DG FEM schemes presented above reduce to a standard continuous Galerkin (CG) FEM. An error estimate of the form presented in Theorem 3.4 then continues to hold (as a special case).

COROLLARY 3.5. *For a CG finite element approximation of the problem we also have*

$$\|u(t_m) - u_m^h\|_0 + \|\boldsymbol{\sigma}(t_m) - \boldsymbol{\sigma}_m^h\|_0 + \left(k\sum_{i=1}^m \|\bar{u}_i - \bar{u}_i^h\|_{\mathcal{A}}^2 + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^h\|_0^2\right)^{1/2} \leqslant C(h^r + k^2)$$

for each $m = 1, \ldots, N$.

REMARK 3.6. *If* $(u, \boldsymbol{\sigma})$ *is a solution of* (12), (13), *then we could use Theorems* 1.1 *and* 3.4 *to show that*

$$\|u_m^h\|_0^2 + \|\boldsymbol{\sigma}_m^h\|_0^2 \leq C(u).$$

This would follow from the triangle inequality and is the closest we can get to a stability estimate. However, to get "data" on the right-hand side we would need stability estimates on higher derivatives of the exact solutions.

Theorem 3.4 naturally contains some regularity assumptions on both u and $\boldsymbol{\sigma}$. Since, via (10), we can replace the system (1) and (2) by the single (11) we can expect that the regularity of σ can be tied into that of u. In this direction, for the case of piecewise linear spatial approximation ($r = 1$), we have the following claim (see [15] for details).

PROPOSITION 3.7. *For* $r = 1$ *the regularity requirements of Theorem* 3.4 *can be replaced by*

$$\begin{aligned}u \in H^2(I; H^2(\Omega)) \cap L_1(I; W_\infty^1(\Omega)) \cap L_\infty(I; W_4^1(\Omega)) \cap W_8^1(I; L_8(\Omega)) \\ \cap W_4^2(I; L_4(\Omega)) \cap H^3(I; L_2(\Omega)) \cap H^1(I; W_4^1(\Omega))\end{aligned}$$

and $\breve{\boldsymbol{\sigma}} \in \boldsymbol{L}_\infty(\Omega) \cap \boldsymbol{H}^1(\Omega)$.

TABLE 1
Tabulated errors for $N = 2$, $\delta = 10^2$, and $\beta = 3$.

	$\kappa = -1$		$\kappa = 1$	
M	$\mathcal{E}$	EOC	$\mathcal{E}$	EOC
1	2.032013		2.026638	
2	1.837297	0.1453	1.837534	0.1413
4	1.532320	0.2619	1.532354	0.2620
8	0.863881	0.8268	0.863874	0.8269
16	0.447608	0.9486	0.447608	0.9486
32	0.225955	0.9862	0.225955	0.9862

TABLE 2
Tabulated errors for $M = 8$, $\delta = 10^4$, and $\beta = 2$.

	$\kappa = -1$		$\kappa = 1$	
N	$\mathcal{E}$	EOC	$\mathcal{E}$	EOC
1	8.845824×10^{-2}		8.845825×10^{-2}	
2	2.561600×10^{-2}	1.7879	2.561600×10^{-2}	1.7879
4	6.598674×10^{-3}	1.9568	6.598677×10^{-3}	1.9568
8	1.660746×10^{-3}	1.9903	1.660747×10^{-3}	1.9903
16	4.166057×10^{-4}	1.9951	4.166061×10^{-4}	1.9951
32	1.049564×10^{-4}	1.9889	1.049566×10^{-4}	1.9889
64	2.707173×10^{-5}	1.9549	2.707180×10^{-5}	1.9549

4. Numerical experiments. We anticipate that the linearized scheme is the one that is of most interest and so first quote from [4] just a few numerical results to illustrate Theorem 3.4 in the case $r = 1$. The data common to these first results are $D = K = \mu = 1$, $\gamma_R = 10$, $\gamma_G = 0.1$, $\Delta = 0.1$, $\Omega = (0,1)^2$, and $I = (0,T)$ for $T = 1$, and in each case the loads and boundary conditions are designed so that the problem has a known exact solution. (To achieve this we added a function $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{x},t)$ to the right of (2).) The resulting errors,

$$\mathcal{E} := k \sum_{i=1}^{N} \left(\|\bar{u}_i - \bar{u}_i^h\|_{\mathcal{A}}^2, + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^h\|_0^2 \right)^{1/2},$$

are tabulated along with the estimated order of convergence (EOC). In the tables, M denotes a uniform $M \times M$ space mesh and N is the number of time intervals.

Table 1 shows results for the solutions

$$u(\boldsymbol{x},t) = t \sin(2\pi x) \sin(2\pi y), \qquad \boldsymbol{\sigma}(\boldsymbol{x},t) = t \begin{pmatrix} \sin(2\pi x) \\ \cos(2\pi y) \end{pmatrix}$$

in the case $\Gamma_D = \{x = 0 \text{ or } y = 0\}$ when $u_{RG} = 0.5$. In this case there is no time discretization error and we observe $O(h)$ convergence.

On the other hand, for the solutions

$$u(\boldsymbol{x},t) = t^3 x, \qquad \boldsymbol{\sigma}(\boldsymbol{x},t) = t^2 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

there is no space discretization error and for $\Gamma_D = \{x = 0\}$ with $u_{RG} = 0.5$ we observe, in Table 2, $O(k^2)$ convergence.

Cohen, White, and Witelski's model [6] produces solutions which exhibit very sharp changes in u, and these fronts become steeper as time advances. It seems that the near-discontinuity in u is driven by the fact that their scalar stress equation is

$$\sigma_t + \gamma(u)\sigma = \mu u,$$

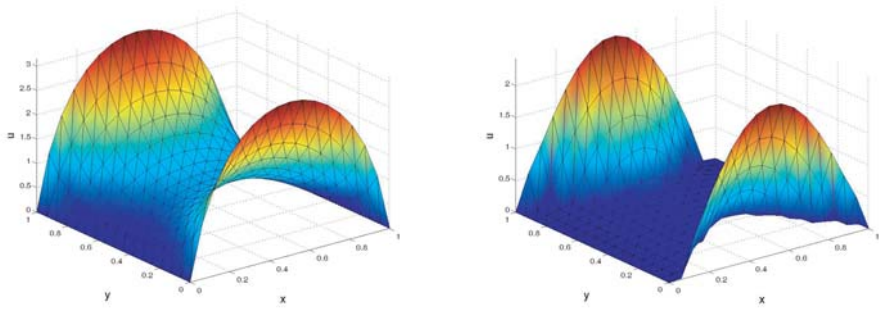


FIG. 1. *Computed surfaces showing* u *at* $t = 10$ *for the data* $\Omega := (0,1)^2$, $f = 0$, $g = 1$, $\Gamma_N := \{y = 0 \text{ or } y = 1\}$, $\breve{u} = -0.5x(x-1)$, $\breve{\boldsymbol{\sigma}} = \mathbf{0}$, $D = 10^{-1}$, $K = 10^{-4}$, $\mu = 10^4$, $\Delta = 10^{-3}$, *and* $u_{RG} = 0.5$. *The figure on the left corresponds to* $\gamma_R = \gamma_G = 5000$ *and the one on the right to* $\gamma_R = 10^4$ *and* $\gamma_G = 10^{-3}$.

whereas ours is a vector equation given by (2) and has the gradient of u on the right. Because of this, solutions to our model exhibit sharp changes in ∇u rather than u itself, and we illustrate this in Figure 1 (16×16 elements, 50 time steps, $\beta = 2$, $\delta = 10^2$, $\kappa = -1$).

The surface plot on the left corresponds to linear non-Fickian behavior where we choose $\gamma_G = \gamma_R$ so as to remove the nonlinear term in (5). The figure on the right shows the effect of the nonlinearity when $\gamma_R \gg \gamma_G$, and we can see steep changes in ∇u.

5. Conclusion. The numerical experiments support the error estimate in Theorem 3.4 and so we conclude that the linearization derived from the extrapolation is an effective method of approximating the solution to this type of problem. Also, on examining the estimates in Lemma 3.3, we see that the linearized scheme does not require any additional regularity assumptions. Hence, we conclude that it should always be preferred over the nonlinear scheme.

As we mentioned earlier in section 1, our model is a simplification of the original model proposed in [6]. Nonetheless, preliminary numerical experiments (not included here) with the CG FEM indicate that it is capable of capturing the same basic phenomena of steep traveling fronts. An error analysis for the model in [6] is currently being undertaken and, when this is complete, we expect to give more extensive numerical demonstrations for both models.

On a closing note, the problem we have studied is a generalization of a parabolic analogue to the dynamic solids problem considered in [17] to the case of nonlinear relaxation time. It is an ongoing project to extend our results to the dynamic case and to other types of nonlinearities (see, for example, the nonlinear relaxation time discussed in [22]).

REFERENCES

[1] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, 2nd ed., Pure and Applied Mathematics 140, Academic Press, New York, 2003.

[2] H. AMANN, *Global existence for a class of highly degenerate parabolic systems*, Japan J. Indust. Appl. Math., 8 (1991), pp. 143–151.

[3] H. T. Banks, G. A. Pintér, and L. K. Potter, *Existence of Unique Weak Solutions to a Dynamical System for Nonlinear Elastomers with Hysteresis*, Tech. Report CRSC–TR98–43, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, www. ncsu.edu/crsc/reports.htm/ (1998).

[4] R. Bultmann and S. Shaw, *Finite Element Discretisation of a Problem in Nonlinear Non-Fickian Viscoelastic Diffusion Using a Discontinuous Galerkin Method in Space*, Technical Report 05/5, BICOM, Brunel University, Uxbridge, England, www.brunel.ac.uk/bicom (2005).

[5] J. R. Cannon and Y. Lin, *A priori L^2 error estimates for finite-element methods for nonlinear diffusion equations with memory*, SIAM J. Numer. Anal., 27 (1990), pp. 595–607.

[6] D. S. Cohen, A. B. White, Jr., and T. P. Witelski, *Shock formation in a multidimensional viscoelastic diffusive system*, SIAM J. Appl. Math., 55 (1995), pp. 348–368.

[7] J. Douglas, R. E. Ewing, and M. F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media*, RAIRO Anal. Numér., 17 (1983), pp. 249–265.

[8] J. D. Ferry, *Viscoelastic Properties of Polymers*, John Wiley and Sons, New York, 1970.

[9] V. Girault, B. Riviere, and M. Wheeler, *A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems*, Math. Comput., 74 (2005), pp. 53–84.

[10] A. R. Johnson, *Modeling viscoelastic materials using internal variables*, The Shock and Vibration Digest, 31 (1999), pp. 91–100.

[11] A. R. Johnson, A. Tessler, and M. Dambach, *Dynamics of thick viscoelastic beams*, Journal of Engineering Materials and Technology, 119 (1997), pp. 273–278.

[12] Ch. Lubich, I. H. Sloan, and V. Thomée, *Nonsmooth data error estimates for approximations of an evolution equation with a positive-type memory term*, Math. Comp., 65 (1996), pp. 1–17.

[13] W. Mclean and V. Thomée, *Numerical solution of an evolution equation with a positive-type memory term*, J. Austral. Math. Soc. Ser. B, 35 (1993), pp. 23–70.

[14] J. W. Nunziato, *On heat conduction in materials with memory*, Quart. Appl. Math., 29 (1971), pp. 187–204.

[15] B. Rivière and S. Shaw, *Discontinuous Galerkin Finite Element Approximation of Nonlinear Non-Fickian Diffusion in Viscoelastic Polymers*, Technical Report 04/3, BICOM, Brunel University, Uxbridge, England, www.brunel.ac.uk/bicom (2004).

[16] B. Rivière, S. Shaw, M. F. Wheeler, and J. R. Whiteman, *Discontinuous Galerkin finite element methods for linear elasticity and quasistatic linear viscoelasticity*, Numer. Math., 95 (2003), pp. 347–376.

[17] B. Rivière, S. Shaw, and J. R. Whiteman, *Discontinuous Galerkin finite element methods for dynamic linear solid viscoelasticity problems*, Numer. Methods Partial Differential Equations, to appear (see also Report 05/7, www.brunel.ac.uk/bicom).

[18] B. Rivière and M. F. Wheeler, *A discontinuous Galerkin method applied to nonlinear parabolic equations*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 1999, pp. 231–244.

[19] B. Rivière and M. F. Wheeler, *Discontinuous Galerkin methods for flow and transport problems in porous media*, Comm. Numer. Methods Engrg., 18 (2002), pp. 63–68.

[20] B. Rivière, M. F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part* I, Comput. Geosci., 3 (1999), pp. 337–360.

[21] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.

[22] S. Shaw, M. K. Warby, and J. R. Whiteman, *Numerical techniques for problems of quasistatic and dynamic viscoelasticity*, in The Mathematics of Finite Elements and Applications (Uxbridge, 1993), J. R. Whiteman, ed., Wiley, Chichester, UK, 1994, pp. 45–68.

[23] I. H. Sloan and V. Thomée, *Time discretization of an integro-differential equation of parabolic type*, SIAM J. Numer. Anal., 23 (1986), pp. 1052–1061.

[24] N. Thomas and A. H. Windle, *Transport of methanol in poly(methyl-methacrylate)*, Polymer, 19 (1978), pp. 255–265.

[25] N. L. Thomas and A. H. Windle, *A theory of Case* II *diffusion*, Polymer, 23 (1982), pp. 529–542.

[26] V. Thomée and L. B. Wahlbin, *Long-time numerical solution of a parabolic equation with memory*, Math. Comp., 62 (1994), pp. 477–496.

SIAM J. NUMER. ANAL.
Vol. 44, No. 6, pp. 2671–2698

AN $H^1(\mathcal{P}^h)$-COERCIVE DISCONTINUOUS GALERKIN FORMULATION FOR THE POISSON PROBLEM: 1D ANALYSIS*

K. G. VAN DER ZEE†, E. H. VAN BRUMMELEN†, AND R. DE BORST†

Abstract. Coercivity of the bilinear form in a continuum variational problem is a fundamental property for finite-element discretizations: By the classical Lax–Milgram theorem, any conforming discretization of a coercive variational problem is stable; i.e., discrete approximations are well-posed and possess unique solutions, irrespective of the specifics of the underlying approximation space. Based on the prototypical one-dimensional Poisson problem, we establish in this work that most concurrent discontinuous Galerkin formulations for second-order elliptic problems represent instances of a generic conventional formulation and that this generic formulation is noncoercive. Consequently, all conventional discontinuous Galerkin formulations are a fortiori noncoercive, and typically their well-posedness is contingent on approximation-space-dependent stabilization parameters. Moreover, we present a new symmetric nonconventional discontinuous Galerkin formulation based on element Green's functions and the data local to the edges. We show that the new discontinuous Galerkin formulation is coercive on the broken Sobolev space $H^1(\mathcal{P}^h)$, viz., the space of functions that are elementwise in the H^1 Sobolev space. The coercivity of the new formulation is supported by calculations of discrete inf-sup constants, and numerical results are presented to illustrate the optimal convergence behavior in the energy-norm and in the $L_2(\Omega)$-norm.

Key words. finite element method, discontinuous Galerkin, elliptic problems, coercivity

AMS subject classifications. 65N30, 65N12

DOI. 10.1137/05063057X

1. Introduction.

The recent renewal of interest in discontinuous Galerkin (DG) methods for second-order elliptic boundary value problems can be attributed to twofold reasons. First, DG methods provide robust finite-element discretizations for hyperbolic conservation laws, as the interelement discontinuities enable an extension of Godunov's method for finite-volume methods. However, to extend these techniques to singularly perturbed elliptic problems, an appropriate treatment for the elliptic part of the operator is required. Second, the absence of interelement-continuity constraints renders DG methods ideally suited for hp adaptivity, e.g., based on a posteriori error estimation; see, for instance, [1, 8]. A comprehensive overview of the historical development of DG methods is provided in [7].

A framework for analyzing DG formulations for elliptic problems has recently been erected in [2]. Although the analysis in [2] clarifies basic properties of the different formulations, it does not seem to warrant a clear preference. The literature on DG methods for elliptic problems is dominated by formulations that possess edge terms composed of linear combinations of the jumps and averages of the test and trial functions and their normal derivatives. That is, denoting by u and v the test and trial functions, and by $[\![\cdot]\!]$ and $\{\cdot\}$ the jump and average of $(\cdot)$ at an interelement edge, these formulations contain terms conforming to $\{\partial_n u\}[\![v]\!]$, $[\![u]\!][\![v]\!]$, $[\![\partial_n u]\!][\![\partial_n v]\!]$, etc., where ∂_n represents the normal derivative. We refer to such formulations as

*Received by the editors May 3, 2005; accepted for publication (in revised form) August 22, 2006; published electronically December 21, 2006.

http://www.siam.org/journals/sinum/44-6/63057.html

†Engineering Mechanics, Faculty of Aerospace Engineering, Delft University of Technology, P.O. Box 5058, 2600 GB Delft, The Netherlands (K.G.vanderZee@TUDelft.nl, E.H.vanBrummelen@TUDelft.nl, R.deBorst@TUDelft.nl). The work of the second author was partially supported by NWO/VENI grant 639.031.305.

conventional DG formulations and to the corresponding edge terms as *conventional edge terms*. Symmetric examples of such formulations are the *global element method* (GEM; see [10, 12]) and the *interior penalty* DG formulation (IPDG; cf. [2, 10, 12]). Nonsymmetric examples are the celebrated *Baumann and Oden* DG formulation (BODG [3]), the *stabilized* DG formulation (SDG [14]), the *nonsymmetric interior penalty* DG formulation (NIPDG [13]), and the family of formulations considered by Larson and Niklasson (LNDG [9]).

The essential deficiency of conventional DG formulations is that their bilinear form is not *strongly coercive* (and simultaneously continuous) on a continuum (infinite-dimensional) broken space, in contrast to the bilinear form in the classical continuous Galerkin (CG) formulation. For conciseness, we say that these methods are *noncoercive*. In particular, this implies that finite-element approximations can be ill-posed, despite well-posedness of the underlying continuum problem. Furthermore, a sequence of nested stable approximations need not converge monotonously, as the constants in the error-estimates are approximation-space-dependent, and cannot be bounded uniformly. Conventional DG formulations can be coercive on *discrete* approximation spaces. However, this generally requires stability parameters which increase unboundedly as the approximation space is refined. For example, for broken polynomial spaces the stability parameters are typically proportional to a monomial of the polynomial degree. Moreover, conventional DG formulations are in general subject to the assumption that the solution resides in $H^2(\Omega)$, whereas a formulation allowing solutions in $H^1(\Omega)$ would be more natural from the classical CG formulation perspective.[1]

Nonsymmetric conventional DG formulations can be well-posed without stability parameters. However, such formulations derive their well-posedness from *weak coercivity*. Moreover, for nonsymmetric formulations the error converges suboptimally in the $L_2(\Omega)$-norm for even-degree broken polynomial spaces. It has been conjectured that this behavior emanates from the nonsymmetry of the formulation; see [3, 9, 10].

Alternatively, DG formulations can be constructed by introducing *nonconventional edge terms*. We remark that the support of such terms is not necessarily restricted to the edges. Examples of such DG formulations are the *lift-operator-based* schemes in [2]. These include, among others, the *Bassi and Rebay* DG formulation (BRDG [4, 5]) and the *local* DG formulation (LDG [6]). However, lift-operators are explicitly defined using a discrete (finite-element) space. As a consequence, the continuum formulation with lift-operators, although consistent at a discrete level, is inconsistent at the continuum level. Therefore, for each approximation space, the edge-traces need to be lifted accordingly and the extension to a consistent continuum formulation is nonobvious.

A recent example of another nonconventional formulation is the discontinuous finite-element formulation based on second-order derivatives in [15]. This formulation resembles a least-squares form. However, it is based on second-order derivatives, thereby implicitly restricting the admissible functions to $H^2(\Omega)$ and, moreover, it is unknown if the bilinear form is simultaneously coercive and continuous.

In this paper, we first establish on the basis of the prototypical one-dimensional Poisson problem that a conventional DG formulation with a coercive bilinear form is *nonexistent*. We then present a new nonconventional symmetric DG formulation based on *element Green's functions* and the data local to the edges. The essential advantage of our new DG formulation is that it is *coercive* on the (infinite-dimensional)

[1] More precisely, only $H^{3/2+\epsilon}(\Omega)$ regularity is required. This ensures that the edge terms in conventional DG formulations are well defined.

broken Sobolev space $H^1(\mathcal{P}^{\mathsf{h}})$, the space of functions that are elementwise in the H^1 Sobolev space. On account of its coercivity, approximations of the new formulation inherit their well-posedness from the continuum formulation; i.e., well-posedness of the approximation problem is ensured for any approximation space and, in particular, for the usual broken polynomial spaces. Furthermore, optimal error estimates hold with constants that can be bounded uniformly independent of the specifics of the approximation space. Finally, we demonstrate that the new DG formulation is equivalent with the classical CG formulation, thus allowing solutions in $H^1(\Omega)$.

The contents of this paper are arranged as follows: section 2 presents the elliptic model problem, viz., the Poisson problem. Furthermore, mathematical preliminaries for DG formulations of the Poisson problem are given. Section 3 reviews elementary existence and uniqueness theorems for linear variational problems, to establish the differences between coercivity and weak coercivity, and to furnish the basis for our analysis in the ensuing sections. In section 4 we present the generic conventional DG formulation, and we prove its noncoerciveness. In section 5 we introduce the new DG formulation and we demonstrate its coercivity. Furthermore, we establish its equivalence with the classical CG formulation. Numerical results are presented in section 6. The coercivity of the new formulation is supported by calculations of discrete inf-sup constants. Moreover, the convergence behavior in the energy-norm and in the $L_2(\Omega)$-norm is investigated. Finally, section 7 contains concluding remarks.

2. Problem statement. In this work, we shall restrict ourselves to the simplest prototypical model problem for second-order elliptic boundary value problems, viz., the linear one-dimensional *Poisson* problem.

2.1. Poisson problem. Let $\Omega \subset \mathbb{R}$ be a bounded open interval. Its two-point boundary $\partial\Omega$ consists of two disjoint parts, $\Gamma_{\mathcal{D}}$ (nonempty) and $\Gamma_{\mathcal{N}}$ (possibly empty) on which Dirichlet and Neumann boundary conditions are imposed, respectively. The unit normal n at the boundary $\partial\Omega$ is defined to be outward with respect to the interval Ω.

Within this one-dimensional setting, we formulate the Poisson problem: Given an arbitrary $\bar{u} \in H^1(\Omega)$ with $\bar{u} = g_{\mathcal{D}}$ on $\Gamma_{\mathcal{D}}$,

$$\boxed{\begin{aligned}&\text{Find } u = \bar{u} + u_0 \in \bar{u} + H^1_{0,\mathcal{D}}(\Omega) : \\ &\qquad\qquad \mathscr{B}_{\mathrm{c}}(u_0, v) = \mathscr{L}_{\mathrm{c}}(v) \qquad \forall v \in H^1_{0,\mathcal{D}}(\Omega)\ ,\end{aligned}} \tag{2.1}$$

where the bilinear form $\mathscr{B}_{\mathrm{c}} : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ and the linear functional $\mathscr{L}_{\mathrm{c}} : H^1(\Omega) \to \mathbb{R}$ are defined as

$$\mathscr{B}_{\mathrm{c}}(u, v) := \int_\Omega \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x\ , \tag{2.2a}$$

$$\mathscr{L}_{\mathrm{c}}(v) := \int_\Omega fv\,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{N}}} \big(g_{\mathcal{N}} v\big)_e - \mathscr{B}_{\mathrm{c}}(\bar{u}, v)\ . \tag{2.2b}$$

We define $H^1_{0,\mathcal{D}}(\Omega)$ to be the subspace of functions in the Sobolev space $H^1(\Omega)$ which vanish on the Dirichlet boundary $\Gamma_{\mathcal{D}}$, i.e.,

$$H^1_{0,\mathcal{D}}(\Omega) := \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_{\mathcal{D}}\}\ .$$

For $f \in L_2(\Omega)$, Problem (2.1) possesses a unique solution $u \in H^2(\Omega)$ which, moreover,

uniquely solves the boundary value problem

$$\boxed{\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} &= f && \text{in } \Omega\ , \qquad (2.3a)\\ u &= g_{\mathcal{D}} && \text{on } \Gamma_{\mathcal{D}}\ , \qquad (2.3b)\\ \partial_n u &= g_{\mathcal{N}} && \text{on } \Gamma_{\mathcal{N}}\ , \qquad (2.3c) \end{aligned}}$$

where $\partial_n(\cdot)$ denotes the normal derivative $\frac{\mathrm{d}}{\mathrm{d}n}(\cdot)$.

The variational problem (2.1) constitutes the classical CG formulation of the Poisson problem. It is *well-posed* (stable); i.e., there *exists* a *unique* solution $u \in H^1(\Omega)$ which, moreover, depends continuously on the auxiliary data. The well-posedness follows from the classical Lax–Milgram theorem on account of *coercivity* of the bilinear functional $\mathscr{B}_{\mathrm{c}}(\cdot,\cdot)$; see section 3. A *conforming approximation* to the *continuum* problem (2.1) is obtained by replacing $H^1_{0,\mathcal{D}}(\Omega)$ by a closed, generally finite-dimensional, *sub*space $\widehat{H}^1_{0,\mathcal{D}}(\Omega) \subset H^1_{0,\mathcal{D}}(\Omega)$. The corresponding approximate solution $\widehat{u} \in \bar{u} + \widehat{H}^1_{0,\mathcal{D}}(\Omega)$ can be extracted by solving the following *approximate* problem:

$$\boxed{\begin{aligned} &\text{Find } \widehat{u} = \bar{u} + \widehat{u}_0 \in \bar{u} + \widehat{H}^1_{0,\mathcal{D}}(\Omega) : \\ &\qquad\qquad \mathscr{B}_{\mathrm{c}}(\widehat{u}_0, v) = \mathscr{L}_{\mathrm{c}}(v) \qquad \forall v \in \widehat{H}^1_{0,\mathcal{D}}(\Omega)\ . \end{aligned}} \qquad (2.4)$$

As the coercivity of the underlying continuum problem transfers to the approximate problem, the approximate problem is automatically well-posed irrespective of the specifics of the approximation space $\widehat{H}^1_{0,\mathcal{D}}(\Omega)$. This is a particularly favorable property which enables, for example, subsequent stable approximations in an hp-adaptive finite element procedure. We emphasize that coercivity is generally lost in a DG formulation.

2.2. Broken Sobolev spaces. To facilitate the ensuing consideration of DG formulations, we introduce a *finite-element partition*. Let $\mathcal{P}^{\mathsf{h}} := \mathcal{P}^{\mathsf{h}}(\Omega)$ denote such a partition of the interval Ω; i.e., $\mathcal{P}^{\mathsf{h}}$ is a finite collection of open nonoverlapping subintervals (*elements*) K, such that

$$\Omega = \mathrm{int}\left(\bigcup_{K \in \mathcal{P}^{\mathsf{h}}} \overline{K} \right) .$$

The mesh parameter h associated with $\mathcal{P}^{\mathsf{h}}$ is defined as

$$\mathsf{h} := \max_{K \in \mathcal{P}^{\mathsf{h}}} h_K\ ,$$

where h_K is the length of element K. The set of all (element) *edges*, $\Gamma := \Gamma(\mathcal{P}^{\mathsf{h}})$, can be divided into complementary subsets:

$$\Gamma = \Gamma_{\mathcal{D}} \cup \Gamma_{\mathcal{N}} \cup \Gamma_{\mathcal{I}}\ ,$$

where $\Gamma_{\mathcal{I}} := \Gamma_{\mathcal{I}}(\mathcal{P}^{\mathsf{h}})$ is the set of *interior* edges. We define a unit normal n_e at each edge $e \in \Gamma$. This normal coincides with the unit outward normal of Ω for *boundary* edges $e \in \partial\Omega = \Gamma_{\mathcal{D}} \cup \Gamma_{\mathcal{N}}$ and we set $n_e := -1$ for interior edges $e \in \Gamma_{\mathcal{I}}$. For example, if e is an interior edge then we have $(\partial_n u)_e = \frac{\mathrm{d}u}{\mathrm{d}n}\big|_e = -\frac{\mathrm{d}u}{\mathrm{d}x}\big|_e$. We will further denote by $\mathcal{K}_e$ the set of elements sharing edge $e \in \Gamma$, that is,

$$\mathcal{K}_e := \left\{ K \in \mathcal{P}^{\mathsf{h}} : \partial K \cap e = e \right\} .$$

Note that for boundary and interior edges e, the set $\mathcal{K}_e$ contains one element and two elements, respectively.

The functional setting of DG formulations is provided by the so-called (partition $\mathcal{P}^{\mathsf{h}}$ dependent) *broken Sobolev spaces* $H^m(\mathcal{P}^{\mathsf{h}})$ [10]. For any positive integer m, the broken Sobolev space $H^m(\mathcal{P}^{\mathsf{h}})$ is defined as

$$H^m(\mathcal{P}^{\mathsf{h}}) := \{v \in L_2(\Omega) : v\big|_K \in H^m(K)\ \forall K \in \mathcal{P}^{\mathsf{h}}\}\ ; \tag{2.5}$$

in other words, $H^m(\mathcal{P}^{\mathsf{h}})$ consists of functions for which the restriction to each element $K \in \mathcal{P}^{\mathsf{h}}$ is in $H^m(K)$. Equipped with the broken inner product

$$(u,v)_{H^m(\mathcal{P}^{\mathsf{h}})} := \sum_{K\in\mathcal{P}^{\mathsf{h}}} (u,v)_{H^m(K)}\ ,$$

$H^m(\mathcal{P}^{\mathsf{h}})$ is a Hilbert space. The corresponding norm will be denoted $\|\cdot\|_{H^m(\mathcal{P}^{\mathsf{h}})}$. Note that functions in a broken Sobolev space are generally discontinuous at the interior edges.

Functions in $H^1(\mathcal{P}^{\mathsf{h}})$ have traces on Γ. These are single-valued at boundary edges and double-valued at interior edges. To handle the traces, we introduce for each boundary edge $e \in \partial\Omega$ the usual boundary trace $(\cdot)_e$ as

$$u_e := \lim_{s\downarrow 0} u(e - n_e\, s)\ ,$$

and we introduce for each interior edge $e \in \Gamma_{\mathcal{I}}$ the $\pm$-trace, $(\cdot)_e^{\pm}$, as

$$u_e^{\pm} := \lim_{s\downarrow 0} u(e \pm s)\ .$$

Furthermore, we define the *average* $\{\cdot\}_e$ and the *jump* $[\![\cdot]\!]_e$ for each interior edge $e \in \Gamma_{\mathcal{I}}$ in the usual manner:

$$\begin{aligned} \{u\}_e &:= \tfrac{1}{2}(u_e^+ + u_e^-)\ , \\ [\![u]\!]_e &:= u_e^+ - u_e^-\ . \end{aligned}$$

These trace operators are bounded in $\mathbb{R}$ for functions in $H^1(\mathcal{P}^{\mathsf{h}})$; that is, trace inequalities hold.

2.3. DG formulations of the Poisson problem. Let $H := H(\mathcal{P}^{\mathsf{h}})$ be a broken space subordinate to the partition $\mathcal{P}^{\mathsf{h}}$ and $\widehat{H} := \widehat{H}(\mathcal{P}^{\mathsf{h}}) \subset H(\mathcal{P}^{\mathsf{h}})$ a finite-dimensional subspace. The generic form of a continuum DG formulation is given by the following abstract Galerkin variational problem:

$$\boxed{\begin{array}{l}\text{Find } u \in H : \\ \qquad\qquad \mathscr{B}(u,v) = \mathscr{L}(v) \qquad \forall v \in H\ .\end{array}} \tag{2.6}$$

Clearly, the continuum DG formulation should be consistent with the Poisson problem; i.e., the solution of (2.1) must comply with (2.6). The generic form of the corresponding approximate DG problem is given by

$$\boxed{\begin{array}{l}\text{Find } \widehat{u} \in \widehat{H} : \\ \qquad\qquad \mathscr{B}(\widehat{u},v) = \mathscr{L}(v) \qquad \forall v \in \widehat{H}\ .\end{array}} \tag{2.7}$$

We will refer to the broken space associated with a particular DG problem as its *DG space*.

The conventional approach to constructing consistent DG formulations premises that $f \in L_2(\Omega)$ and, accordingly, $u \in H^2(\Omega) \subset H^2(\mathcal{P}^{\mathsf{h}}) \subset H$. Multiplication of (2.3a) with $v \in H$, integration on Ω, and elementwise integration by parts then yield

$$\sum_{K\in\mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x - \sum_{e\in\Gamma_{\mathcal{I}}} \left(\partial_n u[\![v]\!]\right)_e - \sum_{e\in\Gamma_{\mathcal{D}}} \left((\partial_n u)v\right)_e = \int_\Omega fv\,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}}v\right)_e \quad \forall v \in H\ .$$

For $u \in H^2(\Omega)$, $(\partial_n u)_e$ is well defined for $e \in \Gamma_{\mathcal{I}}$. However, for u in the DG space H, $(\partial_n u)_e$ is not uniquely defined at the interior edges. Therefore, $(\partial_n u)_e$ is conventionally replaced by $\{\partial_n u\}_e$. On account of $\{\partial_n u\}_e = (\partial_n u)_e$ for $u \in H^2(\Omega)$, this replacement preserves consistency. In addition, the bilinear form can be augmented with other products of edge values and/or edge derivatives, for instance, $\{\partial_n v\}_e[\![u]\!]_e$ for $e \in \Gamma_{\mathcal{I}}$. Most concurrent DG formulations are the result of such an augmentation and, accordingly, we will refer to such augmentations as *conventional edge terms*, and to the corresponding variational statements as *conventional* DG formulations. A precise definition is provided in section 4. Alternatively, the bilinear form can be endowed with other consistency-preserving edge terms, e.g., based on lift operators [4,5,6]. We collectively refer to such terms as *nonconventional edge terms*.

The above exposition furnishes the context for the problem considered in this paper. Our first objective is to establish that all conventional DG formulations are necessarily noncoercive, in contrast to the classical CG formulation. Conventional DG formulations are contingent on weak coercivity for their well-posedness. However, at variance with coercivity, weak coercivity does not transfer to subspaces and, consequently, well-posedness of the continuum DG formulation does not generally imply well-posedness of corresponding approximate DG problems. Moreover, we introduce a new nonconventional symmetric DG formulation based on *element Green's functions* that is coercive on the broken Sobolev space $H^1(\mathcal{P}^{\mathsf{h}})$.

3. Existence and uniqueness theorems for linear variational problems. In this section we review elementary existence theorems pertaining to the well-posedness of linear variational problems. These theorems form the basis for our analysis in section 4 and section 5. Furthermore, a priori error estimates are given for Galerkin approximations.

Section 3.1 is concerned with the generalized Lax–Milgram theorem. This theorem provides the fundament for the classical Lax–Milgram theorem in section 3.2.

3.1. The generalized Lax–Milgram theorem. The generalized Lax–Milgram theorem gives necessary and sufficient conditions for the well-posedness of a generic linear variational problem. Its proof can be found in [11,16].[2]

THEOREM 1 (generalized Lax–Milgram). *Let H be a real Hilbert space with corresponding norm $\|\cdot\|_H$. Consider a continuous bilinear form $\mathscr{B} : H \times H \to \mathbb{R}$; i.e., there exists a positive constant c_b such that*

$$|\mathscr{B}(u,v)| \le c_b\|u\|_H\|v\|_H \qquad \forall u, v \in H\ .$$

[2]The necessity of (3.1) is shown in [16].

If and only if $\mathscr{B}(\cdot,\cdot)$ is weakly coercive on $H \times H$, i.e., there exists a constant $\gamma > 0$ such that

$$\inf_{u\in H\setminus\{0\}} \sup_{v\in H\setminus\{0\}} \frac{\mathscr{B}(u,v)}{\|u\|_H \|v\|_H} \geq \gamma \,, \tag{3.1a}$$

$$\sup_{u\in H} \mathscr{B}(u,v) > 0 \qquad \forall v \in H \setminus \{0\} \,, \tag{3.1b}$$

then for every continuous linear functional $\mathscr{L} : H \to \mathbb{R}$, problem (2.6) *has a unique solution $u \in H$.*

Inequality (3.1a) is known as the *inf-sup condition*, and the supremum over all numbers γ in compliance with (3.1a) is referred to as the *inf-sup constant*.

Let $\widehat{H} \subset H$ be a closed subspace associated with an approximate variational problem. As a closed subspace of a Hilbert space is itself a Hilbert space, well-posedness of the approximate problem on $\widehat{H}$ is settled identically by Theorem 1 with H replaced by $\widehat{H}$. The corresponding inf-sup constant $\widehat{\gamma}$ then generally depends on the approximation space $\widehat{H}$, i.e., $\widehat{\gamma} := \widehat{\gamma}(\widehat{H})$. Moreover, if the approximate problem is well-posed, its solution $\widehat{u} \in \widehat{H}$ complies with the a priori estimate

$$\|u - \widehat{u}\|_H \leq (1+c_b/\widehat{\gamma}) \inf_{v\in\widehat{H}} \|u - v\|_H \,. \tag{3.2}$$

It is to be noted that weak coercivity on $H \times H$ does not imply weak coercivity on $\widehat{H} \times \widehat{H}$. Therefore, well-posedness of the continuum problem does not imply well-posedness of corresponding approximate problems. On account of the dependence of $\widehat{\gamma}$ on $\widehat{H}$, it moreover holds that if we consider a sequence of asymptotically dense nested approximation spaces $\widehat{H}^{(1)} \subset \widehat{H}^{(2)} \subset \cdots \subseteq H$, $\widehat{H}^{(m)} \to H$ as $m \to \infty$, then the corresponding approximations $\widehat{u}^{(m)}$ need not converge, or need not converge monotonously.

3.2. The classical Lax–Milgram theorem. A theorem on the well-posedness of linear Galerkin variational problems with more restrictive conditions and stronger implications is provided by the classical Lax–Milgram theorem (see, e.g., [11, 16]).

THEOREM 2 (classical Lax–Milgram). *Let $\mathscr{B} : H \times H \to \mathbb{R}$ be a continuous, (strongly) coercive bilinear form on H; i.e., there exists a positive constant κ such that*

$$|\mathscr{B}(u,u)| \geq \kappa \|u\|_H^2 \qquad \forall u \in H \,. \tag{3.3}$$

Then for every continuous linear functional $\mathscr{L} : H \to \mathbb{R}$, the variational problem (2.6) *possesses a unique solution $u \in H$.*

Coercivity on H is a sufficient condition for weak coercivity on $H \times H$. As coercivity transfers to subspaces $\widehat{H} \subset H$, it holds that well-posedness of the continuum problem implies well-posedness of approximate problems based on conforming subspaces. Moreover, the subspace approximation $\widehat{u} \in \widehat{H}$ satisfies the a priori estimate

$$\|u - \widehat{u}\|_H \leq c_b/\kappa \inf_{v\in\widehat{H}} \|u - v\|_H \,, \tag{3.4}$$

where c_b and κ denote the continuity and coercivity constant of the bilinear form on $H \times H$, respectively. It is to be noted that the constants in (3.4) are independent of the approximation space. Hence, if $\widehat{H}^{(2)} \subset H$ is a larger approximation space than $\widehat{H}^{(1)} \subset \widehat{H}^{(2)}$, then the error (measured in $\|\cdot\|_H$) of the corresponding approximate

solution $\widehat{u}^{(2)} \in \widehat{H}^{(2)}$ is at most equal to that of the approximate solution $\widehat{u}^{(1)} \in \widehat{H}^{(1)}$. In particular, this implies that if we consider a sequence of asymptotically dense nested approximation spaces $\widehat{H}^{(1)} \subset \widehat{H}^{(2)} \subset \cdots \subseteq H$ and $\widehat{H}^{(m)} \to H$ as $m \to \infty$, then the error $\|u - \widehat{u}^{(m)}\|_H$ converges monotonously to 0 as m increases.

Let us consider an arbitrary variational continuum problem. Under the condition of coercivity of the bilinear form, conforming approximate problems are well-posed if the continuum problem is well-posed. The premise of coercivity not only provides a sufficient condition; it is also *necessary*.

PROPOSITION 3. *Consider a continuous linear functional $\mathscr{L} : H \to \mathbb{R}$ and continuous bilinear form $\mathscr{B} : H \times H \to \mathbb{R}$. If and only if $\mathscr{B}(\cdot,\cdot)$ is coercive on H, then well-posedness of the continuum problem* (2.6) *implies well-posedness of the approximate problem* (2.7).

Proof. (i) Forward implication: By Theorem 2, coercivity on H ensures that the continuum problem (2.6) and the approximate problem (2.7) are well-posed.

(ii) Reverse implication: We show the proof by contradiction. We assume that $\mathscr{B}(\cdot,\cdot)$ is weakly coercive on $H \times H$, but not coercive, and then construct a subspace $\widehat{H} \subset H$ in which the approximate problem is ill-posed. As H is a closed space, noncoercivity implies the existence of a $\bar{u} \in H \setminus \{0\}$ such that $\mathscr{B}(\bar{u},\bar{u}) = 0$. Taking the approximate space as the one-dimensional space $\widehat{H} = \operatorname{span}\{\bar{u}\}$, it follows that

$$\inf_{u \in \widehat{H}\setminus\{0\}} \sup_{v \in \widehat{H}\setminus\{0\}} \frac{\mathscr{B}(u,v)}{\|u\|_H \|v\|_H} = \frac{\mathscr{B}(\bar{u},\bar{u})}{\|\bar{u}\|_H^2} = 0 \,;$$

i.e., weak coercivity does not hold on $\widehat{H} \times \widehat{H}$. By Theorem 1 weak coercivity on $\widehat{H} \times \widehat{H}$ is necessary for well-posedness of the approximate problem. ∎

4. Conventional DG formulations. This section is concerned with an analysis of the generic properties of conventional DG formulations. To this end, we introduce a generic consistent conventional DG formulation in section 4.1. Section 4.2 establishes the existence of well-posed conventional DG formulations. Section 4.3 proves that consistent conventional DG formulations are necessarily noncoercive.

4.1. Generic conventional DG formulation. Consider the following bilinear form $\mathscr{B}_\Lambda(\cdot,\cdot)$ and linear functional $\mathscr{L}_\Lambda(\cdot)$:[3]

$$\mathscr{B}_\Lambda(u,v) := \sum_{K \in \mathcal{P}^h} \int_K \frac{\mathrm{d}u}{\mathrm{d}x} \frac{\mathrm{d}v}{\mathrm{d}x} \,\mathrm{d}x + \sum_{e \in \Gamma} \left(\boldsymbol{u}^\mathsf{T} \Lambda \boldsymbol{v}\right)_e \,, \tag{4.1a}$$

$$\mathscr{L}_\Lambda(v) := \int_\Omega f v \,\mathrm{d}x + \sum_{e \in \Gamma_\mathcal{D}} \left(g_\mathcal{D} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e + \sum_{e \in \Gamma_\mathcal{N}} \left(g_\mathcal{N} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e \,, \tag{4.1b}$$

[3]For notational transparency, in a composition of terms with a subscript $(\cdot)_e$, we suppress the subscript of the individual terms and append it to enclosing parentheses. For example, $(g_\mathcal{D} \bar{\Lambda} \bar{\boldsymbol{v}})_e$ is to be interpreted as $g_{\mathcal{D}e} \bar{\Lambda}_e \bar{\boldsymbol{v}}_e$. Moreover, we occasionally suppress the subscript $(\cdot)_e$ entirely if the dependence is apparent from the context.

where boldfaced variables, such as $\boldsymbol{v}$, and boldfaced overlined variables, such as $\bar{\boldsymbol{v}}$, denote (column-) vectors containing values at edge e according to

$$\boldsymbol{v}_e := \begin{cases} \big(h^{-\frac{1}{2}}\,[\![v]\!],\ h^{\frac{1}{2}}\,\{\partial_n v\},\ h^{\frac{1}{2}}\,[\![\partial_n v]\!],\ h^{-\frac{1}{2}}\,\{v\}\big)_e^{\mathsf{T}}\,, & e\in\Gamma_{\mathcal{I}}\,,\\ \big(h^{-\frac{1}{2}}\,v,\ h^{\frac{1}{2}}\,\partial_n v\big)_e^{\mathsf{T}}\,, & e\in\partial\Omega\,, \end{cases} \tag{4.2a}$$

$$\bar{\boldsymbol{v}}_e := \begin{cases} \big(h^{-1}\,v,\ \partial_n v\big)_e^{\mathsf{T}}\,, & e\in\Gamma_{\mathcal{D}}\,,\\ \big(v,\ h\,\partial_n v\big)_e^{\mathsf{T}}\,, & e\in\Gamma_{\mathcal{N}}\,. \end{cases} \tag{4.2b}$$

The matrices $\Lambda_e\in\mathbb{R}^{4\times4}$ ($e\in\Gamma_{\mathcal{I}}$), and $\Lambda_e\in\mathbb{R}^{2\times2}$, $\bar{\Lambda}_e\in\mathbb{R}^{1\times2}$ ($e\in\partial\Omega$) specify bilinear relations between edge values and edge derivatives of u and v. A conventional edge term can now be precisely defined as any term in the bilinear form conforming to $(\boldsymbol{u}^{\mathsf{T}}\Lambda\boldsymbol{v})_e$ for all $e\in\Gamma$, and any term in the linear form conforming to $(g_{\mathcal{D}}\bar{\Lambda}\bar{\boldsymbol{v}})_e$ for $e\in\Gamma_{\mathcal{D}}$ or $(g_{\mathcal{N}}\bar{\Lambda}\bar{\boldsymbol{v}})_e$ for $e\in\Gamma_{\mathcal{N}}$.

The constants $h:\Gamma\to\mathbb{R}$ in (4.2) are local mesh parameters introduced to minimize the mesh dependence of the matrices. Typically, for $e\in\Gamma_{\mathcal{I}}$, h_e is set to the average of the lengths of the elements sharing edge e and for $e\in\partial\Omega$ it is set to half the length of the element contiguous to edge e; i.e.,

$$h_e = \tfrac{1}{2}\sum_{K\in\mathcal{K}_e} h_K \qquad \forall e\in\Gamma\,. \tag{4.3}$$

To provide a functional setting for conventional DG formulations, we introduce the norm $\|\cdot\|_{H_\Lambda}$,

$$\|u\|_{H_\Lambda}^2 := \sum_{K\in\mathcal{P}^{\mathsf{h}}} |u|_{1,K}^2 + \sum_{e\in\Gamma}\big(\boldsymbol{u}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{u}\big)_e\,, \tag{4.4}$$

where the seminorm $|\cdot|_{1,K}$ is defined by

$$|u|_{1,K}^2 := \int_K \Big(\frac{\mathrm{d}u}{\mathrm{d}x}\Big)^2\,\mathrm{d}x\,,$$

and $\mathrm{D}_\Lambda(=\mathrm{D}_{\Lambda_e})$ is a diagonal matrix in $\mathbb{R}^{4\times4}$ for $e\in\Gamma_{\mathcal{I}}$ and in $\mathbb{R}^{2\times2}$ for $e\in\partial\Omega$ with diagonal entries

$$(\mathrm{D}_\Lambda)_{ii} := \textstyle\sum_j\big(\big|(\mathrm{S}_\Lambda)_{ij}\big|+\big|(\mathrm{A}_\Lambda)_{ij}\big|\big) \quad\text{with}\quad \mathrm{S}_\Lambda := \tfrac{1}{2}\big(\Lambda+\Lambda^{\mathsf{T}}\big),\quad \mathrm{A}_\Lambda := \tfrac{1}{2}\big(\Lambda-\Lambda^{\mathsf{T}}\big); \tag{4.5}$$

i.e., D_Λ is obtained from Λ by lumping the absolute values of its symmetric part S_Λ and its antisymmetric part A_Λ to the diagonal.[4] The matrices Λ and D_Λ are then related by

$$\begin{aligned} \big|\boldsymbol{u}^{\mathsf{T}}\Lambda\boldsymbol{v}\big| &= \big|\textstyle\sum_{i,j}\boldsymbol{u}_i(\mathrm{S}_{\Lambda ij}+\mathrm{A}_{\Lambda ij})\boldsymbol{v}_j\big| \le \sum_{i,j}|\boldsymbol{u}_i|\,(|\mathrm{S}_{\Lambda ij}|+|\mathrm{A}_{\Lambda ij}|)\,|\boldsymbol{v}_j|\\ &\le \sqrt{\textstyle\sum_{i,j}(|\mathrm{S}_{\Lambda ij}|+|\mathrm{A}_{\Lambda ij}|)\boldsymbol{u}_i^2}\sqrt{\textstyle\sum_{i,j}(|\mathrm{S}_{\Lambda ij}|+|\mathrm{A}_{\Lambda ij}|)\boldsymbol{v}_j^2} = \sqrt{\boldsymbol{u}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{u}}\,\sqrt{\boldsymbol{v}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{v}}\,. \end{aligned} \tag{4.6}$$

[4]Strictly speaking, $\|\cdot\|_{H_\Lambda}$ is a norm only if $(\mathrm{D}_\Lambda)_{11}>0$ for $e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}$. This implies that the bilinear form incorporates $[\![u]\!]_e$ and/or $[\![v]\!]_e$ on $\Gamma_{\mathcal{I}}$ and u_e and/or v_e on $\Gamma_{\mathcal{D}}$. However, in Proposition 4 it will be shown that any consistent formulation necessarily contains such terms. Hence, there is no loss of generality in proceeding under the assumption that $\|\cdot\|_{H_\Lambda}$ provides a norm.

The second inequality in (4.6) follows from the discrete Schwartz inequality, viz., $\sum |x_i y_i| \leq \sqrt{\sum x_i^2}\,\sqrt{\sum y_i^2}$. We now define the space H_Λ as the completion of $H^2(\mathcal{P}^\mathsf{h})$ under norm $\|\cdot\|_{H_\Lambda}$:

$$\boxed{H_\Lambda := H_\Lambda(\mathcal{P}^\mathsf{h}) = \overline{H^2(\mathcal{P}^\mathsf{h})}^{\|\cdot\|_{H_\Lambda}}\ .} \tag{4.7}$$

The Hilbert space H_Λ defined in this manner provides the appropriate space for the generic conventional DG formulation:

$$\boxed{\begin{array}{l}\text{Find } u \in H_\Lambda : \\ \qquad\qquad \mathscr{B}_\Lambda(u,v) = \mathscr{L}_\Lambda(v) \qquad \forall v \in H_\Lambda\ .\end{array}} \tag{4.8}$$

The appropriateness of H_Λ is rigorously settled in section 4.2. Let us note that H_Λ is a generalization of the space used in [3].

4.2. Well-posedness results for the continuum formulation. Under certain conditions on the matrices Λ and $\bar{\Lambda}$, (4.8) provides a consistent, well-posed weak formulation of (2.3). The proposition below specifies necessary and sufficient conditions on the matrices Λ and $\bar{\Lambda}$ for consistency with (2.3).

PROPOSITION 4 (consistency of conventional DG). *If and only if the matrices* Λ, $\bar{\Lambda}$ *are of the form*

$$\Lambda_e = \begin{pmatrix} \alpha & \delta & \gamma^\mathrm{u} & \zeta^1 \\ -1 & 0 & 0 & 0 \\ \gamma^\mathrm{l} & \varepsilon & \beta & \zeta^2 \\ 0 & 0 & 0 & 0 \end{pmatrix}_e \qquad \forall e \in \Gamma_\mathcal{I}\ , \tag{4.9a}$$

$$\Lambda_e = \begin{pmatrix} \alpha & \delta \\ -1 & 0 \end{pmatrix}_e\ , \quad \bar{\Lambda}_e = \begin{pmatrix} \alpha & \delta \end{pmatrix}_e \qquad \forall e \in \Gamma_\mathcal{D}\ , \tag{4.9b}$$

$$\Lambda_e = \begin{pmatrix} 0 & 0 \\ \varepsilon & \beta \end{pmatrix}_e\ , \quad \bar{\Lambda}_e = \begin{pmatrix} \varepsilon{+}1 & \beta \end{pmatrix}_e \qquad \forall e \in \Gamma_\mathcal{N} \tag{4.9c}$$

for certain fixed parameters $\alpha_e, \beta_e, \gamma_e^\mathrm{u}, \gamma_e^\mathrm{l}, \delta_e, \varepsilon_e, \zeta_e^1, \zeta_e^2 \in \mathbb{R}$ *(for all* $e \in \Gamma$*), then the corresponding conventional DG formulation* (4.8) *is consistent with* (2.3)*; i.e., the solution* $u \in H^2(\Omega) \subset H_\Lambda$ *of* (2.3) *complies with* (4.8)*.*

Proof. (i) Forward implication: Let $u \in H^2(\Omega)$ solve (2.3). Multiplying (2.3a) by an arbitrary $v \in H_\Lambda$, integrating on Ω, and invoking integration by parts, elementwise, we obtain

$$\sum_{K\in\mathcal{P}^\mathsf{h}} \int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = \int_\Omega fv\,\mathrm{d}x + \sum_{e\in\Gamma_\mathcal{I}} \big(\partial_n u[\![v]\!]\big)_e + \sum_{e\in\partial\Omega} \big((\partial_n u)v\big)_e\ . \tag{4.10}$$

From (4.1a) and (4.10) it follows that

$$\mathscr{B}_\Lambda(u,v) = \int_\Omega fv\,\mathrm{d}x + \sum_{e\in\Gamma_\mathcal{I}} \big(\partial_n u[\![v]\!]\big)_e + \sum_{e\in\Gamma_\mathcal{D}\cup\Gamma_\mathcal{N}} \big((\partial_n u)v\big)_e + \sum_{e\in\Gamma} \big(\boldsymbol{u}^\mathsf{T}\Lambda\boldsymbol{v}\big)_e\ . \tag{4.11}$$

The boundary conditions (2.3b) and (2.3c) imply that $u_e = (g_\mathcal{D})_e$ for $e \in \Gamma_\mathcal{D}$ and $(\partial_n u)_e = (g_\mathcal{N})_e$ for $e \in \Gamma_\mathcal{N}$. Moreover, on account of the C^1-continuity of functions

in $H^2(\Omega)$, the solution u complies with $[\![u]\!]_e = [\![\partial_n u]\!]_e = 0$ and $\{\partial_n u\}_e = (\partial_n u)_e$ for $e \in \Gamma_{\mathcal{I}}$. Hence, upon replacing Λ in (4.11) with (4.9), we obtain

$$\mathscr{B}_\Lambda(u,v) = \int_\Omega fv \,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{D}}} \left(\alpha g_{\mathcal{D}} v/h + \delta g_{\mathcal{D}} \partial_n v\right)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}} v + \varepsilon g_{\mathcal{N}} v + \beta h g_{\mathcal{N}} \partial_n v\right)_e .$$

By (4.1b) and (4.9), for any $v \in H_\Lambda$,

$$\mathscr{L}_\Lambda(v) = \int_\Omega fv \,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{D}}} \left(\alpha g_{\mathcal{D}} v/h + \delta g_{\mathcal{D}} \partial_n v\right)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}} v + \varepsilon g_{\mathcal{N}} v + \beta h g_{\mathcal{N}} \partial_n v\right)_e ,$$

and, hence, $\mathscr{B}_\Lambda(u,v) = \mathscr{L}_\Lambda(v)$ for all $v \in H_\Lambda$.

(ii) Reverse implication: By (4.1b), (4.8), and (4.11),

$$\sum_{e\in\Gamma_{\mathcal{I}}} \left(\partial_n u [\![v]\!]\right)_e + \sum_{e\in\partial\Omega} \left((\partial_n u)v\right)_e + \sum_{e\in\Gamma} \left(\boldsymbol{u}^{\mathsf{T}} \Lambda \boldsymbol{v}\right)_e = \sum_{e\in\Gamma_{\mathcal{D}}} \left(g_{\mathcal{D}} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e \quad \forall v \in H_\Lambda .$$

Upon rearranging the summations, replacing $\boldsymbol{u}_e$ according to its definition (4.2), and invoking the boundary conditions in (2.3) and $[\![u]\!]_e = [\![\partial_n u]\!]_e = 0$, $\{\partial_n u\}_e = (\partial_n u)_e$, and $\{u\}_e = u_e$ for $e \in \Gamma_{\mathcal{I}}$, we obtain

(4.12)
$$\sum_{e\in\Gamma_{\mathcal{I}}} \left(\partial_n u [\![v]\!] + (0,\ h^{\frac{1}{2}}\partial_n u, 0,\ h^{-\frac{1}{2}} u)\Lambda \boldsymbol{v}\right)_e + \sum_{e\in\Gamma_{\mathcal{D}}} \left((\partial_n u)v + (h^{-\frac{1}{2}} g_{\mathcal{D}},\ h^{\frac{1}{2}}\partial_n u)\Lambda \boldsymbol{v}\right)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}} v + (h^{-\frac{1}{2}} u,\ h^{\frac{1}{2}} g_{\mathcal{N}})\Lambda \boldsymbol{v}\right)_e = \sum_{e\in\Gamma_{\mathcal{D}}} \left(g_{\mathcal{D}} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \left(g_{\mathcal{N}} \bar{\Lambda} \bar{\boldsymbol{v}}\right)_e$$

for all $v \in H_\Lambda$. Selecting a $v \in H_\Lambda$ such that $[\![v]\!]_e = 1$ for some edge $e \in \Gamma_{\mathcal{I}}$ and such that all other edge terms vanish, we obtain the identity

$$\left(\partial_n u + (\partial_n u)\Lambda_{21} + u\Lambda_{41}/h\right)_e = 0 .$$

Therefore, $(\Lambda_{21})_e = -1$ and $(\Lambda_{41})_e = 0$. Similarly, by making appropriate choices for the test function $v \in H_\Lambda$ in (4.12), the precise form (4.9) can be established. ∎

To establish the conditions on the matrices $\Lambda, \bar{\Lambda}$ in (4.9) for well-posedness, we appeal to the generalized Lax–Milgram theorem, Theorem 1. In particular, we establish the conditions for continuity of $\mathscr{L}_\Lambda(\cdot)$, and for continuity and weak coercivity of $\mathscr{B}_\Lambda(\cdot,\cdot)$. Continuity of the bilinear form is in fact independent of the precise form of Λ. This is asserted by the following proposition.

PROPOSITION 5 (continuity of $\mathscr{B}_\Lambda$). *The bilinear form $\mathscr{B}_\Lambda(\cdot,\cdot)$ given in* (4.1a) *is continuous on H_Λ, i.e.,*

$$|\mathscr{B}_\Lambda(u,v)| \le c_b \|u\|_{H_\Lambda} \|v\|_{H_\Lambda} \qquad \forall u, v \in H_\Lambda ,$$

with continuity constant $c_b = 1$.

Proof. First, note that

$$|\mathscr{B}_\Lambda(u,v)| \le \sum_{K\in\mathcal{P}^{\mathsf{h}}} \int_K \left|\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\right| \mathrm{d}x + \sum_{e\in\Gamma} \left|\boldsymbol{u}^{\mathsf{T}} \Lambda \boldsymbol{v}\right|_e .$$

From the Schwarz inequality and (4.6) it follows that

$$|\mathscr{B}_\Lambda(u,v)| \leq \sum_{K\in\mathcal{P}^\mathrm{h}} |u|_{1,K}\,|v|_{1,K} + \sum_{e\in\Gamma} \left(\sqrt{\boldsymbol{u}^\mathsf{T}\mathrm{D}_\Lambda\boldsymbol{u}}\sqrt{\boldsymbol{v}^\mathsf{T}\mathrm{D}_\Lambda\boldsymbol{v}}\,\right)_e .$$

Application of the discrete Schwarz inequality then yields

$$|\mathscr{B}_\Lambda(u,v)| \leq \left(\sum_{K\in\mathcal{P}^\mathrm{h}} |u|^2_{1,K} + \sum_{e\in\Gamma} \left(\boldsymbol{u}^\mathsf{T}\mathrm{D}_\Lambda\boldsymbol{u}\right)_e\right)^{1/2} \left(\sum_{K\in\mathcal{P}^\mathrm{h}} |v|^2_{1,K} + \sum_{e\in\Gamma} \left(\boldsymbol{v}^\mathsf{T}\mathrm{D}_\Lambda\boldsymbol{v}\right)_e\right)^{1/2} . \qquad \square$$

In a similar manner it can be shown that for all $f \in [H_\Lambda]'$, $\mathscr{L}_\Lambda(\cdot)$ is a continuous functional on H_Λ. Hence, it remains to derive the conditions on the matrices Λ_e in (4.9) which yield $\mathscr{B}_\Lambda(\cdot,\cdot)$ weakly coercive on $H_\Lambda \times H_\Lambda$. Sufficient conditions for weak coercivity are established in Proposition 6. As the proof is rather elaborate, it is transferred to Appendix A.

PROPOSITION 6 (weak coercivity of $\mathscr{B}_\Lambda$). *If the parameters in the matrices Λ_e in* (4.9) *satisfy the algebraic conditions*

$$\left.\begin{array}{ll} & \alpha \in \mathbb{R}, \\ & \beta,\gamma^\mathrm{u},\gamma^\mathrm{l},\delta,\varepsilon \in \mathbb{R}: \\ & \quad \beta,\gamma^\mathrm{u},\gamma^\mathrm{l},\varepsilon = 0 \ \wedge\ 4 > |\delta| \neq 0, \\ & \textit{or} \quad \delta\beta - \varepsilon\gamma^\mathrm{u} \neq 0 \ \wedge\ 4 > \tfrac{1}{2}|\delta+1| + \tfrac{1}{2}|\delta-1| + |\varepsilon|, \\ & \zeta^1,\zeta^2 = 0 \end{array}\right\} \quad \forall e \in \Gamma_\mathcal{I} , \tag{4.13a}$$

$$\alpha \in \mathbb{R}, \quad 4 > |\delta| \neq 0 \qquad \forall e \in \Gamma_\mathcal{D} , \tag{4.13b}$$

$$\beta \in \mathbb{R}, \quad \varepsilon = 0 \qquad \forall e \in \Gamma_\mathcal{N} , \tag{4.13c}$$

then the corresponding bilinear form $\mathscr{B}_\Lambda(\cdot,\cdot)$ in (4.1a) *is weakly coercive on $H_\Lambda \times H_\Lambda$ with an inf-sup constant $\gamma_\Lambda > 0$.*

Let us note that $\{\delta : 4 > |\delta|\} = \{\delta : 4 > \frac{1}{2}|\delta+1| + \frac{1}{2}|\delta-1|\}$. Proposition 6 generalizes the proof of weak coercivity of the BODG in [3] to any consistent conventional DG formulation. We remark that although the conditions in (4.13) are unrestrictive, they can in fact be further weakened.

In conclusion, by the generalized Lax–Milgram theorem, Theorem 1, if the matrices Λ, $\bar{\Lambda}$ conform to (4.9) and (4.13), then for every $f \in [H_\Lambda]'$ the corresponding conventional DG formulation (4.8) is well-posed and consistent with (2.3). In Table 1, we have summarized the parameter choices for several conventional DG formulations

TABLE 1
Parameters in the matrices Λ, $\bar{\Lambda}$ in (4.9) *for several conventional DG formulations.*

DG formulation		α	β	$\gamma^\mathrm{u},\gamma^\mathrm{l}$	δ	ε	ζ^1,ζ^2
GEM	[10, 12]	0	0	0	-1	0	0
IPDG	[2, 10, 12]	α	0	0	-1	0	0
BODG	[3]	0	0	0	1	0	0
NIPDG	[13]	α	0	0	1	0	0
SDG	[14]	0	β	0	1	0	0
LNDG	[9]	α	0	0	δ	0	0

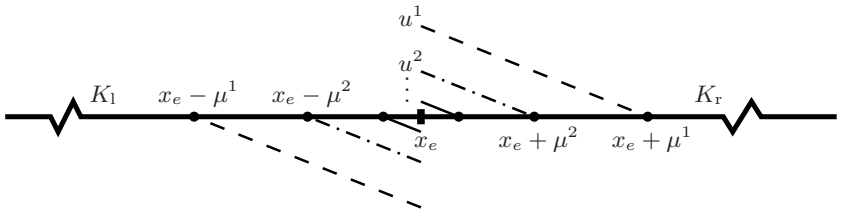


FIG. 1. *Example of a Cauchy sequence* $\{u^i\}$ *in* H_Λ *satisfying* (4.15).

that have appeared in the literature. It can be verified that all formulations, except LNDG, satisfy immediately the conditions in (4.13). LNDG requires the auxiliary condition $4 > |\delta| \neq 0$.

4.3. Noncoercivity of consistent conventional DG formulations. The exposition in section 3 motivates the pursuit of a consistent formulation that is coercive on H_Λ, rather than only weakly coercive. However, the proposition below asserts that a coercive consistent conventional DG formulation is nonexistent; i.e., of the DG formulations in compliance with (4.9), none is coercive.

PROPOSITION 7 (noncoercivity of $\mathscr{B}_\Lambda$). *The bilinear form* $\mathscr{B}_\Lambda(\cdot,\cdot)$ *in* (4.1a) *with* Λ *subject to the consistency requirement* (4.9) *is noncoercive on* H_Λ*; i.e., a positive constant* κ *such that*

$$|\mathscr{B}_\Lambda(u,u)| \geq \kappa \|u\|^2_{H_\Lambda} \qquad \forall u \in H_\Lambda \tag{4.14}$$

is nonexistent.

Proof. We show the existence of a Cauchy sequence $\{u^i\}$ in H_Λ such that $\mathscr{B}_\Lambda(u^i,u^i) \to 0$ and $\|u^i\|_{H_\Lambda} \to c \geq 1$ as $i \to \infty$. Consider an interior edge $e \in \Gamma_\mathcal{I}$ and the left and right elements $K_\mathrm{l}, K_\mathrm{r} \in \mathcal{K}_e$ contiguous to this edge.[5] The Cauchy sequence is chosen such that its elements $u^i \in H_\Lambda$ have local support ($\mathrm{supp}(u^i) \subset \overline{K_\mathrm{l} \cup K_\mathrm{r}}$ with strict inclusion) and, moreover,

$$|u^i|_{1,K_\mathrm{l}}, |u^i|_{1,K_\mathrm{r}} \to 0\,, \tag{4.15a}$$

$$\{u^i\}_e = 0\,,\ [\![u^i]\!]_e \to 0\,, \tag{4.15b}$$

$$h_e^{\frac{1}{2}}\{u^i_n\}_e = 1\,,\ [\![u^i_n]\!]_e = 0\,. \tag{4.15c}$$

An example of a sequence satisfying (4.15) is the sequence $u^1, u^2, \ldots$ depicted in Figure 1. The support of u^i is the closed interval in $\mathbb{R}$ with length $2\mu^i$ centered at e. The length of the support set forms a Cauchy sequence $\{\mu^i\}$ in $\mathbb{R}$ with limit $\lim_{i\to\infty} \mu^i = 0$. Moreover, within the support set, u^i is an asymmetric, piecewise linear function.

From the consistency conditions (4.9) on Λ_e and the properties (4.15) of the sequence $\{u^i\}$ it follows that

$$\begin{aligned}\mathscr{B}_\Lambda(u^i,u^i) &= |u^i|^2_{1,K_\mathrm{l}} + |u^i|^2_{1,K_\mathrm{r}} + \begin{pmatrix} h^{-\frac{1}{2}}[\![u^i]\!] & 1 & 0 & 0\end{pmatrix}_e \begin{pmatrix} \alpha & \delta & \gamma^\mathrm{u} & \zeta^1 \\ -1 & 0 & 0 & 0 \\ \gamma^\mathrm{l} & \varepsilon & \beta & \zeta^2 \\ 0 & 0 & 0 & 0 \end{pmatrix}_e \begin{pmatrix} h^{-\frac{1}{2}}[\![u^i]\!] \\ 1 \\ 0 \\ 0 \end{pmatrix}_e \\ &= |u^i|^2_{1,K_\mathrm{l}} + |u^i|^2_{1,K_\mathrm{r}} + \left(\alpha[\![u^i]\!]^2/h + (\delta-1)h^{-\frac{1}{2}}[\![u^i]\!]\right)_e\,,\end{aligned}$$

[5] If there are no interior edges ($\Gamma_\mathcal{I} = \emptyset$), a proof of noncoercivity can be established similarly by considering a Dirichlet boundary edge.

and, hence, $\mathscr{B}_\Lambda(u^i,u^i)\to 0$ as $i\to\infty$. Furthermore, the norm of u^i reduces to

$$\begin{aligned}\|u^i\|^2_{H_\Lambda} &= |u^i|^2_{1,K_\mathrm{l}} + |u^i|^2_{1,K_\mathrm{r}} + \begin{pmatrix} h^{-\frac{1}{2}}[\![u^i]\!] & 1 & 0 & 0\end{pmatrix}_e \begin{pmatrix} (\mathrm{D}_\Lambda)_{11} & 0 & 0 & \cdots \\ 0 & (\mathrm{D}_\Lambda)_{22} & & \\ 0 & & \ddots & \\ \vdots & & & \end{pmatrix}_e \begin{pmatrix} h^{-\frac{1}{2}}[\![u^i]\!] \\ 1 \\ 0 \\ 0 \end{pmatrix}_e \\ &= |u^i|^2_{1,K_\mathrm{l}} + |u^i|^2_{1,K_\mathrm{r}} + \big((\mathrm{D}_\Lambda)_{11}[\![u^i]\!]^2/h + (\mathrm{D}_\Lambda)_{22}\big)_e\,.\end{aligned}$$

Thus, as $i\to\infty$,

$$\|u^i\|^2_{H_\Lambda} \to (\mathrm{D}_\Lambda)_{22} = \big(\tfrac{1}{2}|\delta{+}1| + \tfrac{1}{2}|\delta{-}1| + |\varepsilon|\big)_e \geq 1\,.$$

The identity follows by replacing $(\mathrm{D}_\Lambda)_{22}$ in accordance with (4.5) and (4.9a). ∎

5. A new symmetric DG formulation with $H^1(\mathcal{P}^\mathsf{h})$-coercivity. In this section we present a new *nonconventional* coercive symmetric DG formulation based on element Green's functions. Section 5.1 presents the variational formulation. In section 5.2 we establish continuity properties of the corresponding bilinear and linear forms. Finally, in section 5.3 we demonstrate consistency and, most importantly, well-posedness on account of coercivity on $H^1(\mathcal{P}^\mathsf{h})$.

5.1. Weak formulation with element Green's functions. By Proposition 7, a coercive DG formulation must contain nonconventional edge terms. Below, we present a formulation based on *element Green's functions*.

Consider an edge $e\in\Gamma_\mathcal{I}\cup\Gamma_\mathcal{D}$, and a contiguous element $K\in\mathcal{K}_e$. The two-point boundary of K is denoted by $\partial K=\{e,\bar{e}\}$. With the pair (K,e) we associate a function $\phi_{K,e}:K\to\mathbb{R}$ by the auxiliary boundary-value problem

$$-\frac{\mathrm{d}^2\phi_{K,e}}{\mathrm{d}x^2} = 0 \qquad \text{in } K, \tag{5.1a}$$

$$\phi_{K,e} = \begin{cases} -n_e\,n_K & \text{on } e\,, \\ 0 & \text{on } \bar{e}\,, \end{cases} \tag{5.1b}$$

where n_K is the unit outward normal of K. For each edge e and each $K\in\mathcal{K}_e$, the solutions of (5.1) are linear functions on K; see Figure 2. Specifically, $\phi_{K,e}$ corresponds to the element *Dirichlet-to-Neumann Green's function* for the one-dimensional Laplacian. To corroborate this assertion, we multiply (5.1a) with $u\in H^2(K)$ and integrate on K. Upon performing integration by parts twice, and invoking the boundary conditions (5.1b), we obtain

$$(\partial_n u)_e = \int_K -\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\,\phi_{K,e}\,\mathrm{d}x - \sum_{\{e,\bar{e}\}}\Big(u\,\frac{\mathrm{d}\phi_K}{\mathrm{d}x}\,n_K\Big)\,,$$

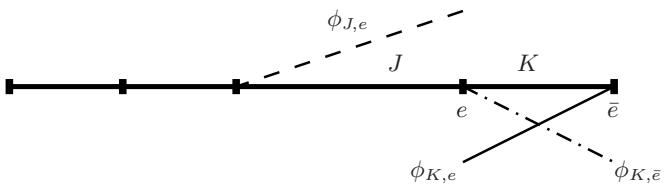


FIG. 2. *Several solutions of auxiliary problem* (5.1).

which shows that the "Neumann value" $\partial_n u$ at e is readily expressed in terms of the Laplacian and "Dirichlet" values of u at e and $\bar{e}$.

For each edge $e \in \Gamma_{\mathcal{I}} \cup \Gamma_{\mathcal{D}}$ we define the functionals $\Phi_e : H^1(\mathcal{P}^{\mathsf{h}}) \to \mathbb{R}$ and $\bar{\Phi}_e : [H^1_{0,\mathcal{D}}(\Omega)]' \to \mathbb{R}$ as

$$\Phi_e(u) := \sum_{K \in \mathcal{K}_e} \theta_{K,e} \int_K \frac{\mathrm{d}u}{\mathrm{d}x} \frac{\mathrm{d}\phi_{K,e}}{\mathrm{d}x} \, \mathrm{d}x \,, \tag{5.2a}$$

$$\bar{\Phi}_e(f) := \sum_{K \in \mathcal{K}_e} \theta_{K,e} \int_K f \phi_{K,e} \, \mathrm{d}x \,. \tag{5.2b}$$

The functionals constitute weighted combinations of contributions of elements that share edge e. The partition-dependent constants $\theta_{K,e} \in \mathbb{R}$ (for $K \in \mathcal{K}_e$) are defined as

$$\theta_{K,e} := h_K / \sum_{J \in \mathcal{K}_e} h_J \,. \tag{5.3}$$

Trivially, $\theta_{K,e} = 1$ for $e \in \Gamma_{\mathcal{D}}$, $K \in \mathcal{K}_e$. It is to be noted that the following partition-of-unity property holds:

$$\sum_{K \in \mathcal{K}_e} \theta_{K,e} = 1 \qquad \forall e \in \Gamma_{\mathcal{I}} \cup \Gamma_{\mathcal{D}} \,. \tag{5.4}$$

Equations (5.2)–(5.4) enable us to condense the new DG formulation into the following variational problem:

(5.5) Find $u \in H^1(\mathcal{P}^{\mathsf{h}})$:

$$\mathscr{B}_{\Phi}(u,v) = \mathscr{L}_{\Phi}(v) \qquad \forall v \in H^1(\mathcal{P}^{\mathsf{h}}) \,,$$

where

$$\begin{aligned} \mathscr{B}_{\Phi}(u,v) := {} & \sum_{K \in \mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}u}{\mathrm{d}x} \frac{\mathrm{d}v}{\mathrm{d}x} \, \mathrm{d}x + \sum_{e \in \Gamma_{\mathcal{I}}} \Big(\alpha [\![u]\!][\![v]\!]/h + [\![u]\!] \Phi(v) + \Phi(u) [\![v]\!] \Big)_e \\ & + \sum_{e \in \Gamma_{\mathcal{D}}} \Big(\alpha uv/h + u\Phi(v) + \Phi(u) v \Big)_e \,, \end{aligned} \tag{5.6a}$$

$$\begin{aligned} \mathscr{L}_{\Phi}(v) := {} & \int_\Omega f v \, \mathrm{d}x + \sum_{e \in \Gamma_{\mathcal{I}}} \Big(\bar{\Phi}(f) [\![v]\!] \Big)_e \\ & + \sum_{e \in \Gamma_{\mathcal{D}}} \Big(\alpha g_{\mathcal{D}} v/h + g_{\mathcal{D}} \Phi(v) + \bar{\Phi}(f) v \Big)_e + \sum_{e \in \Gamma_{\mathcal{N}}} \big(g_{\mathcal{N}} v \big)_e \,. \end{aligned} \tag{5.6b}$$

Note that in a composition of terms with a subscript $(\cdot)_e$, we adhere to the standing notational convention that the subscript of the individual terms is suppressed and appended to the enclosing parenthesis instead.

Let us allude to the fact that the edge terms involving Φ and $\bar{\Phi}$ are nonconventional. The parameters $\alpha_e \in \mathbb{R}$ ($e \in \Gamma_{\mathcal{I}} \cup \Gamma_{\mathcal{D}}$) are associated with conventional edge terms, viz., jumps of u and v at edge e. The rationale for adding these terms is elucidated by the coercivity analysis in section 5.3. The local mesh parameter h_e can in principle be selected in a similar manner as in conventional DG formulations; cf. (4.3). In what follows, we stipulate only that $h_e \leq \frac{1}{2} \sum_{K \in \mathcal{K}_e} h_K$.

5.2. Continuity properties of $\mathscr{B}_\Phi$ and $\mathscr{L}_\Phi$. To facilitate the ensuing analysis, we equip $H^1(\mathcal{P}^\mathsf{h})$ with the energy norm $\|\!|\cdot|\!\|$ according to

$$\|\!|u|\!\|^2 := \sum_{K\in\mathcal{P}^\mathsf{h}} |u|^2_{1,K} + \sum_{e\in\Gamma_\mathcal{I}} \big(\llbracket u\rrbracket^2/h\big)_e + \sum_{e\in\Gamma_\mathcal{D}} \big(u^2/h\big)_e \,.$$

The norm $\|\!|\cdot|\!\|$ is equivalent to $\|\cdot\|_{H^1(\mathcal{P}^\mathsf{h})}$. We then have the following proposition.

PROPOSITION 8 (continuity of $\mathscr{B}_\Phi$). *The bilinear form $\mathscr{B}_\Phi(\cdot,\cdot)$ given in* (5.6a) *is continuous on $H^1(\mathcal{P}^\mathsf{h})$, i.e.,*

$$|\mathscr{B}_\Phi(u,v)| \le c_b \|\!|u|\!\|\|\!|v|\!\| \qquad \forall u,v\in H^1(\mathcal{P}^\mathsf{h}) \,,$$

with, in particular, continuity constant $c_b = \max\big\{2, 1+\max_{e\in\Gamma_\mathcal{I}\cup\Gamma_\mathcal{D}} \alpha_e\big\}$.

Proof. First note that

$$|\mathscr{B}_\Phi(u,v)| \le \sum_{K\in\mathcal{P}^\mathsf{h}} \int_K \Big|\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\Big|\,\mathrm{d}x + \sum_{e\in\Gamma_\mathcal{I}} \Big(\alpha\big|\llbracket u\rrbracket\llbracket v\rrbracket\big|/h + \big|\llbracket u\rrbracket\Phi(v)\big| + \big|\Phi(u)\llbracket v\rrbracket\big|\Big)_e + \sum_{e\in\Gamma_\mathcal{D}} \Big(\alpha\big|uv\big|/h + \big|u\Phi(v)\big| + \big|\Phi(u)v\big|\Big)_e \,.$$

Application of the Schwarz inequality to the first term and subsequent application of the discrete Schwarz inequality yield

$$|\mathscr{B}_\Phi(u,v)| \le \Bigg(\sum_{K\in\mathcal{P}^\mathsf{h}} |u|^2_{1,K} + \sum_{e\in\Gamma_\mathcal{I}} \Big((1+\alpha)\llbracket u\rrbracket^2/h + h\,\Phi(u)^2\Big)_e + \sum_{e\in\Gamma_\mathcal{D}} \Big((1+\alpha)u^2/h + h\,\Phi(u)^2\Big)_e\Bigg)^{1/2} \Big(\cdots\Big)^{1/2} \,,$$

where the dots $(\cdots)$ represent an identical term with u replaced by v. Moreover, by consecutively applying the inequality

$$\big(\theta x + (1-\theta)y\big)^2 \le \theta x^2 + (1-\theta)y^2 \,, \qquad x,y\in\mathbb{R}\,,\ 0\le\theta\le 1\,,$$

the Schwarz inequality, the identity $|\phi_{K,e}|^2_{1,K} = 1/h_K$ for the $|\cdot|_{1,K}$ norm of $\phi_{K,e}$, definition (5.3), and $h_e \le \frac{1}{2}\sum_{K\in\mathcal{K}_e} h_K$, we derive the following important inequality:

$$\Phi_e(u)^2 = \Big(\sum_{K\in\mathcal{K}_e} \theta_{K,e}\int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}\phi_{K,e}}{\mathrm{d}x}\,\mathrm{d}x\Big)^2 \le \sum_{K\in\mathcal{K}_e} \theta_{K,e}\Big(\int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}\phi_{K,e}}{\mathrm{d}x}\,\mathrm{d}x\Big)^2 \le \sum_{K\in\mathcal{K}_e} \theta_{K,e}|u|^2_{1,K}|\phi_{K,e}|^2_{1,K} \le \sum_{K\in\mathcal{K}_e}\Big(\sum_{J\in\mathcal{K}_e} h_J\Big)^{-1}|u|^2_{1,K} \le \frac{1}{2h_e}\sum_{K\in\mathcal{K}_e}|u|^2_{1,K} \,. \tag{5.7}$$

Therefore,

$$|\mathscr{B}_\Phi(u,v)| \le \left(\sum_{K\in\mathcal{P}^\mathsf{h}} |u|_{1,K}^2 + \sum_{e\in\Gamma_\mathcal{I}} \left((1+\alpha)[\![u]\!]^2/h + \tfrac{1}{2}\textstyle\sum_{K\in\mathcal{K}_e} |u|_{1,K}^2 \right)_e \right.$$
$$\left. + \sum_{e\in\Gamma_\mathcal{D}} \left((1+\alpha)u^2/h + \tfrac{1}{2}\textstyle\sum_{K\in\mathcal{K}_e} |u|_{1,K}^2 \right)_e \right)^{1/2} \left(\cdots \right)^{1/2}$$
$$\le \left(2 \sum_{K\in\mathcal{P}^\mathsf{h}} |u|_{1,K}^2 + \sum_{e\in\Gamma_\mathcal{I}} \left((1+\alpha)[\![u]\!]^2/h \right)_e \right.$$
$$\left. + \sum_{e\in\Gamma_\mathcal{D}} \left((1+\alpha)u^2/h \right)_e \right)^{1/2} \left(\cdots \right)^{1/2} . \qquad \square$$

Before addressing the continuity of the linear form $\mathscr{L}_\Phi(\cdot)$, we introduce a function splitting in $H^1(\mathcal{P}^\mathsf{h})$. For any $v \in H^1(\mathcal{P}^\mathsf{h})$, we define its *discontinuous part* $v^\mathrm{d} := v^\mathrm{d}(v) \in H^1(\mathcal{P}^\mathsf{h})$ as

$$v^\mathrm{d} = \sum_{e\in\Gamma_\mathcal{I}} \left([\![v]\!]_e \textstyle\sum_{K\in\mathcal{K}_e} \theta_{K,e} \mathfrak{E}_K(-\phi_{K,e}) \right) + \sum_{e\in\Gamma_\mathcal{D}} \left(v_e \textstyle\sum_{K\in\mathcal{K}_e} \theta_{K,e} \mathfrak{E}_K(-\phi_{K,e}) \right) ,$$

where we have introduced the trivial-extension operators $\mathfrak{E}_K : H^1(K) \to H^1(\mathcal{P}^\mathsf{h})$,

$$\mathfrak{E}_K(\phi) = \begin{cases} \phi & \text{in } K , \\ 0 & \text{in } \Omega \setminus K . \end{cases}$$

Note that v^d is an elementwise linear function. The *continuous part* $v^\mathrm{c} := v^\mathrm{c}(v) \in H^1_{0,\mathcal{D}}(\Omega)$ is now defined as the completion of the splitting

$$v^\mathrm{c} = v - v^\mathrm{d} \qquad \forall v \in H^1(\mathcal{P}^\mathsf{h}) . \tag{5.8}$$

To corroborate that v^d and v^c indeed represent the discontinuous and the continuous parts of v, respectively, we note that

$$[\![v^\mathrm{d}]\!]_e = [\![v]\!]_e , \qquad [\![v^\mathrm{c}]\!]_e = 0 \qquad \forall e \in \Gamma_\mathcal{I} , \tag{5.9a}$$
$$v_e^\mathrm{d} = v_e , \qquad v_e^\mathrm{c} = 0 \qquad \forall e \in \Gamma_\mathcal{D} , \tag{5.9b}$$
$$v_e^\mathrm{d} = 0 , \qquad v_e^\mathrm{c} = v_e \qquad \forall e \in \Gamma_\mathcal{N} . \tag{5.9c}$$

In Figure 3, we illustrate for an example function v the corresponding v^d and v^c.

PROPOSITION 9 (continuity of $\mathscr{L}_\Phi$). *For $f \in [H^1_{0,\mathcal{D}}(\Omega)]'$, the linear functional $\mathscr{L}_\Phi(\cdot)$ in* (5.6b) *is continuous on $H^1(\mathcal{P}^\mathsf{h})$.*

Proof. First note that

$$\sum_{e\in\Gamma_\mathcal{I}} \left(\bar{\Phi}(f)[\![v]\!] \right)_e + \sum_{e\in\Gamma_\mathcal{D}} \left(\bar{\Phi}(f)v \right)_e = \sum_{e\in\Gamma_\mathcal{I}} \left([\![v]\!]_e \textstyle\sum_{K\in\mathcal{K}_e} \theta_{K,e} \int_K f\phi_{K,e} \,\mathrm{d}x \right)$$
$$+ \sum_{e\in\Gamma_\mathcal{D}} \left(v_e \textstyle\sum_{K\in\mathcal{K}_e} \theta_{K,e} \int_K f\phi_{K,e} \,\mathrm{d}x \right) = \int_\Omega f(-v^\mathrm{d}) \,\mathrm{d}x . \tag{5.10}$$

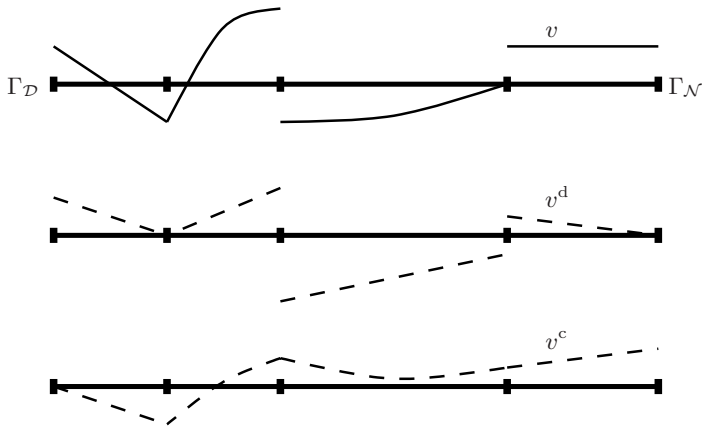


FIG. 3. *Illustration of v^{d} and v^{c} for an example function $v \in H^1(\mathcal{P}^{\mathsf{h}})$ on a domain for which the left boundary is $\Gamma_{\mathcal{D}}$ and the right boundary is $\Gamma_{\mathcal{N}}$.*

As $v - v^{\mathrm{d}} = v^{\mathrm{c}}$, we obtain for $\mathscr{L}_{\Phi}(v)$

$$\mathscr{L}_{\Phi}(v) = \int_{\Omega} f v^{\mathrm{c}} \,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(\alpha g_{\mathcal{D}} v/h + g_{\mathcal{D}}\Phi(v)\Big)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \big(g_{\mathcal{N}} v\big)_e \,, \tag{5.11}$$

which can be bounded as follows

$$|\mathscr{L}_{\Phi}(v)| \le \left| \int_{\Omega} f v^{\mathrm{c}} \,\mathrm{d}x \right| + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(|g_{\mathcal{D}}|\big(\alpha|v|/h + |\Phi(v)|\big)\Big)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \big(|g_{\mathcal{N}}||v|\big)_e.$$

Since $v^{\mathrm{c}} \in H^1_{0,\mathcal{D}}(\Omega)$, the first term is bounded for $f \in [H^1_{0,\mathcal{D}}(\Omega)]'$. The other terms can also be bounded using (5.7) and the usual trace inequalities. ∎

5.3. Well-posedness results for the continuum formulation. At variance with conventional DG formulations, the new DG formulation is consistent with the more general Poisson problem (2.1).

PROPOSITION 10 (consistency with classical CG formulation). *For all f in the dual space $[H^1_{0,\mathcal{D}}(\Omega)]'$, the DG formulation* (5.5) *is consistent with* (2.1)*, i.e., if $u \in H^1(\Omega)$ is the solution of* (2.1)*, then u satisfies* (5.5).

Proof. Let $u \in H^1(\Omega)$ solve (2.1) and let v be an arbitrary function in $H^1(\mathcal{P}^{\mathsf{h}})$. On account of $[\![u]\!]_e = 0$ for $e \in \Gamma_{\mathcal{I}}$ and $u = g_{\mathcal{D}}$ on $\Gamma_{\mathcal{D}}$, it follows from (5.6a) that

$$\mathscr{B}_{\Phi}(u,v) = \sum_{K\in\mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x} \,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{I}}} \Big(\Phi(u)[\![v]\!]\Big)_e + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(\alpha g_{\mathcal{D}} v/h + g_{\mathcal{D}}\Phi(v) + \Phi(u)v\Big)_e \,.$$

Moreover, in analogy with (5.10), it holds that

$$\sum_{e\in\Gamma_{\mathcal{I}}} \Big(\Phi(u)[\![v]\!]\Big)_e + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(\Phi(u)v\Big)_e = \sum_{K\in\mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}(-v^{\mathrm{d}})}{\mathrm{d}x} \,\mathrm{d}x \quad \forall u, v \in H^1(\mathcal{P}^{\mathsf{h}}) \,. \tag{5.12}$$

As $v - v^{\mathrm{d}} = v^{\mathrm{c}}$, we obtain

$$\mathscr{B}_{\Phi}(u,v) = \int_{\Omega} \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v^{\mathrm{c}}}{\mathrm{d}x}\,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(\alpha g_{\mathcal{D}} v/h + g_{\mathcal{D}}\Phi(v)\Big)_e\,,$$

where the sum of integrals is replaced by an integral over Ω, which is admissible because $u \in H^1(\Omega)$ and $v^{\mathrm{c}} \in H^1_{0,\mathcal{D}}(\Omega)$. Recalling from (2.1) that

$$\int_{\Omega} \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v^{\mathrm{c}}}{\mathrm{d}x}\,\mathrm{d}x = \int_{\Omega} f v^{\mathrm{c}}\,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{N}}} \big(g_{\mathcal{N}} v^{\mathrm{c}}\big)_e\,,$$

we finally obtain from (5.9c) that

$$\mathscr{B}_{\Phi}(u,v) = \int_{\Omega} f v^{\mathrm{c}}\,\mathrm{d}x + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(\alpha g_{\mathcal{D}} v/h + g_{\mathcal{D}}\Phi(v)\Big)_e + \sum_{e\in\Gamma_{\mathcal{N}}} \big(g_{\mathcal{N}} v\big)_e\,.$$

Hence, $\mathscr{B}_{\Phi}(u,v)$ can be identified with $\mathscr{L}_{\Phi}(v)$ according to (5.11) for all $v \in H^1(\mathcal{P}^{\mathsf{h}})$. ∎

We remark that consistency can be established for any choice of $\theta_{K,e}$ in the operators Φ and $\bar{\Phi}$ in (5.2), provided that the partition-of-unity property (5.4) holds.

A fundamental property of the bilinear form $\mathscr{B}_{\Phi}(\cdot,\cdot)$ in (5.6a) is its *coercivity* on $H^1(\mathcal{P}^{\mathsf{h}})$.

PROPOSITION 11 (coercivity of $\mathscr{B}_{\Phi}$). *If the parameter $\alpha_e > 1$ for all $e \in \Gamma_{\mathcal{I}} \cup \Gamma_{\mathcal{D}}$, then the bilinear form $\mathscr{B}_{\Phi}(\cdot,\cdot)$ in* (5.6a) *is coercive on $H^1(\mathcal{P}^{\mathsf{h}})$, i.e.,*

$$|\mathscr{B}_{\Phi}(u,u)| \geq \kappa |\!|\!| u |\!|\!|^2 \qquad \forall u \in H^1(\mathcal{P}^{\mathsf{h}})\,,$$

with, in particular, coercivity constant

$$\kappa = \min_{e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}} \tfrac{1}{2}\Big((\alpha_e - 1) + 2 - \sqrt{(\alpha_e-1)^2+4}\Big) \in (0,1). \tag{5.13}$$

Note that α_e can be chosen such that κ in (5.13) is bounded away from 0.

Proof. Consider an arbitrary $u \in H^1(\mathcal{P}^{\mathsf{h}})$. We show that there exists a κ in the interval $0 < \kappa < 1$ such that $\mathscr{B}_{\Phi}(u,u) - \kappa |\!|\!| u |\!|\!|^2 \geq 0$. First, we observe that

$$\begin{aligned}\mathscr{B}_{\Phi}(u,u) - \kappa |\!|\!| u |\!|\!|^2 = (1-\kappa) \sum_{K\in\mathcal{P}^{\mathsf{h}}} |u|^2_{1,K} + \sum_{e\in\Gamma_{\mathcal{I}}} \Big((\alpha-\kappa)[\![u]\!]^2/h + 2[\![u]\!]\Phi(u)\Big)_e \\ + \sum_{e\in\Gamma_{\mathcal{D}}} \Big((\alpha-\kappa)u^2/h + 2u\Phi(u)\Big)_e\,.\end{aligned}$$

Application of the Young inequality yields

$$\begin{aligned}\mathscr{B}_{\Phi}(u,u) - \kappa |\!|\!| u |\!|\!|^2 \geq (1-\kappa) \sum_{K\in\mathcal{P}^{\mathsf{h}}} |u|^2_{1,K} \\ + \sum_{e\in\Gamma_{\mathcal{I}}} \left((\alpha-\kappa)[\![u]\!]^2/h - \frac{[\![u]\!]^2}{(1-\kappa)h} - (1-\kappa)h\,\Phi(u)^2\right)_e \\ + \sum_{e\in\Gamma_{\mathcal{D}}} \left((\alpha-\kappa)u^2/h \quad - \frac{u^2}{(1-\kappa)h} - (1-\kappa)h\,\Phi(u)^2\right)_e\,.\end{aligned}$$

We now invoke (5.7) to obtain

$$\begin{aligned}\mathscr{B}_{\Phi}(u,u)-\kappa|\!|\!|u|\!|\!|^2 \geq (1-\kappa)\sum_{K\in\mathcal{P}^{\mathsf{h}}}|u|^2_{1,K} \\ +\sum_{e\in\Gamma_{\mathcal{I}}}\left(\left(\alpha-\kappa-\tfrac{1}{1-\kappa}\right)[\![u]\!]^2/h-\tfrac{1}{2}(1-\kappa)\textstyle\sum_{K\in\mathcal{K}_e}|u|^2_{1,K}\right)_e \\ +\sum_{e\in\Gamma_{\mathcal{D}}}\left(\left(\alpha-\kappa-\tfrac{1}{1-\kappa}\right)u^2/h\quad-\tfrac{1}{2}(1-\kappa)\textstyle\sum_{K\in\mathcal{K}_e}|u|^2_{1,K}\right)_e .\end{aligned}$$

The summations over the elements cancel, except for the contributions of elements contiguous to Neumann boundaries, and, hence,

(5.14)
$$\begin{aligned}\mathscr{B}_{\Phi}(u,u)-\kappa|\!|\!|u|\!|\!|^2 \geq \sum_{e\in\Gamma_{\mathcal{I}}}\left(\left(\alpha-\kappa-\tfrac{1}{1-\kappa}\right)[\![u]\!]^2/h\right)_e+\sum_{e\in\Gamma_{\mathcal{D}}}\left(\left(\alpha-\kappa-\tfrac{1}{1-\kappa}\right)u^2/h\right)_e \\ +\tfrac{1}{2}(1-\kappa)\sum_{e\in\Gamma_{\mathcal{N}}}\sum_{K\in\mathcal{K}_e}|u|^2_{1,K} .\end{aligned}$$

If $\alpha_e>1$ for all $e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}$ and κ complies with (5.13), then $\alpha_e-\kappa-\frac{1}{1-\kappa}\geq 0$ for all $e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}$. Furthermore, for $0<\kappa<1$ the final term in the right member of (5.14) is nonnegative and, therefore, $\mathscr{B}_{\Phi}(u,u)-\kappa|\!|\!|u|\!|\!|^2\geq 0$. ∎

By the classical Lax–Milgram theorem, Theorem 2, we can now conclude that if $\alpha_e>1$ for all $e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}$, then for all $f\in[H^1_{0,\mathcal{D}}(\Omega)]'$ the new DG formulation (5.5) is well-posed and consistent with (2.1). Moreover, by virtue of the coercivity of the new DG formulation, conforming approximations in $H^1(\mathcal{P}^{\mathsf{h}})$ inherit their well-posedness from the continuum formulation, and optimal error estimates hold with uniformly bounded constants.

6. Numerical experiments. In this section we present numerical results for the new DG formulation. First, we investigate the sharpness of the estimate of the coercivity constant (5.13) by means of discrete inf-sup calculations. Next, we illustrate the optimal convergence behavior of the new formulation in appropriate norms.

6.1. Discrete inf-sup calculations. The estimate of the coercivity constant κ in (5.13) represents a lower bound. That is, a $\bar{\kappa}>\kappa$ possibly exists such that $|\mathscr{B}_{\Phi}(u,u)|\geq\bar{\kappa}|\!|\!|u|\!|\!|^2$ for all $u\in H^1(\mathcal{P}^{\mathsf{h}})$. An upper bound to the coercivity constant can be determined by establishing the discrete coercivity constant, viz., the coercivity constant in a finite-dimensional subspace $\widehat{H}\subset H^1(\mathcal{P}^{\mathsf{h}})$, according to

$$\widehat{\kappa}:=\widehat{\kappa}(\widehat{H})=\inf_{u\in\widehat{H}\setminus\{0\}}\frac{\mathscr{B}_{\Phi}(u,u)}{|\!|\!|u|\!|\!|^2} .$$

For a symmetric bilinear form on a finite-dimensional subspace $\widehat{H}$, the coercivity constant coincides with the discrete inf-sup constant

$$\widehat{\gamma}:=\widehat{\gamma}(\widehat{H})=\inf_{u\in\widehat{H}\setminus\{0\}}\sup_{v\in\widehat{H}\setminus\{0\}}\frac{\mathscr{B}_{\Phi}(u,v)}{|\!|\!|u|\!|\!|\;|\!|\!|v|\!|\!|} ,$$

which can be determined numerically by means of the procedure in [10]. Note that the discrete coercivity constants pertaining to a sequence of nested subspaces $\widehat{H}^{(1)}\subset$

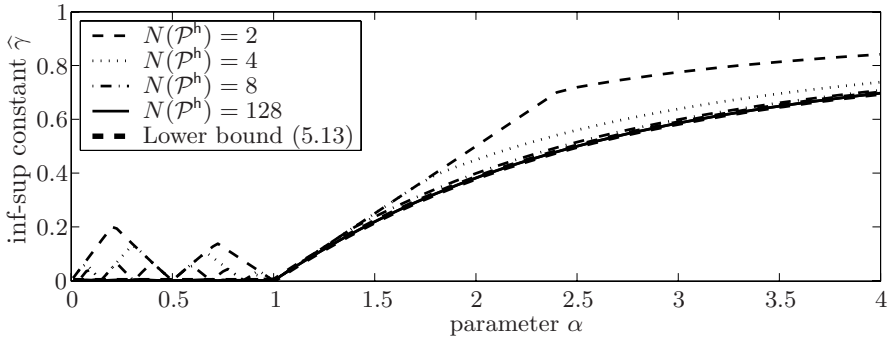


FIG. 4. *Discrete inf-sup constant $\widehat{\gamma}$ versus the parameter α for broken polynomial spaces on uniform partitions with $N(\mathcal{P}^{\mathsf{h}})$ elements. The inf-sup constant $\widehat{\gamma}$ is p independent.*

$\widehat{H}^{(2)} \subset \cdots \subseteq H^1(\mathcal{P}^{\mathsf{h}})$ form a nonincreasing sequence $\widehat{\kappa}^{(1)} \geq \widehat{\kappa}^{(2)} \geq \cdots \geq \bar{\kappa}$. Hence, the discrepancy between the discrete inf-sup constants corresponding to a sequence of nested subspaces and the estimate (5.13) provides a measure of the sharpness of the estimate.

To assess the sharpness of (5.13), we compute the discrete inf-sup constant of the bilinear form in the new DG formulation (5.5) for the Poisson problem on the open unit domain $\Omega = (0,1)$ with Dirichlet boundary conditions, i.e., $\partial\Omega = \Gamma_{\mathcal{D}} = \{0,1\}$. We restrict ourselves to uniform partitions $\mathcal{P}^{\mathsf{h}}$ of $N(\mathcal{P}^{\mathsf{h}})$ elements, and finite-dimensional approximation spaces consisting of broken polynomials with a uniform distribution of the polynomial degree p:

$$\widehat{H} = \mathbb{P}^{\mathsf{p}}(\mathcal{P}^{\mathsf{h}}) := \left\{ u \in L_2(\Omega) : u|_K \in \mathbb{P}^{\mathsf{p}}(K) \; \forall K \in \mathcal{P}^{\mathsf{h}} \right\}.$$

Moreover, we use a uniform distribution of the parameter α_e (=: α) for all $e \in \Gamma_{\mathcal{D}} \cup \Gamma_{\mathcal{I}}$.

The results are displayed in Figure 4. In addition, the figure plots the lower bound κ according to (5.13). The numerical results convey that the computed inf-sup constants are independent of the polynomial degree p (results not displayed). They do, however, depend on $N(\mathcal{P}^{\mathsf{h}})$ and α. It appears that for large $N(\mathcal{P}^{\mathsf{h}})$ the discrete inf-sup constants indeed converge to the lower bound and, hence, the estimate of the coercivity constant κ in (5.13) is apparently sharp.

6.2. Error convergence behavior. We consider the new DG formulation for the Poisson problem (5.5) on the open unit domain $\Omega = (0,1)$ with homogeneous Dirichlet boundary conditions. The prescribed data f is selected such that the solution is $u(x) = \sin(\pi x)$. We consider uniform partitions $\mathcal{P}^{\mathsf{h}}$ with $N(\mathcal{P}^{\mathsf{h}})$ elements. The approximation spaces $\widehat{H}$ are the same as used in the inf-sup calculations above.

Figure 5 plots the error in the approximations. The figure indicates that the approximate solutions $\widehat{u}$ are pointwise exact at the interior and boundary edges. This behavior is characteristic for the classical CG method (2.4). Similarly, it can be proven that if $\mathbb{P}^1(\mathcal{P}^{\mathsf{h}}) \subset \widehat{H}$, i.e., if the approximation space contains the piecewise linear functions, then the DG approximation exhibits the same behavior. In Appendix B we elaborate the pointwise exactness for approximations to the new DG formulation (5.5). In particular, the pointwise exactness implies that

$$\widehat{u} \in \widehat{H} \cap \{ u \in H^1(\Omega) : u = g_{\mathcal{D}} \text{ on } \Gamma_{\mathcal{D}} \}. \tag{6.1}$$

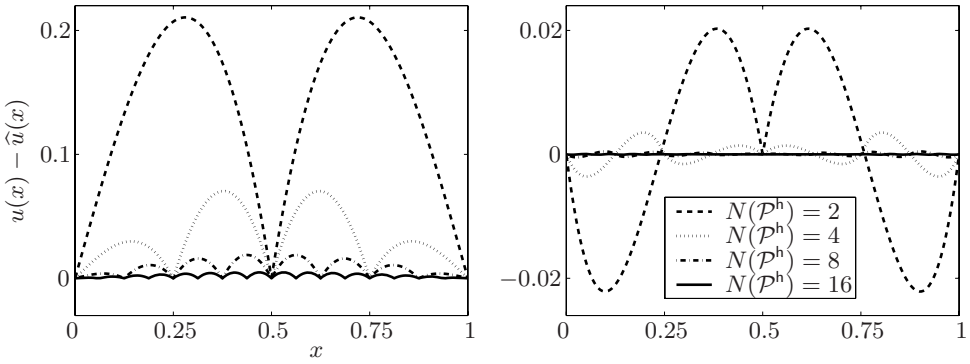


FIG. 5. *Pointwise error for broken polynomial spaces of order* $\mathsf{p} = 1$ *(left) and* $\mathsf{p} = 2$ *(right) on uniform partitions with* $N(\mathcal{P}^{\mathsf{h}})$ *elements.*

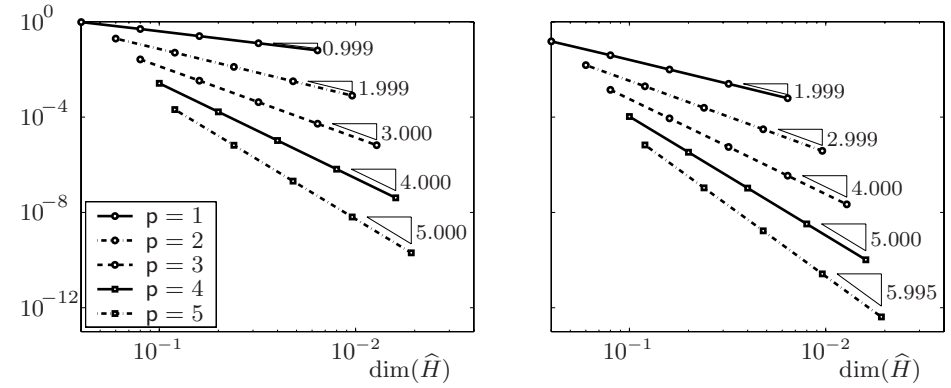


FIG. 6. *Error in the energy-norm (left),* $|\!|\!| u - \hat{u} |\!|\!|$*, and in the* $L^2(\Omega)$*-norm (right),* $\|u - \hat{u}\|_{L_2(\Omega)}$*, versus the dimension of the approximation space* $\dim(\widehat{H})$ *for broken polynomial spaces of order* $\mathsf{p} = 1, \ldots, 5$ *on uniform partitions.*

Moreover, $\widehat{u}$ is then identical to the approximate solution of the classical CG formulation (2.4) on $\widehat{H}^1_{0,\mathcal{D}}(\Omega) = \widehat{H} \cap H^1_{0,\mathcal{D}}(\Omega)$ with $\bar{u} \in \mathbb{P}^1(\Omega)$. Another implication is that the approximations are independent of the parameters α_e (provided that $\alpha_e > 1$ so that the approximate problem is well posed), because the terms associated with α_e vanish from the formulation; cf. Eqs. (5.5) and (5.6).

Figure 6 plots the energy-norm and the $L_2(\Omega)$-norm of the error versus the dimension of the approximation space, $\dim(\widehat{H}) := (\mathsf{p}+1)N(\mathcal{P}^{\mathsf{h}})$, for polynomial orders $\mathsf{p} = 1, 2, \ldots, 5$. The figures corroborate the optimal convergence behavior of the new DG formulation in both norms.

7. Conclusions. We established on the basis of the prototypical Poisson problem that most concurrent DG finite-element methods for second-order elliptic differential equations can be condensed into a generic conventional DG formulation. By means of this generic formulation, we showed that a coercive conventional DG formulation is nonexistent. Conventional DG formulations are contingent on weak coercivity for their well-posedness. However, as weak coercivity does not transfer to subspaces, well-posedness of the continuum problem does not generally imply well-posedness of approximate problems based on conforming subspaces.

We then presented a new nonconventional symmetric DG formulation that is coercive on the broken Sobolev space $H^1(\mathcal{P}^{\mathsf{h}})$. The new formulation is based on element Green's functions and the data local to the edges. On account of its coercivity, conforming approximations of the new formulation inherit their well-posedness from the continuum formulation, and optimal error estimates hold with approximation-space-independent constants. Furthermore, the new DG formulation is consistent with the classical CG formulation in that it admits solutions in $H^1(\mathcal{P}^{\mathsf{h}}) \supset H^1(\Omega)$, rather than $H^2(\mathcal{P}^{\mathsf{h}})$ which is common for conventional DG formulations.

We derived a lower bound for the coercivity constant of the bilinear form in the new formulation. The sharpness of this estimate was confirmed by means of numerical computations of discrete inf-sup constants. Furthermore, numerical experiments were conducted to assess the convergence behavior of the new formulation. The results corroborate that the formulation yields optimal convergence in the energy-norm and in the $L^2(\Omega)$-norm. Moreover, the results demonstrate that discrete approximations in subspaces that contain the piecewise linear functions are identical to classical CG approximations.

It is anticipated that the main attributes of the proposed DG formulation can be extended to higher-dimensional settings. Essentially, the Green's function provides a decomposition of the broken space into the continuous functions and their orthogonal complement. This decomposition can be used to construct a bilinear form that is both consistent and coercive. The generalization of the Green's function to higher dimensions is complex, but there is no fundamental obstacle that precludes such a generalization.

Appendix A. Proof of Proposition 6. The proof is supported by the following lemma.

LEMMA 12. *If there exist a linear continuous operator $v_{(\cdot)} : H_\Lambda \to H_\Lambda$ dependent only on the edge values $\boldsymbol{u}_e$, and a constant $c_1 > \frac{1}{2}$, such that*

$$\big(\Lambda(\boldsymbol{u} + \boldsymbol{v_u})\big)_e = c_1 \big(\mathrm{D}_\Lambda \boldsymbol{u}\big)_e \,, \tag{A.1a}$$

$$\tfrac{1}{2} \sum_{K\in\mathcal{P}^{\mathsf{h}}} |v_u|^2_{1,K} \le \sum_{e\in\Gamma} \big(\boldsymbol{u}^{\mathsf{T}} \mathrm{D}_\Lambda \boldsymbol{u}\big)_e \,, \tag{A.1b}$$

$$\|v_u\|_{H_\Lambda} \le c_\Lambda \|u\|_{H_\Lambda} \tag{A.1c}$$

for all $e \in \Gamma$, then $\mathscr{B}_\Lambda(\cdot,\cdot)$ satisfies the inf-sup condition on $H_\Lambda \times H_\Lambda$.

Note that (A.1c) just expresses the continuity of the operator $v_{(\cdot)}$. Bold-faced variables and the matrix D_Λ are defined in (4.2) and (4.5), respectively.

Proof. By the Young inequality and (A.1a) and (A.1b) it holds that

$$\begin{aligned}
\mathscr{B}_\Lambda(u, u+v_u) &= \sum_{K\in\mathcal{P}^{\mathsf{h}}} \Big(|u|^2_{1,K} + \textstyle\int_K u' v_u' \,\mathrm{d}x \Big) + \sum_{e\in\Gamma} \big(\boldsymbol{u}^{\mathsf{T}} \Lambda(\boldsymbol{u} + \boldsymbol{v_u})\big)_e \\
&\ge \sum_{K\in\mathcal{P}^{\mathsf{h}}} \Big((1 - \tfrac{1}{2\epsilon}) |u|^2_{1,K} - \tfrac{\epsilon}{2} |v_u|^2_{1,K} \Big) + c_1 \sum_{e\in\Gamma} \big(\boldsymbol{u}^{\mathsf{T}} \mathrm{D}_\Lambda \boldsymbol{u}\big)_e \\
&\ge (1 - \tfrac{1}{2\epsilon}) \sum_{K\in\mathcal{P}^{\mathsf{h}}} |u|^2_{1,K} + (c_1 - \epsilon) \sum_{e\in\Gamma} \big(\boldsymbol{u}^{\mathsf{T}} \mathrm{D}_\Lambda \boldsymbol{u}\big)_e
\end{aligned}$$

for all $\epsilon > 0$. Recalling the definition of $\|\cdot\|_{H_\Lambda}$ according to (4.4), we note that for all $c_1 > \frac{1}{2}$ there exists an $\epsilon > \frac{1}{2}$ such that $\mathscr{B}_\Lambda(u, u+v_u) \ge (c_1 - \epsilon)\|u\|^2_{H_\Lambda} > 0$. Using

this in the inf-sup condition, we obtain

$$\sup_{v\in H_\Lambda\setminus\{0\}} \frac{\mathscr{B}_\Lambda(u,v)}{\|u\|_{H_\Lambda}\|v\|_{H_\Lambda}} \geq \frac{\mathscr{B}_\Lambda(u,u+v_u)}{\|u\|_{H_\Lambda}\|u+v_u\|_{H_\Lambda}} \geq \frac{(c_1-\epsilon)\|u\|^2_{H_\Lambda}}{\|u\|_{H_\Lambda}\big(\|u\|_{H_\Lambda}+\|v_u\|_{H_\Lambda}\big)}$$
$$\geq \frac{c_1-\epsilon}{1+c_\Lambda} > 0\,. \qquad \square$$

To prove that the inf-sup condition (3.1a) holds, we establish that under the conditions (4.13) there exists an operator $v_{(\cdot)} : H_\Lambda \to H_\Lambda$ in compliance with the premises of Lemma 12. The existence is verified by construction. Simple linear algebra conveys that if and only if the parameters in the matrices Λ_e in (4.9) satisfy

$$\left.\begin{aligned} \delta\beta-\varepsilon\gamma^{\mathrm{u}}\neq 0 \quad \text{or} \quad \beta,\gamma^{\mathrm{u}},\gamma^{\mathrm{l}},\delta,\varepsilon=0 \\ \zeta^1,\zeta^2=0 \end{aligned}\right\} \quad \forall e\in\Gamma_{\mathcal{I}}\,, \tag{A.2a}$$

$$\delta\neq 0 \qquad \forall e\in\Gamma_{\mathcal{D}}\,, \tag{A.2b}$$

$$\varepsilon=0 \qquad \forall e\in\Gamma_{\mathcal{N}}\,, \tag{A.2c}$$

then for each $u\in H_\Lambda$, (A.1a) admits a (nonunique) solution $(\boldsymbol{v}_u)_e$ for any $c_1\in\mathbb{R}$. Thus, (A.1a) yields the values of v_u at the edges $e\in\Gamma$. The kernel of the matrix Λ in (A.1a) accommodates arbitrary $\{v_u\}_e$ for $e\in\Gamma_{\mathcal{I}}$ and arbitrary $(v_u)_e$ for $e\in\Gamma_{\mathcal{N}}$. We set $(v_u)_e=0$ for $e\in\Gamma_{\mathcal{N}}$.

To facilitate the proof, we introduce an auxiliary operator $\bar{v}_{(\cdot)}$ from H_Λ to $\mathbb{P}^1(\mathcal{P}^{\mathsf{h}})$, viz., the space of piecewise linear functions on the partition $\mathcal{P}^{\mathsf{h}}$. The operator $\bar{v}_{(\cdot)}$ associates with each $u\in H_\Lambda$ the function $\bar{v}_u\in\mathbb{P}^1(\mathcal{P}^{\mathsf{h}})$ such that

$$\begin{aligned} [\![\bar{v}_u]\!]_e &= [\![v_u]\!]_e && \forall e\in\Gamma_{\mathcal{I}}\,,\\ \big(\bar{v}_u\big)_e &= \big(v_u\big)_e && \forall e\in\Gamma_{\mathcal{D}}\,,\\ \big(\bar{v}_u\big)_e &= 0 && \forall e\in\Gamma_{\mathcal{N}}\,, \end{aligned}$$

with $[\![v_u]\!]_e$ the previously determined jumps at edges. Specifically, we define $\bar{v}_u$ as

$$\lim_{K\ni x\to e}\big(\bar{v}_u\big|_K\big)(x) = \begin{cases} \frac{1}{2}n_e\,n_K h_K [\![v_u]\!]_e/h_e & \forall e\in\partial K\cap\Gamma_{\mathcal{I}}\,,\\ \frac{1}{2}h_K(v_u)_e/h_e & \forall e\in\partial K\cap\Gamma_{\mathcal{D}}\,,\\ 0 & \forall e\in\partial K\cap\Gamma_{\mathcal{N}}\,. \end{cases} \tag{A.3}$$

We can now define $v_{(\cdot)}$ as the map $u\to v_u$, where v_u is the limit of a Cauchy sequence $\{v_u^i\}$ in H_Λ with the properties

$$\begin{aligned} v_u^i\big|_K &\to \bar{v}_u\big|_K && \text{in } H^1(K) && \forall K\in\mathcal{P}^{\mathsf{h}}\,,\\ [\![\partial_n v_u^i]\!]_e &\to [\![\partial_n v_u]\!]_e && \text{in } \mathbb{R} && \forall e\in\Gamma_{\mathcal{I}}\,,\\ \{\partial_n v_u^i\}_e &\to \{\partial_n v_u\}_e && \text{in } \mathbb{R} && \forall e\in\Gamma_{\mathcal{I}}\,,\\ \big(\partial_n v_u^i\big)_e &\to \big(\partial_n v_u\big)_e && \text{in } \mathbb{R} && \forall e\in\partial\Omega\,, \end{aligned}$$

where $\{\partial_n v_u\}_e$ and $[\![\partial_n v_u]\!]_e$ refer to the previously determined average derivatives and derivative jumps at edges. Such a Cauchy sequence can be constructed in a similar manner as the sequence in the proof of Proposition 7. The operator $v_{(\cdot)}$ thus defined complies with (A.1a).

To ascertain that $v_{(\cdot)}$ satisfies (A.1b), we note that by (4.9) and (A.2), the second equation in the linear system (A.1a) yields

$$\big(h^{-\frac{1}{2}}[\![v_u]\!]\big)_e = -\big(h^{-\frac{1}{2}}[\![u]\!] + c_1(D_\Lambda)_{22}\, h^{\frac{1}{2}}\{\partial_n u\}\big)_e \qquad \forall e \in \Gamma_{\mathcal{I}}\,, \tag{A.4a}$$

$$\big(h^{-\frac{1}{2}} v_u\big)_e = -\big(h^{-\frac{1}{2}} u \quad + c_1(D_\Lambda)_{22}\, h^{\frac{1}{2}}\partial_n u \quad \big)_e \qquad \forall e \in \Gamma_{\mathcal{D}}\,, \tag{A.4b}$$

with, in particular,

$$1 \le \big((D_\Lambda)_{22}\big)_e = \begin{cases} \big(\frac{1}{2}|\delta+1| + \frac{1}{2}|\delta-1| + |\varepsilon|\big)_e & \forall e \in \Gamma_{\mathcal{I}}\,, \\ \big(\frac{1}{2}|\delta+1| + \frac{1}{2}|\delta-1|\big)_e & \forall e \in \Gamma_{\mathcal{D}}\,. \end{cases} \tag{A.5}$$

Equations (A.3) and (A.4) yield

$$\begin{aligned} \left|\frac{\mathrm{d}\bar{v}_u}{\mathrm{d}x}\Big|_K\right|^2 &= \left|\sum_{e\in\partial K} \big(\bar{v}_u|_K\, n_K\big)_e / h_K\right|^2 \le 2\sum_{e\in\partial K} \big(\bar{v}_u|_K\big)_e^2 / h_K^2 \\ &\le \sum_{e\in\partial K\cap\Gamma_{\mathcal{I}}} \tfrac{1}{2}\Big([\![u]\!]/h + c_1(\mathrm{D}_\Lambda)_{22}\{\partial_n u\}\Big)_e^2 + \sum_{e\in\partial K\cap\Gamma_{\mathcal{D}}} \tfrac{1}{2}\Big(u/h + c_1(\mathrm{D}_\Lambda)_{22}\partial_n u\Big)_e^2 . \end{aligned}$$

From the relation $h_e = \frac{1}{2}\sum_{K\in\mathcal{K}_e} h_K$ in (4.3) it follows that

$$\begin{aligned} \tfrac{1}{2}\sum_{K\in\mathcal{P}^{\mathsf{h}}} |\bar{v}_u|_{1,K}^2 &\le \tfrac{1}{2}\sum_{K\in\mathcal{P}^{\mathsf{h}}} h_K \Big(\textstyle\sum_{e\in\partial K\cap\Gamma_{\mathcal{I}}} \frac{1}{2}\Big([\![u]\!]/h + c_1(\mathrm{D}_\Lambda)_{22}\{\partial_n u\}\Big)_e^2 \\ &\qquad + \textstyle\sum_{e\in\partial K\cap\Gamma_{\mathcal{D}}} \frac{1}{2}\Big(u/h + c_1(\mathrm{D}_\Lambda)_{22}\partial_n u\Big)_e^2\Big) \\ &\le \sum_{e\in\Gamma_{\mathcal{I}}} \Big([\![u]\!]^2/h + c_1^2(\mathrm{D}_\Lambda)_{22}^2 h\{\partial_n u\}^2\Big)_e \\ &\qquad + \sum_{e\in\Gamma_{\mathcal{D}}} \Big(u^2/h + c_1^2(\mathrm{D}_\Lambda)_{22}^2 h(\partial_n u)^2\Big)_e \\ &\le \max\Big\{1, c_1^2 \max_{e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}} \big((\mathrm{D}_\Lambda)_{22}\big)_e\Big\} \sum_{e\in\Gamma} \big(\boldsymbol{u}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{u}\big)_e\,. \end{aligned}$$

Moreover, under the conditions (4.13) it holds that

$$\begin{aligned} \big(4 > \tfrac{1}{2}|\delta+1| + \tfrac{1}{2}|\delta-1| + |\varepsilon| \;\wedge\; \varepsilon \neq 0\big) \quad \text{or} \quad \big(4 > |\delta| \neq 0 \;\wedge\; \varepsilon = 0\big) \qquad & \forall e \in \Gamma_{\mathcal{I}}\,, \\ 4 > |\delta| \qquad & \forall e \in \Gamma_{\mathcal{D}}\,. \end{aligned}$$

Inequality (A.5) then yields $1 \le \max_{e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}}((\mathrm{D}_\Lambda)_{22})_e < 4$. As $v_u^i|_K \to \bar{v}_u^i|_K$ in $H^1(\mathcal{P}^{\mathsf{h}})$ as $i\to\infty$, there exists a $c_1 > \frac{1}{2}$ such that $c_1^2 \max_{e\in\Gamma_{\mathcal{I}}\cup\Gamma_{\mathcal{D}}}((\mathrm{D}_\Lambda)_{22})_e < 1$ and hence (A.1b) holds.

To establish (A.1c), we denote by Λ^- the (square) generalized inverse for the matrix Λ in equation (A.1a). We can then write

$$\boldsymbol{v_u} = \Lambda^-(c_1\mathrm{D}_\Lambda - \Lambda)\boldsymbol{u}\,,$$

and, thus,

$$\boldsymbol{v_u}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{v_u} = \big\|\mathrm{D}_\Lambda^{\frac{1}{2}}\Lambda^-(c_1\mathrm{D}_\Lambda - \Lambda)\boldsymbol{u}\big\|^2 \le \underbrace{\big\|\mathrm{D}_\Lambda^{\frac{1}{2}}\Lambda^-(c_1\mathrm{D}_\Lambda - \Lambda)\mathrm{D}_\Lambda^{-\frac{1}{2}}\big\|^2}_{=:\mathcal{C}_2} \big(\boldsymbol{u}^{\mathsf{T}}\mathrm{D}_\Lambda\boldsymbol{u}\big)\,,$$

where $\|\cdot\|$ represents the usual Euclidian vector norm, and the corresponding matrix norm. Condition (A.1c) can then be verified straightforwardly:

$$\|v_u\|_{H_\Lambda}^2 = \sum_{K\in\mathcal{P}^\mathsf{h}} |v_u|_{1,K}^2 + \sum_{e\in\Gamma} \left(\boldsymbol{v_u}^\mathsf{T} \mathrm{D}_\Lambda \boldsymbol{v_u}\right)_e \leq (2+c_2) \sum_{e\in\Gamma} \left(\boldsymbol{u}^\mathsf{T} \mathrm{D}_\Lambda \boldsymbol{u}\right)_e \leq (2+c_2)\|u\|_{H_\Lambda}^2 \ .$$

The second condition for weak coercivity of $\mathscr{B}_\Lambda(\cdot,\cdot)$, i.e.,

$$\sup_{u\in H_\Lambda} \mathscr{B}_\Lambda(u,v) > 0 \qquad \forall v \in H_\Lambda \setminus \{0\} \ ,$$

is easily established by means of the relation $\mathscr{B}_\Lambda(u,v) = \mathscr{B}_{\Lambda^\mathsf{T}}(v,u)$. Under the conditions in (4.13), we can construct an operator $u_{(\cdot)} : H_\Lambda \to H_\Lambda$, in a similar manner as the operator $v_{(\cdot)}$ above, such that

$$\sup_{u\in H_\Lambda} \mathscr{B}_{\Lambda^\mathsf{T}}(v,u) \geq \mathscr{B}_{\Lambda^\mathsf{T}}(v, v+u_v) > 0 \ .$$

Appendix B. Pointwise exactness of approximations. In this section we establish that the new DG formulation is pointwise exact on all edges $\Gamma = \Gamma_\mathcal{I} \cup \partial\Omega$ if the discrete approximation space contains the piecewise linear polynomials, i.e., $\mathbb{P}^1(\mathcal{P}^\mathsf{h}) \subseteq \widehat{H} \subset H^1(\mathcal{P}^\mathsf{h})$.

Let $\widehat{u} \in \widehat{H}$ be the solution of the approximation problem $\mathscr{B}_\Phi(\widehat{u},v) = \mathscr{L}_\Phi(v)$ for all $v \in \widehat{H}$. First, we show that the jumps of $\widehat{u}$ are zero and that the Dirichlet boundary traces comply with the Dirichlet boundary condition, i.e.,

$$\llbracket \widehat{u} \rrbracket_e = 0 \quad \forall e \in \Gamma_\mathcal{I} \ , \qquad \widehat{u} = g_\mathcal{D} \quad \text{on } \Gamma_\mathcal{D} \ . \tag{B.1}$$

Consider an arbitrary edge $\bar{e} \in \Gamma_\mathcal{I} \cup \Gamma_\mathcal{D}$. We construct a discontinuous test function $w = w(\bar{e}) \in \mathbb{P}^1(\mathcal{P}^\mathsf{h})$ such that $w^\mathrm{c} = 0$, $w^\mathrm{d} = w$ (cf. section 5.2 for the splitting $v = v^\mathrm{c} + v^\mathrm{d}$ into a continuous and a discontinuous part), and

$$\begin{aligned} \alpha\llbracket w \rrbracket/h + \Phi(w) &= 0 \qquad && \forall e \in \Gamma_\mathcal{I} \setminus \{\bar{e}\} \ , \\ \alpha w/h + \Phi(w) &= 0 && \forall e \in \Gamma_\mathcal{D} \setminus \bar{e} \ , \\ \alpha\llbracket w \rrbracket/h + \Phi(w) &= 1 && \text{if } \bar{e} \in \Gamma_\mathcal{I} \ , \\ \alpha w/h + \Phi(w) &= 1 && \text{if } \bar{e} \in \Gamma_\mathcal{D} \ . \end{aligned} \tag{B.2}$$

It can be shown that the system of equations (B.2) admits a unique solution under the (sufficient) condition $\alpha_e > 1$. This condition is satisfied by assumption; see Proposition 11. As $w \in \mathbb{P}^1(\mathcal{P}^\mathsf{h}) \subset \widehat{H}$, it holds that $\mathscr{B}_\Phi(\widehat{u},w) = \mathscr{L}_\Phi(w)$. From (5.11), (5.12), and $w = 0$ on $\Gamma_\mathcal{N}$, it follows that

$$\sum_{e\in\Gamma_\mathcal{I}} \Big(\big(\alpha\llbracket w \rrbracket/h + \Phi(w)\big) \llbracket \widehat{u} \rrbracket \Big)_e + \sum_{e\in\Gamma_\mathcal{D}} \Big(\big(\alpha w/h + \Phi(w)\big) \big(\widehat{u} - g_\mathcal{D}\big) \Big)_e = 0 \ .$$

Equation (B.1) now follows straightforwardly from the conditions (B.2).

We next establish that $\widehat{u}$ is exact on Neumann edges $\Gamma_\mathcal{N}$. Let $e_\mathcal{N}$ denote the Neumann edge and $e_\mathcal{D}$ the complementary Dirichlet edge. Further, let $\varphi_\mathcal{N} \in H^1_{0,\mathcal{D}}(\Omega)$ be the linear function which is $|\Omega|$ at $e_\mathcal{N}$ and which vanishes at $e_\mathcal{D}$. Using $\varphi_\mathcal{N}$ in (2.1), we obtain for the exact solution

$$u(e_\mathcal{N}) = \int_\Omega f\varphi_\mathcal{N} \,\mathrm{d}x + g_\mathcal{D}(e_\mathcal{D}) + |\Omega| g_\mathcal{N}(e_\mathcal{N}) \ . \tag{B.3}$$

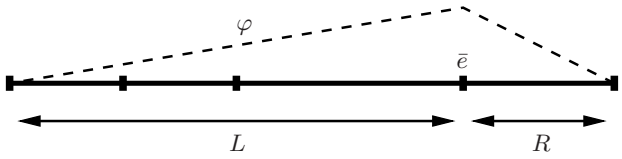


FIG. 7. *Global Green's function φ with respect to edge $\bar{e}$.*

Moreover, as $\varphi_{\mathcal{N}} \in \mathbb{P}^1(\mathcal{P}^{\mathsf{h}}) \subset \widehat{H}$, it holds that $\mathscr{B}_{\Phi}(\widehat{u}, \varphi_{\mathcal{N}}) = \mathscr{L}_{\Phi}(\varphi_{\mathcal{N}})$. On account of $[\![\varphi_{\mathcal{N}}]\!]_e = [\![\widehat{u}]\!]_e = 0$ for $e \in \Gamma_{\mathcal{I}}$, $\varphi_{\mathcal{N}}(e_{\mathcal{D}}) = 0$, and $\widehat{u}(e_{\mathcal{D}}) = g_{\mathcal{D}}(e_{\mathcal{D}})$, this implies

$$\sum_{K \in \mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}\widehat{u}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{\mathcal{N}}}{\mathrm{d}x} \, \mathrm{d}x = \int_\Omega f \varphi_{\mathcal{N}} \, \mathrm{d}x + |\Omega| g_{\mathcal{N}}(e_{\mathcal{N}}) \, .$$

The left side evaluates to $\widehat{u}(e_{\mathcal{N}}) - \widehat{u}(e_{\mathcal{D}})$, which is identical to $\widehat{u}(e_{\mathcal{N}}) - g_{\mathcal{D}}(e_{\mathcal{D}})$ by virtue of the previously established coincidence of $\widehat{u}(e_{\mathcal{D}})$ and $g_{\mathcal{D}}(e_{\mathcal{D}})$. We then conclude from (B.3) that $\widehat{u}(e_{\mathcal{N}}) = u(e_{\mathcal{N}})$.

Finally, we establish that $\widehat{u}$ is exact on interior edges $\Gamma_{\mathcal{I}}$. We consider an arbitrary edge $\bar{e} \in \Gamma_{\mathcal{I}}$ and define $L = L(\bar{e})$ and $R = R(\bar{e})$ to be the open subsets of Ω left and right of edge $\bar{e}$; see Figure 7. Furthermore, we define $\varphi = \varphi(\bar{e}) \in H^1_{0,\mathcal{D}}(\Omega)$ to be the global Green's function corresponding to $\bar{e}$, viz., a hat function for which the jump in the derivative at $\bar{e}$ equals -1. Inserting φ in (2.1), we obtain the following relation for the exact solution at edge $\bar{e}$:

$$u(\bar{e}) = \int_\Omega f\varphi \, \mathrm{d}x + \sum_{e \in \Gamma_{\mathcal{D}}} \vartheta(e) g_{\mathcal{D}}(e) + \sum_{e \in \Gamma_{\mathcal{N}}} \vartheta(e) u(e) \, , \tag{B.4}$$

where $\vartheta(e) := |R|/|\Omega|$ if e is a left edge, and $\vartheta(e) := |L|/|\Omega|$ if e is a right edge. Moreover, the identity $\mathscr{B}_{\Phi}(\widehat{u}, \varphi) = \mathscr{L}_{\Phi}(\varphi)$ yields

$$\sum_{K \in \mathcal{P}^{\mathsf{h}}} \int_K \frac{\mathrm{d}\widehat{u}}{\mathrm{d}x} \frac{\mathrm{d}\varphi}{\mathrm{d}x} \, \mathrm{d}x = \int_\Omega f\varphi \, \mathrm{d}x \, .$$

The left side evaluates to $\widehat{u}(e) - \sum_{e \in \partial\Omega} \vartheta(e) \widehat{u}(e)$. As $\widehat{u}$ is exact on the boundary edges, we finally conclude from (B.4) that $\widehat{u}(e) = u(e)$.

Acknowledgment. The authors thank Dr. Marc Gerritsma for his careful review of the manuscript and the many helpful comments.

REFERENCES

[1] M. AINSWORTH AND J. T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math., Wiley-Interscience, New York, 2000.

[2] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] I. BABUŠKA, C. E. BAUMANN, AND J. T. ODEN, *A discontinuous hp finite element method for diffusion problems: 1-d analysis*, Comput. Math. Appli., 37 (1999), pp. 103–122.

[4] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[5] F. BREZZI, G. MANZINI, D. MARINI, P. PIETRA, AND A. RUSSO, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 365–378.

[6] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[7] G. Cockburn, E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods. Theory, Computations and Applications, G. Cockburn, E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, New York, 2000, pp. 3–50.

[8] P. Houston and E. Süli, *hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems*, SIAM J. Sci. Comput., 23 (2001), pp. 1226–1252.

[9] M. G. Larson and A. J. Niklasson, *Analysis of a family of discontinuous galerkin methods for elliptic problems: The one dimensional case*, Numer. Math., 99 (2004), pp. 113–130.

[10] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.

[11] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, Pure Appl. Math., John Wiley & Sons, New York, 1974.

[12] S. Prudhomme, F. Pascal, J. T. Oden, and A. Romkes, *Review of A Priori Error Estimation for Discontinuous Galerkin Methods*, Tech. report 00-27, Texas Institute for Computational and Applied Mathematics (TICAM), University of Texas, Austin, TX, 2000.

[13] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.

[14] A. Romkes, S. Prudhomme, and J. T. Oden, *A priori error analyses of a stabilized discontinuous Galerkin method*, Comput. Math. Appl., 46 (2003), pp. 1289–1311.

[15] A. Romkes, S. Prudhomme, and J. T. Oden, *Convergence analysis of a discontinuous finite element formulation based on second order derivatives*, Comput. Methods Appl. Mech. Engrg. 195, (2006), pp. 3461–3482.

[16] C. Schwab, *p- and hp-Finite Element Methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 1998.